UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

A multiscale approach for exploring bacterial transcriptional systems

CAUÃ ANTUNES WESTMANN

Ribeirão Preto

Brasil

2018

CAUÃ ANTUNES WESTMANN

A multiscale approach for exploring bacterial transcriptional systems

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Mestre em Ciências - Área de concentração: Biologia Celular e Molecular

Orientador: Prof. Dr. Rafael Silva Rocha

Ribeirão Preto

Brasil

2018

WESTMANN, Cauã Antunes

A multiscale approach for exploring bacterial transcriptional systems. Ribeirão Preto, 2018.

185 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Biologia Celular e Molecular.

Orientador: Prof. Dr. Rafael Silva-Rocha.

1. Regulação Transcricional. 2. Microbiologia. 3. Biologia Sintética. 4. Biologia de Sistemas. 5. Evolução de Sistemas Moleculares. 6. Metabolismo microbiano

Aluno: **Westmann, Cauã Antunes**

Título: **A multiscale approach for exploring bacterial transcriptional systems**

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Mestre em Ciências. Área de concentração: Biologia Celular e Molecular. Orientador: Prof. Dr. Rafael Silva Rocha

Aprovado em: ___/___/___

Banca Examinadora

Prof(a). Dr(a).:_____

Instituição:_____

Assinatura:_____

Prof(a). Dr(a).:_____

Instituição:_____

Assinatura:_____

Prof(a). Dr(a).:_____

Instituição:_____

Assinatura:_____

ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

Primeiramente, **fora Temer**! E que um dia possamos romper com tudo aquilo que anunciava Caetano em suas canções - "Será que nunca faremos senão confirmar, a incompetência da América católica, que sempre precisará de ridículos tiranos? ".

Agora, os **verdadeiros** agradecimentos...

Agradeço às raízes mais profundas que me nutriram nesse mundo com seu amor e dedicação inesgotáveis para que, de galho em galho, eu pudesse crescer, florescer, frutificar e continuar essa emocionante jornada que é a vida. Dedico esta dissertação aos meus pais, **Ana** e **Flavio**, que sempre se fizeram presentes em cada momento, ensinando-me que é somente sendo o todo que nos sentimos parte e que somente nos sentindo parte somos o todo. Obrigado por me mostrarem que é colecionando as pedras no caminho que conseguimos construir, juntos, os mais bonitos castelos. Obrigado por sempre acreditarem. Amo vocês imensamente.

Agradeço aos **meus avós**, tão presentes em minha vida e que marcam minhas memórias (e presente) com as mais bonitas recordações. Em especial, agradeço ao meu avô, **José Antunes**, por me mostrar que há sempre um sorriso escondido em cada canto da vida, apenas esperando para ser encontrado por nós.

Agradeço aos meus orientadores, **Dr. Rafael Silva-Rocha** e **Dra. Maria Eugenia Guazzaroni**, por não serem apenas orientadores, mas sim amigos, verdadeiros líderes e inspirações nesse mundo acadêmico (e fora dele também) que às vezes se mostra tão obscuro. Fico muito feliz em ver que há sim luzes no fim do túnel e espero, um dia, poder juntar-me a elas. Agradeço por todos os ensinamentos e oportunidades, por terem permitido que eu me desenvolvesse tanto e chegasse há lugares que jamais imaginei que conseguiria. A verdade é que só podemos chegar assim tão longe quando estamos apoiados nos ombros de gigantes e fico imensamente grato por esta oportunidade tão única.

Agradeço aos tantos **professores** desde o Jardim 0 até a faculdade que me ensinaram que os verdadeiros prazeres estão nas perguntas e não nas respostas, que me instigaram a pensar criticamente e me mostraram que as portas do conhecimento estão todas ali, somente esperando para serem abertas. Destaco aqui **Therezinha Serzedello**, **Fábio Kassardjian**, **Marcos Fázio** e **Sonale de Oliveira**.

Agradeço aos amigos de longa data, principalmente **Eduardo Siqueira** e **Carlos Neto** que, apesar das muitas veredas da vida, tanto me inspiraram e me motivaram a procurar o melhor de mim.

Agradeço ao **SynBio**, o núcleo no qual conheci não somente uma nova disciplina, mas também pessoas muito especiais que me mostraram como o trabalho em equipe e a dedicação podem ser recompensadores. Agradeço pelos grandes amigos que ali se consolidaram para toda uma vida: **Joana Guiro**, grande companheira de aventuras e que, talvez nem saiba, mas tanto inspirou meus primeiros passos por terras estrangeiras; **Otto Heringer**, um eterno jovem inspirador nesse mundo de SynBio e empreendedorismos; **Pedro Medeiros**, com seu bom humor e acidez pragmática que sempre aponta para as muitas máscaras deste mundo; **Macarena Lopez** que me ensinou a ver a vida de um modo muito mais leve e **Cleandho Souza** que me mostrou que com um bom coração e determinação não há nada que não esteja a nosso alcance. Agradeço ao **Dr. Andres Ochoa** que me introduziu ao mundo da Biologia Molecular/Sintética e me ensinou a dar os meus primeiros passos acadêmicos (sempre com "gentis" empurrões colombianos). Agradeço ao **Dr. Marcelo Boareto** e ao **Dr. João Vitor Molino**, cientistas brilhantes que me ensinaram na sua quietude que a dedicação e o esforço valem infinitamente mais que vãs palavras de marketing científico e egos inflados.

Agradeço aos **amigos do Lison**, pessoas tão especiais que me permitiram ver com clareza que, como diria McCandless, a felicidade só é real quando **compartilhada** (e foi na ausência, no mais frio dos invernos ingleses que a descobri). Descobri com vocês que, antes de se fazer ciência, ou qualquer outra coisa nesta vida tão cheia de rótulos e obrigações, fazemos amigos e são eles que realmente importam. Agradeço ao grande amigo **Bruno Gomes**, por mostrar que cada segunda-feira pode ter uma risada diferente e que a humildade e integridade são virtudes essenciais tanto dentro quanto fora dos meios acadêmicos. Agradeço ao amigo **Gabriel Lovate** (G2) por nossos intensivões científicos no Lab durante os fins de semana, pelo exemplo de dedicação/motivação e pelas divagações sobre a vida, o universo e tudo mais. Agradeço ao **Tiago Borelli** pela amizade e por poder aprender com ele a importância da dedicação no nosso crescimento pessoal. Fico imensamente feliz por tê-lo acompanhado, ainda que brevemente, em sua jornada e vê-lo crescer tanto. Agradeço à **Luísa Nora** pela amizade e por me inspirar com sua alegria diária e dedicação na qual tudo é realmente feito e falado com o coração. Agradeço ao **Murilo Cassiano** pelos Kantarokês, pelas fritações científicas e por nunca me deixar esquecer a importância do brilho nos olhos ao se fazer ciência. Desde um lápis e papel,

até um simples gel de agarose, tudo pode se tornar uma ferramenta poderosa nas mãos daqueles que amam o que fazem. Agradeço à **Ananda Medeiros** pela amizade e companheirismo, pelos Kantarokês compartilhados, ensinamentos organizacionais e por me relembrar que, muitas vezes, o lado bom da vida pode estar literalmente ao lado. Agradeço à **Lummy Monteiro**, **Luana Alves**, **Leonardo Martins** e **Letícia Arruda** (a turma dos bioquímicos de coração) pelos bons momentos no Lab, risadas e ensinamentos. Agradeço à **Juliana Rojas** pela amizade, Marcões/Silvões, pseudo-aulas de Espanhol e por me acompanhar (e salvar) nas atrapalhadas aventuras desta vida dentro e fora do Lab. Agradeço aos novos companheiros de Lab, **Kauan Ribeiro**, **Guilherme de Siqueira**, **Maria Clara**, **Adriano Gomes**, **Lucas Ribeiro** e **Greicy Pereira** pela amizade e momentos descontraídos com muito bolo e café. Agradeço à **Claudinha** por seu esforço diário em manter o Lab organizado e por facilitar tanto nossas vidas com seu trabalho!

Agradeço às moradoras da Casa do Feijão, **Fany Maria** e **Marcella** pela amizade e por trazerem mais amor ao dia-a-dia dessa complexa vida de pós-graduação. Sofremos juntos, mas nos divertimos muito também.

Agradeço ao **Departamento de Biologia Celular e Molecular da FMRP-USP** pelo apoio acadêmico e formação científica.

Agradeço a cada **funcionário** do departamento e fora dele que permitem que tudo o que fazemos no nosso dia-a-dia seja possível. Agradeço à **Gabriela Zamoner** pelo apoio em todas as minhas (muitas) dúvidas durante a pós-graduação.

Agradeço à **CAPES** e à **FAEPA** pelo essencial auxílio financeiro no Brasil.

Agradeço à **Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)** pelo auxílio financeiro no Brasil e no exterior referentes aos processos **2016/05472-6** e **2017/16783-5**, respectivamente.

Agradeço à **sociedade brasileira** que permite, através do seu trabalho diário e pagamento de muitos impostos, o financiamento de nossos sistemas públicos e de nossa ciência. Sem vocês nada disso seria possível e espero que possa retribuir tais esforços ao longo da minha carreira.

Agradeço ao professor **Dr. Orkun Soyer** pela oportunidade de trabalhar em seu laboratório e pelos valiosos insights que pude obter tanto dentro quanto fora do OSS Lab. Agradeço também

ao **Dr. Kalesh Sasidharan** pela orientação, amizade e ensinamentos durante meu tempo em Warwick

E, finalmente, meus sinceros agradecimentos a **todos** aqueles que contribuíram direta ou indiretamente para que este trabalho fosse realizado e eu chegasse até aqui.

“Each person who ever was or is or will be has a song. It isn't a song that anybody else wrote. It has its own melody, it has its own words. Very few people get to sing their song. Most of us fear that we cannot do it justice with our voices, or that our words are too foolish or too honest, or too odd. So people live their song instead.”

— Neil Gaiman, *Anansi Boys*

“I love science, and it pains me to think that so many are terrified of the subject or feel that choosing science means you cannot also choose compassion, or the arts, or be awed by nature. Science is not meant to cure us of mystery, but to reinvent and reinvigorate it.”

— Robert M. Sapolsky, *Why Zebras Don’t Get Ulcers?*

GENERAL INDEX

GENERAL INDEX

INDEX OF FIGURES

INDEX OF TABLES

ABBREVIATIONS

CRP	cAMP receptor protein
°C	Celsius degrees
A.U	Arbitrary Units
bp	base pairs
Cm^R	Chloramphenicol resistance gene
CREs	*Cis*-regulatory elements
DNA	Deoxyribonucleic acid
ESB	Evolutionary Systems Biology
EX	Exchange reaction
FBA	Flux Balance Analysis
Fis	Factor for inversion stimulation
GEMs	Genome Scale Models
GNPs	growth-condition specific nucleoid proteins
GPR	Gene-Protein-Reaction association
GRN	Gene Regulatory Networks
H-NS	Histone-like nucleoid-structuring protein
IHF	Integration host factorTFs
KOs	KEGG Orthologs
LB	Luria-Bertani medium
LVA	LVA protein degradation tag
M9	Minimal medium
mL	millilitre
NAPs	nucleoid-associated proteins
ng	Nanogram
nt	nucleotide
OD_{600}	Optical density at 600nm
PCR	Polymerase Chain Reaction
PWM	Position Weight Matrix
RNA	Ribonucleic acid
RNAP	RNA Polymerase
TF	Transcription Factor

TFBSs	Transcription Factor Binding Sites
U	Enzyme Unit
UNPs	universal nucleoid proteins
URN	Universal Reaction Network
UTR	Untranslated region
WT	wild type
µL	Microliters

RESUMO

RESUMO

WESTMANN, Cauã Antunes. A multiscale approach for exploring bacterial transcriptional systems. [dissertação]. Ribeirão Preto: Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto; 2018.

A vida é um fenômeno intrinsicamente complexo e, para melhor compreender seus princípios fundamentais, devemos ser capazes de investigar todas as camadas organizacionais que a compõem (desde as -ômicas até as populacionais e ecológicas). O estudo de como a informação molecular extra e intracelular flui através dessas camadas bem como estas se interconectam na geração de respostas fenotípicas são essenciais para uma compreensão mais profunda e para (re)engenharia de sistemas biológicos. Além disso, a combinação de diferentes abordagens (*in vivo* e *in silico*) para dissecção destas redes complexas nos permite alcançar uma visão mais holística e preditiva destes sistemas. Nesse contexto, a presente dissertação foca na exploração da camada regulatória transcricional em bactérias, um dos sistemas mais basais na regulação gênica e na integração de estímulos ambientais. Ao combinar uma gama de abordagens diferentes, porém complementares, tais como Biologia Sintética, Biologia de Sistemas Evolutiva e Metagenômica, nós observamos sistema através de diferentes perspectivas para uma compreensão mais geral de seus fundamentos. Adotamos a abordagem da Biologia Sintética para explorar como a arquitetura combinatória de promotores complexos pode originar diferentes lógicas transcricionais e fenômenos emergentes, combinando sítios específicos de ligação de fatores de transcrição (TFBSs) - para os fatores de transcrição (TFs) globais de *E. coli* Fis e IHF. Nossos resultados mostraram que não apenas fenômenos emergentes podem ser observados em promotores sintéticos, mas também respostas específicas que se assemelham à dinâmica de cada um dos componentes. Em seguida, nos concentramos na aplicação Biologia de Sistemas Evolutiva para compreender como a inovação evolutiva poderia surgir em elementos *cis*-regulatórios e quais seriam os principais processos que restringem a diversidade destes. Nossos resultados computacionais baseados em conjuntos de dados de TFBSs para três reguladores globais em *E. coli* - CRP, Fis e IHF - apontaram que o *crosstalk* transcricional (o compartilhamento de TFBSs por diferentes TFs) é não somente muito comum nesses sistemas, mas também um elemento chave em relação à evolução de lógica regulatória e restrição da diversidade de TFBS em bactérias. Por fim, adotamos uma

abordagem Metagenômica para expandir nossa compreensão dos elementos *cis*-regulatórios além de *E. coli*, avaliando e caracterizando a diversidade de promotores constitutivos em amostras ambientais. Esses resultados forneceram dados qualitativos e quantitativos sobre o espaço de sequências naturais de promotores constitutivos em bibliotecas metagenômicas. No capítulo final desta dissertação, investigamos redes metabólicas bacterianas, a camada mais basal de organização molecular em sistemas vivos, que se encontra profundamente entrelaçada com redes transcricionais. Assim, desenvolvemos uma nova série de algoritmos para geração automática de modelos metabólicos estequiométricos a partir de dados (meta)genômicos, que podem, no futuro, ser prontamente integrados com dados transcricionais para a geração de modelos *in silico* de células únicas. Em resumo, o trabalho atual forneceu novas informações sobre muitos aspectos dos sistemas de transcrição em bactérias que, dada uma base teórica adequada, podem ser extrapoladas para sistemas mais complexos, como eucarióticos. Acreditamos, assim, que essa abordagem multi-escala é fundamental tanto para compreensão dos princípios gerais que permeiam o processamento de informações em sistemas vivos quanto para (re)estruturá-los em aplicações biotecnológicas.

Palavras-chave: 1. Informação molecular; 2. Transcrição em bactérias; 3. Biologia Sintética; 4. Biologia de Sistemas Evolutiva; 5. Metagenômica

ABSTRACT

ABSTRACT

WESTMANN, Cauã Antunes. A multiscale approach for exploring bacterial transcriptional systems. [dissertação]. Ribeirão Preto: Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto; 2018.

Life is a complex phenomenon and in order to understand its underlying principles, we must be able to investigate every organizational layer that comprises it (from –omics to ecological ones). Exploring how the molecular information flows from both extra- and intracellular worlds through these layers and how they interact in the generation of phenotypic responses shall provide a more consistent background for both understanding and (re)engineering living systems. Besides, combining different approaches (*in vivo* and *in silico*) for dissecting these complex networks should allow us to achieve a more holistic and predictive view of biological phenomena. In this context, the present dissertation focus on exploring the transcriptional regulatory layer of bacteria, one of the most basal systems in gene regulation and in the integration of environmental stimuli. By merging a range of different yet complementary frameworks such as Synthetic Biology, Evolutionary Systems Biology and Metagenomics we have delved into the different aspects of this system for a more general understanding of its foundations. We have adopted the Synthetic Biology approach to explore how transcriptional logic and emergent phenomena might arise from the combinatorial architecture of complex promoters regarding the combination of specific Transcription Factor Binding Sites (TFBSs) - for the *E. coli* global transcription factors (TFs) Fis and IHF. Our results have shown that not only emergent phenomena might be observed in synthetic promoters, but also specific responses that resemble the dynamics of each of the individual components. Next, we have focused on applying the Evolutionary Systems framework to understand how evolutionary innovation might rise in *cis*-regulatory elements and what would be the main processes constraining their diversity. Our computational results based on datasets of TFBSs for three global regulators in *E. coli* - CRP, Fis and IHF - have pointed that transcriptional crosstalk (the sharing of TFBSs by different TFs) is ubiquitous in these systems and a key element regarding the evolution of regulatory logic and the constraining of TFBS diversity in bacteria. Lastly, we have adopted a Metagenomics approach to expand our understanding of transcriptional *cis*-elements beyond *E. coli*, by assessing and characterizing the diversity of constitutive promoters

in environmental samples. These results have provided both qualitative and quantitative data regarding the natural sequence space of constitutive promoters in metagenomic libraries. In the final chapter of this dissertation, we have investigated bacterial metabolic networks, the most basal layer of molecular organization in living systems, which is deeply intertwined with transcriptional networks. Thus, we have developed a novel series of algorithms for automatic generation of stoichiometric metabolic models from (meta)genomic data, which can, in the future, be readily integrated with transcriptional data for the generation of *in silico* whole-cell models. Altogether, the current work has provided resourceful information regarding many aspects of transcriptional systems in bacteria which, provided the adequate theoretical framework, can be extrapolated to more complex systems such as eukaryotes. We believe this multiscale approach is fundamental for both understanding the general principles underpinning information processing in living systems and (re)engineering them for biotechnological applications.

Keywords: 1. Molecular Information; 2. Bacterial transcriptional systems; 3. Synthetic Biology; 4. Evolutionary Systems Biology; 5. Metagenomics

I. INTRODUCTION

Part of this introduction was published or submitted as:

Westmann, C. A., Guazzaroni, M. E., & Silva-Rocha, R. (2018). Engineering Complexity in Bacterial Regulatory Circuits for Biotechnological Applications. *mSystems*, *3*(2), e00151-17.

Westmann, C. A., de Alves, L. F., Borelli, T. C., Silva-Rocha, R., & Guazzaroni, M. E. (2017). Transcriptional Regulation of Hydrocarbon Efflux Pump Expression in Bacteria. *Cellular Ecophysiology of Microbe*, 1-23.

de Alves, L.F., Westmann, C. A., Lovate, G.L., de Siqueira, G. M. V., Borelli, T.C. & Guazzaroni, M. E. (2018) Metagenomic approaches for understanding new concepts in microbial science – *Submitted in International Journal of Genomics*

1. Biological networks, structure, relevance and the organization of regulatory systems

According to the autopoietic theory of Maturana & Varella (Varela, 1980), an essential characteristic of living beings is their ability to undergo continuous structural changes, while retaining their network-based pattern of organization. As epiphenomena of this adaptive dynamicity of organisms to variations in the outer and intracellular environments, complex, dynamic, plastic and resilient systems arise, capable of persisting, changing and multiplying over time, within the context of Darwinian evolution (Wagner, 2005; Whitacre, 2010).

There are a myriad of mechanisms behind this adaptive dynamicity that can be analysed under different organizational perspectives of the system they are embedded into, which extend from the molecular to the macroscopic/phenotypic levels (Schuster, 2002; Ibáñez-Marcelo and Alarcón, 2014). However, despite the variations among its members, all these systems converge to a common architecture that can be represented in the form of graphs: scale-free biological networks (Albert, 2005). In these networks the distribution of connectivity is not homogeneous, but follows a power-law: there are a few highly connected nodes (hubs) and the rest has low connectivity, in contrast with random networks where connectivity follows a Poisson distribution. The consequence of this mode of organization is the formation of a redundant adaptive system of high robustness, yet substantially fragile - with few deeply interconnected nodes of high biological relevance (Csete and Doyle, 2002; Barabasi and Oltvai, 2004; Kitano, 2004; Albert, 2005)

Within the set of interconnected networks – each network representing an organizational layer of the observed system - that regulate the dynamicity of organisms, molecular-level systems are considered the most basal ones in terms of information processing modules (Farnsworth, Nelson and Gershenson, 2013) - reception and response to stimuli -, especially those in the form of gene expression controlling networks. These regulatory networks can be traditionally subdivided into transcriptional, translational and post-translational, according to the targets of their components, mechanisms of action and sometimes the context of the gene expression process under which they may act. In this context, following the “central dogma” of Molecular Biology, in which the flow of all cellular information begins in DNA, the importance of transcriptional regulation - which directly acts on this nucleic acid and other peripheral

components - is highlighted as one of the initial controlling points for modulating the entire information flow inside the cell (Tkačik and Bialek, 2014; Ledezma-Tejeida, Ishida and Collado-Vides, 2017).

2. Transcriptional Regulation and General Roles of Transcription Factors in bacteria

Evolution has shaped a very compact genomic arrangement in bacteria, in which only about 10% does not correspond to the production of RNAs (coding or non-coding) (Koonin and Wolf, 2008; Koonin, 2009). Remarkably, this relatively small portion of the genome is responsible for most of the gene regulation events that occur in response to the most diverse environmental conditions (Cases, De Lorenzo and Ouzounis, 2003; Cases and de Lorenzo, 2005; Browning and Busby, 2016; Westmann *et al.*, 2018). There are many different mechanisms responsible for these events, however they all converge to regulate the RNA polymerase (RNAP) distribution rates through the transcriptional units of bacterial DNA (Akira, 2000; Bintu, Buchler, Garcia, Gerland, Hwa, Kondev and Phillips, 2005; Bintu, Buchler, Garcia, Gerland, Hwa, Kondev, Kuhlman, *et al.*, 2005; Ishihama, 2009). Such rates are dependent on the affinity of RNAP for the promoter region of the gene and this affinity is, ultimately, subject to modulations by a wide range of mechanisms, such as: association of different sigma (σ) factors; variations in DNA consensus sequences for polymerase and transcription factors binding; folding/looping of the DNA in the promoter region due to thermodynamic constraints; transcription factors that bind to the promoter by activating or repressing gene transcription through interactions with the polymerase or the DNA strand itself (Browning, Grainger and Busby, 2010; Browning and Busby, 2016). Two groups of regulatory proteins are involved in the modulation of the RNAP gene selectivity in *E. coli*: seven types of σ factors and approximately 300 types of transcription factors (Pérez-Rueda, Collado-Vides and Perez-Rueda, 2000; Ishihama, 2010, 2017).

While the σ factors constitute the first stage of gene regulation (Gruber and Gross, 2003; Sharma and Chatterji, 2010; Lee, Minchin and Busby, 2012), transcription factors are the major modulators in the later stages. They are proteins sensitive to internal and external environmental variations, with one or more domains - usually one that interacts with the nucleic acid and another which functions as a sensor of the environmental stimuli -, capable of acting

individually or together on: the DNA molecule itself; adjacent regulators or on RNAP, modulating its affinity to the promoter region (Pabo and Sauer, 1992; Browning and Busby, 2004, 2016). Transcription factors are functionally classified into activators – the ones that promote gene expression and might generally bind in a motif upstream the core promoter region - or repressors - the ones that decrease gene expression and might generally bind in a motif upstream the core promoter region. However, over the past decades, it was discovered that activators and repressors are not immutable functions, they are interchangeable and depend on many factors such as the architecture of the regulatory region (e.g. density and relative position of TFBS), the environmental context etc. (Akira, 2000; Ishihama, 2009; Monteiro, Arruda and Silva-Rocha, 2017).

In addition to their general function, transcription factors can also be classified according to their topology in the regulatory network, as represented in **Figure 1** (Martínez-Antonio *et al.*, 2003). Global regulators are the top nodes of the network, few deeply interconnected nodes with a low degree of edges going into these nodes (not regulated by many genes) and a high degree of edges going to other nodes (they regulate many genes). Hence, they are the regulators with the highest percentage of regulated transcriptional units in the network and usually related to global responses in the cell such as growth phase transitions, catabolite repression etc. Local regulators, on the other hand, are very heterogeneous (many different families) and are related to more specific and adaptive responses. Of the approximately 300 transcription factors described in *E. coli*, it is estimated that only 7 are global (CRP, FNR, IHF, Fis, ArcA, NarL and Lrp), modulating the expression of more than 50% of regulated genes (Martínez-Antonio *et al.*, 2003).

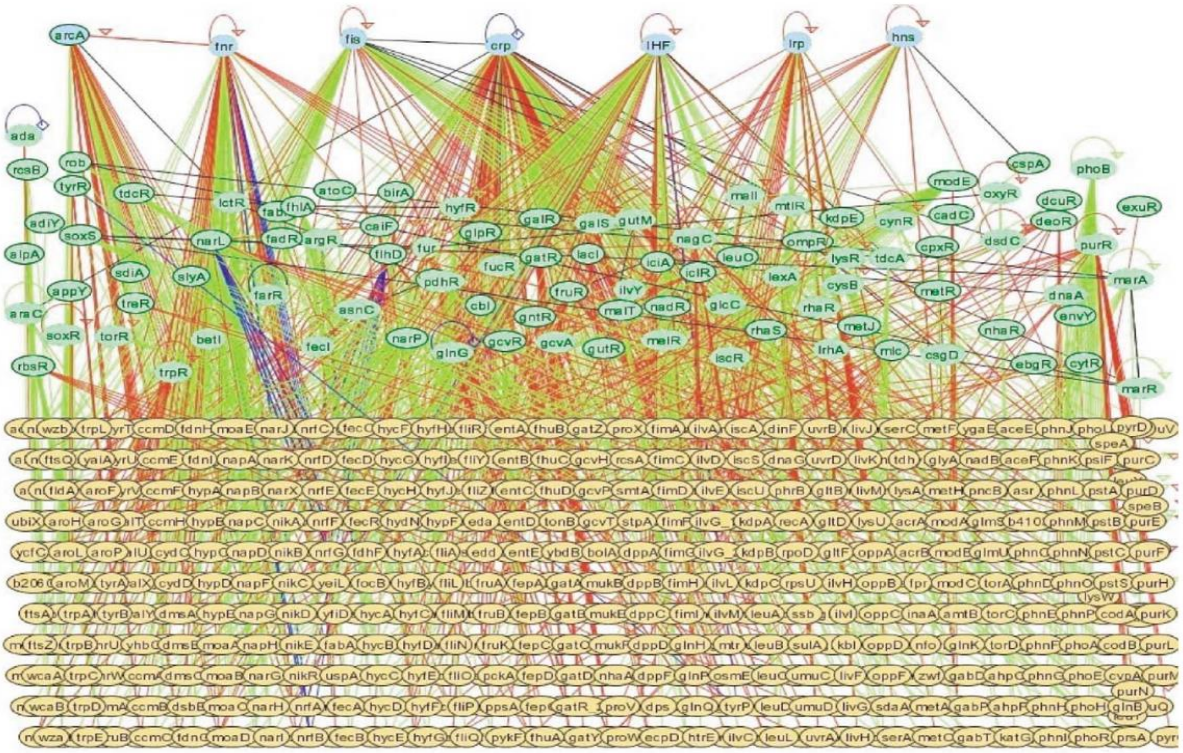


Figure 1. Overview of the regulatory network in *E. coli*. From top to bottom of the image: (i) global blue regulators (ArcA, FNR, Fis, CRP, IHF, Lrp and HNS); (ii) local regulators in green and (iii) yellow modulated genes. Green lines represent activation, red repression and double-blue blending. Retrieved from (Martínez-Antonio *et al.*, 2003).

Within this set of global factors, new subsets can be created between free transcription factors and factors associated with the nucleoid. The latter, in turn, can be subdivided into universal nucleoid proteins (UNPs), which always remain bound to the nucleoid and growth-condition specific nucleoid proteins (GNPs), which appear only at specific stages of growth (Ishihama, 1999, 2017; Dillon and Dorman, 2010). In this context, IHF is a heterodimer UNP that plays both structural role - DNA super-folding and destabilization of DNA duplexes - and regulator of genomic functions - DNA replication, recombination and gene expression (Ishihama, 2009). The Fis protein (inversion stimulation factor), on the other hand, is a GNP associated with cell growth (Ball *et al.*, 1992; Azam *et al.*, 1999; Nowak-Lovato *et al.*, 2013). In optimal growth conditions, Fis is dominant in the nucleoid and regulates the transcription of several genes related to growth and bacterial metabolism - 21% of the genes are modulated by Fis directly or indirectly (Cho *et al.*, 2008).

3. Integration of signals into complex promoters

The activity of most bacterial promoters is dependent on various environmental stimuli. Thus, many promoters are controlled by two or more transcription factors, with each of them conveyed to a specific stimulus (Rydenfelt *et al.*, 2014; Browning and Busby, 2016). Databases indicate that approximately 50% of the *E. coli* promoters are under the control of a single specific regulator (M. Madan Babu, 2013; Rydenfelt *et al.*, 2014; Gama-Castro *et al.*, 2016), whereas the other 50% of the genes are regulated by more than two factors. This complex regulation depends on combinations between repressors and activators or co-dependence of more than one activator (Barnard, Wolfe and Busby, 2004; Hermsen, Tans and Ten Wolde, 2006; Browning and Busby, 2016). For example, at the same time nucleotide-associated factors - Fis, IHF and H-NS - work together to repress transcription in the Nir regulatory region, these are also recruited as transcriptional activators in other promoters (Browning, Cole and Busby, 2000; McLeod and Johnson, 2001). Thus, the analysis of the DNA context in which the regulator is embedded becomes much more relevant to determine its function than an absolute classification.

Within this scenario, regulatory interactions in *E. coli* become much more elaborate than previously thought and probably as complex as those found in some eukaryotes, involving multifactor promoters and multi-target regulators, together, forming hierarchical regulatory networks. As an example of regulatory complexity in *E. coli*, more than 10 factors were characterized as directly involved in regulating the promoter for *csgD* encoding the master regulator of biofilm formation in this organism (**Figure 2**) (Gerstel, Park and Römling, 2003; Ogasawara *et al.*, 2010).

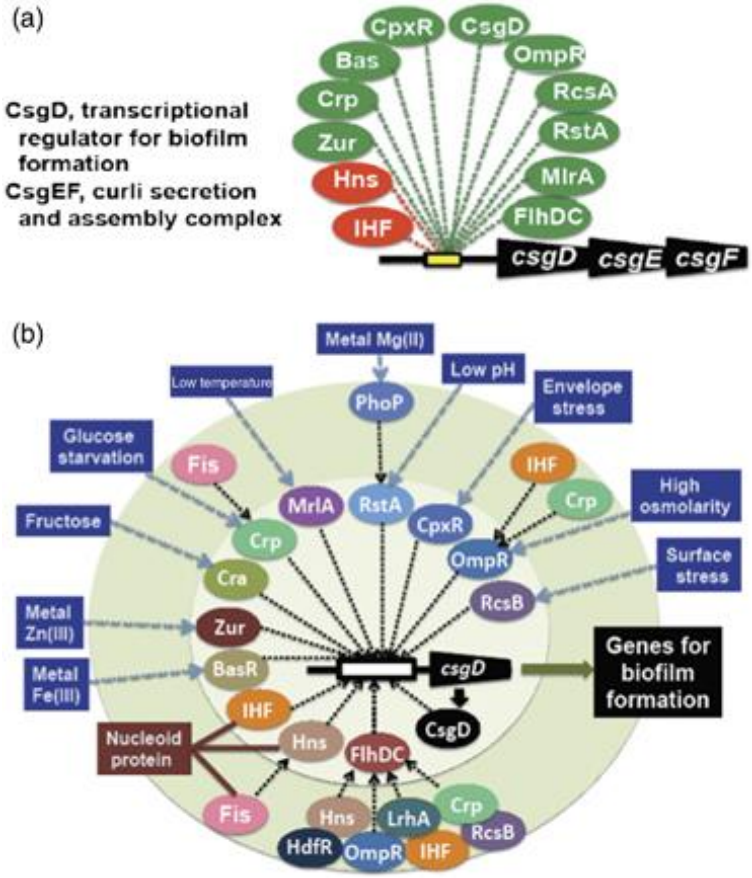


Figure 2. Transcription factors and environmental stimuli involved in regulation of *csgD*. (a) A number of primary transcription factors, indicated in the inner circle, participate directly in the regulation of the *csgD* promoter. **(b)** The genes coding for these transcription factors are organized under the control of secondary transcription factors. The primary and secondary transcription factors together monitor various environmental signals and stresses. Retrieved from (Ishihama, 2010).

4. A multiscale approach for understanding molecular systems

"What I cannot create I do not understand". Many scientists have adopted this argument shaped by Richard Feynman as the reasoning for exploring biological systems through Synthetic Biology. However, life is a complex system continually shaped by evolutionary processes and the gap between designing an organism and fully understanding its complex nature is enormous. Thus, in order to bridge this gap and establish a more holistic framework, there is an urgent need for merging multiple perspectives. Building with "biological parts" to understand through Synthetic Biology and understanding to build through Systems Biology are complementary approaches that have become extremely influential. However, as Theodosius Dobzhansky once wrote: "nothing in Biology makes sense except in the light of Evolution". In this context, one of the goals of the current dissertation is to expand the current views on bacterial transcriptional systems by combining a wide range of experimental strategies and theoretical backgrounds.

These different approaches can act in a synergistic manner for dissecting the basic principles underlying the organization patterns and evolution of living systems (Bayer, 2010; Wagner *et al.*, 2012; Soyer and O'Malley, 2013; Crocker and Ilsley, 2017; Schaerli *et al.*, 2017).

In this context, as explained before, the multi-layered process of decision-making in bacteria (such as biofilm formation, motility, differential expression of catabolic genes etc.) has transcriptional regulation in its core (Browning and Busby, 2004, 2016; Silva-Rocha, Tamames and de Lorenzo, 2012). In bacteria, most of the environmental and intracellular information is embodied inside of the cell by the conformational changes of signal-specific Transcription Factors (TFs) – in one-, two- or three-component systems – and ultimately integrated into short cis-regulatory regions (a few hundred of base pairs) which can bear multiple transcription factor binding sites (TFBSs) in an overlapped or individual architecture. These regions might be analogously compared to electronic microprocessors in the sense of information integration, ultimately comprising the set of logic rules for the expression of a gene or operon (Silva-Rocha, Tamames and de Lorenzo, 2012; Huminiecki and Horbańczuk, 2017; Bashor and Collins, 2018; Westmann, Guazzaroni and Silva-Rocha, 2018).

Here, we wanted to explore bacterial transcriptional systems under different, yet complementary perspectives such as Synthetic Biology, Evolutionary Systems Biology and Metagenomics (**Figure 3**). Each of these fields has its own framework and methods and is focused on specific biological questions. Synthetic Biology is usually based on the deconstruction/re-engineering of natural systems in order to understand its underlying principles (Andrianantoandro *et al.*, 2006; Cameron, Bashor and Collins, 2014). The Evolutionary Systems Biology (Wagner *et al.*, 2012; Soyer and O'Malley, 2013) field is a combination of Systems Biology (Kitano, 2002) and the Evolutionary framework which is concerned with understanding how evolutionary innovation might arise in biological systems in different scales, from molecules to ecosystems. It is a deeply interdisciplinary field that merges both omics data with the modern evolutionary synthesis and population genetics in the search of general trends regarding biological systems. Lastly, the field of Metagenomics is concerned on exploring the universe of uncultivable microbes, using different approaches for studying the structure and function of bacterial communities and allowing the prospection of biological parts (such as genes, promoters etc.) through functional assays (Singh *et al.*, 2009; Delmont *et al.*, 2011). We believe this multiscale approach is fundamental for both

understanding the general principles underpinning information processing in living systems and (re)engineering them for biotechnological applications.

Figure 3. A multiscale approach for understanding and (re)engineering life. The combination of Synthetic Biology, Systems Biology, Evolution and Metagenomics provide a combined strategy for exploring a single molecular system through multiple lenses in a more holistic manner. Credits for central image: Drawing Hands by M. C. Escher (1948).

We also emphasize the relevance of data integration as an important aspect of these multiscale approaches. The gigantic volume of omics data accumulated over the last decades has shown that although high-throughput technologies are able to generate an enormous amount of data, without the correct framework, the real biological meaning beneath it remains hidden (Gehlenborg *et al.*, 2010; Gomez-Cabrero *et al.*, 2014; Dolinski and Troyanskaya, 2015; Stephens *et al.*, 2015). Thus, this is the age of generating novel ways of interpreting and integrating big data for a comprehensive view of biological phenomena as a whole and not through its individual parts. In this context, understanding and developing novel tools for modelling biological networks is essential for simulating and predicting systemic behaviours in whole-cell computational models (**Figure 4**) (Covert *et al.*, 2004; Fondi and Liò, 2015; Karr,

Takahashi and Funahashi, 2015; Yugi *et al.*, 2016). In the next sections, we shall briefly introduce each of the previously cited fields for a better understanding of how they can be applied to the study of transcriptional regulatory systems in bacteria.

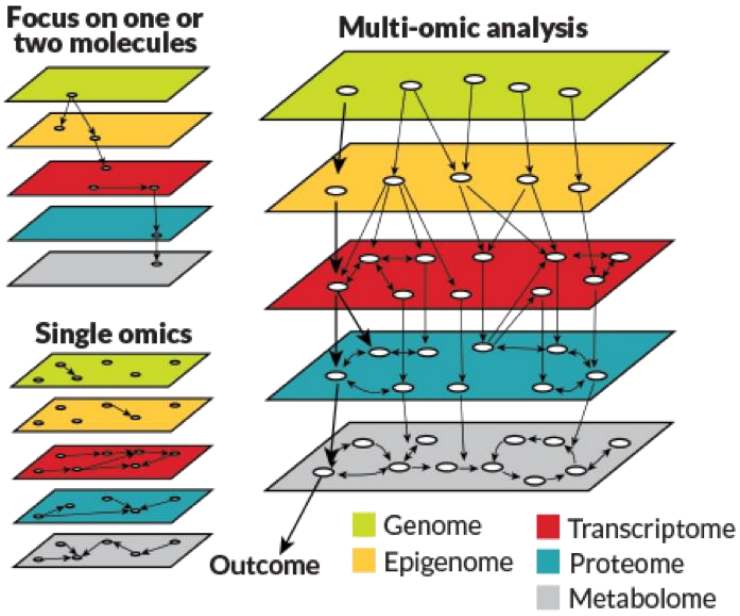


Figure 4. Data integration in biological systems. Currently, there is a lack of a holistic/systemic view of how organizational layers are interconnected in the implementation of cellular functions and information processing mechanisms in living organisms. The literature is surrounded by reductionist studies on small portions of the cellular systems, however, the making of an integrated representation of biological phenomena is still in its embryonic phase. Connections between genes and their outputs or within -omic layers can provide some clues to the molecular organization inside an organism, however, integrating -omic datasets through multi-omic analysis and modelling approaches can provide a much deeper and insightful view of biological phenomena. Modified from (Yugi et al., 2016)

4.1 Synthetic Biology

4.1.1 Advances and challenges in Synthetic Biology

The term Synthetic Biology was first introduced in 1912 by Stéphane Leduc (Leduc, 1912); However, only in the last few years has its meaning come to describe the interface between the areas of Molecular Biology and Engineering (Andrianantoandro *et al.*, 2006; Decoene *et al.*, 2017; Del Vecchio *et al.*, 2018; Westmann, Guazzaroni and Silva-Rocha, 2018). This field seeks to understand and utilize "biological parts" - biobricks - (e.g. genes, promoters, regulators,

etc.) as building blocks in the "engineering" of biological systems, adopting a hierarchical, qualitative and quantitative approach to their standardization (Galdzicki *et al.*, 2014; Beal *et al.*, 2016; Beal, Haddock-Angelli, Baldwin, *et al.*, 2018; Beal, Haddock-Angelli, Farny, *et al.*, 2018) (**Figure 5**). Their applications are vast and range from biotechnology (e.g., drug transporter systems, biofuels, water purification, etc.) to basic sciences (e.g. origins of life, robustness analyzes, molecular architecture etc.). One of the ultimate goals of Synthetic Biology is to recreate a cell as an automaton that can process information algorithmically and perform specific functions (de Lorenzo and Danchin, 2008).

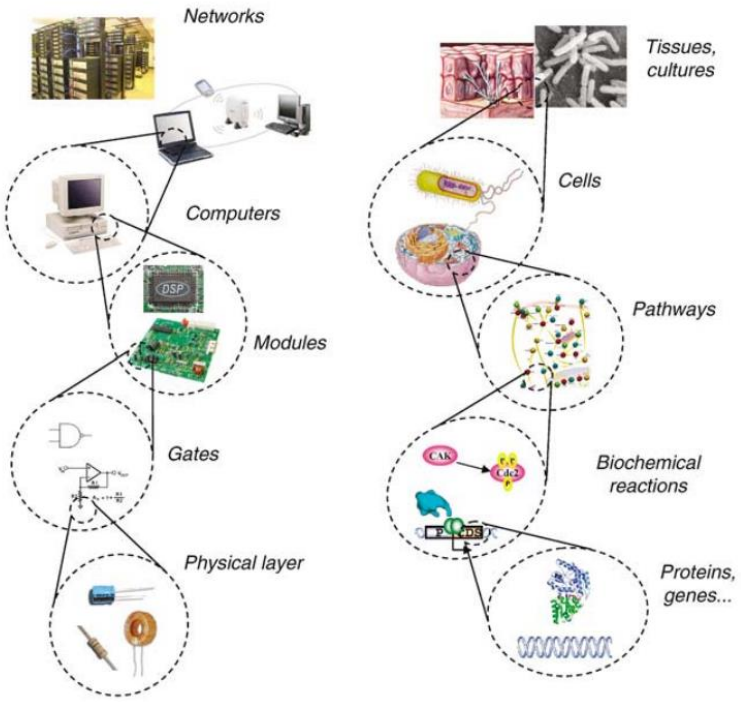


Figure 5 A possible hierarchy for synthetic biology is inspired by computer engineering. Retrieved from (Andrianantoandro *et al.*, 2006)

However, in order to achieve its objectives and become a discipline with strong roots in engineering, Synthetic Biology needs to overcome at least one major limitation: the struggle in executing the rational engineering of dynamic biosystems in a predictive and quantitative manner. Such an approach is difficult to apply even in less complex systems (in the sense of the molecular network connectivity) since there is still a huge gap between the mechanistic study of individual parts and the systemic understanding of the organism and the intrinsic logic embedded in its organizational layers (Danchin, 2012). Thus, without adequate knowledge about the properties of the biological components - and especially of the properties that emerge

from a system with multiple components - the generation of new gene circuits and biological functions intuitively becomes very ineffective. In this context, it is essential not only to study as many systems and molecular components as possible, but also to bring these studies into a Systemic and Synthetic Biology perspective. This interdisciplinary approach will allow deeper insights regarding some of the underlying principles of Life and how to re-engineer them.

4.1.2 Prokaryotic regulatory networks as a model system

Within this scenario, a model of great interest for Synthetic Biology is that of prokaryotic gene regulation. Regulatory networks are phylogenetically highly conserved, robust and hierarchically organized systems (Barabasi and Oltvai, 2004; Aldana *et al.*, 2007; Payne and Wagner, 2015) that can be also be embedded in frameworks such as Information Theory and Engineering (e.g. Boolean Logic and molecular switches). The studies of these networks provide valuable substrates for the understanding and application of logic among the components of the system, as observed in recent works that have uncovered part of the regulatory logic in the development of model eukaryotes (Hart *et al.*, 2012; Peter, Faure and Davidson, 2012; Wunderlich *et al.*, 2012).

On the other hand, the focus on prokaryotes such as *Escherichia coli*, besides taking advantage of their great biotechnological potential, also provides some of the best characterized regulatory systems (Gama-Castro *et al.*, 2016), with less complexity in several biological scales (Lawrence, 1999; Lane and Martin, 2010) and relatively simple genetic manipulation methods (Sambrook, J.; Fritsch, E. F.; Maniatis, 1989) - in relation to eukaryotes -. It is also worth noting that prokaryotes have regulatory elements that are usually larger and less frequent in their genome than in eukaryotes (Koonin and Wolf, 2008; van Hijum, Medema and Kuipers, 2009), usually organized together with functional genomic regions (operons) and with a high degree of informational compression regulatory logic. This is a phenomenon that has been gradually unravelled in recent years (Milo *et al.*, 2002; Silva-Rocha and de Lorenzo, 2008; Bendtsen *et al.*, 2011), bearing enormous potential for the generation of new tools and applications in Synthetic Biology (Bashor and Collins, 2018; Xie and Fussenegger, 2018).

4.2 Evolutionary Systems Biology

4.2.1 An introduction to Evolutionary Systems Biology

At its most basic, evolutionary systems biology (ESB) is the synthesis of system-level approaches to biological function with evolutionary explanations of multilevel properties. "System" in this context refers to dynamically interacting components that produce behaviour not revealed by analyses of isolated components. Cellular interactions of signalling, regulatory, and metabolic components are all considered as systems or networks that can display structural complexity and nonlinear dynamics. In this context, ESB recognizes that the system-level properties of cellular networks are subject to evolutionary change and that evolved network properties will variously influence the future evolutionary course of the organism. It is this interdependency between evolutionary processes and system properties that ESB aims to understand. One of the earliest articles to describe ESB was published in 2005 (Medina, 2005), highlighting that the field goes beyond existing efforts to merge molecular and evolutionary biology (Dean and Thornton, 2007). It focuses on the study of phenotypes as the results of evolving intracellular interaction networks (**Figure 6**), integrating theoretical tools, experimental methods, and extensive datasets within an evolutionary framework. This integration is occurring in a highly pragmatic manner to develop closer insight into evolving genotype-phenotype mappings across different biological scales (**Figure 6**). Researchers with this goal seize upon tools and datasets as they become available (e.g. dynamical models, gene-knockout studies, flux balance analyses, *in silico* evolution, reverse engineering, comparative omics data) to address questions as old as biology or to reformulate new ones in light of system level insight. While specific organizational features may have evolved to cope with environmental perturbations, it is getting a grip on "how" these are implemented at the molecular level rather than "why" that is considered the important task.

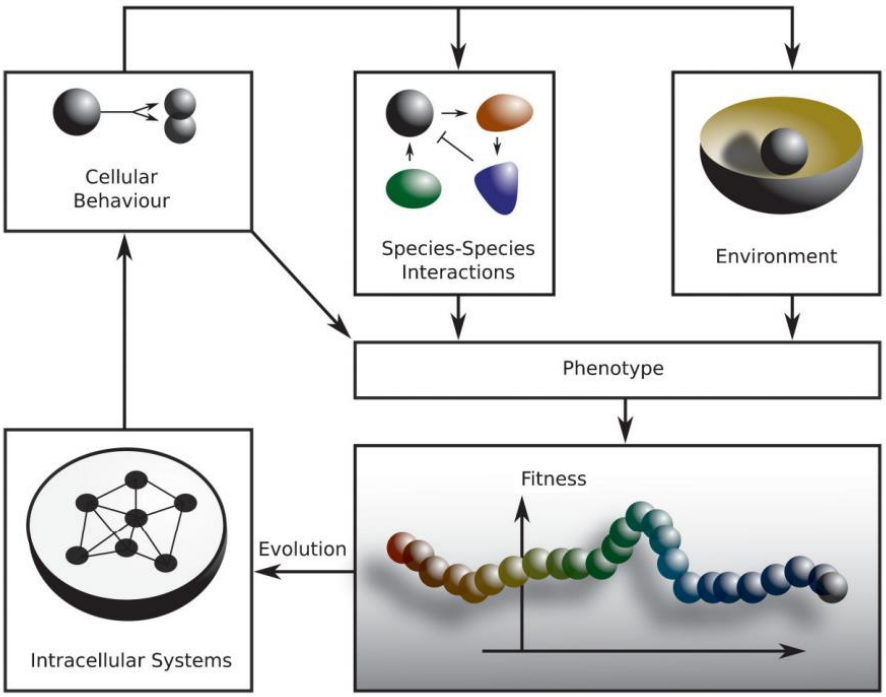


Figure 6. Systems under evolutionary forces. At the core of ESB lies the aim of achieving a deep mechanistic understanding of genotype-phenotype mappings in biological systems. While these mappings can be drawn at different levels in different combinations of ESB, a major area of interest is currently intracellular systems. These systems give rise to cellular physiology, which – in the case of unicellular organisms – directly determines species' interactions with their environment and other organisms. These higher-level interactions are responsible for the fitness of organisms. Evolutionary processes (i.e. neutral drift and adaptation) move populations of these organisms on this dynamic fitness landscape by altering the properties of their intracellular systems. Retrieved from (Soyer and O'Malley, 2013).

4.2.2 An introduction to regulatory complexity

As explained before, the decision-making process in bacterial cells is the result of a myriad of regulatory interactions, which can, in a simplistic manner, be represented as intertwined hierarchical regulatory layers. In this context, gene regulatory networks (GRNs), which represent the coordination of bacterial transcription processes, represent the most fundamental layer in the molecular decision-making process. GRNs consist in Transcription Factors (TFs) that may act as repressors or activators of gene expression depending on environmental signals – which can be extracellular or intracellular (Browning and Busby, 2016). TFs that regulate the largest fraction of genes in a bacterial cell are considered global regulators, whereas TFs that modulate just a few genes are considered local regulators (Martínez-Antonio *et al.*, 2003; Martínez-Antonio, 2011). Both global and local regulators act on specific *cis*-regulatory elements (CREs) modulating the RNA-Polymerase (RNAP) physical access and/or affinity to the core promoter region, modulating the transcription initiation (Browning and Busby, 2016).

Although a great effort has been made over the last decades for elucidating the individual roles of TFs on gene expression, it has become clear that the nature of most bacterial promoters allows the environmental signal integration to be densely compacted in bacterial genomes yet extremely precise in terms of gene expression.

Recent studies have applied high throughput techniques and rational design approaches to evaluate the transcriptional outcomes of different complex promoter architectures (Buchler, Gerland and Hwa, 2003; Hermsen, Tans and Ten Wolde, 2006; Kinkhabwala and Guet, 2008; Gertz, Siggia and Cohen, 2009; van Hijum, Medema and Kuipers, 2009; Sharon *et al.*, 2012; Peeters, Peixeiro and Sezonov, 2013; Rydenfelt *et al.*, 2014; Monteiro, Arruda and Silva-Rocha, 2017; Yuan *et al.*, 2018) - regarding the relative position and nature of their Transcription Factor Binding Sites (**Figure 7**). The general messages from all these studies are: (i) the generation of synthetic complex promoters reveal principles of design for many regulatory functions and (ii) unpredictable emergent properties rise in those systems (Buchler, Gerland and Hwa, 2003; Kinkhabwala and Guet, 2008; Gertz, Siggia and Cohen, 2009; van Hijum, Medema and Kuipers, 2009; Sharon *et al.*, 2012; Peeters, Peixeiro and Sezonov, 2013; Monteiro, Arruda and Silva-Rocha, 2017).

However, most of the conclusions from those studies are restricted to eukaryotes and may not be applicable to prokaryotic systems as their core regulatory logics are intrinsically different (Struhl, 1999). Thus, there is an urgent need for understanding the fundamental rules underlying transcriptional regulation in bacterial systems. In this context, our research group has been providing essential information to fill this gap and to expand the design rule principles for engineering bacterial transcriptional systems (Silva-Rocha and de Lorenzo, 2008; Guazzaroni and Silva-Rocha, 2014; Amores, Guazzaroni and Silva-Rocha, 2015; Monteiro, Arruda and Silva-Rocha, 2017).

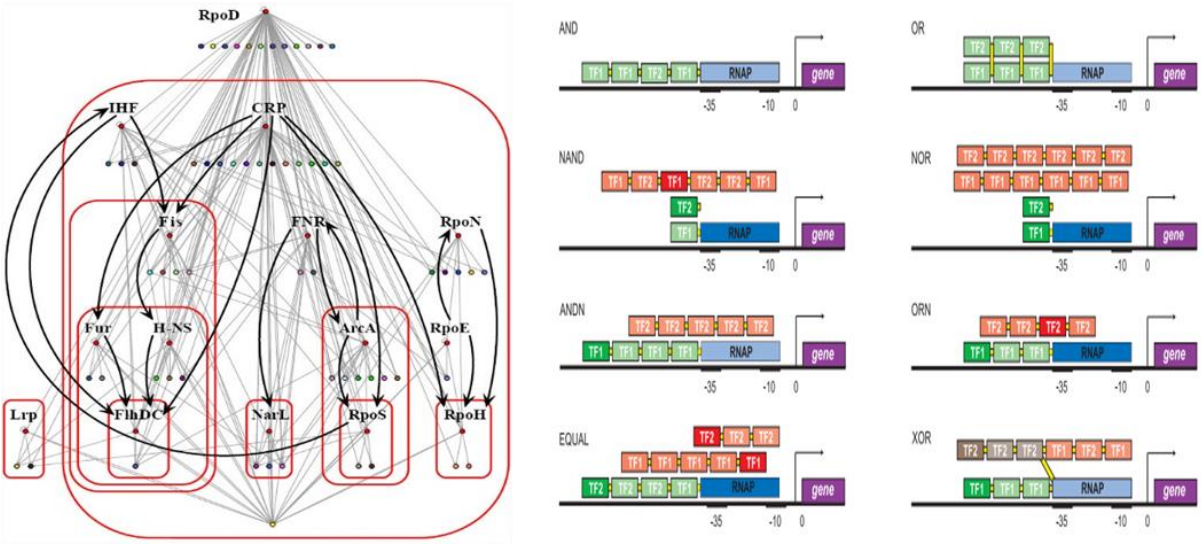


Figure 7. GRNs and complex promoters in *E. coli*. Left. Most global regulators extracted from *E. coli* transcriptional regulatory network, originally published by 2010 Nature Education, All rights reserved. Right. *In silico* prediction of regulatory output functions – as logic gates AND, OR, NOR etc. – in a range of complex promoter architectures. Each small box (red, orange or green) is a representation of a single TFBS within the complex promoters, modified from (Hermsen, Tans and Ten Wolde, 2006).

4.2.3 The rise of innovation in regulatory systems from the perspective of CREs

As Theodosius Dobzhansky once said, "nothing in Biology makes sense except in the light of evolution" (Dobzhansky, 1973) and it also applies to the understanding the evolutionary principles underpinning the diversity of TF binding sites and the architecture of complex promoters. Evolution usually navigates through a small set of possible solutions (genotype space) for locally optimizing specific functions, depending on the selective pressures (Kauffman, 1994; Loewenstein, 2006). Each optimization step relates – directly or indirectly - to a specific phenotype, which, in turn, composes the set of adaptive traits of an organism. Changes in the *cis*-regulation of gene expression have been proposed as a major source of evolutionary innovation (King and Wilson, 1975; Wittkopp and Kalay, 2012; Payne and Wagner, 2015; Lagator *et al.*, 2016). For example, across insect species, there has been increasing evidence for the essential role that *cis*-regulatory changes have in shaping body plan formation (Carroll, 2008; Wittkopp and Kalay, 2012). Changes in the regulation of gene expression can occur through mutations in the transcription factor coding sequence (*trans*-regulatory elements) and/or in CREs, which contain the transcription factor and the RNAP binding sites (Jacob and Monod, 1961). Mutations in CREs may be important targets of selection (Stern and Orgogozo, 2008), as it is hypothesized that, compared with *trans* elements, mutations in CREs have a wider range of effects, giving rise to a greater diversity of phenotypes

that could be selected upon (Wray, 2007). When considering transcriptional systems, the potential set of DNA sequences that are bound by a specific TF can be classified as the available genotype space for evolution's navigation. The combination of multiple TFBSs in a complex promoter generates specific regulatory outputs for gene expressions, considered here as bacterial phenotypes, which will ultimately compose the fitness functions - the paths under selection for evolution's navigation (see **Figure 8**).

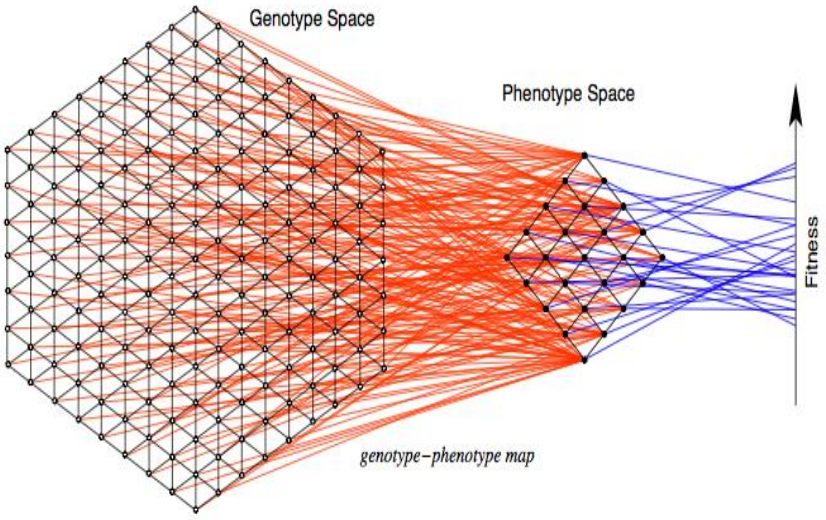


Figure 8. Representation of the connection between genotype, phenotype and fitness. Many genotypes give rise to the same phenotypes, which, in turn, will confer different fitness values depending on the selective pressures upon them. Modified from (Schuster, 2002)

In this context, one way to explore the principles underlying the rise of "innovation" in regulatory complexity is to adopt mutational models, which consider that evolution navigates a nucleotide space through the rise of spontaneous mutations. Indeed, many recent studies that have focused on understanding the dynamics of this process by experimental approaches, accessing parameters as the rate of *de novo* promoter acquisition in random DNA sequences (Yona, Alm and Gore, 2018), the role of epistatic interactions in gene regulation (Payne and Wagner, 2014; Lagator *et al.*, 2016; Aguilar-Rodríguez, Payne and Wagner, 2017) and the role of each nucleotide in determining the binding affinity of TFBSs (Newburger and Bulyk, 2009; Orenstein and Shamir, 2016; Belliveau *et al.*, 2018). A recent computational study using eukaryotic data has also provided essential information regarding evolution's navigability in thousands of adaptive landscapes for eukaryotic TFBSs, suggesting that landscape navigability may have contributed to the enormous success of transcriptional regulation as a source of evolutionary adaptations and innovations (Payne and Wagner, 2014; Aguilar-Rodríguez, Payne

and Wagner, 2017). However, most studies in this area are based on eukaryotic models, which, as described before, possess markedly different regulatory attributes in comparison to prokaryotes (Struhl, 1999; Berg, Willmann and Lässig, 2004; Wunderlich and Mirny, 2009; Stewart, Hannenhalli and Plotkin, 2012; Stewart and Plotkin, 2013) – such as smaller TFBSs/longer promoter regions, higher density of *cis*-regulatory sites, variable information content *per* TFBS.

4.3 Metagenomics

4.3.2 A brief introduction to metagenomics

About thirty years ago, in 1986, Pace and collaborators (Pace et al., 1986) proposed, for the first time, the revolutionary idea of cloning DNA directly from environmental samples to analyse the complexity of natural microbial populations. The adopted strategy was based on shotgun-cloning of 16S rRNA genes using purified DNA from natural samples. At that time, authors stressed that although the DNA was originated from a mixed population of microorganisms, the methodology allowed the recovery and subsequent sequencing of individual rRNA genes. Thus, by evaluating complete or partial rRNA sequences, the composition of the original microbial populations could be retrieved.

Around ten years later, in 1998, the term "metagenome" appeared, when Handelsman and collaborators (Handelsman et al., 1998) described the importance of soil microorganisms as sources for new natural compounds. According to them, a new frontier in science was emerging – the mining for novel chemical compounds from uncultured microorganisms, which comprises more than 99% of the microbial diversity (Sleator, Shortall and Hill, 2008). This new concept in microbial ecology opened the mind of the scientific community in respect to the astonishing large catalogue of biochemical functions available in nature remaining to be discovered.

Currently, metagenomics is subdivided into two major approaches, which target different aspects of the local microbial community associated with a determined environment. In the first one, the so-called structural metagenomic approach, the main focus is to study the structure of the uncultivated microbial population, which can be expanded to other properties, such as the reconstruction of the complex metabolic network established between community members

(Handelsman, 2005; Tringe et al., 2005). In this sense, the microbial community structure can be defined as the population composition and its dynamics in a specific ecosystem, in response to selective pressures and spatiotemporal parameters. The study of the community structure allows a deeper understanding about the relationships between the individual components that build a community and is essential for deciphering ecological or biological functions among its members (Tringe et al., 2005; Vieites et al., 2009). In a different manner, the functional metagenomic approach aims to identify genes that code for a function of interest, which involves the generation of expression libraries with thousands of metagenomic clones followed by activity-based screenings (Schmeisser, Steele and Streit, 2007; Guazzaroni, Silva-Rocha and Ward, 2015).

4.3.3 Metagenomics as a novel approach for exploring transcriptional systems

The study of prokaryotic transcriptional regulation is essential for understanding the molecular mechanisms underlying decision-making processes in microorganisms (Ishihama, 2010), comprising populational, ecological and pathogenic behaviours. The activity of most bacterial promoters is usually dependent on the combined action of transcription factors and sigma factors in response to multiple environmental stimuli (Browning and Busby, 2016). For instance, in *E. coli*, the compilation of decades of experimental data indicate that approximately 50% of its promoters are under the control of a single specific regulator, while all other genes are regulated by at least two transcription factors (Gama-Castro *et al.*, 2016). Moreover, the recent development of experimental and large-scale sequencing techniques, together with powerful computational approaches have allowed both the discovery of insightful information about other bacterial transcriptional systems and the development of novel approaches for studying them n higher depth (Shen-Orr *et al.*, no date; Martínez-Antonio *et al.*, 2003; Covert *et al.*, 2004; Shimada *et al.*, 2005). However, despite technical innovations, most of the studies are still centered on *E. coli*, a single bacterial species among at least 30,000 other already sequenced (Land *et al.*, 2015), in an estimated total of 1 trillion species (Locey and Lennon, 2016; Thompson *et al.*, 2017).

With the advent of Metagenomics (Handelsman *et al.*, 1998), the exploration of unculturable bacteria (approximately 99% of a bacterial community (Amann, Ludwig and Schleifer, 1995) widely expanded genomic information, providing resourceful data about populational

structures and genetic diversity in a myriad of environmental samples (Torsvik and Øvreås, 2002; Venter, 2004; Tringe, 2005). In this context, although a large number of genes/ORFs has been discovered through the previously described approaches, the detection of novel bacterial regulatory elements using high-throughput technologies has been poorly explored. So far, the most relevant researches concerning regulatory elements in metagenomic are: a single well-defined method for the discovery of substrate-inducible regulatory sequences – SIGEX (Uchiyama *et al.*, 2005) -; a direct assay for prospecting promoters for industrial applications (Han et al. 2008) and a recent large-scale metagenomic mining of thousands of natural 5′ regulatory sequences from diverse bacteria, and their multiplexed gene expression characterization in industrially relevant microbes (Johns *et al.*, 2018) . This scarce number of methodologies is directly related to the biased search towards novel enzymatic activities and to a lack of both experimental and computational tools for finding and validating promoter sequences in metagenomic libraries (Guazzaroni, Silva-Rocha and Ward, 2015).

Unravelling novel bacterial promoters is essential for understanding the regulatory diversity of microorganisms, addressing important questions, such as the abundance of both constitutive and inducible elements in a metagenomic library, the bottlenecks regarding host choices (i.e. the constraints limiting the diversity of exogenous promoters that can be recognized by different hosts) and the correlation between promoter strength, transcriptional noise and the functional role of the regulated gene/operon (Ekkers *et al.*, 2012; Silander *et al.*, 2012; Guazzaroni, Silva-Rocha and Ward, 2015; Vester, Glaring and Stougaard, 2015). Furthermore, prospecting and characterizing novel promoters is crucial for expanding the current Synthetic Biology toolbox and generating novel biotechnological applications as there is a high demand for constitutive and inducible promoters responding to process-specific parameters (**Figure 9**) (Uchiyama *et al.*, 2005; Silva-Rocha and de Lorenzo, 2008; Boyle and Silver, 2009; Blount *et al.*, 2012; Guazzaroni, Silva-Rocha and Ward, 2015).

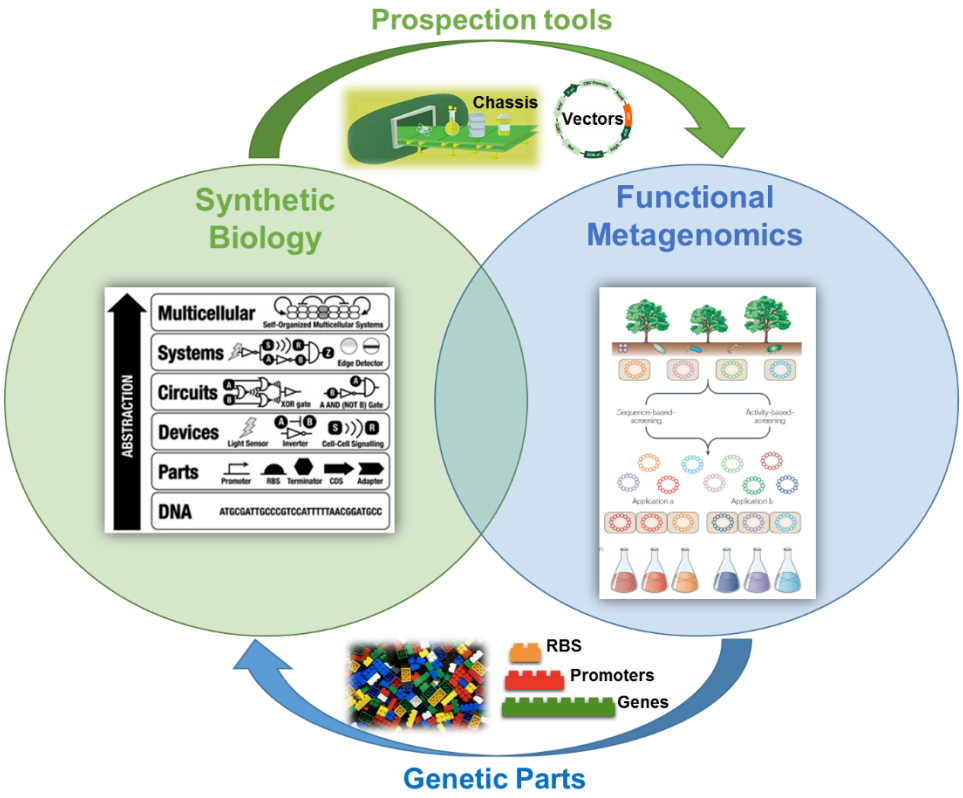


Figure 9. Synergies between Synthetic Biology and Metagenomics. Synthetic Biology is an emerging interdisciplinary field that combines engineering principles and molecular biology. Genetic features such as genes, promoters and ribosome binding sites are conceptually seen as biological parts that can be assembled in the design and building of genetic circuits with specific functions. Thus, one of the greatest challenges is the discovery and characterization of new biological parts. In this context, the parallel progress in the Functional Metagenomics field over the last decade has turned it into a powerful tool for mining new genetic components.

4.4 Data Integration and –omics-based models

The ease at which genomes are currently sequenced has assigned to genomics one of the first steps in microbial systems biology. Regardless of the technique used, assembly and annotation typically follow genome sequencing and return an almost complete picture of the genetic reservoir of a given microorganism. On the other hand, genome sequence only represents a snapshot of the real phenotypic capabilities of an organism, providing very few indications on other crucial aspects of the underlying life cycle such as response to environmental and genetic perturbations, fluctuations in time, gene essentiality and so on (Costanzo *et al.*, 2010; Papp, Notebaart and Pál, 2011; Stallins *et al.*, 2018). To gain a systemic and exhaustive description of living entities, static information deriving from genome sequence is not enough and other levels of knowledge must be taken into consideration. Nowadays, technologies do exist for

measuring, in a large-scale fashion, other crucial aspects of cellular life, including the level of RNA within the cell (transcriptomics), the nature of metabolites present within the cell (metabolomics), the interaction among different proteins (protein–protein interaction) and many others (O'Malley and Soyer, 2012; Gomez-Cabrero *et al.*, 2014; Fondi and Liò, 2015; Goldberg *et al.*, 2018). In addition, metabolic biodiversity of microbial communities can be today evaluated through metagenomics and metatranscriptomics approaches (**Figure 10**). However, no single -omics analysis can fully unravel the complexities of fundamental microbiology (Zhang, Li and Nie, 2010). Multi- and integrated -omics approaches have thus started spreading among several research areas, from bio-based fuel production (Zhu *et al.*, 2013) to biopharmaceuticals processes (Schaub *et al.*, 2012), from medical research (Wiench *et al.*, 2013) to host–pathogen interactions (Ansong *et al.*, 2012). The integration of such diverse data types may be considered one of the key challenges of present-day bioinformatics, due to different data formats, high data dimensionality and need for data normalization.

One of the most important drawbacks associated with the booming of genomics resides in the possibility to (almost) automatically derive the potential metabolic landscape of a strain, given its genome. Bacteria continuously provide industry with novel products/processes based on the use of their metabolism and numerous efforts are being undertaken to deliver new usable substances of microbial origin to the marketplace (Beloqui *et al.*, 2008), including pharmaceuticals, biofuels and bioactive compounds in general (García-Ochoa *et al.*, 2000; Tan, Mccue and Stormo, 2005; Zou *et al.*, 2012). In this context, computational modelling and *in silico* simulations are often adopted by metabolic engineers to quantitatively simulate chemical reactions fluxes within the whole microbial metabolism (Lewis, Nagarajan and Palsson, 2012; Bordbar *et al.*, 2014). To exploit computational approaches, genome annotation-derived metabolic networks are transformed into models by defining the boundaries of the system, a biomass assembly reaction, and exchange fluxes with the environment (Durot, Bourguignon and Schachter, 2009). Also needed are (i) structured (mathematical) representation of that network, (ii) possibly quantitative parameters enabling simulations or predictions on the joint operation of all network reactions in a given environment and, in particular, (iii) predictions on the values of metabolite fluxes and/or concentrations (Papin *et al.*, 2003). A constraint-based modelling framework can then be used to automatically compute the resulting balance of all the chemical reactions predicted to be active in the cell and, in turn, to bridge the gap between

knowledge of the metabolic network structure and observed metabolic processes (Varma and Palsson, 1994).

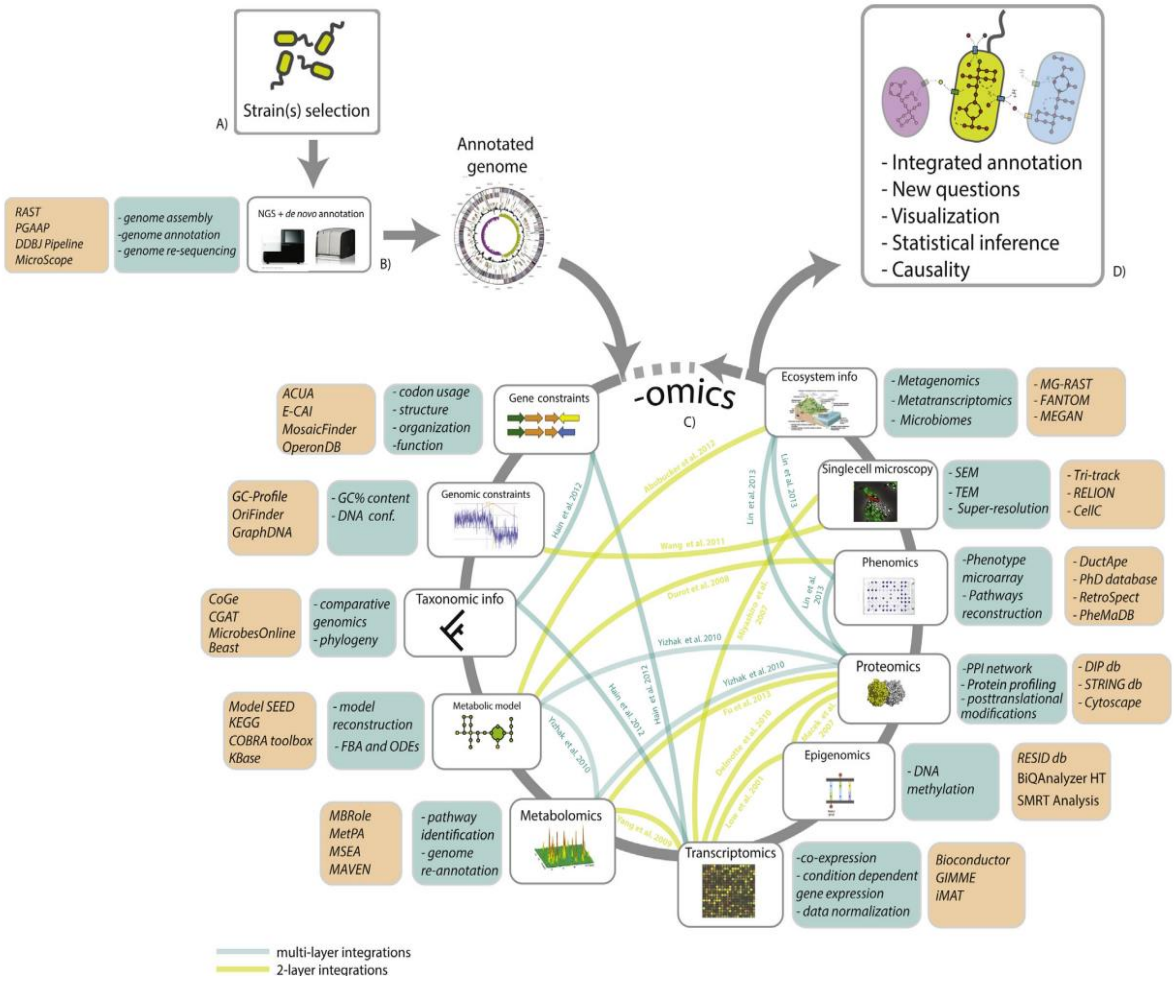


Figure 10. Data integration in microbial -omics pipelines. Nowadays, to gain a systems-level perspective on biotechnologically relevant strains NGS and preliminary genome annotation is usually performed. After this step, information on the presence/absence of metabolic pathways and overall metabolic capabilities of a given microbe is gained. Nevertheless, to obtain a systems-level knowledge, a body of additional information can be mapped onto a genome annotation (the "-omics wheel"). This includes: gene and genomic constraints (derived from a deep inspection of genome properties), taxonomic and metabolic information, "-omics" data (transcriptomics, proteomics, metabolomics, epigenomics, phenomics), other phenotypic information (e.g. high-resolution microscopy), ecosystem information, (microbiome composition, community functional characterization, meta-transcriptomics). Furthermore, these different layers of information can be combined and integrated to merge together datasets resulting from the application of different technologies. Links among -omics represent present-day study cases in which integration among two or more information layers has been performed (see corresponding references). After -omics integration has been performed, a more comprehensive perspective on the microbe(s) under study is gained, providing clues on the possible interactions with the surrounding environment (including metabolic cross-talk with other microbial species), statistically grounded inferences and novel questions to be addressed (possibly re-iterating the pipeline). In this figure, orange boxes include possible softwares, while green-blue boxes specific tasks of general -omics strategy. Modified from (Fondi and Liò, 2015)

4.4.1 The importance of metabolic models for understanding phenotypes

Bottom-up approaches to systems biology rely on constructing a mechanistic basis for the biochemical and genetic processes that underlie cellular functions. Genome-scale network reconstructions of metabolism are built from all known metabolic reactions and metabolic genes in a target organism (Orth, Palsson and Fleming, 2010; Monk, Nogales and Palsson, 2014). Networks are constructed based on genome annotation, biochemical characterization, and the published scientific literature on the target organism. The reactome of a cell is assembled, or reconstructed, from all the biochemical reactions known or predicted to be present in the target microorganism. Importantly, network reconstruction includes an explicit genetic basis for each biochemical reaction in the reactome as well as information about the genomic location of the gene (Orth, Palsson and Fleming, 2010; Monk, Nogales and Palsson, 2014; O'Brien, Monk and Palsson, 2015). Thus, reconstructed networks, or an assembled reactome, for a target organism represent *bi*ochemically, *g*enetically, and *g*enomically structured knowledge bases, or BiGG k-bases (Schellenberger *et al.*, 2010). Network reconstructions have different biological scope and coverage. They may describe metabolism, protein-protein interactions, regulation, signalling, and other cellular processes, but they have a unifying aspect: an embedded, standardized biochemical and genetic representation amenable to computational analysis (O'Brien, Monk and Palsson, 2015).

A network reconstruction can be converted into a mathematical format and thus lend itself to mathematical analysis and computational treatment. *G*enome-scale models, called GEMs, have been under development for nearly 15 years and have now reached a high level of sophistication. The first GEM was created for *Haemophilus influenza* and appeared shortly after this first genome was sequenced (Edwards and Palsson, 1999) and GEMs have now grown to the level where they enable predictive biology (Oberhardt, Palsson and Papin, 2009; McCloskey, Palsson and Feist, 2013; Bordbar *et al.*, 2014). Some of the analysis made possible through GEMs are represented in **Figure 11.**

.

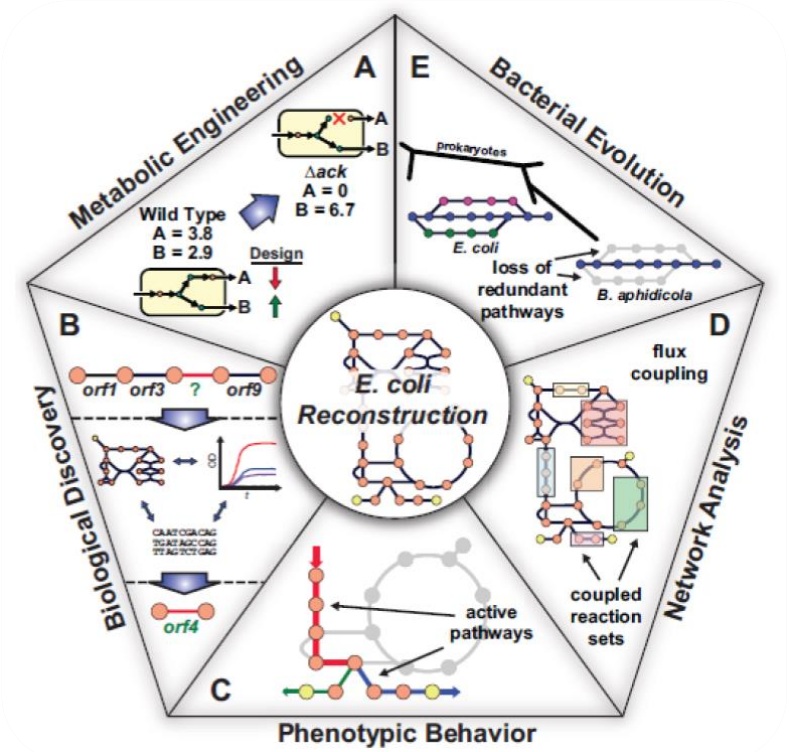


Figure 11. Uses of the *E. coli* reconstructions divided into five categories. (A) A drawing of a predicted effect from a loss of function mutation in a simple system is shown. Metabolic engineering studies have investigated *in silico* strain design using *E. coli* metabolic reconstructions to overproduce desired products. **(B)** Recent studies utilizing the reconstruction in a prospective manner have aimed to use the current biochemical and genetic information included in the metabolic network along with additional data types to drive biological discovery, such as predicting genes encoding for orphan reactions. **(C)** Utilizing the reconstruction in phenotypic studies, computational analyses have examined gene, metabolite, and reaction essentiality along with considering thermodynamics to make better predictions about the physiological state (i.e., the active pathways) of the cell for a given environmental condition. **(D)** The *E. coli* reconstructions have been used to analyze and interpret the intrinsic properties of biological networks. One example being finding coupled reaction activities (as shown in the drawing) across different growth conditions. **(E)** Using the network reconstruction, evolutionary studies have examined the cellular network in the context of adaptive evolution events, horizontal gene transfer and minimal metabolic network evolution (as shown in the drawing). Retrieved from (Feist and Palsson, 2008)

4.4.2 Metabolic modelling for unravelling communities

Metabolic interactions are an emergent property of microbial communities (Morris *et al.*, 2013; Chiu, Levy and Borenstein, 2014). Even the simplest life forms can only be understood in terms of biological consortia characterized by shared metabolic pathways and distributed biosynthetic capacities (Klitgord and Segrè, 2010; McCutcheon and Moran, 2011; Husnik *et al.*, 2013). For example, glucose catabolism to carbon dioxide or methane is a multi-step process often involving several organisms that indirectly exchange intermediate products through their environment (Stams, 1994). Microbial communities are thus complex systems comprising several interacting components that cannot be fully understood in isolation. In fact, metabolic interdependencies between organisms are at least partially responsible for our current inability

to culture the great majority of prokaryotes. Understanding the emergent dynamics of microbial communities is crucial to harnessing these multicomponent assemblages and using synthetic ecology for medical, environmental and industrial purposes (Brenner, You and Arnold, 2008). Genome sequencing has enabled the reconstruction of full-scale cell-metabolic networks (Henry *et al.*, 2010), which have provided a firm basis for understanding individual cell metabolism (Varma and Palsson, 1994; Duarte, 2004; Klitgord and Segrè, 2010). Recent work indicates that multiple cell models can be combined to understand microbial community metabolism and population dynamics (Stolyar *et al.*, 2007; Klitgord and Segrè, 2010; Zengler and Palsson, 2012; Chiu, Levy and Borenstein, 2014; Harcombe *et al.*, 2014) (**Figure 12**). These approaches assume knowledge of all model parameters such as stoichiometric coefficients, maintenance energy requirements or extracellular transport kinetics, a requirement that is rarely met in practice (Feist and Palsson, 2008; Harcombe *et al.*, 2014).

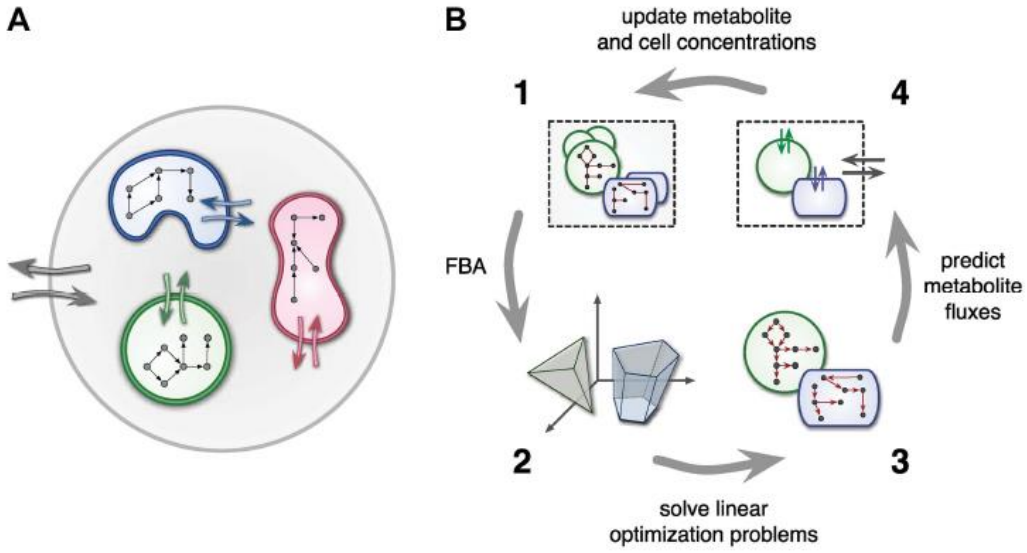


Figure 12. Metabolic modelling for understanding community interactions. (A) Conceptual framework. Cells (colored shapes) optimize their metabolism for maximal growth and influence their environment via metabolite exchange (small colored arrows). Additional external fluxes can also affect the environment (large grey arrows). The environment, in turn, influences each cell's metabolism. **(B)** Computational framework. Each iteration consists of four steps: flux balance analysis (FBA) is used to translate cell-metabolic potentials and environmental conditions **(1)** into a linear optimization problem for the growth rate of each cell species **(2)**. The set of possible reaction rates corresponds to a polytope in high-dimensional space. Solving the optimization problems **(3)** yields predictions on microbial metabolite exchange rates **(4)**. Metabolic fluxes and cell growth rates are used to predict metabolite and cell concentrations in the next iteration **(1)**. Retrieved from (Louca and Doebeli, 2015).

II. OBJECTIVES

General objective:

To explore the different aspects of bacterial transcriptional systems in an integrated manner.

Specific objectives:

1. To understand the basic rules underlying the generation of transcriptional logic and emergent behaviours in complex bacterial promoters by combinatorial design and *in vivo* testing of synthetic promoters with Fis and IHF transcription factor binding sites in *E. coli*.
2. To use *in silico* approaches to explore how innovation rises in transcriptional systems and how evolution navigates in the sequence space of bacterial transcription factor binding sites, using as a case of study *E. coli* sequences bound by the global regulators CRP, Fis and IHF.
3. To explore and characterize the natural diversity of regulatory elements in environmental samples using functional metagenomics approaches.
4. To establish a computational pipeline for *de novo* generation of bacterial metabolic networks as an initial step for data integration in biological informational processing systems

III. RESULTS

Chapter I

A Synthetic Biology approach to engineer and decipher underlying logic rules in complex bacterial promoters

1. Specific Background

The Synthetic Biology framework usually conceptualises gene regulatory interactions as Boolean networks in order to simplify the modelling of cellular behaviours. Boolean networks consist of a set of nodes whose state is binary and determined by other nodes present in the network, through Boolean functions (Wang, Saadatpour and Albert, 2012). Although these networks were originally used in electronic circuits and computer sciences, in the form of digital logic gates, in recent years, their application has extended to several areas of cellular biology, especially in the study of gene regulation (Alon, 2007; Lim, 2010; Morris *et al.*, 2010). The generation of these models depends on multiple approaches (e.g. experimental, mathematical and computational), since only after the experimental elucidation of the individual components and their interactions, a system can be represented in the form of a coherent Boolean network (Wang *et al.*, 2011). Such a representation, although coarse-grained in terms of mathematical models, is very powerful and allows to describe/predict the temporal and qualitative behaviours of the system, as well as its changes through different perturbations (Wang, Saadatpour and Albert, 2012).

Prokaryotic transcriptional networks have a large number of regulatory modules that formally behave like many of the logic gates typical of digital Boolean circuits. The participation of one or more TFs in the regulation of a particular promoter confers the ability to integrate different input signals in a manner, which is not far from those described by Boolean logic gates. An archetypal example in this context is the operon lac of *E. coli* (Jacob and Monod, 1961; Wilson *et al.*, 2007) that had its behaviour described and modelled as an intermediate between AND-OR logical gates (Setty *et al.*, 2003). More complex systems have also been studied from Boolean formalisms and represented as logicomes, such as the TOL system of degradation of toluene and *m*-xylene in *Pseudomonas putida*, depicted in **Figure 13** (Silva-Rocha *et al.*, 2011; Silva-Rocha and de Lorenzo, 2013). These logicomes represent the set of logical operators that compute all information within the TOL pathway, modulating the system response.

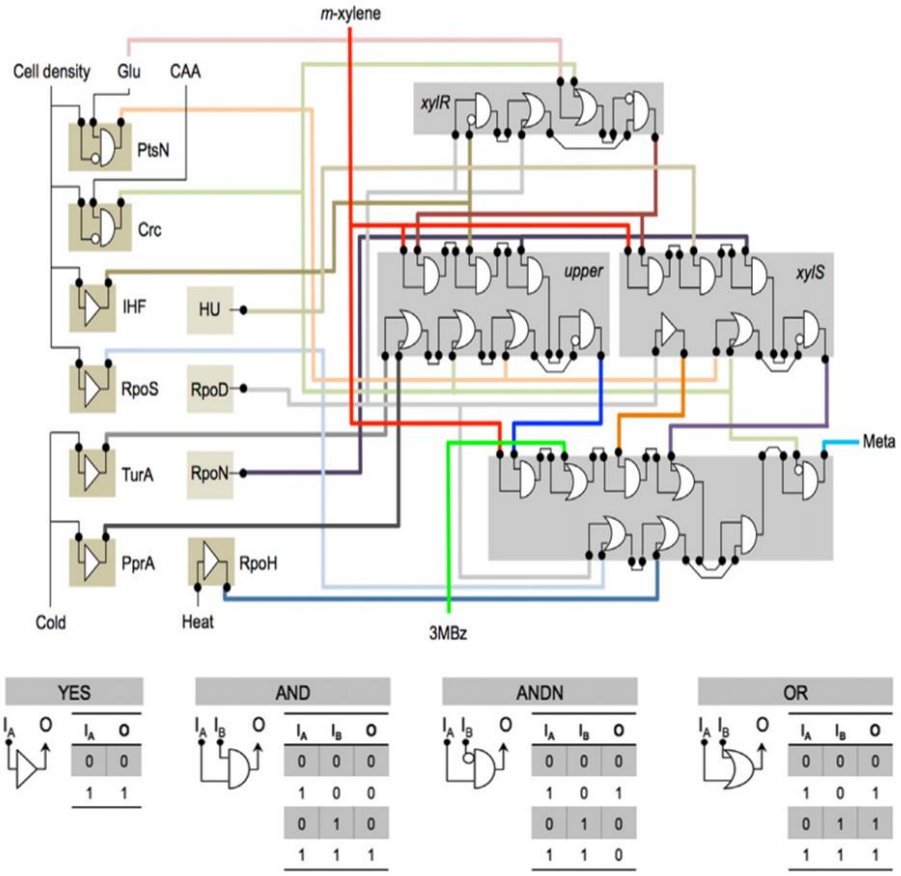


Figure 13. Formalization of the TOL network as a logic circuit. The complete set of logic gates that make up the TOL regulatory network (left) and their interactions with host factors (right) are represented. The logic gates used to create the circuit and their respective truth tables are represented at the bottom. Modified from (Silva-Rocha and de Lorenzo, 2013)

Synthetic Biology tries to use the previously presented concepts to integrate molecular elements as components of a complex biological circuit (De Silva and Uchiyama, 2007; Greber and Fussenegger, 2007) with the construction and / or modification of several genetic logic gates, including those performing AND and NOT functions (Buchler, Gerland and Hwa, 2003; Yashin, Rudchenko and Stojanovic, 2007; Friedland *et al.*, 2009; Dari *et al.*, 2011; Sayut, Niu and Sun, 2011; Siuti, Yazbek and Lu, 2013; Boer *et al.*, 2018). This process has great potential in the elucidation of biological systems and the generation of novel reliable/predictive behaviours in genetically modified organisms (Wang *et al.*, 2015).

Although the current methodologies to (re)engineer these biological systems usually vary depending on the desired goal, they can be roughly divided into two general frameworks:

(i) Those that use pre-existing regulatory elements - promoters and TFs, rewiring them according to a rational design to obtain any desired functions (Silva-Rocha and De Lorenzo, 2011; Tamsir, Tabor and Voigt, 2011).

(ii) Those that create synthetic regulatory elements in order to study the emerging behaviours in the system. This process has presented important results through modifications in sequences of regulatory regions, in TFs and in the architecture of promoters with respect to their number of operators (Gertz, Siggia and Cohen, 2009; Hunziker *et al.*, 2010; Silva-Rocha and De Lorenzo, 2012; Monteiro, Arruda and Silva-Rocha, 2017; Boer *et al.*, 2018).

However, it is important to notice that, within the context of Synthetic Biology, natural regulatory elements often appear to be highly-compressed and under a strong effect of the DNA regulatory architecture, leading to stochastic variations, continuous, non-binary responses, and molecular noise effects (Arkin and Ross, 1994; Elowitz *et al.*, 2007). These are important bottlenecks in the development of a reliable/precise engineering of biological transcriptional systems and in order to overcome them, it is necessary to:

(i) further explore the transcriptional elements within single and multiple species genomes;

(ii) study and improve the orthogonalization and disambiguation of natural regulatory modules

(iii) understand the basic rules underlying the generation of novel behaviours in the interactions between biological components in natural and synthetic systems.

Within the presented panorama, this project seeks to use Synthetic Biology tools to aid in the understanding of the logic - natural and Boolean - present in the transcriptional regulation in complex bacterial promoters. For this, the interactions between two global nucleotide regulators: Fis and IHF - essential in bacterial metabolism and highly associated in the genome will be studied. This study has a great scientific and biotechnological potential, allowing a better elucidation of the molecular system in question, as well as the generation of new tools and design principles in Synthetic Biology (for example, the description of new logic gates and their behaviour, with extrapolation for different biotechnological functions).

2. Objectives

General objective:

To use Synthetic Biology approaches to decipher the logic of signal integration in complex promoters. In this way, we focus on characterizing and understanding the relationship between the architecture of complex promoters and the logic of gene regulation dependent on global regulators in bacteria.

Specific objectives:

1. To construct and characterize a library of synthetic bacterial promoters in which Fis and IHF TFBSs can be combined in different configurations by both random and directed ligation of the two sequences in the promoter scaffold.

2. Characterization of the behaviour of selected promoters in bacterial populations through time-lapse and fluorescence analysis and analysis of emergent behaviours in order to decipher the combinatorial expression logic of Fis and IHF.

3. Materials and Methods

The methodology used in this project was previously established in our laboratory through the generation of promoters libraries containing the following sequences: Neg (inert sequence in the context of TF recognition), CRP and IHF (Monteiro, Arruda and Silva-Rocha, 2017).

Bacterial Strains , Plasmids, Primers, and Growth Conditions

The plasmids, bacterial strains, and primers used in this study are listed in **Table 1**. For cloning procedures, the bacterial strain used was *E. coli* DH5α. *E. coli* BW25113 was used as the wild-type strain (WT) whereas *E. coli* JW1702−1 was used as the mutant for IHF transcription factor, and both were obtained from the Keio collection.(Baba *et al.*, 2006) *E. coli* strains were grown at 37 °C in LB media with chloramphenicol at 34 μg mL−1 or in M9 minimal media (6.4 g L−1 Na2HPO4·7H2O, 1.5 g L−1 KH2PO4, 0.25 g L−1 NaCl, 0.5 g L−1 NH4Cl) supplemented with chloramphenicol at 17 μg mL−1, 2 mM MgSO4, 0.1 mM casamino acids, and 1% glycerol as the sole carbon source.

Table 1. Bacterial Strains , Plasmids and Primers used in this study

Strain	**Description**	**Source**
E. coli DH10B	F⁻ *endA1 deoR*⁺ *recA1 galE15 galK16 nupG rpsL* Δ*(lac)X74* φ80*lacZΔM15 ar aD139* Δ*(ara,leu)7697 mcrA* Δ*(mrr-hsdRMS-mcrBC)* Str^R λ⁻	(Casadaban and Cohen, 1980; Grant *et al.*, 1990)
E. coli DH5α	*F- endA1 gln V44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17(rK- mK+), λ−*	(Grant *et al.*, 1990)
E. coli BW25113	*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, rph-1, Δ(rhaD-rhaB)568, hsdR514.*	(Datsenko and Wanner, 2000)
E. coli JW3229	*E. coli* BW25113 with Δ*fis* mutation	(Baba *et al.*, 2006)
E. coli JW1702	*E. coli* BW25113 with Δ*ihf* mutation	(Baba *et al.*, 2006)
Plasmid	**Description**	**Source**
pMR1	CmR, orip15a; GFPlva promoter probe vector	(Guazzaroni and Silva-Rocha, 2014)
pMR1- NNNN	CmR, orip15a; pMR1- NNNN -GFPlva transcriptional fusion	This study
pMR1- NNNF	CmR, orip15a; pMR1- NNNF -GFPlva transcriptional fusion	This study
pMR1- NNFN	CmR, orip15a; pMR1- NNFN -GFPlva transcriptional fusion	This study
pMR1- NNFF	CmR, orip15a; pMR1- NNFF -GFPlva transcriptional fusion	This study
pMR1- NFNN	CmR, orip15a; pMR1- NFNN -GFPlva transcriptional fusion	This study
pMR1- NFNF	CmR, orip15a; pMR1- NFNF -GFPlva transcriptional fusion	This study
pMR1- NFFN	CmR, orip15a; pMR1- NFFN -GFPlva transcriptional fusion	This study

pMR1- NFFF	CmR, orip15a; pMR1- NFFF -GFPlva transcriptional fusion	This study
pMR1- FNNN	CmR, orip15a; pMR1- FNNN -GFPlva transcriptional fusion	This study
pMR1- FNNF	CmR, orip15a; pMR1- FNNF -GFPlva transcriptional fusion	This study
pMR1- FNFN	CmR, orip15a; pMR1- FNFN -GFPlva transcriptional fusion	This study
pMR1- FFNN	CmR, orip15a; pMR1- FFNN -GFPlva transcriptional fusion	This study
pMR1- FFNF	CmR, orip15a; pMR1- FFNF -GFPlva transcriptional fusion	This study
pMR1- FFFN	CmR, orip15a; pMR1- FFFN -GFPlva transcriptional fusion	This study
pMR1- FFFF	CmR, orip15a; pMR1- FFFF -GFPlva transcriptional fusion	This study
pMR1- FNFF	CmR, orip15a; pMR1- FNFF -GFPlva transcriptional fusion	This study
pMR1- FFNI	CmR, orip15a; pMR1- FFNI -GFPlva transcriptional fusion	This study
pMR1- NFNI	CmR, orip15a; pMR1- NFNI -GFPlva transcriptional fusion	This study
pMR1- NNFI	CmR, orip15a; pMR1- NNFI -GFPlva transcriptional fusion	This study
pMR1- NFFI	CmR, orip15a; pMR1- NFFI -GFPlva transcriptional fusion	This study
pMR1- FNNI	CmR, orip15a; pMR1- FNNI -GFPlva transcriptional fusion	This study
Primer	**Sequence (3'-5')**	**Source**
pMR1-F	CTCGCCCTTGCTCACC	This study
pMR1-R	ACAAGAATTGGGACAACTCC	This study
CoreP-5	CTTGAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAG	This study
CoreP-3	GATCCTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCT	This study
P1−I5	AATTCCAATTTATTGATTTTA	This study
P1−I3	CGCCTAAAATCAATAAATTGG	This study
P2−I5	GGCGCAATTTATTGATTTTA	This study
P2−I3	GCGGTAAAATCAATAAATTG	This study
P3−I5	CCGCCAATTTATTGATTTTA	This study
P3−I3	CCAATAAAATCAATAAATTG	This study
P4−I5	TTGGCAATTTATTGATTTTA	This study
P4−I3	CAAGTAAAATCAATAAATTG	This study
P1−N5	AATTCTCGCCTGCTTGTAGTA	This study
P1−N3	CGCCTACTACAAGCAGGCGAG	This study
P2−N5	GGCGTCGCCTGCTTGTAGTA	This study
P2−N3	GCGGTACTACAAGCAGGCGA	This study
P3−N5	CCGCTCGCCTGCTTGTAGTA	This study
P4−N5	TTGGTCGCCTGCTTGTAGTA	This study
P4−N3	CAAGTACTACAAGCAGGCGA	This study
P1−F5	AATTCTGCTCAAAAATTAAGC	This study
P1−F3	CGCCGCTTAATTTTTGAGCAG	This study
P2−F5	GGCGTGCTCAAAAATTAAGC	This study
P2−F3	GCGGGCTTAATTTTTGAGCA	This study
P3−F5	CCGCTGCTCAAAAATTAAGC	This study
P3−F3	CCAAGCTTAATTTTTGAGCA	This study
P4−F5	TTGGTGCTCAAAAATTAAGC	This study
P4−F3	CAAGGCTTAATTTTTGAGCA	This study

Design of the Minimal Promoter Scaffold and Ligation Reactions

Promoters were constructed by ligation of 5′end phosphorylated oligonucleotides (Cox, Surette and Elowitz, 2007; Kinkhabwala and Guet, 2008) acquired from Sigma-Aldrich (**Table 1**). All single strand nucleotides were designed to carry a discrete 16 bp sequence (Little *et al.*, 1980) containing a Fis binding site (F), IHF binding site (I), one Neutral (Neg) motif with no transcription factor binding, and a core promoter based on the *lac* promoter (boxes -10 / -35)

(**Table 1**), which is a weak promoter and therefore requires activation. All these oligonucleotides were designed to carry three base pair overhangs corresponding to their corrected insertion region on the promoter (**Figure 14A**). The sense and antisense strands corresponding to each position were mixed at equimolar concentrations and annealed by heating at 95 °C followed by gradual cooling to room temperature. External overhangs of the fourth *cis*-element position and the core promoters reassembled on the EcoRI and BamHI digested sites, allowing ligation to a previously digested EcoRI/BamHI pMR1 18 plasmid. The chosen vector, pMR1 (Guazzaroni and Silva-Rocha, 2014), medium copy number based on the p15a origin of replication has a resistance to chloramphenicol and two divergent reporter genes, one mCherry and one GFPlva. Thus, the vector allows the cloning of promoters by controlling the expression of the GFPlva gene, while allowing to investigate if the junction of the *cis*-regulatory elements could cause the appearance of a divergent promoter in the system. All five fragments (four *cis*-elements positions plus core promoter) were mixed at equimolar concentrations in a pool with the final concentration of 5′ phosphate termini fixed at 15 μM. For the ligase reaction, 1 μL of the pooled fragments was added to 50 ng EcoRI/BamHI pMR1 digested plasmid in the presence of ligase buffer and ligase enzyme to a final volume of 10 μL. After 1 h at 16 °C, the ligase reaction was inactivated for 15 min at 65 °C and one aliquot of 2 μL was then electroporated into 50 μL of *E. coli* DH10B competent cells. After 1 h of regenerating in 1 mL LB media, the total volume was plated onto LB solid dishes supplemented with chloramphenicol at 34 μg mL−1. Clones were confirmed by colony PCR using primers pMR1-F and pMR1-R (**Table 1**) using the pMR1 empty plasmid PCR reaction as a further length reference upon agarose gel electrophoresis. Clones with the potential correct length were submitted to Sanger DNA sequencing for confirming the correct promoter assembly.

Fluorescence assay (GFP) and data processing

To measure promoter activity, the library of 21 promoters was analyzed in different genetic backgrounds. For each experiment, a plasmid harbouring the promoter of interest was used to transform *E. coli* wild-type or *E. coli* Δ*fis* or *E. coli* Δ*ihf* mutant cells. Freshly plated single colonies were grown overnight in LB media, centrifuged, and resuspended in fresh M9 media. The culture (10 μL) was then assayed in 96-well microplates in biological triplicates with 170 μL of M9 media or M9 media supplemented with 0.4% glucose whenever required. Cell growth and GFPlva fluorescence were quantified using a Victor X3 plate reader (PerkinElmer). GFPlva is a GFP with a degradation tag. This degradation tag allows us to evaluate the expression over

time in an efficient and robust way. Promoter response was calculated as arbitrary units by dividing the fluorescence levels by the optical density at 600 nm (reported as GFP/OD600) after background correction. The same strain harbouring the pMR1 empty plasmid was used as the threshold background signal during calculations. Fluorescence and absorbance measurements were taken at 30 min intervals over 8 h. Technical triplicates and biological triplicates were included in all experiments. Raw data were processed using *ad-hoc* R script (https://www.r-project.org/) and plots were constructed using R or MeV (www.tm4.org/mev.html).

4. Results

4.1 Library generation and screening for positives

The results described below encompass the main objectives of the initial proposal focusing on the analysis of gene expression in bacterial populations using a commercial plate reader device. Firstly, the construction of a promoter library containing *cis*-regulatory elements controlled only by the Fis regulator was performed - since the library relating to the IHF regulator was previously constructed and validated experimentally in a previous work (Monteiro, Arruda and Silva-Rocha, 2017) -, using a fragment of DNA which is not recognized by any of the regulators of interest (NNNN sequence) as a negative control. The key elements of the construction can be found in **Figure 14**. For the construction of the initial library, the previously described protocol established in the laboratory was used so that the total volume of each ligation reaction yielded a set of more than $9x10^3$ clones. The assembly of the fragments was based on the complementarity of single-stranded DNA fragments (oligonucleotides) with the same size, allowing the formation of double-stranded DNA fragments when subjected to the hybridization reaction under equimolarity. Each complete fragment has a consensus promoter element (promoter core) containing the -10 and -35 boxes based on the *lac* promoter of *E. coli*, and 4 upstream positions which randomly contain sites for Fis, IHF or negative/inert sequences (**Figure 14**).

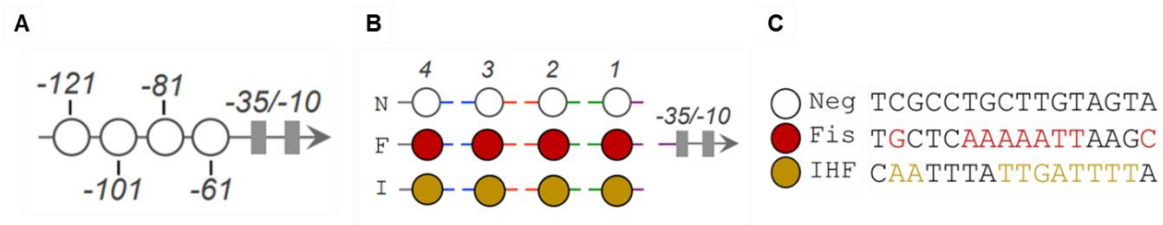


Figure 14. Construction of the complex promoter library. (A) Schematic representation of the promoter library, showing the positions −121, −101, −81, and −61 (white circles) at which *cis*-elements were inserted. The −35 and −10 boxes (grey rectangles) correspond to the core promoter. **(B)** Simplified scaffold scheme for the minimal synthetic promoter library. Motifs positions are identified as 4, 3, 2, and 1 respective to the core promoter, and colored lines represent the cohesive sequences for DNA ligation. **(C)** Nucleotide sequences for Neutral/Negative (N), Fis (F) and IHF (I) *cis*-elements.

Once the libraries were obtained, the colonies that had different levels of fluorescence were isolated by manual screening under a Safe Imager 2.0 (Life Technologies) blue light

transilluminator. The identified colonies were isolated and analyzed for the presence of inserts by PCR (i.e. for identification of correctly assembled promoters). In this context, a protocol for colony checking in the library was standardized through the combination of colony PCR technique and the analysis of fragments sizes by electrophoretic run on agarose gel (**Figure 15**). As a positive control, a "full" sequence with 4-position fragments was used and as a negative control, the pMR1 plasmid without any fragments was used. Only the fragments aligned with the central region of the positive control were selected for sequencing (red horizontal line in the gel on **Figure 15**).

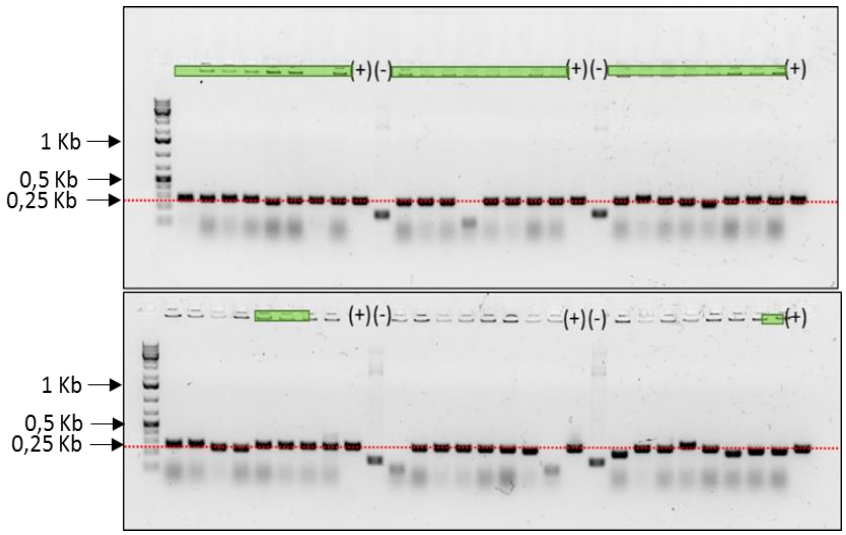


Figure 15. Confirmation of positive sequences by colony-PCR followed by agarose gel electrophoresis. The red dotted line represents the center of the positive control band. Clones with fragments aligned to the positive control were selected for sequencing. (+) positive control; (-) negative control; (green line) clones with GFP expression. DNA ladder (O'GeneRuler 1 kb Plus - Thermo Fischer Scientific) in the first column of the left with arrows indicating the relevant sizes relevant to this analysis.

After sequencing a large number of promoters from the random library containing a combination of only Fis and Neg sequences, 16 variants were obtained. To allow the construction of complex promoters simultaneously containing Fis, IHF and Neg binding sites, a protocol for the direct construction of promoters was employed. It is important to emphasize that previous results from our group with synthetic promoters containing binding sites for CRP and IHF (Monteiro, Arruda and Silva-Rocha, 2017), provided a basis for generating biologically relevant hypotheses in the current work. It has allowed the rational design of specific complex promoters, avoiding the exhaustive strategy of promoter generation by high-throughput methods without defined biological questions. The list of obtained constructs is shown in **Table 2**. After these constructions were obtained, each of them was transferred to the

wild-type *E. coli* (WT) or mutants for the *ihf* (Δ*ihf*) or the *fis* (Δ*fis*) genes for the experimental validation.

Table 2. Subset of complex promoters constructed and tested in this work

Nº	IHF/Neg*	Nº	Fis/Neg	Nº	Fis/IHF/Neg**
1	NNNN	**1**	NNNN	**1**	FFNI
2	NNNI	**2**	NNNF	**2**	NFNI
3	NNIN	**3**	NNFN	**3**	NNFI
4	NNII	**4**	NNFF	**4**	NFFI
5	NINN	**5**	NFNN	**5**	FNNI
6	NINI	**6**	NFNF		
7	NIIN	**7**	NFFN		
8	NIII	**8**	NFFF		
9	INNN	**9**	FNNN		
10	INNI	**10**	FNNF		
11	ININ	**11**	FNFN		
12	IINN	**12**	FFNN		
13	IINI	**13**	FFNF		
14	IIIN	**14**	FFFN		
15	IIII	**15**	FFFF		
16	INII	**16**	FNFF		

* The *IHF*/*Neg* library was previously built and is published as (Monteiro, Arruda and Silva-Rocha, 2017)

** The five constructions were based in previous results for *IHF* and *CRP* (Monteiro, Arruda and Silva-Rocha, 2017)

4.2 Characterization of synthetic promoters

For characterizing the activity of the constructed promoters, the resulting plasmids were transformed into *E. coli* strain BW25113 (wild-type) or *E. coli* Δ*ihf* (mutant for the *ihfA* gene) or *E. coli* Δ*fis* (mutant for the *fis* gene). For the experiments, colonies grown on plates were pre-grown overnight at 37 ° C with shaking in LB medium supplemented with chloramphenicol. After pre-growth, the lines were centrifuged and washed with M9 medium supplemented with 1% glycerol as the sole carbon source. Preliminary results of the libraries were plotted in the form of line graphs through *ad-hoc* scripts of the R software as shown in **Figure 16**. To facilitate the visualization of the fluorescence analyses, the data was plotted as color heatmaps maps - with the use of the MeV software.

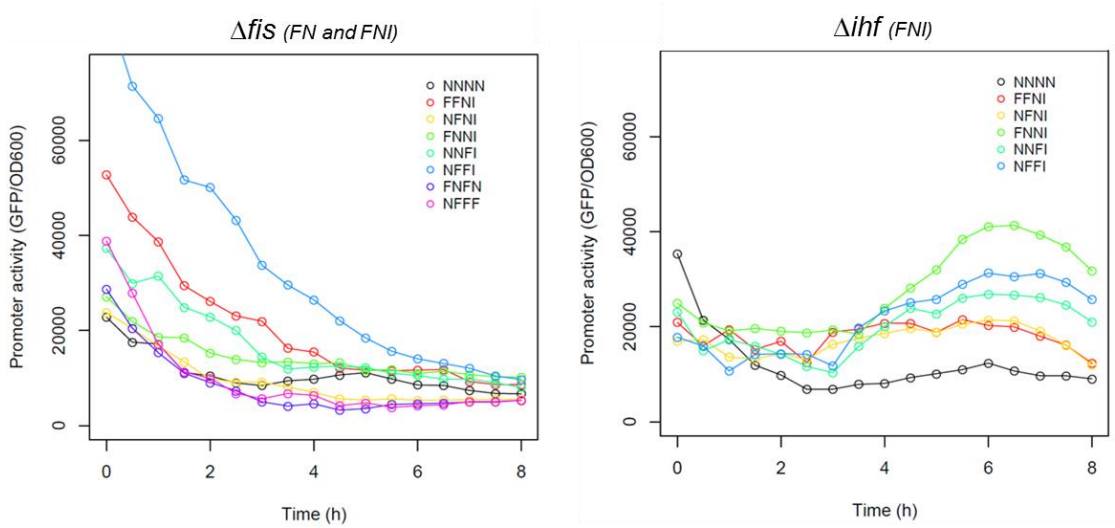


Figure 16. Analysis of Fis vs. Neg and Fis vs. IHF promoters in WT and Δ*fis* *E. coli* strains. Promoter activities are represented as the ratio of the GFP signal divided by the optical density at 600nm of the sample (GFP / OD600).

First, libraries containing only sequences for IHF and Neg (a subset with 16 promoters) were generated. The library was sequenced and transformed into wild and mutant strains for the fis gene of *E. coli* BW25113. After the plate reader experiment with both strains, one can observe the comparison of the expression profiles and identify the effects of IHF for each promoter (**Figure 17**).

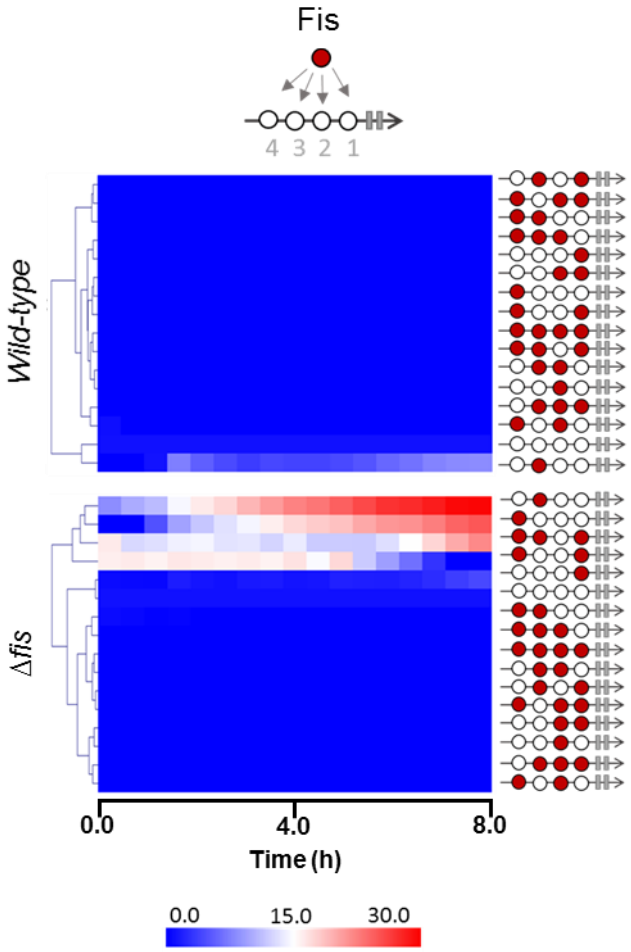


Figure 17. Analysis of Fis vs. Neg promoters in both WT and Δ*fis E. coli* strains. Promoter activities were measured over 8 hours, plotted as horizontal heatmaps of the GFP signal ratio divided by the optical density at 600nm of the sample (GFP / OD600), normalized by the negative control and transformed to log2 scale in order to facilitate the visualisation of subtle activities. The positions of Fis binding sequences (red circles) and inert sequences (white circles) in the synthetic promoters are represented in the right-hand corner of their respective expression profiles.

As can be seen, virtually no construct other than FNFF has GFP expression in the wild-type lineage. However, in the Δ*fis* strain, when the *cis*-regulatory element for Fis is located at positions 3 (NFNN), 4 (FNNN), 4,3,1 (FFNF) and 4,1 (FNNF), there is an increase in gene expression. For other architectures, the results are equal to the negative control (NNNN), with no observed gene expression. The analysis of the architectures that showed activity in the mutant for fis indicated the four promoters presented the 2-position free and at least one site occupied in the 1, 3 and 4 positions. Similarly, 8 of the 10 promoters that did not show significant activity even in the mutant have a cis element for Fis at position 2. These results closely resemble those found for the library of promoters containing IHF sites (Monteiro, Arruda and Silva-Rocha, 2017). The reason why the promoters containing the free 2-position have not shown activity yet is not understood and will be investigated in more detail in the

future. These results indicate that, following the strategy adopted in this project, the only functional promoters in this subset are the ones which are active in the wild-type or Δ*fis* strains, with Fis acting in a repressive manner when the same sequence shows an increase in the Δ*fis* strain expression in comparison with the wild-type strain.

Once the positional effect of the *cis*-regulatory elements for Fis was observed, the next step was to examine whether the IHF protein would have the same preference in modulating the promoter activity. For this, the subset of IHF vs. Neg previously analyzed in wild-type *E. coli* and Δ*ihf* strains (Monteiro, Arruda and Silva-Rocha, 2017) (**Figure 18**) was compared with that obtained for Fis vs. Neg (**Figure 17**).

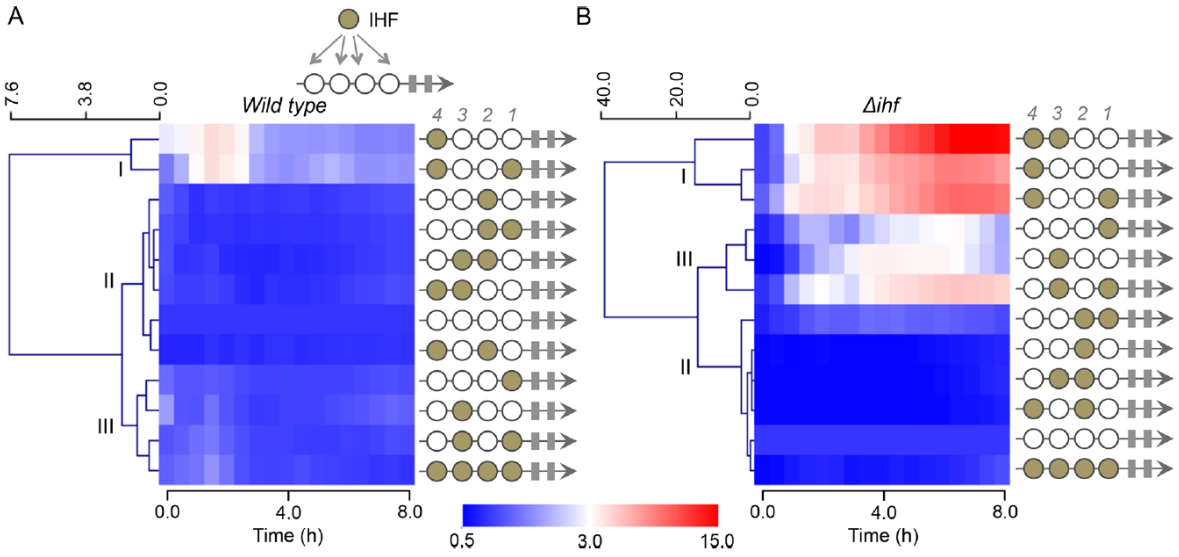


Figure 18. IHF motif enhanced promoter activity in *E. coli* Δ*ihf* strain. A subset of shuffled IHF and Neutral motif promoters were assayed in the wild-type and Δ*ihf* mutant strains and grouped according to their relative activity. Circles in beige represent the positions of IHF sites. **(A)** IHF vs Neutral motifs assayed in the wild-type strain. Synthetic promoters that showed higher promoter activities are clustered in group I, group II is formed of promoters with low activity, whereas group III is formed of promoters with intermediate promoter activity. **(B)** The same set of promoters were assayed in the *E. coli* Δ*ihf* mutant strain, highlighting that in the absence of IHF transcription factor, promoter activity was generally improved for the groups I and III. Relative promoter activity was measured for 8 h, calculated based on the Neutral full promoter, and displayed on an intensity scale from 0.0 to 15.0. Plots were calculated based on the average of three independent experiments. Retrieved from (Monteiro, Arruda and Silva-Rocha, 2017)

As can be seen in **Figure 18**, when the *cis*-regulatory element for IHF is located at positions 4 (INNN) and 4,1 (INNI), in the wild-type strain, an increase in gene expression occurs. However, for other architectures, the results are equal to the negative control (NNNN), with no increase of the gene expression. On the other hand, in the Δ*fis* strain the highest levels of expression

occur when the *cis*-regulatory element for IHF is located at positions 4,3 (IINN), 4 (INNN), 4,1 (INNI), 1 (NNNI), 3 (NIN), 3,1 (NINI) and 2,1 (NNII). The IHF repressive action on gene expression can be observed when the same sequence shows increased gene expression in the Δ*ihf* strain compared to the wild strain. It is important to highlight that some of the architectures that resulted in increased levels of GFP expression in Δ*fis* and Δ*ihf* strains have similarities in the position of their regulatory sites, as shown in **Figure 19**. This fact can be explained by the similar mechanism of action of these proteins that are functionally categorized as NAPS (Nucleoid-Associated-Proteins) (Dillon and Dorman, 2010; Dorman, 2013). In general, they act by modifying the DNA structure, generating folds and kinks that physically modulate RNA polymerase access to the promoter regions regulated by these proteins (Dillon and Dorman, 2010; Dorman, 2013).

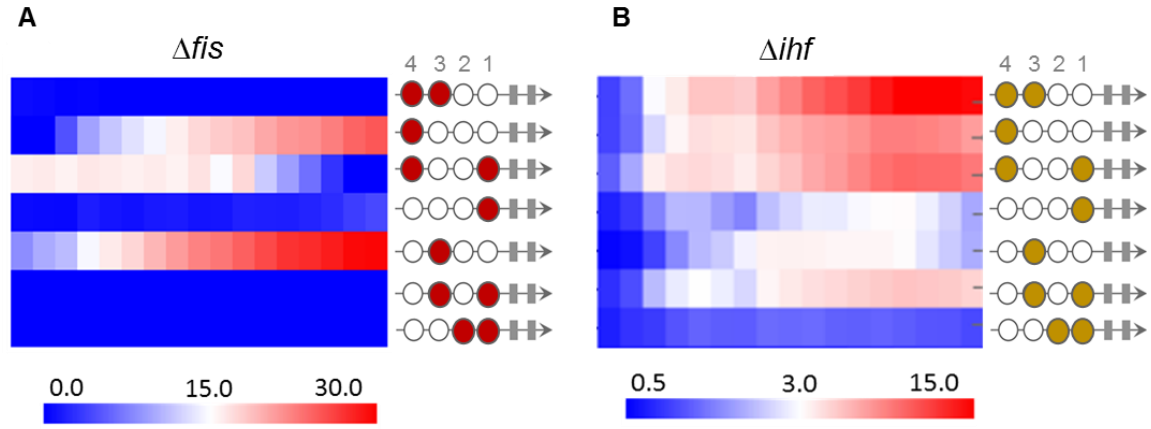


Figure 19. Comparison of similar promoter architectures for Fis and IHF in *E. coli* Δ*fis* and Δ*ihf* strains. Promoter activities were measured over 8 hours, plotted as horizontal heatmaps of the GFP signal ratio divided by the optical density at 600nm of the sample (GFP / OD600), normalized by the negative control and transformed to log2 scale in order to facilitate the visualisation of subtle activities. The positions of IHF or Fis binding sequences are indicated by red and beige circles, respectively and inert sequences (Neg) are indicated as white circles. **(A)** Expression profiles for Fis vs. Neg in Δ*fis*. **(B)** Expression profiles for IHF vs. Neg. in Δ*ihf*.

Taking into account the results for the individual regulators and the previous results for CRP and IHF libraries (Monteiro, Arruda and Silva-Rocha, 2017), five promoters with targeted topology, containing sequences for both IHF and Fis were investigated. It was observed in previous libraries containing CRP and IHF sequences that the synthetic promoters were only functional when they contained sequences for CRP at position 1 (the closest position to the promoter core) and that the addition of IHF sequences resulted in only two non-functional promoters (INNN and INNI). However, the effect of adding secondary CRP sequences to

functional IHF promoters generated unexpected behaviours in the expression dynamics of these promoters (Monteiro, Arruda and Silva-Rocha, 2017). Thus, 5 synthetic promoters were designed and generated in which IHF had a fixed position at 1 while the position and abundance of sites for Fis were shifted (**Figure 20**). As shown in **Figure 20**, the combination of sites in the promoters allowed the observation of different behaviours in GFP expression between the wild-type, Δ*fis* and Δ*ihf* strains. Surprisingly, although regulators maintained their original repressor function in the synthetic promoters, there was a delay in the expression response of Δ*ihf* strain in comparison to the Δ*fis* lineage. As shown in **Figure 20**, while none of the promoters tested showed significant activity in the wild-type strain, the five promoters showed detectable activity in the fis mutant mostly during the first 4 hours of growth. On the other hand, the same promoters showed an even greater activity in the mutant for ihf mostly at the later hours of growth. These results are in agreement with the mechanism of action of these regulators, since Fis is known to act mainly in the transition from lag phase to exponential while IHF coordinates the gene expression from exponential to stationary phase transcription (Azam *et al.*, 1999; Ishihama, 2010). In **Figure 21**, retrieved from (Dorman, 2013), it can be observed that the Fis protein is more active during the exponential phase of growth while IHF becomes more important late. These behaviours can be observed in **Figure 20**.

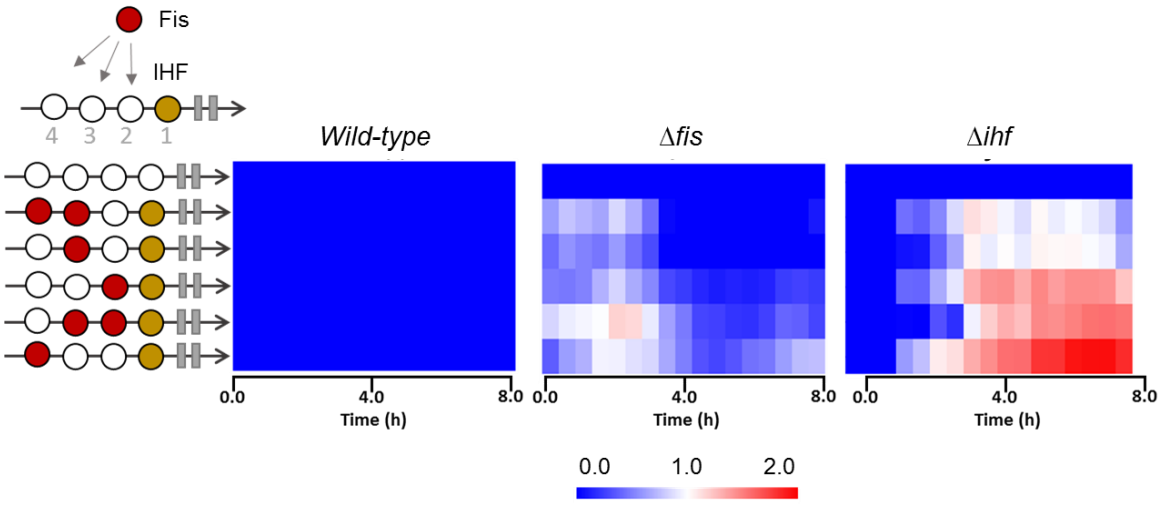


Figure 20. Analysis of IHF vs. Fis promoters in WT, Δ*fis* and Δ*ihf* *E. coli* strains. Promoter activities were measured over 8 hours, plotted as horizontal heatmaps of the GFP signal ratio divided by the optical density at 600nm of the sample (GFP / OD600), normalized by the negative control and transformed to log2 scale in order to facilitate the visualisation of subtle activities. The positions of IHF (beige circles), Fis binding (red circles) and inert (white circles) binding sequences in the synthetic promoters are represented in the left-hand corner of their respective expression profiles.

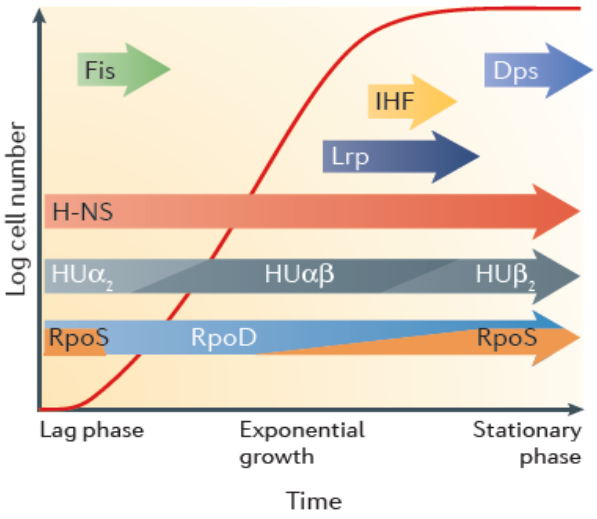


Figure 21. Activity pattern of the major regulators of *E. coli* functionally characterized as NAPS throughout the bacterial growth phases. It is important to highlight the higher Fis activity in the Lag phase and the higher IHF performance during the stationary phase. Such behaviours can also be observed in the synthetic promoters of **Figure 7**. Retrieved from (Dorman, 2013).

5. Discussion

Regulation of gene expression at the level of RNAP recruitment to target promoters is known to be a combinatorial mechanism where multiple transcriptional factors binding to target *cis*-regulatory elements and their interplay defines the timing and intensity of gene expression. This combinatorial control has been extensively described in bacteria and in single-celled and multicellular eukaryotes, and the so-called regulatory code is known to play a major role in the way living organisms develop and interact with the environment (Kinkhabwala and Guet, 2008; Raveh-Sadka *et al.*, 2012; Gama-Castro *et al.*, 2016). However, while classical approaches to understand this code are based on a case-by-case dissection of the *cis*-regulatory elements of particular genes, several studies have now described the systematic investigation of combinatorial promoters through the construction and evaluation of synthetic promoters built from *cis*-regulatory elements.

In this sense, Cox III and colleagues (III, 2008) constructed a library of synthetic promoters for two local activators (AraC and LuxR) and two local repressors (LacI and TetR) at three different promoter positions (upstream, downstream, or overlapping the core −35/−10 box). From this work, the authors described a number of rules for engineering combinatorial promoters for synthetic biology; for instance, activators were only efficient upstream of the core

whereas efficacy of repression was higher at the core and then at the downstream region, with only minor effects at the upstream position (Cox, Surette and Elowitz, 2007). However, this work only used local TFs, which are limited to a few natural targets and, thus, are not found in naturally complex promoter architectures as global regulators are. Moreover, the work by Cox III only explored a single binding site at the upstream promoter regions, which does not allow the investigation of combinatorial effects generated by *cis*-element arrangements and identities in this region.

Therefore, our work addresses a more realistic combinatorial situation by mimicking the manner in which promoters are organized naturally, and indeed, our result of *cis*-element mediated repression of gene expression has not been reported previously. The effect of promoter architecture in gene regulation has also been extensively investigated in single-celled eukaryotes such as yeast, with special interest in the work of Sharon and co-workers (Sharon *et al.*, 2014). In this study, the authors synthesized and analyzed using a high-throughput approach, thousands of different promoters for several TFs of Saccharomyces cerevisiae (Sharon *et al.*, 2014), thus allowing them to investigate the effect of number, position, and affinity of binding sites on gene expression. However, the fundamental difference between transcription initiation in prokaryotes and eukaryotes, due to the sophisticated process of chromatin remodelling required in the latter, makes it impossible to extrapolate the conclusions drawn by Sharon et al. to a bacterial organism. However, the approach used in this study was analogous to the approach used by Sharon et al., since we could inspect the effect of binding site multiplicity, location, and identity.

This result appeared in several promoter architectures tested here and would indicate that the DNA sequence itself was modulating gene expression. It has now been widely demonstrated that DNA can display an allosteric effect on TFs, where the binding of a protein to DNA changes the way this protein interacts with other TFs (Lefstin and Yamamoto, 1998; Chaires, 2008; Kim *et al.*, 2013) Moreover, another type of DNA-based allosteric event has been described where the binding of a protein to DNA can influence the binding of a second protein to an adjacent site independently of protein−protein interaction, and that this influence is transmitted through the DNA molecule (Lefstin and Yamamoto, 1998; Chaires, 2008).

Recently, an increasing number of reports have demonstrated that flanking DNA sequences can strongly affect the binding affinity of eukaryotic TFs for identical binding sites (Khoueiry *et al.*, 2010; Gordân *et al.*, 2013), thus explaining why in vitro and *in vivo* binding assays do not always correlate. In this process, these flanking sequences generate distortions in the local DNA shape that influences the way the TFs interacts with DNA, by altering the groove width and helical parameters of DNA (Gordân *et al.*, 2013). Though we could not find any report of this process influencing bacterial TFs, our results on synthetic complex promoters suggest that a similar process could influence the activity of bacterial promoters, thus explaining the intrinsic repressive activity of the CRP *cis*-element (independently of the presence of CRP protein) at some positions in promoters containing *cis*-elements for IHF (Monteiro, Arruda and Silva-Rocha, 2017). Our findings could thus be extended to naturally complex promoters and indicate that in those systems, not only would the nature of the TF recruited to the target promoter be imperative for gene expression, but also the *cis*-element itself could have a regulatory role in proximal sites. This evidence an unanticipated intrinsic complexity of natural bacterial promoters that should be considered both for synthetic biology projects as well as to understand the regulatory behavior of natural strains. Taken together, our results highlight the appearance of emergent properties in combinatorial control in bacteria, thus opening new venues for understanding combinatorial regulation in bacterial genes and open new venues that could be investigated in future studies.

It is important to highlight that some of the architectures that resulted in increased levels of GFP expression in Δ*fis* and Δ*ihf* strains have similarities in the position of their regulatory sites, as shown before in **Figure 19**. This fact can be explained by the similar mechanism of action of these proteins that are functionally categorized as NAPS (Nucleoid-Associated-Proteins) (Dillon and Dorman, 2010; Dorman, 2013). In general, they act by modifying the DNA structure, generating folds and loops that physically modulate RNA polymerase access to the promoter regions regulated by these proteins (Dillon and Dorman, 2010; Dorman, 2013).

Surprisingly, although regulators maintained their original repressor function in the synthetic promoters, there was a delay in the expression response of Δ*ihf* strain in comparison to the Δ*fis* lineage. In this context, the search for basements in the literature on these behaviours revealed that the synthetic promoters follow the performance profiles of the Fis and IHF proteins during the stages of *E. coli* growth. In **Figure 21**, retrieved from (Dorman, 2013), it can be observed

that the Fis protein is more active during the exponential phase of growth while IHF becomes more important late. These behaviours can be observed in **Figure 20**.

6. Conclusions

By constructing and analyzing 21 synthetic promoters with the sequences for Fis and IHF and comparing them with others previously constructed in our laboratory, it was possible to observe that:

i) in synthetic promoters with only one type of transcription factor binding sites - Fis vs. Neg or IHF vs. Neg -, the regulatory architectures leading to increased expression in their respective mutants are similar. This fact could be explained by the shared function and mechanism of action between these proteins - NAPS (Dillon and Dorman, 2010; Dorman, 2013) - that act normally as repressors of the transcription through the generation of conformational changes in the DNA;

ii) although the functional architectures for the libraries mentioned above are similar, it is possible to notice that there are particular behaviours in each of them that depend on the intrinsic qualities of each regulator involved;

iii) the patterns observed for the individual sequence libraries are not able to explain the expression patterns obtained by combining both the Fis and IHF sequences in complex synthetic promoters. This epistatic phenomenon, in which emergent and unpredictable behaviours arise from the combination of biological parts with known behaviours, appears to be widely distributed in molecular systems and with an important evolutionary role, especially in regulatory systems (Loewe, 2009; Lagator *et al.*, 2016; Aguilar-Rodríguez, Payne and Wagner, 2017; Monteiro, Arruda and Silva-Rocha, 2017).

iv) the study of new architectures of complex promoters allows the generation of information essential to the understanding of the emergent phenomena mentioned above, making the engineering process of regulatory elements more rational and predictable.

Chapter II

Using *in silico* approaches for understanding the evolution of transcription factor binding sites in *Escherichia coli*

1. Specific Background

1.1 A method for studying the evolution of TFBSs in bacterial systems

In the current chapter, we aim to explore the question of how TFBSs might evolve in bacteria and which mechanisms could underlie both regulatory innovation and TFBS diversity. For this purpose, we have assumed five general evolution models based on literature regarding TFBS loss, gain, shift, overlap and promiscuity (see **Figure 22**) (Babu, Balaji and Aravind, 2007; Wolf, 2014; Payne and Wagner, 2015; Friedlander *et al.*, 2016, 2017). Here, we have selected the overlap and promiscuity models as novel frameworks for regulatory evolution (Friedlander *et al.*, 2016, 2017; Rowland *et al.*, 2017). Although these two models have been poorly explored in the literature, preliminary results from our group have shown that using a rather simple genetic algorithm, it was possible to computationally generate TFBSs sequences which could be recognized by up to three TFs simultaneously *in vivo* (Guazzaroni and Silva-Rocha, 2014). Thus, this result has lead us to hypothesize TFBSs promiscuity might be a rather common feature in regulatory systems and could provide a new conceptual framework for understanding and engineering transcriptional behaviours.

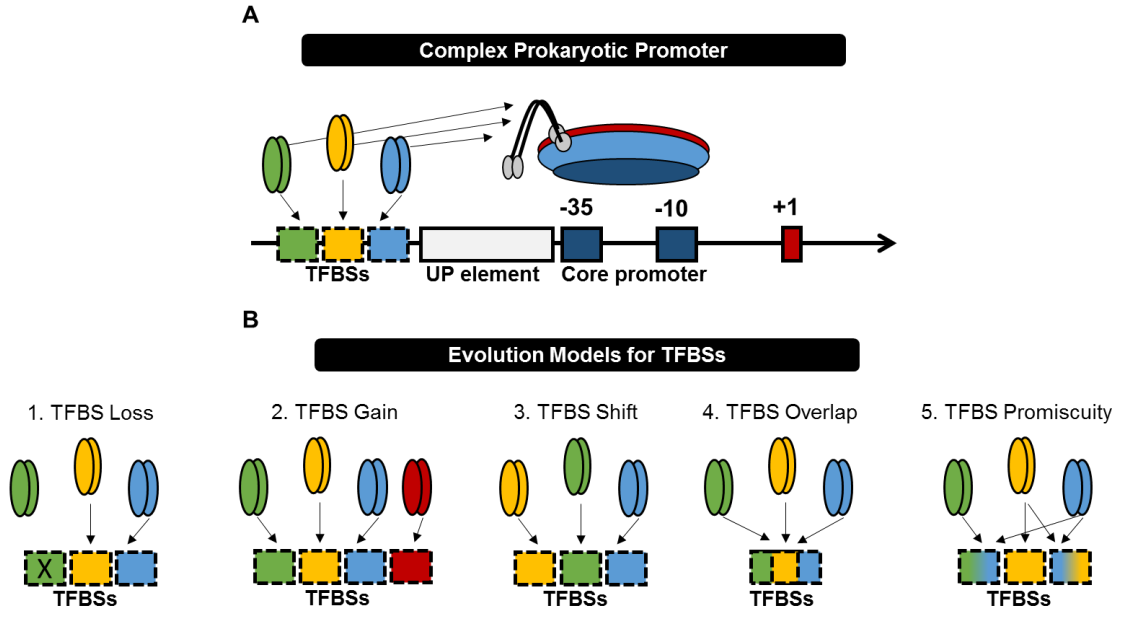


Figure 22. A general complex prokaryotic promoter and the proposed evolution models for *cis*-regulatory elements. (A) A general model of prokaryotic complex promoter with three TFBSs for specific TFs, UP element and core promoter region recognized by the sigma/RNAPol holoenzyme. **(B)** Five models for the evolution of TFBSs. **(1)** loss of TFBs by random mutations; **(2)** gain of a novel TFBS by accumulation of mutations; **(3)** TFBS shift due to DNA translocations or mutations; **(4 and 5)** TFBSs overlap and promiscuity, in which a single regulatory region has enough information for being recognized by multiple TFs.

Thus, we have adopted an *in silico* approach for analysing the evolution of *E. coli* TFBSs related to global transcriptional regulators. We have used experimentally validated datasets from prokaryotic databases (Gama-Castro *et al.*, 2016; Ishihama, Shimada and Yamazaki, 2016) to generate both mutational networks and adaptive landscapes for the TFBSs genotypes related to CRP, the most connected TF in *E. coli*. Every node in the network (TFBS sequence) is connected by a vertex to another node by a small number of mutations and each mutation may elicit a change in the binding affinity of the studied TF, which will be considered as the fitness function in this model. We have used the Position Weight Matrices (PWMs) (Stormo and Hartzell, 1989; Stormo, 2000) scores based on CRP, Fis and IHF sites as proxies for predicting the binding affinities of each TFBS genotype for each of the three TFs. Depending on the environmental conditions, natural selection may act favouring a specific binding affinity, and it might be reflected in the regulated gene expression behaviour. In this context, we might consider two general models for the evolution of a single TFBS (see **Figure 23**): an independent model and a connected model. In the independent model, the genotype spaces for the evolution of a single TFBS do not overlap, posing more boundaries for the rise of innovations – it is more difficult for a sequence initially recognized by a specific TF to become recognized by another TF after a few mutational steps -. Furthermore, this model allows higher rates of transcriptional function loss (mutations that disrupt functional TFBSs), reducing the system's robustness – its ability to keep his regulatory functions in face of random mutations. On the other hand, in the connected model, the genotype spaces of TFBS overlap and a single sequence could be potentially recognized by more than one global TF. The ultimate consequence of this model is the higher rates for the rise of transcriptional innovations.

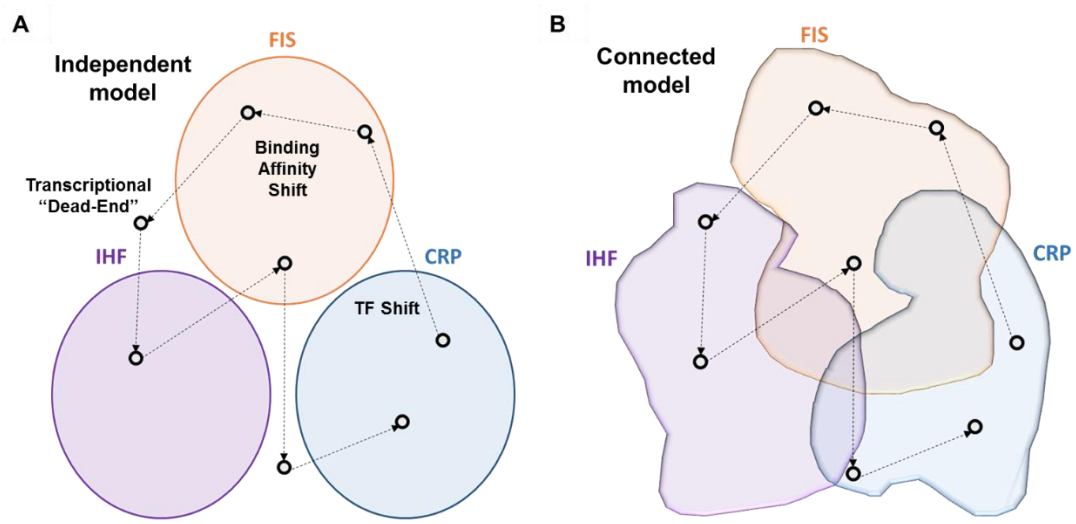


Figure 23. Two models for the evolution of TFBSs proposed in this work. (A) In the independent model, a single genotype, represented by a single node, can mutate to another genotype (dashed arrow) in a non-overlapped space of TFBSs classes, with higher chances for losing its regulatory function by falling into a "transcriptional dead-end". **(B)** On the other hand, the connected model considers that the genotype spaces for TFBSs of different classes are overlapped, leading to the simultaneous walk of a single sequence between all overlapped spaces. In this model, innovation may rise more rapidly than in the former.

1.2 The paradox of global TFBSs diversity in *E. coli*

When considering random mutational networks (the whole set of possible a variants of a DNA sequence), one might expect that a substantial fraction of the artificially generated genotypes for TFBSs would be found in a prokaryotic genome. However, by exploring a large amount of data deposited on prokaryotic databases (e.g. RegulonDB (Gama-Castro *et al.*, 2016), DOOR (Mao *et al.*, 2014), PRODORIC (Münch *et al.*, 2003) etc.) it was possible to conclude this is not the case. On the contrary, just a very small fraction of the whole set of possible genotypes is present in nature – **Table 3**. One explanation for this observation could be that the artificial mutational network of TF binding sites has many genotypes with very low PWM scores, whereas natural networks usually have a well-established cut-off value for binding affinities (i.e. for CRP, in *E. coli*, none of the reported TFs binding sites presented a binding affinity lower than 40% of the maximum PWM score). However, despite the exclusion of sites with low scores from the analysis, a substantial amount of artificial genotypes absent in natural systems remains. Even though those artificial sites might be redundant regarding their PWM scores (many genotypes with the same scores), not all of them are present in bacterial genomes and there seem to be a selection for specific genotypes among all the redundant possibilities (see **Table 3**).

Table 3. The three most Global TFs of *E. coli* and their respective natural and potential number of TFBSs sequences:

Global Regulators	***TF Binding Sites (% of the genomic total)*** *(Gama-Castro et al., 2016)*	***Potential Diversity (Combinatorics)***
CRP	280 (12,1%)	4^{19}
FIS	228 (9,9%)	4^{16}
IHF	102 (4,4%)	4^{15}

1.3 A single TFBS can be designed *in silico* to be recognized by three TFs *in vivo*

Recognition of *cis*-regulatory elements by transcription factors at target promoters is crucial to gene regulation in bacteria. In this process, binding of TFs to their cognate sequences depends on a set of physical interactions between these proteins and specific nucleotides in the operator region. Previously, it has been shown that *in silico* optimization algorithms are able to generate short sequences that are recognized by two different TFs of *E. coli*, namely, CRP and IHF, thus generating an AND logic gate (Guazzaroni and Silva-Rocha, 2014). In a subsequent work (Amores, Guazzaroni and Silva-Rocha, 2015), this approach was expanded in order to engineer DNA sequences that can be simultaneously recognized by three unrelated TFs (CRP, IHF, and Fis) (**Figure 24**). The results demonstrated the potential of *in silico* strategies in bacterial synthetic promoter engineering and how small modifications in *cis*-regulatory elements can drastically affect the final logic of the resulting promoter. Thus, we have extrapolated this conceptual background to the current work with the following assumption: if it is possible to design single TFBSs for many TFs in a straightforward manner, we might assume the same process could have happened in natural living systems along billions of years of evolutionary processes.

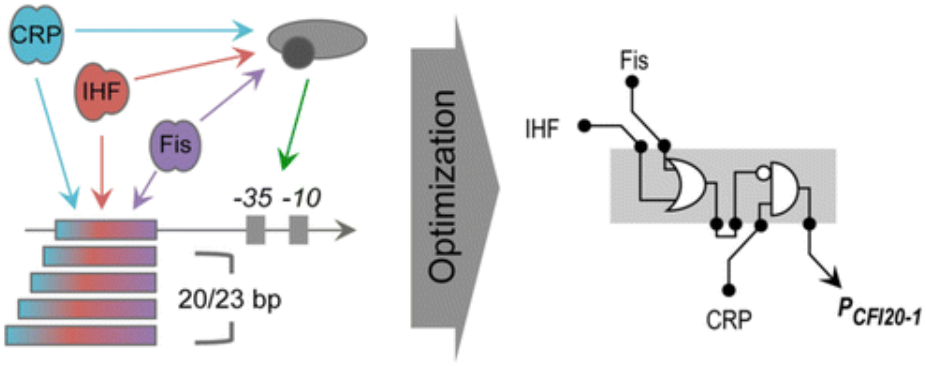


Figure 24. Expanding the optimization algorithm to construct synthetic *cis*-elements and the resulting logic of the system. Starting from a natural system controlled by a single input, the optimization algorithm was able to generate only short sequences efficiently recognized by two TFs (either CRP and IHF, or Fis and IHF). This ensures an AND gate behaviour since the location of the binding sites relative to the - 35/-10 boxes is preserved. In order to generate functional promoters regulated by three TFs, the algorithm needs to consider longer sequences (in this case, 20 and 22 bp) to efficiently accommodate the three binding sites. However, in this case, this seems to generate binding sites placed distantly from optimal activation positions, generating repressive interactions at the resulting promoter. Retrieved from (Amores, Guazzaroni and Silva-Rocha, 2015).

1.4 Synthetic bacterial complex promoters generate novel regulatory outputs

Previous experimental results from our group have highlighted novel principles of design for promoter engineering. Combining TFBSs for Fis and IHF (this work) or CRP and IHF (Monteiro, Arruda and Silva-Rocha, 2017) in synthetic regulatory elements, we were able to describe emergent expression profiles by modulating promoter architecture - shifts in the TFBSs position and frequency in complex promoters (**Figure 20** and **Figure 25**). In this context, we wanted to explore how complexity in natural promoters might evolve in the context of multiple TFBSs. Rydenfelt *et al*., 2014, (Rydenfelt *et al.*, 2014) have provided an extensive analysis on the architecture of *E. coli* promoters based on information gathered from the RegulonDB database and tried to couple this information to gene expression patterns. Combining all the background information provided above, we are focused on understanding the evolution of transcriptional logic under the perspective of single TFBSs, which factors constraint their natural diversity and how evolution "walks" on their sequence landscapes.

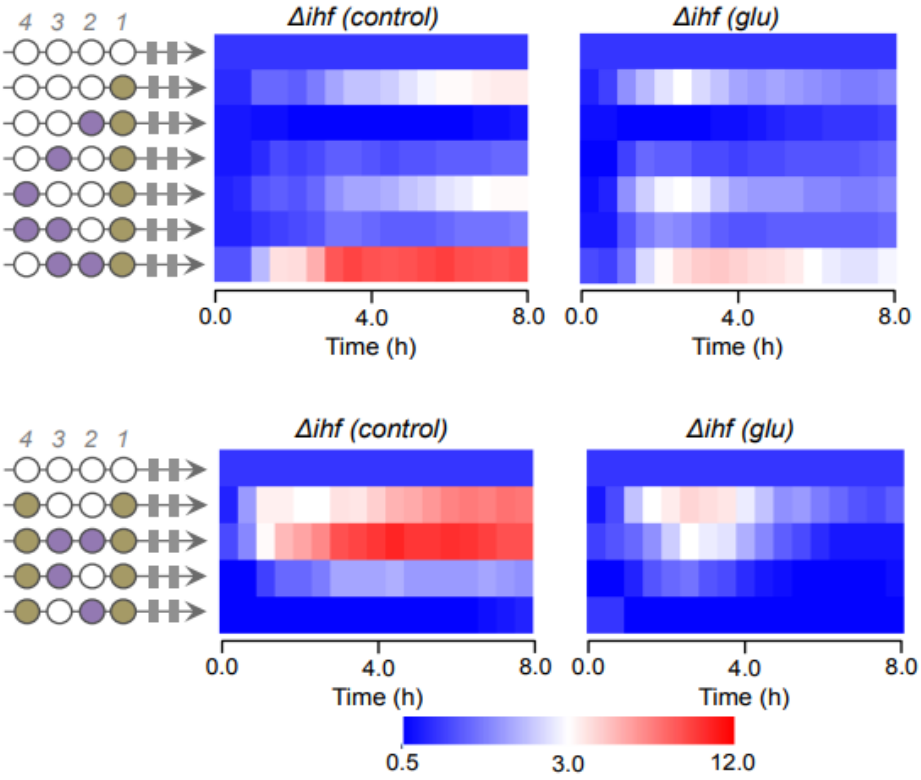


Figure 25. A heatmap representation of GFP expression profiles for different architectures of a complex promoter during 8 hours. White circles represent negative sequences (inert sequences), purple circles represent putative TFBSs for CRP and dark-yellow circles represent putative sites for IHF. Changing the promoter architecture (position of the TFBSs) lead to the rise of emergent expression profiles. The heatmap ranges from blue to red in a crescent scale. Retrieved from (Monteiro, Arruda and Silva-Rocha, 2017)

2. Objectives

General Objective:

To develop a general model of how regulatory innovations might rise in bacterial promoters by evaluating the evolutionary constraints that rule TFBSs diversity in *E. coli* through the combination of *in silico* and *in vivo* approaches. The model should be able to provide resourceful information regarding both transcriptional evolution and principles of design for promoter engineering.

Specific Objectives:

1. To study the natural mutational networks of TFBSs for the three most global TFs in *E. coli* – CRP, Fis and IHF
2. To apply the Position-Weight-Matrix method to access whether TFBS sequences can be recognized by more than its cognate TF
3. To study the navigability of evolution into natural and artificial mutational networks as fitness landscapes in order to explore what kind of constraints would be ruling the TFBSs diversity in *E. coli*

3. Materials and Methods

Below, we will briefly describe the overall computational methods selected for data collection, filtering and analysis in this chapter:

Extraction of TFBSs sequences from the RegulonDB online platform:

Extraction of experimentally validated TFBSs datasets from online databases (such as *RegulonDB*) (Gama-Castro *et al.*, 2016) for the three *E. coli* global regulators (Fis, CRP and IHF). A total of 370 sequences were extracted for CRP, 267 sequences for Fis and 119 sequences for IHF.

Application of the PWM method for addressing scores for CRP, Fis and IHF for each TFBS sequence:

Position Weight Matrices (PWMs) (Stormo and Hartzell, 1989; Stormo, 2000) were used to calculate the multidimensional scores for each TFBS. The PWMs for CRP, Fis and IHF were retrieved from Amores et al., 2015 (Amores, Guazzaroni and Silva-Rocha, 2015) and applied within *ad-hoc Perl* and *Python* scripts for providing scores for each individual sequence. The computational strategy was the following: CRP sequences retrieved from the database had a fixed length of 22bp, thus the CRP PWM (19bp) was used to find the best score in each sequence (the best one considering both forward and reverse strands for each sequence). The region of the original sequence where the best CRP score was found was then fixed and selected for further analysis by the Fis (16 bp) and IHF (15bp) PWMs, allowing all scores to be within the CRP best score (**Figure 26**). We have also generated random artificial networks as control groups, representing a large fraction of the possible mutational variants of a single sequence of n-length (length will be the same as the PWM length) for establishing biologically relevant cut-offs for each PWM.

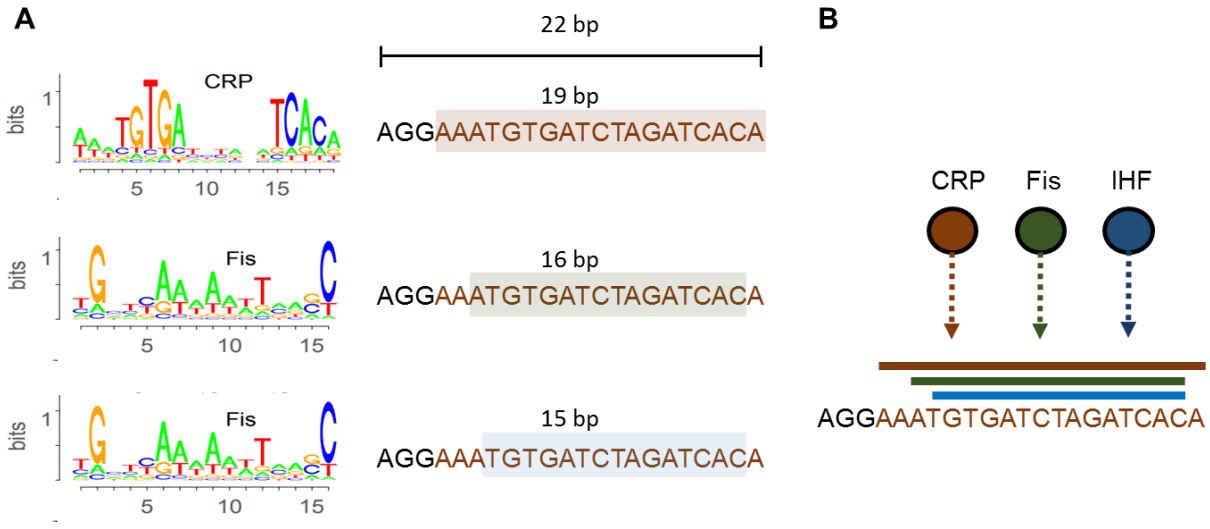


Figure 26. Analysis of TFBSs sequences through multiple PWMs. (A) A single TFBS sequence experimentally validated as belonging to the CRP regulon (right side) is subjected to three PWMs represented as WebLogos (Schneider and Stephens, 1990; Crooks *et al.*, 2004) (left side), one for CRP (top), one for Fis (Middle) and one for IHF (bottom). The crucial step in this analysis is that the best score for CRP should be calculated first as it has the largest consensus sequence (19bp). Once the best score for CRP is found (top-right, in brown), the respective sequence is used for finding the best scores for Fis (middle-right, in green) and IHF (bottom-right, in blue). **(B)** All sequences are then comprised within the best CRP score sequence and can be potentially recognized by the three TFs.

Generation of natural mutational networks with TFBSs from CRP:

We have used the experimentally validated datasets with previously calculated PWM scores to generate both mutational networks and adaptive landscapes for the TFBSs genotypes related to CRP, Fis and IHF the most connected TFs in *E. coli*. Here, we will use CRP TFBSs as a case of study. As sequences had a fixed size of 22 bp, they could be directly compared through an *ad-hoc* algorithm in *Perl* for calculating the Hamming distance between all sequences. Every node in the network is the representation of a single TFBS sequence which is connected by a vertex to another node by a small number of mutations and each mutation may elicit a change in the binding affinity of the studied TF. Here, binding affinities will be considered as the fitness function in this model. As explained before, we have used the Position Weight Matrices (PWMs) (Stormo and Hartzell, 1989; Stormo, 2000) scores based on CRP, Fis and IHF sites as proxies for predicting the binding affinities of each TFBS genotype for each of the three TFs (see **Figure 27**). All mutational networks were built and visualised using the *Gephi* and *Cytoscape* software (Shannon, 2003; Bastian, Heymann and Jacomy, 2009).

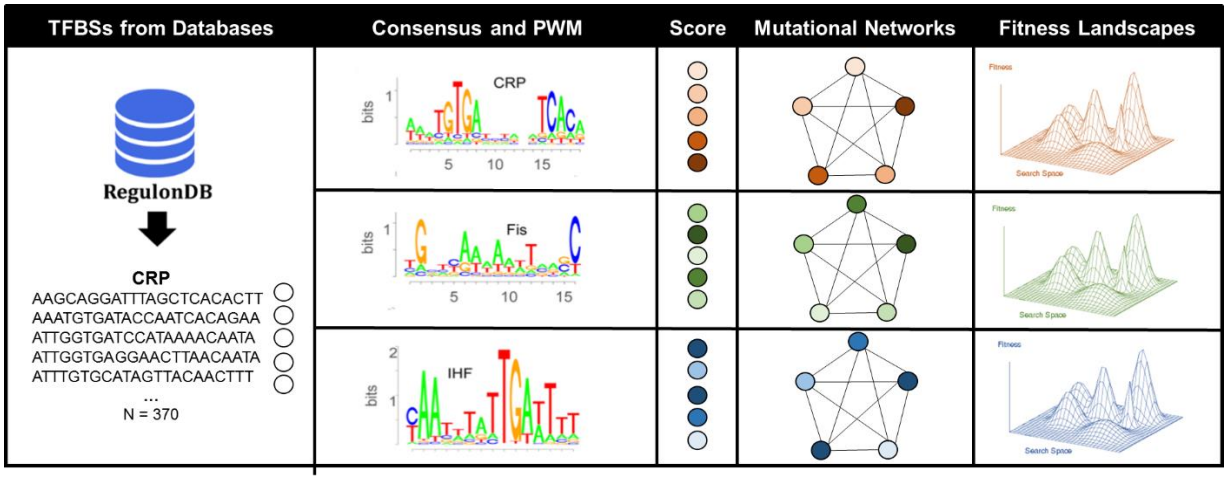


Figure 27. A workflow for generating multidimensional mutational networks. A single set of TFBSs related to a specific TF – genotype space represented as circles - is extracted from online databases and evaluated simultaneously for three different position-weight matrices, which will result in three different scores for every single TFBs, analysed – colour coded. Then, the TFBSs genotypes are organized in mutational networks depending on their mutational distances – the number of differences represented by vertex thickness. Lastly, the resulting network will be analysed as the representation of a multidimensional landscape in which each node has three different scores.

Generation of artificial mutational networks with TFBSs from CRP:

Analysis of navigability inside the natural mutational networks was achieved by the generation of artificial networks between selected nodes (for example, the generation of all possible variants between a peak node with a high score and a bottom node with a low score). Two sequences with contrasting scores for CRP were chosen and their Hamming distance was calculated. *Ad-hoc* scripts in *Perl* and *Python* were generated in order to create all possible variants between these two sequences (n^4 possibilities, being n the Hamming distance value), considering only the mutated positions between them. The shortest path algorithm was then applied using the *Cytoscape* software (Shannon, 2003) for finding the shortest paths between the two natural nodes in the artificial network. A general scheme of the adopted strategy can be seen in **Figure 28**.

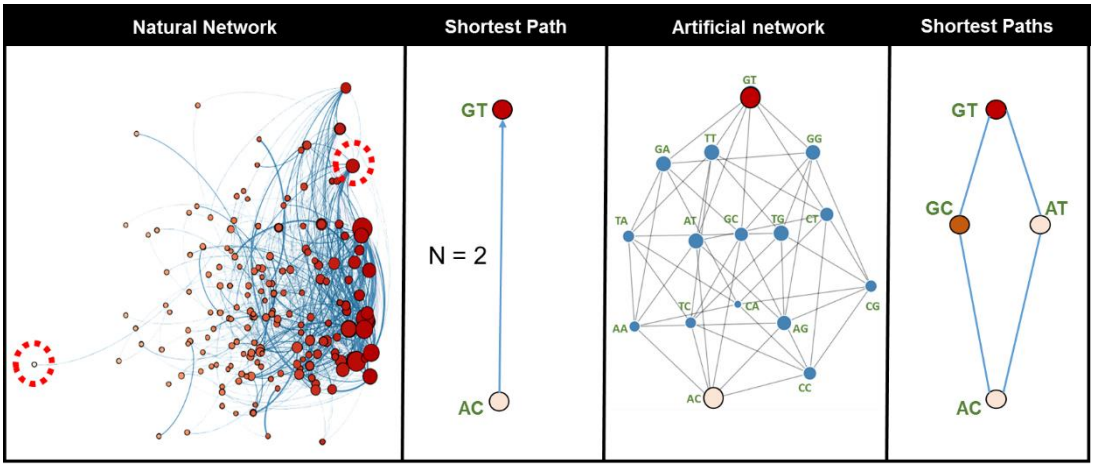


Figure 28. From natural to artificial networks. After generating the natural mutational networks for CRP sequences, the next step was to evaluate the number of potential innovative paths evolution could "navigate" in the sequence space from a natural sequence with low PWM score to a natural sequence with high PWM score. As seen in the illustrative figure A, from left to right: first the two sequences were selected from analysed data, then the number of mutations between then was used to generate an artificial network with all the possible variants that could appear in the evolutionary trajectory between the two sequences. Lastly, scores for CRP, Fis and IHF were calculated for all these sequences and the shortest paths were obtained for further analysis.

4. Results

In this chapter, it was possible to explore the working hypothesis of TFBS promiscuity or regulatory crosstalk by *in silico* analysis of sequence datasets for Fis, CRP and IHF retrieved from the RegulonDB database. The first part of our results is focused on establishing statistical information regarding the distribution of scores while the second part is more focused on understanding evolution's navigability over the natural and artificial adaptive landscapes of these sequences.

4.1 Statistics of natural sequences

Firstly, we have analyzed the distribution of scores for sequences from the CRP (370 sequences), Fis (267 sequences) and IHF (119 sequences) subsets. We have analysed all the 756 sequences comprised by the three groups and plotted them in a 3D dot plot for observing the distribution of each sequence in a 3-dimensional space represented by CRP, Fis and IHF scores (**Figure 29**). In addition, we have found that when subjected to each of the three different PWMs, the mean and the median of the scores for the total set was around 0.6-0.65 regardless the PWM chosen, with the distribution of scores for CRP being more homogeneous than that observed for Fis and IHF scores (**Figure 30**).

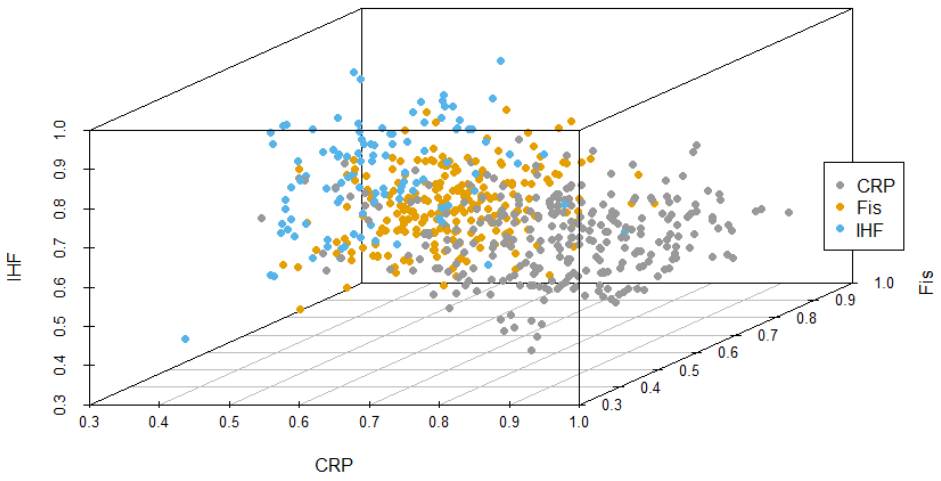


Figure 29. Distribution of CRP, Fis and IHF scores for all the sequence sets extracted from RegulonDB. The 3D dot plot simultaneously shows scores for CRP (x-axis), Fis (y-axis) and IHF (z-axis) for each analysed sequence (n = 756). The data points are coloured according to their original datasets (grey

dots are sequences from the RegulonDB associated with CRP, yellow dots are associated with Fis and blue dots with IHF).

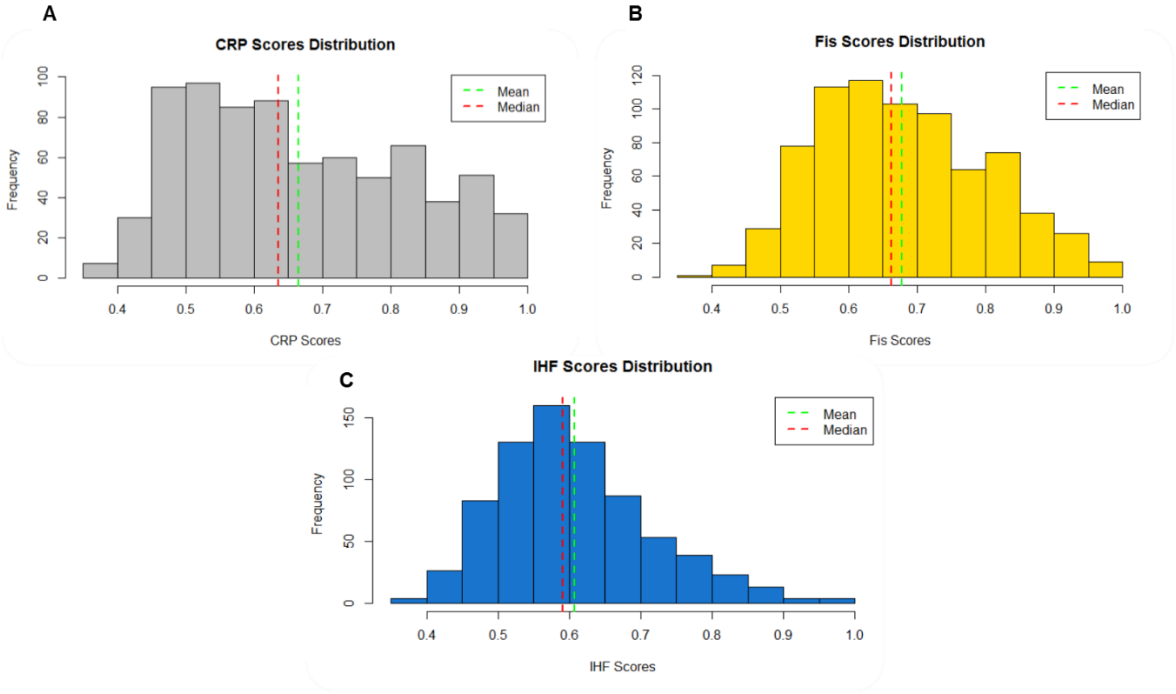


Figure 30. Distribution of scores for all TFBSs. The histograms represent the distribution of scores for all sequences (n=756) when individually subjected to **(A)** CRP PWM, **(B)** Fis PWM and **(C)** IHF PWM. Mean and median are represented by green and red dashed lines, respectively.

Then, we have focused on the CRP subset of TFBSs (370 sequences), analyzing the distribution of CRP, Fis and IHF scores in this subset. We have found that, as expected, this subset presented a higher mean and median for scores related to the CRP PWM than for the other position-weight-matrices (**Figure 31**). The same has happened to the other subsets and their respective PWM scores (for example, sequences from the Fis subset were enriched for higher Fis scores rather than for scores related to CRP and IHF PWMs)

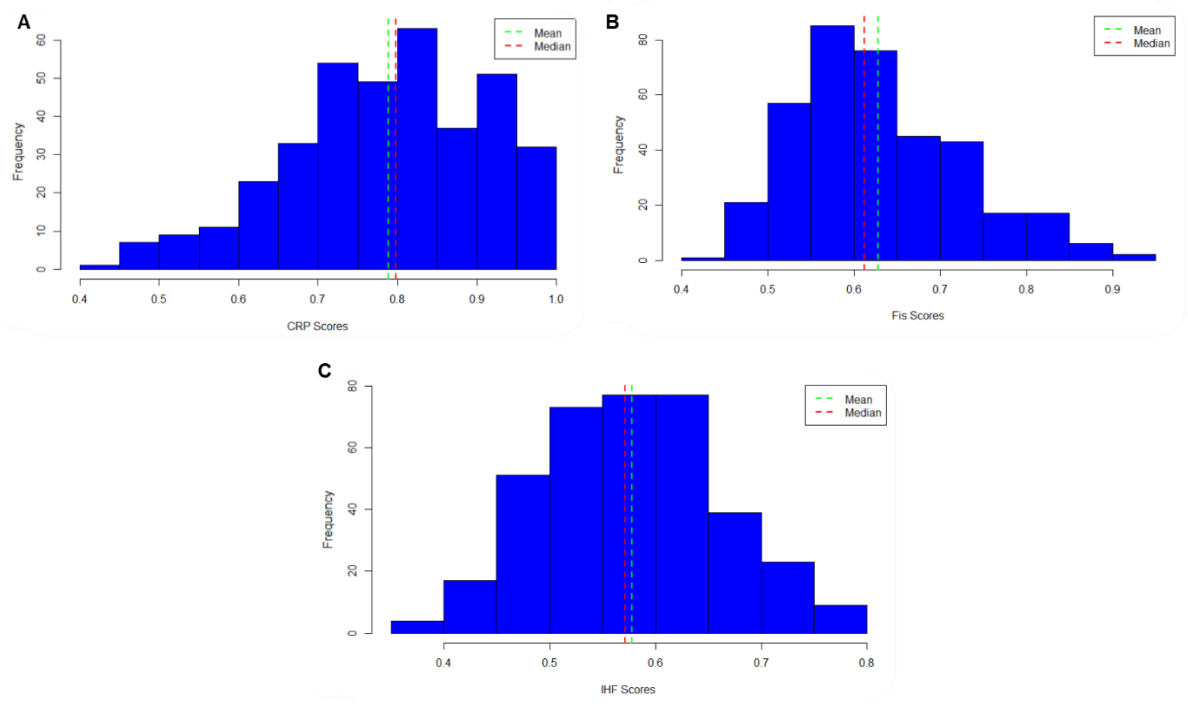


Figure 31. Distribution of Scores in the CRP TFBS subset. The histograms show the distribution of scores for the subset of sequences addressed to CRP in the RegulonDB (n = 370). These sequences were subjected to PWM analysis for **(A)** CRP, **(B)** Fis and **(C)** IHF. Mean and median are represented by green and red dashed lines, respectively.

In order to establish the biologically relevant scores for our analysis, it was important to compare the scores generated for the experimentally validated dataset from RegulonDB to a randomly generated dataset of 100,000 sequences. This comparison has allowed us to set cutoff values for scores from the PWMs which will help us to discriminate between potential biologically relevant and irrelevant score values. In this context, we have created a random dataset and calculated scores for CRP, Fis and IHF PWMs. The distribution of scores was directly compared to the distribution of natural scores for the TFBSs subsets (**Figure 31**). For example, the distribution of scores calculated for CRP in the random dataset was directly compared to the distribution obtained from the natural CRP dataset (370 sequences), showing an overlap of scores from 0.4 to 0.8. Here, we can see that sequences with CRP scores ranging from 0.6-1.0 should be positives with some margin for false-positives around the value of 0.6 (**Figure 32 A**).

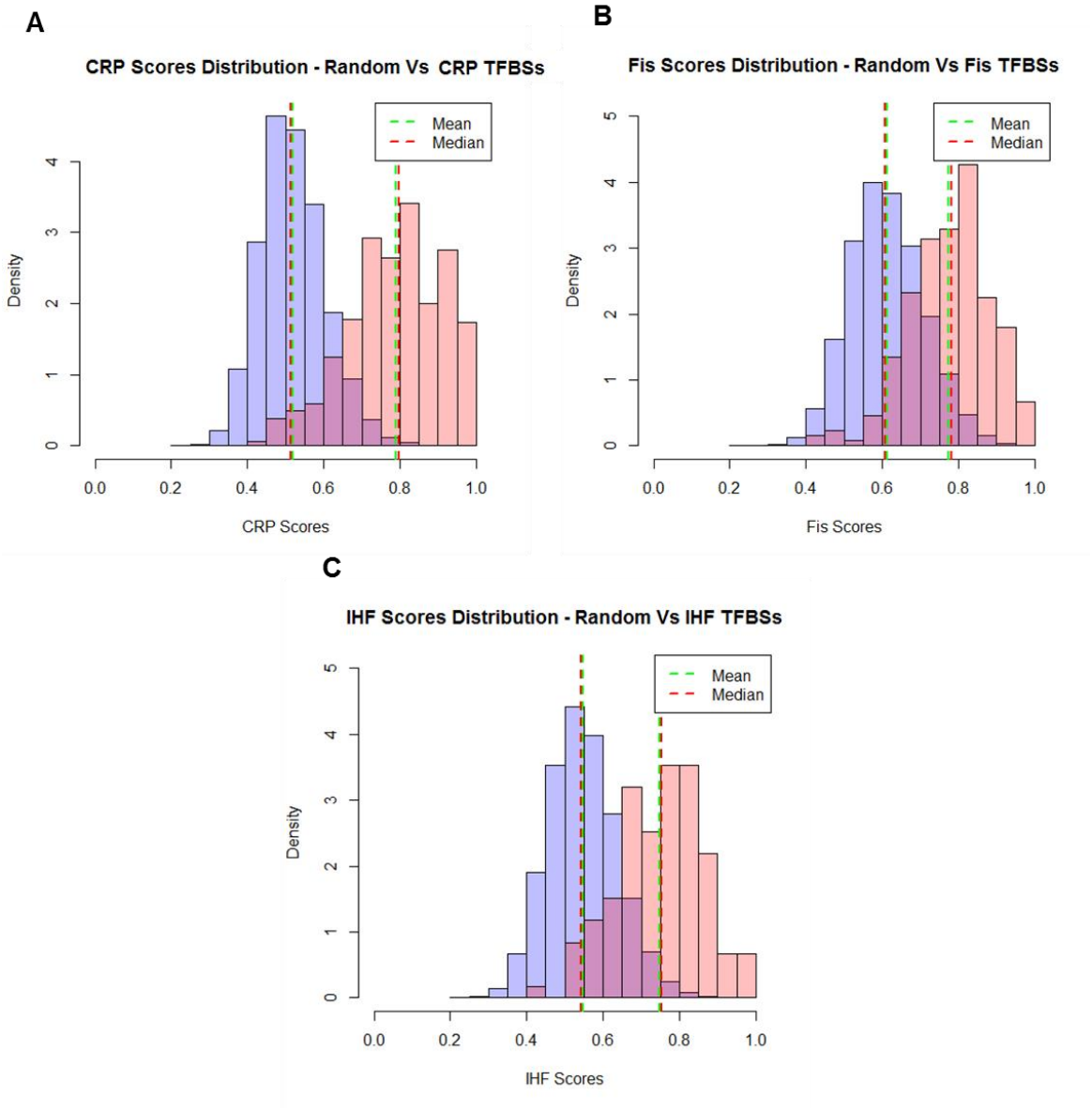


Figure 32. Cut-off values for all three PWM used in this work. In order to evaluate the cut-offs which would separate random values from biologically relevant ones, we have generated a dataset of 100,000 random sequences which were subjected to PWM analysis for CRP, Fis and IHF. In each graphic, the blue bars represent densities of the random dataset while the red bars represent densities for the RegulonDB dataset. The intersection between both datasets is represented as dark-pink bars and they depict the cut-off values for our analysis. Mean and median are represented by green and red dashed lines, respectively. **(A)** Sequences subjected to CRP PWM analysis. Left: random sequences, Middle: intersection, Right: sequence set associated with CRP in Regulon DB, **(B)** Sequences subjected to Fis PWM analysis. Left: random sequences, Middle: intersection, Right: sequence set associated with Fis in Regulon DB and **(C)** Sequences subjected to IHF PWM analysis. Left: random sequences, Middle: intersection, Right: sequence set associated with IHF in Regulon DB

After establishing the cutoff values, we have also checked if there were any correlations between the scores for our sequences. We wanted to check, for example, if variations between scores from a specific PWM would be correlated to variations in scores for a different PWM. Thus, we have analysed all sequences in 2D dot plots in which dimensions were scores for

different PWMs (CRP, Fis and IHF). The coefficient of determination (R^2) calculated for each pair of variables has shown no correlation between scores (**Figure 33**).

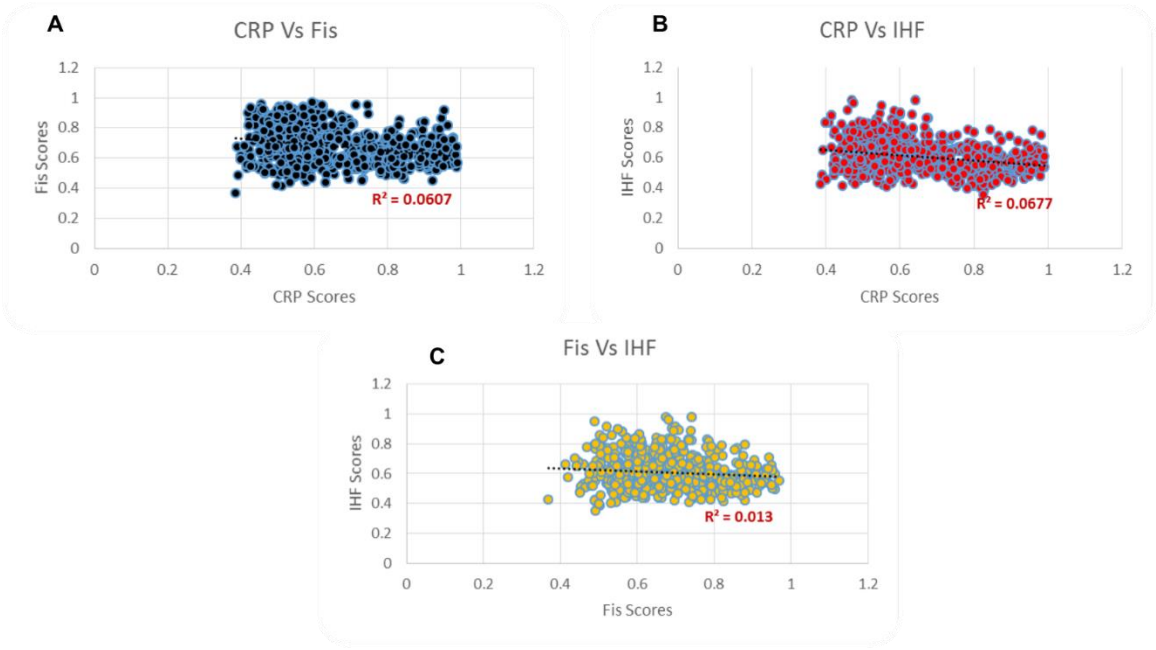


Figure 33. Testing the correlation between scores in the CRP subset. In order to evaluate if there were any correlations between scores for CRP, Fis and IHF in the sequence set belonging to CRP, we have calculated the coefficient of determination (the square of the Pearson correlation coefficient) for each pair of variables. **(A)** CRP Vs Fis scores, **(B)** CRP Vs IHF scores, **(C)** Fis Vs IHF scores.

4.2 Analysing natural and artificial mutational networks for CRP TFBSs

After analysing the statistics for the selected sequences and PWMs, we have focused on establishing a method for understanding how evolution would navigate inside the mutational networks obtained (see **Figure 34**). The generation of a natural mutational network from the CRP dataset, consisting of 269 nodes interconnected as the giant component was subjected to further analysis. Each node in the natural network was connected by an average of 4 mutations. In this scenario, it was impossible to provide a landscape with a reasonable size for understanding general properties such as global and local peaks, valleys and ruggedness, which ultimately comprise the adaptive landscape where evolution would navigate and explore innovation.

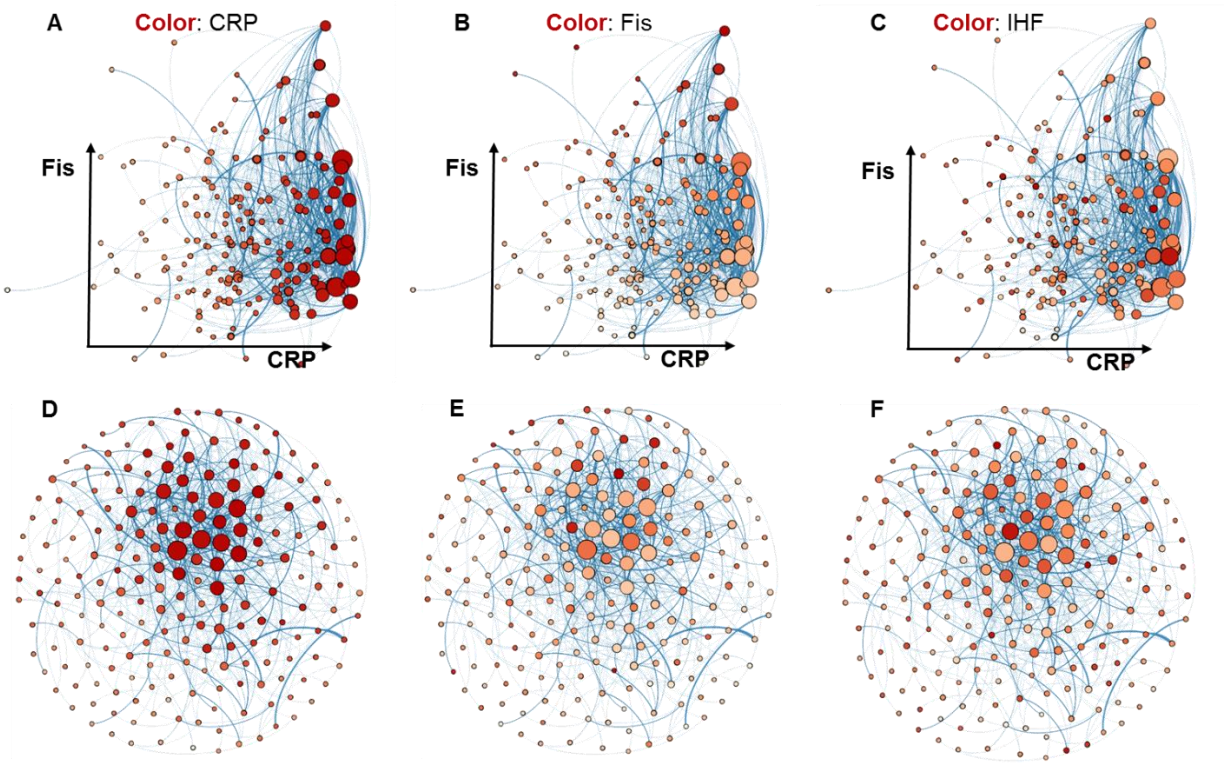


Figure 34. The natural mutational network of TFBSs genotypes experimentally associated to CRP in *E. coli*. Each node represents a genotype and each vertex represents the mutational distance between nodes – the number of differences. All figures represent the same mutational network. Top figures **(A, B and C)** are organized in an x-y plan in which the x-axis represents crescent CRP scores and the y-axis represents crescent Fis scores (e.g. nodes in the top-right of each network would represent genotypes which have high scores for both Fis and CRP). Although all figures represent the same network, each one is coloured differently, as a function of each of the three different PWM tested – for CRP, Fis and IHF. Light colours represent low scores while dark colours represent high scores. **(A and D)** Networks are coloured for CRP scores. **(B and E)** Networks are coloured for Fis scores. **(C and F)** Networks are coloured for IHF scores. The bottom figures **(D, E and F)** represent the same mutational networks as the ones directly above them in a spherical view, in which the highest connected nodes are in the centre of the circle and the lowest connected ones are peripheral.

Thus, as the generation of a full-sized artificial network would be unfeasible (4^{19} variants for CRP), we have decided to explore artificial networks locally. In this sense, our approach can be observed in **Figure 35**. Two sequences with contrasting scores (very low and very high) for CRP were selected from the natural mutational network and the distance between them was calculated as N= 4 (4 mutations). However, as the intermediate variants between these two sequences were absent in the natural sequences, we have generated all the possible variants (4^4 = 256 sequences) and calculated scores for CRP, Fis and IHF for all of them (**Figure 36**).

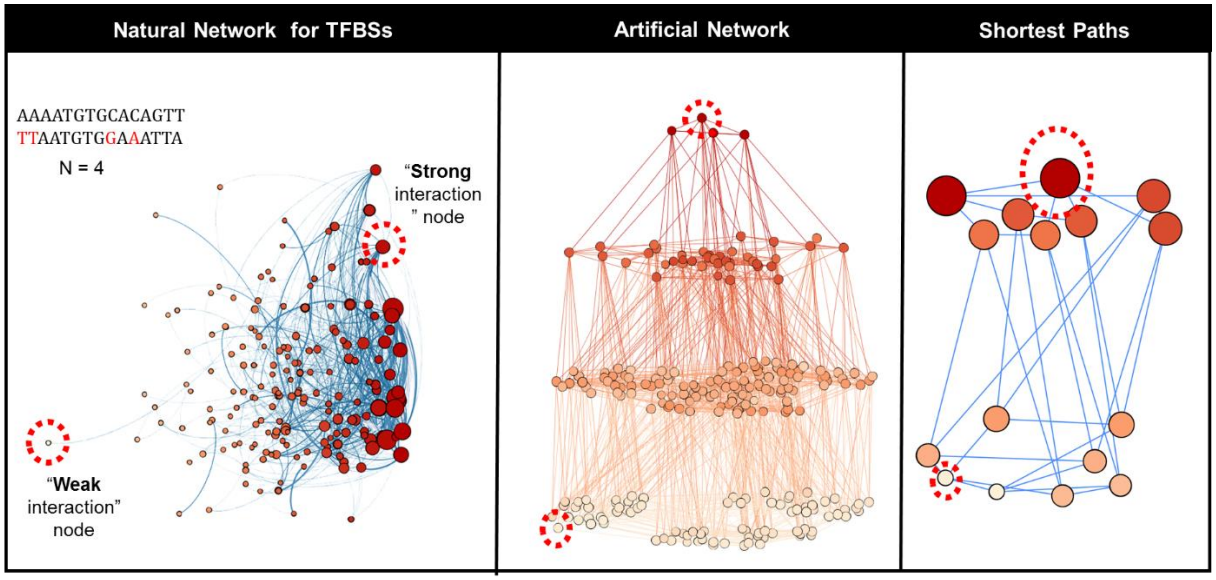


Figure 35. The use of artificial networks in order to examine the possible evolutionary paths between two sequences. After generating the natural mutational networks for CRP sequences, the next step was to evaluate the number of potential innovative paths evolution could "navigate" in the sequence space from a natural sequence with low PWM score to a natural sequence with high PWM score. From left to right: first the two sequences were selected from analysed data, then the number of mutations between then (N= 4) was used to generate an artificial network with all the possible variants that could appear in the evolutionary trajectory between the two sequences (4^4 possibilities = 256 nodes). Lastly, scores for CRP, Fis and IHF were calculated for all these sequences and the shortest paths were obtained for further analysis.

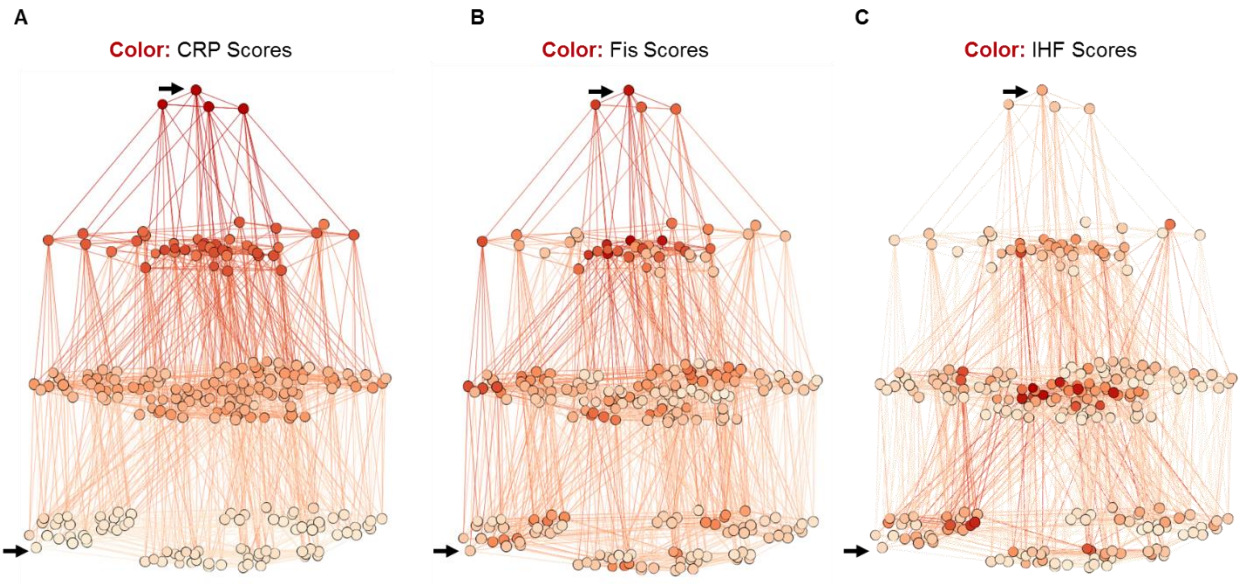


Figure 36. Generation of artificial networks. The artificial network between the two selected CRP sequences (indicated by black arrows) was generated (4^4 nodes) and plotted as a 3D landscape with CRP scores in the z-axis. The network nodes were ranked by heatmap colouring based on the selected PWM scores for **(A)** CRP, **(B)** Fis and **(C)** IHF.

The generation of the artificial networks has allowed the exploration of all the different paths from which a single natural sequence could have become another. In order to simplify the analysis, we have adopted a shortest-path algorithm which has reduced the artificial network to all the shortest trajectories between the two selected nodes (**Figure 37**). This network comprising the shortest paths between the two sequences has been further explored in terms of the scores for each PWM that each intermediate node would have. Thus, it is possible to observe that if a sequence was to evolve from a low-score one to a high-score one (for CRP) it could have evolved through a series of intermediates that would not only monotonically increase the score for CRP, but also simultaneously change its scores for Fis and IHF as well.

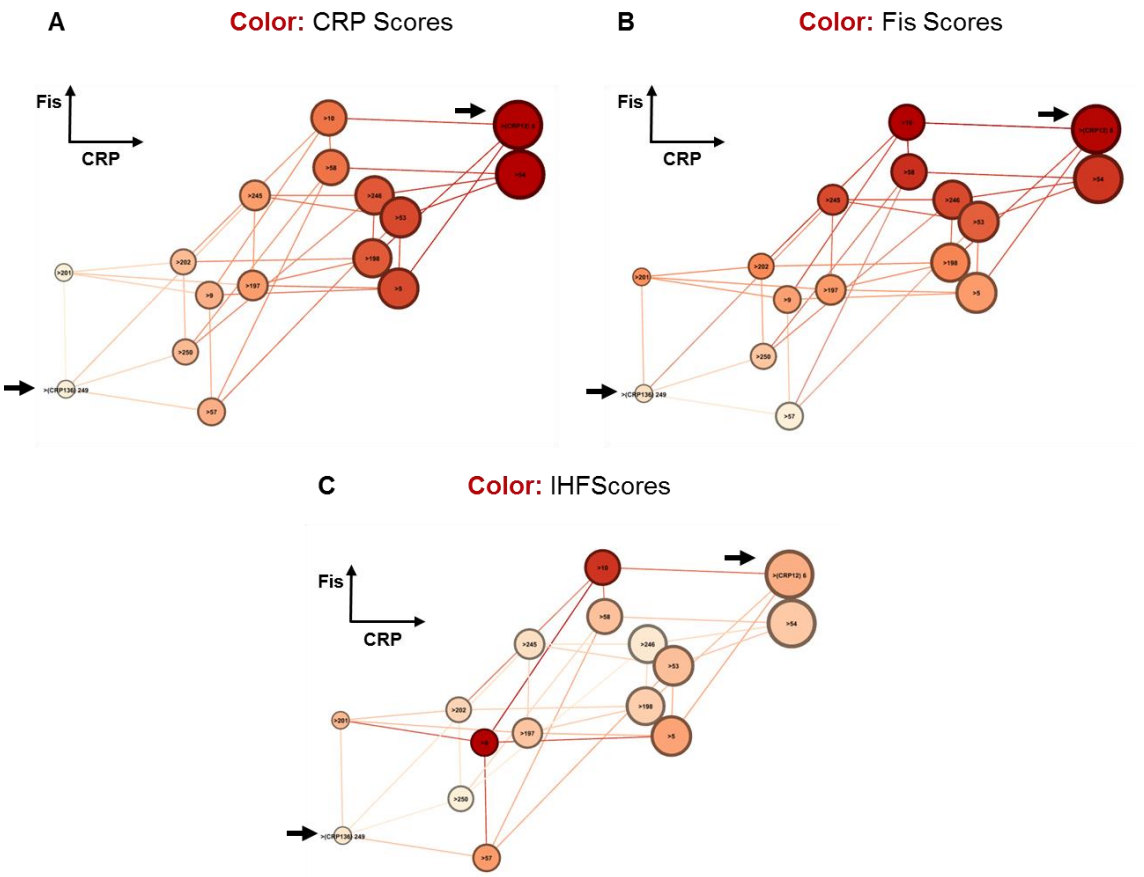


Figure 37. Shortest Paths in artificial networks. The same network is represented three times with a different color mapping in each version representing a different score. The network is topologically organized in a 2D-axis. The x-axis represents nodes with increasing values for CRP scores while the y-axis represents nodes with increasing values for Fis Scores. By calculating the shortest paths between the two natural CRP sequences (indicated by black arrows), it was possible to plot only the minimal set of evolutionary trajectories which would be required for one sequence to evolve into the other. The nodes in this graph were ranked by heatmap colouring based on the selected PWM scores for **(A)** CRP, **(B)** Fis and **(C)** IHF.

5. Discussion and Conclusions

How would evolution navigate through the sequence space of TFBSs? Would it take into account not only the score for a single TF, but also the multidimensional set of scores for different TFs which would potentially bind to this site? The answer is still unclear, however, the analysis presented here has provided us with some directions in this matter. Computationally, our data suggests data evolution might face a multi-dimensional dilemma when walking through the peaks and valleys of TFBSs sequences: mutations in a single might simultaneously affect affinity for multiple TFBSs and evolution might select these mutations according to the expression profiles they generate under specific selection pressures (**Figure 38**). If it really happens, would we be able to show it *in vivo*? The last analysis we have shown were able to provide us with shortest-paths between two sequences. We can now select nodes from these potential evolutionary pathways and clone them as transcriptional fusions to a core promoter followed by a GFP gene. These constructions can then be transformed in WT and mutant strains for CRP, Fis and IHF, revealing the biological impact of these sequences and how far (or close) they are from the computationally predicted behaviours.

Furthermore, very recent literature with both Eukaryotes and Prokaryotes has shown that TFs are able to share TFBSs although its implications are still unknown (Friedlander *et al.*, 2016, 2017; Rowland *et al.*, 2017). Here, we propose that this intrinsic property of regulatory sequences plays an important role in determining the transcriptional logic of complex bacterial promoters, influencing not only in the expression levels, but also in its dynamics and intrinsic noise levels. Further analysis such as "deep scanning" strategies, in which large artificial networks are overlapped with natural networks, would deeply benefit our understanding of the adaptive landscape for TFBSs. Combining this approach with experimental validations shall be the next step in our approach. This multi-scale framework will help us to elucidate the underlying properties of regulatory networks and how evolution might explore them for the rise of regulatory innovation.

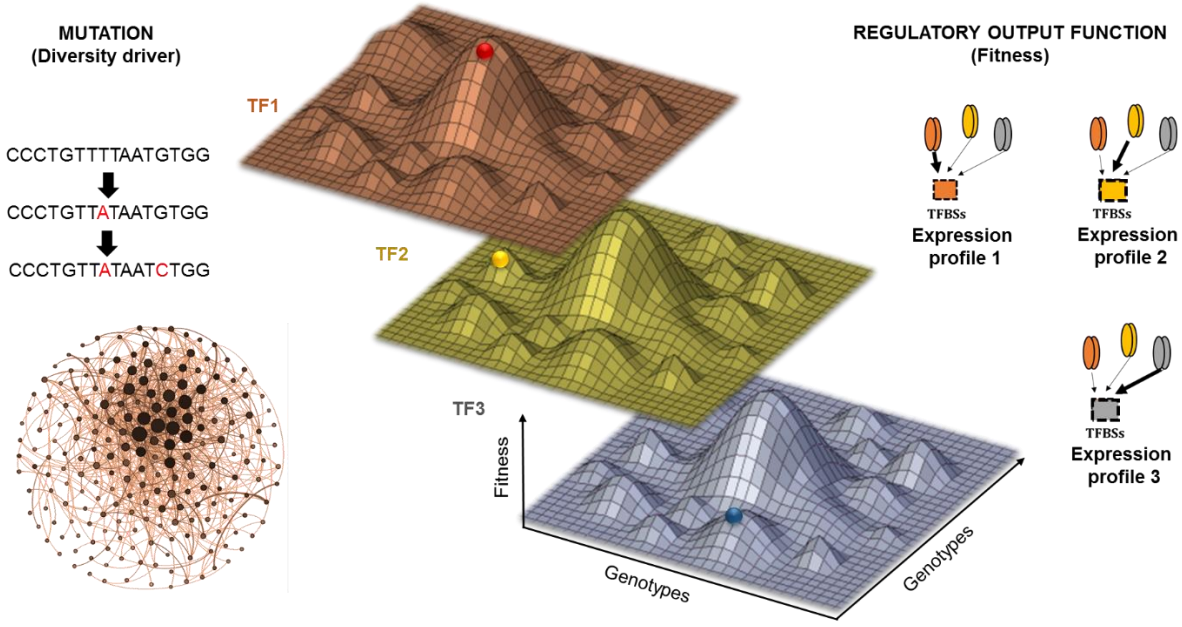


Figure 38. A framework for explaining the evolution of TFBSs in bacteria. A DNA sequence (top left) has a multidimensional space of scores associated with it (middle). In this case, a single sequence has three scores associated with it. Each score is just a single point in its own fitness landscape (here, we have three overlaid fitness landscapes, one for each TF that can bind the sequence). Every time a mutation occurs (top left), changing the sequence space, the scores for each TF will also change, leading to different expression profiles (top right). The selection of specific sequences and their associated expression profiles will depend on the environmental context and the nature of the regulated gene in the organismal context.

Chapter III

Mining novel constitutive promoter elements in soil metagenomic libraries in *Escherichia coli*

This chapter was published as:

Westmann, C. A., Alves, L. D. F., Silva-Rocha, R., & Guazzaroni, M. E. (2018). Mining novel constitutive promoter elements in soil metagenomic libraries in *Escherichia coli*. Frontiers in Microbiology, 9, 1344.

1. Specific Background

Although functional metagenomics has been widely employed for the discovery of genes relevant to biotechnology and biomedicine, its potential for assessing the diversity of transcriptional regulatory elements of microbial communities has remained poorly explored. In this context, the most common strategy for prospecting promoters is the usage of trap-vectors, which consist of transcriptional fusions between DNA fragments and a reporter gene. This method has been widely employed for assessing promoters in genomic DNA (Kubota, Yamazaki and Ishihama, 1991; Dunn and Handelsman, 1999; Lu, Bentley and Rao, 2004; Chen *et al.*, 2007), however its application in metagenomic DNA fragments has remained poorly explored (Uchiyama *et al.*, 2005; Han *et al.*, 2008). Furthermore, most adopted promoter trap-systems are unidirectional, while bacterial genomes present a large variation in the percentage of their leading-strand genes, ranging from ~45% to ~90% (Mao *et al.*, 2012, 2015), suggesting that a bi-directional promoter reporter system would be preferable. Therefore, in the present chapter, we merge this strategy into an integrative approach for exploring bacterial communities through the lens of their regulatory dynamics, focusing on the study of bacterial promoter elements from environmental soil samples.

Although both constitutive and inducible promoters can be potentially detectable by the bi-directional method, we have focused exclusively on the study of the former, as a proof of concept, avoiding substrate-based induction assays (Uchiyama *et al.*, 2005; Williamson *et al.*, 2005; Uchiyama and Miyazaki, 2010; Guazzaroni *et al.*, 2013). We have collected soil samples from two differentially biomass-enriched sites of a Secondary Atlantic Forest in South-eastern Brazil and generated metagenomic libraries in a bi-directional probe vector for primary screenings. We have characterised the expression behaviours of a large set of GFPlva expressing clones from both libraries and narrowed down our selection to 10 clones for an in-depth analysis regarding potential ORFs and endogenous promoters. By cross-validating *in silico* analyses and experimental data of predicted constitutive promoters, we have located and profiled the expression of 33 endogenous promoters within the selected clones, providing resourceful information concerning the architecture and transcriptional dynamics of promoters from metagenomic fragments. Thought the identification of novel constitutive, natural promoters, our work contributes to the expansion of the toolbox of synthetic biology, which, in turn, can be used for genetic modification of microorganisms relevant in Biotechnology.

2. Objectives

General Objective

To explore and characterize *cis*-regulatory elements in environmental samples for assessing the hidden regulatory diversity of uncultured bacteria

Specific Objectives

1. To develop a novel strategy for prospecting and characterizing promoters from metagenomics libraries

2. To analyse the expression profiles of metagenomic libraries

3. To quantify the promoter accessibility of soil metagenomic libraries using *E. coli* as a host

3. Materials and Methods

Bacterial strains, primers, plasmids and general growth conditions

The plasmids, bacterial strains, and primers used in this study are listed in **Table 4.** *E. coli* DH10B (Invitrogen) cells were used for cloning and experimental procedures. *E. coli* strains were routinely grown at 37ºC in Luria-Broth medium or M9 minimal medium (Sambrook, J.; Fritsch, E. F.; Maniatis, 1989) (6.4 g/L $Na_2HPO_4 \cdot 7H_2O$, 1.5 g/L KH_2PO_4, 0.25 g/L NaCl, and 0.5 g/L NH_4Cl) supplemented with 2 mM $MgSO_4$, 0.1 mM casamino acid, and 1% glycerol as the sole carbon source. When required, chloramphenicol (Cm) (34 µg/mL) was added to the medium to ensure plasmid retention. When cells were grown in minimal medium, antibiotics were used at half concentrations. Transformed bacteria were recovered on LB (Luria–Bertani) liquid medium for 1 hour at 37°C and 180 r.p.m, followed by plating on LB-agar plates at 37°C for at least 18 hours. All constructions were cloned into the pMR1 bi-directional-reporter vector (Guazzaroni and Silva-Rocha, 2014), which carries mCherry and GFPlva, a short-lived variant of GFP (**Figure 39**).

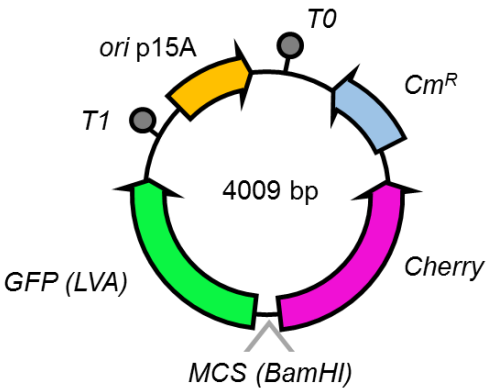


Figure 39. General scheme of the pMR1 vector (Guazzaroni and Silva-Rocha, 2014). The plasmid includes a resistance marker to the antibiotic Chloramphenicol (CmR), a low-medium copy replication origin (ori-p15a) and a multiple cloning site flanked by the *mCherry* and *GFPlva* reporter genes, into which fragments of metagenomic DNA. Transcription terminators T1 and T0 are also shown.

Table 4. Bacterial Strains, Plasmids and Primers used in this study

Strain	**Description**	**Source**
E. coli DH10B	F$^-$ *endA1 deoR*$^+$ *recA1 galE15 galK16 nupG rpsL* Δ*(lac)X74* φ80*lacZΔM15 ar aD139* Δ*(ara,leu)7697 mcrA* Δ*(mrr-hsdRMS-mcrBC)* Str^R λ$^-$	(Casadaban and Cohen, 1980; Grant *et al.*, 1990)
Plasmid	**Description**	**Source**
pMR1	Cm^R, *ori* p15A; dual mCherry GFPlva promoter probe vector. (pRV2 derivative, (Silva-Rocha and De Lorenzo, 2011))	(Guazzaroni and Silva-Rocha, 2014)
pMR1-pCAW1	CmR, orip15a; pMR1-pCAW1-GFPlva transcriptional fusion	This study
pMR1-pCAW2	CmR, orip15a; pMR1-pCAW2-GFPlva transcriptional fusion	This study
pMR1-pCAW3	CmR, orip15a; pMR1-pCAW3-GFPlva transcriptional fusion	This study
pMR1-pCAW4	CmR, orip15a; pMR1-pCAW4-GFPlva transcriptional fusion	This study
pMR1-pCAW5	CmR, orip15a; pMR1-pCAW5-GFPlva transcriptional fusion	This study
pMR1-pCAW6	CmR, orip15a; pMR1-pCAW6-GFPlva transcriptional fusion	This study
pMR1-pCAW7	CmR, orip15a; pMR1-pCAW7-GFPlva transcriptional fusion	This study
pMR1-pCAW8	CmR, orip15a; pMR1-pCAW8-GFPlva transcriptional fusion	This study
pMR1-pCAW9	CmR, orip15a; pMR1-pCAW9-GFPlva transcriptional fusion	This study
pMR1-pCAW10	CmR, orip15a; pMR1-pCAW10-GFPlva transcriptional fusion	This study
pMR1-pCAW1-p1	CmR, orip15a; pMR1-pCAW1-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW1-p2	CmR, orip15a; pMR1-pCAW1-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW1-p3	CmR, orip15a; pMR1-pCAW1-p3-GFPlva transcriptional fusion	This study
pMR1-pCAW1-p4	CmR, orip15a; pMR1-pCAW1-p4-GFPlva transcriptional fusion	This study
pMR1-pCAW2-p1	CmR, orip15a; pMR1-pCAW2-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW2-p2	CmR, orip15a; pMR1-pCAW2-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p1	CmR, orip15a; pMR1-pCAW3-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p2	CmR, orip15a; pMR1-pCAW3-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p3	CmR, orip15a; pMR1-pCAW3-p3-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p4	CmR, orip15a; pMR1-pCAW3-p4-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p5	CmR, orip15a; pMR1-pCAW3-p5-GFPlva transcriptional fusion	This study
pMR1-pCAW3-p6	CmR, orip15a; pMR1-pCAW3-p6-GFPlva transcriptional fusion	This study
pMR1-pCAW4-p1	CmR, orip15a; pMR1-pCAW4-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW4-p2	CmR, orip15a; pMR1-pCAW4-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW4-p3	CmR, orip15a; pMR4-pCAW3-p3-GFPlva transcriptional fusion	This study
pMR1-pCAW4-p4	CmR, orip15a; pMR1-pCAW4-p4-GFPlva transcriptional fusion	This study
pMR1-pCAW5-p1	CmR, orip15a; pMR1-pCAW5-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW5-p2	CmR, orip15a; pMR1-pCAW5-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW5-p3	CmR, orip15a; pMR1-pCAW5-p3-GFPlva transcriptional fusion	This study
pMR1-pCAW5-p4	CmR, orip15a; pMR1-pCAW5-p4-GFPlva transcriptional fusion	This study
pMR1-pCAW5-p5	CmR, orip15a; pMR1-pCAW5-p5-GFPlva transcriptional fusion	This study
pMR1-pCAW6-p1	CmR, orip15a; pMR1-pCAW6-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW6-p2	CmR, orip15a; pMR1-pCAW6-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW7-p1	CmR, orip15a; pMR1-pCAW7-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW7-p2	CmR, orip15a; pMR1-pCAW7-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW8-p1	CmR, orip15a; pMR1-pCAW8-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW8-p2	CmR, orip15a; pMR1-pCAW8-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW8-p3	CmR, orip15a; pMR1-pCAW8-p3-GFPlva transcriptional fusion	This study
pMR1-pCAW8-p4	CmR, orip15a; pMR1-pCAW8-p4-GFPlva transcriptional fusion	This study
pMR1-pCAW9-p1	CmR, orip15a; pMR1-pCAW9-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW9-p2	CmR, orip15a; pMR1-pCAW9-p2-GFPlva transcriptional fusion	This study
pMR1-pCAW10-p1	CmR, orip15a; pMR1-pCAW10-p1-GFPlva transcriptional fusion	This study
pMR1-pCAW10-p2	CmR, orip15a; pMR1-pCAW10-p2-GFPlva transcriptional fusion	This study
Primer	**Sequence (3'-5')**	**Source**

pMR1-F	CTCGCCCTTGCTCACC	This study
pMR1-R	ACAAGAATTGGGACAACTCC	This study

Study Site, Soil Sampling and DNA extraction

Soil samples were obtained from a parcel of southeast region of Brazil (South America), from a Secondary Atlantic Forest at the University of Sao Paulo (Ribeirão Preto, São Paulo, Brazil; 21º09´58.4´´S, 47º51´20.1´´W, at an altitude of 540 m). The soil from those parcels are geologically considered Oxisols (Schaefer, Fabris and Ker, 2008) – clay soil always presenting a red or yellowish color, due to the high concentration of iron (III) and aluminium oxides and hydroxides -. The topsoil from two sections of the parcel (herein referred to as USP1 and USP3) were sampled at a depth of 0–15 cm on July 2015 (soil temperature 23 ºC). Three replicates (0.2 kg each) were collected within a 1m distance, and the samples were stored at −20°C until DNA was extracted. Each sample was differentially enriched regarding tree species abundance on plant-litter composition: (i) enriched in leaves from *Phytolacca dioica* and (ii) from *Anadenanthera* spp. DNA was extracted from soil samples using the UltraClean™ Soil DNA isolation Kit (Mo Bio Laboratories, Solana Beach, CA, USA). DNA was visualized by using 0.7% (w/v) agarose gel electrophoresis and quantified spectrophotometrically (260 nm).

Metagenomic libraries construction and screening for fluorescent clones

For the construction of the libraries, metagenomic DNA was partially digested using Sau3AI, and fragments from 1.5 kb to 7 kb were extracted from an agarose gel for ligation into the dephosphorylated and BamHI-digested pMR1 vector. Ligation mixtures were transformed by electroporation into *E. coli* DH10B cells. To amplify the libraries, they were grown on LB agar plates containing Cm and incubated for 18 h at 37°C. Both green and red clones were manually isolated from LB-agar plates exposed to a blue light wavelength (at approximately 470 nm) by a transilluminator (Safe Imager™ 2.0 Blue Light Transilluminator). Ten fluorescent and twenty non-fluorescent clones were randomly picked from each library and had their plasmids extracted, following digestion with EcoRI and SmaI enzymes for checking presence/absence of inserts and their sizes. Cells from the same library were collected and pooled together in LB supplemented with 10% (wt/vol) glycerol for storing at -80°C. The plasmids from the 10 selected clones were isolated from individual clones and transformed into new *E. coli* DH10B cells to reconfirm expression patterns.

Nucleic acid techniques

DNA preparation, digestion with restriction enzymes, analysis by agarose gel electrophoresis, isolation of DNA fragments, ligations, and transformations were done by standard procedures (Sambrook, J.; Fritsch, E. F.; Maniatis, 1989). Plasmid DNA was sequenced on both strands by primer walking using the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit (PerkinElmer) and an ABI PRISM 377 sequencer (Perkin-Elmer) according to the manufacturer's instructions.

GFP fluorescence assay and data processing

To measure promoter activity, freshly plated single colonies were grown overnight in M9 medium supplemented with required antibiotics. Samples were diluted 1:20 (v/v) in M9 medium for a final volume of 200 µL in 96-well microplates. Cell growth and GFP fluorescence were quantified using a Victor X3 plate reader (PerkinElmer, Waltham, MA, USA). Promoter activities were expressed as the emission of fluorescence at 535 nm upon excitation with 485 nm light and then normalised with the optical density at each point (reported as fluorescence/OD_{600}) after background correction. Background signal was evaluated with non-inoculated M9 medium and used as a blank for adjusting the baseline of measurements. *E. coli* DH10B harbouring the pMR1 empty plasmid was used as a negative control. Three different positive controls were used, consisting in *E. coli* DH10B harbouring pMR1 plasmid with one of the following synthetic constitutive promoters from the iGEM BBa_J23104 Anderson's catalogue (http://parts.igem.org/Promoters/Catalog/Anderson) (Kelly et al. 2009) upstream a GFPlva reporter: J23100, J23106 and J23114 (referred here as p100, p106 and p114, respectively (Sanches-Medeiros, Monteiro and Silva-Rocha, 2018). Unless otherwise indicated, measurements were taken at 30 min intervals over 8 h. All experiments were performed with both technical and biological replicates, being biological triplicates evaluated as independent measurements on different dates. Raw data were processed and plots were constructed using Microsoft Excel. All data was normalised by background values and transformed to a log2 scale for better data visualisation. Heatmap dendrograms with expression profiles were generated by using MeV2 (http://mev.tm4.org/) software.

Small-DNA inserts libraries generation and screening

In order to experimentally find and validate the promoter regions from each of the ten selected metagenomic clones, an experimental technique was developed based on the previously described methodology of metagenomic library construction. All selected clones had their plasmids extracted and pooled together in an equimolar ratio. The pooled sample was amplified through a single PCR reaction using high-fidelity polymerase enzyme (Phusion) and previously described primers flanking the MCS region (Multiple Cloning Site) of the pMR1 vector, into which the metagenomic inserts were cloned. The resulting amplicons were first submitted to an analytical digestion followed by electrophoretic analysis for finding the optimal concentration of Sau3AI enzyme for obtaining fragments size ranging from 0.1 kb to 0.5 kb. Then, the purified pooled samples were fragmented by Sau3AI in preparative digestion and thereafter punctured from a 1% agarose gel in the region between 0.1 kb and 0.5 kb. These small DNA fragments, in turn, were ligated to pMR1 vector. Aliquots of electrocompetent *E. coli* DH10B cells were transformed with ligated DNA. A total of 100 fluorescent clones (80 expressing GFP and 20 expressing mCherry) were isolated under blue light excitation screening and had their plasmids extracted for sequencing reactions. Fluorescent clones were stored at -80°C in LB medium supplemented with required antibiotics and 10% glycerol (v/v).

In silico analysis of ORFs and promoter regions

The inserts of selected clones were sequenced on both strands as previously described. Sequences were manually assembled for the generation of 10 contigs. All sequences were analysed for taxonomic origins by using the *PhylopythiaS* Web Server (Patil, Roune and McHardy, 2012) (http://phylopythias.bifo.helmholtz-hzi.de/index.php?phase=wait), a sequence composition-based classifier that utilizes the hierarchical relationships between clades. Putative ORFs were identified and analysed using the online ORF Finder platform, available at the NCBI website (http://www.ncbi.nlm.nih.gov/gorf/gorf. html). Comparisons of nucleotide and transcribed amino acid sequences were performed against public databases (NCBI) using BlastN, BlastX and BlastP (BLAST, basic local alignment search tool) at the NCBI on-line server. For translation to protein sequences, the bacterial code was selected, allowing ATG, GTG, and TTG as alternative start codons. All the predicted ORFs longer than 270 bp were translated and used as queries in BlastP. Sequences with significant matches were further analysed with PSI-BLAST, and their putative function was annotated based on their similarities to sequences in the COG (Clusters of Orthologous Groups) and Pfam (Protein Families) databases. Predicted general cellular functions were annotated only for known ORFs

based on the MultiFun classification (Serres and Riley, 2000). All sequences with an E-value higher than 0.001 in the BlastP searches and longer than 300 bp were considered to be unknown. Transmembrane helices were predicted with TMprep (http://www.ch. embnet.org/software/TMPRED_form.html) and signal peptides with Signal P3.0 server (http://www.cbs. dtu.dk/services/SignalP/). A complete table can be found at Supplementary **Table S1**. Promoter prediction was based on the analysis of the ten contigs by using both BPROM (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) and bTSSfinder (http://www.cbrc.kaust.edu.sa/btssfinder/) web-based platforms. Both methods searched for *rpoD*-related sequences and we have only considered as valid predictions the ones matched on both approaches. Those filtered sequences were used to cross-validate 23 out of 33 experimentally defined regulatory regions by comparing the positions between predicted and experimental sequences in metagenomic fragments. The positions of the 33 small DNA fragments were obtained by a multiple alignment of the original contigs (queries) against those selected sequences, which has also allowed the validation of the promoter's directionality – forward or reverse - by observing the matched strands (Plus/Plus or Plus/Minus). The consensus Logo sequence was based on the alignment of the 33 experimentally validated promoters, using the WebLogo platform (http://weblogo.berkeley.edu/logo.cgi) (Schneider and Stephens, 1990; Crooks *et al.*, 2004).

Criteria for the choice of sample sizes

The sample sizes chosen in this work were based on a seminal study regarding the characterization of random promoter libraries (Cox, Surette and Elowitz, 2007) in which ~1% (288) of the total set of promoters (22,000) was selected for further analysis. In our study, we have selected a much higher fraction of the population for sampling (~25% of 1,100 screened clones). Furthermore, using classical statistics for determining optimal sample sizes and reducing the uncertainty caused by sampling error (Nakagawa and Cuthill, 2007), we have found that sampling 260 clones from a total of 1,100 clones would result in the confidence level of 99% with a confidence interval of 0.07. Each selected clone was manually streaked in LB-agar and microbiologically purified two times for further validation in plate reader assays – which was done with biological and technical triplicates. Regarding the 10 selected clones at the in-depth analysis, we have adopted the same sample fraction from the study of Cox *et al.*, 2007, (1% of the total number of positive clones – 10 in 1,100 clones). In this context, from

each of the 10 analyzed clones containing metagenomic fragments we have obtained at least three promoters, which were individually characterized in plate reader assays. The choice of 100 clones from the small-fragment library was based on the following rationale: (i) the combined size of the 10 selected clones in this analysis was 30 kb, (ii) each small fragment had an average of 0,4 kb, thus, (iii) 100 fluorescent clones from the small-insert library would represent ~40 kb, providing enough coverage for all 10 original clones. Furthermore, as each fluorescent clone would represent a single promoter sequence at a specific region in the original clones, it was highly improbable that the 100 selected clones would cover the 10 original clones. Thus, our intention in choosing a sample size of 100 clones was to enrich the single promoters. This assumption was further supported by the discovery of only 33 promoters among those 100 sequences (promoter sequences were overrepresented).

4. Results

4.1 Generating metagenomic libraries and screening for fluorescent clones

We have constructed and assessed two metagenomic libraries hosted in *E. coli* DH10B strain for the analysis of bacterial promoters in environmental samples (**Figure 40**). The libraries were generated from soil microbial communities of two sites bearing differential tree litter composition (*Anadenanthera* spp. and *Phytolacca dioica*) within a Secondary semi-deciduous Atlantic Forest zone at the University of Sao Paulo, Ribeirão Preto, Brazil – see Experimental Procedures for further details. Both metagenomic DNA were cloned into the pMR1 (Guazzaroni and Silva-Rocha, 2014) bi-directional reporter vector – which has a *GFPlva* and a *mCherry* reporter gene in opposite directions, flanking a multiple cloning site; chloramphenicol resistance marker and a *p15a* origin of replication for low/medium copy number. Each metagenomic library presented about 250 Mb of environmental DNA distributed into approximately 60.000 clones harbouring insert fragments size ranging from 1.5 kb to 7 kb, with an average size of 4.1 kb (**Table 5**). We have chosen fragments of 1.5-7 kb in order to validate our strategy on standard-sized functional metagenomic libraries based on plasmid vectors (Gabor, Alkema and Janssen, 2004; Uchiyama *et al.*, 2005; Pushpam, Rajesh and Gunasekaran, 2011; Jiménez *et al.*, 2012; Guazzaroni *et al.*, 2013). In total, 1,100 fluorescent clones, resulting in a rate of approximately one fluorescent clone every one hundred fifty clones (USP1) or every ninety clones screened (USP3), were manually selected under blue light exposition. Then, these fluorescent clones were directly recovered from LB agar plates supplemented with chloramphenicol. The direct screening was preferred over the use of metagenomic clone pools from stocks as it reduces the chances of both biased clone enrichment (e.g. clones with higher growth rates, usually clones bearing small inserts or without insert) and dilution of positive clones with impaired growth (e.g. clones with high expression of GFP and/or other exogenous genes), avoiding thus clonal amplification.

Table 5. Features of the generated metagenomic libraries.

Metagenomic Library	USP 1	USP 3
Total number of clones	100,000	90,000
Percentage of clones with insert	60%	70%
Number of clones with insert	60,000	63,000
Total number and rate* of fluorescent clones	400 (1:150)	700 (1:90)
Total number and rate* of green clones	270 (1:220)	400 (1:157)
Total number and rate* of red clones	130 (1:460)	300 (1:210)
Average insert size	4,5 kb	3,7 kb
Total Metagenomic Library Size	270 Mb	233 Mb
Estimated number of genomes**	60	52

* Rate represented by the number of fluorescent clones divided by the total number of clones with inserts.

** Assuming 4.5 Mb per genome (Raes et al. 2007).

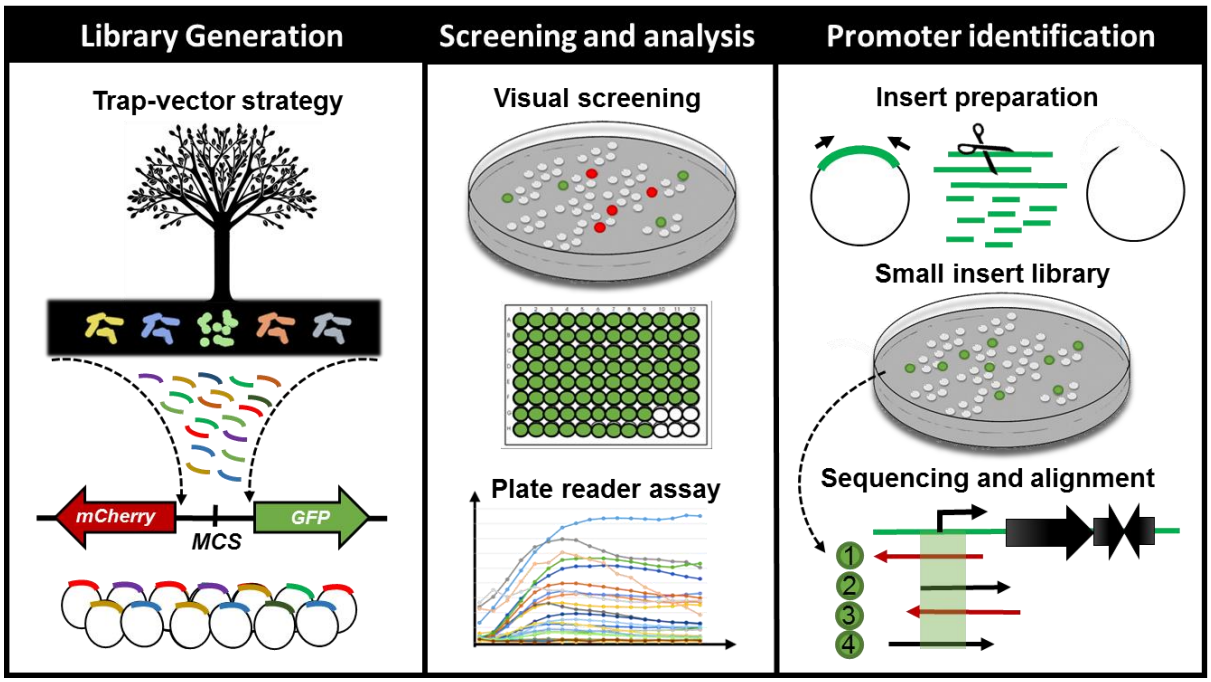


Figure 40. Schematic representation of the workflow for finding, characterising and cross-validating novel bacterial *cis*-regulatory elements in environmental samples. From left to right: firstly, we have generated metagenomic libraries from soil samples in *E. coli* DH10B. The DNA fragments were cloned into a bi-directional reporter trap-vector (bearing *mCherry* and *GFPlva* fluorescent reporters), pMR1, which allowed for the screening of promoters in both DNA strands. Secondly, we have manually screened all visible fluorescent clones from our metagenomic libraries and analysed the expression patterns of all green fluorescent clones on a microplate reader during 8 hours. Lastly, we have selected ten clones based on their GFPlva expression patterns for an in-depth analysis combining experimental (small DNA insert library

generation) and *in silico* promoter prediction. This integrated strategy has allowed us to identify, validate and estimate the accessibility of novel promoter regions from metagenomic libraries.

4.2 Evaluating the expression dynamics of fluorescent clones

In order to analyse the expression patterns of the isolated clones, we evaluated the intrinsic dynamics of GFPlva and mCherry by randomly selecting 20 clones expressing each reporter (as schematically represented in **Figure 40**). As represented in Figures 2B-C, we found that clones expressing mCherry were not suitable for standard microplate 8-hour assays, as the fluorescence intensity values differed dramatically between 8 and 24 hours after the beginning of the experiment. The slow kinetics of mCherry expression has already been reported as a consequence of a two-step oxidation process for protein maturation when compared to the one-step maturation process found in GFP reporters (Hebisch *et al.*, 2013). We highlight that although mCherry clones were not optimized for dynamic profiling, they were essential for quantifying the total number of metagenomic fragments harbouring promoters accessible to *E. coli* – the sum of both green and red fluorescent clones in the library. On the other hand, the clones expressing GFPlva presented the enhanced intrinsic properties for microplate assays, supported by the observation of very similar fluorescence intensities between the two time points tested. Furthermore, the GFPlva has an LVA-degradation tag attached to its C-terminal, which reduces GFP accumulation and increases protein turnover, generating a more precise fluorescence output on analysis of expression patterns (Andersen *et al.*, 1998).

Thus, 260 clones expressing GFPlva – see Experimental Procedures for further information about chosen sample sizes – (160 clones from the USP1 library and 100 from USP3) were selected for further analysis of expression patterns on microplate reader assays with biological and technical triplicates. The dynamic profiles for each clone were converted into heat maps and hierarchically clustered by a Euclidean Distance algorithm into a dendrogram, concisely representing the expression patterns of each metagenomic library. In order to assess the diversity of promoter strengths among the generated metagenomics libraries, three previously characterized constitutive promoters (see Experimental Procedures for further information) positioned upstream a GFPlva reporter were used as standards for strong, medium and weak expression profiles (referred here as p100, p106 and p114, respectively (Sanches-Medeiros, Monteiro and Silva-Rocha, 2018)).

Considering both metagenomics libraries, we have found a total of 30 strong promoters showing a strength similar to the p100 control, 40 medium strength promoters similar to the p106 control, 60 weak promoters similar to the p114 control and a wide range of promoters with particular expression patterns which did not cluster with any of the previously mentioned positive controls (**Figure 41D** and **Supplementary Figure S1**). Moreover, the dynamic expression profiles have allowed us to observe a few clones that, although constitutively active, had their GFPlva expression levels increased during certain time frames (**Figure 41D**). Concerning the hierarchical organization of the expression profiles, the dendrogram of the USP3 library (**Figure 41D**) could be subdivided into at least four well-defined expression clusters comprising: (i) high, (ii) medium, (iii) low and (iv) very low expression profiles. A very similar pattern was identified in the expression dendrogram independently generated for the USP1 metagenomic library (see **Supplementary Figure S1**).

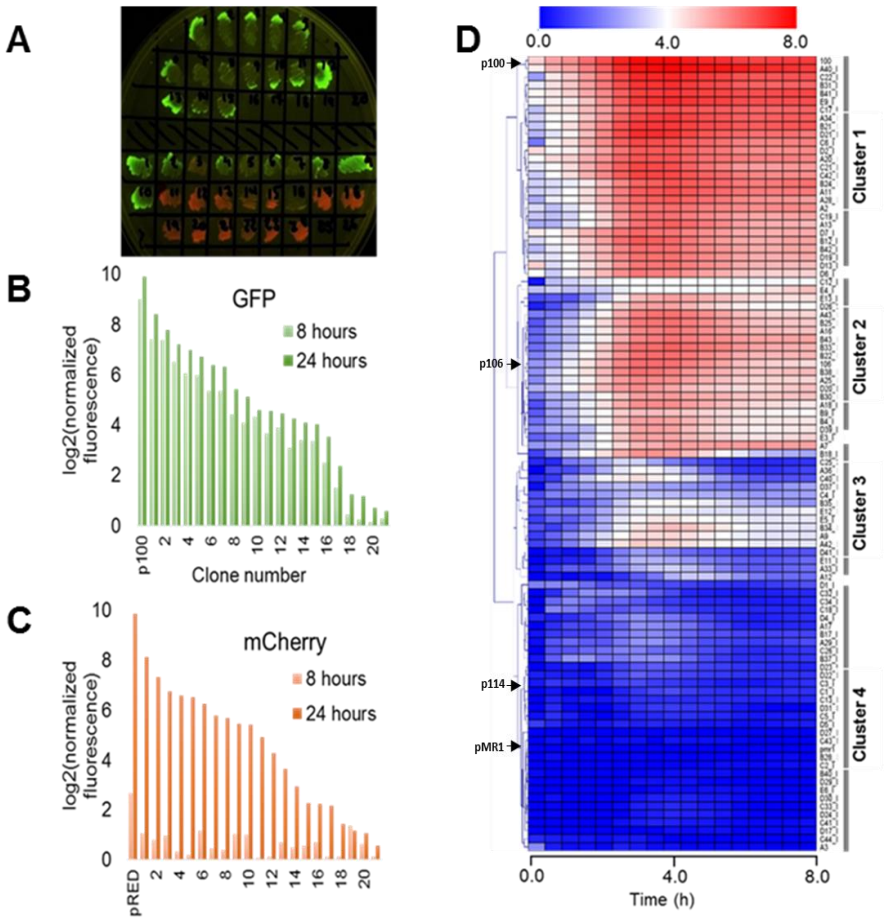


Figure 41. Evaluating the expression dynamics of fluorescent clones. (A) LB-agar plate under blue light excitation comprising a subset of metagenomic isolated clones expressing GFPlva (top) and mCherry (bottom) fluorescent reporters. A few clones were observed to express both reporters. All isolated clones were initially considered to hold at least one endogenous promoter. **(B-C)** Indirect assessment of maturation times from both fluorescent reporters GFPlva (**B**) and mCherry (**C**) after 8 hours (light bars) and 24 hours

(dark bars) of the beginning of the experiment. Maturation times are substantially lower for mCherry than for GFPlva, which excluded the former from further analyses. Positive controls for GFP and mCherry are represented by p100 and pRED, respectively. Fluorescence data has been normalised by OD_{600} values for each sample following normalisation by values from the negative control (empty-pMR1). Data was transformed to log2 scale to allow better visualisation of fluorescence variation. **(D)** Hierarchical representation of a metaconstitutome (i.e. all expression profiles from a single metagenomic library (USP3) in *E. coli*. Fluorescence time-lapse dynamics were measured during 8 hours for each clone and represented as heat maps. Promoter activities (calculated as GFP/OD_{600}) were normalised by the negative control (*E. coli* DH10B harbouring empty pMR1) and transformed to log2 scale in order to facilitate the visualisation of subtle activities. Positive controls (p100, p106 and p114 - strong, medium and low expression, respectively) and negative control (pMR1) expression profiles are indicated by black arrows at the left side of the heatmap. Data are representative of three independent experiments.

4.3 *In silico* analysis of DNA metagenomic fragments from selected clones

From the 260 assessed samples, we have selected 10 clones displaying particular profiles (see **Supplementary Figure S2**) – see Experimental Procedures for further information about chosen sample sizes - depicting the diversity of expression behaviours found in both libraries. The inserts from selected clones were sequenced and analysed for C-G content, taxonomic origins, potential ORFs and RpoD-related promoter regions (-10 and -35 conserved regions). The relative abundance of the guanine-cytosine content of each insert was assessed (**Table 6**), resulting in a median of 54%, varying from 43% to 61%, indicating their diverse phylogenetic affiliation. Using the *PhylopythiaS* sequence classifier for metagenomic sequences (Koonin, 2009; Patil, Roune and McHardy, 2012), the DNA fragments were assigned to their closely related phylum (**Table 6** and **Supplementary Figure S3**). The most abundant assigned phyla were Proteobacteria (46%), followed by Actinobacteria (23%), Verrumicrobia (15%), Chloroflexi (8%) and Bacteroidetes (8%) (**Supplementary Figure S3**).

In the case of the identification of putative genes, twenty-nine ORFs with significant *E-values* (<0,001) were found (**Table 6**) unevenly distributed between both DNA strands, in line with a lack of strong directional trends regarding bacterial genome organization (Koonin, 2009). The ORFs were also classified within a range of functional classes (delineated by MultiFun (Serres and Riley 2000)) and taxonomic groups based on closest similar proteins (**Table 6**). Regarding gene function, the most abundant ORFs were related to unknown functions (31%) and metabolism (31%), followed by stress adaptation cell processes (17%) **(Table 6**).

Table 6. Description of the ORFs contained in plasmids from the selected clones (pCAW1 to pCAW10) and their sequence similarities.

Clone_Sample [insert bp]	G + C %	GenBank accession No.	Phylum[a]	ORF[b]	Strand	Length (aa[c])	Closest similar protein[d] (Length in aa)	Closest Organism / Phylum[e]	Identity (%)	Putative function
pCAW1 (2367bp)	55%	KY939589	Proteobacteria or Verrucomicrobia	1	Minus	131	hypothetical protein (416)	*Bacteriodetes bacterium / Proteobacteria*	68%	Alginate lyase
				2	Plus	271	hypothetical protein (261)	*Acidobacteria bacterium / Acidobacteria*	73%	17-B-hydroxysteroid dehydrogenase
				3[b]	Plus	295	beta-glucosidase (777)	*Caulobacter sp. OV484 / Proteobacteria*	66%	beta-glucosidase
pCAW2 (2069bp)	52%	KY939590	Actinobacteria	1	Plus	304	Unkonwn[c]	*Hyphomicrobium sp. NDB2Meth4 / Proteobacteria*	33%	Unknown
				2	Plus	249	Unkonwn	*Hungatella hathewayi / Firmicutes*	33%	Unknown
pCAW3 (4404bp)	53%	KY939591	Proteobacteria	1	Minus	318	IS4 family transposase (320)	*Escherichia coli / Proteobacteria*	96%	IS4 family transposase
				2	Minus	1011	DNA-directed RNA polymerase subunit beta' (1430)	*Sphingobacteriales bacterium 44-61/ Bacteroidetes*	83%	RNA polymerase - Beta Subunit
				3	Plus	120	Uncharacterised protein (135)	*Bordetella pertussis/ Proteobacteria*	47%	Unknown
				4	Plus	151	Uncharacterised protein (130)	*Bordetella pertussis/ Proteobacteria*	37%	Unknown
				5	Plus	94	Uncharacterised protein (64)	*Bordetella pertussis/ Proteobacteria*	82%	Unknown
				6	Plus	96	Uncharacterised protein (86)	*Vibrio cholerae/ Proteobacteria*	48%	Unknown
				7	Plus	173	predicted protein (585)	*Ruminococcus sp. CAG:403 / Proteobacteria*	26%	Unknown
pCAW4 (4002bp)	61%	KY939592	Proteobacteria	1	Minus	245	nosine monophosphate cyclohydrolase (246)	*Ktedonobacter racemifer/ Chloroflexi*	63%	IMP cyclohydrolase
				2	Minus	214	phosphodiesterase (498)	*candidate division NC10 bacterium/ NC10*	40%	phosphodiesterase

				3	Minus	402	hypothetical protein A2Y08_02680 (625)	*Planctomycetes bacterium GWA2_40_7/ Planctomycetes*	43%	Unknown
				4[b]	Plus	142	gentisate 1,2-dioxygenase (349)	*Pseudomonas sp. 21C1 / Proteobacteria*	60%	gentisate 1,2-dioxygenase
pCAW5 (2724bp)	54%	KY939593	Verrucomicrobia	1[b]	Plus	642	pyruvate:ferredoxin oxidoreductase (1565)	*uncultured bacterium HF770_11D24] / Acidobacterium*	80%	pyruvate:ferredoxin oxidoreductase
pCAW6 (2125bp)	57%	KY939594	Chloroflexi or Proteobacteria	1	Plus	159	hypothetical protein BGO39_33875 (215)	*Chloroflexi bacterium 54-19 / Chloroflexi*	65%	MerR family
				2	Plus	336	hypothetical protein BGO39_33870 (347)	*Chloroflexi bacterium 54-19/ Chloroflexi*	78%	PrsW intramembrane metalloprotease
				3[b]	Plus	163	hypothetical protein BGO39_33865 (173)	*Chloroflexi bacterium 54-19/ Chloroflexi*	75%	chromate transporter
pCAW7 (2558bp)	46%	KY939595	Actinobacteria	1[b]	Minus	391	hypothetical protein A2X07_06330 (480)	*Flavobacteria bacterium GWF1_32_7/ Bacteroidetes*	45%	Por secretion system sorting domain
				2	Minus	250	hypothetical protein (586)	*Chitinophagaceae bacterium PMP191F/ Bacteroidetes*	65%	Polysaccharide Lyase
pCAW8 (4480bp)	57%	KY939596	Actinobacteria	1	Plus	508	hypothetical protein AUH20_02325 (597)	*Rokubacteria bacterium/ Rokubacteria*	76%	5-oxoprolinase / Hydantoinase_B
				2	Minus	348	Oxidoreductase (336)	*Rokubacteria bacterium/ Rokubacteria*	61%	Flavin-utilizing monoxygenases
				3	Plus	314	hypothetical protein ETSY1_46935 (279)	*Candidatus Entotheonella sp. TSY1/ Tectomicrobia*	76%	Cellulose biosynthesis BcsQ
pCAW9 (2573bp)	43%	KY939597	Bacteroidetes or Proteobacteria	1[b]	Minus	81	hypothetical protein (129)	*Janthinobacterium/ Proteobacteria*	50%	Unknown
				2	Minus	303	Formylglycine-generating enzyme (379)	*Mucilaginibacter sp./ Bacteroidetes*	65%	Formylglycine-generating enzyme
				3	Minus	457	acetylglucosamine-6-sulfatase (504)	*Flavihumibacter solisilvae/ Bacteroidetes*	67%	acetylglucosamine-6-sulfatase

pCAW10 (2076bp)	56%	KY939598	Proteobacteria	1	Plus	204	hypothetical protein (195)	*Luminiphilus syltensis/ Proteobacteria*	50%	Unknown

[a]Classification based on *PhylopythiaS* (Patil, Roune and McHardy, 2012) webserver
[b]Truncated proteins.
[c]aa, amino acids.
[d]Sequences with an *E*-value higher than 0.001 in Blastp searches were considered to be unknown proteins.
[e]Classification based on Blastp.

The *in silico* promoter prediction has also provided relevant information concerning the potential number of regulatory regions on each selected fragment. The BPROM software (Solovyev and Salamov, 2011) has been extensively employed in other promoter prediction studies and is based on the analysis of the -35 and -10 consensus sequence of RpoD promoters. The main sigma subunit, sigma-70 encoded by *rpoD*, plays a major role in transcription of growth-related genes, the so-called housekeeping genes (Lonetto, Gribskov and Gross, 1992; Gruber and Gross, 2003; Paget and Helmann, 2003). From the *in silico* analysis, a total of 140 promoters were predicted among the 10 selected clones, suggesting an average of 5 RpoD-related promoters/kb. This led us to question whether most expression profiles previously described (**Figure 41D** and **Supplementary Figure S1**) were representing the dynamics of a single "dominant" promoter or the combined effect of multiple adjacent promoters present in the metagenomic fragment. Considering that, we have delineated a strategy to experimentally assess the number and location of accessible promoters from our selected clones, contrasting experimental results against *in silico* data.

4.4 Experimental identification, characterisation, and cross-validation of promoter regions

In order to explore the potential set of accessible promoter regions from our metagenomic libraries, we developed a small DNA insert library generation approach (**Figure 40**). Firstly, the plasmids from the previously 10 selected clones (original clones) were pooled together for insert amplification in a single PCR reaction. The resulting amplicons were fragmented by Sau3AI digestion and DNA fragments ranging from 0.1 kb to 0.5 kb were selected for subsequent cloning into the pMR1 vector. The generation of this sub-fragment library allowed the screening for both red and green fluorescent colonies as they would represent the accessible set of promoters among the metagenomic DNA fragments studied. It is important to highlight that as the cloning process was not directed, small fragments bearing promoter regions had a 50% chance of being cloned in any direction, thus clones expressing mCherry were also isolated for subsequent sequencing. A total of 100 clones – see Experimental Procedures for further information about chosen sample sizes - coming from the small DNA insert library (80 expressing GFPlva and 20 expressing mCherry) were sequenced and then aligned against the original metagenomic fragments. As a result, we have identified at least 33 promoter regions

within the initial set of the selected metagenomic clones (**Figure 42**, **Supplementary Figure S4** and **Supplementary Table S1**).

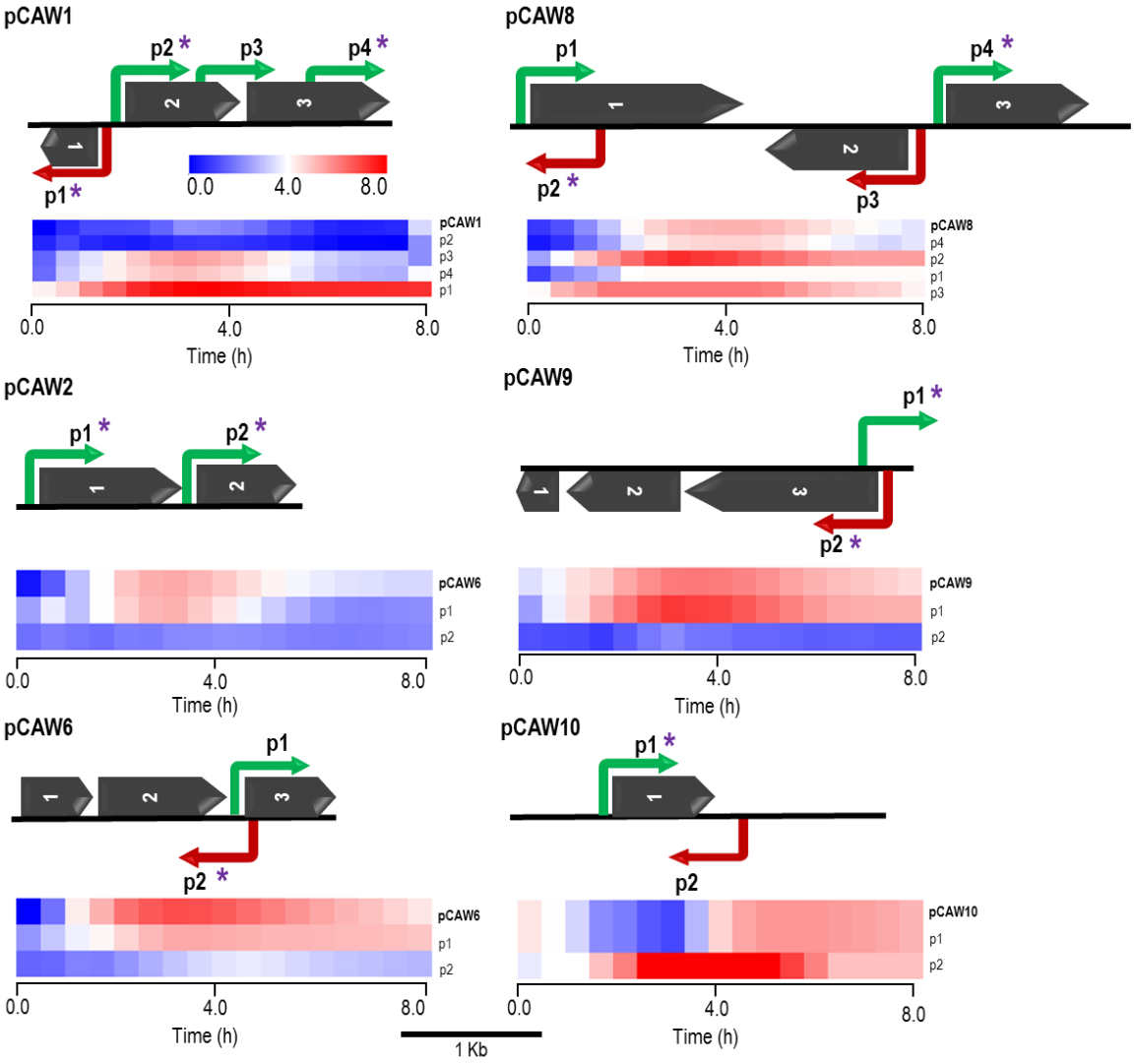


Figure 42. Schematic representation of six metagenomic inserts (contigs) showing predicted ORFs and experimentally validated/characterised promoters. Each contig is identified on the far left of each subfigure. Promoters are indicated by elbow-shaped arrows and name according to their relative position in the contig. Promoter directionality, regarding the leading and lagging strands, is represented by yellow and blue colors, respectively. Asterisks over specific promoters indicate regulatory regions which were cross-validated by matching *in silico* predictions. Dark arrows represent predicted ORFs, according to their relative positions in each contig (see **Table 6** for more information). All genetic features respect their original relative sizes, following the 1 kb scale depicted at the bottom of this figure. Beneath each metagenomic insert, there is a heat map cluster representing the whole set of promoter activities measured during 8-hours fluorescence assays. The first line of each cluster shows the original expression profile initially measured for each metagenomic insert. All other lines represent expression activities from de novo experimentally validated promoters within each contig (small DNA fragments). The second line of each cluster represents the endogenous promoter showing the most similar activity with respect to the original expression profile for each contig. All expression profiles are properly identified at the most rightmost side of each line, following their respective contig/promoter name. For the supplementary set of analysed contigs, see **Supplementary Figure S4**.

Additionally, the current experimental approach allowed us not only to identify novel promoter regions but also to determine promoter directionality. The evaluation of promoter localization within the 10 selected clones revealed that from the 33 experimentally selected small fragments, 7 (21%) were considered intragenic promoters while the remaining 79% (26 promoters) were considered primary promoters, defined as the furthest upstream promoter in a gene/operon (Conway *et al.*, 2014). For the sake of comparison, *E. coli* K-12 genome presents the following proportions: primary (66.3%), secondary (19.6%), intragenic (9.8%), and antisense (4.2%) promoters (Cho *et al.*, 2009; Conway *et al.*, 2014).

Based on the alignment results, we selected a defined set of small fragment clones related to each original sequence for dynamic expression profiling on a microplate reader. The results showed that for each set of small-fragments belonging to a DNA metagenomic clone, there was at least one with an expression pattern corresponding to the original clone previously observed (**Figure 42** and **Supplementary Figure S4**). Similarly, we identified other clones bearing small-inserts with individual profiles different to the primarily observed, representing alternative promoter regions in the original sequence that were not mapped in the initial approach (**Figure 42**). Data has also shown that, in our experimental conditions, it seems that in each case a single promoter (usually the closest to the reporter gene) has the major contribution for the gene expression pattern observed. This can be concluded since, in each case, only one promoter mapped from the small-insert library produced the same expression profile observed for the original full-length fragment.

Regarding *in silico* cross-validation, from the 33 experimentally validated promoters, 23 RpoD-related promoters (70%) were supported by the algorithmic analysis as they were aligned to their respective original sequences (**Figure 42**). On the other hand, the remaining 10 sequences (30%) were considered as promoters exclusively identified by experimental approaches. This could indicate that these promoters that do not match the RpoD consensus are recognized by alternative sigma factors. This hypothesis will be investigated in future studies. Finally, sequences of the above experimentally validated promoters were characterised accordingly to previous studies reported in the literature. For this, we adopted an *in silico* classification proposed by Shimada et al (2014) (Shimada *et al.*, 2014), in which constitutive promoters present a high-level conservation of the consensus sequence for the major sigma factor RpoD, that is, the elements TTGACA (-35) and TATAAT (-10) separated by approximately 17 bp (**Figure 43A and B**). Constitutive promoters are defined as promoters active *in vivo* in all

circumstances, and, on the other hand, inducible promoters are switched ON and OFF by transcription factors depending on the *in vivo* conditions (Shimada, et al. 2014). The Logo pattern (Schneider and Stephens, 1990; Crooks *et al.*, 2004) generated from the alignment of the 33 identified metagenomic promoters (**Figure 43C**) indicated that positions -35 and -34 (-35 box) and positions -8, -7 and -3 (-10 box) were highly conserved. Additionally, when the promoters were analyzed in sub-groups based on the level of strength (high, medium and low), we could notice a variation in the consensus sequence obtained for each group (**Supplementary Figure S5 online**). These variances in the consensus sequences could explain the different promoter expression profiles observed experimentally.

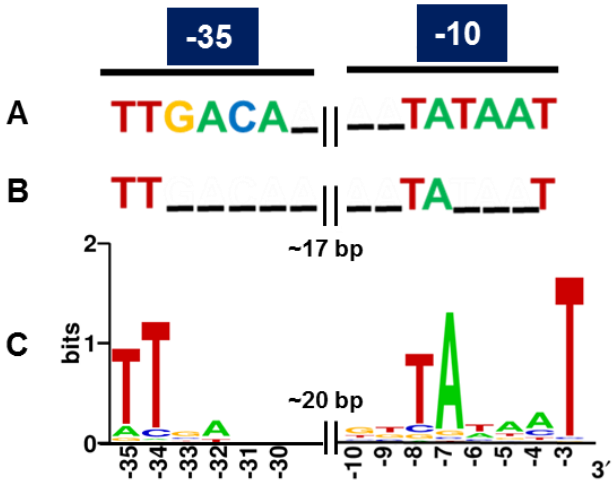


Figure 43. The consensus of RpoD-related metagenomic promoters. (A) Known consensus sequences of the RpoD-dependent promoter determined in vitro, TTGACA (-35) and TATAAT (-10) separated by 17 plus/minus 2 bp in *E. coli* (Shimada, et al. 2014). **(B)** Known consensus sequences of 582 promoters experimentally validated in *E. coli* (Shimada *et al.*, 2014; Gama-Castro *et al.*, 2016; Keseler *et al.*, 2017). **(C)** The sequences of the 33 promoters experimentally validated in this study were aligned and subjected to Logo analysis (Schneider and Stephens, 1990; Crooks *et al.*, 2004). The consensus from the metagenomic set **(C)** is very similar to the one from the experimentally validated set from *E. coli* **(B)**.

5. Discussion

5.1 Meta-expression profiles for studying microbial communities

The similar expression clusters found between the two independent metagenomic libraries might suggest broader trends of organizational expression patterns in nature. Independent studies on microbial communities from aquatic environments have described similar patterns by evaluating gene expression through metatranscriptomic analysis (Frias-Lopez *et al.*, 2008; Stewart, Ulloa and Delong, 2012; Dupont *et al.*, 2015; Fortunato and Crump, 2015), indicating that our observations are not restricted to the assessed soil samples. It has also been computationally demonstrated by Fernandez *et al.* (2014) (Fernandez *et al.*, 2014) that the microbial metaregulome – the whole set of regulons of an environmental sample – is shaped by the physicochemical conditions of the environment as an adaptive process. Thus, we suggest that expression profiling of an environmental sample might bear great potential for revealing insightful trends regarding the transcriptional diversity of microbial communities and for aiding on the design of efficient microbial communities for therapeutic or ecological needs (Fernandez *et al.*, 2014; Fredrickson, 2015; Solé, 2015; Johns *et al.*, 2016).

Regarding the explanation for the diversity expression profiles found among the metagenomic clones, it is important to stress that regulatory patterns have a multifactorial nature, being ruled by many different processes. Firstly, the regulatory dynamic is inherently interconnected with the function of the original regulated gene (e.g. housekeeping, adaptive etc.) (Wolf, Silander and Nimwegen, 2015). Secondly, the transcriptional bias imposed by the *E. coli* molecular machinery might constraint the recognition of promoter elements and/or not necessarily reproduce the original behaviours found in natural hosts (Gabor, Alkema and Janssen, 2004; Gabor, de Vries and Janssen, 2004; Liebl *et al.*, 2014; Guazzaroni, Silva-Rocha and Ward, 2015). Another point to be taken into consideration is that artificial juxtaposition of the exogenous promoter to the ribosome-binding site of the fluorescent reporter might increase expression as a consequence of the cloning process. Finally, another process that could influence the detection of active clones in *E. coli* is that the expression of many heterologous genes are toxic to this host (Kimelman *et al.*, 2012). This would also limit the cloning of some fragments in this host for functional metagenomics approaches.

Our observations also suggested transcriptional regulation beyond the control of the RpoD sigma factor for those clones (i. e. adjacent transcription factors), introducing novel niches for the exploration of regulated promoters. Since the discovery of distinct expression behaviours is essential for expanding the current set of commercial promoters, the diversity of expression profiles highlighted in this study has supported the current framework as a promising strategy for finding novel promoters for downstream applications. We also believe the developed strategy could greatly benefit from the combination with other high-throughput screening methods, such as SIGEX (Uchiyama *et al.*, 2005), providing innovative possibilities for the prospection of both inducible and constitutive promoters. Finally, we emphasize our observations are always constrained, to a certain extent, by the perspective of the chosen microbial host (Guazzaroni, Silva-Rocha and Ward, 2015; Lam *et al.*, 2015; Alves, Silva-Rocha and Guazzaroni, 2017) (i.e. the set of constitutive promoters active in *E. coli*) and might represent only a fraction of the effective environmental metaconstitutome. Future studies systematically applying our methodology to a range of environmental samples and hosts will greatly contribute to understanding this relationship between regulatory diversity and environmental adaptation in bacteria.

5.2 Regulatory architectures and host compatibility for promoter exploration

Through the generation of a small-DNA insert library combined to *in silico* platforms we were able to analyse taxonomic and architectural features of the metagenomic fragments. We have also provided both (i) a consensus of recognizable exogenous constitutive promoters in an *E. coli* host. The analysis of the metagenomic fragments for nucleotide composition were in agreement with previous G-C content diversity analyses of soil samples, which ranged from 50% to 61% (Foerstner *et al.*, 2005; Bohlin *et al.*, 2010; Mann and Chen, 2010) suggesting the environmental influence on G-C content and taxonomic predominance of microbiomes. Although phylogenetic affiliation based on ORFs at the protein level is not suitable as sequence-composition based classifiers – as *PhylopythiaS* - for predicting taxonomic origins, we could observe that there was an agreement between both methods in a few samples (e.g. pCAW3, pCAW6, pCAW9 and pCAW10). Furthermore, the abundance of bacterial groups and gene functions predicted in this work was also similar to previous high-throughput studies in soil microbial communities (Janssen, 2006; Fierer, Bradford and Jackson, 2007; Fierer *et al.*, 2012).

Considering the above, the proposed experimental methodology has allowed us to directly asses the different bacterial groups that had promoters sequence recognizable by the host – as the metagenomic fragments from these predicted taxa have allowed GFP expression in *E. coli*.

Regarding the in-depth search for promoters *in vivo* – small-DNA library - and *in silico*, the experimental finding of at least 33 promoter regions within the initial set of the selected metagenomic clones suggested the *in silico* prediction was overestimated (140 RpoD-related promoters). The above can be explained since it is not uncommon for prediction algorithms to underestimate or overestimate results due to a lack of information regarding diversity and variability of natural *cis*-regulatory sequences (Vanet, Marsan and Sagot, 1999; de Jong *et al.*, 2012; Shahmuradov *et al.*, 2016). Furthermore, the analysis of the metagenomic promoter positions/architectures has slightly diverged from the *E. coli* K-12 genome, suggesting the diversity of genomic architectures in metagenomic libraries and a current underestimation of bacterial intragenic promoters that goes far above the *E. coli* model.

Regarding the promoter consensus obtained from the small-DNA fragments, we hypothesized that these sequences could be either recognized by other sigma factors than RpoD or presented unusual consensus sequences for -10 and -35 boxes which have bypassed the algorithmic analysis. However, experimental validation in *E. coli* strains lacking diverse sigma factors genes should be necessary for a more accurate conclusion. Although the observed logo pattern was distant from the *E. coli* consensus proposed for the RpoD-dependent constitutive promoters identified *in vitro* (**Figure 43A** (Shimada *et al.*, 2014)), it was very similar to the previously described consensus from experimentally validated promoter (Mitchell, 2003) sets from RegulonDB (Gama-Castro *et al.*, 2016) and EcoCyc (Keseler *et al.*, 2017) databases (**Figure 43B**), suggesting a certain degree of degeneracy for the recognition of constitutive promoters in *E.coli*. Thus, it has allowed us to identify a consensus for exogenous promoter recognition in *E. coli*, which can be an important resource for defining host-dependent constraints in functional metagenomics. Yet, it is possible that promoters that do not match the known consensus for RpoD could be recognized by alternative sigma factors, but it should be further exploited in the future.

A seminal study in functional metagenomics provided by Gabor *et al* (2004) (Gabor, Alkema and Janssen, 2004) estimated on a theoretical basis that 40% of the enzymatic activities present in a soil metagenomic library could be readily accessed using *E. coli* as a host in an independent

gene expression mode. This prediction implies that at least 40% of the metagenomic promoters would also be recognized by *E. coli*. Contrastingly, recent empirical studies on *E. coli* and other hosts have shown that functional expression faces a myriad of challenges (Bernstein *et al.*, 2007; Ekkers *et al.*, 2012; Vester, Glaring and Stougaard, 2015), reflecting significantly lower rates than the proposed by Gabor and collaborators (Gabor, Alkema and Janssen, 2004). In agreement with those studies, our work stresses the gap between theoretical estimations and experimental results, as we have observed only a small portion of the whole set of promoters is accessible for *E. coli* in metagenomics libraries (~1% of the clones assayed displayed detectable fluorescence in the plates) - in contrast to the previously predicted enzymatic activities recovery rate (~40%) (Gabor, Alkema and Janssen, 2004). Thus, we remark the importance of generation predictions on a combination of both experimental and computational data.

5.3 Intrinsic challenges in functional metagenomic studies for promoter exploration

In order to address the constraints underlying our observations and predictions, we have selected some caveats raised during this study, which are intrinsic to functional metagenomics and regulatory studies. Firstly, functional metagenomics investigates a system – bacterial community – based on its genetic parts – metagenomic fragments –, thus it is limited to provide blurred (and somewhat biased) depiction of the whole – e.g. some promoters observed as constitutive might be repressed by the structural conformation of bacterial chromatin in the original organism (Dillon and Dorman, 2010), but not in the plasmidial context in the host. Secondly, the metagenomic host will always bias the results as it filters biological information according to its own molecular machinery (Guazzaroni, Silva-Rocha and Ward, 2015; Lam *et al.*, 2015; Alves, Silva-Rocha and Guazzaroni, 2017) – e.g. a promoter might be considered constitutive when its exogenous repressor is not expressed in the host. Another potential limitation of the strategy used here, is that the direct cloning of DNA fragments and screening for fluorescent clones would be biased toward the identification of promoters located near the fluorescent reporter. Yet, since we were able to identify promoters located more than 1kb away from the reporter gene, this potential limitation would not be a concerning issue here. Lastly, the line between constitutive and regulated promoters has become rather arbitrary among studies as it usually relies on the experimental design and concepts adopted by each research

group – e.g. some authors consider constitutive bacterial promoters as those that are active *in vivo* in all circumstances, while others define them as the promoters recognized *in vitro* by RNA polymerase RpoD holoenzyme alone in the absence of additional regulatory proteins (Shimada *et al.*, 2014).

6. Conclusions

In summary, we have focused on integrating experimental and *in silico* approaches to exploit the regulatory diversity from metagenomics DNA fragments by prospecting and characterizing novel promoter sequences in *E. coli*. From this, we were able to identify novel constitutive promoters using real-sized metagenomic DNA fragments, and a further dissection of individual clones allowed us to demonstrate that a number of internal promoters can be recognized by the host to drive gene expression *in vivo*. Further studies could be applied to exploit which type of sigma factors are contributing to the expression of the identifiable active promoter fragments. Despite the intrinsic limitations previously described, our strategy can be further optimized by high-throughput studies, which will be essential for expanding our current estimations into a more holistic landscape. Finally, we highlight that this work should be also suitable for the applied sciences, expanding the current biotechnological toolbox through the discovery and characterisation of novel regulatory features.

Chapter IV

Development of an open-source pipeline for the automatic construction of metabolic models using genomic data

1. Specific Background

1.1 Network reconstruction and biological knowledge

There is a large library of scientific publications that describe different model organisms' specific molecular features. Molecular biology's focus on knowing much about a limited number of molecular events changed once annotated genome sequences became available, leading to the emergence of a genome-scale point of view. Now, putting all available knowledge about the molecular processes of a target organism in context and linking it to its genome sequence has emerged as a grand challenge. Genome-scale network reconstructions were a response to this challenge.

The reconstruction process treats individual reactions as the basic elements of a network. To implement the metabolic reconstruction process, a series of questions need to be answered for each of the enzymes in a metabolic network. (i) What are the substrates and products? (ii) What are the stoichiometric coefficients for each metabolite that participates in the reaction (or reactions) catalyzed by an enzyme? (iii) Are these reactions reversible? (iv) In what cellular compartment does the reaction occur? (v) What gene(s) encode for the protein (or protein complex), and what is (are) their genomic location(s)? Genes are linked to the proteins they encode and the reactions they catalyze using the gene-protein-reaction relationship (GPR) (**Figure 44**). All of this information is assembled from a range of sources, including organism-specific databases, high-throughput data, and primary literature. Establishing a set of the biochemical reactions that constitute a reaction network in the target organism culminates in a database of chemical equations. Reactions are then organized into pathways, pathways into sectors (such as amino acid synthesis), and ultimately into genome-scale networks, akin to reads becoming a full DNA sequence. This process has been described in the form of a 96-step standard operating procedure (Orth, Thiele and Palsson, 2010). Today, after many years of hard work by many researchers, there exist collections of *gen*ome-scale *re*constructions (sometimes called GENREs) for a number of target organisms (Oberhardt, Palsson and Papin, 2009; Bordbar *et al.*, 2014), and established protocols for reconstruction exist (Orth, Thiele and Palsson, 2010) that can be partially automated (Thiele and Palsson, 2010; Agren *et al.*, 2012; Cuevas *et al.*, 2016).

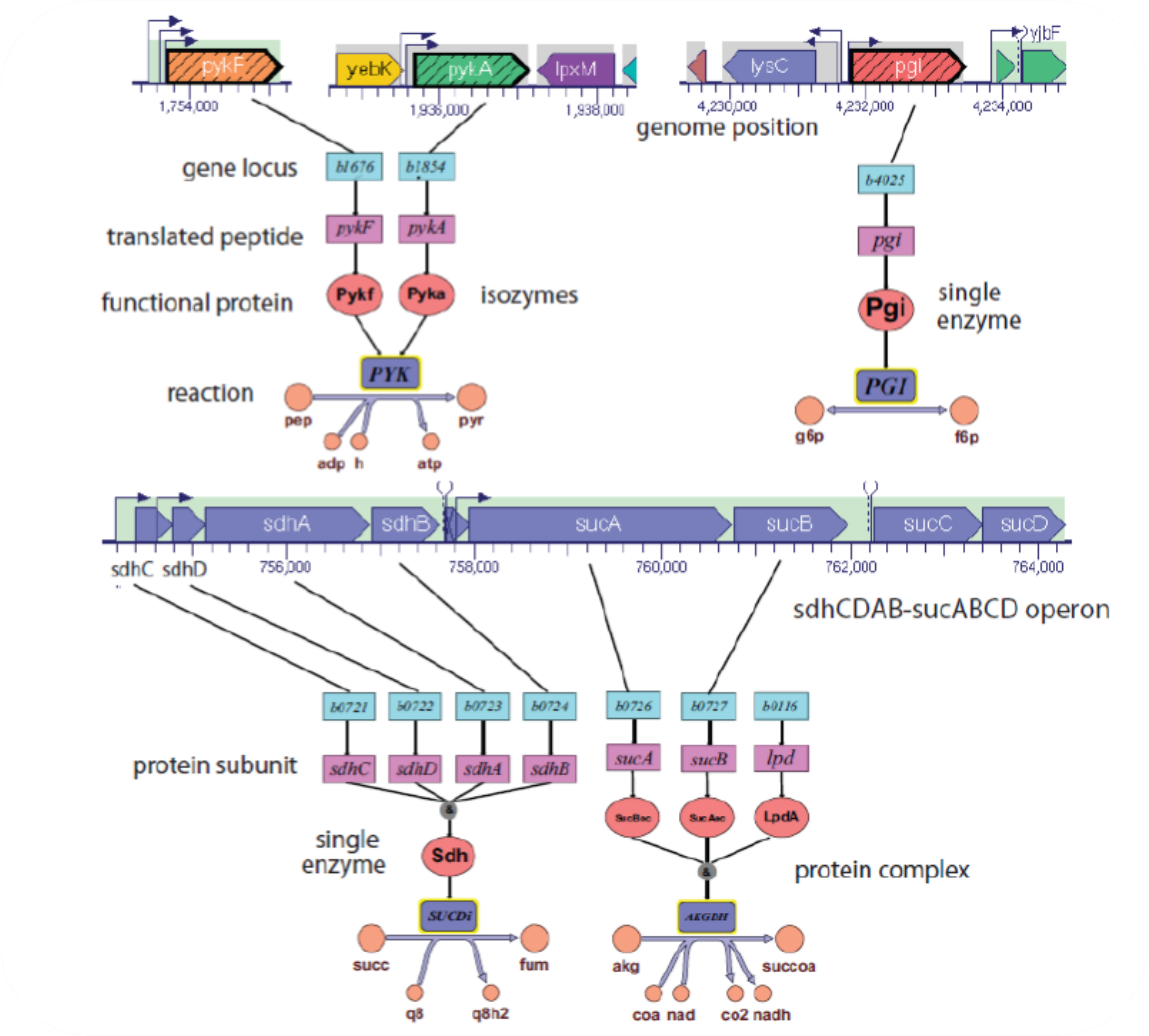


Figure 44 An example of GPRs from the *E. coli* core model. Genes are represented by blue boxes and designated by their locus name, translated peptides are represented by purple boxes, functional proteins are represented by red ovals, and reactions are labelled with blue boxes. For isozymes, two different proteins are connected to the same reaction. For proteins with multiple peptide subunits, the peptides are connected with an "&" above the protein. For complexes of multiple functional proteins, the proteins are connected with an "&" above the reaction. The genomic context of some of these genes is highlighted. Certain genes for the same reaction, e.g., pykF and pykA, are encoded by genes in operons widely separated on the genome. Operons are represented by shaded rectangles around one or more genes. Genes are represented by rectangles with one side pointed to denote the direction of the sense strand. Other operons contain multiple genes that encode protein subunits in a large protein. In this case, the same sdhCDAB-sucABCD operon that codes for the SUCDi proteins also codes for two proteins of the 2-oxoglutarate dehydrogenase enzyme complex, AKGDH. Genome context figures created by use of the Pathway Tools Genome Browser from EcoCyc. Retrieved from (Orth, Palsson and Fleming, 2010).

1.2 From Genome-scale Reconstruction to Computational Models

Before a reconstruction can be used to compute network properties, a subtle but crucial step must be taken in which a network reconstruction is mathematically represented (**Figure 45**). This conversion translates a reconstructed network into a chemically accurate mathematical

format that becomes the basis for a genome-scale model. This conversion requires the mathematical representation of metabolic reactions. The core feature of this representation is tabulation, in the form of a numerical matrix, of the stoichiometric coefficients of each reaction (**Figure 46**). These stoichiometries impose systemic constraints on the possible flow patterns (called a flux map, or flux distribution) of metabolites through the network. These concepts are detailed below. Imposition of constraints on network functions fundamentally differentiates the COBRA approach from models described by biophysical equations, which require many difficult-to-measure kinetic parameters.

Constraints are mathematically represented as equations that represent balances or as inequalities that impose bounds (**Figure 46**). The matrix of stoichiometries imposes flux balance constraints on the network, ensuring that the total amount of any compound being produced must be equal to the total amount being consumed at steady state. Every reaction can also be given upper and lower bounds, which define the maximum and minimum allowable fluxes through the reactions that, in turn, are related to the turnover number of the enzyme and its abundance. Once imposed on a network reconstruction, these balances and bounds define a space of allowable flux distributions in a network—the possible rates at which every metabolite is consumed or produced by every reaction in the network. The flux vector, a mathematical object, is a list of all such flux values for a single point in the space. The flux vector represents a "state" of the network that is directly related to the physiological function that the network produces. Many other constraints such as substrate uptake rates, secretion rates, and other limits on reaction flux can also be imposed, further restricting the possible state that a reconstructed network can take (Reed, 2012). The computed network states that are consistent with all imposed constraints are thus candidate physiological states of the target organisms under the conditions considered. The study of the properties of this space thus becomes an important subject.

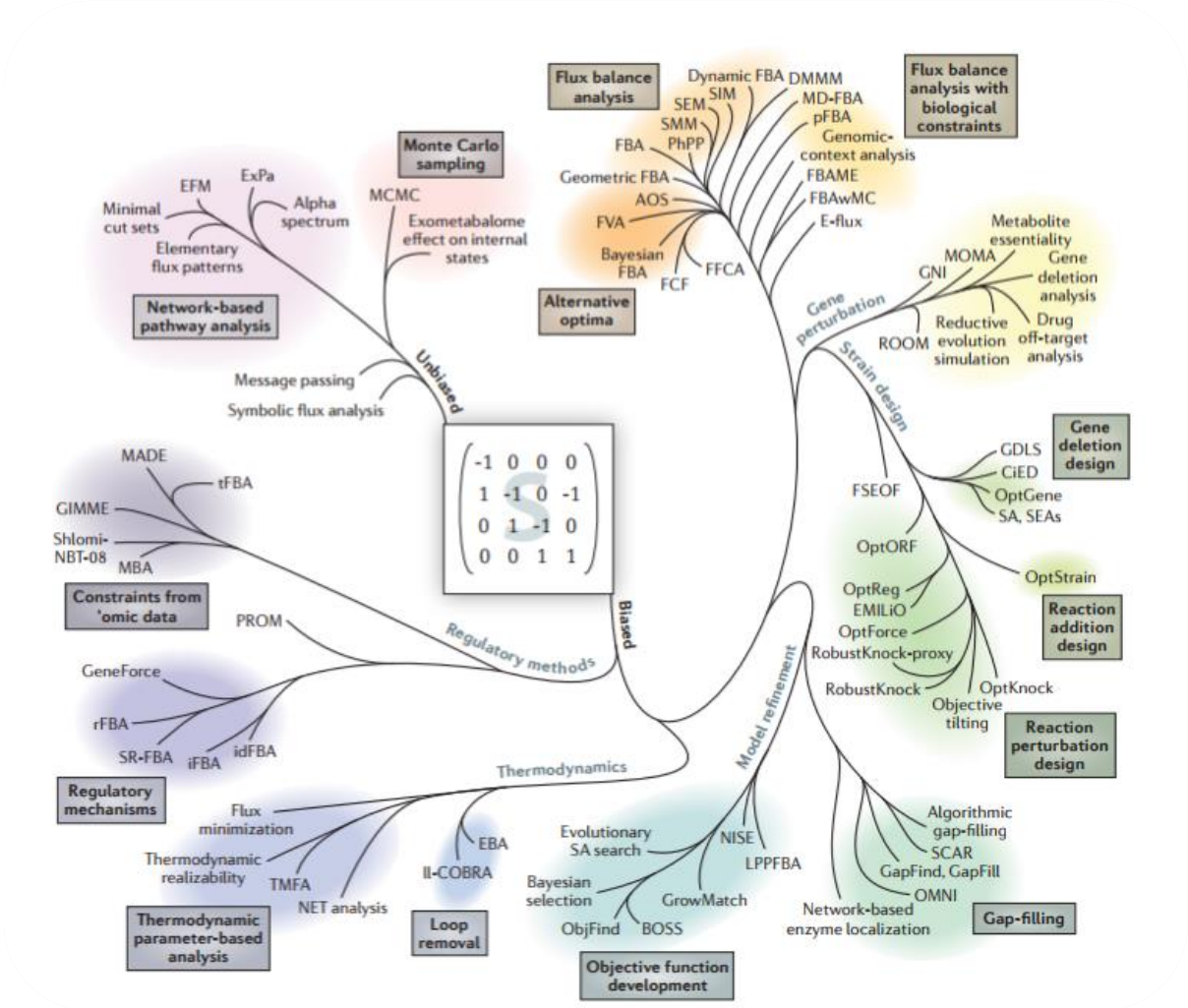


Figure 45. The "phylogeny" of constraint-based modelling methods. Over the past years, the constraint-based modelling community has rapidly expanded. Because of the versatility and scalability of these models, more than 100 methods have been developed for their modelling and analysis, all based on the analysis of the underlying metabolic network structure (i.e., the stoichiometric matrix). A phylogenetic tree is used to depict the similarities between application and use of the methods, and the underlying algorithms for many of the methods. Retrieved from (Lewis, Nagarajan and Palsson, 2012)

1.3 Flux balance analysis - FBA

Flux balance analysis (FBA) is the oldest COBRA method. It is a mathematical approach for analyzing the flow of metabolites through a metabolic network (Orth, Thiele and Palsson, 2010). This approach relies on an assumption of steady-state growth and mass balance (all mass that enters the system must leave). The constraints discussed above take the form of equalities and inequalities to define a polytope (**Figure 46**) that represents all possible flux states of the network given the constraints imposed. Thus, many network states are possible under the given constraints, and multiple solutions exist that satisfy the governing equations. The blue area is

therefore often called the "solution space" to denote a mathematical space that is filled with candidate solutions to the network equations given the governing constraints. FBA uses the stated objective to find the solution(s) that optimize the objective function. The solution is found using linear programming, and, as indicated in **Figure 46**, the optimal solution lies at the edges of the solution space impinging up against governing constraints.

The utility of FBA has been increasingly recognized due to its simplicity and extensibility: it requires only the information on metabolic reaction stoichiometry and mass balances around the metabolites under pseudo-steady state assumption. It computes how the flux map must balance to achieve a particular homeostatic state. However, FBA has limitations. It balances fluxes but cannot predict metabolite concentrations. Except in some modified forms, FBA does not account for regulatory effects such as activation of enzymes by protein kinases or regulation of gene expression.

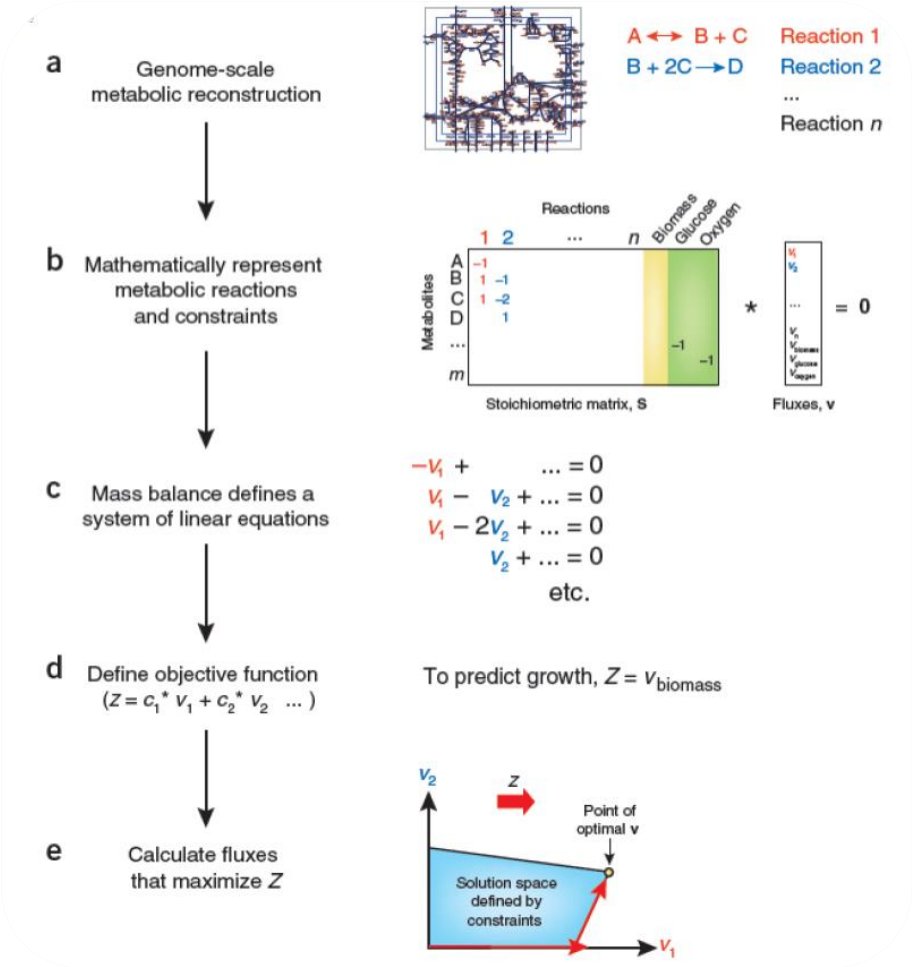


Figure 46. Formulation of an FBA problem. (a) A metabolic network reconstruction consists of a list of stoichiometrically balanced biochemical reactions. **(b)** This reconstruction is converted into a mathematical model by forming a matrix (labeled S), in which each row represents a metabolite and each column represents a reaction. Growth is incorporated into the reconstruction with a biomass reaction (yellow column), which simulates metabolites consumed during biomass production. Exchange reactions (green columns) are used to represent the flow of metabolites, such as glucose and oxygen, in and out of the cell. **(c)** At steady state, the flux through each reaction is given by Sv = 0, which defines a system of linear equations. As large models contain more reactions than metabolites, there is more than one possible solution to these equations. **(d)** Solving the equations to predict the maximum growth rate requires defining an objective function Z = cTv (c is a vector of weights indicating how much each reaction (v) contributes to the objective). In practice, when only one reaction, such as biomass production, is desired for maximization or minimization, c is a vector of zeros with a value of 1 at the position of the reaction of interest. In the growth example, the objective function is Z = vbiomass (that is, c has a value of 1 at the position of the biomass reaction). **(e)** Linear programming is used to identify a flux distribution that maximizes or minimizes the objective function within the space of allowable fluxes (blue region) defined by the constraints imposed by the mass balance equations and reaction bounds. The thick red arrow indicates the direction of increasing Z. As the optimal solution point lies as far in this direction as possible, the thin red arrows depict the process of linear programming, which identifies an optimal point at an edge or corner of the solution space. Retrieved from (Orth, Thiele and Palsson, 2010)

1.4 The importance of constraints

One of the most basic constraints imposed on genome-scale models of metabolism is that of substrate, or nutrient, availability and its uptake rate. Metabolites enter and leave the systems through what are termed "exchange reactions" (i.e., active or passive transport mechanisms) (**Figure 46** and **Figure 47**). These reactions define the extracellular nutritional environment and are either left "open" (to allow a substrate to enter the system at a specified rate) or "closed" (the substrate can only leave the system). Measurements of the rate of exchange with the environment are relatively easy to perform, and they prove to be some of the more important constraints placed on the possible functions of reaction networks internal to the cell. More biological- and data-derived constraints can also be imposed on a model.

The next step in converting a network reconstruction to a model is to define what biological function(s) the network can achieve. Mathematically, such a statement takes the form of an "objective function". For predicting growth, the objective is biomass production — that is, the rate at which the network can convert metabolites into all required biomass constituents such as nucleic acids, proteins, and lipids needed to produce biomass. The objective of biomass production is mathematically represented by a "biomass reaction" that becomes an extra column of coefficients in the stoichiometric matrix (**Figure 46** and **Figure 47**). One can formulate a biomass objective function at an increasing level of detail: basic, intermediate, and advanced

(Feist and Palsson, 2008; Bordbar *et al.*, 2014). The biomass reaction is scaled so that the flux through it represents the growth rate (μ) of the target organism.

It is important to note that the biomass objective function is determined from measurements of biomass composition—the uptake and secretion rates from measuring the nutrients in the medium—and that the model formulation is built on a knowledge-based network reconstruction. Thus, the growth rate optimization problem represents "big data" integrated into a structured format and the hypothesis of a biological objective: grow as fast as possible with the resources available. This is a well-defined optimization problem.

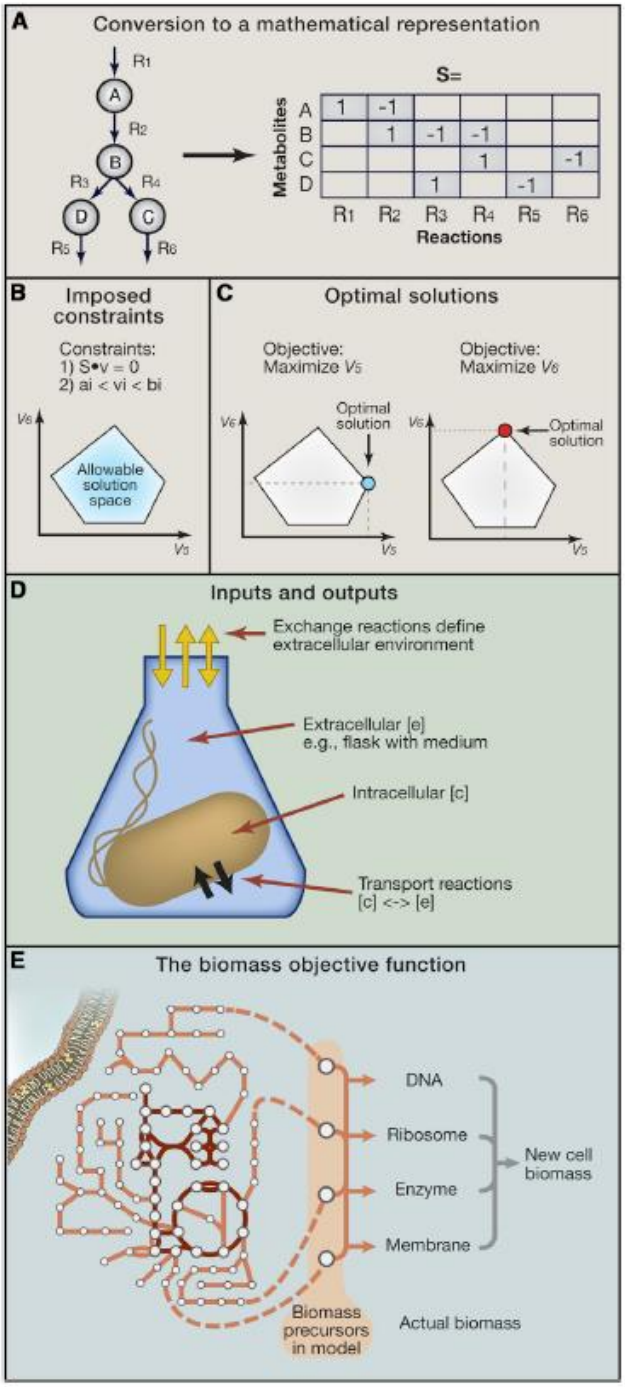


Figure 47. Formulation of a Computational Model (A) After the metabolic network has been assembled, it must be converted into a mathematical representation. This conversion is performed using a stoichiometric (S) matrix in which the stoichiometry of each metabolite involved in a reaction is enumerated. Reactions form the columns of this matrix and metabolites the rows. Each metabolite's entry corresponds to its stoichiometric coefficient in the corresponding reaction. Negative coefficient substrates are consumed (reactants), and positive coefficients are produced (products). Converting a metabolic network reconstruction to a mathematical formulation can be achieved with several of the toolboxes. **(B)** Constraints can be added to the model, such as: (1) enforcement of mass balance and (2) reaction flux (v) bounds. The blue polytope represents different possible fluxes for reactions 5 and 6, consistent with stated constraints. Those outside of the polytope violate the imposed constraints and are thus ''infeasible.'' **(C)** Constraint-based models predict the flow of metabolites through a defined network. The predicted path is determined using linear programming solvers and is termed flux balance analysis (FBA). FBA can be used to calculate the optimal flow of metabolites from a network input to a network output. The desired output

is described by an objective function. If the objective is to optimize flux through reaction 5, the optimal flux distribution would correspond to the levels of flux 5 and flux 6 at the blue point circled in the figure. The objective function can be a simple value or can draw on a combination of outputs, such as the biomass objective shown in (E). It is important to note that alternate optimal flux distributions may exist to reach the optimal state **(D)** Once a network reconstruction is converted to a mathematical format, the inputs to the system must be defined by adding consideration of the extracellular environment. Compounds enter and exit the extracellular environment via ‘‘exchange’’ reactions. The GEM will not be able to import compounds unless a transport reaction from the external environment to the inside of the cell is present. **(E)** In addition to exchange reactions, the biomass objective function acts as a drain on cellular components in the same ratios as they are experimentally measured in the biomass. In FBA simulations, the biomass function is used to simulate cellular growth. The biomass function is composed of all necessary compounds needed to create a new cell, including DNA, amino acids, lipids, and polysaccharides. This is not the only physiological objective that can be examined using COBRA tools. Retrieved from (O’Brien, Monk and Palsson, 2015)

1.5 The importance of generating accurate and consistent models

Ensuring the consistency and accuracy of all of the information available for a target organism is a grand challenge of genome-scale biology. Since model predictions are based on a network reconstruction that represents the totality of what is known about a target organism, such predictions are a critical test of our comprehensive understanding of the metabolism for the target organism. Incorrect model predictions can be used for biological discovery by classifying them and understanding their underlying causes. Performing targeted experiments to understand failed predictions is also a proven method for systematic discovery of new biochemical knowledge (Orth, Thiele and Palsson, 2010).

2. Objectives

General objective

The main objective of this project was to provide a quick, consistent and open-source pipeline for building, manipulating and visualising metabolic models, which can be further integrated into other organizational layers.

Specific objectives

(i) Generation of a quality-check algorithm for filtering and mass-balancing reactions from the KEGG database

(ii) Generation of a Universal Reaction Network (URN), the whole set of pre-checked reactions from KEGG

(iii) Generation of algorithms for using input data as queries against the KEGG URN for the generation of specific metabolic models

(iv) Generation of algorithms for improving the metabolic model consistency and for manipulating its data

(v) Testing the pipeline with toy-model reactions

(vi) Testing the pipeline with real genomic data from the NCBI database

3. Materials and Methods

The present work has been developed *in silico*, and its scripts have been written in one of the following languages: *R, MATLAB® or Python*. The general workflow is depicted below, in **Figure 48**.

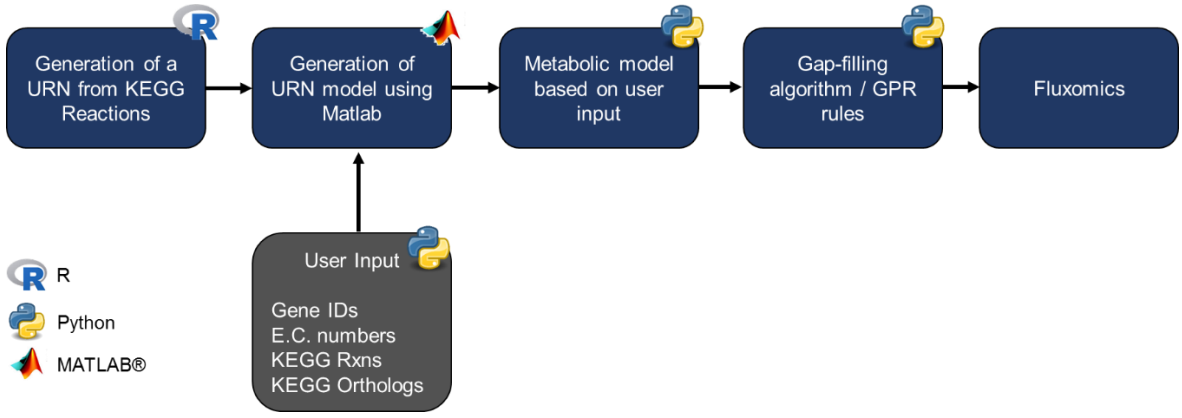


Figure 48. The general workflow of this work. The objective of this project was to provide a quick, consistent and open-source platform for the generation of GEMs. The user input should work as a query against our quality-checked KEGG Universal Reaction Network (URN). The URN should be manipulated for the generation of a draft metabolic model based on the user inputs. The resulting draft model should be structured for the application of constraint-based metabolic modelling methods such as Flux Balance Analysis (FBA) and for generating visual outputs. Additional pruning methods such as gap-filling algorithms and Gene-Protein-Reaction Boolean rules should also be available in order to improve the model consistency (Orth, Palsson and Fleming, 2010; Thiele and Palsson, 2010; Machado, Herrgård and Rocha, 2016). Each step in this workflow has used different programming languages represented by their logos (*R, Python* and *MATLAB®*)

Accessing and manipulating the necessary KEGG files for the generation of a KEGG-based Universal Reaction Network (URN)

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is one of the oldest and most comprehensive collection of databases. Its primary aim has been the digitising of current knowledge on genes and molecules and their interactions (Kanehisa, 2000) and it includes 16 databases and 3 sequence data collections (Kanehisa *et al.*, 2017). At the time of writing, the KEGG database lists 10,947 biochemical reactions and 18,354 compounds and the whole dataset can be downloaded via (paid) FTP access. This reaction-compound information can be represented as a stoichiometric *m* x *n* matrix with *m*-compounds and *n*-reactions. Thus, each reaction (column) will have non-zero values for their respective compounds (rows),

representing their mass-balanced stoichiometric coefficients. This bi-dimensional matrix can then be represented as a hypergraph, where nodes and edges correspond to compounds and biochemical reactions, respectively (McClymont and Soyer, 2013). We can refer to this hypergraph arising from the known universe of biochemical reactions as the 'Universal Reaction Network' (URN). Thus, KEGG "flat files" for KEGG Orthologs (KOs), enzymes, reactions and metabolites were downloaded via paid FTP access and further processed for the generation of a KEGG-based URN.

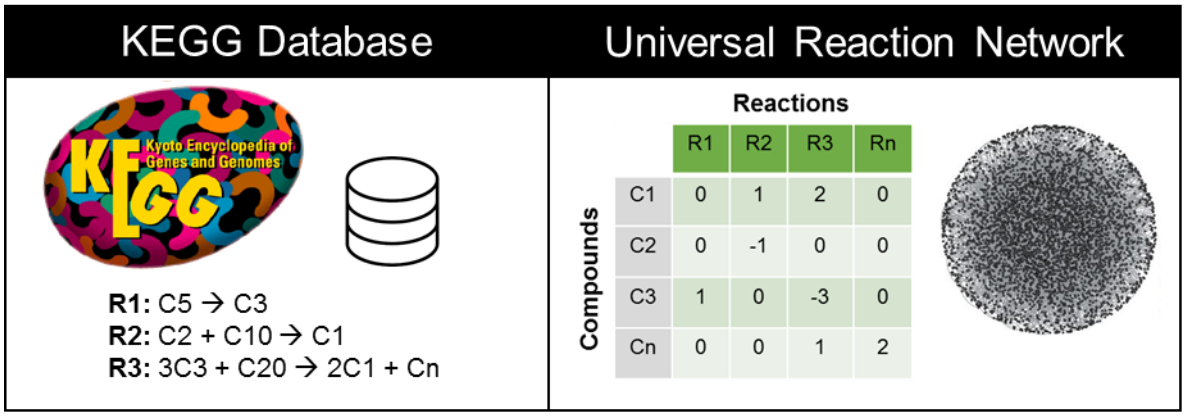


Figure 49. From KEGG Reactions to a KEGG-Based Universal Reaction Network (URN). KEGG "flat files" were downloaded via FTP access and the list of reactions was used to generate a Universal Reaction Network, a stoichiometric matrix containing all reactions and compounds from the database, which can also be represented as a hypergraph. Modified from (Kanehisa, 2000; McClymont and Soyer, 2013).

Generation of *ad-hoc* scripts in *R* for filtering and manipulating KEGG files

One of the most important issues related to the development of GEMs is the quality of the database data obtained. Inconsistencies such as mass-balancing errors in reactions, erroneous structural information and missing chemical formula for compounds and name inconsistencies for both reactions and compounds are quite common in databases and must be flagged and corrected before the development of a metabolic model (Schellenberger *et al.*, 2010; Kumar, Suthers and Maranas, 2012). Thus, before building a KEGG-based URN, we had to generate specific algorithms for checking the quality of KEGG compounds and reactions regarding the presence/absence of chemical formulas for KEGG compounds and the presence/absence of mass-balanced reactions (**Figure 50**).

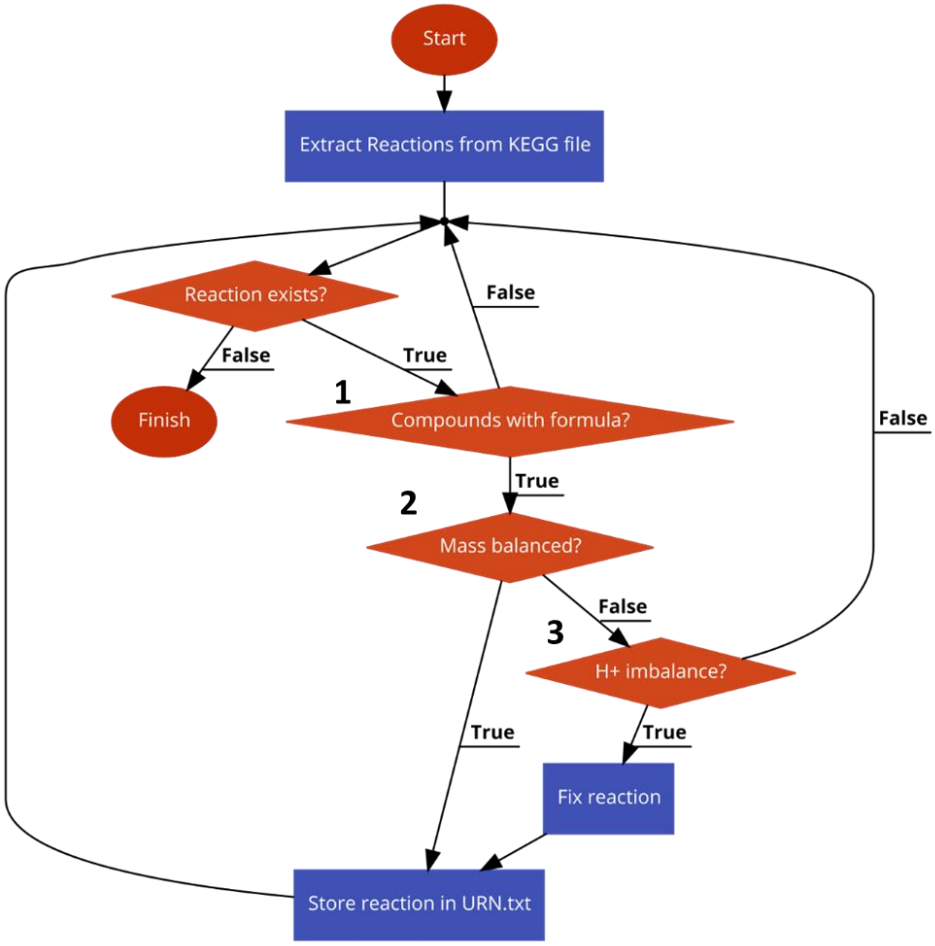
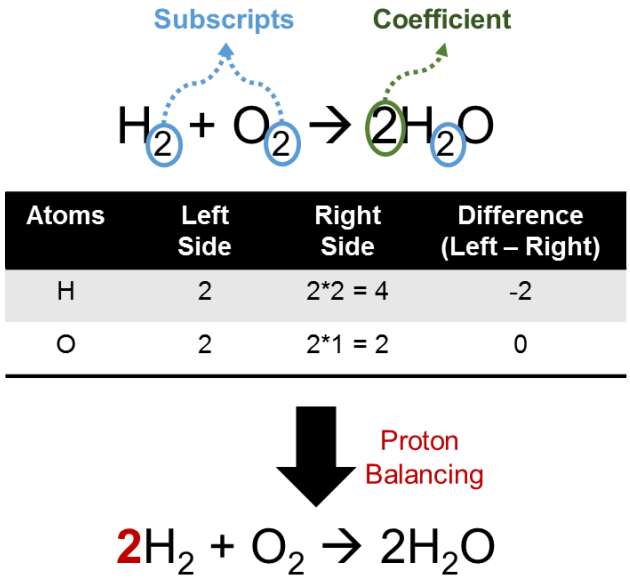


Figure 50. Flowchart representing the developed quality-checking algorithm. The quality-check of KEGG reactions was the first step in the development of the current pipeline. Firstly, the algorithm would search for reactions without compound formulas and remove these reactions; secondly, the algorithm would translate all reactions to their compound formula format and execute a mass-balancing function. Reactions which were not mass-balanced should be removed unless the mass-balancing issue can be solved by adding or removing protons (H^+). All the other kinds of mass-imbalanced reactions are removed. In the end, the URN will be composed by reactions which are both mass-balanced and with their compound formulas accessible.

The algorithm was implemented in *R* and worked as follow, KEGG compounds that did not present a corresponding chemical formula were flagged and all KEGG reactions containing these compounds were excluded from the URN generation process. Subsequently, all reactions with existing compounds' formula were subjected to a mass-balance checking (**Figure 51**). Each reaction was subdivided into left (reactants) and right (products) sides. The compound identifiers for each metabolite were then translated into their respective chemical formula and the subscripts of each atom were stored in a list (when the compound had a stoichiometric coefficient other than 1, the subscript was multiplied by it before being stored). Then, the vector containing all the subscripts for the atoms within the reactants side was compared to the one for

the products side. If each atom had the same value in both vectors the reaction was considered balanced (the difference between atoms from both sides should be zero). If not, the reaction was flagged. An additional step allowed us to check whether the reaction was imbalanced due to a proton imbalance (H^+) or not. If that was the case, the proton was added/removed to the correct side of the equation in order to mass-balance it. Mass-imbalanced reactions that could not be corrected by addition/removal of protons were flagged and excluded from the URN generation process as they would affect the process of metabolic modelling simulations (such as the linear optimization method adopted in FBA), resulting in inaccurate/faulty predictions (Thiele and Palsson, 2010).



Atoms	Left Side	Right Side	Difference (Left – Right)
H	2	2*2 = 4	-2
O	2	2*1 = 2	0

Figure 51. Checking for mass-balanced reactions. Each reaction was subdivided into left and right sides. The atoms of each side had their subscripts extracted (multiplied by the compound stoichiometric coefficient when needed) and compared. If there was any non-zero value in the comparison, the reaction was flagged as mass-imbalanced. If the reaction was proton imbalanced, a proton was added/removed to/from the correct equation side and the reaction was included into the URN generation process.

Using *ad-hoc* scripts in *MATLAB*® for generating of a KEGG-based Universal Reaction Network

As an initial proof of concept and in order to generate an object-structured URN, we have modified some COBRA toolbox (Heirendt *et al.*, 2017) functions in *MATLAB*® to be compatible with our quality-checked files for KEGG reactions and compounds. Thus, we have used these functions to generate not only the KEGG URN, but also a COBRA-structures stoichiometric model based on the previously filtered data (more information in

https://opencobra.github.io/). This model could be saved as a *.mat* file and further manipulated in other platforms such as the *COBRAPy* toolbox, an open-source tool based on the *Python* programming language (Ebrahim *et al.*, 2013).

Manipulating the URN Model through *ad-hoc* python scripts and functions from the *COBRAPy* toolbox (Ebrahim *et al.*, 2013).

The core capabilities of *COBRApy* are enabled by a set of classes (**Figure 52**) that represent organisms (Model), biochemical reactions (Reaction), and biomolecules (Metabolite and Gene). The core code is accessible through either Python or Jython (Python for Java). *COBRApy* contains: *cobra.io*: an input/output package for reading / writing SBML (Hucka *et al.*, 2003) models and reading / writing COBRA Toolbox MATLAB® structures; *cobra.flux_analysis*: a package for performing common FBA operations, including gene deletion and flux variability analysis (Ebrahim *et al.*, 2013); *cobra.topology*: a package for performing structural analysis; *cobra. test*: a suite of unit tests and test data; *cobra.solvers*: interfaces to linear optimization packages. And, *cobra.mlab*: an interface to the COBRA Toolbox for MATLAB®.

In this work, we have adopted the *COBRAPy* toolbox for our initial tests as it provided a practical framework, although our ultimate goal was to completely detach our framework from the object-oriented structures and keep it as simple yet robust as possible. As *COBRAPy* is an open-source toolbox, it has been extensively modified over the years by different users, which has led to a large number of deprecated and defective functions. In this work, we have updated and modified some of these functions in order to improve the efficiency of our mode generation pipeline. First, we have developed new functions for manipulating metabolic models through their stoichiometric matrices, using the *pandas* package (https://pandas.pydata.org/index.html). This has allowed us to develop functions for automatically removing (orphan compounds/reactions) and adding (artificial reactions such as exchange reactions) features into the model. We have also focused on the automatic generation of visual outputs by combining *COBRAPy* and network visualisation scripts provided by both the *D3flux Python* package (https://github.com/pstjohn/d3flux/blob/master/README.md) and the *Gephi* software (Bastian, Heymann and Jacomy, 2009).

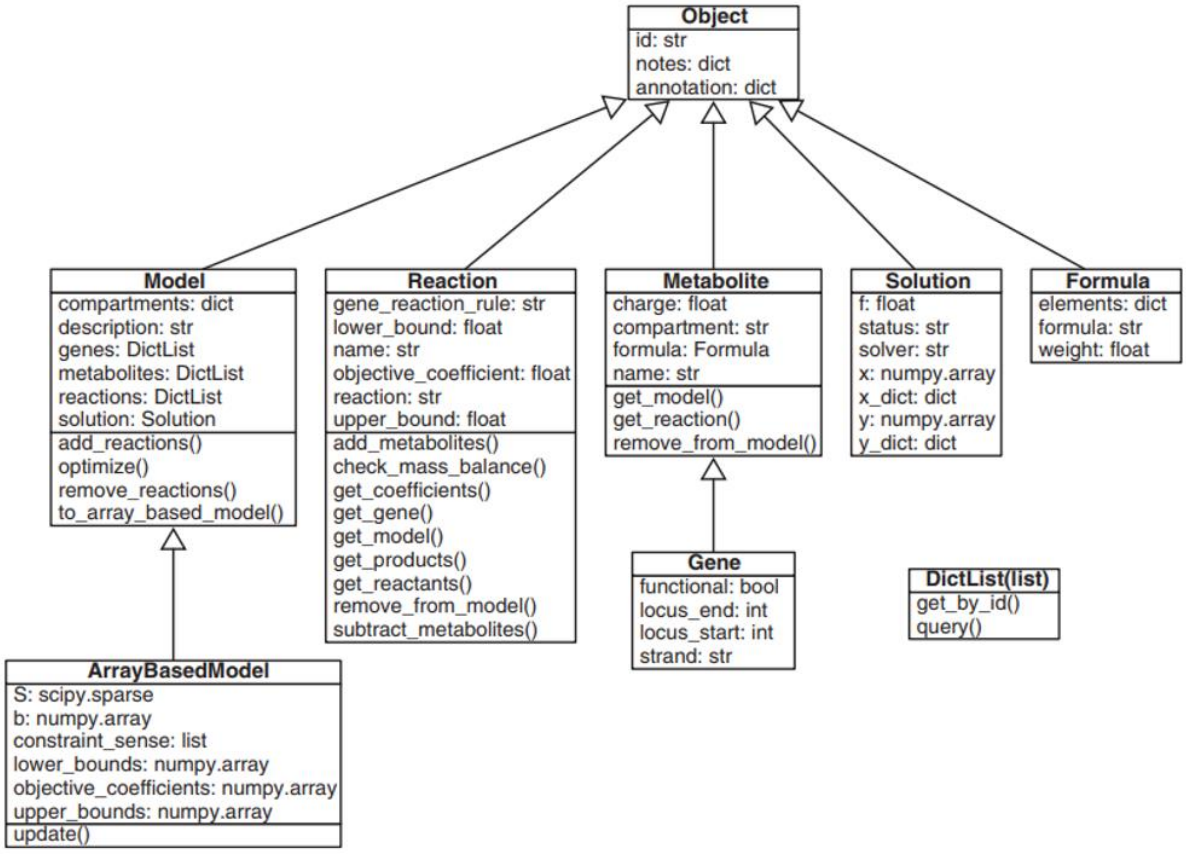


Figure 52 Core classes in COBRA for Python with key attributes and methods listed. The figure depicts a set of core classes that represent organisms (Model), biochemical reactions (Reaction), and biomolecules (Metabolite and Gene). The Solution class stores the numerical solution for FBA and other stoichiometric-based analysis. Retrieved from (Ebrahim *et al.*, 2013)

Testing the model through the adoption of toy reactions

In order to test our pipeline, we have chosen a set of nine toy reactions (artificial reactions) which were added to the original KEGG-based URN:

- **a.** R2: B <=>
- **b.** R3: P ->
- **c.** R4: E->
- **d.** R5: A->B
- **e.** R6: A->C
- **f.** R7: A->D
- **g.** R8: B<=>C
- **h.** R9: B->P
- **i.** R10: C + D -> E + P

The model generation is based on the manipulation of the URN based on user inputs (**Figure 53**). The inputs (gene names, KOs, E.C. numbers, KEGG reaction names) are used as queries

in a relational database for finding all the reaction IDs (reaction names) in the URN related to the user input. Then, all reactions and metabolites unrelated to the queries are removed from the URN and the resulting stoichiometric matrix is used for further pruning the model (**Figure 53**). The next step is to find and remove artefacts such as orphan reactions and compounds. In this sense, orphan compounds can be subdivided into top-node, bottom-node and disconnected orphans. Top-node compounds are only consumed, but never produced, while bottom-node compounds are never consumed. These compounds should be flagged and updated through the generation of artificial exchange reactions (EX_reactions) (**Figure 53**). Artificial reactions have an important role in metabolic models as they allow the flow of the metabolites into and from the model, directly affecting the results of constraint-based analysis such as FBA. It also raises deep implications in the biological relevance of exchange reactions and in the importance of manually curating them once the model has been built for obtaining both biologically consistent behaviours and reliable predictions. The last class of orphans, the disconnected one, consists of compounds which are not connected to any reactions in the metabolic network (**Figure 53**). These metabolites are flagged and the removed from the model in order to avoid numerical issues with subsequent simulations. It is important to highlight that in order to test the pipeline and the automatic identification of orphans, reaction R1 (an exchange reaction for metabolite A) was missing from the previously described set of toy reactions. Without this reaction, the model cannot be fully optimized (see **Figure 57** in the Results section for more information).

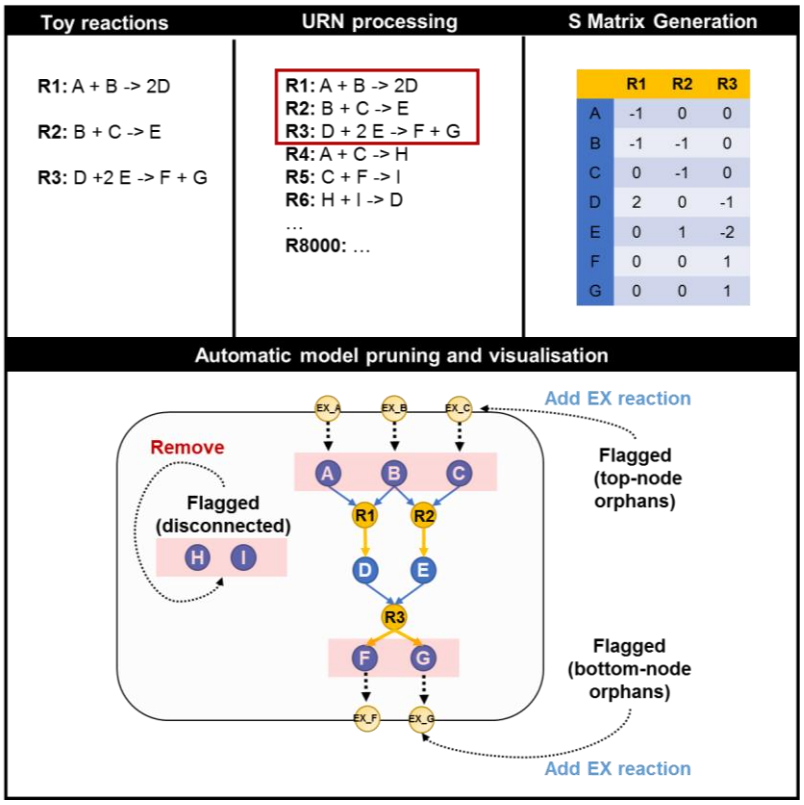


Figure 53. Representation of the toy-model generation for testing the developed pipeline. Top, from left to right: A set of 3 reactions was added to the original URN. The name (ID) of the added reactions were then used as inputs (queries) by the user. All reactions in the URN except for the ones provided by the user were deleted. A series of steps for checking and increasing the model's consistency was then applied: Generation of an S matrix, finding and flagging orphan reactions and metabolites (only produced, only consumed or disconnected nodes), creation of exchange reactions for connected orphan metabolites (top and bottom nodes), removal of unused compounds, linear optimisation and knockout testing followed by the generation of visual outputs representing the network and its current fluxes.

Testing the model generation pipeline through the use of real genomic data

After testing the proof of concept with the toy-reactions network, we have focused on developing genome-scale metabolic models using genomic data from NCBI as inputs (https://www.ncbi.nlm.nih.gov/genome). We have retrieved the genomes of three bacteria:

Escherichia coli K-12

https://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521

Bacillus subtilis strain 168

https://www.ncbi.nlm.nih.gov/genome/665?genome_assembly_id=300274

Mycoplasma genitalium strain G37

https://www.ncbi.nlm.nih.gov/genome/474?genome_assembly_id=300158

Using an *ad-hoc* script written in *R*, the gene IDs and their aliases were extracted and used as queries against our KEGG relational database. This resulted in the extraction of all reactions associated to these aliases, as our KEGG "flat files" have allowed us to link NCBI gene IDs to KEGG Orthologs, reactions and E.C. numbers. Thus, the same process as previously described with the toy-reaction network was adopted for building the metabolic model (**Figure 54**). All reactions and metabolites unrelated to the queries were removed from the KEGG-based URN and the resulting stoichiometric matrix was used for further pruning the model. The next step was to find and remove artefacts such as orphan reactions and compounds. Exchange reactions were created for top-node and bottom-node compounds, while the disconnected compounds were removed from the model in order to avoid numerical issues with subsequent simulations.

An *ad-hoc* script in *Python* was created to directly output the genome-scale stoichiometric matrix as a bipartite network in two different manners. The first one directly used the *D3flux Python* package (https://github.com/pstjohn/d3flux/blob/master/README.md) and the second one generated a *.csv* table already formatted for being used as an input into the *Gephi* software (Bastian, Heymann and Jacomy, 2009).

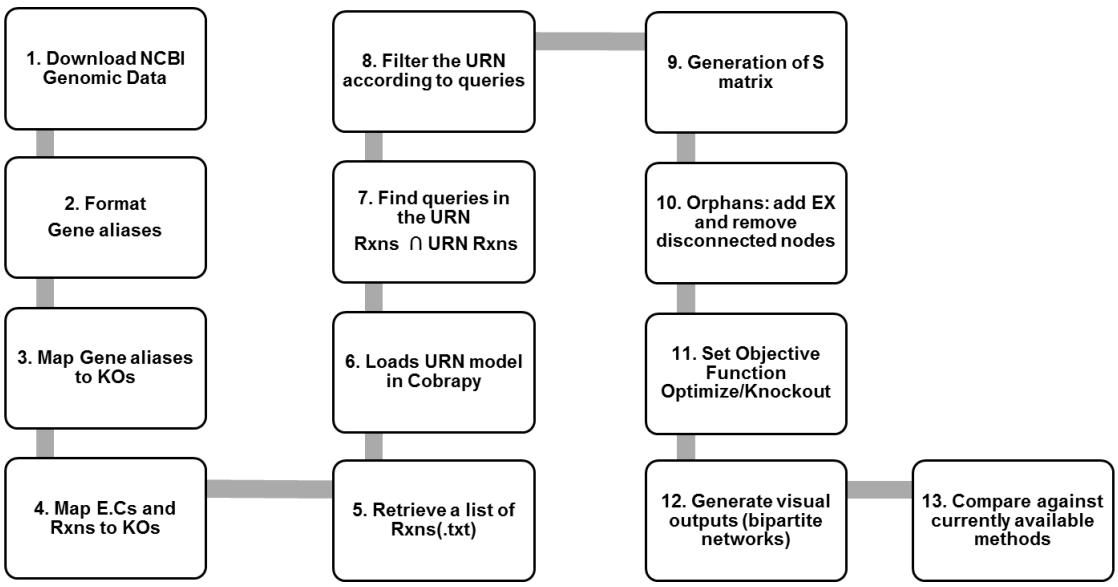


Figure 54. Workflow for generating genome-scale metabolic models. The pipeline for the automatic generation of GEMs from genomic data is represented in 13 steps. Firstly, the desired genomic sequence is downloaded from NCBI and its gene IDs/gene aliases are retrieved. These IDs are mapped to KEGG orthologs in the KEGG files and then to their respective Reactions and E.C.s in the URN. Everything is removed from the URN except the reactions mapped to the genes presented by the user. An S matrix is generated from the remaining reactions and compounds and automatically updated for the generation of Exchange Reactions (EXs) and removal of disconnected metabolic nodes. The user can then set its own objective function for running the FBA method. The user can also generate a bipartite network file that can

be viewed in network visualization software such as *Gephi* or *Cytoscape*. Finally, the generated model can be compared to published models for consistency checking.

Comparing generated and published genomic-based metabolic models

In order to assess the general properties of the generated models and compare them to the literature-based data, we have selected published metabolic models for the three bacterial species previously described. Our generated *E. coli* model was compared to one complete published model – iAF1260 (Feist *et al.*, 2007), to the *E. coli* core model (Orth, Palsson and Fleming, 2010) and to a Model-SEED reconstruction. Our *B. subtilis* model was compared to two published models, iBsu1103 (Henry *et al.*, 2009) and to iYO844 (Oh *et al.*, 2007). Lastly, our *M. genitalium* model was compared to the iPS189 published model (Suthers *et al.*, 2009) and to a Model-SEED reconstructed model.

Here, it is important to highlight a great challenge faced by GEMs: the comparison between models for the same organisms. For example, currently, there are more than ten published metabolic reconstructions for *E. coli* and over 200 peer-reviewed studies assessing *E. coli* capabilities through model-driven methods (McCloskey, Palsson and Feist, 2013). However, it is very difficult to compare these models as different laboratories usually manually curate many reactions, creating artificial exchange reactions and adopting different databases for the reconstructions with diverging IDs for its reactions and compounds. In this context, we have focused on comparing Enzyme Commission numbers between models (**Figure 55**) - a numerical classification scheme for enzymes, based on the chemical reactions they catalyse (International Union of Biochemistry and Molecular Biology, 1992). As a system of enzyme nomenclature, every E.C. number is associated with a recommended name for the respective enzyme. Strictly speaking, E.C. numbers do not specify enzymes, but enzyme-catalyzed reactions. If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same E.C. number. Thus, by comparing E.C.s from different models we might obtain a more general view of the enzymatic functions in each reconstruction.

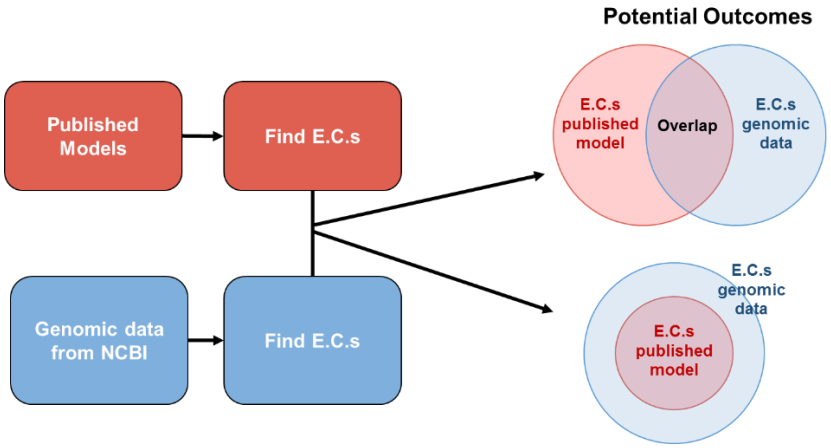


Figure 55 A method for comparing metabolic models from different pipelines. As different published models might have the same reactions, but with different names and different compound identifiers, it turns out to be very difficult to directly compare models generated from different pipelines and laboratories even when they represent the same organism. Thus, a potential solution is to compare the set of E.C.s (Enzyme Comission numbers (International Union of Biochemistry and Molecular Biology, 1992)) in each model as they provide a snapshot of the enzymatic functions present in each model. The analysis of overlapping E.Cs between models can be easily done and ultimately plotted as a Venn diagram for a more visual analysis.

4. Results

The first approach was to apply the pipeline for filtering the KEGG reaction files as shown in **Figures 50 and 51**. The pipeline has successfully removed reactions without compounds (1,126 reactions) and the mass-imbalanced ones (only those which could be mass-balanced for H^+ were fixed). From a total of 10,444 reactions, the dataset has been narrowed down to 8,461 reactions **(Figure 56).**

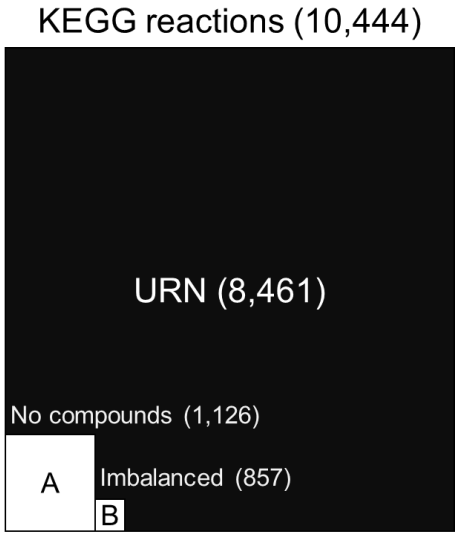


Figure 56. Building the Universal Reaction Network from KEGG. The first step for building the URN was to apply the quality-check algorithm (**Figures 50 and 51**). As a result, the total number of reactions has been narrowed down from 10,444 to 8,461 reactions.

The next step was to test the pipeline with a toy model. A set of 9 artificial reactions was added to the current URN (see Materials and Methods section). Then, the IDs for the 9 reactions were used as inputs into the pipeline as shown in the Materials and Methods section (**Figure 53)**. External reactions were automatically generated and disconnected orphan metabolites/reactions were excluded from the model. The D3flux package (https://github.com/pstjohn/d3flux/blob/master/README.md), a d3.js based visualization tool for *COBRAPy* (Ebrahim *et al.*, 2013) models was integrated into the pipeline, allowing the visualisation of the generated metabolic networks and the reaction fluxes (**Figure 57A**). The model was then successfully optimized for Reaction 4 (R4) (**Figure 57B)** and subsequently knocked out for Reaction 8 (R8) (**Figure 57C)**.

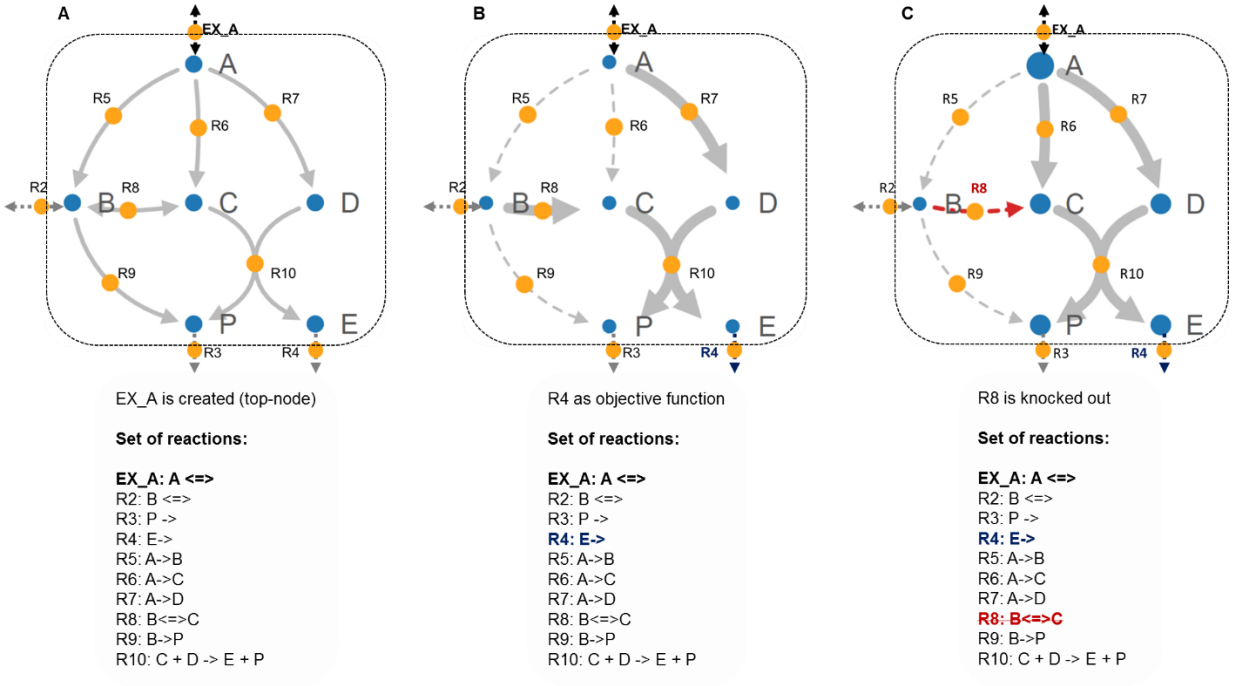


Figure 57. Exploring the pipeline through toy metabolic models. A set of 9 reactions was used as queries or the pipeline and resulted in a metabolic model with a total of 10 reactions (one exchange reaction was automatically added as part of the algorithm testing strategy). Metabolites are represented as blue circles, while reactions are represented as yellow circles. The arrows represent the reaction directionality (from reagents to products) and the arrows width represent how much flux is been passed through each reaction. Dashed arrows represent inactive reactions (distribution of flux is zero) and red dashed arrows represent knocked out reactions (flux is changed to zero). **(A)** Visualisation of the resulting toy model through the integration of the D3flux package into our method without any specific analysis/optimisation process. **(B)** Visualisation of the resulting toy model after optimizing the system towards the R4 reaction. **(C)** Visualisation of the resulting toy model after optimizing the system towards the R4 reaction and knocking out the R8 reaction. It is important to notice how the distribution of fluxes (arrow width) have changed from **(B)** to **(C)**.

After successfully testing the pipeline with the toy reaction set, we focused on further exploring its capabilities by testing it with real genomic data. Thus, we have followed the workflow explained in the Materials and Methods (**Figure 54**) to generate genome-scale metabolic models (GEMs) for three widely studied bacterial species: *E. coli* K-12 strain, *Bacillus subtilis* strain 168 and *Mycoplasma genitalium* strain G37. The generated models and their properties were compared to other published models for these organisms or to models automatically generated from the Model-SEED online platform - the most widely used of the existing frameworks for automated GEM reconstruction (Feist *et al.*, 2007; Oh *et al.*, 2007; Henry *et al.*, 2009; Suthers *et al.*, 2009; Overbeek *et al.*, 2014). Our generated *E. coli* model (1,272 reactions) was compared to one complete and manually curated published model – iAF1260 (Feist *et al.*, 2007) (2,382 reactions), to the *E. coli* core model (Orth, Palsson and Fleming,

2010) (bearing only 95 core metabolic reactions for educational purposes) and to a Model-SEED reconstruction (1,576 reactions). Our *B. subtilis* model (1,028 reactions) was compared to two published models, iBsu1103 (Henry *et al.*, 2009) (1,437 reactions) and to iYO844 (Oh *et al.*, 2007) (1,020 reactions). Lastly, our *M. genitalium* model (178 reactions) was compared to the iPS189 published model (Suthers *et al.*, 2009) (264 reactions) and to a Model-SEED reconstructed model (500 reactions). A summary of the comparison between the different models can be seen in **Table 7**. It is important to notice how the different models for the same species might differ from each other depending on the pipeline that generated it. Usually, published models have more reactions as they have manually curated artificial reactions such as exchange, biomass and maintenance reactions.

Table 7. Comparison between reconstructed and published metabolic models

Models/ Properties	***Escherichia coli K-12***				***Bacillus subtilis 168***			***Mycoplasma genitalium G37***		
	iAF1260	**Ours**	**SEED**	**Core**	**iYO844**	**iBsu1103**	**Ours**	**iPS189**	**SEED**	**Ours**
E.C.s (Enzyme Commission numbers)	522	832	721	37	448	740	638	103	89	103
E.C.s Overlap (against our model)	80%	-	76%	92%	75%	58%	-	50%	77%	-
Reactions	2382	1272	1576	95	1020	1,437	1028	264	499	178
Metabolites	1039	1669	1629	72	988	1140	1102	274	627	245
Genes	1260	1016	1017	137	844	1,103	822	189	222	133

The reconstruction of metabolic networks from genomic data has allowed us to compare our models to the published ones by two main methods: by the degree of E.C.s overlapping and by visually comparing bipartite networks representing both metabolites and reactions in each

model. We preferred the former method as it provided a general overview of the enzymatic functions in each model, while the latter method did not provide a clear distinction between the metabolic networks.

The comparison between the E.Cs of *E. coli* models has shown that our model, although presenting many more E.Cs than the published model was not able to encompass all the E.Cs found on published models, presenting, in average, 82% of the E.Cs found in published models (see **Figure 58**). Our *B. subtilis* was able to accommodate an average of 66,5% of the published models (**Figure 59**) while our *M. genitalium* comprised an average of 63,5% (**Figure 60**) of the published one. It is important to notice that even between published models we could observe large discrepancies such as an overlap of 90% between E.Cs from *B. subtilis* published models, being the iBsu1103 model twice the size of the iYO844 model in regard to E.C. numbers (740 E.C.s Vs 448 E.C.s).

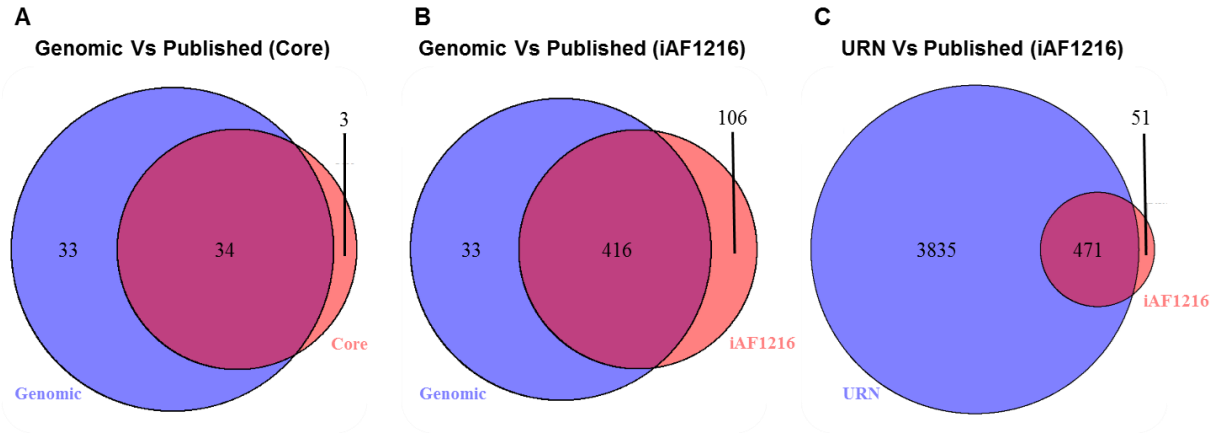


Figure 58. Comparison of E.C.s content between reconstructed and published *E. coli* K-12 stoichiometric metabolic models. (A) A reconstructed core model based on our pipeline (KEGG, in blue) with 67 E.C.s and the published core model (Core, in orange) with 37 E.C.s. **(B)** A reconstructed genomic model based on our pipeline (KEGG, in blue) with 832 E.C.s and the published iAF1260 model (iAF1260, in orange) with 522 E.C.s. **(C)** A positive control for the analysis, consisting in the URN (URN, in blue) with 4,306 E.C.s and the published iAF160 model (iAF1260, in orange) with 522 E.C.s.

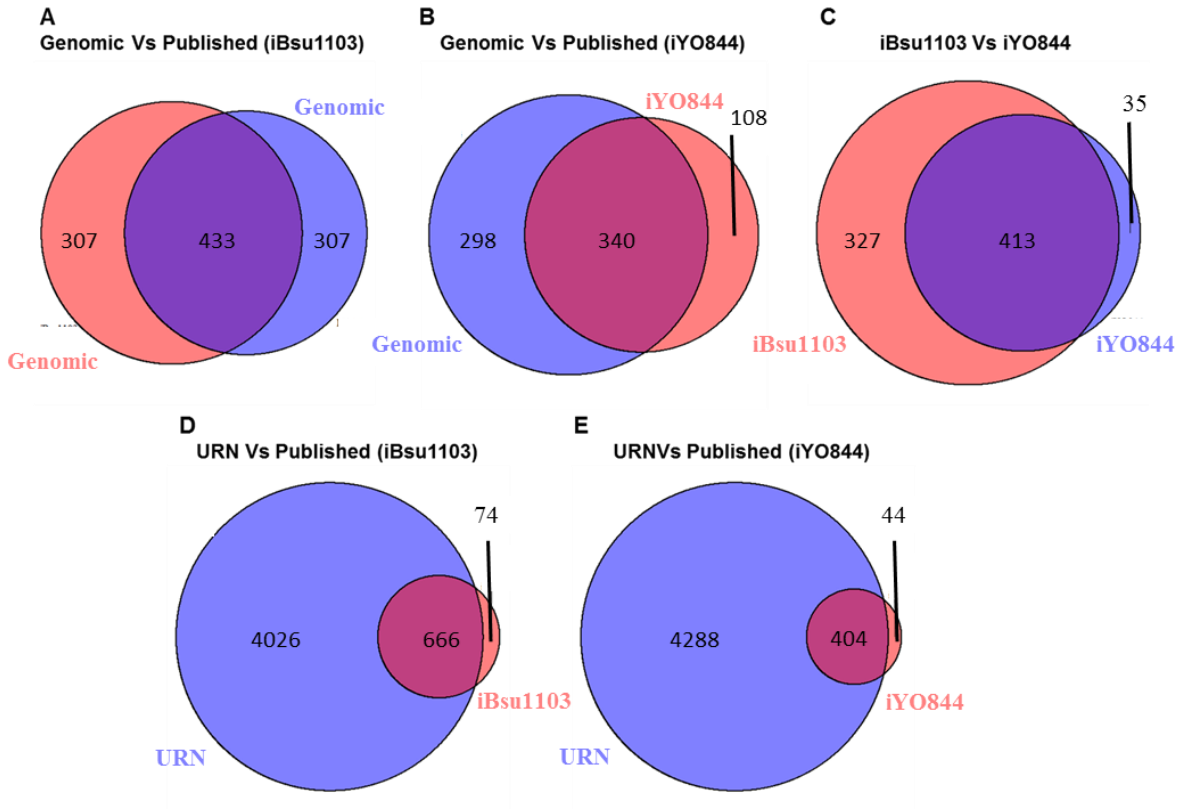


Figure 59. Comparison of E.C.s content between reconstructed and published Bacillus subtilis models **(A)** A reconstructed model based on our pipeline (KEGG, in blue) with 638 E.C.s and the published iBsu1103 model (iBsu1103, in orange) with 740 E.C.s. **(B)** A reconstructed genomic model based on our pipeline (KEGG, in blue) with 638 E.C.s and the published iYO844 model (iYO844, in orange) with 448 E.C.s. **(C)** A comparison between both published models. **(D)** A positive control for the analysis, consisting in the URN (URN, in blue) with 4,306 E.C.s and the published iBsu1103 model (iBsu1103, in orange) with 740 E.C.s. (E) A positive control for the analysis, consisting in the URN (URN, in blue) with 4,306 E.C.s and the published iYO844 model (iYO844, in orange) with 448 E.C.s.

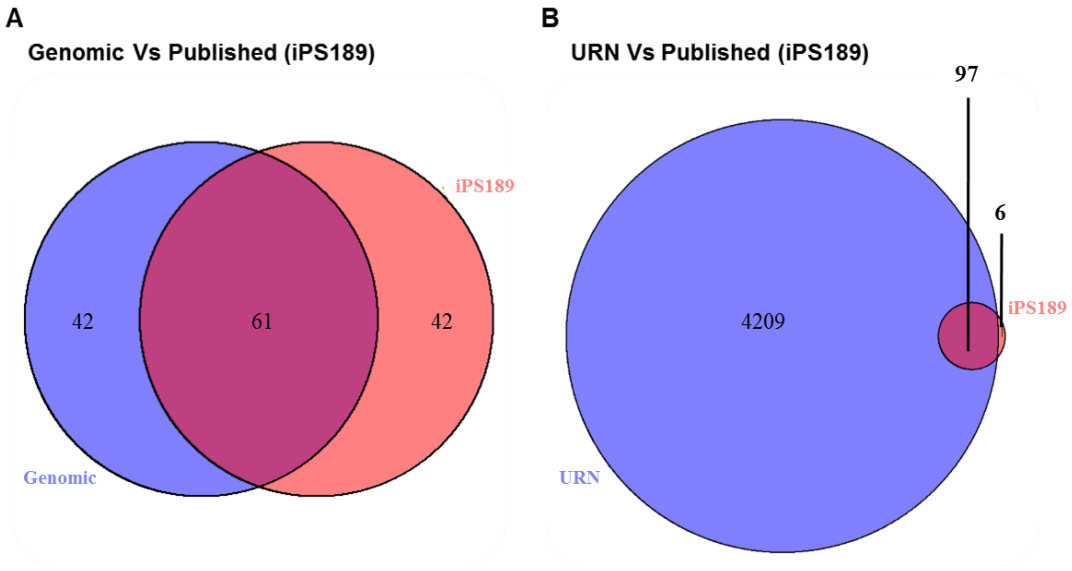


Figure 60. Comparison of E.C.s content between reconstructed and published Mycoplasma genitalium stoichiometric metabolic models. (A) A reconstructed model based on our pipeline (KEGG, in blue) with 103 E.C.s and the published iPS189 model (Published, in orange) with 103 E.C.s. **(B)** A positive control for the analysis, consisting in the URN (URN, in blue) with 4,306 E.C.s and the published iPS189 model (iPS189, in orange) with 103 E.C.s.

The last step of the analysis was to generate visual outputs for the reconstructed models through an automated process (see **Materials and Methods** section for more information). An *ad-hoc* algorithm for the generation of bipartite networks based on the S matrix data was developed, allowing the creation of *-.csv* files compatible to network visualizer software such as *Gephi* and *Cytoscape*. A bipartite graph, also called a bigraph, is a set of graph vertices decomposed into two disjoint sets such that no two graph vertices within the same set are adjacent (Newman, 2010). In the context of this project, a bipartite graph is represented by a directed graph with nodes for both reactions and metabolites. Metabolites are never connected between themselves, but only to reactions (see **Figure 53** in the Materials and Methods section). This visualisation process was coupled to the model generation pipeline, resulting in metabolic networks for all models considered here (see **Figures 61 to 64**). As a proof of concept, we have generated these metabolic networks for our own generated models and for published models of *E. coli* and *M. genitalium*. They have allowed us to visually compare network properties such as the degree of each node - represented as relative node sizes in **Figures 61 to 64** - and the network modularity – represented as the nodes that are more densely connected together than to the rest of the network sharing the same color F**igures 62 and 64** - in a straightforward manner. Furthermore, other network properties such as size, density, average degree, average path length, diameter of

a network, clustering coefficient etc. (Newman, 2010) could be calculated using software such as *Gephi* and *Cytoscape*.

By comparing our models with published ones, we could observe that the network structure was very similar for both *E. coli* and *M. genitalium* in terms of the statistics of the network properties, similar to the general patterns observed by (Oltvai *et al.*, 2000; Mahadevan and Palsson, 2005) in the study of over 40 metabolic reconstructions representing the three domains of Life. In this context, we could also see similarities between our models and the published ones in terms of the nodes with higher degrees, usually related to metabolites such as ATP, H+, H2O and NADH (bigger nodes in **Figures 61-64**). The modularity patterns between our models and the published ones were also quite similar for both organisms (**Figures 62 and 64**), indicating that the general topology and structure of our works are similar to the published ones despite the previously observed variations in the E.C.s content.

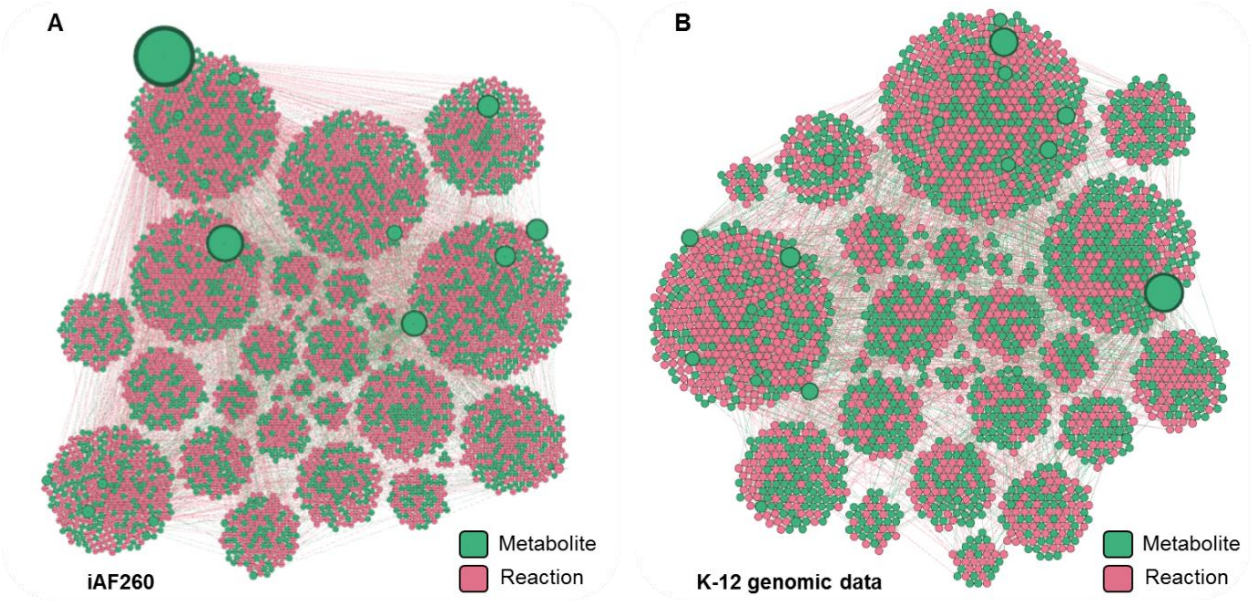


Figure 61. Reconstructed bipartite metabolic networks for *Escherichia coli* K-12. The networks were generated for both the **(A)** iAF1260 model and the **(B)** K-12 reconstructed model. Node colours represent Metabolites or Reactions while node size represents the number of connections of each node. The most connected nodes are usually related to metabolites such as ATP, H^+, H2O and NADH.

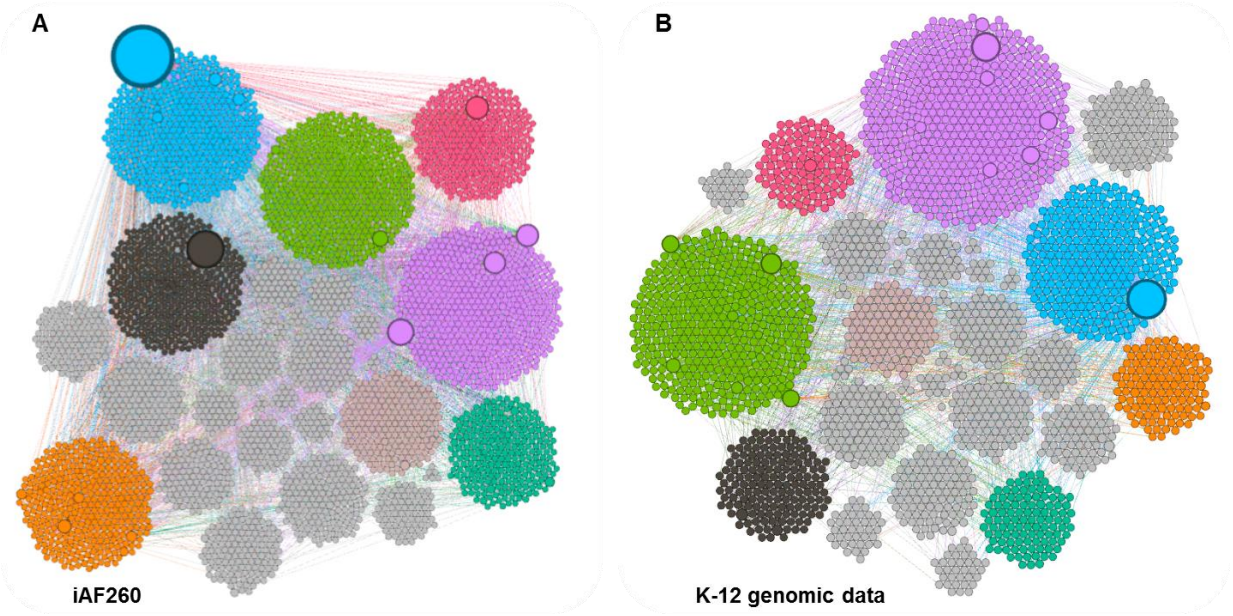


Figure 62. Modularity of reconstructed bipartite metabolic networks for *Escherichia coli* K-12. The networks were generated for both the **(A)** iAF1260 model and the **(B)** K-12 reconstructed model. Node colors represents the network modules (the modularity algorithm implemented in *Gephi* looks for the nodes that are more densely connected together than to the rest of the network - Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules), while node size represents the number of connections of each node. The most connected nodes are usually related to metabolites such as ATP, H^+, H2O and NADH.

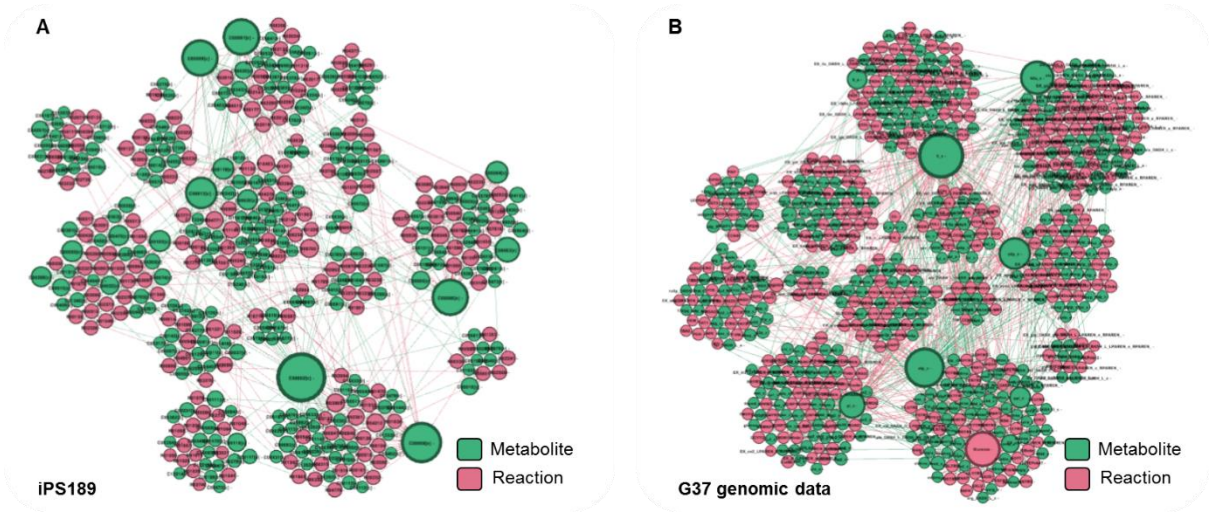


Figure 63. Reconstructed bipartite metabolic networks for *Mycoplasma genitalium*. The networks were generated for both the **(A)** iPS189 model and the **(B)** *M. genitalium* G37 reconstructed model. Node colours represent Metabolites or Reactions while node size represents the number of connections of each node. The most connected nodes are usually related to metabolites such as ATP, H^+, H2O and NADH.

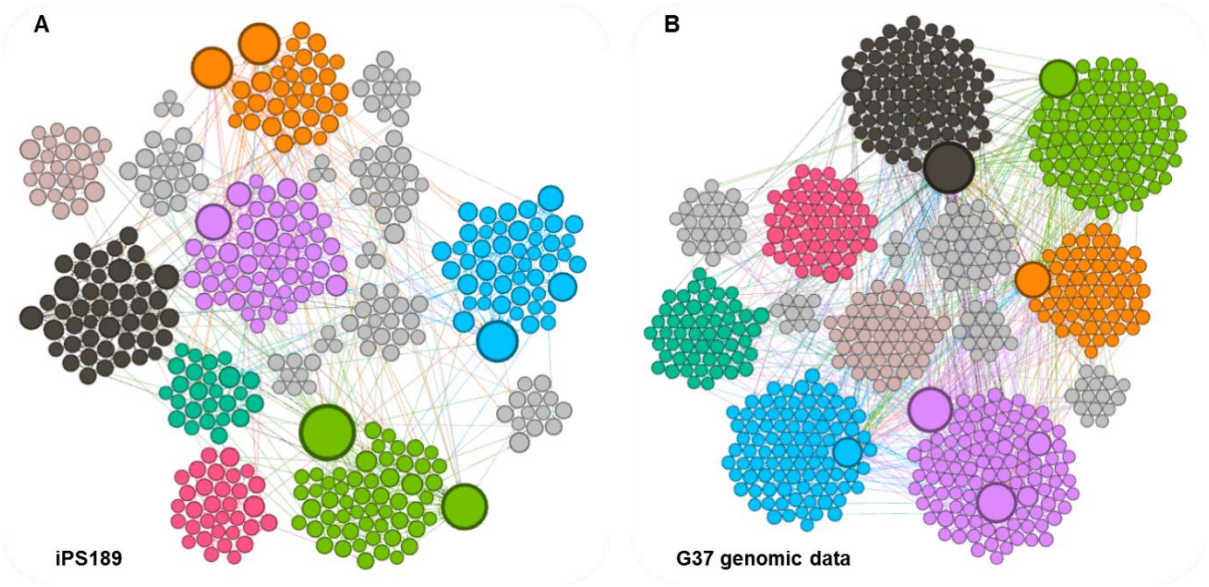


Figure 64. Modularity of reconstructed bipartite metabolic networks for *Mycoplasma genitalium*. The networks were generated for both the **(A)** iPS189 model and the (B) *M. genitalium* G37 reconstructed model. Node colors represents the network modules (the modularity algorithm implemented in *Gephi* looks for the nodes that are more densely connected together than to the rest of the network - Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules), while node size represents the number of connections of each node. The most connected nodes are usually related to metabolites such as ATP, H^+, H2O and NADH.

5. Discussion and Conclusions

As briefly discussed before, a key bottleneck in the pace of reconstruction of new high-quality metabolic models is our inability to directly make use of metabolite/reaction information from biological databases (Reed *et al.*, 2006) (e.g., BRENDA (Schomburg *et al.*, 2002; Placzek *et al.*, 2017), KEGG (Kanehisa, 2000; Kanehisa *et al.*, 2017), MetaCyc, EcoCyc, BioCyc (Caspi *et al.*, 2014), BKM-react (Lang, Stelzer and Schomburg, 2011), UM-BBD (Gao, Ellis and Wackett, 2009), Reactome.org, Rhea, PubChem, ChEBI etc.) or other models due to incompatibilities of representation, duplications and errors. In this context, a major impediment in the generation of genomic metabolic models is the presence of metabolites with multiple names across databases and models, and in some cases within the same resource, which significantly slows down the pooling of information from multiple sources. Therefore, the almost unavoidable inclusion of multiple replicates of the same metabolite can lead to missed opportunities to reveal lethal gene deletions, repair network gaps and quantify metabolic flows.

Moreover, most data sources inadvertently include some reactions that may be stoichiometrically inconsistent (Gevorgyan, Poolman and Fell, 2008) and/or elementally/charge unbalanced (Notebaart *et al.*, 2006; Ott and Vriend, 2006), which can adversely affect the prediction quality of the resulting models if used directly. Finally, a large number of metabolites in reactions are partly specified with respect to structural information and may contain generic side groups (e.g., alkyl groups -R), varying degree of a repeat unit participation in oligomers, or even just compound class identification such as “an amino acid” or “electron acceptor”. Over 3% of all metabolites and 8% of all reactions in the aforementioned databases and models exhibit one or more of these problems (Kumar, Suthers and Maranas, 2012).

There have already been a number of efforts aimed at addressing some of these limitations. The Rhea database, hosted by the European Bioinformatics Institute, aggregates reaction data primarily from IntEnz (Fleischmann, 2004) and ENZYME (Bairoch, 2000), whereas Reactome.org is a collection of reactions primarily focused on human metabolism (Vastrik *et al.*, 2007). Even though they crosslink their data to one or more popular databases such as KEGG, ChEBI, NCBI, KEGG, BRENDA, Ensembl, Uniprot, etc., both retain their own representation formats. More recently, the BKM-react database is a non-redundant biochemical reaction database containing known enzyme-catalyzed reactions compiled from BRENDA,

KEGG, and MetaCyc (Lang, Stelzer and Schomburg, 2011). An important step forward for models was the BiGG database (Schellenberger *et al.*, 2010), which includes seven genome-scale models from the Palsson group in a consistent nomenclature and exportable in SBML format (Hucka *et al.*, 2003). Another important recent development is the web resource Model SEED that can generate draft genome-scale metabolic models drawing from an internal database that integrates KEGG with 13 genome-scale models (including six of the models in the BiGG database) (Overbeek *et al.*, 2014). All of the reactions in Model SEED and BiGG are charge and elementally balanced. In this context, our work has adopted the KEGG database as it is one of the most complete databases in terms of interconnected information (Kanehisa *et al.*, 2017).

We have also observed in our work that metabolic networks for *E. coli* and *M. genitalium* presented similar network properties, raising a question regarding the potential discovery of universal traits in metabolic networks. Although we have focused only on two microorganisms and their reconstructed metabolic networks in this analysis, we have found that our data was in agreement with previous studies exploring this question in a larger scale. Oltvai *et al.*, 2000 (Oltvai *et al.*, 2000), have analysed 43 metabolic reconstructions from all three domains of Life and found that despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties such as network diameter - defined as the shortest biochemical pathway averaged over all pairs of substrates - and show striking similarities to the inherent organization of complex scale-free networks – networks which follow a power-law distribution (Newman, 2010). This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks, and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents (Oltvai *et al.*, 2000).

As discussed in Oltvai's study (Oltvai *et al.*, 2000), the apparent conservation of the network diameter in all living organisms may represent an additional survival and growth advantage, as a larger diameter would attenuate the organism's ability to respond efficiently to external changes or internal errors. It has also been found that that connectivity distribution of non-metabolic pathways in cellular systems may also follow a power-law distribution, indicating that cellular networks as a whole are scale-free networks (Oltvai *et al.*, 2000; Barabasi and

Oltvai, 2004; Mahadevan and Palsson, 2005). Therefore, the evolutionary selection of a robust and error-tolerant architecture may characterize all cellular networks, for which scale-free topology with a conserved network diameter appears to provide an optimal structural organization (Oltvai *et al.*, 2000; Barabasi and Oltvai, 2004; Mahadevan and Palsson, 2005).

In conclusion, this chapter has provided a novel computational pipeline for *de novo* generation of genome-scale metabolic models and for the exploration of intrinsic properties of stoichiometric metabolic models from (meta)genomic data. This pipeline has been extensively tested with both toy models and genomic data, providing an open-source alternative to the current metabolic reconstruction methods and highlighting the challenge of data standardization between different databases. The implementation of additional tools for improving the consistency of this method is still in progress, focusing on the development of a gap-filling algorithm that should greatly improve the applicability of this methodology in future studies. In a more ambitious step, we would like to combine this approach with other computational tools and theoretical frameworks for combining metabolic and transcriptional networks. In this manner, we shall move towards a more holistic approach for understanding Life and its properties.

IV. GENERAL CONCLUSIONS

The results of this dissertation have given rise to the following conclusions:

1. The promoter architecture is essential to its function and the relative distribution and abundance of TFBSs are directly related to the modulation of transcriptional outputs in terms of both transcriptional logic and emergent behaviours. The study of these new architectures of complex promoters also allows the generation of essential information for making the engineering process of regulatory elements more rational and predictable.

   a. in synthetic promoters with only one type of transcription factor binding sites - Fis vs. Neg or IHF vs. Neg -, the regulatory architectures leading to increased expression in their respective mutants are similar. This fact could be explained by the shared function and mechanism of action between these proteins - NAPS (Dillon and Dorman, 2010; Dorman, 2013) - that act normally as repressors of the transcription through the generation of conformational changes in the DNA;

   b. although the functional architectures for the libraries mentioned above are similar, it is possible to notice that there are particular behaviours in each of them that depend on the intrinsic qualities of each regulator involved;

   c. the patterns observed for the individual sequence libraries are not able to explain the expression patterns obtained by combining both the Fis and IHF sequences in complex synthetic promoters. This epistatic phenomenon, in which emergent and unpredictable behaviours arise from the combination of biological parts with known behaviours, appears to be widely distributed in molecular systems and with an important evolutionary role, especially in regulatory systems (Loewe, 2009; Lagator *et al.*, 2016; Aguilar-Rodríguez, Payne and Wagner, 2017; Monteiro, Arruda and Silva-Rocha, 2017).

2. Evolution on single TFBS sequences has the potential to navigate through a multidimensional space of binding affinities for different transcription factors. In complex promoters, this might occur in parallel among multiple TFBSs. The co-evolution of TFBSs complex promoters should shape the promoter architecture and,

ultimately, lead to specific regulatory logic responses as observed with synthetic promoters in the first chapter of this dissertation. The fixation of these mutations might depend on the function of the regulated gene under the specific select pressure imposed by the environment. A few other conclusions might be extrapolated from this model:

a. The rise of innovation is facilitated in the connected model while regulatory "dead-ends" are avoided.

b. There is a trade-off between "crosstalk" levels and regulatory robustness that should be further explored.

c. The diversity of natural TFBSs could be constrained by the multifactorial selective pressure for potentially promiscuous binding sites.

d. A promiscuous TFBS might not be bound by a candidate TF due to molecular constraints such as genomic context, promoter architecture and TFs concentration.

3. In order to achieve a better understanding of the most fundamental principles of transcriptional systems, we have to go beyond the features of model organism *E. coli*. In this context, we have found here that there is an enormous hidden diversity of *cis*-regulatory elements in environmental samples that can only be obtained through for the exploration of uncultured bacteria using metagenomic approaches. We have found a large number of constitutive promoters in soil samples with distinct expression dynamics and strengths, which shall contribute to the understanding of the metaregulomes and how different environments might shape them. Other general outcomes can be found below:

a. The synergy between Metagenomics and Synthetic Biology is essential for understanding Life and for the biotechnological development

b. Metaregulomes and dynamic expression profiles provide a new perspective on microbial communities

c. Only a small set of regulatory elements is currently accessible for prospection through Metagenomics, thus the usage of novel hosts and molecular tools will expand the universe of accessible genetic features.

4. Transcriptional regulatory layers, as any other kind of biological networks, are not isolated. They are deeply intertwined with other organizational layers in the generation of the epiphenomena we know as Life. In this context, the study of metabolic networks is essential for a more comprehensive understanding of transcriptional systems as they regulate themselves in a reciprocal manner. There is an urgent need for the development of novel tools for generating automated consistent metabolic networks in a standardised fashion as the Systems Biology (Hucka *et al.*, 2003) and Synthetic Biology markup languages (Galdzicki *et al.*, 2014). This might not only provide a more comprehensive view of biological systems, but also allow the prediction and (re)engineering cellular behaviours for specific functions.

Altogether, the current work has provided resourceful information regarding many aspects of transcriptional systems in bacteria which, provided the adequate theoretical framework, can be extrapolated to more complex systems such as eukaryotes. We believe this multiscale approach is fundamental for both understanding the general principles underpinning information processing in living systems and (re)engineering them for biotechnological applications.

V. REFERENCES

Agren, R. *et al.* (2012) 'Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT.', *PLoS computational biology*. Edited by C. D. Maranas, 8(5), p. e1002518. doi: 10.1371/journal.pcbi.1002518.

Aguilar-Rodríguez, J., Payne, J. L. and Wagner, A. (2017) 'A thousand empirical adaptive landscapes and their navigability', *Nature Ecology & Evolution*, 1(2), p. 0045. doi: 10.1038/s41559-016-0045.

Akira, I. (2000) 'Functional Modulation of Escherichia Coli RNA polymerase', *Rev. Microbiol*, pp. 499–518. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev.micro.54.1.499 (Accessed: 12 June 2018).

Albert, R. (2005) 'Scale-free networks in cell biology', *J Cell Sci*, 118(Pt 21), pp. 4947–4957. doi: 10.1242/jcs.02714.

Aldana, M. *et al.* (2007) 'Robustness and evolvability in genetic regulatory networks', *Journal of Theoretical Biology*, 245(3), pp. 433–448. doi: 10.1016/j.jtbi.2006.10.027.

Alon, U. (2007) 'An Introduction to Systems Biology: Design Principles of Biological Circuits', *Chapman HallCRC mathematical and computational biology series*. doi: citeulike-article-id:1314150.

Alves, L. de F., Silva-Rocha, R. and Guazzaroni, M.-E. (2017) 'Enhancing Metagenomic Approaches Through Synthetic Biology', in *Functional Metagenomics: Tools and Applications*. Cham: Springer International Publishing, pp. 75–94. doi: 10.1007/978-3-319-61510-3_5.

Amann, R. I., Ludwig, W. and Schleifer, K. H. (1995) 'Phylogenetic identification and in situ detection of individual microbial cells without cultivation.', *Microbiological reviews*, 59(1), pp. 143–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7535888.

Amores, G. R., Guazzaroni, M. E. and Silva-Rocha, R. (2015) 'Engineering Synthetic cis-Regulatory Elements for Simultaneous Recognition of Three Transcriptional Factors in Bacteria', *ACS Synthetic Biology*, 4(12), pp. 1287–1294. doi: 10.1021/acssynbio.5b00098.

Andersen, J. B. *et al.* (1998) 'New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria', *Applied and Environmental Microbiology*, 64(6), pp. 2240–2246. Available at: http://www.ncbi.nlm.nih.gov/pubmed/106306.

Andrianantoandro, E. *et al.* (2006) 'Synthetic biology: new engineering rules for an emerging discipline.', *Molecular systems biology*, 2, p. 2006.0028. doi: 10.1038/msb4100073.

Ansong, C. *et al.* (2012) 'Studying Salmonellae and Yersiniae Host–Pathogen Interactions Using Integrated 'Omics and Modeling', in *Current topics in microbiology and immunology*, pp. 21–41. doi: 10.1007/82_2012_247.

Arkin, A. and Ross, J. (1994) 'Computational functions in biochemical reaction networks', *Biophysical Journal*. doi: 10.1016/S0006-3495(94)80516-8.

Azam, T. A. *et al.* (1999) 'Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid', *Journal of Bacteriology*. doi: .

Baba, T. *et al.* (2006) 'Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection', *Molecular Systems Biology*. European Molecular Biology

Organization, 2, p. 2006.0008. doi: 10.1038/msb4100050.

Babu, M. M., Balaji, S. and Aravind, L. (2007) 'General Trends in the Evolution of Prokaryotic Transcriptional Regulatory Networks', in *Gene and Protein Evolution*. Basel: KARGER, pp. 66–80. doi: 10.1159/000107604.

Bairoch, A. (2000) 'The ENZYME database in 2000', *Nucleic Acids Research*, 28(1), pp. 304–305. doi: 10.1093/nar/28.1.304.

Ball, C. A. *et al.* (1992) 'Dramatic changes in Fis levels upon nutrient upshift in Escherichia coli', *Journal of Bacteriology*. doi: 10.1128/jb.174.24.8043-8056.1992.

Barabasi, A. L. and Oltvai, Z. N. (2004) 'Newtowk Biology: Understanding the Cells Functional Organization', *Nature Reviews, Genetics*, 5, pp. 101–113. Available at: https://www.nature.com/articles/nrg1272 (Accessed: 12 June 2018).

Barnard, A., Wolfe, A. and Busby, S. (2004) 'Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes', *Current Opinion in Microbiology*, 7(2), pp. 102–108. doi: 10.1016/j.mib.2004.02.011.

Bashor, C. J. and Collins, J. J. (2018) 'Understanding Biological Regulation Through Synthetic Biology.', *Annual review of biophysics*, 47(March), pp. 399–423. doi: 10.1146/annurev-biophys-070816-033903.

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi: An Open Source Software for Exploring and Manipulating Networks', *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362. doi: 10.1136/qshc.2004.010033.

Bayer, T. S. (2010) 'Using synthetic biology to understand the evolution of gene expression', *Current Biology*. Elsevier Ltd, 20(17), pp. R772–R779. doi: 10.1016/j.cub.2010.06.049.

Beal, J. *et al.* (2016) 'Reproducibility of fluorescent expression from engineered biological constructs in E. coli', *PLoS ONE*, 11(3). doi: 10.1371/journal.pone.0150182.

Beal, J., Haddock-Angelli, T., Baldwin, G., *et al.* (2018) 'Quantification of bacterial fluorescence using independent calibrants', *PLOS ONE*. Edited by D. Olson. Public Library of Science, 13(6), p. e0199432. doi: 10.1371/journal.pone.0199432.

Beal, J., Haddock-Angelli, T., Farny, N., *et al.* (2018) 'Time to Get Serious about Measurement in Synthetic Biology', *Trends in Biotechnology*. doi: 10.1016/j.tibtech.2018.05.003.

Belliveau, N. M. *et al.* (2018) 'A Systematic and Scalable Approach for Dissecting the Molecular Mechanisms of Transcriptional Regulation in Bacteria', *Biophysical Journal*. Elsevier, 114(3), p. 151a. doi: 10.1016/j.bpj.2017.11.849.

Beloqui, A. *et al.* (2008) 'Recent trends in industrial microbiology.', *Current opinion in microbiology*, 11(3), pp. 240–8. doi: 10.1016/j.mib.2008.04.005.

Bendtsen, K. M. *et al.* (2011) 'Direct and indirect effects in the regulation of overlapping promoters.', *Nucleic acids research*, 39(16), pp. 6879–6885. doi: 10.1093/nar/gkr390.

Berg, J., Willmann, S. and Lässig, M. (2004) 'Adaptive evolution of transcription factor binding sites', *BMC Evolutionary Biology*. BioMed Central, 4(6769), pp. 564–567. doi: 10.1186/1471-2148-4-42.

Bernstein, J. R. *et al.* (2007) 'Directed Evolution of Ribosomal Protein S1 for Enhanced

Translational Efficiency of High GC Rhodopseudomonas palustris DNA in Escherichia coli', *Journal of Biological Chemistry*, 282(26), pp. 18929–18936. doi: 10.1074/jbc.M701395200.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., *et al.* (2005) 'Transcriptional regulation by the numbers: Applications', *Current Opinion in Genetics and Development*, pp. 125–135. doi: 10.1016/j.gde.2005.02.006.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J. and Phillips, R. (2005) 'Transcriptional regulation by the numbers: Models', *Current Opinion in Genetics and Development*, pp. 116–124. doi: 10.1016/j.gde.2005.02.007.

Blount, B. A. *et al.* (2012) 'Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology', *PLoS ONE*. Edited by M. F. Tuite, 7(3), p. e33279. doi: 10.1371/journal.pone.0033279.

Boer, C. de *et al.* (2018) 'Deciphering cis-regulatory logic with 100 million random promoters', *bioRxiv*. Cold Spring Harbor Laboratory, p. 224907. doi: 10.1101/224907.

Bohlin, J. *et al.* (2010) 'Analysis of intra-genomic GC content homogeneity within prokaryotes', *BMC Genomics*, 11(1), p. 464. doi: 10.1186/1471-2164-11-464.

Bordbar, A. *et al.* (2014) 'Constraint-based models predict metabolic and associated cellular functions', *Nature Reviews Genetics*, 15(2), pp. 107–120. doi: 10.1038/nrg3643.

Boyle, P. M. and Silver, P. A. (2009) 'Harnessing nature's toolbox: regulatory elements for synthetic biology', *Journal of The Royal Society Interface*, 6(Suppl_4), pp. S535–S546. doi: 10.1098/rsif.2008.0521.focus.

Brenner, K., You, L. and Arnold, F. H. (2008) 'Engineering microbial consortia: a new frontier in synthetic biology', *Trends in Biotechnology*, 26(9), pp. 483–489. doi: 10.1016/j.tibtech.2008.05.004.

Browning, D. D. F. and Busby, S. J. W. S. (2004) 'The regulation of bacterial transcription initiation', *Nature Reviews Microbiology*, 2(1), pp. 57–65. doi: 10.1038/nrmicro787.

Browning, D. F. and Busby, S. J. W. (2016) 'Local and global regulation of transcription initiation in bacteria', *Nature Reviews Microbiology*. Nature Publishing Group, 14(10), pp. 638–650. doi: 10.1038/nrmicro.2016.103.

Browning, D. F., Cole, J. A. and Busby, S. J. W. (2000) 'Suppression of FNR-dependent transcription activation at the Escherichia coli nir promoter by Fis, IHF and H-NS: Modulation of transcription by a complex nucleo-protein assembly', *Molecular Microbiology*. Wiley/Blackwell (10.1111), 37(5), pp. 1258–1269. doi: 10.1046/j.1365-2958.2000.02087.x.

Browning, D. F., Grainger, D. C. and Busby, S. J. W. (2010) 'Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression', *Current Opinion in Microbiology*, pp. 773–780. doi: 10.1016/j.mib.2010.09.013.

Buchler, N. E., Gerland, U. and Hwa, T. (2003) 'On schemes of combinatorial transcription logic', *Proceedings of the National Academy of Sciences*, 100(9), pp. 5136–5141. doi: 10.1073/pnas.0930314100.

Cameron, D. E., Bashor, C. J. and Collins, J. J. (2014) 'A brief history of synthetic biology.', *Nature reviews. Microbiology*. Nature Publishing Group, 12(5), pp. 381–90. doi: 10.1038/nrmicro3239.

Carroll, S. B. (2008) 'Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution', *Cell*, 134(1), pp. 25–36. doi: 10.1016/j.cell.2008.06.030.

Casadaban, M. J. and Cohen, S. N. (1980) 'Analysis of gene control signals by DNA fusion and cloning in Escherichia coli', *Journal of Molecular Biology*, 138(2), pp. 179–207. doi: 10.1016/0022-2836(80)90283-1.

Cases, I. and de Lorenzo, V. (2005) 'Promoters in the environment: Transcriptional regulation in its natural context', *Nature Reviews Microbiology*, pp. 105–118. doi: 10.1038/nrmicro1084.

Cases, I., De Lorenzo, V. and Ouzounis, C. A. (2003) 'Transcription regulation and environmental adaptation in bacteria', *Trends in Microbiology*, pp. 248–253. doi: 10.1016/S0966-842X(03)00103-3.

Caspi, R. *et al.* (2014) 'The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases', *Nucleic Acids Research*, 38(Database issue), pp. 459–471. doi: 10.1093/nar/gkp875.

Chaires, J. B. (2008) 'Allostery: DNA Does It, Too', *ACS Chemical Biology*, 3(4), pp. 207–209. doi: 10.1021/cb800070s.

Chen, S. *et al.* (2007) 'Characterization of Strong Promoters from an Environmental Flavobacterium hibernum Strain by Using a Green Fluorescent Protein-Based Reporter System', *Applied and Environmental Microbiology*, 73(4), pp. 1089–1100. doi: 10.1128/AEM.01577-06.

Chiu, H.-C., Levy, R. and Borenstein, E. (2014) 'Emergent Biosynthetic Capacity in Simple Microbial Communities', *PLoS Computational Biology*. Edited by C. A. Ouzounis. Public Library of Science, 10(7), p. e1003695. doi: 10.1371/journal.pcbi.1003695.

Cho, B.-K. *et al.* (2008) 'Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts', *Genome Research*, 18(6), pp. 900–910. doi: 10.1101/gr.070276.107.

Cho, B.-K. *et al.* (2009) 'Elucidation of the transcription unit architecture of the Escherichia coli K-12 MG1655 genome', *Nature Biotechnology*. Nature Publishing Group, 27(11), pp. 1043–1049. doi: 10.1038/nbt.1582.

Conway, T. *et al.* (2014) 'Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing', *mBio*, 5(4), pp. 1–12. doi: 10.1128/mBio.01442-14.

Costanzo, M. *et al.* (2010) 'The genetic landscape of a cell', *Science*, 327(5964), pp. 425–431. doi: 10.1126/science.1180823.

Covert, M. W. *et al.* (2004) 'Integrating high-throughput and computational data elucidates bacterial networks', *Nature*, 429(6987), pp. 92–96. doi: 10.1038/nature02456.

Cox, R. S., Surette, M. G. and Elowitz, M. B. (2007) 'Programming gene expression with combinatorial promoters', *Molecular Systems Biology*. European Molecular Biology Organization, 3, p. 145. doi: 10.1038/msb4100187.

Crocker, J. and Ilsley, G. R. (2017) 'Using synthetic biology to study gene regulatory evolution', *Current Opinion in Genetics and Development*. Elsevier Ltd, 47, pp. 91–101. doi: 10.1016/j.gde.2017.09.001.

Crooks, G. E. *et al.* (2004) 'WebLogo: A Sequence Logo Generator', *Genome Research*, 14(6), pp. 1188–1190. doi: 10.1101/gr.849004.

Csete, M. and Doyle, J. (2002) 'Reverse Engineering of Biological Complexity', *Science*, 295(5560), pp. 1664–1669. doi: doi: 10.1126/science.1069981.

Cuevas, D. A. *et al.* (2016) 'From DNA to FBA: How to build your own genome-scale metabolic model', *Frontiers in Microbiology*, 7(JUN), pp. 1–12. doi: 10.3389/fmicb.2016.00907.

Danchin, A. (2012) 'Scaling up synthetic biology: Do not forget the chassis', *FEBS Letters*. Federation of European Biochemical Societies, 586(15), pp. 2129–2137. doi: 10.1016/j.febslet.2011.12.024.

Dari, A. *et al.* (2011) 'Logical stochastic resonance with correlated internal and external noises in a synthetic biological logic block', *Chaos*. doi: 10.1063/1.3660159.

Datsenko, K. A. and Wanner, B. L. (2000) 'One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products', *Proceedings of the National Academy of Sciences*, 97(12), pp. 6640–6645. doi: 10.1073/pnas.120163297.

Dean, A. M. and Thornton, J. W. (2007) 'Mechanistic approaches to the study of evolution: The functional synthesis', *Nature Reviews Genetics*, pp. 675–688. doi: 10.1038/nrg2160.

Decoene, T. *et al.* (2017) 'Standardization in synthetic biology: an engineering discipline coming of age.', *Critical reviews in biotechnology*, pp. 1–10. doi: 10.1080/07388551.2017.1380600.

Delmont, T. O. *et al.* (2011) 'Metagenomic mining for microbiologists', *The ISME Journal*, 5(12), pp. 1837–1843. doi: 10.1038/ismej.2011.61.

Dillon, S. C. and Dorman, C. J. (2010) 'Bacterial nucleoid-associated proteins, nucleoid structure and gene expression.', *Nature reviews. Microbiology*. Nature Publishing Group, 8(3), pp. 185–95. doi: 10.1038/nrmicro2261.

Dobzhansky, T. (1973) 'Nothing in Biology Makes Sense except in the Light of Evolution', *The American Biology Teacher*. University of California Press , 35(3), pp. 125–129. doi: 10.2307/4444260.

Dolinski, K. and Troyanskaya, O. G. (2015) 'Implications of Big Data for cell biology', *Molecular Biology of the Cell*. Edited by K. G. Kozminski, 26(14), pp. 2575–2578. doi: 10.1091/mbc.E13-12-0756.

Dorman, C. J. (2013) 'Genome architecture and global gene regulation in bacteria: making progress towards a unified model?', *Nature reviews. Microbiology*, 11(5), pp. 349–55. doi: 10.1038/nrmicro3007.

Duarte, N. C. (2004) 'Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model', *Genome Research*, 14(7), pp. 1298–1309. doi: 10.1101/gr.2250904.

Dunn, A. K. and Handelsman, J. (1999) 'A vector for promoter trapping in Bacillus cereus', *Gene*, 226(2), pp. 297–305. doi: 10.1016/S0378-1119(98)00544-7.

Dupont, C. L. *et al.* (2015) 'Genomes and gene expression across light and productivity

gradients in eastern subtropical Pacific microbial communities.', *The ISME journal*. Nature Publishing Group, 9(5), pp. 1076–92. doi: 10.1038/ismej.2014.198.

Durot, M., Bourguignon, P.-Y. and Schachter, V. (2009) 'Genome-scale models of bacterial metabolism: reconstruction and applications', *FEMS Microbiology Reviews*, 33(1), pp. 164–190. doi: 10.1111/j.1574-6976.2008.00146.x.

Ebrahim, A. *et al.* (2013) 'COBRApy: COnstraints-Based Reconstruction and Analysis for Python', *BMC Systems Biology*. BioMed Central, 7(1), p. 74. doi: 10.1186/1752-0509-7-74.

Edwards, J. S. and Palsson, B. O. (1999) 'Systems properties of the Haemophilus influenzae Rd metabolic genotype.', *The Journal of biological chemistry*, 274(25), pp. 17410–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10364169 (Accessed: 29 May 2018).

Ekkers, D. M. *et al.* (2012) 'The great screen anomaly—a new frontier in product discovery through functional metagenomics', *Applied Microbiology and Biotechnology*, 93(3), pp. 1005–1020. doi: 10.1007/s00253-011-3804-3.

Elowitz, M. B. *et al.* (2007) 'Stochastic Gene Expression in a Single Cell', *Science*, 297(2002), pp. 1183–6. doi: 10.1126/science.1070919.

Farnsworth, K. D., Nelson, J. and Gershenson, C. (2013) 'Living is Information Processing: From Molecules to Global Systems', *Acta Biotheoretica*, 61(2), pp. 203–222. doi: 10.1007/s10441-013-9179-3.

Feist, A. M. *et al.* (2007) 'A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information', *Molecular Systems Biology*, 3, p. 121. doi: 10.1038/msb4100155.

Feist, A. M. and Palsson, B. Ø. (2008) 'The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.', *Nature biotechnology*. NIH Public Access, 26(6), pp. 659–67. doi: 10.1038/nbt1401.

Fernandez, L. *et al.* (2014) 'Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities.', *BMC genomics*, 15, p. 877. doi: 10.1186/1471-2164-15-877.

Fierer, N. *et al.* (2012) 'Cross-biome metagenomic analyses of soil microbial communities and their functional attributes', *Proceedings of the National Academy of Sciences*, 109(52), pp. 21390–21395. doi: 10.1073/pnas.1215210110.

Fierer, N., Bradford, M. A. and Jackson, R. B. (2007) 'Toward an ecological classification of soil bacteria', *Ecology*, 88(6), pp. 1354–1364. doi: 10.1890/05-1839.

Fleischmann, A. (2004) 'IntEnz, the integrated relational enzyme database', *Nucleic Acids Research*, 32(90001), p. 434D–437. doi: 10.1093/nar/gkh119.

Foerstner, K. U. *et al.* (2005) 'Environments shape the nucleotide composition of genomes.', *EMBO reports*, 6(12), pp. 1208–13. doi: 10.1038/sj.embor.7400538.

Fondi, M. and Liò, P. (2015) 'Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology', *Microbiological Research*. Urban & Fischer, 171, pp. 52–64. doi: 10.1016/J.MICRES.2015.01.003.

Fortunato, C. S. and Crump, B. C. (2015) 'Microbial gene abundance and expression patterns

across a river to ocean salinity gradient', *PLoS ONE*, 10(11), pp. 1–22. doi: 10.1371/journal.pone.0140578.

Fredrickson, J. K. (2015) 'Ecological communities by design', *Science*, 348(6242), pp. 1425–1427. doi: 10.1126/science.aab0946.

Frias-Lopez, J. *et al.* (2008) 'Microbial community gene expression in ocean surface waters', *Proceedings of the National Academy of Sciences*, 105(10), pp. 3805–3810. doi: 10.1073/pnas.0708897105.

Friedland, A. E. *et al.* (2009) 'Synthetic gene networks that count.', *Science (New York, N.Y.)*. doi: 10.1126/science.1172005.

Friedlander, T. *et al.* (2016) 'Intrinsic limits to gene regulation by global crosstalk', *Nature Communications*, 7. doi: 10.1038/ncomms12307.

Friedlander, T. *et al.* (2017) 'Evolution of new regulatory functions on biophysically realistic fitness landscapes', *Nature Communications*, 8(1). doi: 10.1038/s41467-017-00238-8.

Gabor, E. M., Alkema, W. B. L. and Janssen, D. B. (2004) 'Quantifying the accessibility of the metagenome by random expression cloning techniques', *Environmental Microbiology*, 6(9), pp. 879–886. doi: 10.1111/j.1462-2920.2004.00640.x.

Gabor, E. M., de Vries, E. J. and Janssen, D. B. (2004) 'Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases', *Environmental Microbiology*, 6(9), pp. 948–958. doi: 10.1111/j.1462-2920.2004.00643.x.

Galdzicki, M. *et al.* (2014) 'The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology.', *Nature biotechnology*, 32(6), pp. 545–50. doi: 10.1038/nbt.2891.

Gama-Castro, S. *et al.* (2016) 'RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond', *Nucleic Acids Research*, 44(D1), pp. D133–D143. doi: 10.1093/nar/gkv1156.

Gao, J., Ellis, L. B. M. and Wackett, L. P. (2009) 'The University of Minnesota Biocatalysis/Biodegradation Database: Improving public access', *Nucleic Acids Research*, 38(SUPPL.1). doi: 10.1093/nar/gkp771.

García-Ochoa, F. *et al.* (2000) 'Xanthan gum: production, recovery, and properties.', *Biotechnology advances*, 18(7), pp. 549–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14538095 (Accessed: 29 May 2018).

Gehlenborg, N. *et al.* (2010) 'Visualization of omics data for systems biology', *Nature Methods*, pp. S56–S68. doi: 10.1038/nmeth.1436.

Gerstel, U., Park, C. and Römling, U. (2003) 'Complex regulation of csgD promoter activity by global regulatory proteins', *Molecular Microbiology*. doi: 10.1046/j.1365-2958.2003.03594.x.

Gertz, J., Siggia, E. D. and Cohen, B. A. (2009) 'Analysis of combinatorial cis-regulation in synthetic and genomic promoters.', *Nature*, 457(7226), pp. 215–8. doi: 10.1038/nature07521.

Gevorgyan, A., Poolman, M. G. and Fell, D. A. (2008) 'Detection of stoichiometric

inconsistencies in biomolecular models', *Bioinformatics*, 24(19), pp. 2245–2251. doi: 10.1093/bioinformatics/btn425.

Goldberg, A. P. *et al.* (2018) 'Emerging whole-cell modeling principles and methods', *Current Opinion in Biotechnology*, 51, pp. 97–102. doi: 10.1016/j.copbio.2017.12.013.

Gomez-Cabrero, D. *et al.* (2014) 'Data integration in the era of omics: current and future challenges', *BMC systems biology*, p. I1. doi: 10.1186/1752-0509-8-S2-I1.

Gordân, R. *et al.* (2013) 'Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape', *Cell Reports*, 3(4), pp. 1093–1104. doi: 10.1016/j.celrep.2013.03.014.

Grant, S. G. *et al.* (1990) 'Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants.', *Proceedings of the National Academy of Sciences*, 87(12), pp. 4645–4649. doi: 10.1073/pnas.87.12.4645.

Greber, D. and Fussenegger, M. (2007) 'Mammalian synthetic biology: Engineering of sophisticated gene networks', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2007.05.014.

Gruber, T. M. and Gross, C. A. (2003) 'Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space', *Annual Review of Microbiology*, 57(1), pp. 441–466. doi: 10.1146/annurev.micro.57.030502.090913.

Guazzaroni, M.-E. E. and Silva-Rocha, R. (2014) 'Expanding the Logic of Bacterial Promoters Using Engineered Overlapping Operators for Global Regulators', *ACS Synthetic Biology*, 3(9), pp. 666–675. doi: 10.1021/sb500084f.

Guazzaroni, M.-E., Silva-Rocha, R. and Ward, R. J. (2015) 'Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening', *Microbial Biotechnology*, 8(1), pp. 52–64. doi: 10.1111/1751-7915.12146.

Guazzaroni, M. E. *et al.* (2013) 'Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment', *Environmental Microbiology*, 15(4), pp. 1088–1102. doi: 10.1111/1462-2920.12021.

Han, S. S. *et al.* (2008) 'Screening of promoters from metagenomic DNA and their use for the construction of expression vectors', *Journal of Microbiology and Biotechnology*, 18(10), pp. 1634–1640. doi: 8104 [pii].

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.', *Chemistry & biology*, 5(10), pp. R245–R249. doi: 10.1016/S1074-5521(98)90108-9.

Handelsman, J. (2005) 'Metagenomics: Application of Genomics to Uncultured Microorganisms', *Microbiology and Molecular Biology Reviews*, 69(1), pp. 195–195. doi: 10.1128/MMBR.69.1.195.2005.

Harcombe, W. R. *et al.* (2014) 'Metabolic Resource Allocation in Individual Microbes Determines Ecosystem Interactions and Spatial Dynamics', *Cell Reports*, 7(4), pp. 1104–1115. doi: 10.1016/j.celrep.2014.03.070.

Hart, Y. *et al.* (2012) 'Design principles of cell circuits with paradoxical components', *Proceedings of the National Academy of Sciences*, 109(21), pp. 8346–8351. doi: 10.1073/pnas.1117475109.

Hebisch, E. *et al.* (2013) 'High Variation of Fluorescence Protein Maturation Times in Closely Related Escherichia coli Strains', *PLoS ONE*. Edited by A. Hofmann, 8(10), p. e75991. doi: 10.1371/journal.pone.0075991.

Heirendt, L. *et al.* (2017) 'Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0'. Available at: http://arxiv.org/abs/1710.04038 (Accessed: 29 May 2018).

Henry, C. S. *et al.* (2009) 'iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.', *Genome biology*, 10(6), p. R69. doi: 10.1186/gb-2009-10-6-r69.

Henry, C. S. *et al.* (2010) 'High-throughput generation, optimization and analysis of genome-scale metabolic models', *Nature Biotechnology*, 28(9), pp. 977–982. doi: 10.1038/nbt.1672.

Hermsen, R., Tans, S. and Ten Wolde, P. R. (2006) 'Transcriptional regulation by competing transcription factor modules', *PLoS Computational Biology*. Sinauer, 2(12), pp. 1552–1560. doi: 10.1371/journal.pcbi.0020164.

van Hijum, S. A. F. T., Medema, M. H. and Kuipers, O. P. (2009) 'Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation', *Microbiology and Molecular Biology Reviews*, 73(3), pp. 481–509. doi: 10.1128/MMBR.00037-08.

Hucka, M. *et al.* (2003) 'The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models', *Bioinformatics*, 19(4), pp. 524–531. doi: 10.1093/bioinformatics/btg015.

Huminiecki, Ł. and Horbańczuk, J. (2017) 'Can We Predict Gene Expression by Understanding Proximal Promoter Architecture?', *Trends in Biotechnology*, 35(6), pp. 530–546. doi: 10.1016/j.tibtech.2017.03.007.

Hunziker, A. *et al.* (2010) 'Genetic flexibility of regulatory networks', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0915003107.

Husnik, F. *et al.* (2013) 'Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis', *Cell*, 153(7), pp. 1567–1578. doi: 10.1016/j.cell.2013.05.040.

Ibáñez-Marcelo, E. and Alarcón, T. (2014) 'The topology of robustness and evolvability in evolutionary systems with genotype-phenotype map', *Journal of Theoretical Biology*, 356, pp. 144–162. doi: 10.1016/j.jtbi.2014.04.014.

III, R. C. (2008) *Transcriptional regulation and combinatorial genetic logic in synthetic bacterial circuits*. Available at: http://search.proquest.com/openview/40cd3c113e9c6790e009f914a87139cd/1?pq-origsite=gscholar&cbl=18750&diss=y (Accessed: 12 June 2018).

International Union of Biochemistry and Molecular Biology (1992) 'Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes', *Academic, San Diego, CA*, p. 862.

Ishihama, A. (1999) 'Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival', *Genes to Cells*. doi: 10.1046/j.1365-2443.1999.00247.x.

Ishihama, A. (2009) 'The Nucleoid: an Overview', *EcoSal Plus*, 3(2). doi:

10.1128/ecosalplus.2.6.

Ishihama, A. (2010) ‘Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks’, *FEMS Microbiology Reviews*, 34(5), pp. 628–645. doi: 10.1111/j.1574-6976.2010.00227.x.

Ishihama, A. (2017) ‘Building a complete image of genome regulation in the model organism *Escherichia coli*’, *The Journal of General and Applied Microbiology*, 63, pp. 311–324. doi: 10.2323/jgam.2017.01.002.

Ishihama, A., Shimada, T. and Yamazaki, Y. (2016) ‘Transcription profile of Escherichia coli: Genomic SELEX search for regulatory targets of transcription factors’, *Nucleic Acids Research*, 44(5), pp. 2058–2074. doi: 10.1093/nar/gkw051.

Jacob, F. F. and Monod, J. (1961) ‘Genetic regulatory mechanisms in the synthesis of proteins’, *Journal of Molecular Biology*. Elsevier, 3(3), pp. 318–356. doi: 10.1016/B978-0-12-460482-7.50042-7.

Janssen, P. H. (2006) ‘Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes MINIREVIEWS Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes’, *Applied and Environmental Microbiology*, 72(3), pp. 1719–1728. doi: 10.1128/AEM.72.3.1719.

Jiménez, D. J. *et al.* (2012) ‘A novel cold active esterase derived from Colombian high Andean forest soil metagenome’, *World Journal of Microbiology and Biotechnology*, 28(1), pp. 361–370. doi: 10.1007/s11274-011-0828-x.

Johns, N. I. *et al.* (2016) ‘Principles for designing synthetic microbial communities’, *Current Opinion in Microbiology*, 31, pp. 146–153. doi: 10.1016/j.mib.2016.03.010.

Johns, N. I. *et al.* (2018) ‘Metagenomic mining of regulatory elements enables programmable species-selective gene expression’, *Nature Methods*. Nature Publishing Group, 15(5), pp. 323–329. doi: 10.1038/nmeth.4633.

de Jong, A. *et al.* (2012) ‘PePPER: a webserver for prediction of prokaryote promoter elements and regulons.’, *BMC genomics*, 13(1), p. 299. doi: 10.1186/1471-2164-13-299.

Kanehisa, M. (2000) ‘KEGG: Kyoto Encyclopedia of Genes and Genomes’, *Nucleic Acids Research*, 28(1), pp. 27–30. doi: 10.1093/nar/28.1.27.

Kanehisa, M. *et al.* (2017) ‘KEGG: new perspectives on genomes, pathways, diseases and drugs’, *Nucleic Acids Research*, 45(D1), pp. D353–D361. doi: 10.1093/nar/gkw1092.

Karr, J. R., Takahashi, K. and Funahashi, A. (2015) ‘The principles of whole-cell modeling’, *Current Opinion in Microbiology*. Elsevier Ltd, 27, pp. 18–24. doi: 10.1016/j.mib.2015.06.004.

Kauffman, S. A. (1994) ‘The origins of order; self organization and selection in evolution’, *International Journal of Biochemistry*. Oxford University Press, 26(6), p. 855. doi: 10.1016/0020-711X(94)90119-8.

Keseler, I. M. *et al.* (2017) ‘The EcoCyc database: reflecting new knowledge about Escherichia coli K-12’, *Nucleic Acids Research*, 45(D1), pp. D543–D550. doi: 10.1093/nar/gkw1003.

Khoueiry, P. *et al.* (2010) ‘A cis-Regulatory Signature in Ascidians and Flies, Independent of Transcription Factor Binding Sites’, *Current Biology*, 20(9), pp. 792–802. doi:

10.1016/j.cub.2010.03.063.

Kim, S. *et al.* (2013) 'Probing allostery through DNA', *Science*, 339(6121), pp. 816–819. doi: 10.1126/science.1229223.

Kimelman, A. *et al.* (2012) 'A vast collection of microbial genes that are toxic to bacteria', *Genome Research*, 22(4), pp. 802–809. doi: 10.1101/gr.133850.111.

King, M.-C. and Wilson, A. C. (1975) 'Evolution at two levels in humans and chimpanzees', *Science*, 188(4184), pp. 107–116. doi: 10.1126/science.1090005.

Kinkhabwala, A. and Guet, C. C. (2008) 'Uncovering cis regulatory codes using synthetic promoter shuffling.', *PloS one*, 3(4), p. e2030. doi: 10.1371/journal.pone.0002030.

Kitano, H. (2002) 'Systems biology: a brief overview.', *Science (New York, N.Y.)*, 295(5560), pp. 1662–4. doi: 10.1126/science.1069492.

Kitano, H. (2004) 'Biological robustness', *Nature Reviews Genetics*, pp. 826–837. doi: 10.1038/nrg1471.

Klitgord, N. and Segrè, D. (2010) 'Environments that induce synthetic microbial ecosystems.', *PLoS computational biology*. Edited by J. A. Papin, 6(11), p. e1001002. doi: 10.1371/journal.pcbi.1001002.

Koonin, E. V. (2009) 'Evolution of genome architecture', *International Journal of Biochemistry and Cell Biology*, 41(2), pp. 298–306. doi: 10.1016/j.biocel.2008.09.015.

Koonin, E. V. and Wolf, Y. I. (2008) 'Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world', *Nucleic Acids Research*, 36(21), pp. 6688–6719. doi: 10.1093/nar/gkn668.

Kubota, M., Yamazaki, Y. and Ishihama, A. (1991) 'Random screening of promoters from Escherichia coli and classification based on the promoter strength.', *The Japanese Journal of Genetics*, 66(4), pp. 399–409. doi: 10.1266/jjg.66.399.

Kumar, A., Suthers, P. F. and Maranas, C. D. (2012) 'MetRxn: A knowledgebase of metabolites and reactions spanning metabolic models and databases', *BMC Bioinformatics*, 13(1). doi: 10.1186/1471-2105-13-6.

Lagator, M. *et al.* (2016) 'Epistatic Interactions in the Arabinose Cis-Regulatory Element.', *Molecular biology and evolution*. Oxford University Press, 33(3), pp. 761–9. doi: 10.1093/molbev/msv269.

Lam, K. N. *et al.* (2015) 'Current and future resources for functional metagenomics.', *Frontiers in microbiology*, 6, p. 1196. doi: 10.3389/fmicb.2015.01196.

Land, M. *et al.* (2015) 'Insights from 20 years of bacterial genome sequencing.', *Functional & integrative genomics*, 15(2), pp. 141–61. doi: 10.1007/s10142-015-0433-4.

Lane, N. and Martin, W. (2010) 'The energetics of genome complexity.', *Nature*. Nature Publishing Group, 467(7318), pp. 929–934. doi: 10.1038/nature09486.

Lang, M., Stelzer, M. and Schomburg, D. (2011) 'BKM-react, an integrated biochemical reaction database', *BMC Biochemistry*, 12(1). doi: 10.1186/1471-2091-12-42.

Lawrence, J. (1999) 'Selfish operons: The evolutionary impact of gene clustering in prokaryotes

and eukaryotes', *Current Opinion in Genetics and Development*, pp. 642–648. doi: 10.1016/S0959-437X(99)00025-8.

Ledezma-Tejeida, D., Ishida, C. and Collado-Vides, J. (2017) 'Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in Escherichia coli', *Frontiers in Microbiology*, 8(AUG). doi: 10.3389/fmicb.2017.01466.

Leduc, S. (1912) 'La biologie synthetique', *A. Poinat*.

Lee, D. J., Minchin, S. D. and Busby, S. J. W. (2012) 'Activating Transcription in Bacteria', *Annual Review of Microbiology*. Annual Reviews, 66(1), pp. 125–152. doi: 10.1146/annurev-micro-092611-150012.

Lefstin, J. A. and Yamamoto, K. R. (1998) 'Allosteric effects of DNA on transcriptional regulators', *Nature*, pp. 885–888. doi: 10.1038/31860.

Lewis, N. E., Nagarajan, H. and Palsson, B. O. (2012) 'Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.', *Nature reviews. Microbiology*, 10(4), pp. 291–305. doi: 10.1038/nrmicro2737.

Liebl, W. *et al.* (2014) 'Alternative hosts for functional (meta)genome analysis', *Applied Microbiology and Biotechnology*, 98(19), pp. 8099–8109. doi: 10.1007/s00253-014-5961-7.

Lim, W. A. (2010) 'Designing customized cell signalling circuits', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 11(6), pp. 393–403. doi: 10.1038/nrm2904.

Little, J. W. *et al.* (1980) 'Cleavage of the Escherichia coli lexA protein by the recA protease.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 77(6), pp. 3225–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6447873 (Accessed: 10 June 2018).

Locey, K. J. and Lennon, J. T. (2016) 'Scaling laws predict global microbial diversity', *Proceedings of the National Academy of Sciences*, 113(21), pp. 5970–5975. doi: 10.1073/pnas.1521291113.

Loewe, L. (2009) *A framework for evolutionary systems biology.*, *BMC systems biology*. doi: 10.1186/1752-0509-3-27.

Loewenstein, W. R. (2006) *The Touchstone of Life : Molecular Information, Cell Communication, and the Foundations of Life*. Oxford University Press. Available at: https://books.google.com.br/books?hl=en&lr=&id=g82cx_99NbQC&oi=fnd&pg=PR7&dq=Loewenstein++evolution+information&ots=h7jO6bSzZ-&sig=ERIKvWOOzSczyUIHrpvlzsFaw3A#v=onepage&q=Loewenstein evolution information&f=false (Accessed: 25 June 2017).

Lonetto, M., Gribskov, M. and Gross, C. A. (1992) 'The sigma 70 family: sequence conservation and evolutionary relationships.', *Journal of bacteriology*, 174(12), pp. 3843–9. doi: 1597408.

de Lorenzo, V. and Danchin, A. (2008) 'Synthetic biology: Discovering new worlds and new words. The new and not so new aspects of this emerging research field', *EMBO Reports*, 9(9), pp. 822–827. doi: 10.1038/embor.2008.159.

Louca, S. and Doebeli, M. (2015) 'Calibration and analysis of genome-based models for microbial ecology.', *eLife*. eLife Sciences Publications, Ltd, 4, p. e08208. doi:

10.7554/eLife.08208.

Lu, C., Bentley, W. E. and Rao, G. (2004) 'A High-Throughput Approach to Promoter Study Using Green Fluorescent Protein', *Biotechnology Progress*, 20(6), pp. 1634–1640. doi: 10.1021/bp049751l.

M. Madan Babu (2013) *Bacterial Gene Regulation and Transcriptional Networks - Google Books*. Caister Academic Press. Available at: https://books.google.lk/books?id=8dUAAgAAQBAJ&pg=PA67&dq=operon+definition&hl=en&sa=X&ved=0ahUKEwjj9aO8-dfPAhUEFiwKHfsvDC0Q6AEIJTAC#v=onepage&q=operon definition&f=false (Accessed: 12 June 2018).

Machado, D., Herrgård, M. J. and Rocha, I. (2016) 'Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction', *PLoS Computational Biology*, 12(10). doi: 10.1371/journal.pcbi.1005140.

Mahadevan, R. and Palsson, B. O. (2005) 'Properties of metabolic networks: Structure versus function', *Biophysical Journal*, 88(1). doi: 10.1529/biophysj.104.055723.

Mann, S. and Chen, Y. P. P. (2010) 'Bacterial genomic G + C composition-eliciting environmental adaptation', *Genomics*. Elsevier Inc., 95(1), pp. 7–15. doi: 10.1016/j.ygeno.2009.09.002.

Mao, X. *et al.* (2012) 'The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces', *Nucleic Acids Research*, 40(17), pp. 8210–8218. doi: 10.1093/nar/gks605.

Mao, X. *et al.* (2014) 'DOOR 2.0: presenting operons and their functions through dynamic and integrated views', *Nucleic Acids Research*, 42(D1), pp. D654–D659. doi: 10.1093/nar/gkt1048.

Mao, X. *et al.* (2015) 'Revisiting operons: an analysis of the landscape of transcriptional units in E. coli.', *BMC bioinformatics*. BMC Bioinformatics, 16(1), p. 356. doi: 10.1186/s12859-015-0805-8.

Martínez-Antonio, A. *et al.* (2003) 'Identifying global regulators in transcriptional regulatory networks in bacteria', *Current Opinion in Microbiology*, 6(5), pp. 482–489. doi: 10.1016/j.mib.2003.09.002.

Martínez-Antonio, A. (2011) 'Escherichia coli transcriptional regulatory network', *Network Biology*, 1(1), pp. 21–33. Available at: www.iaees.org (Accessed: 25 June 2017).

McCloskey, D., Palsson, B. Ø. and Feist, A. M. (2013) 'Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli.', *Molecular systems biology*, 9(1), p. 661. doi: 10.1038/msb.2013.18.

McClymont, K. and Soyer, O. S. (2013) 'Metabolic tinker: An online tool for guiding the design of synthetic metabolic pathways', *Nucleic Acids Research*. Oxford University Press, 41(11), pp. e113–e113. doi: 10.1093/nar/gkt234.

McCutcheon, J. P. and Moran, N. A. (2011) 'Extreme genome reduction in symbiotic bacteria.', *Nature reviews. Microbiology*, 10(1), pp. 13–26. doi: 10.1038/nrmicro2670.

McLeod, S. M. and Johnson, R. C. (2001) 'Control of transcription by nucleoid proteins',

Current Opinion in Microbiology. doi: 10.1016/S1369-5274(00)00181-8.

Medina, M. (2005) 'Genomes, phylogeny, and evolutionary systems biology', *Proceedings of the National Academy of Sciences*, 102(Supplement 1), pp. 6630–6635. doi: 10.1073/pnas.0501984102.

Milo, R. *et al.* (2002) 'Network motifs: Simple building blocks of complex networks', *Science*, 298(5594), pp. 824–827. doi: 10.1126/science.298.5594.824.

Mitchell, J. E. (2003) 'Identification and analysis of "extended -10" promoters in Escherichia coli', *Nucleic Acids Research*, 31(16), pp. 4689–4695. doi: 10.1093/nar/gkg694.

Monk, J., Nogales, J. and Palsson, B. O. (2014) 'Optimizing genome-scale network reconstructions', *Nature Biotechnology*, pp. 447–452. doi: 10.1038/nbt.2870.

Monteiro, L. M. O., Arruda, L. M. and Silva-Rocha, R. (2017) 'Emergent Properties in Complex Synthetic Bacterial Promoters', *ACS Synthetic Biology*, p. acssynbio.7b00344. doi: 10.1021/acssynbio.7b00344.

Morris, B. E. L. *et al.* (2013) 'Microbial syntrophy: interaction for the common good', *FEMS Microbiology Reviews*, 37(3), pp. 384–406. doi: 10.1111/1574-6976.12019.

Morris, M. K. *et al.* (2010) 'Logic-based models for the analysis of cell signaling networks', *Biochemistry*. doi: 10.1021/bi902202q.

Münch, R. *et al.* (2003) 'PRODORIC: prokaryotic database of gene regulation.', *Nucleic acids research*. Oxford University Press, 31(1), pp. 266–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12519998 (Accessed: 25 June 2017).

Nakagawa, S. and Cuthill, I. C. (2007) 'Effect size, confidence interval and statistical significance: a practical guide for biologists', *Biological Reviews*, 82(4), pp. 591–605. doi: 10.1111/j.1469-185X.2007.00027.x.

Newburger, D. E. and Bulyk, M. L. (2009) 'UniPROBE: an online database of protein binding microarray data on protein-DNA interactions', *Nucleic Acids Research*. Oxford University Press, 37(Database), pp. D77–D82. doi: 10.1093/nar/gkn660.

Newman, M. E. J. (Mark E. J. . (2010) *Networks : an introduction*. Oxford University Press.

Notebaart, R. A. *et al.* (2006) 'Accelerating the reconstruction of genome-scale metabolic networks', *BMC Bioinformatics*, 7. doi: 10.1186/1471-2105-7-296.

Nowak-Lovato, K. *et al.* (2013) 'Binding of Nucleoid-Associated Protein Fis to DNA Is Regulated by DNA Breathing Dynamics', *PLoS Computational Biology*, 9(1). doi: 10.1371/journal.pcbi.1002881.

O'Brien, E. J., Monk, J. M. and Palsson, B. O. (2015) 'Using Genome-scale Models to Predict Biological Capabilities', *Cell*, 161(5), pp. 971–987. doi: 10.1016/j.cell.2015.05.019.

O'Malley, M. a. and Soyer, O. S. (2012) 'The roles of integration in molecular systems biology', *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*. Elsevier Ltd, 43(1), pp. 58–68. doi: 10.1016/j.shpsc.2011.10.006.

Oberhardt, M. A., Palsson, B. Ø. and Papin, J. A. (2009) 'Applications of genome-scale metabolic reconstructions.', *Molecular systems biology*. European Molecular Biology

Organization, 5, p. 320. doi: 10.1038/msb.2009.77.

Ogasawara, H. *et al.* (2010) 'Regulation of the Escherichia coli csgD promoter: Interplay between five transcription factors', *Microbiology*. doi: 10.1099/mic.0.039131-0.

Oh, Y.-K. *et al.* (2007) 'Genome-scale Reconstruction of Metabolic Network in *Bacillus subtilis* Based on High-throughput Phenotyping and Gene Essentiality Data', *Journal of Biological Chemistry*, 282(39), pp. 28791–28799. doi: 10.1074/jbc.M703759200.

Oltvai, Z. N. *et al.* (2000) 'The large-scale organization of metabolic networks', *Nature*. Nature Publishing Group, 407(6804), pp. 651–654. doi: 10.1038/35036627.

Orenstein, Y. and Shamir, R. (2016) 'Modeling protein?DNA binding via high-throughput in vitro technologies', *Briefings in Functional Genomics*, 16(August), p. elw030. doi: 10.1093/bfgp/elw030.

Orth, J. D., Palsson, B. Ø. and Fleming, R. M. T. (2010) 'Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide', *EcoSal Plus*, 4(1). doi: 10.1128/ecosalplus.10.2.1.

Orth, J. D., Thiele, I. and Palsson, B. Ø. (2010) 'What is flux balance analysis?', *Nature Biotechnology*, 28(3), pp. 245–248. doi: 10.1038/nbt.1614.

Ott, M. A. and Vriend, G. (2006) 'Correcting ligands, metabolites, and pathways', *BMC Bioinformatics*, 7. doi: 10.1186/1471-2105-7-517.

Overbeek, R. *et al.* (2014) 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic Acids Research*, 42(D1), pp. 206–214. doi: 10.1093/nar/gkt1226.

Pabo, C. O. and Sauer, R. T. (1992) 'Transcription Factors: Structural Families and Principles of DNA Recognition', *Annual Review of Biochemistry*. doi: 10.1146/annurev.bi.61.070192.005201.

Pace, N. R. *et al.* (1986) 'The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences', in *Advances in Microbial Ecology*, pp. 1–55. doi: 10.1007/978-1-4757-0611-6_1.

Paget, M. S. B. and Helmann, J. D. (2003) 'The sigma70 family of sigma factors.', *Genome biology*, 4(1), p. 203. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12540296.

Papin, J. A. *et al.* (2003) 'Metabolic pathways in the post-genome era.', *Trends in biochemical sciences*, 28(5), pp. 250–8. doi: 10.1016/S0968-0004(03)00064-1.

Papp, B., Notebaart, R. a and Pál, C. (2011) 'Systems-biology approaches for predicting genomic evolution.', *Nature reviews. Genetics*. Nature Publishing Group, 12(9), pp. 591–602. doi: 10.1038/nrg3033.

Patil, K. R., Roune, L. and McHardy, A. C. (2012) 'The PhyloPythiaS Web Server for Taxonomic Assignment of Metagenome Sequences', *PLoS ONE*. Edited by S. K. Highlander. Public Library of Science, 7(6), p. e38581. doi: 10.1371/journal.pone.0038581.

Payne, J. L. and Wagner, A. (2014) 'The Robustness and Evolvability of Transcription Factor Binding Sites', *Science*, 343(6173). Available at: http://science.sciencemag.org/content/343/6173/875 (Accessed: 25 June 2017).

Payne, J. L. and Wagner, A. (2015) 'Mechanisms of mutational robustness in transcriptional

regulation', *Frontiers in Genetics*. Frontiers Media SA, 6(OCT), p. 322. doi: 10.3389/fgene.2015.00322.

Peeters, E., Peixeiro, N. and Sezonov, G. (2013) 'Cis-regulatory logic in archaeal transcription.', *Biochemical Society transactions*, 41(1), pp. 326–31. doi: 10.1042/BST20120312.

Pérez-Rueda, E., Collado-Vides, J. and Perez-Rueda, E. (2000) 'The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12', *Nucleic Acids Research*, 28(8), pp. 1838–1847. doi: 10.1093/nar/28.8.1838.

Peter, I. S., Faure, E. and Davidson, E. H. (2012) 'Predictive computation of genomic logic processing functions in embryonic development', *Proceedings of the National Academy of Sciences*, 109(41), pp. 16434–16442. doi: 10.1073/pnas.1207852109.

Placzek, S. *et al.* (2017) 'BRENDA in 2017: New perspectives and new tools in BRENDA', *Nucleic Acids Research*, 45(D1), pp. D380–D388. doi: 10.1093/nar/gkw952.

Pushpam, P., Rajesh, T. and Gunasekaran, P. (2011) 'Identification and characterization of alkaline serine protease from goat skin surface metagenome', *AMB Express*, 1(1), p. 3. doi: 10.1186/2191-0855-1-3.

Raveh-Sadka, T. *et al.* (2012) 'Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast', *Nature Genetics*, 44(7), pp. 743–750. doi: 10.1038/ng.2305.

Reed, J. L. *et al.* (2006) 'Systems approach to refining genome annotation', *Proceedings of the National Academy of Sciences*, 103(46), pp. 17480–17484. doi: 10.1073/pnas.0603364103.

Reed, J. L. (2012) 'Shrinking the metabolic solution space using experimental datasets.', *PLoS computational biology*. Edited by J. A. Papin, 8(8), p. e1002662. doi: 10.1371/journal.pcbi.1002662.

Rowland, M. A. *et al.* (2017) 'Crosstalk and the Dynamical Modularity of Feed-Forward Loops in Transcriptional Regulatory Networks', *Biophysical Journal*. ElsevierCompany., 112(8), pp. 1539–1550. doi: 10.1016/j.bpj.2017.02.044.

Rydenfelt, M. *et al.* (2014) 'The influence of promoter architectures and regulatory motifs on gene expression in Escherichia coli', *PLoS ONE*, 9(12), pp. 1–31. doi: 10.1371/journal.pone.0114347.

Sambrook, J.; Fritsch, E. F.; Maniatis, T. (1989) *Molecular cloning: a laboratory manual*. 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sanches-Medeiros, A., Monteiro, L. M. O. and Silva-Rocha, R. (2018) 'Calibrating Transcriptional Activity Using Constitutive Synthetic Promoters in Mutants for Global Regulators in Escherichia coli', *International Journal of Genomics*. Hindawi, 2018, pp. 1–10. doi: 10.1155/2018/9235605.

Sayut, D. J., Niu, Y. and Sun, L. (2011) 'Engineering the logical properties of a genetic AND gate.', *Methods in molecular biology (Clifton, N.J.)*, 743, pp. 175–184. doi: 10.1007/978-1-61779-132-1_14.

Schaefer, C. E. G. R., Fabris, J. D. and Ker, J. C. (2008) 'Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review', *Clay Minerals*, 43(1), pp. 137–154. doi:

10.1180/claymin.2008.043.1.11.

Schaerli, Y. *et al.* (2017) 'Mechanistic causes of constrained phenotypic variation revealed by synthetic gene regulatory networks', *Doi.Org*, p. 184325. doi: 10.1101/184325.

Schaub, J. *et al.* (2012) 'Advancing biopharmaceutical process development by system-level data analysis and integration of omics data.', *Advances in biochemical engineering/biotechnology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 127, pp. 133–63. doi: 10.1007/10_2010_98.

Schellenberger, J. *et al.* (2010) 'BiGG: A Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions', *BMC Bioinformatics*, 11(1), p. 213. doi: 10.1186/1471-2105-11-213.

Schmeisser, C., Steele, H. and Streit, W. R. (2007) 'Metagenomics, biotechnology with non-culturable microbes', *Applied Microbiology and Biotechnology*, pp. 955–962. doi: 10.1007/s00253-007-0945-5.

Schneider, T. D. and Stephens, R. M. (1990) 'Sequence logos: A new way to display consensus sequences', *Nucleic Acids Research*. Oxford University Press, 18(20), pp. 6097–6100. doi: 10.1093/nar/18.20.6097.

Schomburg, I. *et al.* (2002) 'BRENDA: A resource for enzyme data and metabolic information', *Trends in Biochemical Sciences*, pp. 54–56. doi: 10.1016/S0968-0004(01)02027-8.

Schuster, P. (2002) 'A testable genotype-phenotype map: modeling evolution of RNA molecules', *Biological Evolution and Statistical Physics*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 55–81. doi: 10.1007/3-540-45692-9_4.

Serres, M. H. and Riley, M. (2000) 'MultiFun, a Multifunctional Classification Scheme for Escherichia coli K-12 Gene Products', *Microbial & Comparative Genomics*, 5(4), pp. 205–222. doi: 10.1089/mcg.2000.5.205.

Setty, Y. *et al.* (2003) 'Detailed map of a cis-regulatory input function', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1230759100.

Shahmuradov, I. A. *et al.* (2016) 'bTSSfinder: a novel tool for the prediction of promoters in Cyanobacteria and *Escherichia coli*', *Bioinformatics*, 33(September 2016), p. btw629. doi: 10.1093/bioinformatics/btw629.

Shannon, P. (2003) 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks', *Genome Research*, 13(11), pp. 2498–2504. doi: 10.1101/gr.1239303.

Sharma, U. K. and Chatterji, D. (2010) 'Transcriptional switching in Escherichia coli during stress and starvation by modulation of σ70 activity', *FEMS Microbiology Reviews*. doi: 10.1111/j.1574-6976.2010.00223.x.

Sharon, E. *et al.* (2012) 'Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters', *Nature Biotechnology*, 30(6), pp. 521–530. doi: 10.1038/nbt.2205.

Sharon, E. *et al.* (2014) 'Probing the effect of promoters on noise in gene expression using thousands of designed sequences', *Genome Research*, 24(10), pp. 1698–1706. doi: 10.1101/gr.168773.113.

Shen-Orr, S. *et al.* (no date) 'Network motifs in the transcriptional regulation network of Escherichia coli', *nature.com*. Available at: https://www.nature.com/articles/ng881 (Accessed: 12 June 2018).

Shimada, T. *et al.* (2005) 'Systematic search for the Cra-binding promoters using genomic SELEX system', *Genes to Cells*, 10(9), pp. 907–918. doi: 10.1111/j.1365-2443.2005.00888.x.

Shimada, T. *et al.* (2014) 'The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of Escherichia coli', *PLoS ONE*, 9(3). doi: 10.1371/journal.pone.0090447.

Silander, O. K. *et al.* (2012) 'A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli', *PLoS Genetics*, 8(1). doi: 10.1371/journal.pgen.1002443.

Silva-Rocha, R. *et al.* (2011) 'The logicome of environmental bacteria: Merging catabolic and regulatory events with Boolean formalisms', *Environmental Microbiology*, pp. 2389–2402. doi: 10.1111/j.1462-2920.2011.02455.x.

Silva-Rocha, R. and de Lorenzo, V. (2008) 'Mining logic gates in prokaryotic transcriptional regulation networks', *FEBS Letters*, 582(8), pp. 1237–1244. doi: 10.1016/j.febslet.2008.01.060.

Silva-Rocha, R. and De Lorenzo, V. (2011) 'Implementing an OR-NOT (ORN) logic gate with components of the SOS regulatory network of Escherichia coli', *Molecular BioSystems*. doi: 10.1039/c1mb05094j.

Silva-Rocha, R. and De Lorenzo, V. (2012) 'Broadening the signal specificity of prokaryotic promoters by modifying cis-regulatory elements associated with a single transcription factor', *Molecular BioSystems*. doi: 10.1039/c2mb25030f.

Silva-Rocha, R. and De Lorenzo, V. (2013) 'The TOL network of Pseudomonas putida mt-2 processes multiple environmental inputs into a narrow response space', *Environmental Microbiology*, 15(1), pp. 271–286. doi: 10.1111/1462-2920.12014.

Silva-Rocha, R., Tamames, J. and de Lorenzo, V. (2012) 'The Logic of Decision Making in Environmental Bacteria', *Biomolecular Information Processing: From Logic Systems to Smart Sensors and Actuators*, pp. 279–302. doi: 10.1002/9783527645480.ch15.

De Silva, A. P. and Uchiyama, S. (2007) 'Molecular logic and computing', *Nature Nanotechnology*. doi: 10.1038/nnano.2007.188.

Singh, J. *et al.* (2009) 'Metagenomics: Concept, methodology, ecological inference and recent advances', *Biotechnology Journal*, 4(4), pp. 480–494. doi: 10.1002/biot.200800201.

Siuti, P., Yazbek, J. and Lu, T. K. (2013) 'Synthetic circuits integrating logic and memory in living cells', *Nature Biotechnology*. doi: 10.1038/nbt.2510.

Sleator, R. D., Shortall, C. and Hill, C. (2008) 'Metagenomics', *Letters in Applied Microbiology*, pp. 361–366. doi: 10.1111/j.1472-765X.2008.02444.x.

Solé, R. (2015) 'Bioengineering the biosphere?', *Ecological Complexity*, 22, pp. 40–49. doi: 10.1016/j.ecocom.2015.01.005.

Solovyev, V. and Salamov, A. (2011) 'Automatic Annotation of Microbial Genomes and Metagenomic Sequences', in R.W., L. (ed.) *Metagenomics and its Applications in Agriculture,*

Biomedicine and Environmental Studies. Nova Science Publishers, pp. 61–78.

Soyer, O. S. and O'Malley, M. a. (2013) 'Evolutionary systems biology: What it is and why it matters', *BioEssays*, 35(8), pp. 696–705. doi: 10.1002/bies.201300029.

Stallins, J. A. *et al.* (2018) 'Geography and postgenomics: how space and place are the new DNA', *GeoJournal*. Springer Netherlands, 83(1), pp. 153–168. doi: 10.1007/s10708-016-9763-6.

Stams, A. J. M. (1994) 'Metabolic interactions between anaerobic bacteria in methanogenic environments', *Antonie van Leeuwenhoek*, 66(1–3), pp. 271–294. doi: 10.1007/BF00871644.

Stephens, Z. D. *et al.* (2015) 'Big data: Astronomical or genomical?', *PLoS Biology*. Public Library of Science, 13(7), p. e1002195. doi: 10.1371/journal.pbio.1002195.

Stern, D. L. and Orgogozo, V. (2008) 'The loci of evolution: How predictable is genetic evolution?', *Evolution*, pp. 2155–2177. doi: 10.1111/j.1558-5646.2008.00450.x.

Stewart, A. J., Hannenhalli, S. and Plotkin, J. B. (2012) 'Why Transcription Factor Binding Sites Are Ten Nucleotides Long', *Genetics*. Genetics Society of America, 192(3), pp. 973–985. doi: 10.1534/genetics.112.143370.

Stewart, A. J. and Plotkin, J. B. (2013) 'The evolution of complex gene regulation by low-specificity binding sites', *Proceedings of the Royal Society B: Biological Sciences*, 280(1768), pp. 20131313–20131313. doi: 10.1098/rspb.2013.1313.

Stewart, F. J., Ulloa, O. and Delong, E. F. (2012) 'Microbial metatranscriptomics in a permanent marine oxygen minimum zone', *Environmental Microbiology*, 14(1), pp. 23–40. doi: 10.1111/j.1462-2920.2010.02400.x.

Stolyar, S. *et al.* (2007) 'Metabolic modeling of a mutualistic microbial community', *Molecular Systems Biology*, 3. doi: 10.1038/msb4100131.

Stormo, G. D. (2000) 'DNA binding sites: Representation and discovery', *Bioinformatics*, pp. 16–23. doi: 10.1093/bioinformatics/16.1.16.

Stormo, G. D. and Hartzell, G. W. (1989) 'Identifying protein-binding sites from unaligned DNA fragments.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 86(4), pp. 1183–1187. doi: 10.1073/pnas.86.4.1183.

Struhl, K. (1999) 'Fundamentally different logic of gene regulation in eukaryotes and prokaryotes', *Cell*, 98(1), pp. 1–4. doi: 10.1016/S0092-8674(00)80599-1.

Suthers, P. F. *et al.* (2009) 'A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189.', *PLoS computational biology*. Edited by H. M. Sauro, 5(2), p. e1000285. doi: 10.1371/journal.pcbi.1000285.

Tamsir, A., Tabor, J. J. and Voigt, C. A. (2011) 'Robust multicellular computing using genetically encoded NOR gates and chemical "wiresg"', *Nature*. Nature Publishing Group, 469(7329), pp. 212–215. doi: 10.1038/nature09565.

Tan, K., Mccue, L. A. and Stormo, G. D. (2005) 'Making connections between novel transcription factors and their DNA motifs Making connections between novel transcription factors and their DNA motifs', *Genome Research*, 15(1), pp. 312–320. doi: 10.1101/gr.3069205.

Thiele, I. and Palsson, B. Ø. (2010) 'A protocol for generating a high-quality genome-scale metabolic reconstruction.', *Nature protocols*. NIH Public Access, 5(1), pp. 93–121. doi: 10.1038/nprot.2009.203.

Thompson, L. R. *et al.* (2017) 'A communal catalogue reveals Earth's multiscale microbial diversity', *Nature*, 551(7681), pp. 457–463. doi: 10.1038/nature24621.

Tkačik, G. and Bialek, W. (2014) 'Information processing in living systems'. doi: 10.1146/annurev-conmatphys-031214-014803.

Torsvik, V. and Øvreås, L. (2002) 'Microbial diversity and function in soil: from genes to ecosystems.', *Current opinion in microbiology*, 5(3), pp. 240–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12057676.

Tringe, S. G. *et al.* (2005) 'Comparative metagenomics of microbial communities', *Science*, 308(5721), pp. 554–557. doi: 10.1126/science.1107851.

Tringe, S. G. (2005) 'Comparative Metagenomics of Microbial Communities', *Science*, 308(5721), pp. 554–557. doi: 10.1126/science.1107851.

Uchiyama, T. *et al.* (2005) 'Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes.', *Nature biotechnology*, 23(1), pp. 88–93. doi: 10.1038/nbt1048.

Uchiyama, T. and Miyazaki, K. (2010) 'Product-Induced Gene Expression, a Product-Responsive Reporter Assay Used To Screen Metagenomic Libraries for Enzyme-Encoding Genes', *Applied and Environmental Microbiology*, 76(21), pp. 7029–7035. doi: 10.1128/AEM.00464-10.

Vanet, A., Marsan, L. and Sagot, M. F. (1999) 'Promoter sequences and algorithmical methods for identifying them', *Research in Microbiology*, 150(9–10), pp. 779–799. doi: 10.1016/S0923-2508(99)00115-1.

Varela, H. R. M. and F. J. (1980) *Autopoiesis and cognition the realization of the living. With a pref. to autopoiesis by Sir Stafford beer*. D. Reidel Pub. Co. Available at: https://books.google.com.br/books?hl=en&lr=&id=nVmcN9Ja68kC&oi=fnd&pg=PR17&dq=Autopoiesis+and+cognition:+The+realization+of+the+living&ots=_oq56RAg3k&sig=ruzmgV4LQNfeU1Lsma1R7n_UDQk#v=onepage&q=Autopoiesis and cognition%3A The realization of the living&f= (Accessed: 12 June 2018).

Varma, A. and Palsson, B. O. (1994) 'Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.', *Applied and environmental microbiology*, 60(10), pp. 3724–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7986045 (Accessed: 29 May 2018).

Vastrik, I. *et al.* (2007) 'Reactome: A knowledge base of biologic pathways and processes', *Genome Biology*, 8(3). doi: 10.1186/gb-2007-8-3-r39.

Del Vecchio, D. *et al.* (2018) 'Future systems and control research in synthetic biology', *Annual Reviews in Control*, May. doi: 10.1016/j.arcontrol.2018.04.007.

Venter, J. C. (2004) 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74. doi: 10.1126/science.1093857.

Vester, J. K., Glaring, M. A. and Stougaard, P. (2015) 'Improved cultivation and metagenomics

as new tools for bioprospecting in cold environments', *Extremophiles*, 19(1), pp. 17–29. doi: 10.1007/s00792-014-0704-3.

Vieites, J. M. *et al.* (2009) 'Metagenomics approaches in systems microbiology', *FEMS Microbiology Reviews*, pp. 236–255. doi: 10.1111/j.1574-6976.2008.00152.x.

Wagner, A. (2005) *Robustness and evolvability in living systems*. 1st edn. Princeton University Press. Available at: https://books.google.com.br/books?hl=en&lr=&id=pRFYAQAAQBAJ&oi=fnd&pg=PP1&dq=Robustness+and+Evolvability+in+Living+Systems+pmid+wagner&ots=2TNsLi8XN4&sig=YOWeI1olCmDgQpKEgK2xovymGwY#v=onepage&q&f=false (Accessed: 12 June 2018).

Wagner, A. *et al.* (2012) 'Evolutionary Systems Biology', *Advances in Experimental Medicine and Biology*, 751, pp. 29–53. doi: 10.1007/978-1-4614-3567-9.

Wang, B. *et al.* (2011) 'Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology', *Nature Communications*. doi: 10.1038/ncomms1516.

Wang, L. *et al.* (2015) 'SynBioLGDB: A resource for experimentally validated logic gates in synthetic biology', *Scientific Reports*. doi: 10.1038/srep08090.

Wang, R. S., Saadatpour, A. and Albert, R. (2012) 'Boolean modeling in systems biology: An overview of methodology and applications', *Physical Biology*. doi: 10.1088/1478-3975/9/5/055001.

Westmann, C. A. *et al.* (2018) 'Transcriptional Regulation of Hydrocarbon Efflux Pump Expression in Bacteria', in *Cellular Ecophysiology of Microbe: Hydrocarbon and Lipid Interactions*. Cham: Springer International Publishing, pp. 177–200. doi: 10.1007/978-3-319-50542-8_4.

Westmann, C. A., Guazzaroni, M. and Silva-Rocha, R. (2018) 'Engineering Complexity in Bacterial Regulatory Circuits for Biotechnological Applications', *mSystems*, 3(2), pp. e00151-17. doi: 10.1128/mSystems.00151-17.

Whitacre, J. M. (2010) 'Degeneracy: A link between evolvability, robustness and complexity in biological systems', *Theoretical Biology and Medical Modelling*, 7(1), p. 6. doi: 10.1186/1742-4682-7-6.

Wiench, B. *et al.* (2013) 'Integration of Different "-omics" Technologies Identifies Inhibition of the IGF1R-Akt-mTOR Signaling Cascade Involved in the Cytotoxic Effect of Shikonin against Leukemia Cells', *Evidence-Based Complementary and Alternative Medicine*, 2013, pp. 1–11. doi: 10.1155/2013/818709.

Williamson, L. L. *et al.* (2005) 'Intracellular Screen To Identify Metagenomic Clones That Induce or Inhibit a Quorum-Sensing Biosensor', *Applied and Environmental Microbiology*, 71(10), pp. 6335–6344. doi: 10.1128/AEM.71.10.6335-6344.2005.

Wilson, C. J. *et al.* (2007) 'The lactose repressor system: Paradigms for regulation, allosteric behavior and protein folding', *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-006-6296-z.

Wittkopp, P. J. and Kalay, G. (2012) 'Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence', *Nature Reviews Genetics*. Nature Publishing Group, 13(1), pp. 59–69. doi: 10.1038/nrg3095.

Wolf, L. (2014) ‘Evolution of transcriptional regulation in Escherichia coli’.

Wolf, L., Silander, O. K. and Nimwegen, E. van (2015) ‘Expression noise facilitates the evolution of gene regulation’, *eLife*, 4, p. e05856. doi: 10.7554/eLife.05856.

Wray, G. A. (2007) ‘The evolutionary significance of cis-regulatory mutations’, *Nature Reviews Genetics*. Nature Publishing Group, 8(3), pp. 206–216. doi: 10.1038/nrg2063.

Wunderlich, Z. *et al.* (2012) ‘Dissecting sources of quantitative gene expression pattern divergence between Drosophila species’, *Molecular Systems Biology*, 8. doi: 10.1038/msb.2012.35.

Wunderlich, Z. and Mirny, L. A. (2009) ‘Different gene regulation strategies revealed by analysis of binding motifs.’, *Trends in genetics : TIG*. NIH Public Access, 25(10), pp. 434–40. doi: 10.1016/j.tig.2009.08.003.

Xie, M. and Fussenegger, M. (2018) ‘Designing cell function: assembly of synthetic gene circuits for cell biology applications’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/s41580-018-0024-z.

Yashin, R., Rudchenko, S. and Stojanovic, M. N. (2007) ‘Networking particles over distance using oligonucleotide-based devices’, *Journal of the American Chemical Society*. doi: 10.1021/ja074335t.

Yona, A. H., Alm, E. J. and Gore, J. (2018) ‘Random sequences rapidly evolve into de novo promoters’, *Nature Communications*. Springer US, 9(1), p. 1530. doi: 10.1038/s41467-018-04026-w.

Yuan, Y. *et al.* (2018) ‘Regulation by competition: a hidden layer of gene regulatory network’, *bioRxiv*, p. 258129. doi: 10.1101/258129.

Yugi, K. *et al.* (2016) ‘Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple “Omic” Layers’, *Trends in Biotechnology*, 34(4), pp. 276–290. doi: 10.1016/j.tibtech.2015.12.013.

Zengler, K. and Palsson, B. O. (2012) ‘A road map for the development of community systems (CoSy) biology’, *Nature Reviews Microbiology*, 10(5), pp. 366–372. doi: 10.1038/nrmicro2763.

Zhang, W., Li, F. and Nie, L. (2010) ‘Integrating multiple “omics” analysis for microbial biology: application and methodologies’, *Microbiology*, 156(2), pp. 287–301. doi: 10.1099/mic.0.034793-0.

Zhu, H. *et al.* (2013) ‘Integrated OMICS guided engineering of biofuel butanol-tolerance in photosynthetic Synechocystis sp. PCC 6803.’, *Biotechnology for biofuels*, 6(1), p. 106. doi: 10.1186/1754-6834-6-106.

Zou, W. *et al.* (2012) ‘Reconstruction and analysis of a genome-scale metabolic model of the vitamin C producing industrial strain Ketogulonicigenium vulgare WSH-001’, *Journal of Biotechnology*, 161(1), pp. 42–48. doi: 10.1016/j.jbiotec.2012.05.015.

VI. ANNEXES

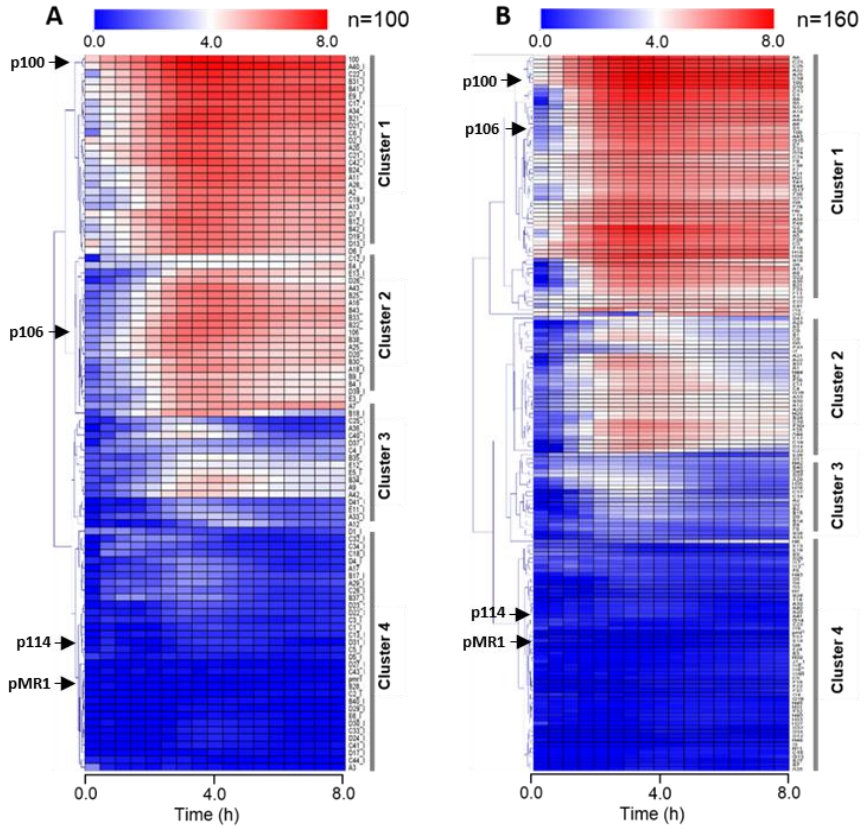


Figure S1. Hierarchical representation of metaconstitutomes for both metagenomic libraries highlighting expression trends as clusters. Fluorescence time-lapse dynamics were measured during 8 hours for each clone and represented as heat maps. Promoter activities (calculated as GFP/OD_{600}) were normalised by the negative control (*E. coli* DH10B harbouring empty pMR1) and transformed to log2 scale in order to facilitate the visualisation of subtle activities. Positive controls (p100, p106 and p114 - strong, medium and low expression, respectively) and negative control (pMR1) expression profiles are indicated by black arrows at the left side of the heatmap. Data are representative of three independent experiments. **(A)** Dendrogram for USP3 metagenomic library composed by 100 fluorescent clones. **(B)** Dendrogram for USP1 metagenomic library composed by 160 fluorescent clones.

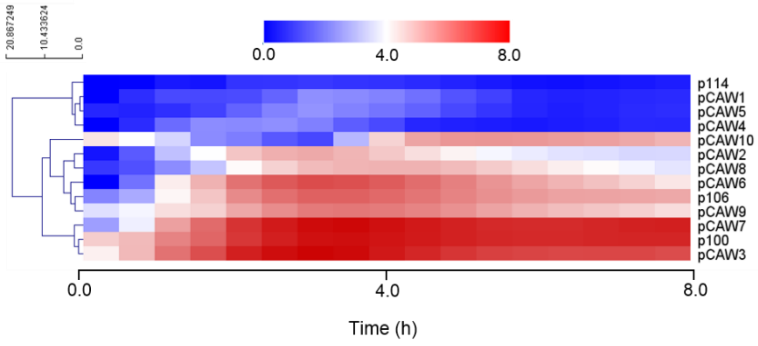


Figure S2. Expression profiles for the ten selected clones (pCAW1-pCAW10). Fluorescence time-lapse dynamics were measured during 8 hours for each clone and represented as heat maps. Promoter activities (calculated as GFP/OD600) were normalised by negative control (E. coli DH10B harbouring empty pMR1) and transformed to log2 scale in order to facilitate the visualisation of subtle activities. Data are representative of three independent experiments. Clones p100, p106 and p114 are positive controls for different promoter strengths, representing strong, medium and low expression, respectively. Hierarchical clustering of the selected clones according to their expression profiles.

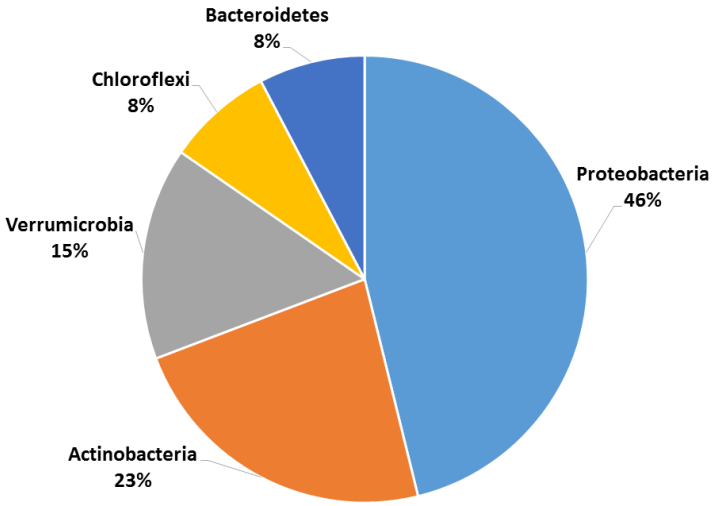


Figure S3. Abundance of microbial phyla with recognizable regulatory sequences in *E. coli* DH10B. The ten sequenced metagenomic fragments (pCAW1-pCAW10) were submitted to the *PhylopythiaS Web Server* and the assigned taxonomic origins were used for identifying the potential set of phyla recognizable by *E. coli* regarding exogenous promoter sequences.

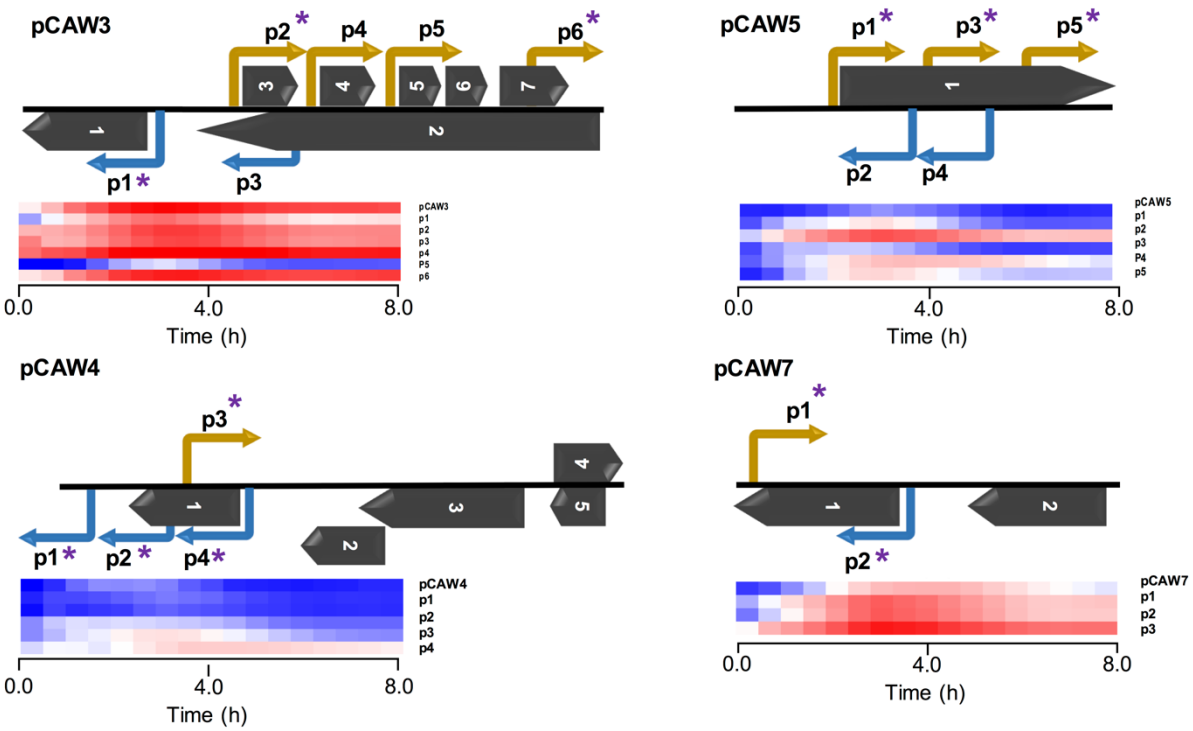


Figure S4. Schematic representation of the supplementary set of sequenced contigs showing predicted ORFs and validated/characterised promoters used in this work. Each contig is identified on the far left of each subfigure. Promoters are indicated by elbow-shaped arrows and name according to their relative position in the contig. Promoter directionality, regarding the leading and lagging strands, is represented by green and red colours, respectively. Asterisks over specific promoters indicate regulatory regions which were cross-validated by matching *in silico* predictions. Dark arrows represent predicted ORFs, according to their relative positions in each

contig (see **Table 5** for more information). Beneath each metagenomic insert, there is a heat map cluster representing the whole set of promoter activities measured during 8-hours fluorescence assays. A colour scale for all heat maps is provided in the first figure. The first line of each cluster shows the original expression profile initially measured for each metagenomic insert. All other lines represent expression activities from de novo experimentally validated promoters within each contig. The second line of each cluster represents the endogenous promoter showing the most similar activity with respect to the original expression profile for each contig. All expression profiles are properly identified at the most rightmost side of each line, following their respective contig/promoter name.

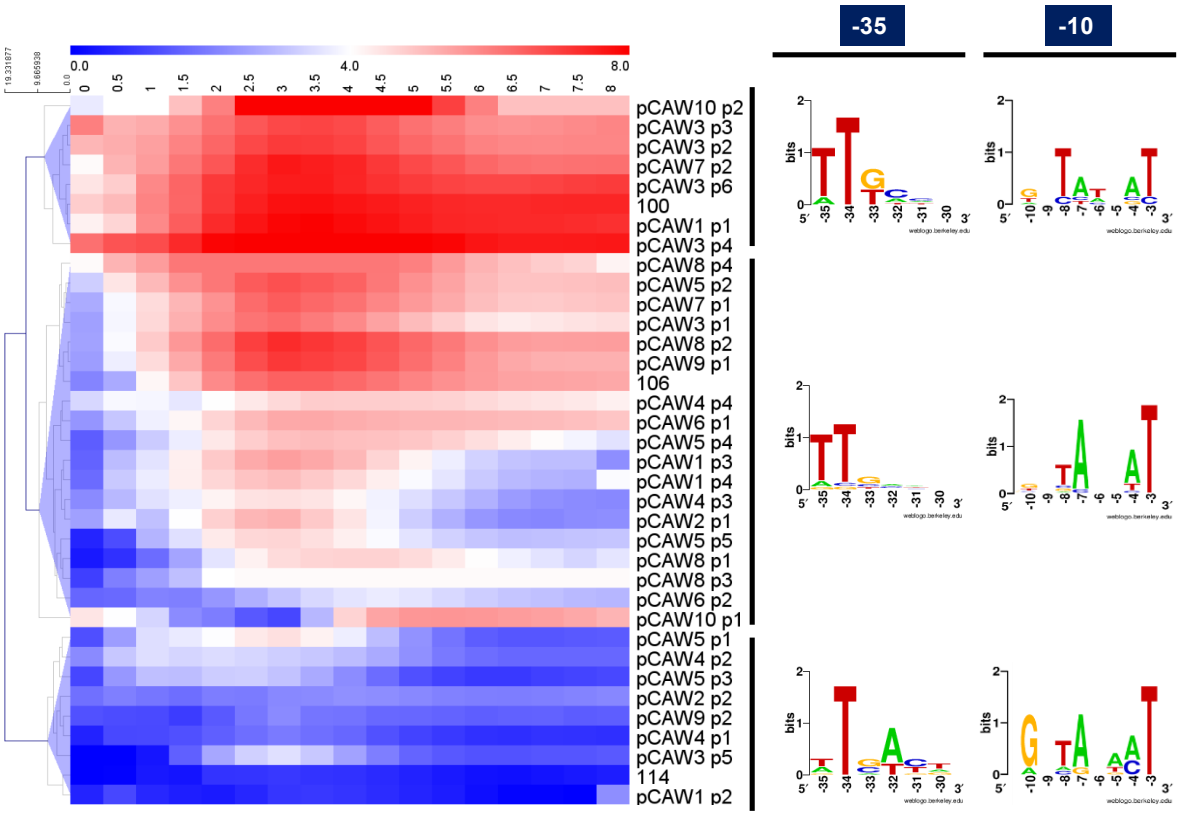


Figure S5. Consensus sequences for hierarchically clustered sets of experimentally validated promoters. Fluorescence time-lapse dynamics were measured during 8 hours for each clone and represented as heat maps. Promoter activities (calculated as GFP/OD600) were normalized by negative control (*E. coli* DH10B harboring empty pMR1) and transformed to log2 scale in order to facilitate the visualization of subtle activities. Data are representative of three independent experiments. Clones p100, p106 and p114 are positive controls for different promoter strengths, representing strong, medium and low expression strengths, respectively. All thirty-three experimentally validated promoters were organized by a hierarchical clustering method, revealing three general categories: strong (top), medium (middle) and weak (bottom) promoters. For each category, its respective set of promoter sequences was aligned using ClustalW (http://www.genome.jp/tools-bin/clustalw) and subjected to the *WebLogo* platform (https://weblogo.berkeley.edu/logo.cgi) for the generation of consensus sequences (right side).

Table S1: Experimentally validated metagenomic promoters found in this study

Clone_ Sample	Promoter ID	Sequence	Orientation	*In silico* validation
pCAW1	p1	TACCGGTAACGACTTAGATGGGAGGCCGACACTGTACAACGTCGGTGTGTAGTTGG	Reverse	Yes
	p2	ACGGGATTCTATTGACTGCGGCTGCGGCTGTCAACAGTCAAAATTCGGTAATCGGCGCCGTGA	Forward	Yes
	p3	TTGCACTCGTCCGACAAAACTGCACCAACTACCGGCATTGATTAGAGTTTTGAAAATAGAGTTTAACCACGAT	Forward	No
	p4	GTTGATCGGTGAGATTGGCCGCATCACCGCGGCTGAGGCGCGCGCGCTTGGCTACACGAACGTCTA	Forward	Yes
pCAW2	p1	TTGCTCACCATACCAACCTCCCTTGCGAATTTTAATTAAGGCTGAATTCAAGTGGAT	Forward	Yes
	p2	AGCGCATTCAATGACCTGTTCAACGATGTCCCGTCCTCTCGAAAAACTTTCGCCGGTCGGTTCGCGACAATTTCGAAGGCCAGAATGGATATCTAATGACCGT	Forward	Yes
pCAW3	p1	ATCGTCACCTCCACAAAGAGCGACTCGCTGTATACCGTTGGCATGCTAGCTTTATCTGT	Reverse	Yes
	p2	CTAAGCACCTTCGGTAGTTTCCTGGAACGAAGCCGCTGAAATCCAGCTCTGCGTACCCAGTGAAGCC	Forward	Yes
	p3	GCGCGCAGGCCAGATCGTTAGCCTGAGGGAAGTGAGAGAAGAAAACTCGATCCTCCGCAGGAACGATAAGAAAC	Reverse	No
	p4	GGTAAACTTCCTGTATCTCGTTCACAACATATTCCTGAACCGCGAATGGTCCTTTGA	Forward	No
	p5	CAGATCGGTTGTCCTTTGTCTACCTTCTGACCATCTTTCACCAGCAGTGTAGCTCCGTAAGGAACGTTGTTGGTGA	Forward	No
	p6	CGCGTCCAACAGTTGTTTCAATCAGTTTTCTTTTCAACTCACCACCTTCACGCACTATCGTTTTTACTTTGATGTGTGCGTGCAGGTCAACCTGTG	Forward	Yes
pCAW4	p1	TTCTGCATTGGCTACAGCAGGAACTGCGCCGCCGCCAGCGCGCTGGCGACGCTGGGATGGTACCCTACCCT	Reverse	Yes
	p2	ACCCGGTCGATCCGGTCTGCACAGACCGCTACACCTATCAGCCAGCCATACCAGCCGAGGGGTTTGGAGTCGGCCTGACCACGT	Reverse	Yes
	p3	ATACGTGGTCAGGCCGACTCCAAACCCCTCGGCTGGTATGGCTGGCTGATAGGTGTAGCGGTCTGTGCAGACCGGATCGACCGGGTTC	Forward	Yes
	p4	GGCAGCATCCATGCATCATTCCTCGGTAAAAGCCAGCCAGTAGGGGGTATGCCGTCTTGGAGTTTCACGACGT	Reverse	Yes
pCAW5	p1	ACCGGGCAGGACCGTCCCAAGCCAAAATATCCCGGCATCCCGGTGACCTGTAACGGCAATCAACTCGTCGCCCAATACGTTGA	Forward	Yes
	p2	GAGCATCCCGTCCTGGATATTCATCCCGGGATTGAGGGTGAGCTCGTTGACCTTGCGTAAAATAATCCCTTGAT	Reverse	No
	p3	TCCCGGGATGAATATCCAGGACGGGATGCTCACCACCCATTCGGAGCGAACCTACCTGGCGCCG	Forward	Yes
	p4	AAGTCAGGATTGGTTCATTGAAATTGTTGCGGCGGGCCACGACGCCGTTCATGTGATGTTCCTGGTTTTGCACTGGGCCGAGCA	Reverse	No
	p5	CGCCGCAACAATTTCAATGAACCAATCCTGACTTTCCTAAGCCGCGCCTACGAAGAATTCGGAAATCTCACCGGTCGCTATTACGG	Forward	Yes
pCAW6	p1	CCCGACTCCTTTGAAAGTATTCTTTTCCTGGTTTAGCATTGGCGCTCAATCATTTGGCGGCGGACCGTCCACCCTGCAACTAATTCAGAAC	Forward	Yes
	p2	ATCGGCCGGAGCAGAAGAATCGAGCAGGTCAGGGGTTTCCGAGGTCAGTTTAGGTCACTCCTTTCGGTG	Reverse	No

pCAW 7	p1	GTGGATCGTTTGGGTTATATTACCCTCAAAAAGGTTCGCAAACGCCCAATTGCCGTGTAACACGATATCAGGAGTATT	Forward	Yes
	p2	ACGATCTACTTCTCGTTTTGCCTTCTTTTGTCGCTACACTCCACAATCGGCTTAAGCCAGAGCATACCAACAGACCGGGTAGTTAACTGGAAACTA	Reverse	Yes
pCAW 8	p1	CTCTCGTCTTGCTCACCATACCAACCTCCCTTGCGAATTTTAATTAAGGCTGAATTCAAGTGGATCCCTATGATGGGGGTACAC	Forward	No
	p2	CCCTCGCACTTGTGCACTGGTGCCGGTGAAATCGAAGATGAGCGACGAGCCCCGGACAGTCACGGCGACGGCTAGCTTCACCGGCTGGTCGA	Reverse	Yes
	p3	ATCTTGTTCGCGGGCTTTTCCATTCGGTGGCTAAGCGCTATAGTTCGGCCCTATAGGAGAACATGCCA	Reverse	No
	p4	GCAAATGCTCCTTCTTCAATTCTATCCGCGCCTATGATCCCACCCGACAAATAGAACATAGACAAATTCGCGC	Forward	Yes
pCAW 9	p1	CTTAATTTCTCCTCTTTAATTCTAGGTACCCGGGGATCGCGATCGCAAGGATCATCGCTATGATGCCATGGGCTTCATGAA	Forward	Yes
	p2	TGGCATCATAGCGATGATCCTTGCGATCGCGATCCCCGGGTACCTAGAATTAAAGAGGAGA	Reverse	Yes
pCAW 10	p1	CATTTTTCGTTCAGGTTGCGTGCGCTTCGGCAGGCTCAGTGAGAACGAAGGCACCGATTGGTATGGA	Forward	Yes
	p2	GCACGCAAGGTTTCGCTATTGTGTGATTGACGCGGGTTCCGGCTTCGCCGGCCTGTCGCCTCAACTCAGTCCGACCAGCGACAATGCG	Reverse	No