UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXX CICLO - 2017

COMPUTATIONAL AND EXPERIMENTAL METHODS FOR METABOLIC ENGINEERING: APPLICATIONS IN *Escherichia coli* AND *Bacillus subtilis*

PhD Thesis by
Ilaria Massaiu

Advisors:
Prof. Paolo Magni
Prof. Lorenzo Pasotti

PhD Program Chair:
Prof. Riccardo Bellazzi

The mind once enlightened
cannot again become dark.

T. Paine

Abstract (English)

Metabolic engineering was defined more than 25 years ago as the directed modulation of metabolic pathways, using methods of recombinant DNA technology, for the purpose of overproducing high-value compounds, such as pharmaceutical products, food additives and fuels. Given the increasing need of more sustainable processes for the production of value-added chemicals and materials from renewable resources, metabolic engineering became a powerful tool for the development of highly efficient microbial cell factories.

The main innovation introduced from metabolic engineering, compared to traditional trial-and-error approaches, is the use of predictive modelling methods to study the behaviour of cellular metabolism and to guide the rational strain design. In this context, the cellular metabolism is described by the complete set of biochemical reactions that occur in the target microorganism, known as *genome-scale metabolic model*, and can be analyzed in terms of flux distributions, namely the reaction rates. Differently from gene expression levels or protein and metabolite concentration, the metabolic flux profiles are able to reflect the consequences of cellular component interactions.

Despite a variety of in-silico modelling approaches have been developed for the study of cellular metabolism, only those requiring a limited number of readily available parameters can be successfully applied to genome-scale models. Currently, *constraint-based* modelling approach is the best methodology by which genome-scale models are constructed and analyzed. This approach identifies a set of allowable solutions, by the assumption of steady-state conditions and limiting the fluxes, and then finds an unique flux distribution, by an optimization problem that maximizes or minimizes a biological objective function. Several methods based on different objective functions, and therefore appropriate for specific study goals, were developed. *Flux Balance Analysis* (FBA)

is the most popular method, which determines the flux through the metabolic network that maximizes growth rate. However, in some contexts the reliability of such models in the quantitative prediction of cellular phenotypes and fluxes through biochemical reactions can be low. The integration of additional biological information in the model, e.g., genome-scale transcriptomic or proteomic profiles, has been recently proposed as an attempt to improve prediction accuracy.

The last and crucial step for strain improvement is the application of genetic manipulations for the control of metabolic fluxes through recombinant DNA technologies. The perturbations, identified by the in-silico design phase, are implemented through the synthetic biology techniques for the tight control of gene expression levels, namely over-, down-expression and deletion. Synthetic biology is an emerging discipline, closely coupled with metabolic engineering field, that promotes the optimization of microorganisms using toolkits of pre-characterized regulatory elements. In particular, regulatory parts such as promoters or ribosome binding sites (RBS) are commonly used for the over- or down-regulation of transcriptional and translational processes of target genes, respectively, whereas gene knockouts are implemented using homologous recombination or silencing the gene via the new proposed techniques.

This thesis work includes both in-silico and in-vivo investigations on different metabolic engineering tools on *Escherichia coli* and *Bacillus subtilis*.

In Chapter 1 the key concepts of metabolic engineering field are introduced. An overview of the main computational approaches and experimental tools used in this field are presented.

In Chapter 2 a general analysis of the most widely used constraint-based methods applied for the study of *E. coli* metabolism is presented. The impact of input data, required for modelling the environmental conditions, on results is described and the prediction capability, under different genetic and environmental conditions, is then reported. Moreover, the integration of transcriptomic data into genome-scale metabolic model is implemented with the aim to increase the prediction accuracy.

In Chapter 3 the construction of an enzyme-constrained *B. subtilis* metabolic model is described and its performance, both for wild type and mutant strains, are reported. The target perturbations, identified using the developed model, for increasing the production of a relevant biopolymer are presented.

In Chapter 4 the design, construction and evaluation of small RNAs for the silencing of target genes expression in *E. coli* are described. Their repression performance is quantitatively evaluated, also with the help of mathematical

modelling.
In Chapter 5 a novel allelic replacement vector for chromosomal gene deletion in *E. coli*, based on the colorimetric XylE assay and the *BioBrick* standard, is proposed and used for the disruption of three genes encoding the production of organic acids, that compete for pyruvate utilization in ethanologenic strains.
Finally, in Chapter 6 the overall conclusions of this thesis work are drawn, considering the improvements obtained in the described studies, and the reliability of the investigated in-vivo and in-silico methods, based on high-impact case studies.
The studies illustrated in Chapter 2, 4 and 5 have been performed in the Cell Culture Laboratory of the Center for Health Technologies (CHT) and in the Laboratory of Bioinformatics, Mathematical Modelling and Synthetic Biology (BMS) University of Pavia, Italy, while the study reported in Chapter 3 has been carried out at the Novo Nordisk Foundation Center for Biosustainability (CFB), Technical University of Denmark.

Abstract (Italian)

L'ingegneria Metabolica è stata definita più di 25 anni fa come la disciplina finalizzata alla modifica diretta di pathway metabolici, attraverso l'utilizzo delle tecnologie del DNA ricombinante, in modo da aumentare la produzione di sostanze di interesse, come farmaci, additivi alimentari e biocarburanti. Data la crescente richiesta di processi biosostenibili per la produzione di composti chimici a partire da risorse rinnovabili, l'ingegneria metabolica si è affermata un potente strumento per lo sviluppo di efficienti *cell factory.*

La principale innovazione che distingue l'ingegneria metabolica, rispetto ai tradizionali approcci di mutazione genica trial-and-error, è l'utilizzo di metodi modellistici per lo studio del comportamento metabolico cellulare e per guidare la progettazione razionale di microorganismi. Il metabolismo cellulare viene descritto attraverso il completo set di reazioni identificate per il microorganismo in studio, noto come *modello metabolico genome-scale*, il quale può essere poi analizzato in termini di flussi, ossia la velocità con cui avviene una reazione. Questi ultimi, a differenza dei dati di espressione genica o concentrazione di proteine e metaboliti, hanno il vantaggio di descrivere le interazioni tra le diverse componenti cellulari.

Nonostante i molteplici approcci proposti per la modellizzazione in-silico del comportamento cellulare, solamente quelli per cui viene richiesto un numero limitato di parametri facilmente reperibile risultano adatti per l'applicazione ai modelli metabolici su scala genomica. Attualmente, l'approccio di modellizzazione *constraint-based* è riconosciuto come il miglior metodo per la costruzione e l'analisi dei modelli metabolici. Tale approccio identifica uno spazio di soluzioni possibili, attraverso l'assunzione di stato stazionario e vincolando il range di variazione di ciascun flusso, e successivamente trova un'unica soluzione, applicando un problema di ottimizzazione che massimizza o mini-

mizza una funzione obiettivo. Sono stati sviluppati differenti metodi basati sull'approccio *constraint-based*, il più famoso dei quali, chiamato *Flux Balance Analysis* (FBA), massimizza la velocità di crescita del microorganismo in studio. Tuttavia, in alcuni contesti tali modelli presentano una bassa affidabilità, in termini di predizioni quantitative del fenotipo cellulare e dei flussi attraverso le reazioni biochimiche. Recentemente, sono stati proposti nuovi approcci che prevedono l'integrazione di aggiuntive informazioni biologiche nel modello, come profili trascrittomici o proteomici su scala genomica, al fine di migliorare l'accuratezza di predizione.

L'ultima e fondamentale fase per l'ottenimento di microorganismi metabolicamente ottimizzati, consiste nell'applicazione delle manipolazioni geniche per il controllo dei flussi metabolici attraverso le tecnologie del DNA ricombinante. Tali perturbazioni, identificate dalla precedente fase di progettazione in-silico, vengono implementate tramite le avanzate tecniche di biologia sintetica, che permettono il controllo in maniera fine e predicibile dei livelli di espressione genica, ossia sovra-, sotto-espressione e delezione. La biologia sintetica è una disciplina emergente, strettamente legata all'ingegneria metabolica, che favorisce l'ottimizzazione di microorganismi mediante specifici elementi di regolazione. In particolare, collezioni di elementi di regolazione pre-caratterizzati quantitativamente, come promotori e siti di legame al ribosoma (RBS), vengono generalmente utilizzati per sovra- o sotto-regolare i processi di trascrizione e traduzione dei geni target, mentre la delezione genica può essere effettuata mediante ricombinazione omologa oppure mediante silenziamento genico, sfruttando tecniche recentemente proposte.

In questo lavoro di tesi verranno investigati diversi metodi in-vivo e in-silico di ingegneria metabolica in *Escherichia coli* e *Bacillus subtilis*.

Nel Capitolo 1 verranno introdotti i concetti chiave dell'ingegneria metabolica. In particolare verranno presentati i principali approcci computazionali e gli strumenti sperimentali utilizzati in questo campo.

Nel Capitolo 2 verrà riportata l'analisi dei metodi *constraint-based* maggiormente utilizzati per lo studio di *E. coli*. Per prima cosa verrà discusso l'impatto dei dati in input, per la modellizzazione delle condizioni ambientali, sui risultati e successivamente verranno analizzate le capacità predittive al variare delle condizioni ambientali e genetiche. Inoltre, l'integrazione di dati trascrittomici verrà implementata con lo scopo di aumentare l'accuratezza dei risultati.

Nel Capitolo 3 verrà descritta la costruzione del modello metabolico basato su vincoli enzimatici del *B. subtilis* e verrà valutata l'accuratezza delle predizioni ottenute per ceppi wild type e mutanti. Inoltre saranno presentate le

manipolazioni target identificate attraverso il nuovo modello per migliorare la produzione di un importante biopolimero.
Nel Capitolo 4 verrà descritta la progettazione, lo sviluppo e la valutazione di *small RNA* per il silenziamento dell'espressione di geni target in *E. coli*. L'efficienza di silenziamento verrà quantitativamente analizzata anche attraverso l'utilizzo di un modello matematico.
Nel Capitolo 5 verrà presentato un nuovo metodo per delezione genica a livello cromosomiale in *E. coli*, basato su saggio colorimetrico XylE e standard *BioBrick*, ed utilizzato per l'eliminazione dei geni responsabili della produzione di acidi organici che competono per l'uso di piruvato in ceppi etanologenici.
Infine, nel Capitolo 6 verranno tratte le conclusioni di questo lavoro di tesi, considerando i miglioramenti ottenuti negli studi descritti, e l'affidabilità dei metodi in-silico e in-vivo investigati, sulla base di alcuni casi di studio ad alto impatto.
Gli studi illustrati nei Capitoli 2, 4 e 5 sono stati svolti presso il Laboratorio di Colture Cellulari del Centro di Tecnologie per la Salute (CHT) e il Laboratorio di Bioinformatica e Biologia Sintetica (BMS), Università degli Studi di Pavia, mentre lo studio riportato nel Capitolo 3 è stato condotto presso Novo Nordisk Foundation Center for Biosustainability (CFB), Technical University of Denmark.

Contents

List of Figures

List of Tables

Chapter 1

Metabolic Engineering

Metabolic engineering is an emergent discipline aimed at direct improving product formation or cellular proprieties of microorganisms through the modification of specific biochemical reaction(s) or the introduction of new one(s) with the use of recombinant DNA technology [1]. The main innovation, respect to the traditional concept of metabolic pathway manipulation, is the rational approach used to identify the genetic modifications required to obtain the desired phenotype.

Over the past few years, due to the limited fossil energy and environmental problems, the need of an economic and ecological production of chemicals, drugs and fuels from renewable resources is increased. Therefore, in order to avoid money- and time-consuming trial-and-error approaches or random mutagenesis techniques, the new approach for the development of cell factories to produce high-value metabolites is based on a preliminary study of the complete biochemical reaction networks, called *genome-scale metabolic models* (GEMs), with the aim to investigate the properties of integrated metabolic pathways. The analysis of GEMs with an appropriate mathematical model leads to the prediction and in-silico optimization of metabolic capacity, expressed in terms of fluxes through each reaction. Despite the variety of in-silico modelling approaches in biology, the GEMs are mainly studied using stoichiometry-based models. These methods have the main advantage of requiring easily available information in large-scale.

Once the optimal genetic configuration is predicted, this is experimentally applied in the target strain through synthetic biology techniques, such as gene deletion, gene addition, gene knockdown, or gene over-expression. Synthetic biology aims to create novel biological functions and systems not found in nature by combining biology with engineering and plays an important role in metabolic engineering field. Indeed, its powerful tools and approaches, based on three engineering principles, namely standardization/modularity, decoupling, and abstraction, are crucial to reduce the time and cost required for the development of cell factories.

In order to support a commercial process, namely the ultimate purpose of this field, the optimization of titer (final concentration in the fermentation medium), rate (production per unit of time) and yield (units of product synthesized per unit of raw material consumed) is essential and represents a current challenge [2]. The construction of a strain that meets these industrial requirements is the last but most intense part of developing a novel bioprocess, involving many years of costly development time, due to the large number of genetic modifications, that typically can only be done in a serial fashion, and subsequent phenotypic characterization [3]. This final process involves several rounds of the so-called *design-build-test* cycle, in which a specific metabolic design is implemented and thereafter tested. The different genetic perturbations are needed not only to enhance the production of the target metabolite, but also to maintain the metabolic equilibrium. In particular, the production of cellular components has to be balance with energy production and consumption in order to ensure the homeostasis even when the microorganism is exposed to varying environmental and nutritional conditions. The use of microorganisms for which the genome was sequenced and the molecular and genetic methodologies for their cultivation and manipulation are well established, for example *Escherichia coli*, *Saccharomyces cerevisiae*, *Bacillus subtilis* and *Corynebacterium glutamicum*, promotes the scale-up to the industrial level. However, other nonstandard microorganisms can be considered on the basis of their ability to grow well in specific environments, use a variety of alternative feedstocks or naturally produce and tolerate amounts of the desired product. Moreover, there is a need for new tools of synthetic biology that can facilitate the genetic manipulations, such as CRISPR/Cas9 systems, that allow the engineering of nearly any host that is transformable [4], the insertion of many genes into many target sites [5], knockout or down-regulation of competing pathways [6], and up-regulation of beneficial pathways, in addition to the use of well-characterized promoters and ribosome binding sites.

In this work, the study and optimization of *E. coli* and *B. subtilis* metabolism were carried out, which are the best-characterized members of the Gram-positive and Gram-negative bacteria, respectively. In the following subsections, an overview of computational methods commonly used to provide targets for metabolic engineering on the basis of genome-scale metabolic models and the main synthetic biology tools applied for its realization are presented.

1.1 Computational strain design methods

The reconstruction of the genome-scale metabolic model is the first essential step for the in-silico study of metabolic behaviour and the design of target cell factories. The metabolic phenotype of an organism, mathematically represented in the respective GEM, can be characterized in terms of either fluxes (metabolite mass per dry organism mass per unit time) through each reaction or network topology. The first approach, based on the concept of flux balance analysis, provides quantitative predictions of metabolic networks by solving optimization problems, while the second one, based on the topology analysis, allows a qualitative description of small networks, subsection of GEM, due to its computational complexity.

In order to facilitate the implementation of the several computational methods used in metabolic engineering field, different software tools with specific characteristics [7] are implemented. These software applications can be classified on the basis of their platform and software dependencies as toolbox-based, such as Metatool [8] and COBRA toolbox [9] based on MATLAB, COBRApy [10] and CAMEO [11] written in Python, stand-alone, such as OptFlux [12] and web-based such as FAME [13].

1.1.1 Genome-scale metabolic models

A GEM contains the complete set of chemical reactions that occurs in the target microorganism, including the biomass reaction, and their relations to the genome and proteome. The reactions are represented as a set of stoichiometric equations and the main cellular compartments are modelled in order to distinguish intracellular from extracellular reactions, which can be reversible or irreversible. As for biomass reaction, it is composed of the key components for growth (i.e., cell wall, proteins, lipids, carbohydrates, DNA and RNA), which are determined on the basis of experimental measurements, and its flux represents the growth rate (μ). Moreover, the relationships of genes, proteins and reactions are described in the model by using Boolean logical rules, called *gene-protein-reaction* (GPR) relationships, and the genetic states are assigned classifying each gene as either “on” (variable equal to 1) or “off” (variable equal to 0). In particular, for the reactions catalyzed by more than one enzyme (or codified from more than one gene) the GPR relationships use the standard operators “and” and “or”. For example, when two or more enzymes are required to catalyze a reaction, as in the case of multi-protein complexes, these are linked with the “and” operator, while if a reaction is catalyzed by any of several enzymes, as in case of isonzymes, these are linked with the “or” operator.

The general work-flow for the development of a high-quality GEM [14] (Fig. 1.1) consists of a first draft reconstruction based on the genome annotation of the considered microorganism and information retrieved from biochemical databases, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) [15] and BioCyc [16]. The draft model is then evaluated against organism-specific data from literature and it is refined. The reconstruction is converted into a mathematical format and the environmental growth conditions are defined. In particular, the m internal metabolites and the n reactions of the model are represented within a stoichiometric matrix S [m x n]. Furthermore, for each reaction, the upper and lower flux bounds [mmol ${g_{DW}}^{-1}$ h^{-1}] can be set in order to include information about the directionality and the environmental condition, and these are included in two different flux vectors (v_{lb} and v_{ub}). Finally, with the aim to ensure an appropriate accuracy of model predictions, a validation step is carried out.

The improvement in genome-annotation technologies and the easy availability of omics data promoted the GEMs reconstruction for both prokaryotic and eukaryotic organisms [18, 19, 20, 21, 22, 23], which are integrated in BIGG (Biochemical Genetic and Genomic) Models database [24]. Moreover, recently several tools to facilitate the GEM reconstruction were proposed, such as Model SEED [25], RAVEN [26], or merlin [27]. The standard format for the representation of GEMs is SBML (Systems Biology Markup Language) [28], which is based on XML format and contains all the information to be directly used in different software tools.

Most up-to-date GEM versions of the same microorganism are usually developed on the basis of the previous model and with the integration of new information, such as reactions, genes or thermodynamic data. For example, five GEM versions of *E. coli* K-12 MG1655 [29, 30, 31, 18, 32] and four of *B. subtilis* 168 [20, 33, 34, 35] are currently available. In order to distinguish between the different models, a naming convention in form of *iXXxxx* was proposed for *E. coli* models and subsequently adopted for other organisms. The first letter *i* of this name stands for in-silico, *XX* is for the initials of the first

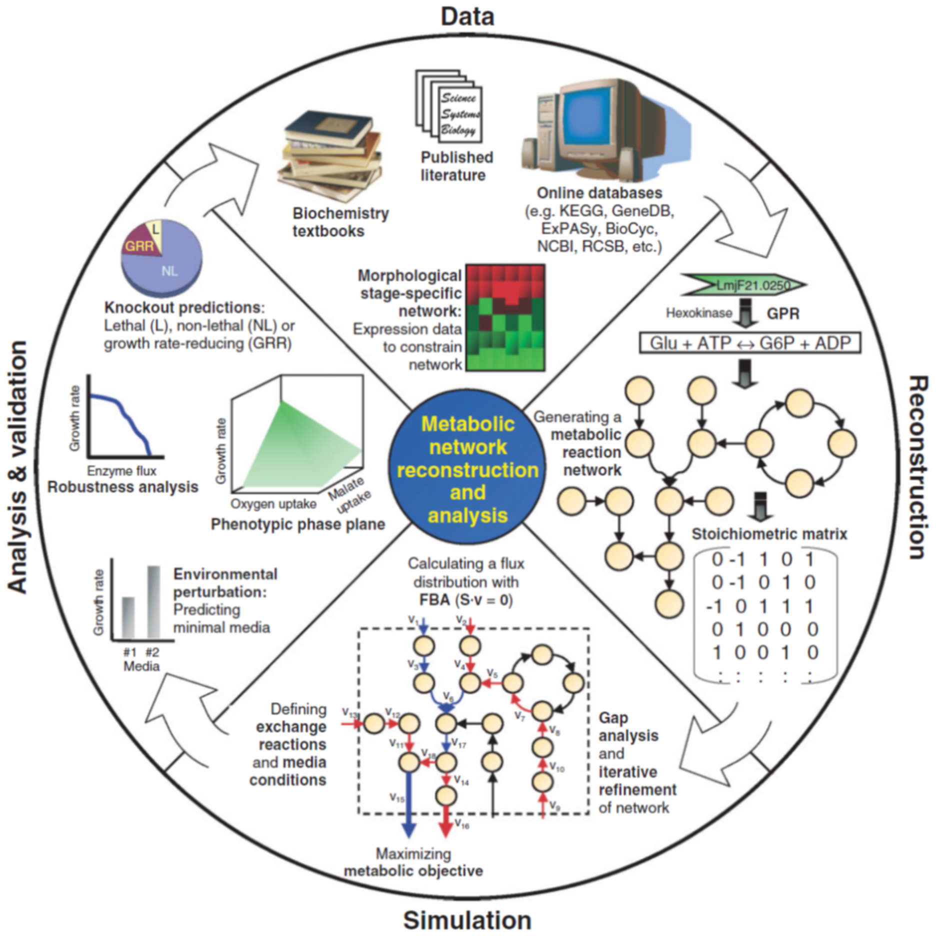

Data
Science
Systems
Biology
Published literature
Biochemistry textbooks
Online databases (e.g. KEGG, GeneDB, ExPASy, BioCyc, NCBI, RCSB, etc.)
Morphological stage-specific network: Expression data to constrain network
LmjF21.0250
Hexokinase
GPR
Glu + ATP ↔ G6P + ADP
Generating a metabolic reaction network
Stoichiometric matrix
0 -1 1 0 1
0 -1 0 1 0
-1 0 1 1 1
0 1 0 0 0
1 0 0 1 0
Reconstruction
Metabolic network reconstruction and analysis
L
GRR
NL
Knockout predictions: Lethal (L), non-lethal (NL) or growth rate-reducing (GRR)
Growth rate
Enzyme flux
Robustness analysis
Growth rate
Oxygen uptake
Malate uptake
Phenotypic phase plane
Analysis & validation
Growth rate
Environmental perturbation: Predicting minimal media
#1 #2
Media
Calculating a flux distribution with FBA (S·v = 0)
Defining exchange reactions and media conditions
Gap analysis and iterative refinement of network
Maximizing metabolic objective
Simulation

Figure 1.1: **The general work-flow for the reconstruction of GEMs.** After that the list of biochemical reactions and of their associated genes are defined, on the basis of the genome annotation and curated biochemical databases, the network is converted to mathematical form. The draft reconstruction is then refined through a comparison with experimental data from literature sources. Finally, through the implementation of computational methods, the evaluated model is used to predict the metabolic phenotype of the target microorganism and optimize its metabolic capacity. Reference: [17].

author and *xxx* for the number of genes included in the model. Alternative syntaxes were used, as the insertion of organism initials instead of those of author.

1.1.2 Principles of constraint-based modelling

Stoichiometric models are studied under the steady-state assumption, which corresponds to consider the total amount of consumed internal metabolites equal to the total amount of produced internal metabolites. Therefore, the set of mass balances becomes a system of linear algebric equations:

$$\sum_{j=1}^{n} S_{ij} v_j = 0 \qquad for \;\; i = 1...m \tag{1.1}$$

where S_{ij} is the stoichiometric coefficient of i-th metabolite in the j-th reaction and v_j is the flux of j-th reaction.
This steady-state condition is experimentally encountered in *chemostat* operation, in which the volume of liquid cell cultures is maintained constant by simultaneously adding fresh medium and removing cultured broth. Therefore, the cells remain indefinitely in their exponential growth phase, during which their growth rate is constant and maximum.
The space of allowable solutions obtained using this constraint, also known as the *convex polyhedral cone*, is further restricted by adding thermodynamic constraints to describe the reaction directionality and fixing the maximum and minimum flux through each reaction on the basis of experimental data. In particular, the lower flux bound of irreversible reactions is fixed to 0 mmol ${g_{DW}}^{-1}$ h^{-1}, while the environmental condition is modelled imposing the uptake rate to 0 mmol ${g_{DW}}^{-1}$ h^{-1} for the exchange reactions of nutrients that are not available in the medium and to 100 mmol ${g_{DW}}^{-1}$ h^{-1} for the unlimited nutrients or, when known, to experimental values:

$$v_j \geq 0 \qquad if \;\; j \in irrev \tag{1.2}$$

$$v_{lb_j} < v_j < v_{ub_j} \qquad for \;\; j = 1...n \tag{1.3}$$

The approach based on these physico-chemical constraints is so-called *COnstraint-Based Reconstruction and Analysis* (COBRA) and is frequently used in the field of microbial metabolic engineering to predict the metabolic phenotype using genome-scale models. Over the past decade more of 100 methods were developed on the basis of COBRA approach [36].

Since the number of reactions (n), which represent the unknown variables, is higher than the number of metabolites (m), namely the equations, the obtained equation system (Equations 1.1, 1.2 and 1.3) is often under-determined. Therefore, additional constraints are required in order to obtain an unique steady-state flux distribution:

- **Optimization-based methods**

A large family of COBRA methods is based on the optimization of a biologically relevant objective function (Z), in order to identify the optimal flux distribution, with respect to the constraints previously defined (Equations 1.1, 1.2 and 1.3):

$$max/min\ \ Z = f(\sum_{j=1}^{n}(c_j \cdot v_j)^k) \tag{1.4}$$

where the c_j coefficients are the weights indicating how much each reaction contributes to the objective function. Considering the vector notation, c has 1 at the position of the reaction whose flux must be optimized, while the other values are equal to zero. The k constant represents the objective function degree, for example equal to 1 for linear optimization problems, or 2 for quadratic optimization problems.
The choice of objective function depends on the target study and the desired goal. A variety of methods, based on the constraints of steady-state, reversibility, flux capacity and with a specific optimization problem, were proposed and successfully applied to different organisms.

Flux Balance Analysis
The most popular constraint-based method in metabolic engineering field is *flux balance analysis* (FBA) [37], appropriate for the metabolic behaviour study of wild type organisms or mutants that have evolved over a large number of generations, for example by *Adaptive Laboratory Evolution* (ALE). FBA uses a linear programming (LP) to find the flux distribution that maximizes the objective function (Z), commonly defined as the flux through the biomass reaction ($v_{biomass}$), i.e., the growth rate (μ), on the basis of evolutionary pressure concept:

$$max\ \ Z = v_{biomass} \tag{1.5}$$

However, different objective function can be considered in function of target study, such as the maximization of ATP production.
Over the past few years, variants of FBA method were developed, such as *parsimonious enzyme usage* FBA (pFBA) [38]. In particular, since FBA problems often have non-unique solutions, pFBA method uses a bilevel linear programming optimization to find a flux distribution with the minimal sum of fluxes and maximal growth rate ($v_{biomass}$):

$$max \;\; v_{biomass}, \quad min \; \sum_{j=1}^{n} |v_j| \tag{1.6}$$

The alternative flux distributions computed by FBA can be analyzed in order to test the robustness. In fact, *Flux Variability Analysis* (FVA) [39] method finds the variation range of each flux by solving a pair of LP problems that computes the maximum and minimum fluxes through each reaction for which the objective function remains optimal and with respect for the other defined constraints:

$$max \;\; v_j \quad for \;\; j = 1...n, \quad j \neq Z \tag{1.7}$$

$$min \;\; v_j \quad for \;\; j = 1...n, \quad j \neq Z \tag{1.8}$$

Dynamic Flux Balance Analysis

An extension of FBA method, called *Dynamic Flux Balance Analysis* (dFBA) [40], was developed in 1994 with the aim to study the dynamic metabolic behaviour of the target organism, for example to describe the diauxic growth. dFBA can be implemented using nonlinear programming, for a dynamic optimization problem, or by using linear programming on a series of short time intervals, for a static optimization. In the static optimization approach, the time period is divided into several time intervals and the initial substrate concentration S_{c0} (mmol/L) is used (for t=t_0) to determine the substrate concentration S_c for the successive time interval:

$$S_c = S_{c0} \tag{1.9}$$

A system of the ordinary differential equations (ODEs) is used to describe the evolution of the target system:

$$\frac{dX(t)}{dt} = \mu \cdot X(t) \tag{1.10}$$

$$\frac{dS_c(t)}{dt} = S_u \cdot X(t) \tag{1.11}$$

where X(t) (g_{DW}/L) is the biomass concentration, μ (h^{-1}) the growth rate and S_u (mmol/g_{DW} h) is the vector of substrate uptake.
With the integration of Eq. 1.10 and Eq. 1.11 in a specific time interval the following equations are obtained:

$$X(t) = X(t_0) \cdot e^{\mu \Delta t} \tag{1.12}$$

$$S_c(t) = S_c(t_0) + \frac{S_u}{\mu} X(t_0) \cdot (1 - e^{\mu \Delta t}) \tag{1.13}$$

Finally, with the assumption of steady-state condition for each time interval, the growth, nutrient uptake and by-product secretion rates can be predicted.

Minimization of Metabolic Adjustments and Regulatory On/Off Minimization

Alternative methods to FBA principles were developed in order to study the metabolic phenotype of mutant strains affected by genetic perturbations, in particular gene deletions. In fact, these strains assume a suboptimal flux distribution intermediate to wild type and mutant optimum, due to the inability to immediately adapt their metabolic network to achieve the wild type objective function.
In this regard, the method of *Minimization of Metabolic Adjustment* (MoMA) [41] is based on the constraint-based approach and considers as objective function the minimization of Euclidean distance between the flux distributions in the wild type (v^{WT}) and mutant (v^M) strains, which is mathematically formalized as quadratic programming (QP) problem:

$$min \;\; Z(v^{WT}, v^M) = \sqrt{\sum_{j=1}^{n} (v_j^{WT} - v_j^M)^2} \tag{1.14}$$

Despite the success achieved with MoMA method for the prediction of mutant phenotypes, it is unable to model large modifications in single fluxes. Therefore a new method, known as *Regulatory On/Off Minimization* (ROOM) [42], was proposed few years later. ROOM method implements the minimization of the total number of significant flux changes from the wild type flux distribution, on the basis of the assumption that the adaptation cost, in terms of genetic regulatory changes, after gene deletions is minimized, and it is independent of the magnitude of flux change:

$$min\ \ Z = \sum_{j=1}^{n} (y_j), \qquad y_j \in \{0, 1\} \tag{1.15}$$

where y_j is a binary variable defined for each flux, which is equal to 1 when the respective flux change is significant, namely the difference between the wild type and mutant strains is higher than a user-defined threshold value, and is equal to 0 otherwise.

Since the end goal of metabolic engineering field is the identification of optimal genetic configuration of a target organism to improve the production of the desired byproduct, different methods, to guide the rational design of microbial cell factories, were developed.

OptKnock and OptGene

OptKnock [43] was the first method based on the simultaneous optimization of cellular growth rate and the target chemical production, namely the so-called *growth-coupled design* approach. The bilevel optimization can be reformulated as a single mixed integer linear programming (MILP) problem and its structure consists of an inner problem that identifies the possible flux distribution (usually using FBA method, based on the maximization of biomass yield, but also MoMA or ROOM) and an outer problem that finds the reaction eliminations for maximizing the bioengineering objective, namely the desired product:

$$
\begin{aligned}
\max_{y_j} \quad & v^{M}_{chemical} \\
\text{subject to} \quad & \max_{v_j} \quad v^{M}_{biomass} \\
& \quad \text{subject to} \quad S \cdot v^{M}_{j} = 0 \\
& \qquad\qquad\qquad v^{M}_{biomass} \geq \gamma (v^{WT}_{biomass})_{max} \\
& \qquad\qquad\qquad v_{lb_j} \cdot y_j \leq v_j \leq v_{ub_j} \cdot y_j \qquad for \;\; j = 1...n \\
& y_j \in \{0, 1\} \\
& \sum_{j=1}^{n} (1 - y_j) \leq K
\end{aligned}
\tag{1.16}
$$

where y_j is a binary variable defined for each reaction to indicate its active (equal to 1) or the inactive state (equal to 0) and K represents the maximum number of allowable reaction eliminations. Parameter γ determines the minimum value which the biomass flux can assume, that is the percentage of the maximum biomass yield obtained for wild type. Different extended and improved versions of *OptKnock* were developed to overcome its limitations and improve the results. For example, since for some pathways the coupling between maximum production of the target compound and maximal biomass flux is not possible, a new method, called *RobustKnock* [44], identifies the reaction eliminations for which the minimal production rate is maximized.

OptGene [45] is a further alternative optimization strategy based on genetic algorithms, which was implemented to avoid the high computational cost obtained with MILP formulation, when a large numbers of deletions are accounted. Moreover, OptGene allows the optimization of non-linear objective functions. As for its work-flow, in the first step a population of individuals, which is represented through a specific set of genes, is initialized by assigning a status (present/absent) to each gene randomly. Subsequently, the best individuals are selected on the basis of a fitness score, which is calculated using the desired objective function value (FBA, MoMA or ROOM) and then crossed to produce a new offspring. Finally, the individuals obtained from the crossover are perturbed in terms of gene deletions. The steps of computation of fitness score, crossover and mutation are repeated until an individual with desired phenotype characteristics is found.

Flux Scanning based on Enforced Objective Flux

The main metabolic engineering strategies for the overproduction of a desired metabolite are the deletion and amplification of target genes. As described in the previous section, several methods for identifying gene deletion targets are available. Whereas, a limited number of methods for identifying gene amplification targets have been developed, given the complexity of the metabolic phenotype predictions after this manipulation. A strategy to identify gene amplification targets was implemented in the method called *Flux Scanning based on Enforced Objective Flux* (FSEOF) [46]. In particular, the initial flux distribution ($v_j^{initial}$) was predicted using FBA method and the production flux of target compound ($v_{chemical}^{initial}$) is extracted. Then, the maximal production flux of target compound ($v_{chemical_{max}}$) is computed with FBA but changing the standard objective function. Finally, FBA is again implemented considering the cell growth as objective function and fixing the production flux of target compound at increasing values for each step, from the initial flux value to 90% of the maximum theoretical value. A reaction is selected if its maximum flux assumed during the different steps is higher than the initial value without changing the direction:

$$|v_j|^{max} > \left|v_j^{initial}\right| \quad \text{and} \quad v_j^{max} \times v_j^{min} \geq 0 \tag{1.17}$$

where v_j^{max} and v_j^{min} are the maximum and minimum fluxes of the j-th reaction.

- **Pathway-based methods**

An alternative approach based on the constraint-based modelling (Equations 1.1, 1.2 and 1.3) aims at identifying the topology of cellular metabolism without solving the optimization of an objective function. This type of analysis has been successfully applied for the study of network structure and robustness and for rational strain designs. Since the pathway-based methods fully describe the steady-state solution space, their computational complexity increases with the size of the metabolic network. Therefore, in this context, the analysis of small GEMs or their subsections is preferred.

Despite pathway-based methods were not applied in this work, the key principles and applications will be described to provide a complete description.

$$T^{(0)} = \left[\begin{array}{ccccc:cccc}
\multicolumn{5}{c:}{S^T} & \multicolumn{4}{c}{I} \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & \dots & 0 \\
0 & -1 & 0 & 2 & 0 & 0 & 1 & \dots & 0 \\
-1 & 0 & 0 & 2 & 1 & 0 & 0 & \dots & 0 \\
-2 & 0 & 2 & 1 & -1 & 0 & 0 & \dots & 0 \\ \hdashline
0 & 0 & 0 & -1 & 0 & 0 & 0 & \dots & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & \dots & 1
\end{array}\right]
\begin{array}{l}
\left.\begin{array}{c}\\ \\ \\ \\ \end{array}\right\}\text{reversible reactions} \\
\left.\begin{array}{c}\\ \\ \\ \\ \\ \end{array}\right\}\text{irreversible reactions}
\end{array}$$

Figure 1.2: **Initial tableau used for elementary modes computations.** The transposed stoichiometric matrix (S^T) is augmented with the identity matrix to build the initial tableau.

Elementary Mode Analysis

The most popular method for metabolic pathway analysis is based on the concept of *elementary flux modes* (EMFs) as building blocks of network and has the main purpose to identify the set EMFs, namely metabolic pathways required to obtain the desired phenotype [47]. Therefore, the *elementary mode analysis* is particularly suited for the creation of a minimal cell that is able to self-assemble and to self-replicate and is specialized for the formation of the desired product with a limited number of genes [48].

In order to identify a finite set of solutions, the *non-decomposability* constraint, also called *genetic independence*, is added to the assumption of steady-state and thermodynamic constraints. Under this new constraint each elementary mode is unique and is composed of a minimal set of enzymes (catalyzing reactions) necessary to operate as a functional unit under steady-state condition. “Minimal” indicates that if any enzyme is eliminated, the resulting elementary mode cannot operate as a functional unit. As for the work-flow, in the first step the stoichiometric matrix S, in which irreversible and reversible reactions are distinct, is transposed and combined with an identity matrix, to obtain a matrix called initial tableau (Fig. 1.2). New tableaux are obtained by pairwise linear combination of rows, reversible or irreversible reactions, so that the columns consecutively become null vectors successively and the steady-state assumption is assured. In general, linear combinations of two rows corresponding to the same type of directionality are inserted into the part of the respective type in the new tableau, whereas linear combinations of rows corresponding to different types are inserted into the “irreversible” part. Moreover, a linear combination of “irreversible” rows can be carried out only using a positive coefficient in order that the appropriate reaction direction is considered. Therefore, if the combination of two rows leads nonelementary modes, duplicate modes or flux modes violating the sign restriction for the irreversible reactions, this candidate pair is rejected. In the final tableau all the columns in the left-hand side are null vectors and the rows in the right-hand side represent the elementary modes, namely the irreversible reactions or pathways.

Extreme Pathway Analysis
An alternative approach for structure analysis of a metabolic network that links the cellular phenotype to the corresponding genotype, is based on the identification of *extreme pathways* (ExPas) [49]. ExPas represent a subset of elementary modes, namely only the internal reactions. *Extreme pathway analysis*, in addition to the set of constraints considered for elementary mode analysis (that are the state-state assumption, the thermodynamic and non-decomposability constraints) introduces the principle of *systematic independence*. This new constraint requires that none of the ExPas can be expressed as a non-negative combination of at least two other ExPas.

1.1.3 Integration of omics data into genome-scale metabolic models

The main reason for the popularity of constraint-based methods is the high availability of stoichiometric data required for the reconstruction of GEMs and for the prediction of cellular metabolism in terms of biochemical reaction rates. But on the other hand this simplified modelling strategy, in which the predicted fluxes are strongly based on the constraint of carbon source uptake rate, leads to unrealistic predictions, especially when the environmental and genetic conditions are different from the optimal state. The optimal metabolic phenotype, predicted by FBA method, differs significantly from the experimental behaviour measured under different conditions. More accurate predictions could be computed using experimental measurements of nutrient uptake rates as upper bound constraints, which are rarely available due to the complexity for the high-throughput quantification. Therefore, different approaches for integration of additional biological information [50], such as gene expression and proteomic data and enzyme kinetic parameters, were developed in order to further reduce the solution space and refine predictions both of growth rate and internal fluxes without using measurements of nutrient uptake rates. In particular, transcriptomic and proteomic data can be readily integrated into GEMs using GPR relationships. Thanks to the advent of high-throughput

technologies, a large amount of "omics" data are available for different organisms. However, the data integration usually covers a limited subset of genes and reactions that are included in a GEM.
In the following parts, the methods used in this work are described.

- **Transcriptomic data**

Transcriptomic profiles of different organisms under specific environmental conditions are available in literature. Therefore, several methods to integrate gene expression data into GEMs were developed and evaluated [51].
The techniques usually used to measure the transcripts are DNA microarray [52], which allows to compare thousands of genes at once, and RNA-Seq [53], with clear advantages in terms of amount of sequence coverage. The first approach measures the amount of the mRNA (or cDNA) in terms of fluorescent intensity. Indeed, it is fluorescently labeled and bound to a specific DNA template on the array. Whereas, with RNA-Seq technology the mRNA is used to generate a DNA library, which is analyzed using *next-generation sequencing* (NGS) methods.
Some of these computational methods use transcriptomic data to predict the flux distribution by maximizing the correlation, whereas others reconstruct a new model with only the reactions classified as active and predict the flux distribution based on FBA approach. We evaluated the methods based on FBA, which differ from the rule used to classify active and inactive reactions.
Åkesson and co-workers [54] proposed one of the earliest methods to tailor the GEMs to the specific context by using gene expression data. This approach is based on a very simple rule: a reaction is considered as inactive if the associated gene expression is lower than a specific threshold. More complex rules were then adopted to improve the building of context-specific models. In particular, the method used in this work is presented in the next paragraph.

GIMME

The *Gene Inactivity Moderated by Metabolism and Expression* (GIMME) [55] approach aims to reconstruct context-specific models, in which the reactions with gene expression data below a specific threshold and that are not needed to achieve the presupposed metabolic objective function are removed. Moreover, the flux distribution is predicted in order to minimize the utilization of reactions with gene expression data below the threshold but that are needed for the objective function. The fluxes of these "inactive" reactions are computed by a two-step procedure, in which first the FBA method is used to find the maximum possible flux through the objective function and then under the constraint that the flux of objective function can be equal or higher than a percentage of the maximum value, the following linear programming problem is solved:

$$
\begin{aligned}
&\min_{v_j} && \sum c_j \cdot |v_j| \\
&\text{subject to} && S \cdot v = 0 \\
& && v_{lb_j} \leq v_j \leq v_{ub_j} \\
&\text{where} && c_j = \begin{cases} x_{cutoff} - x_j, & \text{if } \; x_{cutoff} > x_j \\ 0, & \text{otherwise} \end{cases} \\
& && \text{for} \quad j = 1....n
\end{aligned}
\tag{1.18}
$$

- **Proteomic and kinetic data**

Since a finite protein pool with a limited efficiency is used to complete the different cellular processes, the integration of protein or enzyme levels and kinetic parameters are fundamental to model the cellular metabolism.
Proteome analysis is usually obtained through mass spectrometry measures. Despite thousands of proteins can be quantify, thanks to the advancement of omics technologies, the acquisition of good quality data is difficult. Constraints of protein abundance, measured in relative or absolute value, can be integrated into the GEM for each protein or for a protein pool. For what concerns the enzyme kinetic, it is commonly represented by the maximal turnover rate of the enzyme (k_{cat}). k_{cat} values are generally measured through in-vitro enzyme assays and they can be collected from enzyme databases, such as BRENDA [56] and SABIO-RK [57]. However, these values are usually available for a small number of enzymes. This data scarcity leads to include kinetic information only for a small part of the total number of enzymes and limits the improvement of prediction accuracy.
Here we described the two methods implemented and evaluated in this work.

MOMENT

MetabOlic Modeling with ENzyme kineTics (MOMENT) [58] requires turnover numbers and molecular weights for each enzyme and the total mass of proteins respect to the dry weight mass (C) as new inputs to predict the concentrations (g) for each enzyme needed to catalyze the predicted metabolic flux rates:

$$\sum g_i \cdot MW_i \leq C \tag{1.19}$$

The GPR relationships are used to associate the enzyme parameters with the reaction fluxes and specific flux constraints are imposed for isozymes, protein complexes and multifunctional enzymes.
For the j-th reaction catalyzed by the i-th single enzyme:

$$v_j \leq k_{cat_j} \cdot g_i \tag{1.20}$$

For the j-th reaction catalyzed by two isozymes a or b:

$$v_j \leq k_{cat_j} \cdot (g_a + g_b) \tag{1.21}$$

For the j-th reaction catalyzed by an enzyme complex a and b:

$$v_j \leq k_{cat_j} \cdot min(g_a, g_b) \tag{1.22}$$

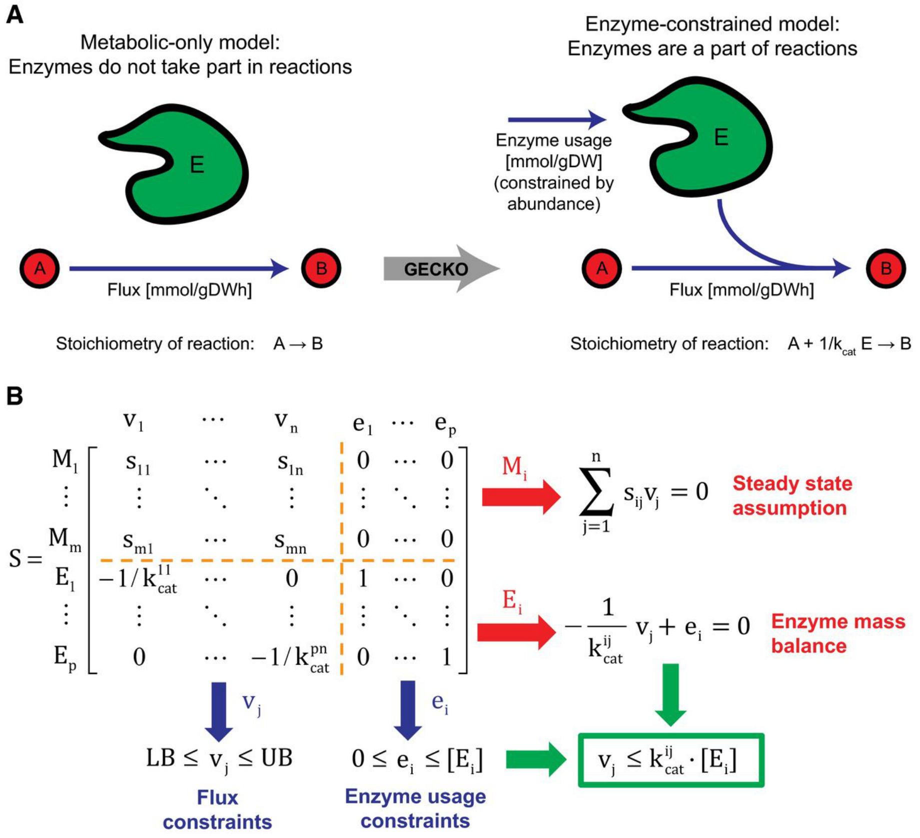
GECKO

GECKO (*Enzymatic Constraints using Kinetic and Omics data*) [59] is the most recent method based on the integration of enzymatic data and it was implemented in this work to improve the predictions of *B. subtilis* flux distribution. The key principle of this approach is the addition of a new constraint into the GEMs to ensure that each metabolic flux (v_j) does not exceed its maximum capacity (v_{max}), corresponding to the product of turnover number (k_{cat}^{ij}) and enzyme's abundance ($[E_i]$) :

$$v_j \leq k_{cat}^{ij} \cdot [E_i] \tag{1.23}$$

This constraint (Eq. 1.23) represents the simplest scenario, but different constraints are considered for more complicated relationships.

Since the enzymes are considered as pseudo-metabolites and their import



$$S = \begin{array}{c|ccc|ccc} & v_1 & \cdots & v_n & e_1 & \cdots & e_p \\ \hline M_1 & s_{11} & \cdots & s_{1n} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ M_m & s_{m1} & \cdots & s_{mn} & 0 & \cdots & 0 \\ \hline E_1 & -1/k_{cat}^{11} & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ E_p & 0 & \cdots & -1/k_{cat}^{pn} & 0 & \cdots & 1 \end{array}$$

$$\sum_{j=1}^{n} s_{ij} v_j = 0$$

$$-\frac{1}{k_{cat}^{ij}} v_j + e_i = 0$$

$$\boxed{v_j \leq k_{cat}^{ij} \cdot [E_i]}$$

Figure 1.3: **Implementation of GECKO approach.** (A) The stoichiometry of reactions reported in GEMs is modified through the inclusion of the enzymes as pseuo-metabolites with a specific stoichiometric coefficient, equal to the inverse of k_{cat} value, and with a limited concentration. (B) k_{cat}^{ij} values are integrated in the stoichiometric matrix (S) by adding new rows that represent the enzymes and new columns that represent each enzyme's usage. After the integration the new stoichiometric matrix consistes of 4 submatrices: the upper left submatrix is equivalent to the original stoichiometric matrix, the upper right submatrix is a matrix of zeros, the lower left submatrix has the kinetic information, and the lower right submatrix is the identity matrix. While the enzyme concentrations are integrated in the vector containing upper bounds on the metabolic fluxes (UB). *Reference: [59].*

as pseudo-reaction (Fig. 1.3A), the kinetic information is readily integrated into the model through the expansion of the stoichiometric matrix (S) and the vector of the upper flux bounds (UB) to include k_{cat} values (expressed as h^{-1}) and the enzyme concentrations ($mmol/g_{DW}$), respectively (Fig. 1.3B).

1.2 Synthetic biology tools

Once the design phase of a cell factory is completed, synthetic biology tools are used to genetically edit metabolism and reroute metabolic flux towards a given native or non-natural metabolic pathway. There are many synergies between metabolic engineering and synthetic biology, and the two fields need one another [2]. Synthetic biology aims to build novel synthetic biological systems to carry out specific user-defined tasks, laying its foundations on key principles from the engineering world, like abstraction, modularity and model-based design. This field can be applied for different goals, such as the high-throughput chemical synthesis of DNA and the construction of genetic control circuits, however, the design and construction or manipulation of metabolic pathways has received the most attention due to its industrial relevance [2].

The fundamental idea is to consider the basic cell components, namely specific DNA sequences, as separate modules and embed them in a more complex system [60]. In order to achieve a physical standardization, introduction of standardized assembly techniques is required. The BioBrick™Standard Assembly [61] is the most popular procedure to construct complex circuits from basic parts via an easy and iterative assembly procedure. The basic parts are DNA sequences that can be classified on the basis of their function and they represent the functional components (modules) of genetic circuits. The main modules are defined through the following four categories.

Promoters

Promoters are DNA sequences located upstream of the coding sequence and are responsible for transcription process. The RNA polymerase complex recognizes and binds the promoter site to transcribe the genetic information of DNA into a new molecule of messenger RNA (mRNA). Each promoter can be quantitatively characterized on the basis of the affinity between its specific sequence and RNA polymerase, which defines its *strength*, namely the rate of transcription initiation. Moreover, a promoter is classified as *inducible* if its activity can be regulated by transcriptional factors and/or chemicals, conversely as *constitutive* if it works with a constant activity.

Ribosome Binding Sites

Ribosome Binding Sites (RBSs) are small RNA sequences, located between the coding sequence and the promoter at 5' untraslated region (UTR), to which ribosomes bind in order to initiate translation. In prokaryotes, the RBS, also called *Shine Dalgarno* sequence, is complementary to the 16S ribosomal RNA (rRNA) and is directly responsible for the efficiency of translation initiation. However, the real translation efficiency of RBSs is not easy to predict since it could affect the stability of the mRNA and is highly gene sequence-dependent. In the last few years many efforts have been carried out to develop computational tools able to predict the activity of RBSs starting from their sequence, such as *RBS Calculator* [62], *UTR Designer* [63] and *RBSDesigner* [64], based on a thermodynamic model of bacterial translation initiation. Although none of the three evaluated tools shows a high prediction accuracy, RBS Calculator performs better than the others.

Coding Sequences

A coding sequence (CDS) is a DNA sequence located downstream of RBS sequence and bounded by a *start* (usually ATG) and *stop* (TAA, TGA ot TAG) sequence. The CDS is first transcribed into mRNA and then is translated by ribosomes to produce proteins. Each amino acid, the building block of the protein, is encoded by a nucleotide triplet, called codon. *Reporter genes* are essential tools for the study of biological systems. They encode proteins that can be readily assayed, for example fluorescent proteins (such as GFP or RFP) or enzymes that can be quantified using specific fluorimetric/colorimetric assays (such as β-galactosidase).

Terminators

Terminators are DNA sequences located at the end of a gene or of a set of genes regulated by a single promoter (i.e. operon) and are responsible for triggering the end of transcription process. In prokaryotes, terminators can be classified as ρ *-independent* or ρ *-dependent* terminators based on their sequence. ρ -independent terminators consist of a G-C rich stem loop, followed by a T-rich sequence, while in the latter a protein called ρ *factor* is required to unbind the RNA polymerase from the DNA fragment which is transcribing.

These BioBrick™parts are properly incorporated in a special circular and double-stranded DNA molecule, called *plasmid*, with standardized features

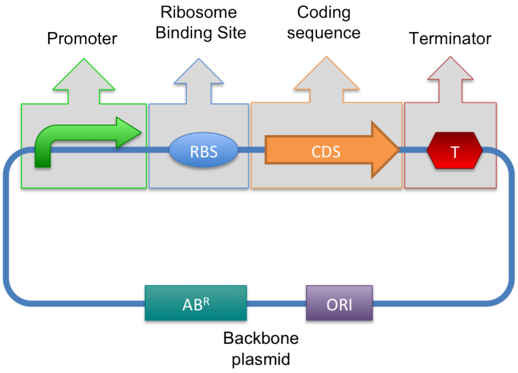


Figure 1.4: **Synthetic genetic circuit architecture.** The simplest genetic circuit is composed by a promoter (green), a ribosome binding site (blue), a coding sequence (orange) and a terminator (red). These components are inserted in a plasmid backbone with selection marker (ABR, antibiotic resistance) and an origin of replication (ORI), which ensure the plasmid maintenance and propagation in the host cell.

(Fig. 1.4). Indeed, in order to simplify the assembly procedure, the modules are put in the *cloning site*, in a special plasmid backbone, that is flanked by a *prefix* sequence upstream and a *suffix* sequence downstream. The prefix is composed by *EcoRI* (E) and *XbaI* (X) restriction sites, while the suffix includes the *SpeI* (S) and *PstI* (P) sites. Each of these restriction sites must be unique in the plasmid. Moreover, an antibiotic resistance gene, used as a selection marker, and a replication origin, which determines the number of copies of the plasmid per cell, are necessary features in plasmids.

BioBrick parts are collected in an open access repository, called *Registry of Standard Biological Parts* [65], founded in 2003 by the *Massachusetts Institute of Technology* (MIT).

Chapter 2

Evaluation of constraint-based methods in *Escherichia coli*

In-silico design based on genome-scale metabolic models is the first step of the rational metabolic engineering. Constraint-based methods provide a quantitative description of cellular metabolism of wild type or mutant strains, in terms of flux distributions, by optimization of an objective function, using the steady-state assumption and limiting the variation range of each flux. Their accuracy depends on the flux constraints, through which physio-chemically and biologically infeasible results are eliminated.

In this chapter, the prediction performance of constraint-based methods for *E. coli*, under different environmental and genetic conditions, will be described. An introduction on the state of the art of constraint-based methods and the alternative approaches proposed to overcome their limits will be reported (Sec. 2.1). The experimental datasets, collected from the literature and measured in this work, will be presented and the implementation of computational methods and parameters used for the accuracy evaluation will be explained (Sec. 2.2). The impact of realistic flux constraints, derived from experimental data, on the predictions (Sec. 2.3.1) and the ability to predict the phenotype of metabolically optimized strains, using both standard methods and the transcriptomic data integration, will be shown (Sec. 2.3.2). Based on the results obtained in this study, the quantitative prediction performance of *E. coli* models, as a function of the available biological knowledge, will be discussed in Sec. 2.4.

2.1 Introduction

Within the past 25 years, different computational methods have been developed and applied to study the metabolism of organisms, mainly prokaryotes, and to guide their engineering as cell factories for the chemical of interest. All the available information useful to describe the organism metabolism is integrated into GEMs. Their reconstruction is primarily based on the genome annotation and, thanks to advances in sequencing technologies, GEMs for a wide variety of microorganisms are available. In particular, *E. coli* is the most suitable organism to be used for metabolic engineering purpose, because of its ability to grow in different environmental conditions and to be easily manipulate in the laboratory. In the last few years, several updated versions of *E. coli* GEM were reconstructed [29, 30, 31, 18] and applied for different in-silico studies [46, 66].

GEMs are commonly modelled by *constraint-based* approaches and, as described in Chapter 1, FBA is the most used method. It identifies the flux distribution through all the reactions reported into the GEM, such that the growth rate is maximized [37]. Extensions of FBA (pFBA [38], dFBA [40] and FVA [39]) and alternative approaches specific for the simulation of mutant strains (MoMA [41] and ROOM [42]) were subsequently proposed.

Although the traditional constraint-based approaches are widely used to capture the genotype-phenotype relationship, they suffer from some intrinsic limitations. Generally, they showed a low accuracy, especially in terms of intracellular flux distributions, to simulate a metabolic model with perturbed genetic and environmental conditions [67, 51], both due to the limited information that can be used for their modelling and the unrealistic objective function in some contexts, based on an optimal-yield metabolism. Indeed, the genetic deletions are modelled just by setting the flux of respective encoded reaction(s) to zero and the specific environmental context constraining the fluxes of nutrient uptakes. Therefore, new approaches for integration of additional biological information, such as gene expression and proteomic data, were developed in order to improve the prediction performance of an organism metabolism under different substrate conditions and genetic perturbations [68, 58, 59].

In particular, the advancements in high-throughput sequencing methods promoted the development of new methods for integration of transcriptomic data into GEMs, based on FBA principles. *Åkesson et al.* [54] proposed the earliest and simplest approach for the addition of transcriptomic data, that considers only the reactions with relative gene expression higher than a specified thresh-

old. Subsequently, other variations of the former method were developed with a more complex rule for the classification of inactive reactions. One approach, GIMME [55] relies on the usage minimization of low-expression reactions while keeping the growth rate, namely the objective function, above a certain value. Alternative methods are not based on FBA, but use gene expression data to infer metabolic fluxes, such as Lee-12 [69] and EXAMO [70].
Given the recent development of constraint-based approaches and their dependence on the simulation context, a limited number of evaluation studies, changing the genetic and environmental conditions, are available. In this work, we tested the accuracy of constraint-based methods to predict the metabolic flux distribution of *E. coli* wild type and mutant strains grown in minimal medium with different elements, such as the carbon source and culture mode. In particular, a sensitivity analysis of predictions respect to the uptake rate constraints of key nutrients was presented to evaluate how the accuracy can increase when experimental values are known. Furthermore, both pFBA and MoMA were applied for the simulation of strains engineered to improve the production of the target high-value metabolite, namely ethanol, pyruvate and acetate and the accuracy predictions were analyzed. Finally, the former simulations were repeated with the integration of transcriptomic data into *E. coli* GEM by GIMME method, to test if an improvement of prediction can be achieve.

2.2 Materials and Methods

2.2.1 Experimental datasets

We retrieved the experimental fluxes (single value for each reaction, expressed as mmol/g_{DW}h) measured for *E. coli* wild type and knockout strains grown in minimal media with different carbon sources, culture modes (batch or chemostat) and oxygen conditions, from seven studies found in literature, that are briefly described below.
Edwards et al., 2001 [71]: growth experiments of *E. coli* MG1655 in aerobic batch conditions and using succinate or acetate minimal M9 media were carried out, changing both the carbon source concentration (0.05-4 g/L) and the temperature (27.5-37 °C). Measurements of growth rates, uptake rates of carbon source and oxygen during the growth exponential phase were reported for the two conditions.
Causey et al. 2003 [72]: fermentations in glucose-minimal medium and

micro-aerobic conditions of *E. coli* W3110 wild type and three mutant strains, developed with the aim to improve the acetate yields, were conducted. Measurements of specific growth rates, maximum specific glucose utilization rates and maximum specific acetate production rates were reported.
Causey et al. 2004 [73]: fermentations in glucose-minimal medium and micro-aerobic conditions of *E. coli* W3110 wild type and four mutant strains, developed with the aim to improve the pyruvate yield, were conducted. Measurements of maximum growth rates, maximum specific glucose utilization rates and maximum specific pyruvate production rates were reported.
Kayser et al., 2005 [74]: growth experiments of *E. coli* K-12 strain TG1 in aerobic glucose-limited continuous cultures were carried out at dilution rates ranging from 0.044 to 0.415 h^{-1}. The glucose uptake rate, carbon dioxide evolution rate, oxygen uptake rate, acetate formation rate and ammonium uptake rate were measured during steady-state growth at the various dilution rates.
Ishii et al., 2007 [75]: aerobic growth experiments of *E. coli* K-12 strain BW25113 wild type and 23 single-gene knockout mutants in glucose-limited chemostat cultures were carried out. Production rate of lactate, acetate, succinate, pyruvate and formate were determined by the measurements of respective metabolite concentrations and the internal fluxes through the central carbon reactions were measured by ^{13}C-labeling experiments for each condition.
Kim et al., 2007 [76]: specific growth rates under aerobic and anaerobic conditions and in glucose minimal medium were reported for *E. coli* K-12 strain W3110 wild type and three mutant strains, developed with the aim to improve the ethanol yield.
Orencio-Trejo et al., 2008 [77]: *E. coli* C was grown under anaerobic batch conditions, in M9 minimal media with glucose. Measurements of glucose uptake rate and production rates of acetate, formate, succinate and ethanol during the exponential phase were reported.
In addition, we measured the growth rate for three mutant strains of *E. coli* W, developed by allelic replacement vector (see Cap. 5), with the aim to improve the ethanol production from lactose, growing in aerobic and anaerobic batch culture with glucose M9 media.

2.2.2 Transcriptomic data

We collected five transcriptomic datasets of *E. coli*, four of these measured in anaerobic conditions and one in micro-aerobic conditions, that are described below. They were properly analyzed and processed for the integration into the

GEM.
Covert et al., 2004 [78]: expression data of 4202 genes were measured, at least in triplicate, by microarray in anaerobic glucose minimal medium conditions.
Park et al., 2013 [79]: expression data of 4150 genes were measured, in triplicate, under anaerobic conditions.
Bordbar et al., 2014 [80]: expression data of 4295 genes were measured, in triplicate, by RNA-seq analysis in anaerobic glucose minimal medium conditions.
von Wulffen et al., 2016 [81]: expression data of 3539 genes were measured, at least in triplicate, by RNA-seq in anaerobic batch cultures.
Singh et al., 2010 [82]: a single measure of expression data were reported for 4344 genes under micro-aerobic conditions.
For the datasets with two or more replicates, we evaluated their degree of agreement, through the coefficients of variation, to test the reliability of measures and the mean value was computed and used for the integration into GEM. The mean gene expressions among the replicates were compared for each dataset measured under the same conditions, by normalizing with the median expression of measured genes, and the correlation was assessed. Moreover, the over-expressed genes, namely with highest fold-change, respect to the data measured in aerobic conditions, were identified and verified with the experimental evidence.
For each dataset, the expression levels of each reaction were determined by mapping the data of each associated gene using the gene-protein-reaction (GPR) association rules reported in the GEM. In particular, for the reactions catalyzed by enzyme complexes (*and* operator) the relative expression level was set equal to the minimum value of associated genes, wheres for reactions catalyzed by isozymes (*or* operator) the relative expression level was set equal to the sum of the values of associated genes.

2.2.3 Simulations

All the simulations were based on iJO1366 GEM [18] and performed using the available implementation in the COBRA Toolbox [9] using MATLAB R2012a. In general, the different growth conditions were simulated setting, when required, appropriate values for uptake of the key nutrients, namely for oxygen 0 mmol/g_{DW}h, 100 mmol/g_{DW}h, 5 mmol/g_{DW}h when anaerobic, aerobic and micro-aerobic conditions were considered, whereas glucose uptake rate was constrained to the experimental value or, if it was not available, to the values reported by Varma and Palsson [83] both for anaerobic (18.5 mmol/g_{DW}h) and aerobic (10.5 mmol/g_{DW}h) batch conditions and continuous cultures in function of dilution rate. Instead for mutant strains, the glucose uptake rate was imposed equal to the specific experimental value or to the value measured for wild type, with the assumption that the implemented deletions did not lead relevant changes.

The metabolic flux distributions of wild type were simulated by pFBA method [38], whereas for the mutant strains, both pFBA and MoMA [41] were used. For simulation of gene deletions, the respective gene(s) or reaction(s) was blocked prior to simulation.

In order to understand the influence of imposed nutrient uptake rates on the prediction accuracy of the constraint-based methods, some simulations were repeated setting the available experimental measure for glucose, oxygen and ammonium salt, one at a time. These results were compared with predictions obtained with standard uptake rate (0 or 100 mmol/g_{DW}h for oxygen, 100 mmol/g_{DW}h for ammonium salt and 18.5, 10.5 mmol/g_{DW}h or the value as a function of dilution rate for glucose), namely unlimited for oxygen and ammonium, and fixed to previously measured values for carbon source.

The GIMME implementation in the COBRA Toolbox was used for the integration of transcriptomic data. For the mutant strains, we assumed that expression levels are unchanged respect to wild type for all genes except for the gene knockout(s), imposed equal to zero. A simulation using each of the four transcriptomic datasets was carried out, in which the gene expression cutoff value was imposed equal to the 90th quantile of each given dataset, whereas the minimum fraction of growth rate was set to 90% (default value) of the maximum growth rate.

2.2.4 Evaluation of prediction accuracy

The prediction accuracy for each simulation was evaluated by the comparison with available experimental fluxes, usually growth rate, secretion rates of main products, nutrient uptake rates and internal fluxes of central carbon reactions. The prediction error was computed for each simulation of a strain grown under a specific condition by the normalized Euclidean distance between the experimental and the respective predicted fluxes:

$$PRED\ ERROR = \frac{\|exp\ \ flux - pred\ \ flux\|}{\|exp\ \ flux\|} \tag{2.1}$$

2.3 Results

In order to evaluate the predictive capability of constraint-based methods, under different experimental conditions, we simulated *E. coli* grown anaerobically or aerobically in minimal medium with a different carbon source and different culture mode, and the obtained results were compared with experimental datasets, taken from the literature. This work aims to investigate if the modelling of study-specific growth conditions, in terms of uptake rates, leads to increased prediction accuracy, or the use of literature values already provides a reasonable prediction accuracy. We tested such features using pFBA and MoMA on different increasingly complex studies of metabolic engineering applications. Finally, we integrated transcriptomic data as an attempt to improve prediction accuracy.

2.3.1 Impact of nutrient uptake rate constraints

We analyzed the differences between the predictions of *E. coli* metabolic flux distributions, under different environmental and genetic conditions, obtained using or not experimental measures to constraint the nutrient uptake rates. The key nutrients considered in this work are glucose (GLC) and ammonium salt (NH_4), the only source of carbon and nitrogen in a minimal medium, respectively, and oxygen (O_2), responsible for the activation of fermentative or respiratory pathways.
As for the evaluation of oxygen and ammonium uptake rate impact, datasets published by Kayser et al. [74], Ishii et al. [75] and Edwards et al. [71] were used. For both the metabolites, two simulations were implemented fixing the respective uptake rate to experimental or standard value (100 mmol/g_{DW}h, i.e., unlimited), and the experimentally measured glucose uptake rate. The error distributions, obtained using the two different constraints, (Fig. 2.1A and 2.1C) are very similar. In particular, the median values of prediction errors are equal to 0.22 and 0.26, changing the constraint of oxygen uptake rate, and equal to 0.18 and 0.11, changing the constraint of ammonium uptake rate. From the scatter plots of the experimental and predicted uptake rate values for each condition, we observed a very high correlation (see regression lines in Fig. 2.1B and 2.1D; correlation coefficient of 0.99 and 0.79, respectively). However, bisector lines in both panels show a low-entity offset (Fig. 2.1B) and slope discrepancy (Fig. 2.1D) compared to regression lines, demonstrating that the model is able to accurately capture the experimentally observed variation of

uptake values, though with small systematic errors in both oxygen and ammonium, as highlighted by the deviation of data points from bisector. This analysis shows the low impact on model predictions using specific measures of oxygen or ammonium uptake rates, demonstrating the ability of the method, in most cases, to properly set these values only with the knowledge of the constraint of carbon source uptake rate.

Despite unlimited values can be used as maximum rate of oxygen and ammonia uptake, the uptake rate of carbon source, which is chosen from the method in order to optimize the growth rate, must be constraint.

We evaluated if the values of glucose utilization rates indicated by Varma and Palsson [83], for aerobic and anaerobic (10.5 and 18.5 mmol/g_{DW}h) batch conditions and continuous cultures at different dilution rates, are sufficiently reliable, or if the predictions can be improved using their experimental measure. For this analysis, datasets published by Kayser et al. [74], Ishii et al. [75] and Orencio-Trejo et al. [77] were considered. The prediction errors for each simulated growth condition and with or without context-specific experimental glucose uptake were reported in Figure 2.2A. We obtained a small decrease of median error (from 0.25 to 0.15) constraining the glucose uptake to the experimental value, however a good accuracy with the values found in literature can be reached. Indeed, the correlation between the specific experimental glucose uptake rates and the collected values from analogous studies shows that the latter can be used as a reference (Fig. 2.2B).

Moreover, the evaluation of glucose uptake rate impact was tested also in mutant strain predictions. The metabolic flux distributions of 23 single-gene knockout strains grown in glucose-limited chemostat cultures, reported by Ishii et al. [75], were simulated using as constraint for maximum glucose uptake rate the specific measure for each mutant or the value measured for wild type. The prediction errors for the two simulations were computed and the distributions are shown in Figure 2.3. Differently from the results for wild type, the predictions for mutant strains do not improve when experimental values are used for setting the glucose uptake rate. Indeed, the obtained median errors are analogous and equal to 0.3. We observed that the glucose uptake rates specific for each mutant are included in a range between 2.7 and 4.5 mmol/g_{DW}h, and they are not very different from the value measured for wild type, equal to 2.93 mmol/g_{DW}h, grown at the same dilution rate.

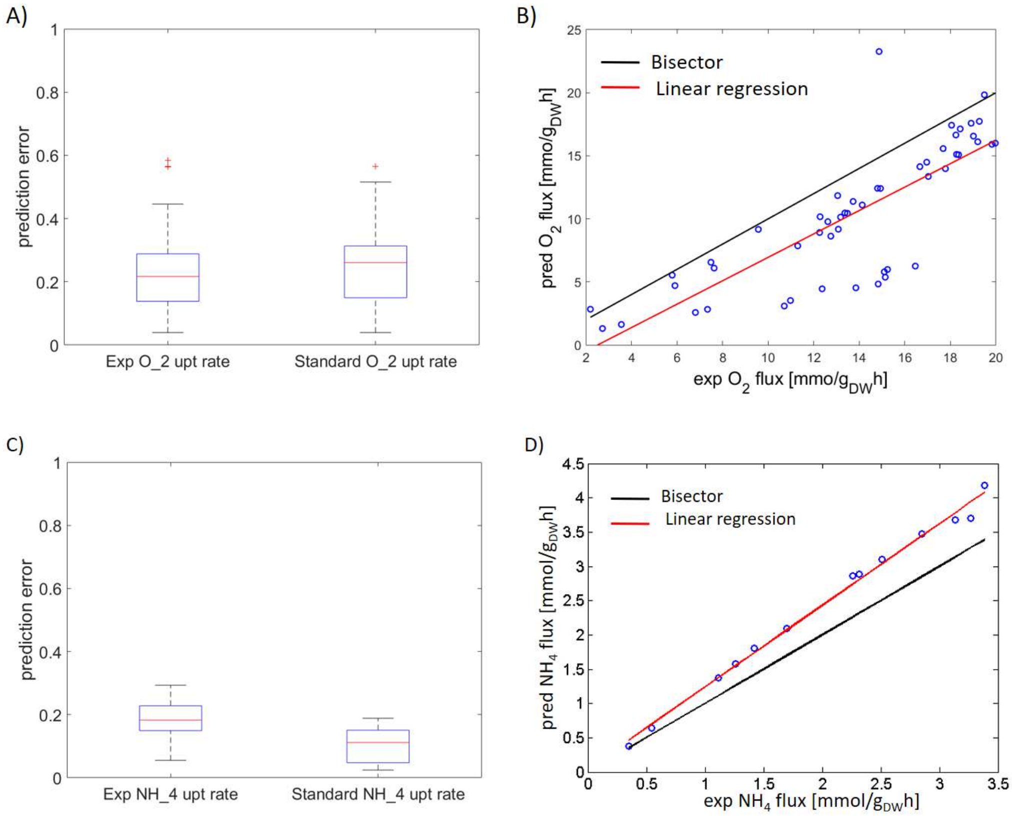


Figure 2.1: **Impact of oxygen and ammonium salt uptake rates on predictions of wild type *E. coli*.** Distributions of prediction errors for each experimental condition constraining the oxygen A) or ammonium C) uptake rate to experimental value and standard value reported in iJO1366 model (100 mmol/g_{DW}h, i.e., unlimited). Correlation analysis between experimental and predicted uptake rates of oxygen B) and ammonium D) through linear regression and bisector. For the analysis of oxygen uptake, the growth in aerobic batch or continuous conditions and with a different carbon source were considered. For the analysis of ammonium uptake, the growth in aerobic glucose-limited continuous cultures was considered.

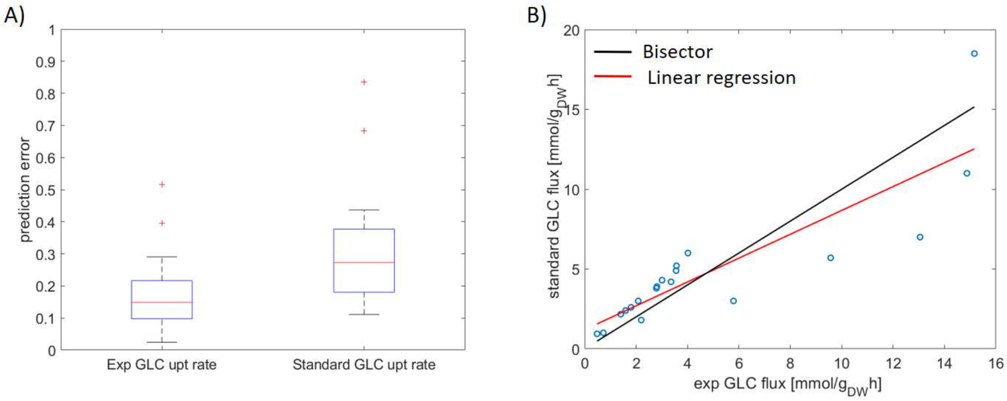


Figure 2.2: **Impact of glucose uptake rate on predictions of wild type *E. coli*.** A) Distributions of prediction errors for each experimental condition constraining the glucose uptake rate to specific measures and standard values reported in literature [83] for analogous growth conditions. B) Correlation analysis between experimental uptake rates of glucose and the respective value reported in literature through linear regression and bisector. The aerobic or anaerobic growth in batch or continuous conditions were considered.

2.3.2 Metabolic engineering applications

So far, different successful results were obtained by modifying microorganisms for a sustainable production of high-value metabolites. The constraint-based methods are generally used to accelerate the design process of cell factories.
Ethanol is one of most widely used metabolites as renewable and sustainable energy source. Numerous research groups have focused on the development of new ethanologenic *E. coli* strains for fermentation of different substrates [84, 85, 86, 76, 48]. In this context, we tested the ability of constraint-based methods, pFBA and MoMA, to predict the growth rate of four different mutant strains, WL, WP, WLF and WLP (see Tab. 2.1), developed with the aim to improve the ethanol yield. In particular, the growth rates under aerobic and anaerobic conditions in glucose minimal medium reported by Kim et al. [76] and measured in this work were considered. Comparing the experimental growth rates, we observed a different growth phenotype, under anaerobic conditions, when *pfl*B and *ldh*A genes were simultaneously deleted. Indeed, whereas Kim et al. [76] reported no growth for WLP strain, according to the elementary mode analysis presented by Trinh et al. [48], because the synthesis of acetyl-CoA, required for biomass synthesis, is blocked, we measured a

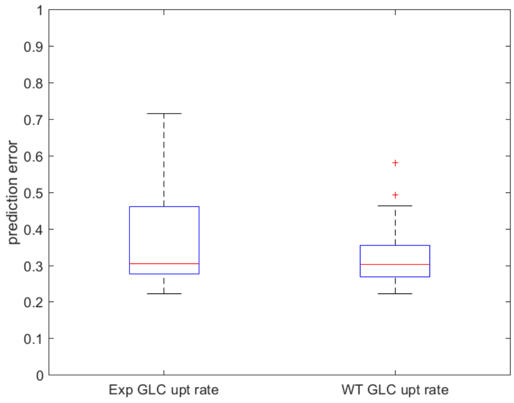


Figure 2.3: **Impact of glucose uptake rate on predictions of *E. coli* single-gene knockout mutants.** Distributions of prediction errors for each experimental condition constraining the glucose uptake rate to specific measures and using the value of wild type. The aerobic growth in continuous cultures was considered.

Strain	Genotype	Ref.
WL	Δ*ldh*A	[76], This work
WP	Δ(*pfl*B-*foc*A)	[76]
WLP	Δ*ldh*A Δ(*pfl*B-*foc*A)	[76], This work
WLF	Δ*ldh*A Δ*frd*AB	This work
SZ47	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A	[72]
TC24	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A Δ*atp*(FH)	[72]
TC36	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A Δ*atp*(FH) Δ*adh*E Δ*suc*A	[72]
TC38	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A Δ*atp*(FH) Δ*adh*E Δ*suc*A, Δ*ack*A	[73]
TC42	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A Δ*atp*(FH) Δ*adh*E Δ*suc*A, Δ*pox*B	[73]
TC44	Δ(*foc*A-*pfl*B) Δ*frd*BC Δ*ldh*A, Δ*atp*(FH) Δ*adh*E Δ*suc*A Δ*ack*A Δ*pox*B	[73]

Table 2.1: **Strains of *E. coli* simulated in this work.**

growth rate different to zero although it was lower than the other mutants. The experimental and predicted growth rates obtained by pFBA and MoMA for the different strains under aerobic and anaerobic conditions are represented in Figure 2.4. A good prediction accuracy was obtained by the both methods, except for WP and WLP strains grown under anaerobic conditions. As noted through the experimental measurements, the deletion of *pfl*B gene, encoding pyruvate formate-lyase, leads to a controversial situation in the absence of oxygen. Considering WP, with the single deletion of *pfl*B, pFBA method computed a growth rate equal to 0.38 h^{-1}, that is lower than the predicted value for WT and WL strains (0.47 h^{-1} for both of us) but higher than the experimental measure (0.17 h^{-1}) reported by Kim et. al, instead MoMA computed no growth phenotype. When the deletion of *ldh*A gene, encoding lactate dehydrogenase, is carried out in addition to *pfl*B gene (WLP strain), pFBA predicted an unchanged behaviour respect to WP strain, namely a growth rate of 0.38 h^{-1}, in agreement with our measure (0.13 h^{-1}), whereas, also in this case no growth was computed by MoMA method due to the *pfl*B deletion, as reported by Kim et al.

For a more comprehensive evaluation of computational constraint-based methods, alternative fluxes in addition to growth rate should be considered. We retrieved two experimental works [72, 73] in which six mutant strains: SZ47, TC24, TC36, TC38, TC42 and TC44 (Tab. 2.1), were developed to improve the production of acetate (the first three mutants) and pyruvate (the last three mutants), widely used as food additive, from glucose. The growth rate and secretion flux of the respective target metabolite were predicted for each strain grown micro-aerobically and compared to the experimental values. In these cases, the micro-aerobic condition introduces a new variable that can

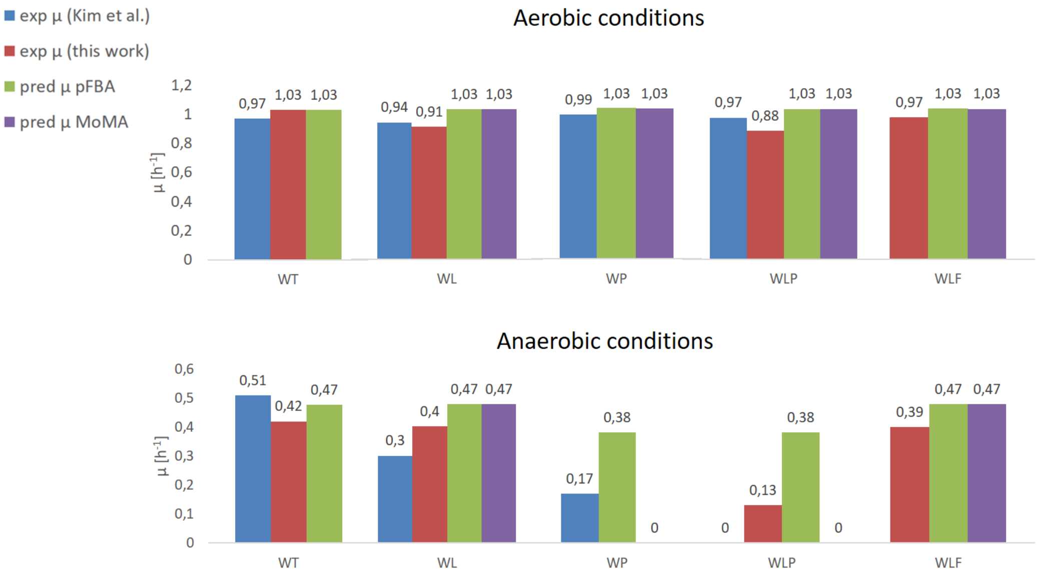


Figure 2.4: **Experimental and predicted growth rates for *E. coli* wild type and mutant strains engineered to improve the ethanol production.** The experimental values measured in this work and by Kim et al. [76] and the predicted values by pFBA and MoMA methods under aerobic and anaerobic glucose minimal medium conditions are represented. We measured the growth rate for each strain except for WP, in both conditions. The genotype characteristics for each mutant strain are reported in Table 2.1.

not be specifically controlled because the relative value of oxygen uptake rate is undefined. In order to consider a good compromise between aerobic and anaerobic conditions, we chose to limit the oxygen uptake to 5 mmol/g_{DW}h, for micro-aerobic growth, as previously carried out by Zsolt et al. [87].

The results confirmed that pFBA is able to properly predict the growth rate also for mutants with multiple gene deletions, and we can observe that, due to the inactivation of PFL reaction, MoMA predicted no growth for these strains also under micro-aerobic conditions (Fig. 2.5A). As for the predictions of internal or external fluxes, obtaining a good accuracy is more complicated respect to the growth rate. The experimental data indicate that, the inactivation of oxidative phosphorylation (ΔatpFH) and cyclic function of the tricarboxylic acid pathway (ΔsucA), in addition to native fermentation pathways, leads to an increase of acetate production. However, very similar acetate yields are computationally predicted for all strains (Fig. 2.5B). The pyruvate secretion fluxes predicted by simulating three mutants (TC38, TC42 and TC44) with larger number of gene deletions are not in agreement with the experimental values. Indeed, the computed rates of pyruvate secretion are equal or close to zero for each engineered strain, and the positive impact, shown through the experimental data, of the perturbation that eliminates acetate production (ΔackA) was not obtained.

As expected, the predictions of acetate and pyruvate secretion rates of the engineered strains change in function of oxygen uptake rate. We observed that, a predicted acetate secretion rate in agreement with the experimental value is achieved when the oxygen uptake rate increased from 5 to 7 mmol/g_{DW}h for SZ47 strain, and to 17.37 mmol/g_{DW}h for TC24 and TC36. Whereas for the pyruvate secretion rate, the predicted value increases only in TC44 strain but for an oxygen uptake close to aerobic condition (16-20 mmol/g_{DW}h).

Through the in-silico studies, the elimination of some reactions may lead to the activation of others, known as inactive, in order to optimize the growth rate (objective function). For this reason, we tried to prevent an uncorrected configuration of active reactions with the integration of omics data.

GECKO was recently proposed by Sánchez et al. [59] as a powerful tool for improving the predictive performance of GEMs by the integration of enzymatic data, in terms of turnover numbers and protein abundance. But since proteomic datasets for *E. coli* grown anaerobically or micro-aerobically in minimal medium with glucose are not available in literature, we cannot apply GECKO tool for improving the prediction of engineered strains reported in Table 2.1. However a context-specific model was created by integrating transcriptomic

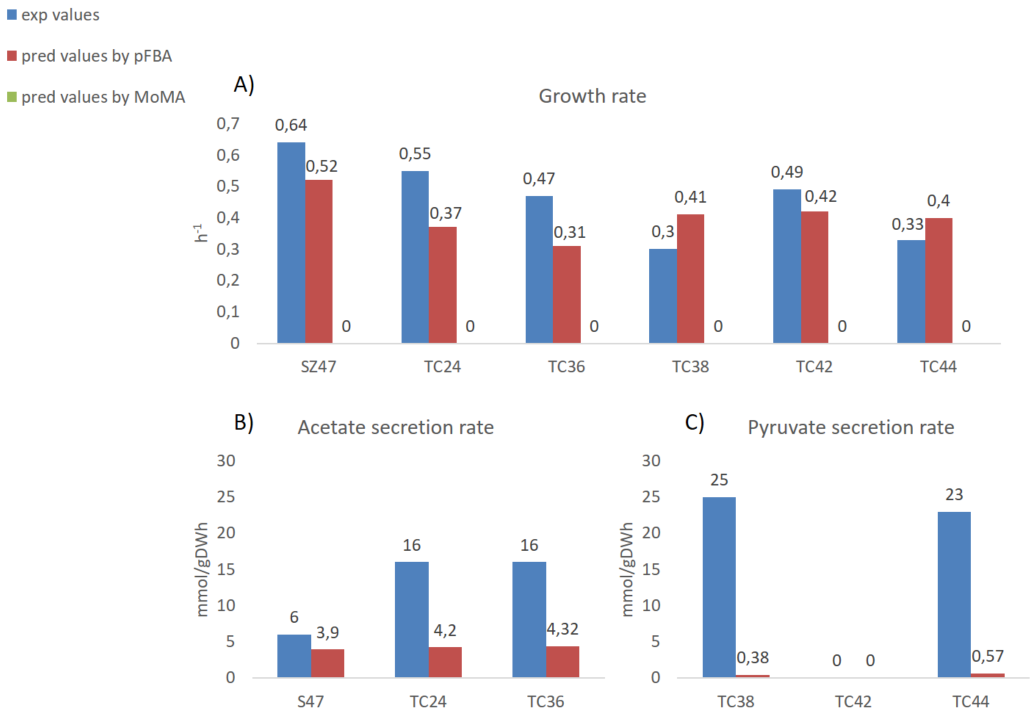


Figure 2.5: **Experimental and predicted values of growth rate and secretion rate of target metabolite for the *E. coli* mutant strains engineered to improve the acetate and pyruvate production.** The experimental values reported by Causey et al. [72, 73] and the predictions obtained by pFBA are reported. The predictions obtained by MoMA are reported only for the growth rate, equal to zero. The growth under micro-aerobic conditions in glucose minimal medium was considered. For the simulation of micro-aerobic conditions the oxygen uptake rate was constraint to 5 mmol/g_{DW}h. The genotype characteristics for each mutant strain are reported in Table 2.1.

data. From the evaluation of methods for integration of transcriptomic data, published by Machado and Herrgard [51], we selected GIMME [55], among the approaches based on FBA, because it combines widely available gene expression data with presupposed cellular function to predict the subset of active reactions under particular conditions.

Once the robustness of transcriptomic datasets, used as input, was verified, the predicted results obtained with GIMME were compared with the experimental values. We observed that, unchanged results were obtained for each of simulated mutant strains, compared to the implementation of pFBA in GEM (data not shown). This inefficiency of transcriptomic data on the prediction accuracy can be due to their inability to model the modification of expression at post-transcriptional and post-translational level. Indeed, we observed that PDH reaction is predicted active both under anaerobic and micro-aerobic conditions, despite its enzymatic complex, in the absence of oxygen, is not sufficiently active to support the flux.

2.4 Conclusion

In this work the performance of constraint-based methods, under different genetic and environmental conditions, were investigated.

First, we provided a general description, useful for the user, about the input data, in terms of nutrient uptake rate constraints, required to obtain accurate predictions also changing the default configuration of the GEM. From the shown results, the experimental uptake rate of carbon source is the input with higher impact on the predictions of metabolic flux distribution of wild type under the different environmental conditions, which is able to decrease the error from 0.25 to 0.15. The other inputs, such as oxygen and ammonium uptake rates, are automatically adjusted only as a function of the this last constraint. However, we observed a good prediction accuracy also using the values of glucose uptake rates reported by Varma and Palsson [83] to simulate the wild type stain under different growth conditions. Whereas for mutant strains, the knowledge of specific glucose uptake rates does not improve the predictions. This can be due to experimental noise on the measurements or because the mutant phenotypes are not properly modelled.

Constraint-based methods are important tools to study the steady-state cellular metabolic phenotype of microorganisms requiring only simple physical-chemical constraints, however, they present some limits when used for metabolic engineering applications. We showed that, the predictions of actual growth rate, and especially of target metabolite secretion rate, are difficult to compute for engineered strains with a large number of gene deletions, grown anaerobically or micro-aerobically, both by pFBA and MoMA methods.

In particular, we described the controversial situation due to the deletion of PFL reaction when the oxygen level is low or equal to zero. Despite PFL reaction was defined essential [76, 48], a growth rate different to zero was experimentally measured in this work for the strain with *ldh*A and *pfl*B deletion (WLP). For this disagreement there was no proper motivation, but could be due to different wild type strains used for the construction of mutants, a spontaneous activation of pyruvate dehydrogenase (PDH) in our strain, as reported by Singh et al. [82], or a growth under no strict anaerobic condition in our experiments. Similarly, when these deletions were in-silico simulated, pFBA predicted a growth different to zero and analogous to the value of wild type, instead with MoMA no growth was obtained.

The limited ability of constraint-based methods to represent context-specific phenotypes is due both to the starting metabolic model and the computational method. Indeed, the metabolic model includes all reactions implied by the genome annotation, some of which are not active under specific conditions, and the constraint-based methods are not able to identify the specific pathway configuration, but find the one that optimizes the objective function, namely the maximum growth rate for pFBA and the minimum changes between wild type and mutant strain for MoMA.

In order to tailor the GEM into a context-specific network and promote more accurate predictions for these engineered strains, different methods for the integration of additional omics data, in particular transcriptomic and proteomic, have been developed. Due to a lack of proteomic data for *E. coli* under anaerobic and micro-aerobic conditions, we focused on the integration of specific transcriptomic datasets, found in literature, by using GIMME approach. However, we obtained the same inaccurate predictions also considering transcriptomic data, that in this context are not able to properly simulate the real metabolic phenotype. As an example, the PDH reaction is predicted active, both under anaerobic and micro-aerobic conditions. However, it is catalyzed by a complex of three enzymes that under anaerobic growth conditions is transcribed but its activity is negligible due to the inhibition by NADH. For this reason, since transcriptomic data are not able to model the modification of expression at post-transcriptional and post-translational level, PDH reaction is wrongly modelled also with GIMME method. Therefore, despite the integration of this type of omics data adds more specific information on GEM, under similar complex situations these are unable to reflect the proper refinements at reaction level and therefore this approach does not always lead to an increase of prediction accuracy.

Chapter 3

Integration of enzymatic data in *Bacillus subtilis* genome-scale metabolic model to improve the phenotype predictions

The predictions obtained from constraint-based methods rely on the flux constraints, that define the space of all feasible solutions and exclude physiochemically and biologically infeasible behaviours. Commonly, the uptake rate of carbon source is the only constraint imposed into the model, that, in some cases, is not enough to obtain a good prediction accuracy, also if the experimental value is known. Different methods limiting the metabolic fluxes using enzymatic data were proposed.

Here, an enzyme-constraint model of *B. subtilis* has been developed to improve the in-silico design process useful for the metabolic engineering of this bacterium, frequently used as cell factory. The genome-scale metabolic models available for *B. subtilis* and the main approaches for the integration of enzyme levels will be introduced (Sec. 3.1). The construction of the new model and the conditions used for the performance evaluation will be described in Sec. 3.2 and then, the results obtained using it for the study of metabolic phenotype of wild type and mutant strains (Sec. 3.3.1) and for a direct metabolic engineering goal (Sec. 3.3.2) will be presented. Finally, the advantages of this approach will be discussed (Sec. 3.4).

3.1 Introduction

B. subtilis is the best-characterized bacterium among all the Gram-positive. It is able to produce large amounts of proteins, enzymes, vitamins and food supply fermentation by efficient secretion pathways [88, 89] and it has been defined as Generally Recognized as Safe (GRAS) by the FDA.

For its advantageous characteristics, in both research and industrial applications, *B. subtilis* is one of the most studied microorganisms in the metabolic engineering field. Different works were carried out, aimed at optimizing *B. subtilis* as cell factory for relevant medical, agricultural, pharmaceutical and other industrial bioproducts [90, 91, 92, 93].

Despite industrially-attractive results were achieved in the metabolic engineering of different strains, the rational identification of target manipulations to optimize the metabolism of microorganisms and the knowledge of the respective perturbed behaviour are the current challenges in the metabolic engineering field, due to the complexity of the biological systems and the growth conditions dependency. Therefore, in order to avoid a full trial-and-error approach, an in-silico approach to study metabolic behaviour, described via GEMs, was recently adopted for different microorganisms, especially bacteria, to drive the genomic optimization process before strain construction.

During the last ten years, three different versions of *B. subtilis* GEM [20, 33, 34, 35] were reconstructed and analyzed with the final aim to identify the genetic perturbations required to optimize the production of a target high-value metabolite, such as riboflavin, cellulase, (R,R)-2,3-butanediol and isobutanol [35]. The first *B. subtilis* GEM [20], called iYO844, is based on the annotated genome sequence [94], biochemical [95] and high-throughput phenotyping data and consists of 844 genes, 1020 biochemical reactions and 988 metabolites. However, since the genome annotation used for the iYO844 reconstruction is incomplete, two years later a new model (*i*Bsu1103) was reconstructed from an up-to-date annotation generated by the SEED Project [96]. *i*Bsu1103 model was further refined in 2013 [34] and the final GEM includes 1108 genes, associated to 1700 reactions, and 1390 metabolites and showed an improvement of accuracy, compared to iYO844, in terms of growth phenotype predictions (growth/no-growth), under different environmental and genetic conditions. Another *B. subtilis* GEM was proposed in 2013, called *i*Bsu1147, and is primarily based on the first reconstruction of *i*Bsu1103 with the addition of genomic information retrieved from KEGG and Uniprot and it was refined according to the simulations on biomass and ATP synthesis. This last *B. subtilis* GEM is characterized by 1147 genes, 1742 reactions and 1198 metabolites. It represents the most detailed version among the available *B. subtilis* GEMs, with superior accuracy in the prediction of growth phenotypes under different conditions [35].

Although the available *B. subtilis* GEMs were validated based on the growth rate under different substrates and gene essentiality analysis, these models show a low accuracy on the prediction of fluxes through the central carbon reactions and of secreted metabolites, and the integration of additional information on its metabolism could improve the prediction capability.

Since the GEMs are generally analyzed through constraint-based methods, commonly the predicted growth rate and the production of the target metabolite are mainly limited by the carbon source uptake rate constrained in the model. However, each metabolic flux is highly dependent on the concentration and kinetics of the enzyme catalyzing the reaction. For this reason, the predictions based only on the flux constraints may not agree with the experimental behaviour. Therefore, different approaches for the integration of enzyme levels in the metabolic model were developed to reduce the solution space and leave out the infeasible predictions. FBAwMC [97] was one of the first approaches using the concentration constraints for enzymes within the crowded cytoplasm, to improve the prediction of growth rates of *E. coli* under different growth media, without using the measurements of nutrients uptake rates. Other methods were proposed as an extension of FBAwMC, such as MOMENT [58], which utilizes the kinetic parameters under the limitations of the total enzyme pool available. Similarly, Nilsson et al. [98] used an extension of FBA to predict the metabolic trad-offs in yeast, in which the sum of fluxes was constrained to the sum of the product of the maximum in-vitro activity and the total enzyme mass, respect to a saturation factor. An alternative approach integrates quantitative measurements of protein and metabolite levels into GEM, by associating them with metabolic fluxes by using Michaelis Menten-like rate equations [99]. The most recent method, referred to as GECKO, uses enzymatic data, in the form of protein abundance and turnover number, as new constraints for each metabolic flux, so that it does not exceed its maximum capacity [59]. GECKO was applied to *S. cerevisiae* GEM and more realistic predictions than FBA were obtained under different carbon sources in excess, temperature stress, for the simulation of Crabtree effect and of metabolic behaviour of a single-gene knockout strain.

In this work, we integrated enzyme constraints for the reactions of central carbon and poly-γ-glutamic acid (γ-*pga*) production pathways into iYO844

GEM of *B. subtilis*, following the principles of GECKO method [59], with the aim to improve the predictions in terms of fluxes through the central carbon reactions and secretion rate of the main products, in addition to the growth rate, without the use of nutrient uptake rates. Subsequently, the accuracy of the developed enzyme-constrained model was evaluated under different genetic conditions and considering minimal medium supplemented with glucose as growth medium. Finally, we tested the potential of the *B. subtilis* enzyme-constrained model in a metabolic engineering application aiming to improve the production of the promising biopolymer γ-*pga*.

3.2 Materials and Methods

3.2.1 Data collection

The kinetic data, in the form of k_{cat} values $[s^{-1}]$, for the enzymes of *B. subtilis* central carbon and γ-*pga* production pathways, and of other two enzymes connected to these, were manually collected from BRENDA [56] and SABIO-RK [57] databases and literature, with the respective molecular weights $[kDa]$. In particular, we focused on active reactions in glucose minimal medium and aerobic condition. In literature, when k_{cat} values are not directly reported for the characterized enzymes, their activity can be expressed as *specific activity* (SA). This value is defined as the number of micromoles of product formed per milligram of enzyme per minute, under given temperature and pH. In this case, SA values were converted into the respective k_{cat} using the molecular weight of enzyme [100], with the assumption that the prepared enzyme is 100% pure and that the number of subunits is equal to the number of active sites [101]:

$$k_{cat}[s^{-1}] = \frac{SA[\mu mol \cdot mg^{-1} \cdot min^{-1}] \times MW[mg \cdot \mu mol^{-1}]}{60[s^{-1}]} \tag{3.1}$$

In particular, a manual search of parameters about the kinetic activity, in terms of k_{cat} value or specific activity, for the 43 enzymes of central carbon pathway and for the 5 enzymes of γ-*pga* production pathway reported in iYO844 model [20] was carried out. When no measure was available specifically for *B. subtilis*, k_{cat} value for *E. coli* was retrieved from the collection reported by Davidi et al. [100], resulting in a final data collection for 15 enzymes. In fact, since *E. coli* is the most widely studied species of bacteria, the number of enzymes with measured k_{cat} values is larger than in other organisms and moreover its central carbon pathway is very similar to those of *B. subtilis.* However, since the preliminary simulation of the model obtained by the integration of this set of enzymatic data predicted the activation of two new reactions (MICITL and PGCDr) experimental reported as inactive, the addition of data for these reactions was required.

We considered only the reactions catalyzed by a unique enzyme, except for the CS and OXADC reactions, for which 2 enzymes are associated but they are mainly catalyzed by *citZ* [102] and *oxdC* [103], respectively, while the second enzyme of both reactions (citA and oxdD, respectively) does not give a relevant contribution.

As for the absolute protein quantifications, the data reported by Goelzer et al. were considered [104]. These values are expressed in number of molecules per cells and are obtained from LC/MS^E analysis [105] for most of the cytosolic proteins in *B. subtilis* 168 strain, growing under aerobic batch conditions and in minimal media. In particular, the measurements in minimal medium with glucose were extracted for this study and converted into mmol/g_{DW} by assuming 6.3 x 10^8 cells per ml per OD [105]. For each enzyme with known k_{cat}, we used the upper limit of the 95% confidence interval of abundance value to allow flexibility in case of variable measurements (Table 3.1). When the quantification of a specific enzyme was not available, we assumed that the measure is under detection limit, and the minimum value among all the measurements in the same condition was considered.

The total amount of proteins (P_{total}) in the cell was considered equal to 0.55 g/g_{DW}, corresponding to the value measured for *E. coli* [106], and the mass fraction of proteins (*f*) was computed equal to 0.0191, by summing the abundance, expressed as parts per million (ppm), of the 17 considered proteins, retrieved from PaxDB database [107].

3.2.2 Integration of enzymatic data in the model

The enzymatic data reported in Table 3.1 were integrated into the iYO844 model, consisting of 1020 reactions, in order to obtain a enzyme-constrained model [59] easy to use with all the computation methods of metabolic engineering. This method was implemented via MathWorks MATLAB R2012a and run with COBRA toolbox [9]. The construction of the enzyme-constrained model was based on the same formalism of FBA [37]. To implement the enzymatic data integration following GECKO approach (see section 1.1.3), an additional constraint was considered so that the metabolic flux through the j-th reaction (R_j), reported in Table 3.1, does not exceed its maximum capacity (v_{max}),

Reaction name	**Gene name**	**Equation**	k_{cat} $[s^{-1}]$	**[E]** **[mmol/g_{DW}]**	**Organism of k_{cat} data**	**Ref. for k_{cat} data**
PGI	*pgi*	g6p→f6p	126	1.55×10^{-5}	*E. coli*	[108]
TPI	*tpiA*	dhap→g3p	150	1.28×10^{-5}	*E. coli*	[109]
GAPD_NAD	*gapA*	g3p+nad+pi→13dpg+h+nadh	70	5.77×10^{-5}	*B. subtilis*	[110]
PGK	*pgk*	13dpg+adp→3pg+atp	329	3.60×10^{-5}	*E. coli*	[111]
PGM	*pgm*	3pg→2pg	765.9	8.85×10^{-6}	*B. subtilis*	[112]
ENO	*eno*	2pg→h2o+pep	130.4	3.17×10^{-5}	*B. subtilis*	[113]
G6PDH	*zwf*	g6p+nadp→6pgl+h+nadph	174	8.05×10^{-6}	*E. coli*	[114]
CS	*citZ*	accoa+h2o+oaa→cit+coa+h	49	2.51×10^{-5}	*B. subtilis*	[102]
ICDHy	*icd*	icit+nadp→akg+co2+nadph	82	1.10×10^{-4}	*B. subtilis*	[115]
FUM	*citG*	fum+h2o→mal-L	283.3	7.29×10^{-6}	*E. coli*	[116]
MDH	*mdh*	mal-L+nad→h+nadh+oaa	177.1	1.06×10^{-4}	*B. subtilis*	[117]
PTAr	*pta*	accoa+pi → actp+coa	651.6	8.49×10^{-6}	*B. subtilis*	[118]
LDH_L	*ldh*	lac-L+nad→h+nadh+pyr	6416.6	3.60×10^{-6}	*B. subtilis*	[119]
PGCDr	*serA*	3pg+nad→3php+h+nadh	14.56	1.90×10^{-5}	*B. subtilis*	[120]
OXADC	*oxdC*	h+oxa→co2+for	59	6.21×10^{-7}	*B. subtilis*	[121]
MICITL	*yqiQ*	micit→pyr+succ	19	6.80×10^{-8}	*E. coli*	[122]
OXGDC	*menD*	akg+h→co2+sucsal	20	6.80×10^{-8}	*B. subtilis*	[123]

Table 3.1: **List of k_{cat} values and protein quantifications integrated in iYO844 model.** For each reaction reported in the table, the encoding gene, the equation, the $k_{cat}$and concentration of catalyzing enzyme are reported. Moreover, the organism for which the k_{cat} value was measured and the paper in which it was found are specified. The reaction names, with the associated gene names and equations, correspond to the annotations used in iYO844 model.

corresponding to the product between the k_{cat} value (expressed as h^{-1}) of the enzyme E_i (that catalyzes the j-th reaction) and its abundance (expressed as mmol/g_{DW}, as described above):

$$v_j \leq k_{cat}^{ij} \cdot [E_i] \qquad for\ \ i\ \ and\ \ j = 1....17 \tag{3.2}$$

Since we considered reactions catalyzed by a unique enzyme, the number of enzyme-constrained reactions (17) is equal to the number of enzymes.

In summary, each constrained metabolic reaction R_j includes a pseudo-metabolite representing enzyme usage, which is limited by protein abundance.

To implement the described method, as a first step, the iYO844 model was converted into an irreversible model and the constraint for uptake rate of glucose, the sole carbon source, was removed. Then, the stoichiometric matrix and the upper bound vector of the model were expanded by adding the k_{cat} values and the protein abundance known (Table 3.1).

Specific proteomic data for the mutant strains tested in this work were not available. For this reason, the approach shown above was applied under the assumption that enzyme concentrations in wild type and mutant strains are the same, except for the enzyme associated to the deleted reaction, whose concentration was fixed to zero. An alternative approach was tested for mutant strains simulations, in which only the total amount of enzymes was constrained, similar to previously proposed approaches [58, 97]:

$$\sum_{i}^{17} MW_i e_i \leq f \cdot P_{total} \tag{3.3}$$

The mass fraction of accounted proteins (*f*) and the total protein measured in the cell (P_{total}) were retrieved from PaxDB database and literature [106], respectively. In addition, the kinetics data, in terms of turnover number, were integrated in the stoichiometric matrix (S) as in the standard approach.

3.2.3 Simulations

For the in-silico simulation of each mutant strain, the reaction encoded by the knocked-out gene was set as inactive, namely with a flux equal to zero. All strains were simulated using both the previously-published iYO844 model and the new enzyme-constrained model obtained in this work. The upper bound of glucose uptake rate was fixed to the specific experimental value for the iYO844 model, based only on stoichiometric reactions and directionality, whereas to an unlimited value (1000 mmol/g_{DW}h) for the enzyme-constrained iYO844 model. The metabolic phenotypes of *B. subtilis* wild type and mutant strains were predicted by pFBA [38]. Mutant strains were also simulated using MoMA method [41], which minimizes the distance between the flux distributions in the wild type and mutant strains. The pFBA and MOMA methods were run in Matlab using the available packages in COBRA Toolbox [9].

3.2.4 Identification of deletion and over-expression targets

The genetic perturbations that are required to optimize the production of γ-*pga* were identified by using the iYO844 model and its enzyme-constrained version, via two different algorithms, specific for the prediction of gene deletion and amplification targets, respectively. The first method uses MoMA to find the single or multiple gene deletions corresponding to the best trade-off between growth rate and secretion rate of the target metabolite [66]. The second one, FSEOF (see section 1.1.2), selects the fluxes that increase when the flux towards product formation is enforced as an additional constraint during flux analysis [46]. In particular, we considered the reactions whose flux profile increases monotonically with the enforced objective flux and with highest fold-change respect to its initial value. Since the *pgs* operon, including the enzyme-encoding genes responsible for γ-*pga* production, is not expressed in the laboratory strain modelled in iYO844, we added the production reaction reported in the GEM of *Bacillus licheniformis* [124] (*0.77 glu-D + 0.23 glu-L* $\rightarrow \gamma$-*pga*). Moreover, we considered, as an alternative approach to optimize the γ-*pga* production, the maximization of three of its well-characterized precursors: 2-Oxoglutarate (*akg*), D-Glutamate (*glu-D*) and L-Glutamate (*glu-L*). In this case, the secretion reactions of these precursors were added into the model.

3.2.5 Evaluation of prediction accuracy

For the evaluation of the prediction accuracy of flux distributions for *B. subtilis* wild type and single-gene deletion strains, grown under M9 minimal medium with glucose, the respective experimental growth rate $[h^{-1}]$ and measured external and internal fluxes [mmol/g_{DW}h] from the literature were considered (single value for each reaction) [125, 126]. The internal fluxes are reported for the main reactions of the central carbon pathway and were measured by ^{13}C-labeling experiments. Furthermore, 95 single-gene deletions, experimentally found to be lethal [125, 127], were simulated via pFBA and MoMA methods with standard and integrated models, in minimal medium with glucose, and the percentage of correctly predicted essential genes (i.e., yielding a predicted growth rate lower than 0.05) was calculated and compared.
The prediction error was computed for each simulation of a strain grown under a specific condition by the normalized Euclidean distance between the experimental and the respective predicted fluxes:

$$PRED\ \ ERROR = \frac{\|exp\ \ flux - pred\ \ flux\|}{\|exp\ \ flux\|} \tag{3.4}$$

The distribution of prediction errors for the five mutant strains was evaluated to analyze the accuracy of the models under perturbed genetic conditions.

3.2.6 Sensitivity analysis

The prediction errors, obtained from 10,000 simulations of wild type strain using the enzyme-constrained model with k_{cat} values randomly extracted from the list of measured values (Tab. 3.1), were computed. Furthermore, by following a stepwise inclusion procedure, we selected the minimum set of reactions that must be constrained with the respective enzymatic data in order to achieve the final accuracy of the new model. In particular, the prediction errors for wild type and mutant strains together with the accuracy of gene essential predictions were taken into account as indexes for the final accuracy.

3.3 Results

From the comparison of the three *B. subtilis* models [20, 34, 35], we noted that the latest published GEM, iBsu1147, describes the glucose transport from external to internal compartment via proton symport (GLC-Dt2), instead of using the transport via PEP and the phosphotransferase system (GLCpts) [128], as is correctly predicted by the other two models. Finally, considering the predicted central carbon pathway fluxes and acetate external flux, iYO844 showed higher accuracy than iBsu1103V2, with a prediction error equal to 0.37 and 0.65, respectively. Therefore, we decided to used iYO844 to model the *B. subtilis* phenotype.

3.3.1 Evaluation of prediction performance

The metabolic behaviour of wild type *B. subtilis*, grown in minimal medium with glucose, was predicted with pFBA method both using iYO844 model, with glucose uptake rate fixed to the experimentally measured value (7.71 mmol/g_{DW}h) for a more realistic simulation, and using the developed enzyme-constrained model. The comparison of available experimental fluxes with the corresponding predicted values (Fig. 3.1) showed that, despite iYO844 model is able to accurately predict growth rate, the flux distribution for central carbon metabolism is not consistent with experimental values, especially for the reactions of pentose phosphate pathway (PPP) and acetate secretion. A significant improvement was achieved with the integration of enzymatic data, that corrects the largely inaccurate predictions of the PPP and acetate flux distribution. The overall increase of prediction accuracy using the enzyme-constrained model is confirmed by the decreasing of the prediction error from 0.47 to 0.27.

Moreover, the prediction performance of the new model was tested also under perturbed genetic conditions. In particular, the metabolic behaviours of five mutant strains, constructed by single-gene deletions, (Δpgi, Δzwf, ΔsdhABC, Δmdh, ΔserA) were simulated with pFBA and MoMA methods using the two models. The experimental fluxes of four central carbon reactions (PGI, G6PDH, PYK and CS) and acetate production [125] were compared with the

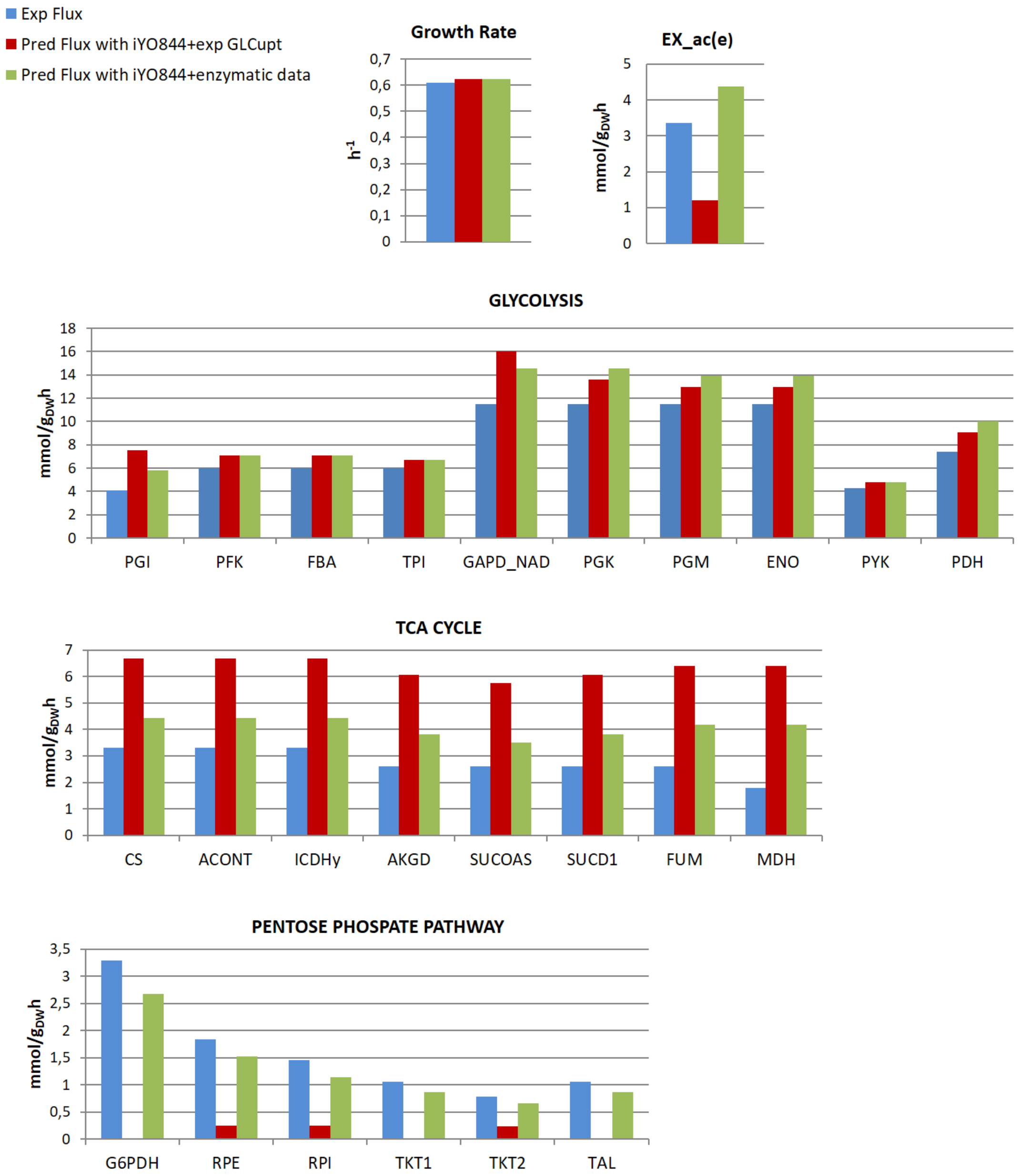


Figure 3.1: **Experimental and predicted fluxes for wild type *B. subtilis*.** The predictions of growth rate (h^{-1}), acetate secretion rate (mmol/g_{DW}h), fluxes through the reactions of glycolysis, TCA cycle and pentose phosphate pathway (mmol/g_{DW}h) for wild type, using iYO844 model with the experimental value of glucose uptake rate and the enzyme-constrained model (red and green bars, respectively), are compared to the experimental measures (blue bars) to analyzed the respective accuracy.

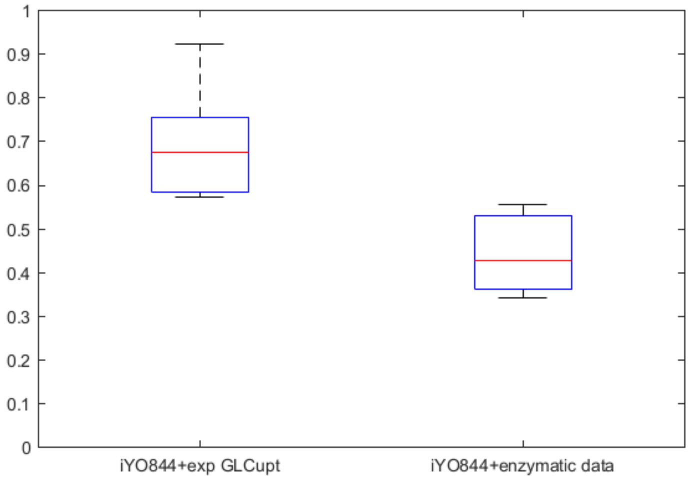


Figure 3.2: **Distribution of prediction errors for *B. subtilis* mutant strains.** The normalized prediction errors obtained using iYO844 with the experimental value of glucose uptake rate and the enzyme-constrained model were computed (see section 3.2) for each of 5 mutant strains considered in this work.

corresponding predicted values. First, we noted that MoMA gives considerably lower error than pFBA in the prediction of flux distribution of mutants, especially when the initial iYO844 model is used (data not shown). For this reason, we considered the predictions obtained from MoMA method to compare the two models. The results show an overall improvement of prediction accuracy when enzymatic data are integrated also for mutant strains: considering the distribution of prediction errors, for all mutants the median value decreases from 0.67, using the initial model, to 0.43 (Fig. 3.2). In particular, from the comparison between each flux predicted by the two models and the experimental measure (Fig. 3.3), a significant improvement of accuracy for the acetate secretion rate and for the flux of G6PDH reaction is observed. The rates of acetate production computed by analysis of iYO844 are lower than the measures for each strain, especially for ΔmdH and ΔserA, in which MDH and SUCD1 reaction is blocked respectively, with predicted value closed to zero. However, with the integration of enzymatic data into the model, the predicted rates become similar to the experimental values. Similar results are obtained for the G6PDH reaction, the first one of PPP. G6PDH flux is measured active in Δpgi, ΔsdhABC, Δmdh, ΔserA strains, with maximum value

of 6.5 mmol/g_{DW}h, when the PGI reaction is blocked (ΔPgi strain), and with minimum value of 0.6, when PGDHr reaction is blocked (ΔserA strain). But, except for ΔsdhABC strain, only using the new model the predicted fluxes through G6PDH reaction are in agreement with the measurements.
Moreover, with the model developed in this work, the fluxes through CS reaction of TCA cycle decrease for each mutant strain, resulting more similar to experimental values, thanks to the respective enzymatic data that limit the maximum flux.
We compared the performance of developed enzyme-constrained model with those of model in which, similarly to MOMENT method [58], the k_{cat} values for the 17 reactions were integrated and only the total amount of enzymes (g/g_{DW}) was constrained (see section 3.2). The results showed that, also in this context, the predictions obtained by GECKO approach are the most accurate. Indeed, the median value of the prediction errors for the five mutants obtained with this alternative method is computed equal to 0.55, namely lower than the error obtained with the initial model (0.67), but higher than the developed model with the integration of concentration for each of the 17 enzymes (0.43).

Finally, the new model was evaluated by gene essentiality analysis, a commonly performed step for the validation of new GEMs. As before, we used both pFBA and MoMA methods to simulate the growth phenotype of single-gene deletions, but only the results obtained with MoMA were showed since it has a higher prediction accuracy with both models, as obtained above for the flux distribution analysis of mutants. From a list of 95 essential genes, the enzyme-constrained model is able to correctly predict the lethal effect of each deletion on growth with 75% accuracy, which is 3% higher than the accuracy using the initial model (Tab. 3.2). Considering only the 11 knockout strains with single deletion in central carbon metabolism genes (*eno*, *pgm*, *ywlF*, *pycA*, *pdhA*, *odhB*, *fbaA*, *tpiA*, *pgk*, *tkt* and *pfkA*), no growth phenotype was predicted by iYO844 only for two strains (Δ*pgk* and Δ*tkt*), whereas the developed model was able to predict the essentiality of 3 additional central carbon genes (*pfkA*, *eno* and *pgm*), for a total of 5 out of 11 correctly predicted strains. No difference was observed in the prediction of the remaining 84 strains, in which deletions are present in genes belonging to other pathways.
The sensitivity of the model respect to k_{cat} values was analyzed considering the variation of prediction accuracy obtained with these values randomly assigned (see section 3.2). Results (Fig. 3.4) showed that the median value of prediction errors using random k_{cat} (0.64) is significantly higher than the value obtained

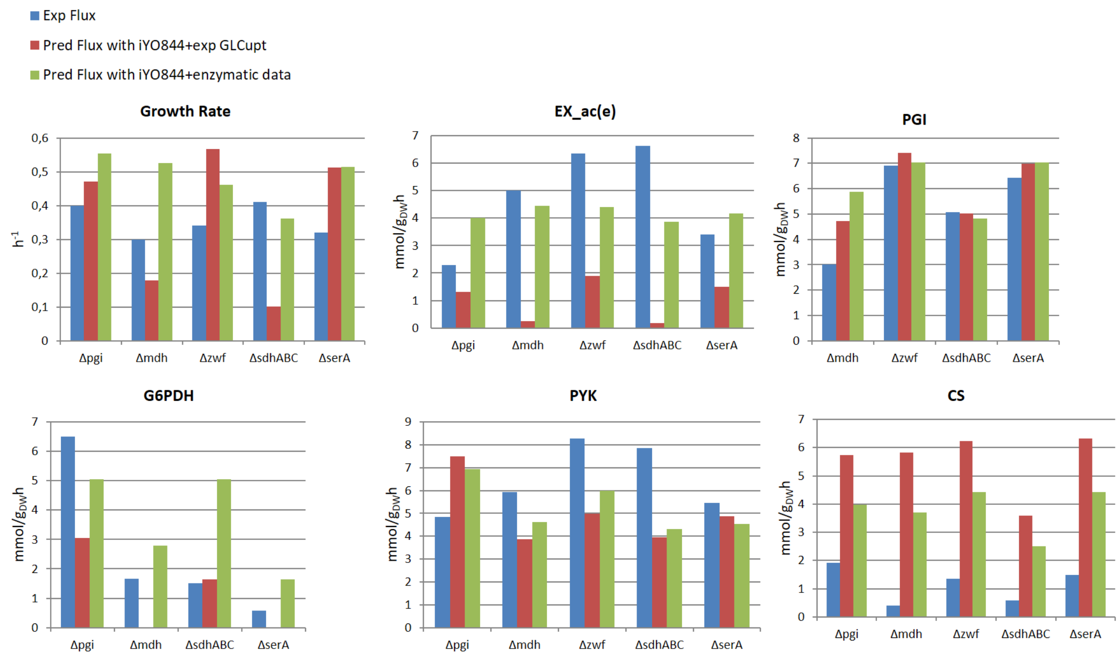


Figure 3.3: **Experimental and predicted fluxes for *B. subtilis* mutant strains.** Five different mutants strains (Δpgi, Δmdh, Δzwf, ΔsdhABC and ΔserA) were simulated using iYO844 model, with the experimental value of glucose uptake rate, and the enzyme-constrained model (red and green bars, respectively). The predictions of growth rate (h^{-1}), acetate secretion rate (mmol/g_{DW}h) and fluxes through PGI, G6PDH, PYK, CS reactions (mmol/g_{DW}h) are compared to the experimental measures (blue bars) to analyzed the respective accuracy.

Essential genes			
	CC	**OTHERS**	**TOT**
Experimental data	11	84	95
Predicted with GEM	2	67	69
Predicted with enzyme-constrained model	5	67	72

Table 3.2: **Gene essentiality analysis.** The number of essential genes reported by [125, 127], and predicted using iYO844 GEM and the enzyme-constrained model are reported. The number of genes encoding the central carbon (CC) pathway and the others are specified.

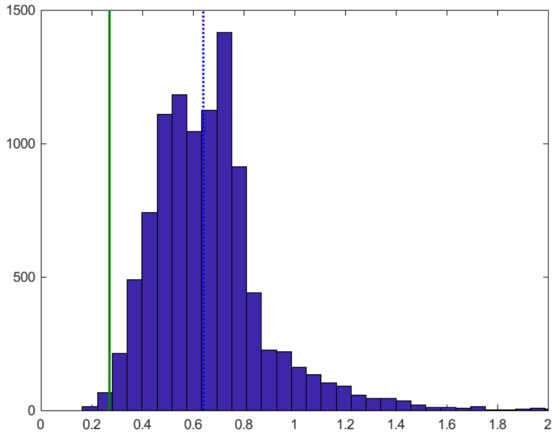


Figure 3.4: **Distribution of predicted errors obtained from enzyme-constrained model with mixed k_{cat} values.** Considering the wild type simulations, the dashed blue line indicates the median value of the prediction errors obtained using k_{cat} values randomly assigned and the solid green line the prediction error obtained using k_{cat} values properly assigned.

by enzyme-constrained model (0.27). The knowledge of specific k_{cat} values is therefore essential to obtain accurate predictions. Moreover, among the 17 enzyme-constrained reactions, we identified TPI, GAPD_NAD, CS, PGCDr as the minimum set required to achieve the final prediction performance of the developed model, namely a prediction error equal to 0.27 for wild type, equal to 0.43 for mutants and the 75% of gene essentiality accuracy. This result, however, depends on the study on which the model is applied and as the number of enzyme-constrained reactions increases, the accuracy of the predictions improves.

3.3.2 Metabolic engineering application

Once an overall improvement of prediction performance was found with the new model, it was used to identify the genetic perturbations required to increase the production of γ-*pga*. The former is a biodegradable, water-soluble, non-toxic, and edible biopolymer that has a large number of biotechnological applications, ranging from biomedicine to bioremediation. Since the *pgs* operon, including the enzyme-encoding genes responsible for the synthesis of this biopolymer, is not expressed in laboratory strains, several experimental studies were performed to improve γ-*pga* production by using derivatives of *B.*

subtilis 168 strain [129, 93].

For the in-silico modelling, iYO844 and the enzyme-constrained model were modified to properly take into account the γ*-pga* production process (γ-PGA reaction) and both gene deletion (with MoMA method) and amplification (with FSEOF method) targets were predicted for its production optimization (see section 3.2), in addition to three of its precursors (*akg*, *glu-L*, *glu-D*).

The identified deletions are the same for each of target metabolites with both the two models. Using GEM the suggested single-reaction deletions are AKGD (*akg + coa + nad* → *co2 + nadh + succoa*) and SUCOAS (*atp + coa + succ* ⇔ *adp + pi + succoa*) reactions (Tab. 3.3). In particular, the elimination of SUCOAS reaction results more efficient than AKGD deletion, with a production rate of target metabolite about two-fold higher (data not shown). However, the best performance of target production is predicted by the double deletion of AKGD and OXGDC or SSALy reactions (*akg + h* → *co2 + sucsal/h2o + nadp + sucsal* → *(2) h + nadph + succ*), for which is associated a growth rate lower than the predicted single deletions. SUCOAS and AKGD reactions are downstream of the *akg* production in the TCA cycle (see Fig. 3.5), whereas OXGDC and SSALy are consecutive reactions forming a bypass pathway for the production of *succ* from *akg*. OXGDC and SSALy reactions are predicted to be active only when the AKGD reaction is blocked, however experimental studies reported an undetectable concentration for the protein associated with OXGDC (menD). This incorrect prediction was settled using the enzyme-constrained model developed in this work, in which OXGDC is one of the enzyme-constrained reactions and therefore the bypass pathway was not classified as a competitor for the use of *akg*. In this way, the single deletions of AKGD and SUCOAS were identified also using the new model (Tab. 3.4), but the first one achieves alone the same performance of the double deletion predicted with the GEM.

Similarly, simulating the best deletion predicted with MoMA on the relative model (SUCOAS and AKGD, respectively), we identified by FSEOF the reaction fluxes whose increase promotes the production rate of γ*-pga* and each of its precursors. In particular, using GEM we selected PC (*atp + hco3 + pyr* → *adp + h + oaa + pi*) as the best candidate reaction to be over-expressed to increase the production rate of *akg* and GLUR reaction for *glu-D* and γ*-pga* production, whereas no reaction was selected with our criteria for *glu-L* (see Tab. 3.5). PC reaction is the first of TCA cycle and GLUR reaction produces *glu-D* from *glu-L*, which are the metabolites used for the γ*-pga* production (Fig. 3.5). With the integration of enzymatic data into the model, CS, ACONT and

GEM	
Target metabolite	**Suggested deletions**
akg	AKGD, SUCOAS, AKGD+OXGDC/SSALy
glu-L	AKGD, SUCOAS, AKGD+OXGDC/SSALy
glu-D	AKGD, SUCOAS, AKGD+OXGDC/SSALy
γ-pga	AKGD, SUCOAS, AKGD+OXGDC/SSALy

Table 3.3: **Deletions identified with GEM to improve the production of target metabolite in *B. subtilis*.** Deletions with the best trade-off between growth rate and secretion rate of the target metabolite based on MoMA approach are indicated. For the reaction names, the nomenclature reported into iYO844 model is reported.

Enzyme-constrained model	
Target metabolite	**Suggested deletions**
akg	AKGD, SUCOAS
glu-L	AKGD, SUCOAS
glu-D	AKGD, SUCOAS
γ-pga	AKGD, SUCOAS

Table 3.4: **Deletions identified with enzyme-constrained model to improve the production of target metabolite.** Deletions with the best trade-off between growth rate and secretion rate of the target metabolite based on MoMA approach are indicated. For the reaction names, the nomenclature reported into iYO844 model is reported.

GEM	
Target metabolite	**Suggested over-expressions**
akg	PC
glu-L	—
glu-D	GLUR
γ-pga	GLUR

Table 3.5: **Over-expressions identified with GEM to improve the production of target metabolite.** The results specific for each target metabolite are indicated. No reaction was identified considering glu-L as target metabolite. The strain with the best deletion identified with MoMa was considered. For the reaction names, the nomenclature reported into iYO844 model is reported.

Enzyme-constrained model	
Target metabolite	**Suggested over-expressions**
akg	CS, ACONT, ICDHy, R_HIST
glu-L	CS, ACONT, ICDHy
glu-D	GLUR, CS, ACONT, ICDHy
γ-PGA	GLUR, CS, ACONT, ICDHy

Table 3.6: **Over-expressions identified with enzyme-constrained model to improve the production of target metabolite.** The results specific for each target metabolite are indicated. The strain with the best deletion identified with MoMa was considered. For the reaction names, the nomenclature reported into iYO844 model is reported, except for R_HIST through which a set of 8 reactions of histidine pathway are summarized.

ICDHy reaction fluxes showed an increasing pattern with increasing the production of each target metabolite considered in this work (Tab. 3.6). These three reactions are located in the TCA cycle and are responsible for *akg* production. In addition, with the enzyme-constrained model, 8 reactions of histidine pathway (R_HIST) were selected to increase the *akg* production and, as found also with GEM, GLUR reaction when the maximization of *glu-D* and γ*-pga* production is considered. R_HIST reactions describe the degradation process of L-histidine and the corresponding production of L-glutamate.

3.4 Discussion

In this work, an enzyme-constrained model of *B. subtilis* was developed with the aim to improve prediction power about the metabolic behaviour under different conditions, without using the experimental uptake rate of the

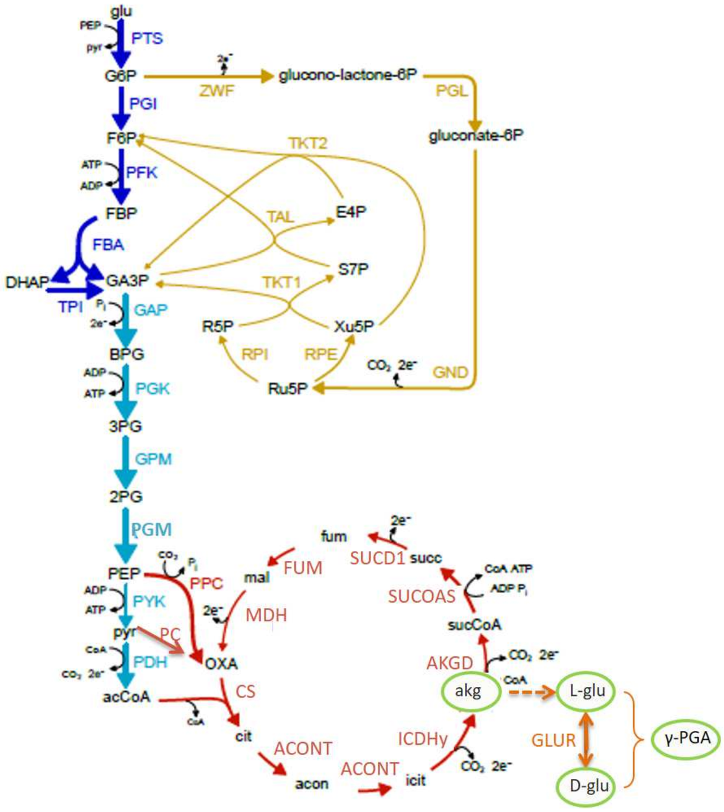


Figure 3.5: **Central carbon and γ-PGA production pathways of *B. subtilis.*** The central carbon pathway that includes the pathway of glycolysis (blue and cyan arrows), the pentose phosphate pathway (yellow arrows) and the TCA cycle (red arrows), and the pathway of γ-*pga* synthesis (orange arrows) are represented. The target metabolites considered in this work are marked with a green circle.

carbon source, and facilitate the rational design process for its metabolic optimization. In this context, for its advantageous characteristics, in both research and industrial applications, *B. subtilis* is one of most engineered Gram-positive bacteria as cell factory for commercially interesting products.

The principles proposed by GECKO method [59] were applied to the first GEM published for *B. subtilis* [20], namely the maximum flux of reactions was constrained by the product between the k_{cat} value and the concentration of enzyme. Since k_{cat} measures are available for a small number of enzymes, we focused on retrieving enzymatic data for the reactions of central carbon and γ-PGA production pathway, and finally, 17 of these reactions were enzymatically constrained into the iYO844 model.

Using the developed model to simulate the metabolic phenotype of *B. subtilis*, under different genetic conditions, in terms of growth rate, flux distribution through central carbon reactions and acetate secretion rate, we showed that the integration of enzymatic data is essential to increase the prediction capability of GEM, despite in this work the number of enzyme-constrained reactions is limited. It decreases the prediction error from 0.47 to 0.27 for wild type, and from 0.67 to 0.43 for five mutant strains considered in this study. In particular, a significant improvement was obtained for the prediction of acetate production rate and of fluxes through the pentose phosphate pathway, which using GEM are computed much lower than the experimental measures. For mutant strains, since the specific protein abundance is not available, we used the data of wild type changing only the concentration of enzyme associated to the deleted reaction, fixed equal to zero, assuming that protein abundances are not significantly perturbed by the single knockout. Alternatively, MOMENT, based on a less restrictive constraint for enzyme concentrations, was implemented only for mutant strains. We showed that GECKO approach was superior in terms of simulation accuracy, with a prediction error of 0.43, while an error of 0.55 was obtained with the MOMENT approach. Moreover, 3 additional essential genes of central carbon pathway were properly identified using the model developed by GECKO method, with a final gene essentiality accuracy of 75%, higher than the percentage obtained with GEM (72%).

Finally, we presented the results obtained using *B. subtilis* enzyme-constrained model on a specific metabolic engineering application. In particular, the knockout and amplification targets were identified to optimize the production of γ-*pga*, a biopolymer with many useful proprieties. We showed that, AKGD and SUCOAS reactions are predicted by the developed model as the best candidates for the deletion, in agreement with the results obtained with GEM.

However, with the only deletion of AKGD reaction the enzyme-constrained model is able to compute a production rate of target metabolite similar to the performance achieved by the double deletion (AKGD+OXGCD/SSALy) predicted with GEM. In this way, the integration of enzymatic data prevents the prediction of knockout for reactions with very low enzymatic activity (OXGCD and SSALy) and therefore without significant impact. The inverse correlation of *γ-pga* production with the SUCOAS and AKGD reactions was confirmed by Yu et al. [130] in *B. licheniformis*. Whereas for the amplification targets, the integration of enzymatic data leads to identify as candidates new reactions closely connected with the production of *γ-pga*, *akg*, *glu-L* and *glu-D*. Experimental validations are required to specifically evaluate the power of new model to identify the target manipulations for the optimization of a considered metabolic capacity.

According to the result reported by S*á*nchez et al. [59], the predictions obtained from the developed model largely rely on the assigned k_{cat} values, which are therefore essential to achieve good accuracy.

In general, since enzymatic information was integrated considering the central carbon metabolism and γ-PGA production pathway, the improvement of prediction performance in this pathway is consistent with our expectations; other pathways, currently not improved, may benefit from the GECKO approach once enzymatic information is available at genome scale.

Chapter 4

Quantification of the gene silencing performances of rationally-designed synthetic small RNAs[1]

Synthetic biology provides different techniques and tools for specific gene expression control to achieve desired phenotypes of various bacteria. A traditional metabolic engineering strategy for the overproduction of the target chemical consists in the inactivation of competing pathways. Experimentally, gene deletion is commonly obtained via homologous recombination, that alters the host cell's genotype permanently. Despite the successful applications of this gene deletion technique, a laborious procedure and a previous accurate study of essential genes are required for the permanent inactivation, and, moreover, multiple deletions are hard to be obtained simultaneously. Therefore, methods to repress gene expression levels, without modification of genome sequences, were developed.

In this chapter, small RNAs (sRNAs), tools for the silencing of target genes, will be characterized in *E. coli*, also with the support of a mathematical model. First, the three general categories of naturally occurring sRNAs will be introduced and the previous studies and applications of artificial sRNAs will be reported (Sec. 4.1). An accurate description of design and construction pro-

[1] The contents of this chapter are published in *Massaiu I, Pasotti L, Casanova M, Politi N, Zucca S, Cusella De Angelis MG and Magni P* Systems and synthetic biology

cesses implemented in this study and the mathematical model used for the characterization will be presented (Sec. 4.2). The silencing performance of the developed sRNAs, at different copy numbers, targeting the reporter gene RFP, expressed at different transcription levels, under the control of different promoters, in different strains, and in single-gene or operon architecture, will be shown (Sec. 4.3). Moreover, in this section, the specific silencing of the endogenous *ldh*A gene, encoding lactate dehydrogenase (LDH) involved in the fermentation pathway of *E. coli*, will be described. Finally, the overall silencing capability of the two synthetic sRNAs, designed with recently proposed guidelines, will be discussed in Sec. 4.4. Additional figures about the results of evaluation study are reported in the App. A.
This study has been published in [131].

4.1 Introduction

Naturally occurring sRNAs are short non-coding RNAs, typically between 50-250 nucleotide long, able to control the expression of target genes in bacteria, predominantly at the post-transcriptional level [132, 133, 134, 135]. Hundreds of sRNAs have been identified in different bacteria so far, especially in *E. coli*. These sRNAs can be sorted in three general categories: sRNAs that have intrinsic catalytic activity or are components of ribonucleoproteins, sRNAs that affect protein activity and sRNAs that regulate gene expression by base-pairing to a target mRNA [133]. The third sRNAs category is the best-characterized and the most abundant in Gram-negative bacteria. The post-transcriptional control system of sRNAs belonging to the latter category is based on a trans-acting mechanism, in which sRNAs bind to the 50 untranslated region (50-UTR) or to the translation initiation region (TIR) of single or multiple target messenger RNAs (mRNAs) through imperfect base-pairing [136], although sRNA binding in regions downstream of the TIR have also been reported [137]. The regulation of gene expression is carried out, upon binding, by the modulation of translation or transcript stability [135, 138, 139, 140]. In particular, specific sRNAs are known to change ribosome accessibility, mainly repressing translation, although examples of positive regulation have also been reported [141, 135]. Conversely, the final effect of other sRNAs is to change the stability of the target mRNA, mainly by accelerating its degradation [141, 135, 142]. Morita et al. [143] analyzed the repression effects of two sRNAs, SgrS and RyhB, which experimentally showed to degrade their target transcript. By inhibiting the mRNA degradation machinery of the host strain, the authors showed that SgrS and RyhB could act as translation inhibitors without affecting transcript stability. This double final effect of the investigated sRNAs, i.e., translation repression and mRNA degradation, could be motivated by the necessity to rid the cell of translationally inactive mRNAs [143]. A major class of sRNAs requires (or are strengthened by) the RNA chaperone Hfq for efficient gene silencing [136, 141, 135].

In nature, sRNAs are involved in the regulation of disparate functions in bacteria, such as stress response, outer membrane protein biogenesis, quorum sensing, virulence, iron and sugar metabolism [141, 133]. Due to their importance, several studies have been recently carried out to discover sRNAs, identify their targets and characterize their regulation mechanisms, also with the help of mathematical models. Among all the research investigations, high-throughput analyses have been performed to search for novel sRNAs or specific targets of known sRNAs by microarray studies [141, 138, 144] and, more recently, by the sort-seq approach [135]. Reporter gene fusions with the initial part of a target gene (including the 5'-UTR) have been adopted to quantitatively study the contribution of sRNAs and their specificity [145]. Mathematical models, generally studied at the steady-state, were developed to quantify the effects of parameter changes, such as mRNA/sRNA levels and their half life [142, 146]. Such kinetic models were able to capture the observed behaviour of artificially-constructed systems where sRNA and mRNA levels were tuned [142, 146]. Moreover, the proposed models are sufficiently general to describe the contribution of sRNAs affecting translation and/or transcript degradation [134]. One of these models was recently refined to investigate the effects of ribosome binding site (RBS) strength on the RyhB, DsrB and OmrA sRNA efficiency towards their target gene, whose expression was studied via gene fusion and quantitative PCR [147]. Comparison between model simulations and experimental data demonstrated that, in the context of the investigated sRNAs, increasing translation rate can lead to increased repression [147].

Inspired by the features of natural control systems, sRNAs can also play an important role in the design of synthetic biological systems. In metabolic engineering studies, metabolic fluxes towards the target bioproduct can be optimized by the simultaneous expression of heterologous genes, over-expression and down-regulation of endogenous genes. In this case, sRNAs can be used to down-regulate the expression of the target genes involved in the desired pathway. The use of an sRNA-based approach has several advantages over the common gene knockout method: 1) sRNA expression plasmids, which actuate the silencing of target genes, can be incorporated in the host strain by simple bacterial transformation and it makes sRNA systems highly portable to different hosts; 2) several combinations of sRNAs can be simultaneously tested by co-transforming different expression plasmids or assembling multiple cassettes in the same vector; 3) a scalable, sRNA sequence-dependent, repression efficiency can be obtained; 4) sRNAs can be used to down-regulate essential genes, since sRNAs can be placed under the control of inducible promoters [133].

It is worth noting that other post-transcriptional control systems, such as antisense RNAs (asRNAs) also have many of the advantages described above [148, 149]. However, asRNA-based systems are generally characterized by a lower efficiency than sRNAs [133], although efforts have been recently carried out to improve their activity [150].

Another recently proposed method for programmable silencing of gene expression in bacteria is CRISPR interference (CRISPRi). It uses an engineered clustered regularly interspaced palindromic repeats (CRISPR) pathway, where a customizable single guide RNA (sgRNA) forms a complex with a catalitically inactive Cas9 protein (dCas9) that can bind DNA [151, 152]. Differently from sRNAs, which act at post-transcriptional level, CRISPRi relies on transcriptional regulation by steric block of promoter binding or transcription elongation [151]. Activation of gene expression in bacteria has also been reported via this method, upon dCas9 protein engineering [153]. The regulation mechanism of CRISPRi is highly promising for genome-wide control of gene expression and it is complementary to post-transcriptional regulation elements, like sRNAs and asRNAs. CRISPRi is characterized by high expression modulation efficiency, it has been shown to work in many species, and guidelines for sgRNA rational design have been proposed [151, 152]; compared to post-transcriptional element, an intrinsic drawback of CRISPRi is that the selective repression of an individual gene in polycistronic transcript cannot be easily achieved [154].

Many bacterial genes are organized in operon architecture. The clustering of genes in operons is an important context in *E. coli* and other prokaryotic organisms, allowing to coordinately express proteins that are involved in common processes, while greatly facilitating the ability to efficiently respond to environmental changes. Because transcription and translation are physically coupled in prokaryotes, operons provide a highly efficient method of regulating the transfer of genetic information from DNA to protein [155]. Genes in polycistronic transcripts can be naturally targeted for repression by sRNAs. Intuitively, the silencing mechanism is important when targeting operon genes, since transcript degradation leads to the down-regulation of all the operon genes, while translation repression can specifically target individual genes within an operon. Specific *E. coli* operons were analyzed in different works [156, 139, 144, 157]. The *sdh*CDAB operon, involved in succinate metabolism, is targeted by the RyhB sRNA, which binds the transcript between the first and the second gene of the operon, resulting in mRNA degradation [139, 144]. Conversely, the *gal*ETKM operon, involved in galactose metabolism, is targeted by the Spot 42 sRNA, which binds the mRNA upstream of the third gene of the operon (*gal*K), but it does not result in transcript degradation. Only the *gal*K gene is down-regulated by Spot 42, which acts as a translator inhibitor [156]. Other, more complex, operon regulation mechanisms have recently been reported. For example, the iscRSUA transcript, involved in the Fe-S clusters biosynthesis, is targeted for degradation by the RyhB sRNA, but the first operon gene (*isc*R) is not down-regulated, thanks to a strong repetitive extragenic palindromic secondary structure (between *isc*R and *isc*S) which may protect the gene against ribonucleases degradation [157].

Motivated by the attractive features of sRNA-based control systems, after the discovery of natural sRNAs in bacteria many efforts were carried out to design synthetic sRNAs that can repress the desired target genes. Inspired by the natural architecture of the discovered bacterial sRNA, the synthetic sRNA are composed of two functional parts: a target-binding sequence and a scaffold sequence [133, 158, 159]. The first part is a sequence complementary to the TIR of its target mRNA, which specifically binds to the target and actuates the gene silencing. The scaffold sequence recruits the RNA chaperone Hfq, a highly abundant protein that facilitates the binding of the sRNA to the target mRNA at a much faster rate than that of the binding of ribosomes [159].

However, owing to a lack of full understanding of the sRNA silencing mechanism in prokaryotic, the first studies on synthetic sRNA design mainly focused on random screening methods [133, 158]. Man et al. [133] developed a semi-rational strategy for sRNA design based on the sequence of well-known trans-encoded *E. coli* sRNAs.The target-binding sequence was complementary to the 5'-UTR of the target mRNA and then appropriately adjusted to have a secondary structure with least two stem loops. The Hfq-binding scaffold sequence and the transcriptional terminator were extracted from a list of well-studied endogenous sRNAs and randomly combined to the target-binding part. These candidate sRNAs were finally filtered according to their secondary structure and a shorter list of candidates was obtained. This method resulted in the design of successful sRNAs, although out of the 16 initially selected candidates, only two repressed the target gene expression by 70% or more.

Sharma et al. [158] developed a screening strategy that can identify synthetic sRNAs capable of regulating endogenous genes. They constructed a large library of artificial sRNAs by fusing a randomized antisense domain to a scaffold sequence from four natural sRNAs that interact with the Hfq protein. In order to select the sRNA actually targeting the mRNA of interest, expression plasmids including the random sRNA library were co-transformed with and expression plasmid including a fluorescent reporter gene (GFP) fused with the 5' leader sequence of the mRNA of interest. Fluorescence detection by visual inspection of transformation plates identified the colonies containing the desired sRNAs. The described method enabled the obtainment of sRNAs that repressed the *omp*F target gene by 45- to 145-fold, but the approach required the screening of a large number of clones ($>10^5$) and was characterized by a low probability to find a clone where fluorescence was repressed (0.03%).

Guidelines for the rational design of customized sRNA were recently proposed by Na et al. [159]. The authors used reporter genes to test different features of sRNA expression systems, by investigating the repression capability as a function of different scaffold sequences, hybridization energy, binding position of sRNA within the transcript and target-binding sequence length. From their investigations, they selected the MicC [145, 160] sequence as the best scaffold among four candidates, 24 nucleotides as the optimal length of target-binding sequence and an hybridization energy lower than -20 kcal/mol. In their work, the proposed guidelines were successfully applied to metabolic engineering, demonstrating that complex pathways can be optimized via large libraries of rationally designed sRNAs and such strategy can be easily adopted to search the best producer among a collection of candidate *E. coli* strains.

The aim of this work is to quantitatively evaluate the performance of synthetic sRNAs designed with guidelines proposed by Na and colleagues. We designed the sRFP silencer, which represses the expression of reporter target gene RFP, encoding the Red Fluorescent Protein, to evaluate its performance on different ad-hoc constructed model systems, in two *E. coli* strains, as a function of sRNA and mRNA levels, also with the help of mathematical modelling for data interpretation. Since the operon context has never been quantitatively tested before using rationally designed sRNAs [161], in this work we studied the down-regulation of a target gene in a synthetic operon. Finally, we present data on the silencing of an endogenous gene, ldhA, which has a crucial role in the fermentation pathway of *E. coli* and in metabolic engineering studies, by means of another rationally designed sRNA. The quantitative study performed in this work elucidated interesting performance-related and context-dependent features of synthetic sRNAs that have never been investigated before. The obtained results and data will strongly support predictable gene silencing in disparate basic or applied research studies via novel designed sRNAs.

4.2 Materials and methods

4.2.1 Strains and plasmids

The *E. coli* TOP10 (Invitrogen) strain was used for cloning. The TOP10 and W (ATCC 9637; [162]) strains were used for quantitative experiments.
We designed the sRFP, sLDH, sACK, sFRD and sPFL synthetic sRNAs, targeting the RFP, *ldh*A, *ack*A, *frd*A and *pfl*B genes, respectively, according to the guidelines proposed by [159]. The target-binding sequence was designed as the reverse complement of the first 24 bp of the coding sequence included in the target mRNA. The hybridization energy of the target-binding sequence was computed via the UNAfold web server ([163]; [164]), to verify that it was lower than -20 kcal/mol. The MicC scaffold sequence was included downstream of the target-binding sequence to obtain the final sRNA. This sRNA sequence is placed between the strong promoter P_R upstream and the T1/TE transcriptional terminator downstream. The genomic sequences of the DH10B (NC_010473.1), closely related to TOP10 and with the same genotype, and the W (NC_017664.1) strains in the NCBI database were used to retrieve the *ldh*A, *ack*A, *frd*A and *pfl*B gene sequences. All the genes had identical nucleotides in the initial 24 bp of their coding sequences between the two strains. The sequence of the RFP gene was retrieved from the BBa_E1010 entry in the MIT Registry of Standard Biological Parts (Registry) ([65]; [61]).
The sRNA expression cassettes were de-novo synthesized by the GenScript gene synthesis service (Piscataway, NJ, USA). They were designed with the standard Bio-Brick™prefix upstream and suffix downstream [61] to facilitate their transfer in different plasmid vectors. These cassettes were delivered in the pUC57-Simple shipping vector and they were subsequently transferred, upon EcoRI/PstI digestion, both into the pSB3K3 and into the pSB1A2 BioBrick™vectors [165].
All the other parts were either physically retrieved from the Registry DNA Distribution 2008-2011 or assembled in this study from existing BioBrick™parts,

by using the BioBrick™Standard Assembly procedure [61].

All the strains were grown in 5 ml of L-broth (LB; [166]) at 37 °C, 220 rpm. When required, ampicillin (100 mg/l), kanamycin (20 mg/l) and chloramphenicol (12.5 mg/l) were added to cultures to maintain plasmids. Long-term glycerol stocks, stored at -80 °C, were prepared for all the recombinant strains by mixing 750 μl of bacterial culture and 250 μl of sterile 80% glycerol. Plasmids were extracted from overnight cultures through the NucleoSpin Plasmid kit (Macherey-Nagel). DNA was digested as appropriate, with the EcoRI/XbaI/SpeI/PstI enzymes, and the fragments of interest were extracted from 1% agarose gel by NucleoSpin Extract II kit (Macherey-Nagel) before proceeding with ligation. DNA-modifying enzymes were purchased from Roche Diagnostics and used according to manufacturer's instructions.

We constructed model systems to quantitatively evaluate the performance and the specificity of the synthetic sRNAs (see Fig. 4.1). In particular, these systems include synthetic circuits expressing RFP and/or GFP in single gene or operon, which are driven by promoters and RBSs with different strength. All these constructs were placed in the pSB4C5 low-copy plasmid and they were co-transformed with an sRNA expression cassette, placed in the pSB3K3, pSB1A2 or pUC57-Simple plasmid. This expression systems design allows to study genes transcribed/translated at different levels (through promoter/RBS/inducer concentration changes) in combination with sRNAs expressed at different levels (through plasmid copy number changes). A similar experimental design, including a two-plasmid expression system for reporter gene and silencer, respectively, has been used by Levine et al. [142] and Lavi-Itzkovitz et al. [147] to characterize the effects of transcription, translation and RNA degradation parameters change. We use copy number change to tune the sRNA level in order to reproduce the same expression system design proposed by Na et al. [159], which proved to be successful as sRNA production cassette. The J101-R (Fig. 4.1A) and J101-R32 (Fig. 4.1B) constructs have the same constitutive promoter (BBa_J23101) upstream of RFP, but different RBSs (the BBa_B0034 RBS is stronger than BBa_B0032 when placed upstream of RFP; [167]). The Plux-R (Fig. 4.1C) circuit contains RFP driven by the P_{lux} inducible promoter. P_{lux} in the induced state is about eightfold stronger than BBa_J23101 [168]. These circuits allow to characterize RFP silencing as a function of mRNA level (J101-R and Plux-R constructs) and RBS strength (J101-R and J101-R32 constructs).

We studied the unspecific silencing by comparing the output of the Plux-R and Plux-G (Fig. 4.1D) circuits, where the latter includes an inducible expression

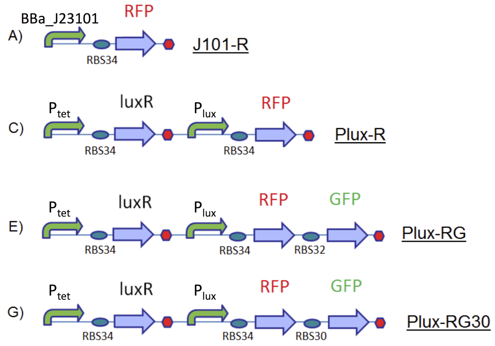


Figure 4.1: **Synthetic circuits used in this study. The BioBrick™codes are reported in brackets and can be found in the Registry of Standard Biological Parts with their nucleotide sequences.** A) J101-R (BBa_J107029): single-gene cassette for the expression of RFP driven by the constitutive BBa_J23101 promoter, with the BBa_B0034 RBS upstream of the RFP coding sequence. B) J101-R32 (BBa_K516132): single-gene cassette for the expression of RFP driven by the constitutive BBa_J23101 promoter, with the BBa_B0032 RBS upstream of the RFP coding sequence. C) Plux-R (BBa_J107032): single-gene cassette for the expression of RFP driven by the inducible P_{lux} promoter, with the BBa_B0034 RBS upstream of the RFP coding sequence. D) Plux-G (BBa_T9002): single-gene cassette for the expression of GFP driven by the P_{lux} promoter, with the BBa_B0032 RBS upstream of the GFP coding sequence. E) Plux-RG (BBa_J107042): RFP-GFP operon driven by the P_{lux} promoter, with the BBa_B0034 and BBa_B0032 RBSs upstream of RFP and GFP, respectively. F) Plux-GR (BBa_J107043): GFP-RFP operon driven by the P_{lux} promoter, with the BBa_B0032 and BBa_B0034 RBSs upstream of GFP and RFP, respectively. G) Plux-RG30 (BBa_J107044): RFP-GFP operon driven by the P_{lux} promoter, with the BBa_B0034 and BBa_B0030 RBSs upstream of RFP and GFP, respectively. H) Plux-G30R (BBa_J107045): GFP-RFP operon driven by the P_{lux} promoter, with the BBa_B0030 and BBa_B0034 RBSs upstream of GFP and RFP, respectively. All the described constructs are present in the pSB4C5 low-copy vector. Curved arrows represent promoters, ovals represent RBSs, straight arrows represent genes and hexagons represent transcriptional terminators. RBS34, RBS32 and RBS30 are the BBa_B0034, BBa_B0032 and BBa_B0030 BioBrick™RBSs.

cassette for GFP, which is not targeted by the designed sRNAs. We also studied unspecific silencing by characterizing the output of the above constructs in presence of sLDH, sACK, sFRD or sPFL, which do not target RFP and GFP. The RFP-GFP and GFP-RFP operons (circuits Plux-RG, Plux-GR, Plux-RG30 and Plux-G30R, see Fig. 4.1E-H) allow the study of the specific and unspecific gene silencing in polycistronic mRNA by red and green fluorescence quantification. The RFP and GFP genes of these circuits are in different positions and the GFP gene is placed under two different RBSs (the BBa_B0030 RBS is stronger than BBa_B0032 when placed upstream of GFP).
Transformation was carried out in TOP10 and W by heat shock at 42 °C. Both the ampicillin-resistant high-copy plasmids pSB1A2 and pUC57-Simple have a ColE1-based replication origin, but it has a single nucleotide mismatch (according to their sequence in the Registry and in the provided GenScript document, respectively), which could contribute to a different copy number. Finally, the pSB3K3 plasmid has a p15A replication origin and the pSB4C5 plasmid has a pSC101 origin [165].

4.2.2 Fluorescence assays

Recombinant strains were grown in 2-ml tubes at 37 °C, 220 rpm for 16-20 h in 0.5 ml of M9 supplemented medium (11.28 g/l M9 salts-M6030, Sigma Aldrich, 2 mM MgSO4, 0.1 mM CaCl2, 2 g/l casamino acids, 1 mM thiamine hydrochloride and 4 ml/l glycerol; [166]) with antibiotics as required, inoculated with a single colony from a streaked selective LB-agar plate (at least 3 independent biological replicates were carried out for each recombinant strain). The grown cultures were 100-fold diluted in 200 μl of M9 in a 96-well microplate. When required, 2 μl of the N-3-oxohexanoyl-L-homoserine lactone (HSL) inducer (K3007, Sigma Aldrich) were added to reach the desired final concentration. Unless differently indicated, 100 nM of HSL were used to induce the P_{lux} promoter. The microplate was incubated at 37 °C in the Infinte F200 reader (Tecan) and the following kinetic cycle, programmed via the i-control software (Tecan), was carried out: linear shaking 15 s (amplitude 3 mm), wait 5 s, absorbance measurement (600 nm), fluorescence measurement (excitation 485 nm, emission 540 nm for GFP; excitation 535 nm, emission 620 nm for RFP, gain of 50 or 80), sampling time 5 min [167, 169].

4.2.3 Data analysis for fluorescence assays

Matlab R2010a (MathWorks) and Microsoft Excel were used to analyze the absorbance and fluorescence time series to obtain doubling time and average RFP or GFP synthesis rate per cell (S_{cell}; [170, 171]). S_{cell} is expressed in arbitrary units (AU), proportional to the average per-cell protein synthesis rate. In each experiment, the absorbance of M9 without bacteria (background absorbance) and the fluorescence of the TOP10 and W strains without reporter genes (background fluorescence) were measured. The background absorbance time series was subtracted from the absorbance of each culture of interest to obtain a time series (OD_{600}) proportional to bacterial cell density (see Fig. A.1A-B; [170, 171]). Similarly, the RFP fluorescence background (which is not characterized by a relevant OD_{600}-dependent autofluorescence, see Fig. A.1B-C) time series was substracted from the raw RFP fluorescence of each culture to yield a time series proportional to the total RFP proteins in the microplate well. Since GFP shows a relevant OD_{600}-dependent autofluorescence (see Fig. A.1D), a different background subtraction procedure was carried out: a standard curve was obtained by fitting GFP background fluorescence against OD_{600} via linear regression for each of the two strains (see Fig. A.1E; [172]); the fitted standard curve was used to subtract GFP background fluorescence from the raw GFP fluorescence of each culture at the same OD_{600}, yielding a time series proportional to the total GFP proteins in the microplate well [172]. Raw and background-subtracted absorbance and fluorescence data are shown in Fig. A.1F-I (for RFP-expressing cultures) and in Fig. A.1K-N (for GFP-expressing cultures). The slope of the $\ln(OD_{600})$ time series in the OD600 range 0.05-0.18 (exponential growth phase) was computed, via linear regression, to calculate the cell growth rate. Doubling time was computed as ln(2) divided by the slope. A signal proportional to the RFP or GFP synthesis rate per cell was computed as the numerictime derivative of RFP or GFP time series, divided by OD_{600} (see Fig. A.1J and O for representative data of RFP- and GFP-expressing cultures, respectively). This signal was averaged over the exponential growth phase and the obtained value was divided by the average synthesis rate per cell of a reference culture expressing RFP or GFP, to compute S_{cell}. RFP and GFP reference cultures were recombinant strains (TOP10 or W) bearing an RFP (BBa_I13507) and a GFP (BBa_E0240) expression system under the control of the constitutive BBa_J23101 promoter, in pSB4C5. The silencing capability (Eff%) for a given gene in each of the above illustrated constructs was computed as reported in Eq. 4.1.

$$Eff\% = 100 * \left(1 - \frac{S_{cell\ \ with\ \ silencer}}{S_{cell\ \ without\ \ silencer}}\right) \tag{4.1}$$

However, in case S_{cell} with silencer was higher than without silencer, Eff% was set to zero. Assuming that pSB4C5 is present at 5 copies per cell, the per-cell copy numbers of pSB3K3, pSB1A2 and pUC57-Simple in the TOP10 and W strains were estimated as reported in Eqs. 4.2, 4.3, 4.4 [173, 168].

$$Copy\ number\ pSB3K3 = \left(\frac{5}{S_{cell_{4C5}}}\right) * S_{cell_{3K3}} \tag{4.2}$$

$$Copy\ number\ pSB1A2 = \left(\frac{5}{S_{cell_{4C5}}}\right) * S_{cell_{1A2}} \tag{4.3}$$

$$Copy\ number\ pUC57_{Simple} = \left(\frac{5}{S_{cell_{4C5}}}\right) * S_{cell_{pUC}} \tag{4.4}$$

where $S_{cell_{4C5}}$, $S_{cell_{3K3}}$, $S_{cell_{1A2}}$ and S_{cell} values of cultures bearing the J101-R construct (Fig. 4.1A) in the pSB4C5, pSB3K3, pSB1A2 and pUC57-Simple vectors, respectively, assuming that no metabolic overload affects cells at the highest copy numbers [168].

4.2.4 Lactate dehydrogenase assay

The activity of lactate dehydrogenase (LdhA) was determined by a specific enzymatic assay. LdhA catalyzes the conversion of pyruvate and NADH to lactate and NAD^+, respectively. The decrease of NADH concentration is measured by absorbance (340 nm) in order to compute the reaction rate, which is proportional to the enzyme concentration in the sample [174].
2 ml of LB with 100 mM phosphate buffer and 40 g/l of glucose were inoculated with 5 μl of recombinant bacteria from a glycerol stock and incubated at 37 °C, 220 rpm for 16-20 h. The grown cultures were 100-fold diluted in 9 ml of the same medium and incubated as before for 4 h. One ml of sample was taken, centrifuged at 13,000 rpm for 1 min and the supernatant was removed. The bacterial pellet was resuspended with 1 ml of Tris-HCl 100 mM pH 7.3, the vial was centrifuged and the supernatant discarded as before. The CelLytic B (Sigma Aldrich) lysis buffer, supplemented with protease inhibitor cocktail, was used to resuspend the pellet, and the vial was incubated at room temperature for 10 min under slow shaking conditions. Cell debris were separated by centrifugation (13,000 rpm, 5 min). 20 μl of supernatant, which includes the

intracellular content, were transferred into the well of a microplate and mixed with 180 ll of a solution containing Tris-HCl pH 7.3 100 mM, NADH 0.4 mM and sodium pyruvate 10 mM. The 96-well microplate was incubated at 25 °C in the Infinite F200 reader and absorbance (340 nm) was read every 5 min. The enzymatic activity of the sample in the well was computed by linear regression of the absorbance time series. Since this activity depends from cell lysis efficiency and initial amount of cells in the sample, we computed the specific enzymatic activity by dividing the enzymatic activity of the sample by the milligrams of total proteins extracted during lysis, quantified with the Micro BCA Protein Assay Kit (Thermo Scientific). The specific activity of all the bacterial cultures analyzed is divided by the wild type activity.

4.2.5 Statistical tests

Statistical analysis was performed on S_{cell} values via the Kruskal-Wallis (KW) nonparametric test to evaluate the statistical significance of repression in the assayed conditions. When the KW test detects at least a significantly different S_{cell} value (p value <0.05) among groups, the least significant difference (LSD) method was used to evaluate the significantly different conditions by multiple comparisons. We implemented the test by the *kruskalwallis* Matlab function. In such multiple comparisons, we focused on the significance of silencing (specific or non-specific) of each condition compared to the recombinant strain without sRNA. For this reason, we only evaluated the contexts where S_{cell} was lower than the reference context, by one-sided test. An analogous procedure was used to analyze the statistical significance of specific LdhA enzymatic activity among the tested contexts.

4.2.6 Mathematical modelling

The kinetic model of Eqs. 4.5, 4.6, 4.7, 4.8, 4.9 was considered [142, 146, 175, 170] and a summary of species and parameters is reported in Table 4.1.

$$\frac{dm}{dt} = \alpha_m - \beta_m \cdot m - k_+ \cdot s \cdot m + k_- \cdot c \tag{4.5}$$

$$\frac{ds}{dt} = \alpha_s - \beta_s \cdot m - k_+ \cdot s \cdot m + k_- \cdot c \tag{4.6}$$

$$\frac{dc}{dt} = k_+ \cdot s \cdot m + k_- \cdot c - \beta_c \cdot c \tag{4.7}$$

$$\frac{di}{dt} = \rho \cdot m - (z + \mu) \cdot i \tag{4.8}$$

$$\frac{dr}{dt} = z \cdot i - \mu \cdot r \tag{4.9}$$

The equations above describe the transcription process of the target mRNA (m) and of the sRNA (s), assuming constant transcription rates (α_m and α_s, respectively) and linear degradation rates (β_m and β_s, respectively). The c state variable represents the mRNA-sRNA complex, which is formed upon m and s interaction with kinetic constant k_+; the complex releases m and s with kinetic constant k_- and its degradation rate is linear β_c. The immature (i.e., non-fluorescent; [170]) protein (i) synthesis process is described by a linear production term ($\rho \cdot m$, where ρ is the translation rate per mRNA unit) and a linear extinction rate ($z + \mu$, where z is the maturation rate to yield the fluorescent form, and μ is the cell growth rate which represents the protein dilution due to cell division). The last equation describes protein maturation, to yield the fluorescent form r of the reporter protein. The described model assumes that RNA degradation rate is much faster than cell growth rate, while protein degradation rate is negligible and cell division is the only responsible of the intracellular protein extinction rate.
Considering the steady-state of the system, $S_{cell} = q \cdot z \cdot \overline{i}$ (where the bar indicates the steady-state) is the experimentally observable variable [175, 176], already defined above, where q represents the unit conversion constant between the actual protein synthesis rate per cell and the S_{cell} values (in AU) obtained in the experiments described above. S_{cell} is also proportional to the steady-state mRNA level: $S_{cell} = b \cdot \overline{m}$, where $b = \frac{q \cdot z \cdot \rho}{z+\mu}$.

The solution of the system, which is S_{cell}, can be analytically computed as [142]:

$$S_{cell} = \frac{1}{2} \cdot \left(a - a_s - a_\lambda + \sqrt{(a_s + a_\lambda - a)^2 + 4 \cdot a \cdot a_\lambda} \right) \tag{4.10}$$

This equation describes the Scell value of an RFP expression system in presence of the sRFP silencer. In this equation, $a = \frac{b \cdot \alpha_m}{\beta_m}$ is proportional to the transcription rate of the target mRNA and it is identical to the S_{cell} value in absence of sRNA, $a_s = \frac{b \cdot \alpha_s}{\beta_m}$ is proportional to the transcription rate of sRNA, and $a_\lambda = \frac{b \cdot \lambda}{\beta_m}$, where $\lambda = \frac{\beta_m \cdot \beta_s}{k}$ has been previously defined as the leakage rate (since its value affects the threshold-linear response of the target mRNA, as a function of its transcription rate for a given sRNA level value; [142]) and $k = \frac{\beta_c \cdot k_+}{k_- + \beta_c}$.

Species or parameter	**Description**	**Units**
m	mRNA per-cell level	Molecules
s	sRNA per-cell level	Molecules
c	mRNA-sRNA complex per-cell level	Molecules
i	Immature reporter protein per-cell level	Molecules
r	Mature reporter protein per-cell level	Molecules
α_m	mRNA transcription rate per cell	Molecules $time^{-1}$
α_s	sRNA transcription rate per cell	Molecules $time^{-1}$
β_m	mRNA degradation rate	$time^{-1}$
β_s	sRNA degradation rate	$time^{-1}$
k_+	Kinetic constant for mRNA and sRNA association to form the mRNA-sRNA complex	$Molecules^{-1}$ $time^{-1}$
k_-	Kinetic constant for mRNA-sRNA complex dissociation	$time^{-1}$
β_c	mRNA-sRNA complex degradation rate	$time^{-1}$
ρ	Protein translation rate per RNA unit per cell	$time^{-1}$
z	Protein maturation rate	$time^{-1}$
μ	Cell growth rate	$time^{-1}$

Table 4.1: **Description of kinetic model species and parameters.** Parameters refer to the ordinary differential equation model (Eqs. 4.5, 4.6, 4.7, 4.7, 4.8, 4.9).

The analytical solution of the model allows to study the Scell output value, at the steady-state, of systems including a target gene (RFP) and a specific sRNA (sRFP) for different RFP (a) and sRFP levels (a_s). In this study, the RFP level is tuned by inducing the P_{lux} promoter upstream of the RFP gene through different HSL concentrations (with the Plux-R construct), while the sRFP level is tuned by changing the copy number of the plasmid containing the sRNA expression cassette. Equation 4.10 was used to fit experimental data (S_{cell} value in the non-repressed condition in the x-axis, Scell value in the repressed conditions tested in the y-axis, for different HSL concentrations) via the *lsqnonlin* Matlab routine. A different as parameter value was estimated for each sRNA level tested, while a single ak parameter value for all the sRNA levels was estimated, as described in [142].

4.3 Results

4.3.1 RFP silencing in a single-gene cassette

The constructs with reporter target gene RFP driven by the P_{lux} inducible promoter (Plux-R, Figure 4.1C) were tested, both in the TOP10 and W strains, in presence of no silencer and the sRFP silencer in medium-copy (pSB3K3 vector) and high-copy (pSB1A2 and pUC57-Simple vectors) contexts. The RFP synthesis rate per cell was measured as systems output that reflects gene silencing. Such tests allowed the study of specific silencing as a function of intracellular concentration of sRNA, which is regulated by changing the intracellular copy number of the sRNA expression system.

Results, reported in Fig. 4.2A, show that the sRFP silencer works as expected in both strains, since RFP is repressed only in presence of its silencer. Statistical analysis of the S_{cell} values showed a significant difference between the condition with the silencer and the condition without the silencer, in both strains and in all conditions (p-value<0.05, Kruskal-Wallis and multiple comparisons). Figure 4.2A shows that the silencer represses RFP by up to 92% in TOP10 and 68% in the W strain. Repression values were systematically higher in TOP10 than in W, with sRFP in medium-copy giving the lowest Eff% value and the pUC context giving the highest value, for a given strain. Doubling times were similar among the tested conditions for each of the two strains (see Fig. A.2A).

The per-cell copy numbers of pSB3K3, pSB1A2 and pUC57-Simple in the TOP10 and W strains were estimated to investigate the difference of repres-

Vector	Replication origin	Copy number in TOP10 strain	Copy number in W stain
pSB4C5	pSC101	5 (fixed)	5 (fixed)
pSB3K3	p15A	26	12
pSB1A2	Mutated pMB1	39	16
pUC57-Simple	Mutated pMB1	52	28

Table 4.2: **Estimated copy numbers for the pSB4C5 lowcopy (from the literature), pSB3K3 medium-copy, pSB1A2 and pUC57-Simple high-copy vectors.** Both high-copy replication origins are noted as ”Mutated pMB1”, but their sequences are different in a single nucleotide mismatch, therefore they can be considered as different origins, since they can be characterized by a quantitatively different copy number.

sion values between strains and among the conditions of a given strain. Results are shown in Table 4.2.

The obtained copy number values for TOP10 strain are comparable to the values of the literature for similar laboratory strains [173, 177], whereas for the W strain data are not available. Copy numbers are systematically higher in the TOP10 strain than in the W strain. Such copy number values are highly correlated with the RFP repression efficiencies in the tested conditions with a Pearson correlation coefficient of 0.98. These results suggest that the silencer copy number is the main responsible of the repression efficiency variation among the tested sRFP plasmid contexts and strains. To evaluate if the obtained repression efficiencies were actually due to the specific action of sRFP, we studied the effect of different unspecific silencers on the target genes RFP and GFP driven by the Plux inducible promoter (Plux-R and Plux-G, Figure 4.1C-D), both in TOP10 and W. In particular, RFP or GFP repression was tested in presence of: 1) no silencer, 2) the highest-copy number vector without expression cassettes (pUC-RING), and 3) a set of silencers (sLDH, sACK, sFRD and sPFL, see section 4.2), in different copy numbers, designed to target specific genes involved in the E. coli fermentation pathway.

Results (Figure 4.2B) showed that unspecific silencers have a low repression capability towards both the RFP and GFP gene. In particular, the highest unspecific repression values of RFP (31% in the TOP10 strain and 26% in the W strain) were obtained when sRNA expression systems are in the high

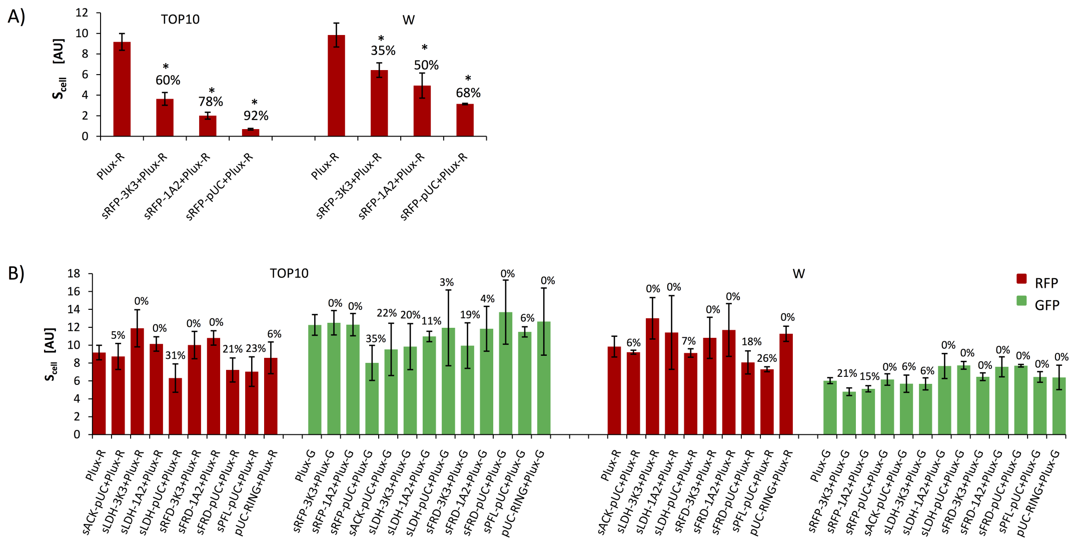


Figure 4.2: **Silencing results for RFP or GFP expressed by a single-gene cassette driven by P_{lux} (Plux-R or Plux-G).** A) Specific silencing of the target gene (RFP) via sRFP in TOP10 and W. B) Unspecific silencing of RFP or GFP via different sRNAs in TOP10 and W. Red and green bars correspond to RFP and GFP, respectively. Bars represent the mean Scell value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the Scell value in the condition is statistically different from the Scell of the expression cassette without sRNA (Plux-R and Plux-G conditions for RFP and GFP, respectively). Percentages represent the Eff% values.

copy number pUC57-Simple vector. On the other hand, the highest repression values observed for GFP were 35% for the TOP10 strain (sRFP silencer in the pUC57-Simple vector) and 21% for the W strain (sRFP silencer in the pSB1A2 vector). In general, the entity of the observed unspecific repressions is lower than the specific silencing percent values obtained above (compare Figure 4.2A and Figure 4.2B) and the highest values corresponded to conditions where silencer is placed in a high copy plasmid. The doubling times of the recombinant strains in the illustrated conditions did not significantly correlate with unspecific repression values (see Fig. A.2B) and are similar to the ones shown above for specific silencing (Fig. A.2A). Statistical analysis of the unspecific silencing data showed that no significant difference between the condition without sRNA and the conditions with plasmid-borne sRNA occurs.

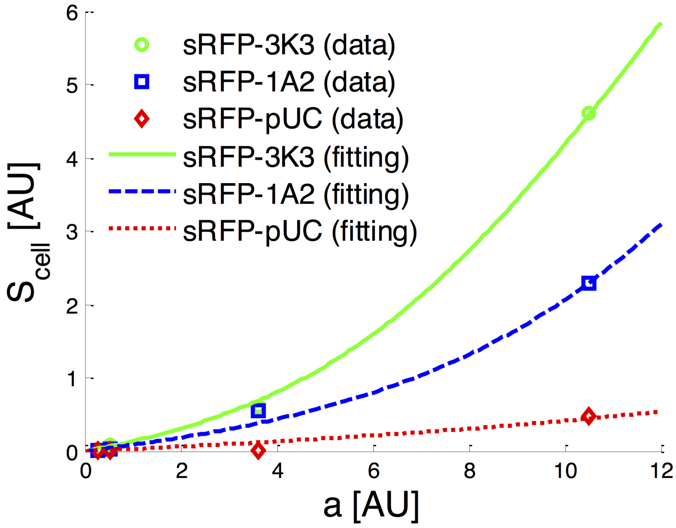


Figure 4.3: **RFP silencing as a function of target mRNA and sRNA levels in TOP10 with the Plux-R construct.** Data points represent average S_{cell} values in presence of sRFP in three different copy number conditions (corresponding to three different sRFP levels), as a function of a, corresponding to the average S_{cell} in absence of sRFP that is proportional to the mRNA level. Lines represent model fitting. Estimated parameters are reported in Table 4.3. Average S_{cell} values are computed on at least three biological replicates.

4.3.2 Model-based characterization of RFP silencing

A mathematical model, previously developed to describe silencing efficiency as a function of sRNA and mRNA levels, was used to support the characterization of the sRFP-dependent RFP gene repression. The sRFP level and the RFP transcript level were varied by means of plasmid copy number (as in previous section) and by tuning the P_{lux} promoter transcriptional activity, respectively. The resulting data are shown in Figure 4.3 and they were fitted with the steady-state solution of the kinetic model (Equation 4.10). The fitted curves showed that the model was able to describe RFP silencing according to different sRNA/mRNA levels, with the trend reported previously [142]. The estimated model parameters are reported in Table 4.3. The a_s parameters, corresponding to the sRNA levels for each of the three copy number conditions, were consistent with the estimated plasmid copy number reported in the previous section (see Table 4.2): the estimated a_s value in the pSB1A2 context was

Parameter	Description	Estimated value (AU)
a_λ	Leakage rate	1.0198
a_{s-3K3}	Transcription rate of sRNA placed in the high-copy vector (pSB3K3)	7.2256
a_{s-1A2}	Transcription rate of sRNA placed in the high-copy vector (pSB1A2)	11.8460
a_{s-pUC}	Transcription rate of sRNA placed in the high-copy vector (pUC-Simple)	33.1899

Table 4.3: **Definitions and estimated values of model parameters.** Parameters refer to the analytical solution of the model (Eq. 4.10).

1.6-fold higher than in the pSB3K3 context, consistent with the data of Table 4.2 where a 1.5-fold variation is observed. The estimated a_s value in the pUC context was higher than the values of the two other plasmids, as expected, but it was 2.8- and 4.6-fold higher than a_s in pSB1A2 and pSB3K3, while data of Table 4.2 showed a smaller fold-change (1.3 and 2, respectively). This could be due to saturation phenomena in the measurement of the copy number through RFP; in fact, protein expression may not change in a linear fashion at high per-cell copy numbers and measured values can be underestimated [168].

The obtained a_s and a_λ parameter units depend on our acquisition system (see section 4.2) and for this reason they are not immediately comparable with published values. In order to enable such comparisons, we computed the α_s (per DNA copy) and k values in absolute units as $\frac{RNAmolecules}{s}$ and $\frac{1}{nM \cdot min}$, respectively. We considered the a_s value in the medium copy context ($\alpha_s = 65$) and we assumed: a transcriptional activity of 0.03 $\frac{RNAmolecules}{s}$ per DNA copy for $a = 1$ (corresponding to the activity of the BBa_J23101 promoter; [170]), plasmid copy numbers of 5 for pSB4C5 and 26 for pSB3K3, an *E. coli* cell volume of 1 μm^3 [106], and a half-life of 6.8 min [178] for both mRNA and sRNA molecules.

We found an sRNA transcription rate of about 0.042 $\frac{RNAmolecules}{s}$, which is consistent with the activity of the P_R promoter, previously found to have an about 2.5-fold higher activity than the BBa_J23101 promoter [167]. We found a k value of 0.0007 $\frac{1}{nM \cdot min}$, which is about 30-fold lower than typical k values found in literature for naturally occurring regulatoryRNAs (0.02 $\frac{1}{nM \cdot min}$; [142].

This result highlights that, under the hypotheses above, the sRNA designed in this work following the guidelines of [159] resulted to be functional but with a lower binding rate, k, than observed in nature, thus showing a lower repression efficiency [142].

4.3.3 RFP silencing in different expression systems

While the previous sections focused on RFP repression when produced by a Plux-driven expression system, here we considered a different expression cassette for RFP, which is driven by the constitutive BBa_J23101 promoter (J101-R construct of Fig. 4.1A). As above, the sRFP-mediated silencing was tested in TOP10 and W in different copy number contexts and the RFP expression system was kept in a low-copy plasmid. Results, shown in Fig. 4.4A, were consistent with the ones obtained for the Plux-driven cassette (see Fig. 4.2A): silencing efficiencies were copy number- and strain-dependent, with systematically higher repression for TOP10 than W. Again, repression efficiencies were highly correlated with the estimated copy numbers of Table 4.2 (Pearson correlation coefficient of 0.96). Statistical analysis showed that RFP repression in most of the conditions with the sRFP silencer was significantly different from the sRNA-free condition. In principle, the tested condition is equivalent to setting the target mRNA to 1 (i.e., the S_{cell} value of the J101-R construct) in the mathematical model, while leaving all the other parameters unchanged, since mRNA and sRNA sequences were the same, and the tested copy number context and strains were identical. However, according to the prediction of the described mathematical model (i.e., the Scell values of the silenced systems as a function of mRNA level and for three different sRNA levels), reported in Fig. 4.3, higher repression efficiencies were expected for this mRNA level (see Fig. 4.4B). The observed differences could be due to the slightly different mRNA sequences between the BBa_J23101- and the Plux-driven expression cassettes; in fact, the transcription start site (TSS) of the two promoters is different [179, 180]. For this reason, the mRNA sequence of the BBa_J23101-driven mRNA has the ACTAGAG sequence upstream of the BBa_B0034 RBS, while the corresponding sequence for Plux has 4 additional nucleotides and it is AAATACTAGAG. Despite this structural difference, the local secondary structure free energy was predicted to be the same between the two sequences and the reasons determining the unexpected difference are unclear. Free energy was computed as described in [180], by analyzing the entire 50-UTR and 30 nucleotides of the RFP coding sequence, via the *UNAfold* web server [164].

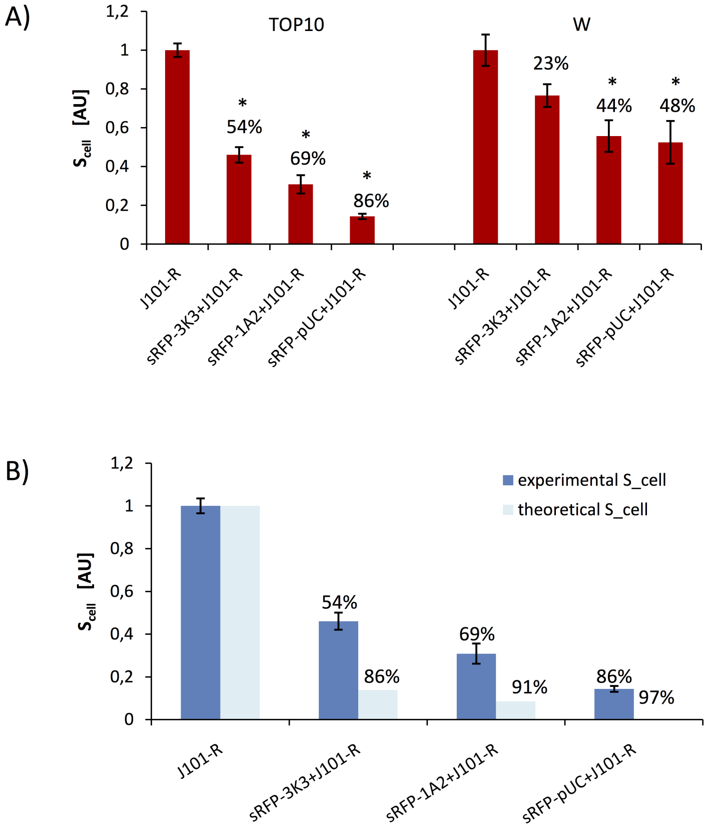


Figure 4.4: **Silencing results for RFP expressed by a single-gene cassette driven by BBa_J23101 (J101-R).** A) Specific silencing of the target gene (RFP) via the specific silencing device (sRFP) in TOP10 and W. B) Experimental S_{cell} results and values predicted by the mathematical model in the different copy number conditions of sRFP in TOP10. Bars represent the mean S_{cell} value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the S_{cell} value in the condition is statistically different from the S_{cell} of the expression cassette without sRNA (J101-R). Percentages represent the Eff% values.

Doubling times for the conditions tested above are reported in Fig. A.3 and are consistent with the ones obtained for the other tested systems (see Fig. A.2A and Fig. A.2B).
A similar RFP expression system (J101-R32 construct of Fig. 4.1B), which has the BBa_B0032 RBS upstream of RFP instead of the stronger BBa_B0034 RBS, was tested in the TOP10 strain. Results (see Fig. A.4) demonstrated that sRFP was also functional in a different RBS context and the quantitative repression values were similar to the ones obtained with J101-R. Since the RBS sequence has been recently shown to exert a complex effect, the tested context of J101-R32 could not be used to draw strong conclusions on the RBS-dependent functioning of rationally designed sRNAs and further investigations are needed.
The obtained results showed that an sRNA designed with the guidelines of [159] can work in several contexts (different promoters and RBSs for the target gene, and different strains), with qualitatively expected strain and copy number dependence, although the precise repression values could not be predicted, thus highlighting the need for additional studies and the importance of evaluating sRNA efficiency on different measurement constructs.

4.3.4 Silencing of a target gene in a synthetic operon

We used synthetic two-gene operons, including RFP and GFP under the control of the Plux promoter, to complete the characterization of the sRFP silencer. Specifically, the repression capability of the sRNA designed in this work was evaluated when targeting a specific gene present in a polycistronic transcript. The Plux-RG and Plux-GR constructs (Fig. 4.1E, 4.1F), in a low-copy plasmid, were used as model systems in the TOP10 and W strains, and the sRFP expression system, in pSB3K3, pSB1A2 or pUC57-Simple, was co-transformed. Red and green fluorescence signals were simultaneously quantified to study the protein synthesis rate for the target and non-target gene, respectively. The fluorescence acquisition system used in this study was previously characterized and a negligible crosstalk was found to occur between the red and green fluorescence signals [167]. Considering the TOP10 strain, as observed for single-gene cassettes, RFP repression correctly worked (although statistical differences were detected only for the pSB1A2 and pUC57-Simple contexts in both operon systems) and it was dependent on the copy number of the specific sRNA (see Fig. 4.5A). Such experiments showed that RFP repression occurred when the RFP target gene was present as both the first

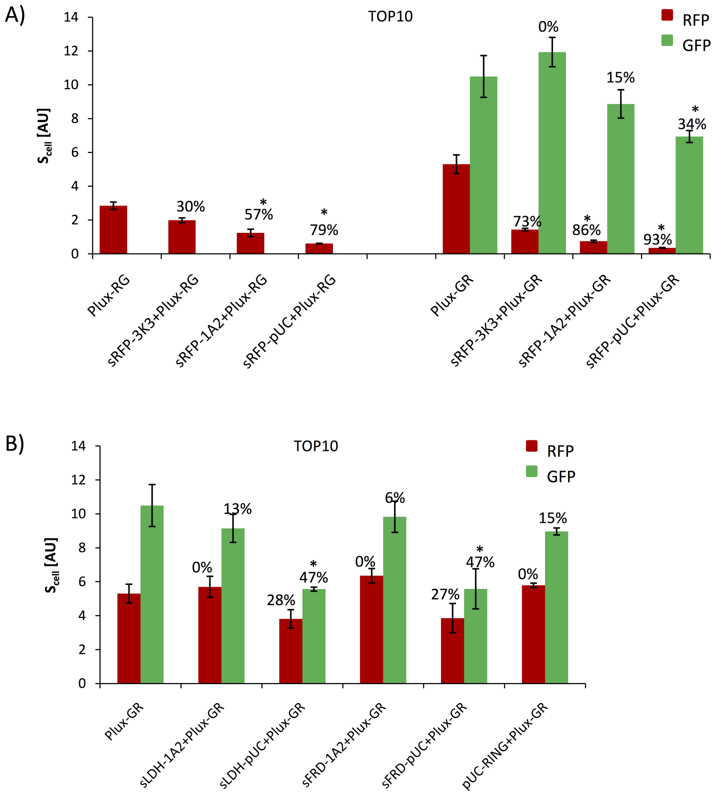


Figure 4.5: **Silencing results for RFP and GFP expressed by the Plux-RG and Plux-GR synthetic operons in the TOP10 strain.** A) Silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP, in the Plux-GR construct, via different sRNAs. Bars represent the mean S_{cell} value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the Scell value of RFP or GFP in the condition is statistically different from the Scell of the operon without sRNA (Plux-RG or Plux-GR). Percentages represent the Eff% values. When S_{cell} in a given condition is higher than Scell without sRNA, Eff% value is set to zero.

(Plux-RG construct) and the second (Plux-GR construct) operon gene. In particular, repression values were systematically higher when RFP was present in the second position. A direct comparison among repression values in operons and single-gene cassette contexts is not trivial to carry out, since the stability of mRNA molecules with different sequence can be different [181]. GFP was expected not to be repressed as the RFP target protein. Unfortunately, the GFP signal could not be detected for Plux-RG. The operon context is known to be highly unpredictable [182, 181], thus preventing synthetic biological systems designers to infer the translation efficiency of operon genes when their order is changed. GFP could be successfully detected when present as the first gene of the operon (Plux-GR construct). Results showed that an sRFP expression-dependent repression occurred for GFP, although it was much smaller than the one observed for RFP (see Fig. 4.5A). In particular, while RFP in Plux-GR was repressed by 73, 86 and 93% in the pSB3K3, pSB1A2 and pUC57-Simple contexts, a repression was observed for GFP in the pSB1A2 and pUC57-Simple conditions (15 and 34% respectively), although only the latter was found to be statistically significant.

As in the case of single-gene cassettes, the analysis of unspecific silencing was carried out by testing a set of sRNAs, in different plasmids, targeting genes that were different from RFP and GFP. The GFP-RFP operon (Plux-GR construct) was considered for the unspecific silencing study. Results (see Fig. 4.5B) showed that RFP production was repressed up to 28%, in the sLDH-pUC context, although a statistically significant RFP repression was not detected for any of the tested conditions. On the other hand, GFP was significantly repressed, up to 47%, in the two tested conditions where an sRNA (sLDH and sFRD) expression cassette was present in the pUC57-Simple vector.

The obtained results indicate that the RFP target gene is specifically repressed also in operon context, while the non-target gene in the operon mRNA was not affected. According to the unspecific silencing data (Fig. 4.5B), the observed repression of GFP was most probably due to the metabolic overload of the host strain, caused by the presence of two plasmids with an operon and an sRNA expression cassettes. The observation of a considerable repression of similar entity for both RFP (27-28%) and GFP (47%, statistically significant) by unspecific silencers only in the conditions in which sRNAs are expressed in the pUC57-Simple plasmid supports this statement. Doubling times, reported in Fig. A.5, were consistent with the ones reported for single-gene cassettes. Moreover, since RFP is efficiently repressed while GFP is not, the data suggest that the designed sRFP silencer acts as a repressor of protein synthesis, not increasing the mRNA decay rate like other natural sRNAs, although sRNAs can always affect target mRNA stability.

The GFP-RFP operon (Plux-GR construct) was also tested in the W strain in presence of the specific RFP target gene repressor, sRFP, in pSB3K3, pSB1A2 and pUC57-Simple, or with unspecific sRNAs. Specific silencing results (see Fig. A.6A) showed that RFP silencing in operon also works in W, with similar repression efficiency to the single-gene context (see Figs. 4.2A, 4.4A). Repression values of RFP were systematically lower than in TOP10 tested with the same plasmid (see Fig. 4.5A), as expected from the lower sRNA plasmid copy number in the W strain. While RFP is repressed in a copy number-dependent fashion, as expected, reaching up to 62% repression, GFP is never repressed by sRFP. GFP expression unexpectedly increased up to 1.7-fold when the operon was co-transformed with the sRFP-pUC context, compared to the operon without sRNA expression cassettes. Unspecific silencing experiments (see Fig. A.6B) showed that RFP and GFP were not repressed by sRNAs different from sRFP. However, these data showed a highly variable RFP and GFP expression, which were highly correlated. Importantly, doubling times analysis showed that in all the conditions with Plux-GR in W growth is clearly slower than in conditions with single-gene cassettes (see Fig. A.7). This slow growth occurs even when the operon was tested without sRNAs, demonstrating that the operon itself is responsible of the high doubling time and this was not due to the presence of a co-transformed sRNA expression cassette. This effect suggests a metabolic burden of recombinant strains with the operon. The highly variable RFP and GFP expression may be explained by the metabolic burden of the strains in such conditions, which could result in copy number variation of the medium- and high-copy number plasmids, as previously reported [183]. Overall, the results obtained in W were consistent with the ones obtained in TOP10 and confirmed the conclusions drawn above.

Finally, we attempted to overcome the GFP detection limit problem in the RFP-GFP operon (Plux-RG) by constructing and studying novel operons with a stronger RBS (BBa_B0030 instead of BBa_B0032) upstream of GFP (Plux-RG30 and Plux-G30R constructs of Fig. 4.1G, H). Unfortunately, they resulted in slow, highly variable doubling times (see Fig. A.8) and significant unspecific silencing (see Fig. A.9). For this reason, the obtained results cannot be considered to draw robust conclusions.

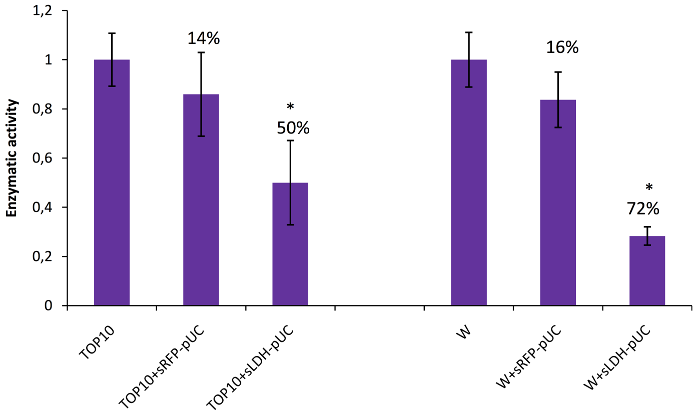


Figure 4.6: **Lactate dehydrogenase assay results for sLDH characterization in TOP10 and W.** Specific enzymatic activity of LdhA in the strain without sRNA, in the strain with an unspecific sRNA (sRFP) and in the strain with the sRNA targeting LdhA (sLDH). Bars represent the mean activity value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the value in the condition is statistically different from the value without sRNA (first bar). Percentages represent the Eff% values computed on activity values.

4.3.5 Silencing of the endogenous lactate dehydrogenase

One of the sRNAs used to evaluate the RFP and GFP unspecific silencing, sLDH, was used to study the specific silencing of the endogenous *ldh*A gene, encoding for a lactate dehydrogenase (LdhA) involved in the fermentation pathway of *E. coli*.
The change of LdhA activity in presence of the sLDH silencer was studied through enzymatic assay (see section 4.2) in the TOP10 and W strains. In this case, the sRFP silencer was used as unspecific sRNA to evaluate the nonspecific repression of LdhA activity. Both sLDH and sRFP were tested in the high-copy number pUC57-Simple vector. Results (see Figure 4.6) showed that sLDH significantly repressed LdhA activity, with 50 and 72% repression values in TOP10 and W, respectively. LdhA repression was very low and not statistically significant (14 and 16%, respectively) for TOP10 and W with the unspecific silencer sRFP. Such results demonstrated that the sLDH silencer, designed in this study according to the guidelines of [159], is functional. The

observed LdhA repression difference between the TOP10 and W strains was unexpected, since the pUC57-Simple plasmid is maintained at higher copy number in TOP10 than in W. However, differences in the *ldh*A gene expression and regulation between the two strains, not investigated in this work, can occur and might explain the observed repression values. Such differences can be due to differences in nucleotide sequences of the *ldh*A gene or promoter in the genomes of TOP10 and W, even if the 24-bp binding sequence of sLDH is identical.

4.4 Discussion

In this work, the silencing capability of sRNAs designed with recently proposed guidelines was characterized. In particular, key features that are typically investigated in quantitative studies on natural sRNA have been herein measured to evaluate the performance of synthetic sRNAs in several contexts and to enable the comparison with natural sRNAs. Since synthetic sRNAs designed with the guidelines of [159] have never been tested in different contexts, such as different target mRNA/sRNA levels, when the target gene is in operon architecture and in gene expression cassettes driven by different promoters [161], this study is of wide importance in the bottom-up design of artificial sRNAs. Our study is mainly focused on quantitative performance evaluation for a synthetic sRNA targeting a reporter gene, but data on another sRNA, targeting a gene of interest in metabolic engineering, are also reported.

A synthetic sRNA targeting the reporter gene RFP, here called sRFP, was designed and used in most of the performed experiments. sRFP was expressed at different levels by tuning the copy number of the plasmid bearing the sRNA expression system, while the target gene was produced via different constitutive or inducible expression systems. Statistical analysis was carried out for all the performed experiments to highlight the conditions where significant repression was present, compared to the recombinant strain without sRNA. Unspecific silencing analysis was also carried out to decouple specific gene silencing from other non-specific silencing mechanisms. To this aim, sRNAs designed to target genes different from RFP were used, as well as the GFP gene, which is not targeted by sRFP and is easily detectable.

When RFP was present in a single-gene expression cassette, driven by the Plux promoter at full induction, sRFP showed to work as expected in two different *E. coli* strains (TOP10 and W) and repressed red fluorescence in a copy

number-dependent fashion, reaching silencing levels up to 92%, in the highest copy number condition (pUC57- Simple plasmid in TOP10) of sRFP. The same construct (called Plux-R) was used to study RFP silencing for different mRNA (tuned via P_{lux} induction with HSL) and sRNA (tuned via plasmid copy number, as above) levels and results showed the expected mRNA-dependent trend [142]. By using a previously proposed kinetic model of target gene silencing with sRNA, we fitted the experimental data and estimated the binding affinity parameter of sRFP, under different assumptions and considering the Plux-R construct. It resulted to be about 30-fold lower than the one of RyhB, an extensively studied natural sRNA involved in iron metabolism in *E. coli*.
When RFP was constitutively expressed via a single-gene cassette driven by the BBa_J23101 promoter, sRFP also worked as expected in both strains, but the repression values were lower than the ones predicted by the mathematical model that was trained on data from Plux-R. Such observed deviation could be due to the different target mRNA sequence, which is 4 nucleotide longer in transcripts produced by P_{lux} than by BBa_J23101. Synthetic operons including the target RFP gene and a non-target reporter gene (GFP) were used to study gene silencing, via sRFP, in polycistronic transcripts. RFP repression successfully worked, in an sRNA copy number dependent fashion, and it was higher when RFP was present as the second operon gene. This result could be due to different mRNA decay rates of the two operons, which affects the steady-state level, or by the mRNA folding which differently exposes the binding sequence. RFP repression level reached values up to 93% in the highest copy number condition (pUC57-Simple plasmid in TOP10), while the non-target gene, GFP, was not considerably repressed. Experimental data of operon systems suggested that sRFP silencing acts only at the translation level, not by enhancing the decay rate of the whole transcript. GFP signal could not be detected in one of the tested operons (Plux-RG), while in the other operon (Plux-GR) specific and unspecific silencing could be fully studied. The latter operon was also successfully tested in the W strain, although it caused a slow growth for the host. Additional operons were also constructed (Plux-RG30 and Plux-G30R) with a stronger RBS upstream of GFP, than Plux-RF and Plux-GR. However, although both of them worked as expected, they resulted in slow growth and could not be used to draw sound conclusions, since unspecific silencing was considerable probably due to the metabolic burden exerted by the operon.
Overall, the obtained results on reporter genes demonstrated the importance of the target sequence that could affect gene silencing, and the difficulty of

characterizing gene silencing in operon via ad-hoc constructed model systems. Experimental data were also presented for a second sRNA, targeting the endogenous *ldh*A gene in the TOP10 and W strains. Repression efficiency was 50 and 72%, respectively, while unspecific silencing, via sRFP, did not result in significant repression values, thus confirming that another rationally designed silencer works as expected. In this case, the quantitative repression values obtained could be explained by measuring the expression level of the target *ldh*A gene.

The rational design of sRNAs with predictable performance is a key feature in synthetic biology and the guidelines proposed by Na et al. can be successfully used without relying on trial-and-error searches. Quantitative characterization studies, like the one proposed in this work, will strongly support the predictability of sRNA performances in different contexts. Although the measured variables can have an sRNA- and target gene-specific behaviour, the reported procedure and results support the future characterization of novel sRNAs, confirm the effectiveness of the design via the used guidelines, and elucidate quantitative performance-related and context-dependent features never investigated before for such synthetic silencers. Synthetic sRNAs will enable to face different problems in synthetic biology, such as the simultaneous silencing of different pathways in metabolic engineering studies [159], the silencing of essential or non-essential genes for bacterial physiology research studies [133, 184] and the tuning of synthetic circuits [159] to engineer repression systems with low-fluctuations or noise in the regulation of target proteins [142], which is an important feature to control cell-to-cell variability [185, 186].

Chapter 5

A BioBrick™-Compatible Vector for Allelic Replacement Using the *XylE* Gene as Selection Marker[1]

Once the reactions that compete for the production of the target chemical were selected for the elimination, their total inactivation can be achieved only by permanent deletion of encoding gene(s). Different experimental methods for the gene deletion have been proposed over the years.

A novel allelic replacement vector for chromosomal gene deletion in *E. coli*, based on the colorimetric XylE assay and the BioBrick standard, will be presented in this chapter and used for the disruption of three genes encoding the production of organic acids, that compete for pyruvate utilization in ethanologenic strains.

A general introduction of different genome engineering methods in *E. coli* and other bacteria and their common limitations will be described (Sec. 5.1). Then, the characteristics of the developed vector and the analysis of strain with *ldh*A gene deletion, encoding lactate dehydrogenase (LDH), will be reported (Sec. 5.2). The capacity and advantages of this BioBrick™-compatible vector, compared with other plasmid-based solutions will be discussed in Sec. 5.3.

This study has been published in [187], and subsequently this deletion tool

[1] The contents of this chapter are published in *Casanova M, Pasotti L, Zucca S, Politi N, Massaiu I, Calvio C, Cusella De Angelis MG and Magni P* Biological Procedures Online

has been used for the further disruption of *frd*AB and *pfl*B-*focA* genes, encoding the production of fumarate reductase (FRD) and pyruvate formatelyase (PFL), respectively.
The materials and methods sections and supplementary notes, results, figures and tables are reported in the App. B.

5.1 Background

A large number of methods, recently reviewed by Song et al. [188], are available for the efficient genome engineering of *E. coli* and other bacteria. Among them, circular plasmid-mediated homologous recombination is commonly used for marker-less allelic replacement, exploiting the endogenous recombination machinery of the host. In such method, a mutated version of the target locus is cloned in a conditional-replication plasmid, together with the two DNA sequences flanking it. Upon transformation, a first cross-over event integrates the plasmid in the target chromosomal region and a second one excises the integrated plasmid, leaving the allele with the desired modifications without any plasmid DNA sequences. While clones in which the first cross-over successfully occurs are easily selected via antibiotic resistance, the second cross-over is a rare event and clones that have lost the plasmid are usually screened via a counter-selection method [188]. Finally, the counter-selected clones, which have the same theoretical probability (50%) to contain the desired modified allele or to maintain the original state, need to be screened by PCR [189]. The counter-selection gene most widely used in this type of plasmids is *sacB*, which converts sucrose into a toxic product, thus enabling the selection of clones in growth media containing this sugar [190]. Apart from the requirement of specific media, a reported drawback of such popular method is the spontaneous mutation that can occur in *sac*B, resulting in false positive clones [191]. Other counter-selection methods available, such as those based on the *rps*L, *gal*K, *thy*A, *tet*A and *tol*C genes, also present strong strain and/or medium limitations [192, 193]. The I-*Sce*I counter-selection system has been proposed to overcome such issues [194], but false positive clones due to mutations can still occur at high frequency [195]. This is a common feature of synthetic kill switches implemented via toxic genes [196], although combination of multiple counter-selection systems has been reported to decrease the false positive rate [193]. Methods have been proposed that use temperature-sensitive vectors without toxic genes, exploiting the integrated replication origin to stimulate the second

recombination event in permissive (replicative) conditions [197]. This strategy, coupled with a *lac*Z gene-mediated blue/white screening, is successfully used in Gram-positive bacteria [198], although its use in *E. coli* would be limited to specific *lacZ*-mutant strains.
In this work, we propose a new vector (pBBknock, see Fig. 5A) for allelic replacement in *E. coli* that exploits a temperature-sensitive replication origin and the *xyl*E gene from *Pseudomonas putida*, coding for the catechol 2,3-dioxygenase enzyme [199]. This enzyme is not toxic for *E. coli* (data not shown) and converts the colourless substrate catechol into the yellow product 2-hydroxymuconic semialdehyde within seconds, resulting in a cheap and fast colorimetric assay to identify clones in which the second recombination event, i.e., plasmid excision, has not occurred. Although the *xylE* gene has previously been used as a reporter for gene expression in different microorganisms, such as *B. subtilis*, *Actinosynnema pretiosum* and *Streptomyces* spp [199, 200, 201] its application as selection marker in marker-less genome engineering protocols for *E. coli* represents a novel aspect of this work. XylE is encoded by a single 0.9-kbp gene and its activity can be detected without the requirement of specific strains or media. It was preferred over other available reporter systems for coloured product formation because the latter have less attractive features for pBBknock: violacein and carotenoid pathways are encoded by large multigenic constructs [202]; the single gene for melanin production requires specific medium formulation [203]. Finally, fluorescent reporters can be hard to detect when expressed from low or single DNA copies.

Since the development of standard genetic tools is one of the hallmarks of synthetic biology, strongly facilitating and speeding up the recombinant strain construction process [182, 204], we designed a vector that is compatible with commonly used BioBrick™standards (RFC10, RFC12 and RFC23) [65]. This novel plasmid for allelic replacement represents an advanced genetic tool in the ready-to-use BioBrick™-compatible vectors for genome engineering that have been recently proposed by our group [169], which, although enabling markerless genome engineering, still introduce plasmid-derived sequences surrounding the target locus.

5.2 Results

The pBBknock vector includes a pSC101ts temperature-sensitive origin (BBa_J107112) derived from pAH123 [205] [GenBank: AY048726] (see sec-

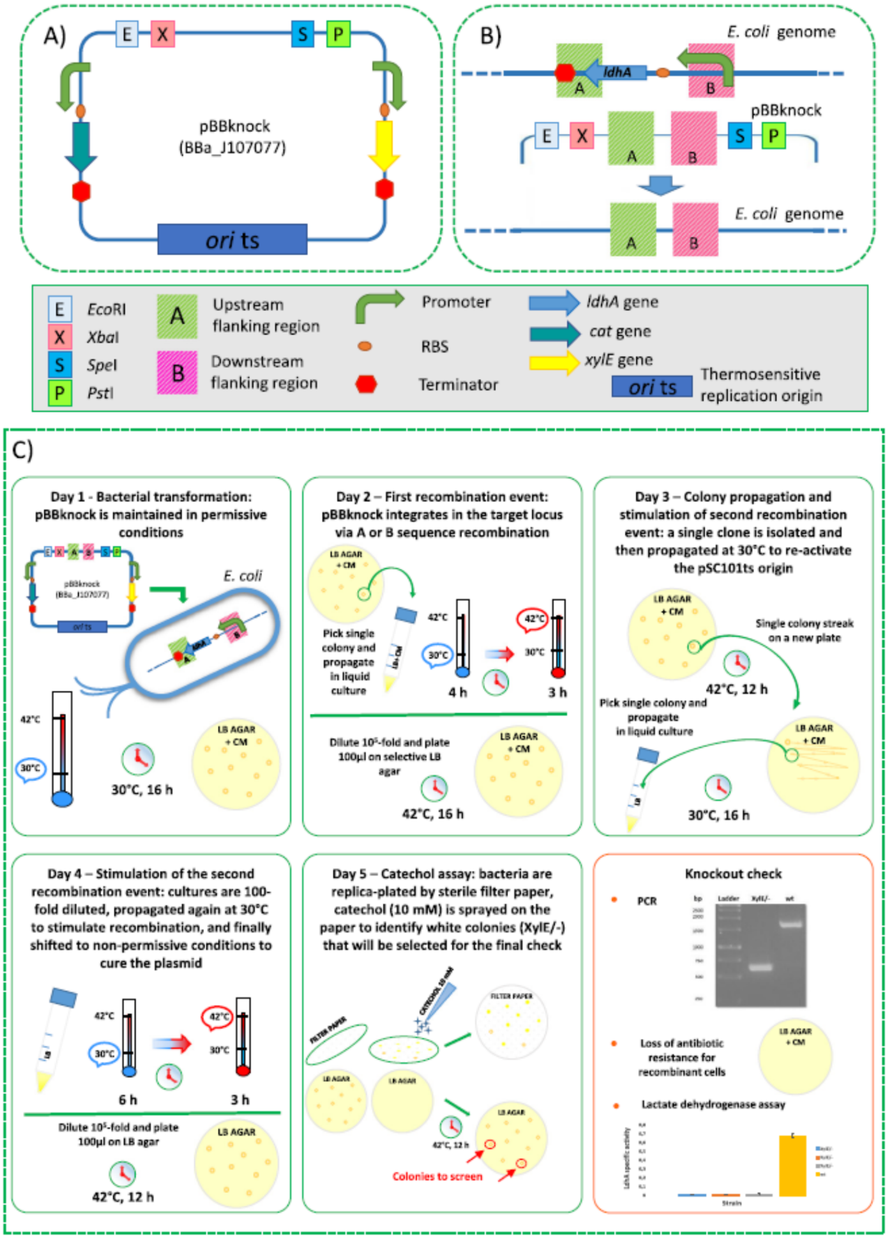

A)
E
X
S
P
pBBknock
(BBa_J107077)
ori ts
B)
E. coli genome
ldhA
A
B
pBBknock
E
X
A
B
S
P
E. coli genome
A
B
E EcoRI
X XbaI
S SpeI
P PstI
A Upstream flanking region
B Downstream flanking region
Promoter
RBS
Terminator
ldhA gene
cat gene
xylE gene
ori ts Thermosensitive replication origin
C)
Day 1 - Bacterial transformation: pBBknock is maintained in permissive conditions
pBBknock (BBa_J107077)
ori ts
E. coli
42°C
30°C
30°C, 16 h
LB AGAR + CM
Day 2 – First recombination event: pBBknock integrates in the target locus via A or B sequence recombination
LB AGAR + CM
Pick single colony and propagate in liquid culture
LB+CM
42°C
30°C
4 h
42°C
30°C
3 h
Dilute 10^5-fold and plate 100µl on selective LB agar
LB AGAR + CM
42°C, 16 h
Day 3 – Colony propagation and stimulation of second recombination event: a single clone is isolated and then propagated at 30°C to re-activate the pSC101ts origin
LB AGAR + CM
Single colony streak on a new plate
42°C, 12 h
Pick single colony and propagate in liquid culture
LB AGAR + CM
LB
30°C, 16 h
Day 4 – Stimulation of the second recombination event: cultures are 100-fold diluted, propagated again at 30°C to stimulate recombination, and finally shifted to non-permissive conditions to cure the plasmid
LB
42°C
30°C
6 h
42°C
30°C
3 h
Dilute 10^5-fold and plate 100µl on LB agar
LB AGAR
42°C, 12 h
Day 5 – Catechol assay: bacteria are replica-plated by sterile filter paper, catechol (10 mM) is sprayed on the paper to identify white colonies (XylE/-) that will be selected for the final check
FILTER PAPER
CATECHOL 10 mM
FILTER PAPER
LB AGAR
LB AGAR
LB AGAR
42°C, 12 h
Colonies to screen
Knockout check
PCR
bp
Ladder
XylE/-
wt
Loss of antibiotic resistance for recombinant cells
LB AGAR + CM
Lactate dehydrogenase assay
LdhA specific activity
Strain

Figure 5.1: **Description of the pBBknock vector, knockout experimental design and protocol.** A) Vector description; all the elements are described in the box below the panel. B) The AB DNA sequence is assembled in the pBBknock vector and the resulting plasmid is used to carry out chromosomal gene deletion via two successive recombination events, described in panel C). After the two recombination events, the resulting genomic target sequence is shown: it has about 50% probability to be successfully modified or to revert to the wild type state (not shown). C) Allele replacement protocol description. Notes on protocol development are reported in B.2.2

tion B.2.1). The vector also carries a chloramphenicol resistance cassette (BBa_P1004) including the *cat* gene with its own promoter and ribosome binding site (RBS) and the *xyl*E gene with its own RBS (BBa_J33204) under the control of the BBa_J23101 constitutive promoter [65]. BBa_J23101 is a medium-strength promoter that is widely used in synthetic biology studies and often serves as a standard reference in promoter characterization experiments [169, 170, 131, 186]. We used BBa_J23101 to drive the xylE expression in preliminary experiments in different strains and plasmid copy numbers and, according to catechol plate assay, the resulting expression cassette was functional and did not significantly reduce bacterial growth rate (data not shown). The L3S2P42 and L3S3P22 synthetic transcriptional terminators [206] are used downstream of the cat and *xyl*E cassette, respectively. Properly-placed unique *Eco*RI, *Xba*I, *Spe*I and *Pst*I restriction sites constitute the BioBrick™-compatible cloning site. The vector was fully constructed via the GenScript (Piscataway, NJ, USA) gene synthesis service.

The design specifications described above, including heterologous and synthetic components, allowed us to obtain a BioBrick™-compatible vector with a significantly low level of similarity to the *E. coli* genome, thus minimizing the off-target integration probability. The pBBknock sequence (see Fig. 5.1) can be accessed as BBa_J107077 in the Registry of Standard Biological Parts [65] and its DNA is available upon request.

As expected, the resulting vector replicates in *E. coli* at 30 °C and not at 42 °C. The copy number of pBBknock is very similar to the one of pSB4C5, demonstrating that in permissive conditions the pSC101ts origin is maintained at a copy number comparable with the one of a vector with the non-ts pSC101 low-copy number origin (see section B.2.3).

We used pBBknock to delete the lactate dehydrogenase (ldhA) gene in the chromosome of *E. coli* W, a widely used strain in metabolic engineering studies [162]. In particular, A and B sequences were designed, constructed and ligated to pBBknock to delete the chromosomal sequence comprised between the ldhA core promoter region (annotated in [EcoCyc: G592]) and the last 7 codons of the coding sequence (see Fig. 5.1B).

The process followed to achieve the gene knockout, inspired by Hamilton et al. [197] and Arnaud et al. [198], is described in Fig. 5.1C. Among 6 independent experiments, white colonies (i.e., with successful vector excision) ranged from 1% to 11% of the total colonies, with a 4% mean occurrence. Ten white clones were screened by colony PCR: three of them were successful knockout strains, while the others maintained the original allele (see Fig. 5.1C). Gene deletion was also confirmed by the absence of lactate dehydrogenase activity in the three ldhA- strains (see Fig. 5.1C).

5.3 Discussion

This work develops a novel allelic replacement vector, merging physical standardization concepts and a screening procedure based on a simple colorimetric assay, never applied before in marker-less allelic replacement methods for *E. coli*, that can be virtually used with any growth medium and host. The false positive rate is expected to be lower than in counter-selection systems based on toxic genes, which can frequently mutate (see section B.2.2). However, allelic replacement efficiency may vary in different strains and experiments, according to the host recombination capability, allele-dependent fitness, and flanking sequence length and homology [207]. Homologous sequences can be retrieved from a specific collection of BioBrick™parts [169] or can be easily constructed via PCR (as it was carried out in this work). BioBrick™parts can also be assembled between the two homologous DNA regions to be integrated in the target locus. Since pBBknock is replicated at low copy, it is particularly suited to deliver difficult parts (toxic when present in high copy) in the chromosome, for which other plasmid-based methods, e.g., the ones using the conditional R6K origin which is replicated at medium or high copy, may not be successful [192, 205]. Although novel promising techniques for large-scale genome editing have been developed [188], the modification of a single gene via the plasmid-based *sacB* method is still commonly carried out in many laboratories [208, 209, 210]. Efficient one-step methods based on linear DNA are also commonly used [188, 211], but they require a helper plasmid expressing specific recombinases and are applicable only to limited bacterial strains, since others might suffer from poor transformation efficiency with linear fragments. In this view, we expect that pBBknock will represent a versatile solution both for practitioners, also among the iGEM competition teams, and for research laboratories that use BioBrick™-based assembly procedures.

Chapter 6

Overall conclusions

The shift of production from petrochemical-based processes to bioprocesses is an important emerging challenge to establish a sustainable society. Metabolic engineering is the practice of optimizing cellular proprieties to obtain high production yield of high-value metabolites through analysis and rational manipulation of metabolic pathways, using genetic engineering techniques. Currently, it is obtained thanks to the development and advances of several computational tools for the in-silico design and sophisticated recombinant DNA techniques for the targeted genetic manipulation of microbial metabolic systems suitable for bioproduction.

In this thesis work, in-silico and in-vivo tools for metabolic engineering have been investigated, in *E. coli* and *B. subtilis*, at different levels. Starting from a preliminary study about the impact of experimental nutrient uptake rates on the predictions obtained with the constraint-based approach, different computational methods have been evaluated on metabolic engineering applications, considering several experimental datasets, and then applied to identify the target perturbations required for increasing the production of a high-value metabolite. Finally, experimental tools have been developed and characterized to implement the manipulations for obtaining the desirable metabolic phenotype, both at genomic and transcriptomic levels.

In Chapter 2 constraint-based methods have been analyzed in *E. coli*, under different genetic and environmental conditions, by means of comparison between the predicted and experimental flux distributions. In particular, since the predictions rely heavily on the flux constraints imposed in the metabolic model, defining the solution space, a preliminary study was focused on the impact of values used for constraining the main nutrient uptake rates. Moreover, the predictions of different engineered strains, under different growth conditions, have been evaluated and transcriptomic data have been integrated as an attempt to improve it. The carbon source uptake rate is the key constraint on which the predictions are based. The results demonstrated that, for wild type strains the knowledge of its experimental value improves the prediction performance, whereas for mutants it does not impact and the values measured for wild type can be used. FBA and MOMA methods showed limited prediction performance, both in terms of growth rate and production rate of main metabolite, when used to simulate *E. coli* strains subject to a large number of genetic deletions under anaerobic and micro-aerobic conditions. The same inaccurate results have been predicted also with the integration of transcriptomic data by GIMME method, which has been shown to be inefficient in complex contexts.

Recently, a new powerful tool, known as GECKO, for the integration of enzymatic data, in terms of k_{cat} values and protein abundance, has been proposed to further reduce the space of allowable solutions and improve the predictions. This method has not been applied for improving the former predictions, because the proteomic data for *E. coli* in anaerobic or micro-aerobic conditions are not currently available in literature. Chapter 3 describes the development and evaluation of an enzyme-constrained model for *B. subtilis*, based on the principles of GECKO. Given the small number of available k_{cat} measures, this integration has been focused on a small set of enzymes, namely for central carbon pathway and some connected reactions of glutamate pathway. Through this new model, an increase of accuracy has been demonstrated considering both quantitative predictions for wild type and mutant strains, especially for the fluxes of pentose phospate reactions and the acetate production rate, and in terms of gene essentiality predictions. However, a greater improvement can be achieved with a genome-scale integration of enzymatic data into the model. The increased accuracy shown for the predicted fluxes, has been obtained also for the identification of modifications required to improve the target metabolite production. Indeed, the developed enzyme-constrained model of *B. subtilis*, after properly modifications, has subsequently been applied as engineering guide for improving the production of an emerging polyvalent natural product, poly-γ-glutamic, whose enzyme-encoding genes are not expressed in laboratory strains. Differently from the deletions identified by GEM analysis, the results obtained through the integration of enzymatic data showed the advantage of predicting only the deletions of active reactions under the specific considered condition. Moreover, a set of reactions, directly connected with γ-*pga* production, has been selected for the over-expression. However, for an appropriate evaluation of identified target manipulations, an experimental validation is necessary.

Finally, the following two chapters focus on the experimental applications of genetic manipulations, through advanced synthetic biology techniques, to achieve the final goal of metabolic engineering.

Synthetic sRNAs are genetic tools able to control the target gene expression in bacteria. Chapter 4 describes the quantitative evaluation of sRNAs designed and constructed in *E. coli* on the basis of recently proposed guidelines. Despite developed synthetic sRNA showed to properly work under different contexts, a complete repression of target pathway, generally required for the reactions that compete for production of the desirable metabolite, was not achieved. Indeed the higher repression efficiency obtained for sRNA targeting the endogenous lactate dehydrogenase gene (*ldh*A) has been shown equal to 72%.

In this context, despite the different advantages of methods for down-regulation of the target gene expression, such as sRNAs, without modification of genome sequences, once the elimination of a specific reaction was established, its total inactivation can be achieved only by permanent deletion of encoding gene(s). Chapter 5 reports a novel allelic replacement vector for chromosomal gene deletion in *E. coli*, based on the colorimetric XylE assay and compatible with commonly used BioBrick™standards. It has been used for the disruption of lactate dehydrogenase (*ldh*A), fumarate reductase (*frd*AB) and pyruvate formatelyase (*pfl*B-*foc*A) genes, that compete for pyruvate utilization in ethanologenic strains. These gene deletions have been confirmed by properly experimental assay, in this chapter the absence of lactate dehydrogenase activity in $ldhA^{-}$ strain has been shown.

Taken together, the results show the relevance of constraint-based methods in metabolic engineering applications and that, differently from the transcriptomic data, which, in the considered contexts, were not able to model the modification of expression at post-transcriptional and post-translational level, the integration of enzymatic data is a promising approach to improve the predictions based on genome-scale metabolic models. These in-silico approaches can be efficiently supported by the two in-vivo tools for the repression or completely elimination of target genes, constructed in this thesis using the synthetic biology toolkits of pre-characterized regulatory elements.

Appendix A

Supplementary Figures for Ch. 4

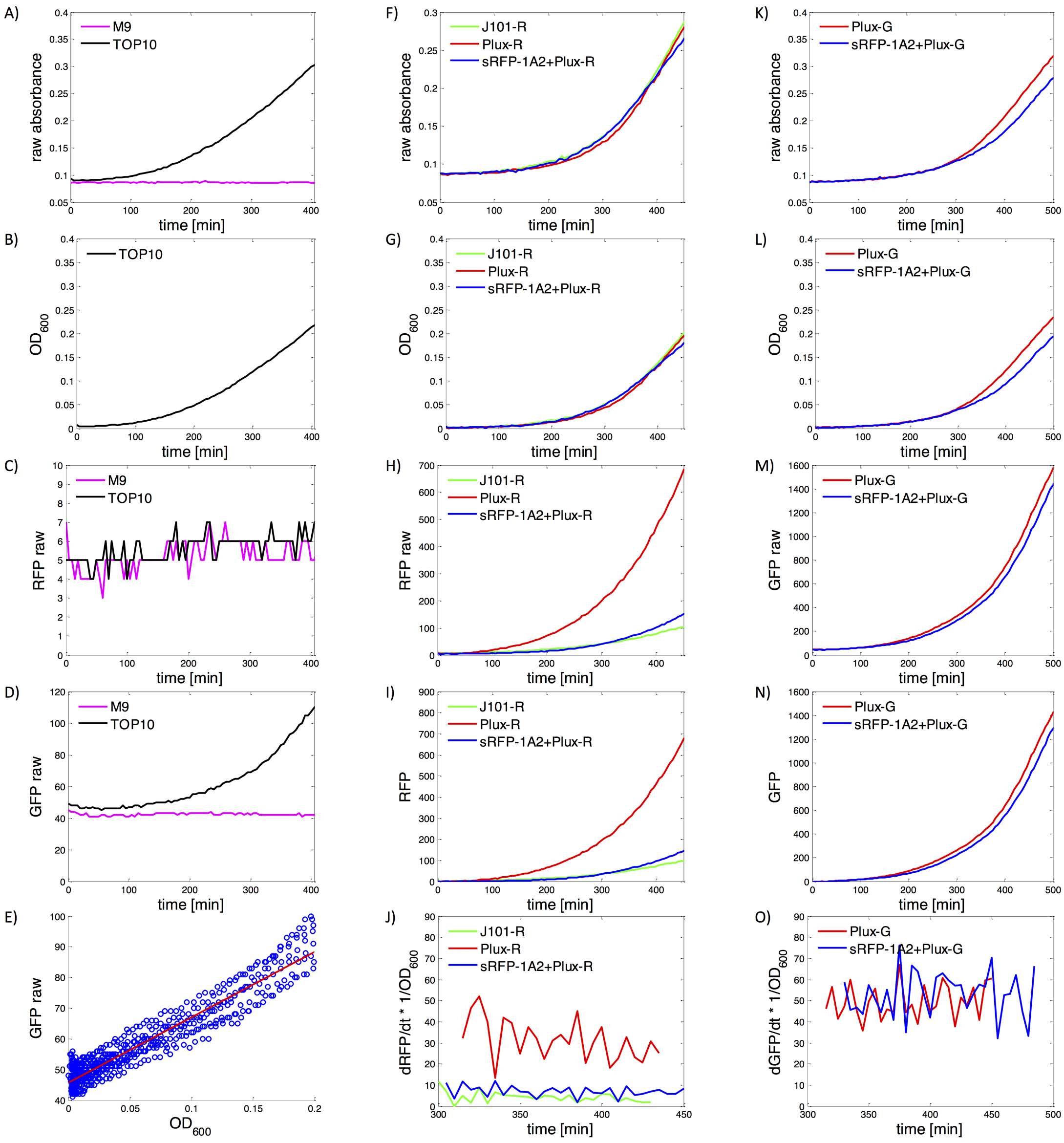


Figure A.1: **Raw data and processed time series for cultures in representative experiments.**

Figure A.1: A)Raw absorbance time series of sterile medium (M9) and a non-fluorescent culture (TOP10). B) Background-subtracted absorbance (OD_{600}) time series of a non-fluorescent culture (TOP10). C) Raw red fluorescence of sterile medium (M9) and a non-fluorescent culture (TOP10) time series, showing that the auto-fluorescence of bacteria is comparable with the fluorescence of medium and it is not OD_{600}-dependent. D) Raw green fluorescence of sterile medium (M9) and a non-fluorescent culture (TOP10) time series, showing that the auto-fluorescence of bacteria is higher than the fluorescence of medium and it is OD_{600}-dependent. E) Raw green fluorescence as a function of OD_{600} for several non-fluorescent strains assayed in the same experiment; such data are used to compute the OD_{600}-dependent auto-fluorescence function (by linear regression), which represents the background green fluorescence at a given OD_{600}; circles represent data points and solid line represents the regression line. F) Raw absorbance time series of three RFP-expressing cultures: Plux-R, sRFP-1A2+Plux-R and J101-R in TOP10. G) Background-subtracted OD_{600} time series of the three RFP-expressing cultures. H) Raw red fluorescence time series of the three RFP-expressing cultures. I) Background subtracted red fluorescence time series of the three RFP-expressing cultures, yielding a time series proportional to the total RFP proteins in the microplate well. J) Numeric time derivative of RFP divided by OD_{600}, yielding a signal proportional to the RFP synthesis rate per cell at the steady-state; the time series in the exponential growth phase (OD_{600} between 0.05 and 0.18, assumed) is shown for the three RFP-expressing cultures; for each culture, this time series is averaged and divided by the average RFP synthesis rate per cell of the reference culture (see Methods section in the main text). K)-O) the same time series as panels F)-J) are shown for two GFP-expressing cultures: Plux-G and sRFP-1A2+Plux-G.

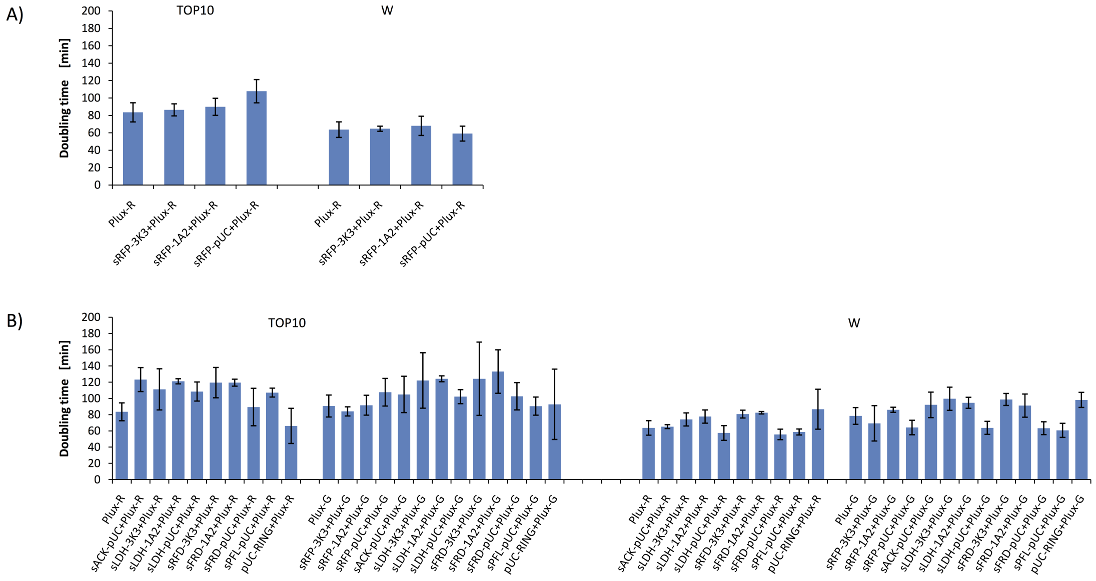


Figure A.2: **Doubling times of recombinant strains bearing a single-gene RFP or GFP expression system driven by P_{lux} (Plux-R or Plux-G).** A) Specific silencing of the target gene (RFP) via sRFP in TOP10 and W. B) Unspecific silencing of RFP or GFP via different sRNAs in TOP10 and W. Bars represent the mean doubling time value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value.

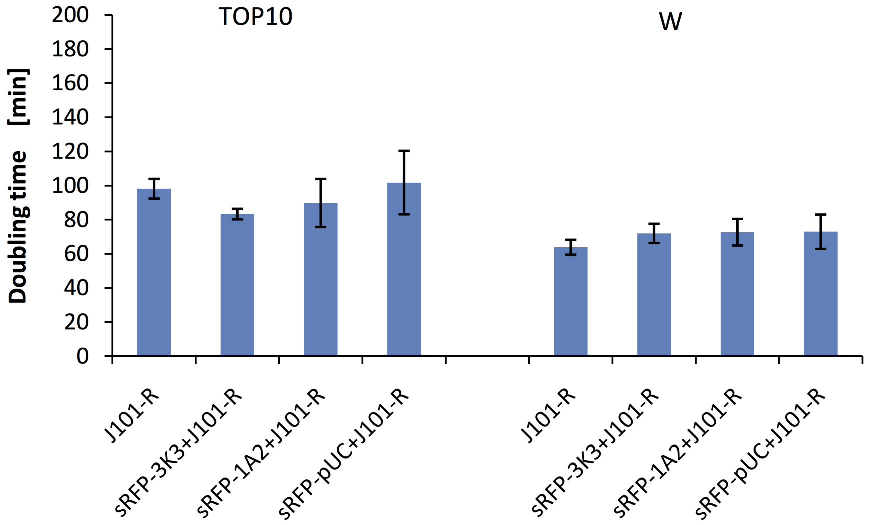
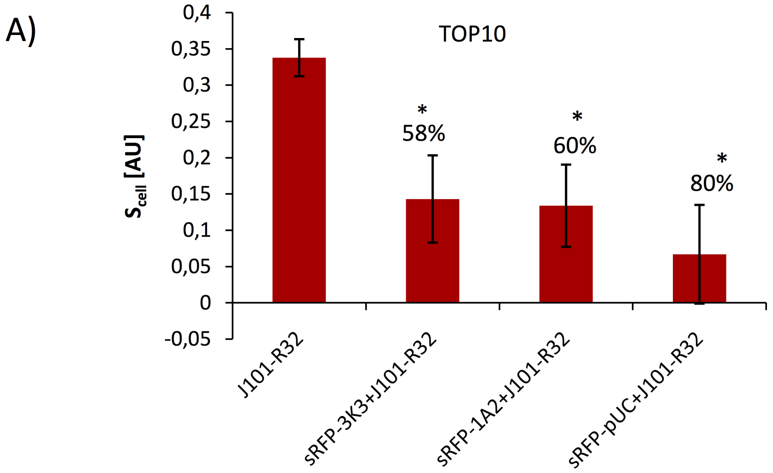


Figure A.3: **Doubling times of recombinant strains bearing an RFP expression system driven by BBa_J23101 (J101-R).** Bars represent the mean doubling time value computed on at least three biological replicates in the indicated conditions. Error bars represent the 95% confidence intervals of the mean value.



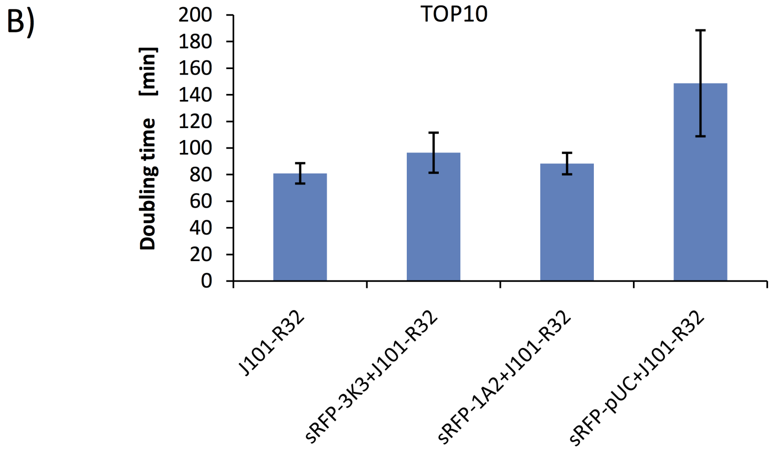


Figure A.4: **Silencing results for RFP expressed by a single-gene cassette (J101-R32) driven by BBa_J23101 with the BBa_B0032 RBS upstream of the RFP gene.** A) Specific silencing of the target gene (RFP) via sRFP in TOP10. Bars represent the mean S_{cell} value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the S_{cell} value in the condition is statistically different from the S_{cell} of the expression cassette without sRNAs (J101-R32). Percentages represent the Eff% values. B) Doubling times.

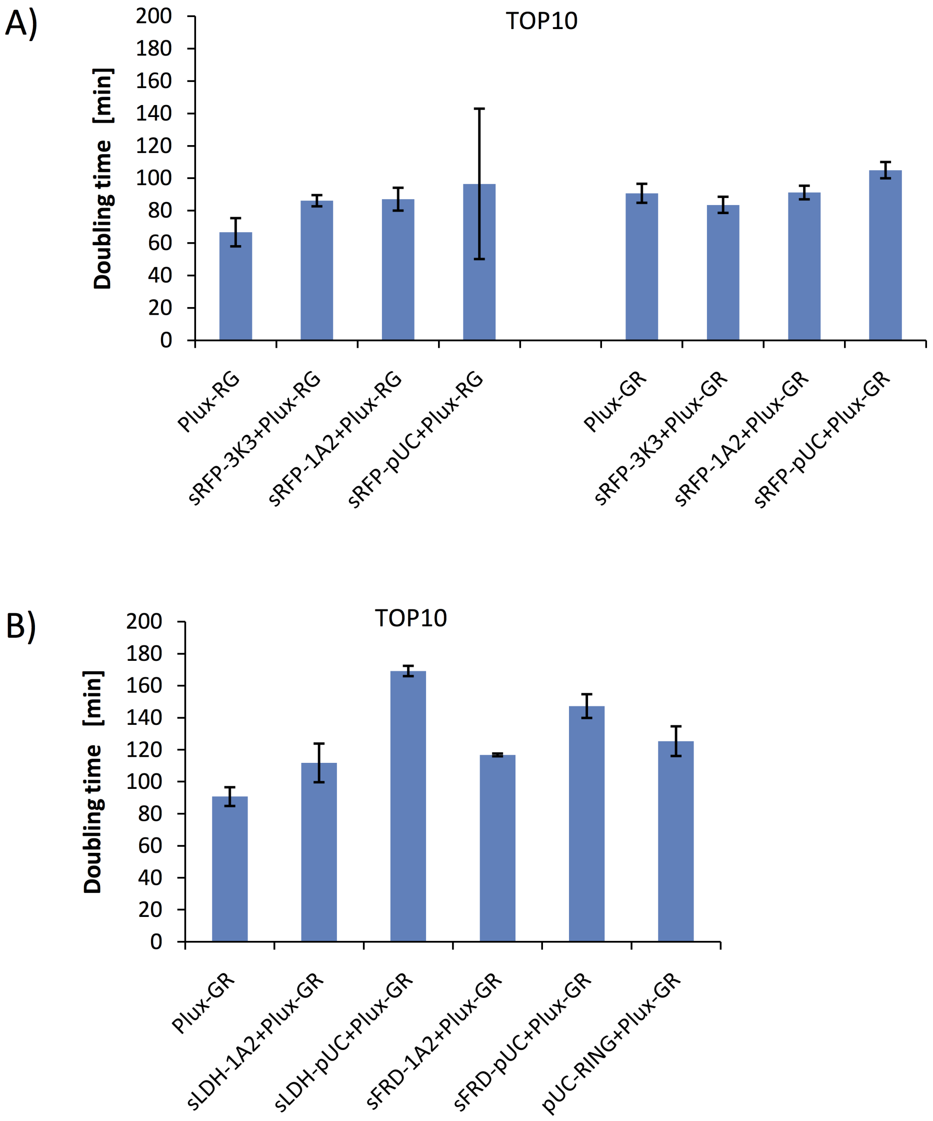


Figure A.5: **Doubling times of recombinant TOP10 bearing the Plux-RG and Plux-GR synthetic operons.** A) Silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP, in the Plux-GR construct, via different sRNAs. Bars represent the mean doubling time value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value.

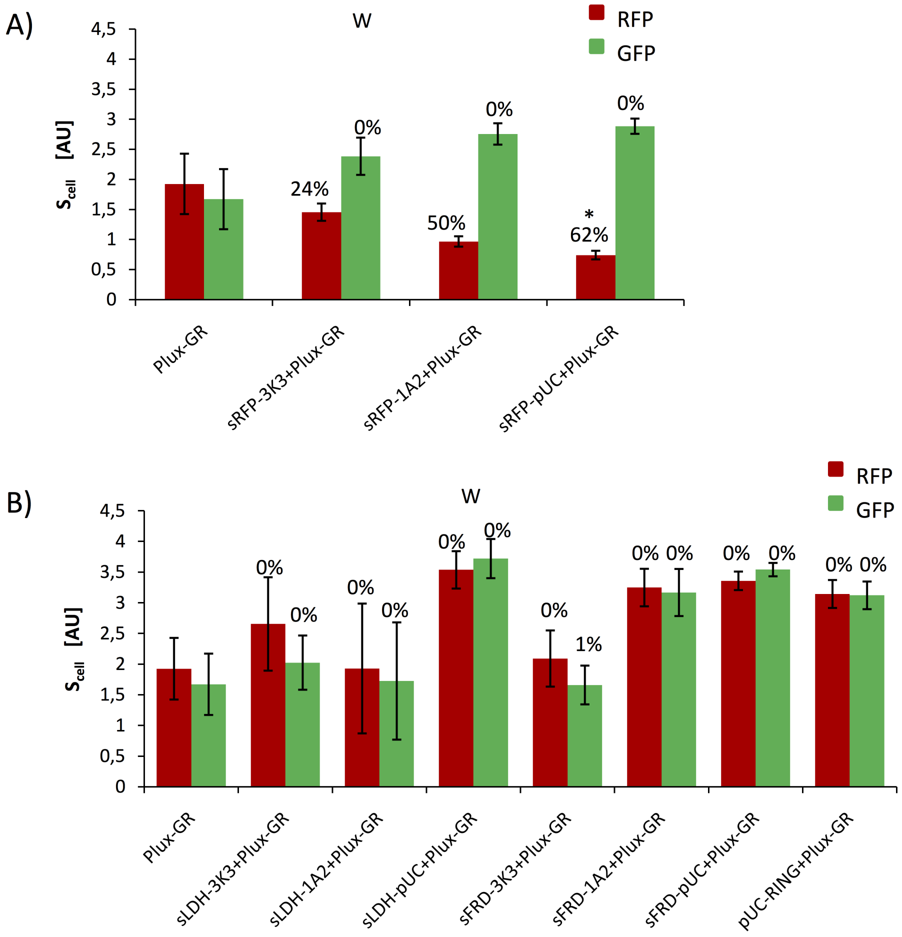


Figure A.6: **Silencing results for RFP and GFP expressed by the Plux-GR synthetic operon in the W strain.** A) Silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP via different sRNAs. Bars represent the mean S_{cell} value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the Scell value of RFP or GFP in the condition is statistically different from the Scell of the operon without sRNAs (Plux-GR). Percentages represent the Eff% values. When S_{cell} in a given condition is higher than Scell without sRNA, Eff% value is set to zero.

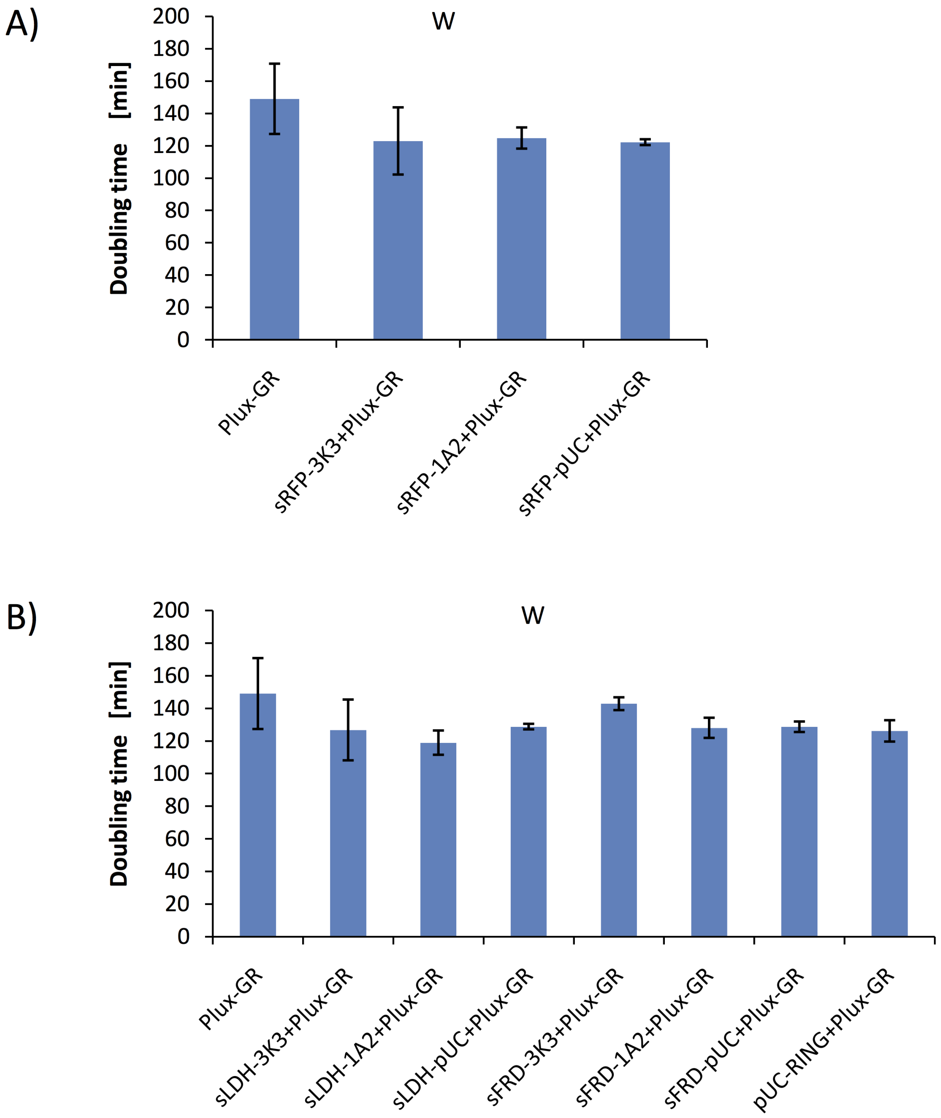


Figure A.7: **Doubling times of recombinant W bearing the Plux-GR synthetic operon.** A)Specific silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP via different sRNAs. Bars represent the mean doubling time value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value.

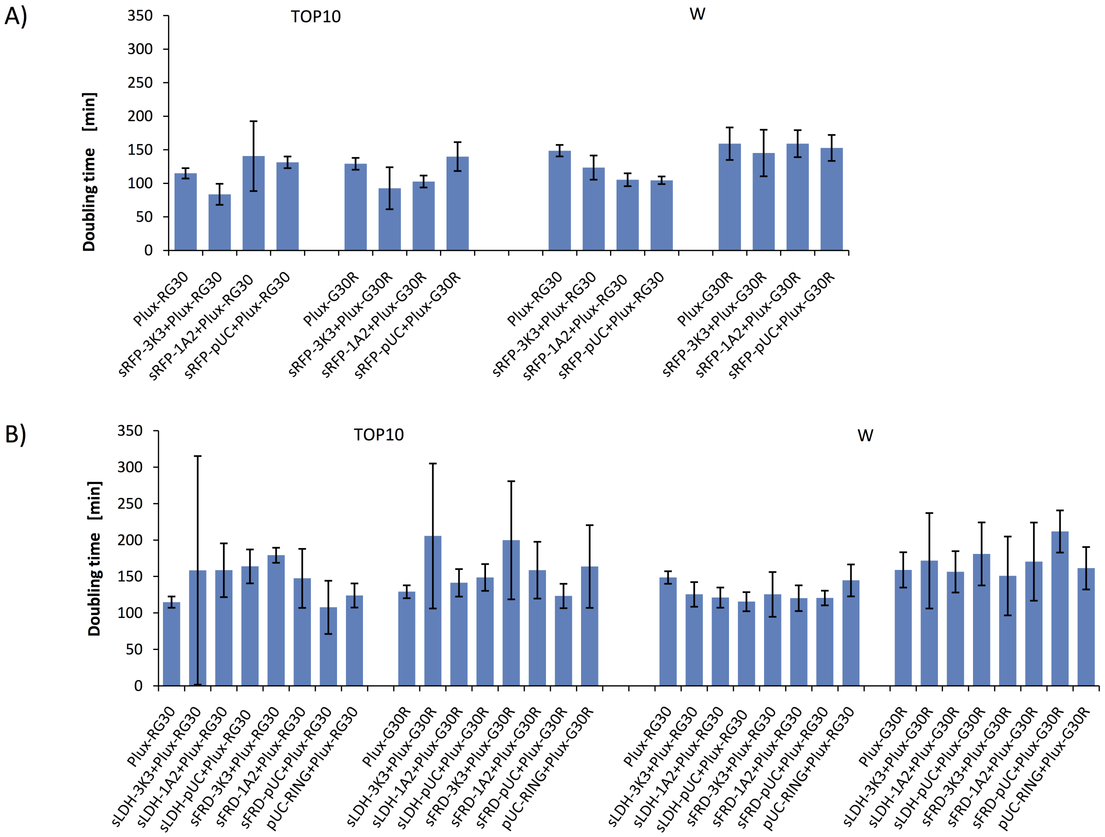


Figure A.8: **Doubling times of recombinant strains bearing the Plux-RG30 and Plux-G30R synthetic operons in TOP10 and W.** A)Silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP via different sRNAs. Bars represent the mean doubling time value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value.

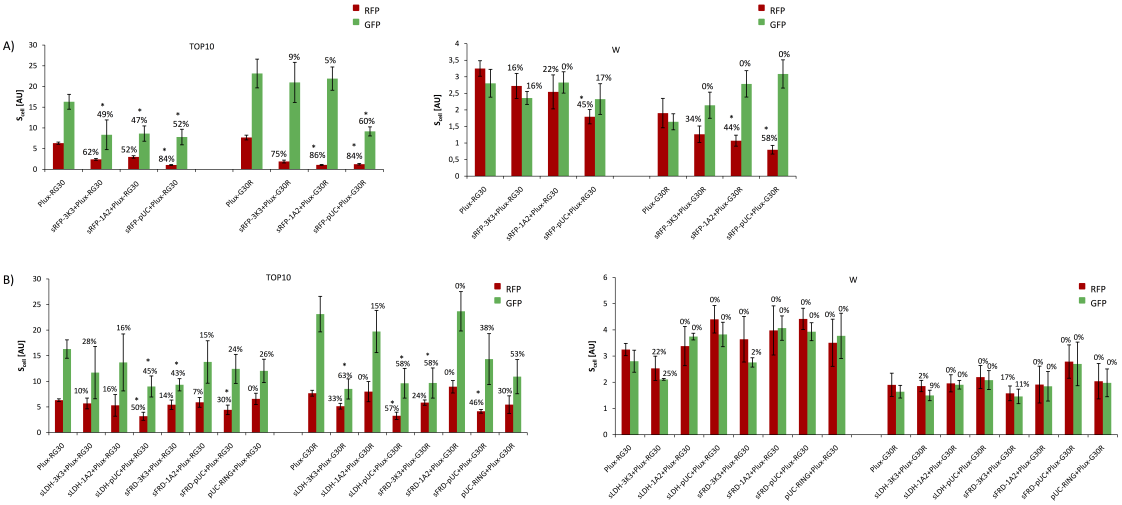


Figure A.9: **Silencing results for RFP and GFP expressed by the Plux-RG30 and Plux-G30R synthetic operons in TOP10 and W.** A)Silencing of the target gene (RFP) and the non-target gene (GFP) via the silencing device sRFP. B) Unspecific silencing of RFP and GFP via different sRNAs. Bars represent the mean S_{cell} value computed on at least three biological replicates. Error bars represent the 95% confidence intervals of the mean value. Asterisks indicate that the S_{cell} value of RFP or GFP in the condition is statistically different from the S_{cell} of the operon without sRNA (Plux-RG30 or Plux-G30R). Percentages represent the Eff% values. When S_{cell} in a given condition is higher than S_{cell} without sRNA, Eff% value is set to zero.

Appendix B

Methods and supplementary information for Ch. 5

B.1 Materials and Methods

B.1.1 *E. coli* Strains, Reagents and Cloning

TOP10 (Invitrogen) were used for cloning according to manufacturer's instructions. For gene knockout experiments, the W strain (ATCC 9637) was transformed by a standard heat shock protocol [166]. Strains were routinely grown in LB medium; chloramphenicol (12.5 mg/l) or ampicillin (100 mg/l) were added as required. Catechol (C9510, Sigma Aldrich) was dissolved in deionized water to obtain a 10 mM stock that was prepared fresh every day. Primers used in this work are listed in Tab. B.1.

The pBBknock vector was specialized to delete the *ldh*A gene of *E. coli* W by assembling the *ldh*A flanking DNA fragments (A and B, both 0.9 kbp-long, see Fig. 5.1B) in the cloning site. A and B regions were separately amplified from the genome of E. coli W with primer pairs PAtail_F/PAtail_R and PBtail_F/PBtail_R, respectively, with Phusion Hot Start Flex polymerase (New England Biolabs). Each PCR product was purified (NucleoSpin Extract II, Macherey-Nagel), digested with EcoRI and PstI (Roche), purified again, and finally individually ligated (T4 ligase, Roche) into the *Eco*RI-*Pst*I-digested pSB1A2 vector [65]. Each construct was sequence-verified with standard BioBrick™primers VF2 and VR. The A and B fragments in pSB1A2 were then digested with *Spe*I-*Pst*I and *Xba*I-*Pst*I, respectively, and ligated according to the

BioBrick™Standard Assembly to yield the AB construct (in pSB1A2), which was sequence-verified and, upon *Eco*RI-*Pst*I digestion, finally ligated into pBBknock.

B.1.2 Lactate Dehydrogenase Assay

The assay was performed as described by Massaiu et al. [131]. Cultures grown to saturation at 37 °C at 220 rpm in 2 ml of LB with 100 mM phosphate buffer and 40 g/l glucose, were 100-fold diluted in 9 ml of the same medium and grown for 4 h. One ml of culture was centrifuged (13,000 rpm, 1 min), washed with 1 ml of 100 mM Tris-HCl pH 7.3 and the pellet was resuspended with 0.4 ml of CelLytic B (Sigma Aldrich), supplemented with a protease inhibitor cocktail, to lyse the cells. After 10 min at room temperature, cell debris were removed by centrifugation (13,000 rpm, 5 min) and the supernatant was assayed. Reaction mix (180 μl), containing 100 mM Tris-HCl pH 7.3, 0.4 mM NADH and 10 mM sodium pyruvate, was mixed with 20 Î¼l of lysate and absorbance at 340 nm (OD_{340}) was monitored at 25 °C every 5 min in an Infinite F200 (Tecan) microplate reader. The slope of the absorbance time series, proportional to enzymatic activity of the sample, was computed via linear regression. Protein quantification in the lysate was obtained via Micro BCA Protein Assay Kit (Thermo Scientific). Specific enzymatic activity was calculated by dividing the total enzymatic activity by protein level and expressed as 10^4 *OD_{340}/min/μg of cell protein.

B.1.3 Copy Number Estimation for pBBknock

The copy number of pBBknock was estimated by comparing it to the one of pSB4C5 [65], which carries a non-ts pSC101 origin. To this aim, the BBa_J107029 part containing a constitutive promoter driving the Red Fluorescent Protein (RFP) expression, was assembled in both vectors upon *Eco*RI-*Pst*I digestion. Transformed TOP10 cells were assayed both in selective LB and M9 supplemented medium (11.28 g/l M9 salts - M6030, Sigma Aldrich, 2 mM MgSO4, 0.1 mM CaCl2, 2 g/l casamino acids, 1 mM thiamine hydrochloride and 4 ml/l glycerol) as previously reported [169], except that cultures were always incubated at 30 °C. RFP synthesis rate per cell (S_{cell}), expressed in arbitrary units (AU), was computed and assumed to be proportional to the plasmid copy number. S_{cell} and cell growth rate were computed as previously described [169]. Results were expressed as average S_{cell} values of at least three biological replicates and the confidence intervals of S_{cell} mean were reported.

B.2 Supplementary notes, results, figures and tables

B.2.1 Details about pBBknock thermosensitive sequence design

The pBBknock vector was obtained via de-novo synthesis. In order to design it, the sequence of the pSC101ts temperature-sensitive replication origin was retrieved and modified to remove unwanted restriction sites (*Spe*I). In this process we found inconsistencies among the available sequences for some widely used temperature-sensitive vectors, which are described below.
BBa_I50052 (pSC101ts) is a thermo-sensitive version of BBa_I50042 (pSC101), both present in the Registry; the mutation conferring the temperature-sensitive phenotype of BBa_I50052 is annotated.
Consistently, this mutation is also found in the sequence of the temperature-sensitive pKOV vector [189] [https://www.addgene.org/25769]; conversely, in the deposited sequence of another widely used temperature-sensitive vector, pAH123 [205] [GenBank: AY048726], such mutation is not reported. Because of such inconsistency, we sequenced pAH123 with primers P1_F, P2_R, P3_F, P4_F, P5_R (reported in Tab. B.1) and we could confirm the presence of the temperature-sensitive mutation described in pSC101ts and pKOV; thus, the sequence given in the [GenBank: AY048726] entry does not include the temperature-sensitive nucleotide change. Since comparison between our sequence of pAH123 and the one of BBa_I50052 showed additional mismatches (not shown), we decided to rely on pAH123 origin to design our vector. We extended the origin region to the upstream *Nco*I restriction site and to the downstream stop codon of the ampicillin resistance gene in pAH123. We also modified the origin region by removing the *Spe*I site, following the strategy used to modify pSC101 (BBa_I50042), present in the widely used pSB4C5 vector, to ensure that the nucleotide changes do not affect the replication origin functioning. The resulting sequence is annotated in the BBa_J107077 entry of the Registry as temperature-sensitive replication origin pSC101ts and has been submitted as part BBa_J107112.

B.2.2 Notes on protocol development

The second recombination is a rare event and the corresponding step of the protocol (DAY 3-4, Fig. 5.1C) is critical. No white colonies were present in catechol-stained plates if the incubation at 30 ◦C was carried out for 6 hours only; an overnight incubation, followed by a 100-fold dilution and additional 6-hour incubation, was essential to obtain positive clones in the strain and conditions tested. Also, no white colonies were present in catechol-stained plates if the incubation temperature of DAY 3-4 (Fig. 5.1C) was set at 42 ◦C instead of 30 ◦C, demonstrating that the second recombination event is stimulated only at permissive temperature for the pSC101ts origin. Since no white colonies were present in such conditions, these results also suggest that the false positive rate (i.e., the occurrence of white colonies in which the second recombination did not happen) is negligible in the host strain and condition tested.

B.2.3 Copy number characterization

Measured fluorescence values (used to estimate plasmid copy number) were similar between pBBknock and pSB4C5: S_{cell}=1.52 ± 0.17 AU and 1.48 ± 0.29 AU, respectively, in LB, and 3.34 ± 0.03 AU and 2.65 ± 0.05 AU, respectively, in M9. Doubling times of recombinant strains bearing pBBknock and pSB4C5 were 78 min and 54 min, respectively, in LB, and 161 min and 112 min, respectively, in M9. Figure B.1 shows the growth curves of *E. coli* bearing pBBknock or pSB4C5.

B.2.4 Occurrence of illegal restriction sites in homologous fragments

Here we report the probability of finding at least one illegal restriction site (*EcoR*I, *Xba*I, *Spe*I or *Pst*I) when designing a homologous fragment to be cloned, like A or B in this work. To perform this task, we considered the *EcoR*I, *Xba*I, *Spe*I or *Pst*I restriction sites in the ATCC 9637 genome sequence. By using a nucleotide window moving along the genome, we found via *Per*l script this probability, shown in Figure B.2 as a function of the nucleotide window length. The considered lengths have been chosen in accordance with the length of homologous fragments used in previous studies [189, 190, 197]. Although a high probability of finding at least one site (about 26%) is present for a length of 0.9 Kb, i.e., the one used in this work, this probability dramatically decreases when a lower length is considered, e.g., about 16% for 0.5 Kb, successfully used in several applications [189].

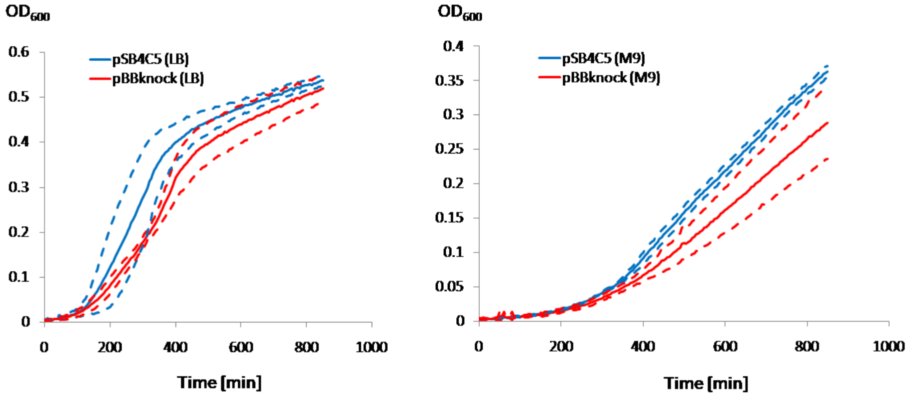


Figure B.1: **Growth curves for TOP10 strain bearing pBBknock or a control vector (pSB4C5) with pSC101 replication origin.** Both vectors have BBa_J107029 as insert. Data are relative to the copy number characterization experiment carried out in microplate reader [169]. Data represent background-subtracted absorbance (OD_{600}) over time in cultures grown in LB or M9 supplemented medium. Solid lines represent the mean of four independent clones, while dotted lines are the 95% confidence intervals of the mean.

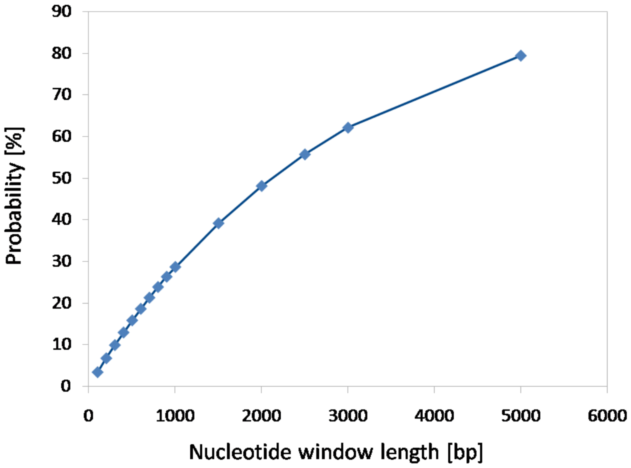


Figure B.2: **Probability of finding at least one illegal BioBrick™restriction site in a nucleotide window of variable length in the genome of *E. coli* W.** Points represent probability values for different length values.

PAtail_F (*Eco*RI)[a]	CTTC*GAATTC*GCGGCCGCTTCTAGAGGAATGTTTTGATCAAACAGAGGGC
PAtail_R (*Pst*I)	CTTC*CTGCAG*CGGCCGCTACTAGTATGCCCGAACGAACTGGTTTA
PBtail_F (*Eco*RI)	CTTC*GAATTC*GCGGCCGCTTCTAGAGCATCAACAACTATGCTTAGTGTAG
PBtail_R (*Pst*I)	CTTC*CTGCAG*CGGCCGCTACTAGTACATCGCTTACGGTCAATTGTTGAC
VF2	TGCCACCTGACGTCTAAGAA
VR	ATTACCGCCTTTGAGTGAGC
PA_F	TTACACATCCCGCCATCAGC
PB_R	GCAATTTCGCCAGACAAGCA
P1_F	TAGCCAGTCTGAATGACCTGTCAC
P2_R	CCTCAGATCCTTCCGTATTTAGCC
P3_F	CAAACAGCGTTTGCGACATCCT
P4_F	GCCCGACTGATACGTTGATTTTCC
P5_R	AAGGCTTAAGTAGCACCCTCGCAA

Table B.1: **Primers used in this study.** a) The restriction sites used for cloning are reported in brackets and their sequence is underlined.

Bibliography

[1] Stephanopoulos G, Aristidou AA and Nielsen J. *Metabolic engineering: principles and methodologies.* Academic press., 1998.

[2] Stephanopoulos G. Synthetic biology and metabolic engineering. *ACS synthetic biology*, 1(11):514–525, 2012.

[3] Nielsen J and Keasling JD. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

[4] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, and Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, 2012.

[5] Jakočiūnas T, Bonde I, Herrgård M, Harrison SJ, Kristensen M, Pedersen LE, Jensen MK, and Keasling JD. Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae. Metabolic engineering*, 28:213–222, 2015.

[6] Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *cell*, 154(2):442–451, 2013.

[7] Lakshmanan M, Koh G, Chung BKS, and Lee D-Y. Software applications for flux balance analysis. *Briefings in bioinformatics*, 15(1):108–122, 2012.

[8] Kamp AV and Schuster S. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, 22(15):1930–1931, 2006.

[9] Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 6(9):1290, 2011.

[10] Ebrahim A, Lerman JA, Palsson BØ, and Hyduke DR. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC systems biology*, 7(1):74, 2013.

[11] Cardoso J, Jensen, Lieven, Hansen ASL, Galkina S, Beber ME, Ozdemir E, Herrgard M, Redestig H, and Sonnenschein N. Cameo: A Python Library for Computer Aided Metabolic Engineering and Optimization of Cell Factories. *bioRxiv*, page 147199, 2017.

[12] Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, and Rocha M. OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC systems biology*, 4(1):45, 2010.

[13] Boele J, Olivier BG, and Teusink B. FAME, the flux analysis and modeling environment. *BMC systems biology*, 6(1):8, 2012.

[14] Thiele I and Palsson B Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.*, 5:93, 2010.

[15] Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28:27–30, 2012.

[16] Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldvsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, and Lopez-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160. *Nucleic acids research*, 2005(19):6083–6089, 33.

[17] Gianchandani EP, Chavali AK, and Papin JA. The application of flux balance analysis in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3):372–382, 2010.

[18] Orth JL, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, and Palsson BØ. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Molecular systems biology*, 7:535, 2011.

[19] Förster J, Famili I, Fu P, Palsson BØ, and Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research*, 13(2):244–253, 2003.

[20] Oh Y-K, Palsson BO, Park SM, Schilling CH, and Mahadevan R. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *Journal of Biological Chemistry*, 282:28791–28799, 2007.

[21] Papoutsakis ET Senger RS. Genome-scale model for *Clostridium acetobutylicum*: Part I. Metabolic network resolution and analysis. *Biotechnology and bioengineering*, 101(5):1036–1052, 2008.

[22] Kjeldsen KR and Nielsen J. *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnology and bioengineering*, 102(2):583–597, 2009.

[23] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, and Palsson BØ. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.

[24] BIGG Models. [http://bigg.ucsd.edu/].

[25] DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, and Best A. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC bioinformatics*, 8(1):139, 2007.

[26] Agren R, Liu Land Shoaie S, Vongsangnak W, Nookaew I, and Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum. PLoS computational biology*, 9(3):e1002980, 2013.

[27] Ferreira EC Dias O, Rocha M and Rocha I. Reconstructing genome-scale metabolic models with merlin. *Nucleic acids research*, 43(8):3899–3910, 2015.

[28] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, KH, Arkin AP, Bornstein BJ, BD, Cornish-Bowden A, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[29] Edwards JS and Palsson BØ. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533, 2000.

[30] Reed JL, Vo TD, Schilling CH, and Palsson BØ. An expanded genome-scale model of *Escherichia coli* K-12 (i JR904 GSM/GPR). *Genome biology*, 4(9):R54, 2003.

[31] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, and Palsson BØ. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology*, 3(1):121, 2007.

[32] Orth JD, Fleming R MT, and Palsson BØ. Reconstruction and use of microbial metabolic networks: the core *Escherichia coli* metabolic model as an educational guide. *EcoSal plus*, 2010.

[33] Henry CS, Zinner JF, Cohoon MP, and Stevens RL. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome biology*, 10(6):R69, 2009.

[34] Tanaka K, Henry CS, Zinner JF, Jolivet E, Cohoon MP, Xia F, Bidnenko V, Ehrlich SD, Stevens RL, and Noirot P. Building the repertoire of dispensable chromosome regions in *Bacillus subtilis* entails major refinement of cognate large-scale metabolic model. *Nucleic acids research*, 41(1):687–699, 2012.

[35] Hao T, Han B, Ma H, Fu J, Wang H, Wang Z, Tang B, Chen T, and Zhao X. *In silico* metabolic engineering of *Bacillus subtilis* for improved production of riboflavin, Egl-237,(R, R)-2, 3-butanediol and isobutanol. *Molecular BioSystems*, 9(8):2034–2044, 2013.

[36] Lewis NE, Nagarajan H, and P BØ. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature reviews. Microbiology*, 10(4):291, 2012.

[37] Orth JD, Thiele I, and Palsson BØ. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[38] Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390, 2010.

[39] Mahadevan R and Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276, 2003.

[40] Mahadevan R, Edwards JS, and Doyle FJ. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysical journal*, 83(3):1331–1340, 2002.

[41] Segre D, Vitkup D, and Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.

[42] Shlomi T, Berkman O, and Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7695–7700, 2005.

[43] Burgard AP, Pharkya P, and Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[44] Tepper N and Shlomi T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics*, 26(4):536–543, 2009.

[45] Patil KR, Rocha I, Förster J, and Nielsen J. Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC bioinformatics*, 6(1):308, 2005.

[46] Choi HS, Lee SY, Kim TY, and Woo HM. *In silico* identification of gene amplification targets for improvement of lycopene production. *Applied and environmental microbiology*, 76(10):3097–3105, 2010.

[47] Schuster S and Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(02):165–182, 1994.

[48] Trinh CT, Unrean P, and Srienc F. Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and environmental microbiology*, 74(12):3634–3643, 2008.

[49] Schilling CH, Letscher D, and Palsson BØ. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of theoretical biology*, 203(3):229–248, 2000.

[50] Nielsen J. Systems Biology of Metabolism. *Annual Review of Biochemistry*, 86:245–275, 2017.

[51] Machado D and Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*, 10(4):e1003580, 2014.

[52] Heller MJ. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

[53] Wang Z, Gerstein M, and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[54] Åkesson M, Förster J, and Nielsen J. Integration of gene expression data into genome-scale metabolic models. *vMetabolic engineering*, 2004.

[55] Becker SA and Palsson BØ. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.

[56] Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids research*, 41(D1):D764–D772, 2012.

[57] Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A, Sauer-Danzwith H, Mir S, et al. SABIORKdatabase for biochemical reaction kinetics. *Nucleic acids research*, 40(D1):D790–D796, 2011.

[58] Adadi R, Volkmer B, Milo R, Heinemann M, and Shlomi T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS computational biology*, 8(7):e1002575, 2012.

[59] Sánchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, and Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*, 13(8):935, 2017.

[60] Purnick PEM and Weiss R. The second wave of synthetic biology: from modules to systems. *Nature reviews Molecular cell biology*, 10(6):410–422, 2009.

[61] Knight T. Idempotent vector design for standard assembly of biobricks. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 2003.

[62] Salis HM, Mirsky EA, and Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–950, 2009.

[63] Seo SW, Yang J-S, Kim I, Yang J, Min BE, Kim S, and Jung GY. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metabolic engineering*, 15:67–74, 2013.

[64] Na D and Lee D. RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics*, 26(20):2633–2634, 2010.

[65] MIT: Registry of Standard Biological Parts. [http://partsregistry.org].

[66] Park JH, Lee KH, Kim TY, and Lee SY. Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proceedings of the National Academy of Sciences*, 104(19):7797–7802, 2007.

[67] Maia P, Rocha M, and Rocha I. *In silico* constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiology and Molecular Biology Reviews*, 80(1):45–67, 2016.

[68] Blazier AS and Papin JA. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in physiology*, 3, 2012.

[69] Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell D B, Mendes P, and Swainston N. Improving metabolic flux predictions using absolute gene expression data. *BMC systems biology*, 6(1):73, 2012.

[70] Rossell S, Huynen MA, and Notebaart RA. Inferring metabolic states in uncharacterized environments using gene-expression measurements. *PLoS computational biology*, 9(3):e1002988, 2013.

[71] Edwards JS, Ibarra RU, and Palsson BO. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125, 2001.

[72] Causey TB, Zhou S, Shanmugam KT, and Ingram LO. Engineering the metabolism of *Escherichia coli* W3110 for the conversion of sugar to redox-neutral and oxidized products: homoacetate production. *Proceedings of the National Academy of Sciences*, 100(3):825–832, 2003.

[73] Causey TB, Shanmugam KT, Yomano LP, and Ingram LO. Engineering *Escherichia coli* for efficient conversion of glucose to pyruvate. *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2235–2240, 2004.

[74] Kayser A, Weber J, Hecht V, and Rinas U. Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state. *Microbiology*, 151(3):693–706, 2005.

[75] Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, et al. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316(5824):593–597, 2007.

[76] Kim Y, Ingram LO, and Shanmugam KT. Construction of an *Escherichia coli* K-12 mutant for homoethanologenic fermentation of glucose or xylose without foreign genes. *Applied and Environmental Microbiology*, 73(6):1766–1771, 2007.

[77] Orencio-Trejo M, Flores N, Escalante A, Hernández-Chávez G, Bolívar F, Gosset G, and Martinez A. Metabolic regulation analysis of an ethanologenic *Escherichia coli* strain based on RT-PCR and enzymatic activities. *Biotechnology for biofuels*, 1(1):8, 2008.

[78] Covert MW, Knight EM, Reed JL, Herrgard MJ, and Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92, 2004.

[79] Park D, Acktar S, Ansari A, Landick R, and Kiley P. Expression analysis of *Escherichia coli* MG1655 K-12 WT and Δ*arc*A mutant. *data accessible at NCBI GEO database, accession GSE46412*, 2013.

[80] Bordbar A, Nagarajan H, Lewis NE, Latif H, Ebrahim A, Federowicz S, Schellenberger J, and Palsson BO. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular systems biology*, 10(7):737, 2014.

[81] von Wulffen J, Sawodny O, Feuer R, et al. Transition of an anaerobic *Escherichia coli* culture to aerobiosis: balancing mRNA and protein levels in a demand-directed dynamic flux balance analysis. *PloS one*, 11(7):e0158711, 2016.

[82] Amarjeet Singh, Anis Karimpour-Fard, and Ryan T Gill. Increased mutation frequency in redox-impaired *Escherichia coli* due to RelA-and RpoS-mediated repression of DNA repair. *Applied and environmental microbiology*, 76(16):5463–5470, 2010.

[83] Varma A and Palsson BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and environmental microbiology*, 60(10):3724–3731, 1994.

[84] Ingram LO, Conway T, Clark DP, Sewell GW, and Preston JF. Genetic engineering of ethanol production in *Escherichia coli. Applied and Environmental Microbiology*, 53(10):2420–2425, 1987.

[85] Ohta K, Beall DS, Mejia JP, Shanmugam KT, and Ingram LO. Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. *Applied and Environmental Microbiology*, 57(4):893–900, 1991.

[86] Dien BS, Nichols NN, Obryan PJ, and Bothast RJ. Development of new ethanologenic *Escherichia coli* strains for fermentation of lignocellulosic biomass. *Applied biochemistry and biotechnology*, 84(1):181–196, 2000.

[87] Bodor Z, Fazakas A, and Abraham B Lanyi S. *In silico* modelling and metabolic engineering of *Escherichia coli* to succinic acid productio. *U.P.B. Sci. Bull*, 76(4):59–70, 2014.

[88] Adams BL. The Next Generation of Synthetic Biology Chassis: Moving Synthetic Biology from the Laboratory to the Field, 2016.

[89] Guiziou S, Sauveplane V, Chang H-J, Clerté C, Declerck N, Jules M, and Bonnet J. A part toolbox to tune genetic expression in *Bacillus subtilis*. *Nucleic acids research*, 44(15):7495–7508, 2016.

[90] Perkins JB, Sloma A, Hermann T, Theriault K, Zachgo E, Erdenberger T, Hannett N, Chatterjee NP, Williams II V, Rufo Jr GA, et al. Genetic engineering of *Bacillus subtilis* for the commercial production of riboflavin. *Journal of Industrial Microbiology and Biotechnology*, 22(1):8–18, 1999.

[91] Navaneeth S, Bhuvanesh S, Bhaskar V, Vijay KP, Kandaswamy SKJ, and Achary A. Optimization of medium for the production of subtilisin from *Bacillus subtilis* MTCC 441. *African Journal of Biotechnology*, 8(22), 2009.

[92] Gilbert C, Howarth M, Harwood CR, and Ellis T. Extracellular self-assembly of functional and tunable protein conjugates from *Bacillus subtilis*. *ACS Synthetic Biology*, 2017.

[93] Scoffone V, Dondi D, Biino G, Borghese G, Pasini D, Galizzi A, and Calvio C. Knockout of pgdS and ggt genes improves γ-PGA yield in *B. subtilis*. *Biotechnology and bioengineering*, 110(7):2006–2012, 2013.

[94] Kunst F, Ogasawara N, Moszer I, Albertini AM, et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, 390(6657):249–256, 1997.

[95] Sonenshein AL, Hoch JA, and Losick R. *Bacillus subtilis* and Its Closest Relatives. *American Society of Microbiology*, 2002.

[96] Overbeek R, Disz T, and Stevens R. The SEED: a peer-to-peer environment for genome annotation. *Communications of the ACM*, 47:46–51, 2004.

[97] Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási A-L, and ZN Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proceedings of the National Academy of Sciences*, 104(31):12663–12668, 2007.

[98] Nilsson A and Nielsen J. Metabolic trade-offs in yeast are caused by F1F0-ATP synthase. *Scientific reports*, 6:22264, 2016.

[99] Yizhak K, Benyamini T, Liebermeister W, Ruppin E, and Shlomi T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–i260, 2010.

[100] Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, Barenholz U, Goldenfeld M, Shlomi T, and Milo R. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k_{cat} measurements. *Proceedings of the National Academy of Sciences*, 113(12):3401–3406, 2016.

[101] Davidi D and Milo R. Lessons on enzyme kinetics from quantitative proteomics. *Current Opinion in Biotechnology*, 46:81–89, 2017.

[102] Jin S and Sonenshein AL. Characterization of the major citrate synthase of *Bacillus subtilis*. *Journal of bacteriology*, 178(12):3658–3660, 1996.

[103] Costa T, Steil L, Martins LO, Völker U, and Henriques AO. Assembly of an oxalate decarboxylase produced under σK control into the *Bacillus subtilis* spore coat. *Journal of bacteriology*, 186(5):1462–1474, 2004.

[104] Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, Mariadassou M, Aymerich S, Hecker M, Noirot P, et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic engineering*, 32:232–243, 2015.

[105] Muntel J, Fromion V, Goelzer A, Maaβ S, Mäder U, Büttner K, Hecker M, and Becher D. Comprehensive absolute quantification of the cytosolic proteome of *Bacillus subtilis* by data independent, parallel fragmentation in liquid chromatography/mass spectrometry (LC/MSE). *Molecular & Cellular Proteomics*, 13(4):1008–1019, 2014.

[106] Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, 35(12):1050–1055, 2013.

[107] Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, and von Mering C. PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & Cellular Proteomics*, 11(8):492–500, 2012.

[108] Gao H, Chen Y, and Leary JA. Kinetic measurements of phosphoglucose isomerase and phosphomannose isomerase by direct analysis of phosphorylated aldose–ketose isomers using tandem mass spectrometry. *International Journal of Mass Spectrometry*, 240(3):291–299, 2005.

[109] Wierenga RK, Kapetaniou EG, and Venkatesan R. Triosephosphate isomerase: a highly evolved biocatalyst. *Cellular and molecular life sciences*, 67(23):3961–3982, 2010.

[110] Fillinger S, Boschi-Muller S, Azza S, Dervyn E, Branlant G, and Aymerich S. Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *Journal of Biological Chemistry*, 275(19):14031–14037, 2000.

[111] D'Alessio G and Josse J. Glyceraldehyde Phosphate Dehydrogenase, Phosphoglycerate Kinase, and Phosphoglyceromutase of *Escherichia coli* SIMULTANEOUS PURIFICATION AND PHYSICAL PROPERTIES. *Journal of Biological Chemistry*, 246(13):4319–4325, 1971.

[112] Watabe K and Freese E. Purification and properties of the manganese-dependent phosphoglycerate mutase of *Bacillus subtilis*. *Journal of bacteriology*, 137(2):773–778, 1979.

[113] Brown CK, Kuhlman PL, Mattingly S, Slates K, Calie PJ, and Farrar WW. A model of the quaternary structure of enolases, based on structural and evolutionary analysis of the octameric enolase from *Bacillus subtilis*. *Journal of protein chemistry*, 17(8):855–866, 1998.

[114] Olavarría K, Valdes D, and Cabrera R. The cofactor preference of glucose-6-phosphate dehydrogenase from *Escherichia coli*–modeling the physiological production of reduced cofactors. *The FEBS journal*, 279(13):2296–2309, 2012.

[115] Singh SK, Miller SP, Dean A, Banaszak LJ, and LaPorte DC. *Bacillus subtilis* Isocitrate Dehydrogenase A SUBSTRATE ANALOGUE FOR ESCHERICHIA COLI ISOCITRATE DEHYDROGENASE KINASE/PHOSPHATASE. *Journal of Biological Chemistry*, 277(9):7567–7573, 2002.

[116] Ueda Y, Yumoto N, Tokushige M, Fukui K, and Ohya-Nishiguchi H. Purification and characterization of two types of fumarase from *Escherichia coli*. *The Journal of Biochemistry*, 109(5):728–733, 1991.

[117] Smith K, Sundaram TK, Kernick M, and Wilkinson AE. Purification of bacterial malate dehydrogenases by selective elution from a triazinyl dye affinity column. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 708(1):17–25, 1982.

[118] Shin B-S, Choi S-K, and Park S-H. Regulation of the *Bacillus subtilis* Phosphotransacetylase Gene. *The Journal of Biochemistry*, 126(2):333–339, 1999.

[119] Garvie EI. Bacterial lactate dehydrogenases. *Microbiological reviews*, 44(1):106, 1980.

[120] SASKI R and PIZER LI. Regulatory Properties of Purified 3-Phosphoglycerate Dehydrogenase from *Bacillus subtilis*. *The FEBS Journal*, 51(2):415–427, 1975.

[121] Svedružić D, Liu Y, Reinhardt LA, Wroclawska E, Cleland WW, and Richards NGJ. Investigating the roles of putative active site residues in the oxalate decarboxylase from *Bacillus subtilis*. *Archives of biochemistry and biophysics*, 464(1):36–47, 2007.

[122] Liu S, Lu Z, Han Y, Melamud E, Dunaway-Mariano D, and Herzberg O. Crystal structures of 2-methylisocitrate lyase in complex with product and with isocitrate inhibitor provide insight into lyase substrate specificity, catalysis and evolution. *Biochemistry*, 44(8):2949–2962, 2005.

[123] Dawson A, Chen M, Fyfe PK, Guo Z, and Hunter WN. Structure and reactivity of *Bacillus subtilis* MenD catalyzing the first committed step in menaquinone biosynthesis. *Journal of molecular biology*, 401(2):253–264, 2010.

[124] Guo J, Zhang H, Wang C, Chang J-W, and Chen L-L. Construction and analysis of a genome-scale metabolic network for *Bacillus licheniformis* WX-02. *Research in microbiology*, 167(4):282–289, 2016.

[125] Fischer E and Sauer U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nature genetics*, 37(6):636–640, 2005.

[126] Chubukov V, Uhr M, Le Chat L, Kleijn RJ, Jules M, Link H, Aymerich S, Stelling J, and Sauer U. Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Molecular systems biology*, 9(1):709, 2013.

[127] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al. Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences*, 100(8):4678–4683, 2003.

[128] Inaoka T and Ochi K. Glucose uptake pathway-specific regulation of synthesis of neotrehalosadiamine, a novel autoinducer produced in *Bacillus subtilis*. *Journal of bacteriology*, 189(1):65–75, 2007.

[129] Osera C, Amati G, Calvio C, and Galizzi A. SwrAA activates poly-γ-glutamate synthesis in addition to swarming in *Bacillus subtilis*. *Microbiology*, 155(7):2282–2287, 2009.

[130] Yu W, Chen Z, Shen L, Wang Y, Li Q, Yan S, Zhong C-J, and He N. Proteomic profiling of *Bacillus licheniformis* reveals a stress response mechanism in the synthesis of extracellular polymeric flocculants. *Biotechnology and bioengineering*, 113(4):797–806, 2016.

[131] Massaiu I, Pasotti L, Casanova M, Politi N, Zucca S, Cusella De Angelis MG, and Magni P. Quantification of the gene silencing performances of rationally-designed synthetic small RNAs. *Systems and synthetic biology*, 9(3):107–123, 2015.

[132] Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. *TRENDS in Genetics*, 21(7):399–404, 2005.

[133] Man S, Cheng R, Miao C, Gong Q, Gu Y, Lu X, Han F, and Yu W. Artificial trans-encoded small non-coding RNAs specifically silence the

selected gene expression in bacteria. *Nucleic acids research*, 39(8):e50–e50, 2011.

[134] Jost D, Nowojewski A, and Levine E. Small RNA biology is systems biology. *BMB reports*, 44(1):11–21, 2011.

[135] Peterman N, Lavi-Itzkovitz A, and Levine E. Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. *Nucleic acids research*, 42(19):12177–12188, 2014.

[136] Aiba H. Mechanism of RNA silencing by Hfq-binding small RNAs. *Current opinion in microbiology*, 10(2):134–139, 2007.

[137] Pfeiffer V, Papenfort K, Lucchini S, Hinton JCD, and Vogel J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nature structural & molecular biology*, 16(8):840–846, 2009.

[138] Wassarman KM, Repoila F, Rosenow C, Storz G, and Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes & development*, 15(13):1637–1651, 2001.

[139] Massé E and Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 99(7):4620–4625, 2002.

[140] Massé E, Escorcia FE, and Gottesman S. Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. *Genes & development*, 17(19):2374–2383, 2003.

[141] Gottesman S. The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, 58:303–328, 2004.

[142] Levine E, Zhang Z, Kuhlman T, and Hwa T. Quantitative characteristics of gene regulation by small RNA. *PLoS biology*, 5(9):e229, 2007.

[143] Morita T, Mochizuki Y, and Aiba H. Translational repression is sufficient for gene silencing by bacterial small noncoding RNAs in the absence of mRNA destruction. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):4858–4863, 2006.

[144] Massé E, Vanderpool CK, and Gottesman S. Effect of RyhB small RNA on global iron use in *Escherichia coli. Journal of bacteriology*, 187(20):6962–6971, 2005.

[145] Urban JH and Vogel J. Translational control and target recognition by *Escherichia coli* small RNAs *in vivo. Nucleic acids research*, 35(3):1018–1037, 2007.

[146] Shimoni Y, Friedlander G, Hetzroni G, Niv G, Altuvia S, Biham O, and Margalit H. Regulation of gene expression by small non-coding RNAs: a quantitative view. *Molecular Systems Biology*, 3(1):138, 2007.

[147] Lavi-Itzkovitz A, Peterman N, Jost D, and Levine E. Quantitative effect of target translation on small RNA efficacy reveals a novel mode of interaction. *Nucleic acids research*, 42(19):12200–12211, 2014.

[148] Tummala SB, Welker NE, and Papoutsakis ET. Design of antisense RNA constructs for downregulation of the acetone formation pathway of *Clostridium acetobutylicum. Journal of bacteriology*, 185(6):1923–1934, 2003.

[149] Nakashima N, Tamura T, and Good L. Paired termini stabilize antisense RNAs and enhance conditional gene silencing in *Escherichia coli. Nucleic acids research*, 34(20):e138–e138, 2006.

[150] Nakashima N and Tamura T. Conditional gene silencing of multiple genes with antisense RNAs and generation of a mutator strain of *Escherichia coli. Nucleic acids research*, 37(15):e103–e103, 2009.

[151] Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, and Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature protocols*, 8(11):2180, 2013.

[152] Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, and Lim WA. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

[153] Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, and Marraffini LA. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic acids research*, 41(15):7429–7437, 2013.

[154] Choudhary E, Thakur P, Pareek M, and Agarwal N. Gene silencing by CRISPR interference in mycobacteria. *Nature communications*, 6:6267, 2015.

[155] Keene JD and Tenenbaum SA. Eukaryotic mRNPs may represent post-transcriptional operons. *Molecular cell*, 9(6):1161–1167, 2002.

[156] Møller T, Franch T, Udesen C, Gerdes K, and Valentin-Hansen P. Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes & development*, 16(13):1696–1706, 2002.

[157] Desnoyers G, Morissette A, Prévost K, and Massé E. Small RNA-induced differential degradation of the polycistronic mRNA iscRSUA. *The EMBO journal*, 28(11):1551–1561, 2009.

[158] Sharma V, Yamamura A, and Yokobayashi Y. Engineering artificial small RNAs for conditional gene silencing in *Escherichia coli*. *ACS synthetic biology*, 1(1):6–13, 2011.

[159] Na D, Yoo SM, Chung H, Park H, Park JH, and Lee SY. Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nature biotechnology*, 31(2):170–174, 2013.

[160] Chen S, Zhang A, Blyn LB, and Storz G. MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*. *Journal of bacteriology*, 186(20):6689–6697, 2004.

[161] Yoo SM, Na D, and Lee SY. Design and use of synthetic regulatory small RNAs to control gene expression in *Escherichia coli*. *Nature protocols*, 8(9):1694, 2013.

[162] Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, and Nielsen LK. The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. *BMC genomics*, 12(1):9, 2011.

[163] UNAFold main page. [http://mfold.rna.albany.edu/].

[164] Markham NR and Zuker M. DINAMelt web server for nucleic acid melting prediction. *Nucleic acids research*, 33(suppl_2):W577–W581, 2005.

[165] Shetty RP, Endy D, and Knight TF. Engineering BioBrick vectors from BioBrick parts. *Journal of biological engineering*, 2(1):5, 2008.

[166] Sambrook J, Fritsch EF, Maniatis T, et al. *Molecular cloning: a laboratory manual.* Number Ed. 2. 1989.

[167] Pasotti L, Politi N, Zucca S, Cusella De Angelis MG, and Magni P. Bottom-up engineering of biological systems through standard bricks: a modularity study on basic parts and devices. *PloS one*, 7(7):e39407, 2012.

[168] Zucca S, Pasotti L, Mazzini G, Cusella De Angelis MG, and Magni P. Characterization of an inducible promoter in different DNA copy number conditions. *BMC bioinformatics*, 13(4):S11, 2012.

[169] Zucca S, Pasotti L, Politi N, Cusella De Angelis MG, and Magni P. A standard vector for the chromosomal integration and characterization of BioBrick™parts in *Escherichia coli*. *Journal of biological engineering*, 7(1):12, 2013.

[170] Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, and Endy D. Measuring the activity of BioBrick promoters using an *in vivo* reference standard. *Journal of biological engineering*, 3(1):4, 2009.

[171] Politi N, Pasotti L, Zucca S, Casanova M, Micoli G, Cusella De Angelis MG, and Magni P. Half-life measurements of chemical inducers for recombinant gene expression. *Journal of biological engineering*, 8(1):5, 2014.

[172] Mutalik VK, Guimaraes JC, Cambray G, Mai Q-A, Christoffersen MJ, Martin L, Yu A, Lam C, Rodriguez C, Bennett G, et al. Quantitative estimation of activity and quality for collections of functional genetic elements. *Nature methods*, 10(4):347–353, 2013.

[173] Lutz R and Bujard H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic acids research*, 25(6):1203–1210, 1997.

[174] Gay RJ, McComb RB, and Bowers GN. Optimum reaction conditions for human lactate dehydrogenase isoenzymes as they affect total lactate dehydrogenase activity. *Clinical chemistry*, 14(8):740–753, 1968.

[175] Canton B, Labno A, and Endy D. Refinement and standardization of synthetic biological parts and devices. *Nature biotechnology*, 26(7):787, 2008.

[176] Zucca S Pasotti L and Magni P. *Modelling for synthetic biology. In: Modeling methodology for physiology and medicine: second edition.* 2013.

[177] Guido NJ, Wang X, Adalsteinsson D, McMillen D, Hasty J, Cantor CR, Elston TC, and Collins JJ. A bottom-up approach to gene regulation. *Nature*, 439(7078):856, 2006.

[178] Selinger DW, Saxena RM, Cheung KJ, Church GM, and Rosenow C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome research*, 13(2):216–223, 2003.

[179] Wang B, Kitney RI, Joly N, and Buck M. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature communications*, 2:508, 2011.

[180] Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, and Church GM. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli. Proceedings of the National Academy of Sciences*, 110(34):14024–14029, 2013.

[181] Levin-Karp A, Barenholz U, Bareia T, Dayagi M, Zelcbuch L, Antonovsky N, Noor E, and Milo R. Quantifying translational coupling in *E. coli* synthetic operons using RBS modulation and fluorescent reporters. *ACS synthetic biology*, 2(6):327–336, 2013.

[182] Pasotti L and Zucca S. Advances and computational tools towards predictable design in biological engineering. *Computational and mathematical methods in medicine*, 2014, 2014.

[183] Lee TS, Krupa RA, Zhang F, Hajimorad M, Holtz WJ, Prasad N, Lee SK, and Keasling JD. BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *Journal of biological engineering*, 5(1):12, 2011.

[184] Sharma V, Sakai Y, Smythe KA, and Yokobayashi Y. Knockdown of recA gene expression by artificial small RNAs in *Escherichia coli. Biochemical and biophysical research communications*, 430(1):256–259, 2013.

[185] Politi N, Pasotti L, Zucca S, and Magni P. Modelling the effects of cell-to-cell variability on the output of interconnected gene networks in bacterial populations. *BMC systems biology*, 9(3):S6, 2015.

[186] Zucca S, Pasotti L, Politi N, Casanova M, Mazzini G, Cusella De Angelis MG, and Magni P. Multi-faceted characterization of a novel LuxR-repressible promoter library for *Escherichia coli*. *PloS one*, 10(5):e0126264, 2015.

[187] Casanova M, Pasotti L, Zucca S, Politi N, Massaiu I, Calvio C, Cusella De Angelis MG, and Magni P. A BioBrick ™-Compatible Vector for Allelic Replacement Using the XylE Gene as Selection Marker. *Biological procedures online*, 18(1):6, 2016.

[188] Song CW, Lee J, and Lee SY. Genome engineering and gene expression control for bacterial strain development. *Biotechnology journal*, 10(1):56–68, 2015.

[189] Link AJ, Phillips D, and Church GM. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. *Journal of bacteriology*, 179(20):6228–6237, 1997.

[190] Blomfield IC, Vaughn V, Rest RF, and BI Eisenstein. Allelic exchange in *Escherichia coli* using the Bacillus subtilis *sac*B gene and a temperature-sensitive pSC101 replicon. *Molecular microbiology*, 5(6):1447–1457, 1991.

[191] Heermann R, Zeppenfeld T, and Jung K. Simple generation of site-directed point mutations in the *Escherichia coli* chromosome using *Red/ET* Recombination. *Microbial cell factories*, 7(1):14, 2008.

[192] Philippe N, Alcaraz J-P, Coursange E, Geiselmann J, and Schneider D. Improvement of pCVD442, a suicide plasmid for gene allele exchange in bacteria. *Plasmid*, 51(3):246–255, 2004.

[193] Li X-t, Thomason LC, Sawitzke JA, Costantino N, and Court DL. Positive and negative selection using the *tet*A-*sac*B cassette: recombineering and P1 transduction in *Escherichia coli*. *Nucleic acids research*, 41(22):e204–e204, 2013.

[194] Pósfai G, Kolisnychenko V, Bereczki Z, and Blattner FR. Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic acids research*, 27(22):4409–4415, 1999.

[195] Warming S, Costantino N, Court DL, Jenkins NA, and Copeland NG. Simple and highly efficient BAC recombineering using galK selection. *Nucleic acids research*, 33(4):e36–e36, 2005.

[196] Mandell DJ, Lajoie MJ, Mee MT, Takeuchi R, Kuznetsov G, Norville JE, Gregg CJ, Stoddard BL, and Church GM. Biocontainment of genetically modified organisms by synthetic protein design. *Nature*, 518(7537):55–60, 2015.

[197] Hamilton M, Aldea M, Washburn BK, Babitzke P, and Kushner SR. New method for generating deletions and gene replacements in *Escherichia coli*. *Journal of Bacteriology*, 171(9):4617–4622, 1989.

[198] Arnaud M, Chastanet A, and Débarbouillé M. New vector for efficient allelic replacement in naturally nontransformable, low-GC-content, gram-positive bacteria. *Applied and environmental microbiology*, 70(11):6887–6891, 2004.

[199] Choffnes ER, Relman DA, Pray L, et al. *The science and applications of synthetic and systems biology: workshop summary.* 2011.

[200] Goh S, Camattari A, Ng D, Song R, Madden K, Westpheling J, and Wong VVT. An integrative expression vector for *Actinosynnema pretiosum*. *BMC biotechnology*, 7(1):72, 2007.

[201] Ingram C, Brawner M, Youngman P, and Westpheling J. *xylE* functions as an efficient reporter gene in Streptomyces spp.: use for the study of galP1, a catabolite-controlled promoter. *Journal of bacteriology*, 171(12):6617–6624, 1989.

[202] Cambridge 2009 iGEM Team. [http://2009.igem.org/team:cambridge].

[203] Santos CN S and Stephanopoulos G. Melanin-based high-throughput screen for L-tyrosine production in *Escherichia coli*. *Applied and environmental microbiology*, 74(4):1190–1197, 2008.

[204] Porcar M, Danchin A, and de Lorenzo V. Confidence, tolerance, and allowance in biological engineering: the nuts and bolts of living things. *BioEssays*, 37(1):95–102, 2015.

[205] Haldimann A and Wanner BL. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *Journal of bacteriology*, 183(21):6384–6393, 2001.

[206] Chen Y-J, Liu P, Nielsen AAK, Brophy JAN, Clancy K, Peterson T, and Voigt CA. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature methods*, 10(7):659–664, 2013.

[207] Emmerson JR, Gally DL, and Roe AJ. Generation of gene deletions and gene replacements in *Escherichia coli* O157: H7 using a temperature sensitive allelic exchange system. *Biological procedures online*, 8(1):153–162, 2006.

[208] Horiyama T and Nishino K. AcrB, AcrD, and MdtABC multidrug efflux systems are involved in enterobactin export in *Escherichia coli*. *PLoS One*, 9(9):e108642, 2014.

[209] Mahalik S, Sharma AK, and Mukherjee KJ. Genome engineering for improved recombinant protein expression in *Escherichia coli*. *Microbial cell factories*, 13(1):177, 2014.

[210] Ginesy M, Belotserkovsky J, Enman J, Isaksson L, and Rova U. Metabolic engineering of *Escherichia coli* for enhanced arginine biosynthesis. *Microbial cell factories*, 14(1):29, 2015.

[211] Datsenko KA and Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, 2000.

List of publications

Articles in peer reviewed journals

- Casanova M, Pasotti L, Zucca S, Politi N, Massaiu I, Calvio C, Cusella De Angelis MG and Magni P. A BioBrick™-Compatible Vector for Allelic Replacement Using the *XylE* Gene as Selection Marker. *Biological Procedures Online*, 18:6, 2016.

- Massaiu I, Pasotti L, Casanova M, Politi N, Zucca S, Cusella De Angelis MG and Magni P. Quantification of the gene silencing performances of rationally-designed synthetic small RNAs. *Systems and synthetic biology*, 9:107-123, 2015.

Contributions to conference proceedings

- Pasotti L, Zucca S, Massaiu I, Casanova M, Bellato M, Mazzini G, Cusella De Angelis MG, Calvio C and Magni P. Efficient conversion of industrial bio-waste into biofuels and bioproducts through *E. coli* and *B. subtilis* synthetic biology. *Synthetic Biology, Engineering, Evolution & Design (SEED).* Canada, June 20-23, 2017.

- Pasotti L, Zucca S, Massaiu I, Casanova M, Bellato M, Mazzini G, Cusella De Angelis MG, Calvio C and Magni P. Efficient conversion of industrial bio-waste into biofuels and bioproducts through *E. coli* and *B. subtilis* synthetic biology. *The Seventh International Meeting on Synthetic Biology (SB7.0).* Singapore, June 13-16, 2017.

- Bellato M, Pasotti L, Casanova M, Massaiu I, Cusella De Angelis MG and Magni P. Rational engineering of protein- and CRISPRi- mediated regulation devices to design predicatable interconnected circuits with reduced cell load. *The Seventh International Meeting on Synthetic Biology (SB7.0).* Singapore, June 13-16, 2017.

- Pasotti L, Zucca S, Casanova M, Bellato M, Massaiu I, Arbuschi M, Murgiano M, Serra A, Cusella De Angelis MG and Magni P. Definition and in-vivo evaluation of mathematical models to predict the effect of copy number variations and cell burden in interconnected synthetic circuits. *Synthetic Biology, Engineering, Evolution & Design (SEED).* Chicago, USA, July 18-21, 2016.

- Massaiu I, Maestri S, Zucca S, Pasotti L, Cusella De Angelis MG and Magni P. Evaluation of constraint-based methods to predict the metabolic phenotype of E. coli under different environmental and genetic conditions. *3rd International Synthetic and Systems Biology Summer School: Abstract Book.* Volterra, IT, July 17-21, 2016.

- Bellato M, Pasotti L, Castronuovo F, Politi N, Casanova M, Massaiu I, Zucca S, Cusella De Angelis MG and Magni P. Study of a genetic negative feedback controller via bottom-up approach and mathematical modelling. *Congress of Italian National Bioengineering Group (GNB).* Naples, IT, June 20-22, 2016.

- Borella E, Carrara L, Lavezzi SM, Massaiu I, Sauta E, Tosca EM, Vitali F, Zucca S, Pasotti L, De Nicolao G and Magni P. Methods and tools for multiscale modelling in Systems Pharmacology: a review. *Twenty-fifth Population Approach Group Europe (PAGE) meeting.* Lisbon, PT, June 7-10, 2016.

- Pasotti L, Zucca S, Casanova M, Massaiu I, Mazzini G, Micoli G, Calvio C, Cusella De Angelis MG and Magni P. Conversion of industrial biowaste into biofuels through syntjetic biology supported by flow cytometry. *XXXIII Conferenza Nazionale di Citometria della SocietÃ Italiana di Citometria (GIC).* Lucca, IT, September 22-25, 2015.

- Pasotti L, Zucca S, Casanova M, Politi N, Massaiu I, Micoli G, Calvio C, Cusella De Angelis MG and Magni P. Methods for genetic optimization of biocatalysts for biofuel production from dairy waste through synthetic biology. *IEEE Engineering in Medicine and Biology Society (EMBC).* Milan, IT, August 25-29, 2015.

- Pasotti L, Zucca S, Casanova M, Politi N, Massaiu I, Cusella De Angelis MG and Magni P. Predictable design in biological engineering: Debugging of synthetic circuits by in vivo and in silico approaches. *Synthetic Biology, Engineering, Evolution & Design (SEED).* Boston, USA, June 10-13, 2015.

- Massaiu I, Pasotti L, Zucca S, Politi N and Magni P. Evaluation of sequence-to-function predictive tools to support the bottom-up design of complex genetic circuits in synthetic biology. *Bioinformatics Italian Society (BITS) annual meeting.* Milan, IT, June 3-5, 2015.

- Casanova M, Zucca S, Pasotti L, Politi N, Massaiu I, Bellato M, Mazzini G, Cusella De Angelis MG and Magni P. Mathematical model-based modular design of genetic circuits to express a target protein at a desired mean and cell-to-cell variability level. *Bioinformatics Italian Society (BITS) annual meeting.* Milan, IT, June 3-5, 2015.

- Massaiu I, Pasotti L, Zucca S, Politi N, Casanova M, and Cusella De Angelis MG Magni P. Gene silencing via small RNAs: fine tuning of synthetic biological circuits and metabolic pathways. *IV Congresso del Gruppo Nazionale di Bioingegneria (GNB).* Pavia, IT, June 25-27, 2014.