

Bayesian Design for Whole Cell Synthetic Biology Models

C. Angione¹, E. Bartocci², L. Bortolussi^{3,4}, P. Lió¹, A. Occhipinti¹, G. Sanguinetti^{5,6}

¹ Cambridge University, UK

² Vienna University of Technology, Austria

³ DMG, University of Trieste, Italy

⁴ CNR/ISTI, Pisa, Italy

⁵ School of Informatics, University of Edinburgh

⁶ SynthSys, Centre for Synthetic and Systems Biology, University of Edinburgh

Abstract. We present a novel approach to the automated design of large scale biological systems. Our model connects multiple levels of cellular organization through a hybrid continuous/discrete and deterministic/stochastic approach. At the core of our model, a hybrid dynamical system represents interactions between gene expression and DNA topology. Gene expression then influences the metabolic level of the cell by affecting the flux distribution in a genome-scale metabolic model of *Escherichia coli*. We implement a Bayesian optimization strategy to select optimal parameters which maximally satisfy a temporal logic formula specifying the desired behavior of the system. Finally, we show preliminary results of a study on the interaction of the master regulator FIS with supercoiling.

1 Introduction

Synthetic Biology [1, 2] is an emerging discipline that aims at the rational design of artificial living systems with a predictable behavior, either by creating new biological entities that do not exist in nature or by redesigning the existing ones. Even though important technological developments have been achieved in this field, *de novo* design of biological models implementing a desired behavior is a major challenge. One of the most formidable obstacles for systems biology is represented by the complex interplay between synthetic networks and the environment in which they are implanted: the non-linear nature of the interactions reflects in the emergence of systemic behavioral properties, not directly derivable from the knowledge of the individual parts. Therefore, it is often impossible to determine whether a theoretically predicted (and possibly *in vitro* verified) behavior of a genetic circuit will be maintained once the circuit is embedded within a cellular context.

To address this challenge, an ambitious research program is being pursued about the development of *whole cell models*, namely dynamical models connecting various functional levels within the cell (genomic, transcriptomic, metabolomic, etc.), with the aim to capture the global effects of perturbation at any single level. Whole cell models vary greatly in the level of detail and biological fidelity: coarse grained models coupling cell growth with gene expression have recently established strikingly simple relationships to emerge [3], however they are necessarily limited in the detail of their predictions. At the other end of the scale, Karr et al. [4] proposed an exceptionally detailed model for one of the simplest organisms, the bacterium *Mycobacter pneumoniae*: while striking predictions could emerge from this model, its calibration required a prodigious amount of data which is unlikely to be available for any other organism in the medium term.

In this paper, we strike a compromise approach to whole cell modelling, retaining some of the detail of large scale models but substantially coarse graining other aspects of cellular function. Specifically, we focus on one particular aspect of cell growth regulation in *E. coli*, i.e. regulation of its energy metabolism by the master regulator FIS. FIS regulates a large number of *E. coli* genes by modulating the coiling of the chromosome: intuitively, opening or closing the DNA regulates access of genetic material to the transcriptional machinery. This mechanism, in turns, self regulates the expression of FIS in a negative feedback loop. The core of our approach is represented by a dynamical hybrid model of FIS regulation: expression of the FIS protein (modelled as an ordinary differential equation) regulates the twisting of the DNA (modelled as a discrete birth/death process representing the number of twists in the DNA) which in turn controls the production of FIS. FIS expression is further connected with the whole cell metabolism by regulating metabolic fluxes through the expression of its targets: thus, the hybrid dynamical system is coupled with a large scale, flux balance analysis (FBA) type model describing the whole cell metabolism. In order to show the relevance to synthetic biology of this approach, we show how phenotypic traits can in principle be controlled by tuning the parameters of the hybrid dynamical

system. To this end, we employ a recently proposed algorithm [5] that enables the design of stochastic models optimally satisfying temporal logic requirements.

The preliminary results of our pipeline, obtained with optimal parameters and with an ODE for the FIS dynamics, show how FIS is responsible for the growth rate and the synthetic yield of *E. coli*. This is achieved by acting on the topology and supercoiling of the bacterial chromosome, modelled by the regulation of 104 gene expression levels. Our approach can be easily generalized adopting different proteins and different types of dynamics for the gene expression levels.

The rest of the paper is organized as follows: in the next section, we present the modelling techniques used, briefly reviewing temporal logic and the optimization techniques used. We then move on to describe in detail the biological case study, justifying our approach from the biological application point of view. Initial results demonstrate the feasibility of this approach in optimizing the parameters in order to ensure stable cell division. We conclude the paper by discussing the methodological and biological significance of our contribution, as well as sketching plans for future work.

2 Methodology

2.1 Models

Here we describe the dynamics of the process regulated by an ODE that we use to drive the FBA, which connects gene expression data, protein amount and flux balance analysis to control the growth rate and the acetate secretion of *E. coli* (see Figure 1). We assume that the parameters of the ODE regulate the boundary conditions of a constrained optimization problem representing the behavior of a large number of variables which are fast compared to gene expression.

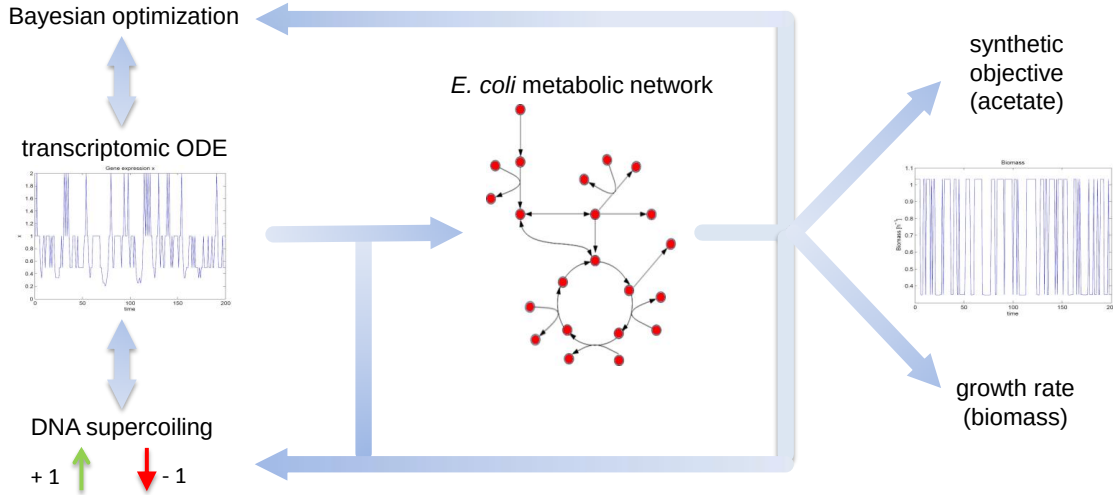


Fig. 1. The dynamical process involving the gene expression level controls the lower- and upper-bounds of the *E. coli* metabolic network (graph in the center; red circles represent metabolites, arrows represent reactions). Using bilevel flux balance analysis, we optimize the network towards the production of biomass and acetate. The output of the model feeds back on the process simulating supercoiling and on the Bayesian optimization algorithm, both used to define the gene expression dynamics. For the sake of simplicity, the DNA supercoiling is modeled as a discrete birth/death rate stochastic process involving the number of *supercoiling packets*, a discrete number associated with the number of twists in the DNA, and therefore correlated with the DNA structure and topology. The birth and death rates depend on the current expression level and on the biomass output. We adopt a Bayesian optimization algorithm to design the stochastic models optimally with the aim of satisfying temporal logic requirements involving the output of the biological model.

We run the coupled model with a set of optimal parameters able to match given system requirements, e.g. maximizing the biomass in a bilevel fashion involving also acetate production [6], and ensuring that the *E. coli* produces at least a given amount of biomass for a certain amount of time. In genome-scale

metabolic reconstructions, the biomass represents the exponential growth rate of the bacterium, while the acetate secretion is often an objective to be maximized and validates the metabolic model by comparison with the biomass [7]. In the case studies of our pipeline, we introduce an ODE coupled with a stochastic process to describe the evolution in time of the expression level of the genes responsible for ATP production. Furthermore, using the same dynamics, we model the upregulation and downregulation effects of the FIS protein on 104 genes. Translating the gene expression levels to the metabolism allows to investigate the biomass formation as function of the amount of FIS available. Then, we use the flux balance analysis to maximize the biomass and obtain feedback onto the metabolism. The differential equation depends on a birth/death stochastic process involving supercoiling of the DNA structure, where an excessive supercoiling slows down the transcription machinery.

Assuming that the rate of change of the gene expression level is proportional to the difference between the current size and some target size, we suppose that the gene expression dynamics follows the monomolecular model. Hence, we assume that a gene with expression level x_m (a continuous variable) satisfies the following ODE:

$$\frac{dx_m}{dt} = a - kx_m. \quad (1)$$

Equation 1 defines the evolution in time of the expression level x_m , i.e. it is produced at a constant rate a (from some reactions which are kept going). Nevertheless, an increased transcription activity needs more enzymes to relax the DNA structure and avoid supercoiling. We control this feedback through the rate k . Supercoiling has been recently proved to affect the regulation of gene expression by changing the shape of DNA and facilitating or repressing transcription initiation. As in most bacteria, DNA in *E. coli* is negatively supercoiled, i.e. twisted against the helical conformation. An excess supercoiling can significantly change gene expression levels [8]. A more convenient form of the above equation to describe the progression of x_m is

$$x_m = (x_0 - a/k)e^{-kt} + a/k. \quad (2)$$

A highly transcribed gene requires more enzymes able to relax and untwist the DNA structure. Therefore, supercoiling can be thought of as a control mechanism that reduces and stabilizes the expression level of a highly expressed gene. In our model, we suppose that the gene expression level is decreased linearly with rate k due to DNA supercoiling. Through the rate $k = Knt$, we modulate the degradation rate so that we model the effect of DNA coiling on the gene expression level. We adopt a birth/death stochastic process with birth rate α and death rate β to regulate the dynamics of the number of supercoiling packets (associated with the number of twists) nt in the DNA structure, where nt represents the difficulty of gene transcription. We set $\alpha = x_m/x_m^{max} + \eta$ and $\beta = 1 - \alpha$, in order to account for the fact that a high gene expression level tends to be decreased because it requires more enzymes to be untwisted; we represent this process by increasing the number of supercoiling packets as the gene expression level increases, due to lack of enzymes to untwist the DNA. This represents a feedback loop on the input: the more the gene expression level, the higher the request for enzymes to relax the helix (simulated by an increase in the number of supercoiling packets); but if these enzymes are not available, the gene expression tends to decrease with a rate depending on the number of supercoiling packets.

2.2 Problem statement

The emergent behavior of the complex dynamics presented in the previous section can be captured using linear temporal logic [9], a very intuitive specification language to express formally the requirement of a system, where the truth of a proposition (i.e. the concentration of a species being greater or smaller than a threshold) can vary in time. In the last years, temporal logic has been enriched with important extensions to deal with dense-time [10], real values [11] and oscillatory patterns [12, 13, 14], and it has become nowadays a key tool to specify complex behavioral patterns emerging in the simulation of biological models [15, 16, 12].

We consider here the Signal Temporal Logic (STL) [11] as our requirement specification language. STL is suitable to characterize behavioral patterns in time series of real values generated during the simulation of a dynamical system. STL provides two different semantics: a Boolean semantics that returns yes/no if the observed trace satisfies or not the STL specification, and a quantitative semantics that also returns a measure of the robustness of the specification. We refer the reader to [17] for further details on the syntax and semantics.

Several recent research works [5, 18, 19, 20] employ the qualitative and quantitative semantics of this logic as a behavioral discrimination function to guide the design of models reproducing properties

observed in the experimental data [5, 18, 19] or to *de novo design* of biological circuits in synthetic biology [20]. The *system design* problem consists in tuning properly the uncertain parameters of biological systems in order to reproduce the desired behavioral properties. Here, we provide a case study by adopting a technique recently developed by some co-authors in [5] where the problem of designing a model from requirements expressed in terms of temporal logic formulae, is recast in terms of learning and maximizing an unknown function representing the likelihood of the parameters.

We now consider the property for the biomass B of *eventually* (**F** means finally) ending up over a certain threshold θ and remaining *always* there (**G** means globally) for an interval of time T :

$$\varphi : \mathbf{F}_{[0, T_{max}-T]} \mathbf{G}_{[0, T]} (B \geq \theta). \quad (3)$$

We want to optimize the birth and death rates (α and β) such that this property results to be the more robust as possible. The objective function of this optimization can be written, according to the STL quantitative semantics, as

$$\max_{t \in [0, T_{max}-T]} (\min_{s \in [t, t+T]} (B(s) - \theta)), \quad (4)$$

where $B(s)$ is the biomass of the bacterium at time s . In other words, we aim at seeking the time t that maximizes the minimum gap between the biomass and the threshold θ in the $[t, t+T]$ time slot. Coupled with the condition $B(s) - \theta \geq 0$, expressed by Equation 3, this ensures that the bacterium achieves an acceptable growth rate throughout the interval $[t, t+T]$.

2.3 Bayesian optimization

While our modelling approach aims at a compact description of cellular behaviour, the dynamics of the coarse grained system are still remarkably complex and are not amenable to naive analysis tools. In particular, the probability of a certain property being satisfied by the system (the *satisfaction function* [21]) is an intractable function of the model parameters, which can only be evaluated approximately through simulation based techniques [22]. Optimizing such a function, i.e. solving the system design problem, cannot therefore be solved using standard optimization techniques.

Recently, some of the authors proposed a statistical solution to this class of design problems, based on a Bayesian approach to estimating the underlying unknown satisfaction function [5]. Here we use this approach to optimize the dynamics of a gene expression level (Figure 1). Briefly, one can abstract the true satisfaction function using a statistical model which provides a probability distribution over continuous functions, a *Gaussian Process* (GP). GPs are a natural functional extension of the multivariate Gaussian distribution; the fundamental property of GPs is that a function is a sample from a GP if and only if all of its finite dimensional marginals (i.e. the distributions over the random variables obtained by evaluating the function at a finite set of points) are Gaussian. This implies that, given noisy observations of the function at a finite set of points, one can obtain an analytic expression for the distribution of the function at any other point, a process called *GP regression* [23].

GP regression provides us with a tool to estimate the satisfaction function at any value of the parameter space, given observations of its values at a finite number of points. However, it falls short of providing a method for optimizing the unknown function: in fact, it can be easily shown that optimizing the *mean* estimate is a suboptimal optimization method which is likely to remain trapped in a local maximum. An elegant solution to the above problem is provided by *active learning*, in particular by Bayesian optimization techniques. The key idea is to also consider the *uncertainty* of the estimated function (which is also computed analytically in GP regression): intuitively, one should explore regions where the maximum *could plausibly be*, i.e. regions in parameter space where there is substantial posterior probability mass for the function to take a high value. We formalize these ideas in a recursive search rule, the so called Gaussian Process Upper Confidence Bound (GP-UCB) algorithm: assume we have computed the average robustness at N parameter values (so that we have N input output pairs). Let $\mu_N(\theta)$ and $\nu_N(\theta)$ be the mean and variance of the GP emulator at a given point in input space θ (recall that the marginal at any point will be Gaussian)¹. We select the parameter value θ_{N+1} for the next function evaluation according to the following rule

$$\theta_{N+1} = \operatorname{argmax}_{\theta} (\mu_N(\theta) + \beta_{N+1} \nu_N(\theta)), \quad (5)$$

where β_{N+1} is a parameter. Thus, the next point for exploration does not maximize the emulated function, but an upper confidence bound at a certain confidence level specified by the parameter β_{N+1}

¹ We now denote the input as θ to emphasize that they are the parameters of a stochastic process

(the quantile can be obtained by applying the inverse probit transform to the parameter). [24] proved that this algorithm converges to the global maximum of the unknown function with high probability (which can be adjusted by varying the algorithm’s parameters).

3 Case study

Growth and adaption in *E. coli* result from the structural coupling between the anabolic and catabolic gene expression, DNA topology and metabolism. These represent the components underlying the organizational complexity of the bacterium. The coordination of transcriptional response is coupled with the DNA topology, whereas both respond to changes of metabolism. In the next sections we define the parameterization of the gene expression, metabolic and topological information we use (see Figure 1 for the general scheme).

3.1 Flux Balance Analysis models of whole cell metabolism.

We implement the gene expression fluctuations by taking into account the *E. coli* genome-scale metabolic reconstruction by Orth et al. [25], as represented in Figure 1. The model has three compartments (cytoplasm, periplasm and extracellular space), 1805 metabolites, and 2583 reactions (including exchange and biomass reactions). The biomass reaction is scaled so that its flux rate equals the exponential growth rate of the bacterium, and represents the production rate of the metabolites necessary for the growth.

Our idea is to use FBA to control *in silico* under- and over-expression of the genes responsible for FIS and ATP synthase, a key precursor for the production of biomass. By taking into account also a synthetic objective (e.g. acetate), our analysis suggests potential regulatory steps to maximize both the natural and the synthetic yield. The FBA framework is based on the steady state condition $dX/dt = Sv = 0$, where X represents the vector of the concentrations of all metabolites, S is the stoichiometric matrix, and v the vector of fluxes. In the *E. coli* model by Orth et al. [25], we module the ATP synthase lower and upper bound according to the variations of x_m . The gene set responsible for the ATP synthase is²:

{((b3736 AND b3737 AND b3738) AND (b3731 AND b3732 AND b3733 AND b3734 AND b3735) AND b3739) OR (b3736 AND b3737 AND b3738) AND (b3731 AND b3732 AND b3733 AND b3734 AND b3735))}.

The AND/OR relation between genes is formalized using a real-valued function that defines the “gene set expression” as function of the expression level of the genes involved. Let x_i and x_j be the gene expression levels of the genes s_i and s_j respectively. We define the expression level of two gene sets s_i AND s_j and s_i OR s_j using the following map:

$$s_i \text{ AND } s_j \mapsto \min\{x_i, x_j\}, \quad (6)$$

$$s_i \text{ OR } s_j \mapsto \max\{x_i, x_j\}. \quad (7)$$

If a gene set requires the activation of both its genes to be activated, its expression level is defined as the lowest expression level among its genes. Conversely, if one of its genes is sufficient to activate the gene set, the gene set expression level is defined as the largest gene expression level. As a result, the bounds of a reaction catalyzed by an enzyme complex associated with the gene products s_1 AND s_2 will be function of $\min\{x_1, x_2\}$, while the bounds of a reaction catalyzed by two isozymes s_3 OR s_4 will be function of $\max\{x_3, x_4\}$. The expression level of more complex gene sets is defined using compositions of function (6) and (7).

Given that the AND relation is translated into the *min* operator, without loss of generality, we choose the expression of the genes b3731, b3732, b3733, b3734, b3735, b3736, b3737, b3738, to be regulated by x_m . These genes are responsible for the subunits of the F_0 and F_1 sectors of the membrane-bound ATP synthase. Formally, we solve the following bilevel problem:

$$\begin{aligned} \max \quad & g^T v \\ \text{such that} \quad & \max \quad f^T v \\ & \text{such that} \quad Sv = 0 \\ & V_i^{\min} h(y_i) \leq v_i \leq V_i^{\max} h(y_i), \end{aligned} \quad (8)$$

² Gene IDs can be located in the KEGG database, see http://www.genome.jp/kegg-bin/show_genomemap_top?org_id=eco

where V_i^{min} and V_i^{max} are the minimum and maximum flux of the wild-type configuration of the model, f and g are n -dimensional arrays of weights defining the natural and synthetic objectives respectively. We set f and g as Boolean vectors, and we select the biomass as natural objective, and acetate as synthetic objective. The function h is a piecewise multiplicative function defined as

$$h(y_i) = (1 + |\log(y_i)|)^{\text{sgn}(y_i-1)}, \quad (9)$$

where $\text{sgn}(y_i-1) = (y_i-1)/|y_i-1|$ and y_i is the gene set expression level of the i th gene set, responsible for the i th reaction in the *E. coli* model. In $y_i = 1$, we set $h(y_i) = 1$. The lower and upper bounds of the i th flux v_i are defined according to the expression level of the genes involved in the i th reaction. The properties $h(y_i) \xrightarrow{y_i \rightarrow 0} 0$ and $h(1) = 1$ ensure that the approaches performing Boolean gene knockouts are a particular case of our approach. The outer maximization problem is subject to the inner one. More specifically, the inner maximization finds the distributions of flux in the network such that the growth rate is maximized. In the outer maximization, all the unregulated fluxes are then distributed such that the synthetic objective is maximized. We solve the bilevel problem using the GLPK solver. As a result, the ATP synthase flux follows the gene expression dynamics x_m .

3.2 Linking FIS and supercoiling: biology and topological model order reduction

The bacterial transcriptional machinery is strongly positively correlated with the growth rate and both of them are tightly controlled in different ways. The control is based on the coordinated action of nucleoid associated proteins (among which the most relevant is the DNA bending protein FIS) and the changes of the negatively supercoiled state of DNA that is unwound relatively to its unconstrained state. The availability of nutrients favors the exponential growth of bacteria, which is sustained by the high production rate of ribosomal RNA and transfer RNA. The rates slow down when bacteria enter into stationary phase. The FIS protein (factor for inversion stimulation) is one of the most abundant DNA-bending proteins in *E. coli* under rapid growth conditions, and is encoded by the *yhdG-fis* operon. This protein facilitates the initiation of the transcription by making contacts with the α and σ subunits of RNA polymerase and by binding to helically phased sites (*global positive control on the metabolism*). The transcription from the FIS main promoter is repressed by the FIS protein itself (*local gene-protein negative autoregulation*) and stimulated by DNA negative supercoiling. There are evidences suggesting that the increase of FIS produces an increase in ppGpp concentration which in turn generates a substantial decrease of global DNA superhelicity, via topoisomerases (*global negative control on the metabolism*).

The direct influence of FIS on supercoiling is subtle and of different types; for example, FIS is a negative regulator of the *gyrA* and the *gyrB* genes that encode the A and B subunits of DNA gyrase and it is also involved in the complex regulation of the *topA* gene that encodes the DNA-relaxing topoisomerase I [26, 27, 28]. Therefore, FIS acts as a topological buffer, constraining DNA supercoiling. Note that the supercoiling, and in general any torsional stress, results in two forms of DNA molecule twisting: a change in the helical repeat, that is a change in the number of double helical turns, (called twists, Tw), and also a twisting of the long axis of the DNA molecule, which is a different type of turn (a writhe, Wr). They are topologically equivalent (called toroidal and plectonemic forms), so that in absence of any DNA strand breakage or topoisomerase activity, there is interconvertibility of twists and writhes and the following expression stands as a constant: $\delta Lk = \delta Tw + \delta Wr$ [29]. Here Lk is the deficit of linking, that provides a meaningful way to express the supercoiling through the number of times one chain crosses the other. Defined in this way, the absolute linking number of a domain is quantized and must be integer.

Changes in supercoiling affect promoter activities and alter the gene expression of many genes. Supercoiling responds to environmental changes and cellular energetics. Changes in cellular energetic cause perturbations of the balance between relaxing and supercoiling activities of gyrase, e.g. the ratio ATP/ADP changes the gyrase activity. This results in lowering the negative supercoiling which in turn increases twists. Although further topological complexity has been investigated, for sake of model order reduction we simply consider a positive discrete quantity, termed *twist* or *supercoiling packet* (to emphasize that it is discrete). This quantity sums all the supercoiling contributions from FIS (and other similar DNA bending proteins).

3.3 Dynamic hybrid model: FIS regulation of chromatin conformation in *E. coli*

On a genomic scale, the expression level of 231 genes was shown to be significantly altered by FIS during one or more growth stages of *E. coli* [30], including a large number of negatively regulated genes. In

our analysis, we consider the 104 genes implemented in the *E. coli* model. We assume a linear relation between the amount of FIS protein and the fold regulation of the gene expression of these 104 genes, 76 of which are upregulated while 28 are downregulated in presence of FIS.

The final bounds for the flux rates are determined using Equation 9, and also involve a coefficient γ representing the importance of the genes involved in the reaction. The γ value of each gene is inversely proportional to the variance of the gene expression level from the compendium by Faith et al. [31]. When more than one gene appears in a gene set (i.e., a Boolean expression involving genes), the coefficient γ for the gene set is computed using the same approach used to map gene expression to gene set expression (Equations 6 and 7): we start from the gene set, then we substitute the gene name with its γ value, and finally we substitute the AND and OR Boolean relations into *max* and *min*.

The concentration of FIS also affects the organization of bacterial chromatin, causing changes in many gene expression levels. For the values of y_i , here we adopt a simple linear model of gene expression $y_i = F \cdot \bar{y}_i$, where F is the amount of FIS and $\bar{y}_i$ is the fold change measured experimentally between the wild type and the FIS strains. As a result, the last constraint for the bilevel problem (8) becomes

$$V_i^{min} h(F \cdot \bar{y}_i) \leq v_i \leq V_i^{max} h(F \cdot \bar{y}_i), \quad (10)$$

where

$$h(y_i) = (1 + \gamma |\log(y_i)|)^{\text{sgn}(y_i - 1)}. \quad (11)$$

In $y_i = 1$, we set $h(y_i) = 1$. The amount of FIS is regulated by the stochastic dynamical process controlled by Equation 1. Furthermore, in this case study we add a negative feedback loop starting from the biomass output onto the amount of FIS. Specifically, the biomass plays a role in the death rate of the stochastic process. In *E. coli*, FIS is able to contrast the excessive levels of supercoiling by changing the topology of DNA [32]. Since the changes of FIS are optimized towards the production of biomass, we discretize the amount of FIS and we let the amount of biomass feed back onto the death rate of the stochastic process regulating supercoiling in such a way that, at each step, the feedback is achieved by defining the death rate $\beta = B/B_{max}$, where B represents the current biomass and $B_{max} = 1.26 \text{ h}^{-1}$ is the maximum biomass that can be reached in aerobic conditions by the *E. coli* model [33].

4 Results and Discussion

In this section, we present an early case study on the approach reported in the previous sections. Since we have yet fully investigated the models, we show some preliminary results. Our aim is to find the optimal values of the birth and death rates (α and β respectively) that maintain the biomass value above a given threshold θ for a fixed time T . Hence, we use the method in Section 2.3 to find the values of a and η optimizing the objective function (4). For the initial tests, we fixed $\theta = 1$, $T = 3$ and analyzed the models with three different initial numbers of supercoiling packets nt_i (associated with DNA twists) equal to 1, 10 and 20. In Figure 2 (a), we plot the variation of ATP gene expression level that after an initial increase, oscillates around 6 for $nt_i = 10$ and around 9 for $nt_i = 20$. We also plot in Figure 2 (b) the evolution of the number of supercoiling packets. When the initial number of packets is equal to 10 or 20, nt oscillates around the initial value. For $nt_i = 1$ the progression of the number of packets shows a constant increase up to 21 twists. In conclusion, for $nt_i = 10$ and 20 we observe a similar trend with oscillations around a fixed value, while for $nt_i = 1$ the trend is totally different. More investigations are required to better understand the influence of the initial number of supercoiling packets, but we can suppose that for $nt_i = 10$ and 20 the model oscillates around an equilibrium point. In our analysis we also focused on the FIS variations; in Figure 3 (a) and (b), we plot the quantity of FIS against the maximum biomass and acetate reached by the bacterium. We report the results for different initial conditions based on the initial number of supercoiling packets varying in $\{1, 10, 20\}$. As shown in Figure 3 (a), the biomass increases almost exponentially for each value of nt_i , reaching quickly the maximum value for $nt_i = 10$. The trends of the plots for $nt = 10$ and 20 are again similar in this analysis and totally different from the case $nt_i = 1$ which gets the lowest maximum value. Analyzing the variation of acetate against FIS, Figure 3 (b) shows that for $nt_i = 1$ acetate is almost constant as FIS increases. For $nt_i = 10$ and $nt_i = 20$, the quantity of acetate is initially constant, but decreases when the amount of FIS is higher than 10, reaching the minimum value of 4.4.

These preliminary analyses advance a possible relation between the number of initial supercoiling packets (associated with DNA twists) and the variation of biomass and acetate, opening a new scenario for more and deeper tests. It is noteworthy that the behavior of the acetate curves is opposite to that of

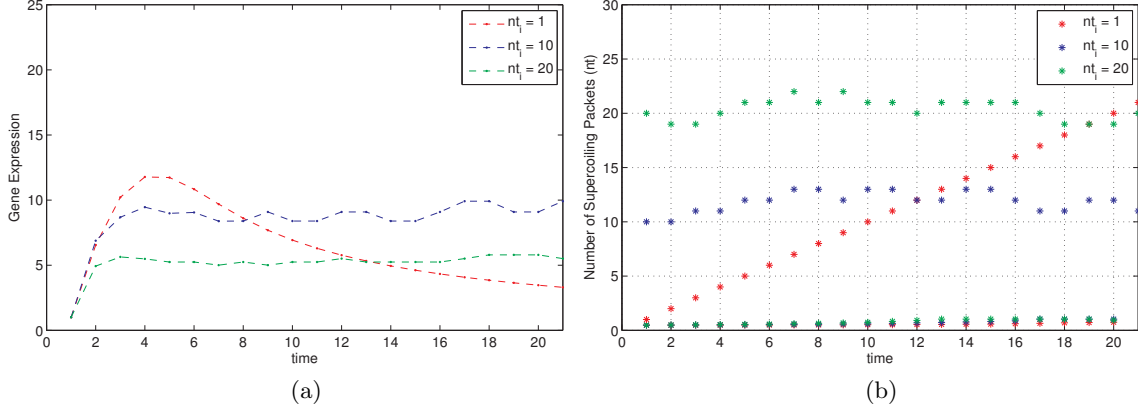


Fig. 2. (a) ATP Gene expression level and (b) number of supercoiling packets against time. The optimal a and η were fixed and the number of initial supercoiling packets nt_i (associated with DNA twists) varied in $\{1, 10, 20\}$. As nt_i increases, the maximum of the gene expression level decreases. In terms of supercoiling packets, if nt_i equals 10 or 20, the number of packets remains almost constant, while for $nt_i = 1$ it increases linearly, indicating that a non-supercoiled DNA structure is not a stable configuration for the cell and for synthetic biology applications.

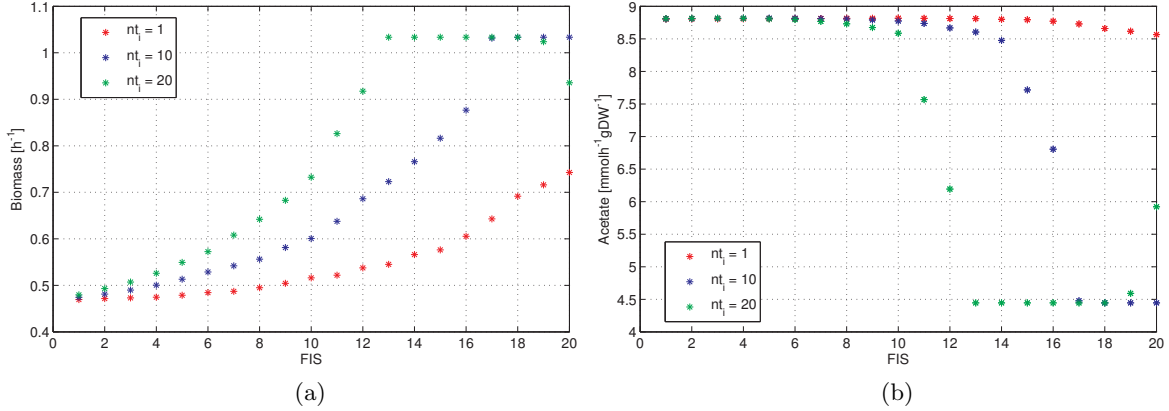


Fig. 3. (a) Variation of FIS against biomass and (b) variation of FIS against acetate. The optimal a and η were fixed and the number of initial supercoiling packets nt_i (associated with DNA twists) varied in $\{1, 10, 20\}$. For every value of nt_i , the biomass increases as the FIS value rises, reaching the maximum value for $nt_i = 10$. The acetate gets the highest values for $nt_i = 1$.

biomass. This is in agreement with experimental evidences and with the results obtained by Palsson's group [7]. Note that the behaviors of the growth rate become more linear and reach quicker the plateau when the initial number of supercoiling packets is large enough (see Figure 3 (a)), suggesting that synthetic biology applications may be conducted with higher nt . This work, although based on FIS, could be extended to other regulatory proteins that generate temporal gradients and trigger the overall transcriptional program. We stress that our framework is general and could be used to study interaction and interferences among temporal gradients of regulatory proteins.

Interestingly, the framework opens the space to reverse engineering, i.e. the identification of those genes that may be altered to produce the desired STL-based design circuit. Modifications of promoters for different genes may be coupled with different optimization formulas, leading to building up very flexible timer circuits for synthetic biology toolboxes with different performances, e.g. synchronous/asynchronous signaling protocols.

Bibliography

- [1] P. Fu and Panke S. *Systems Biology and Synthetic Biology*. John Wiley & Sons, 2009.
- [2] Z. Szallasi, J. Stelling, and Periwal V. *System Modelling in Cellular Biology: from concepts to nuts and bolts*. The Mit Press, 2006.
- [3] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330:1099–1102, 2010.
- [4] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garci, John I. Glass, and Markus W. Cover. A whole cell model predicts phenotype from genotype. *Cell*, 150:389–401, 2012.
- [5] L. Bortolussi and G. Sanguinetti. Learning and Designing Stochastic Processes from Logical Constraints. In *Proc. of QEST 2013: the 10th International Conference on Quantitative Evaluation of Systems, Buenos Aires, Argentina, August 27-30*, volume 8054 of *Lecture Notes in Computer Science*, pages 89–105. Springer-Verlag, 2013.
- [6] C. Angione, G. Carapezza, J. Costanza, P. Lio, and G. Nicosia. Pareto optimality in organelle energy metabolism analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99):1–1, 2013.
- [7] Amit Varma and Bernhard O Palsson. 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.
- [8] Marc Drolet. Growth inhibition mediated by excess negative supercoiling: the interplay between transcription elongation, R-loop formation and DNA topology. *Molecular microbiology*, 59(3):723–730, 2006.
- [9] Amir Pnueli. The temporal logic of programs. *IEEE Annual Symposium on Foundations of Computer Science*, pages 46–57, 1977.
- [10] R. Alur, T. Feder, and T. A. Henzinger. The Benefits of Relaxing Punctuality. *J. ACM*, 43(1):116–146, 1996.
- [11] O. Maler and D. Nickovic. Monitoring Temporal Properties of Continuous Signals. In *Proc. of FORMATS-FTRTFT 2004: the Joint International Conferences on Formal Modeling and Analysis of Timed Systmes and Formal Techniques in Real-Time and Fault -Tolerant Systems, Grenoble, France, September 22-24*, volume 3253 of *Lecture Notes in Computer Science*, pages 152–166. Springer-Verlag, 2004.
- [12] A. Donzé, O. Maler, E. Bartocci, D. Nickovic, R. Grosu, and S. A. Smolka. On Temporal Logic and Signal Processing. In Supratik Chakraborty and Madhavan Mukund, editors, *Proc. of ATVA 2012: 10th International Symposium on Automated Technology for Verification and Analysis, Thiruvananthapuram, India, October 3-6*, volume 7561 of *Lecture Notes in Computer Science*, pages 92–106. Springer-Verlag, 2012.
- [13] E. Bartocci, F. Corradini, E. Merelli, and L. Tesei. Model checking biological oscillators. *Electr. Notes Theor. Comput. Sci.*, 229(1):41–58, 2009.
- [14] Petr Dluhos, Lubos Brim, and David Safránek. On expressing and monitoring oscillatory dynamics. In *Proc. of HSB 2012: the First International Workshop on Hybrid Systems and Biology*, volume 92 of *EPTCS*, pages 73–87, 2012.
- [15] E. Bartocci, P. Liò, E. Merelli, and N. Paoletti. Multiple Verification in Complex Biological Systems: The Bone Remodelling Case Study. *T. Comp. Sys. Biology*, 14:53–76, 2012.
- [16] R. Grosu, G. Batt, F. Fenton, J. Glimm, C. Le Guernic, S. Smolka, and E. Bartocci. From Cardiac Cells to Genetic Regulatory Networks. In *Proc. of CAV 2011: the 23rd International Conference on Computer Aided Verification*, volume 6806 of *Lecture Notes in Computer Science*, pages 396–411. Springer Berlin / Heidelberg, 2011.
- [17] A. Donzé and O. Maler. Robust satisfaction of temporal logic over real-valued signals. In *Proc. of FORMATS 2010: the 8th International Conference on Formal Modeling and Analysis of Timed Systems, Klosterneuburg, Austria, September 8–10*, volume 6246 of *Lecture Notes in Computer Science*, pages 92–106. Springer-Verlag, 2010.
- [18] A. Donzé, E. Fanchon, L. M. Gattepaille, O. Maler, and P. Tracqui. Robustness analysis and behavior discrimination in enzymatic reaction networks. *PLoS One*, 6(9):e24246, 2011.
- [19] E. Bartocci, L. Bortolussi, L. Nenzi, and G. Sanguinetti. On the Robustness of Temporal Properties for Stochastic Models. In *Proc. of HSB 2013: Second International Workshop on Hybrid Systems and Biology, Taormina, Italy, 2nd September 2013*, volume 125 of *EPTCS*, pages 3–19, 2013.

- [20] E. Bartocci, L. Bortolussi, and L. Nenzi. A Temporal Logic Approach to Modular Design of Synthetic Biological Circuits. In *Proc. of CMSB 2013: the 11th International Conference on Computational Methods in Systems Biology, IST Austria, Klosterneuburg, Austria, September 23-25*, volume 8130 of *Lecture Notes in Computer Science*, pages 164–178. Springer-Verlag, 2013.
- [21] L. Bortolussi and G. Sanguinetti. Smoothed model checking for uncertain CTMC. In *arXiv preprint 1402.1450*, 2014.
- [22] S. K. Jha, E. M. Clarke, C. J. Langmead, A. Legay, A. Platzer, and P. Zuliani. A bayesian approach to model checking biological systems. In *Proc. of CMSB 2009: the 7th International Conference on Computational Methods in Systems Biology, CMSB '09*, pages 218–234, Berlin, Heidelberg, 2009. Springer-Verlag.
- [23] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [24] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.
- [25] J.D. Orth, T.M. Conrad, J. Na, J.A. Lerman, H. Nam, A.M. Feist, and B.Ø. Palsson. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. *Molecular systems biology*, 7(1), 2011.
- [26] Andrew DS Cameron, Daniel M Stoebe, and Charles J Dorman. DNA supercoiling is differentially regulated by environmental factors and FIS in *Escherichia coli* and salmonella enterica. *Molecular microbiology*, 80(1):85–101, 2011.
- [27] Robert Schneider, Andrew Travers, Tamara Kutateladze, and Georgi Muskhelishvili. A DNA architectural protein couples cellular physiology and DNA topology in *Escherichia coli*. *Molecular microbiology*, 34(5):953–964, 1999.
- [28] Dalit Weinstein-Fischer and Shoshy Altuvia. Differential regulation of *Escherichia coli* topoisomerase I by Fis. *Molecular microbiology*, 63(4):1131–1144, 2007.
- [29] Chris R Calladine and Horace Drew. *Understanding DNA: the molecule and how it works*. Academic press, 1997.
- [30] Meranda D Bradley, Michael B Beach, AP Jason de Koning, Timothy S Pratt, and Robert Osuna. Effects of fis on escherichia coli gene expression during different growth stages. *Microbiology*, 153(9):2922–2940, 2007.
- [31] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.
- [32] Andrew Travers, Robert Schneider, and Georgi Muskhelishvili. DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. *Biochimie*, 83(2):213–217, 2001.
- [33] Jole Costanza, Giovanni Carapezza, Claudio Angione, Pietro Lió, and Giuseppe Nicosia. Robust design of microbial strains. *Bioinformatics*, 28(23):3097–3104, 2012.