University of Nebraska - Lincoln

[DigitalCommons@University of Nebraska - Lincoln](#)

2016

Mutual Information Upper Bound of Molecular Communication Based on Cell Metabolism

Massimiliano Pierobon

Zahmeeth Sakkaff

Jennie L. Catlett

Nicole R. Buan

Mutual Information Upper Bound of Molecular Communication Based on Cell Metabolism

Massimiliano Pierobon and Zahmeeth Sakkaff
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska 68588 USA
Email: {pierobon,zsayedsa}@cse.unl.edu

Jennie L. Catlett and Nicole R. Buan
Department of Biochemistry
University of Nebraska-Lincoln
Lincoln, Nebraska 68588 USA
Email: {jcatlett,nbuan}@unl.edu

Abstract—Synthetic biology is providing novel tools to engineer cells and access the basis of their molecular information processing, including their communication channels based on chemical reactions and molecule exchange. Molecular communication is a discipline in communication engineering that studies these types of communications and ways to exploit them for novel purposes, such as the development of ubiquitous and heterogeneous communication networks to interconnect biological cells with nano and biotechnology-enabled devices, *i.e.*, the Internet of Bio-Nano Things. One major problem in synthetic biology stands in the development of reliable techniques to control the engineered cells from the external environment. In this paper, molecular communication concepts are applied to study the potential of cell metabolism, and its regulation, to serve this purpose. In particular, a communication system abstraction is introduced to obtain a binary encoder model of the mechanisms underlying the regulation of the chemical reactions activity within the cell metabolism in function of the chemical composition of the external environment. Stemming from this model, an upper bound to the information theoretic mutual information is obtained through the use of a well-known and computationally efficient metabolic simulation technique. This upper bound stands as a theoretical limit of the ability of a particular cell to internally represent the information contained in the chemical composition of the external environment. Numerical results related to the metabolism of the *E. coli* bacterium are presented to evaluate the proposed approach.

Index Terms—Molecular communication, information theory, cell metabolism, regulation of gene expression, flux balance analysis

I. INTRODUCTION

Molecular communication is one of the latest frontiers in communication engineering [1], where tools from computer communications, information theory, signal processing, and wireless networking are applied to the domain of chemical reactions and molecule exchange. Recent results in molecular communication research range from theoretical studies of the communication channels and the expression of their communication capacity [4], [10], [24], [27], to the more practical design of suitable modulation and coding techniques [19], and networking protocols [11].

Synthetic biology is today providing novel tools for the design, realization, and control of biological processes through the programming of cells' genetic code [13]. These tools are allowing engineers to study and access the basis of molecular information processing in biological cells, which can be potentially utilized for the realization of practical molecular communication systems [17], [23]. The future pervasive deployment of genetically engineered cells and their interaction with other bio, micro and nano-technology enabled devices through molecular communication systems and networks has been recently envisioned as the novel paradigm of the Internet of Bio-Nano Things [2]. These ubiquitous and heterogeneous communications will enable advanced applications in many fields, including medicine (*e.g.*, developing bio-compatible diagnosis and treatment systems), industry (*e.g.*, biologically-controlled food production), and agriculture (*e.g.*, monitoring and control of soil chemical and microbiological status).

One major problem in synthetic biology stands in the control from the external environment of the internal functionalities of genetically-engineered cells. Various techniques to realize this control have been explored, such as the use of light, i.e., optogenetics [29], magnetic fields, i.e., magnetic nanoparticles [8], and dedicated signaling circuits [15].

In this paper, we apply molecular communication concepts to study the potential of cell metabolism, and its regulation, to channel information from the outside environment into the cell. In particular, we develop a communication system abstraction of the mechanisms underlying the regulation of the chemical reactions activity within the cell metabolism in function of the chemical composition of the external environment. According to this abstraction, we model the enzyme expression regulation as a binary encoder of the information contained in the concentration of input chemical compounds. Inspired by [25], we apply information theory tools to express the performance of this encoder in terms of steady-state mutual information. Subsequently, we define an upper bound to this mutual information that can be easily quantified through the use of a well-known and computationally efficient metabolic simulation technique, which relies only on the *a priori* knowledge of the cell DNA code (genome). Finally, we present numerical results obtained by analyzing with the proposed method the *E. coli* bacterium metabolism and its regulation with respect to two different input chemical compounds.

The paper is organized as follows. In Sec. II we briefly review the basic concepts in cell metabolism and present the proposed molecular communication abstraction. In Sec. III we express the steady-state mutual information of the proposed

system and we define an upper bound based on the Flux Balance Analysis (FBA) simulation technique. In Sec. IV we present numerical results obtained for the *E. coli* bacterium, and in Sec. V we conclude the paper.

II. REGULATION OF CELL METABOLISM AS A MOLECULAR COMMUNICATION SYSTEM

A. Background

Cell metabolism is the complex network of chemical reactions that underlie cell growth and reproduction [21]. These chemical reactions are chained into pathways where input chemical compounds are broken down generating energy, and at the same time cellular components, or biomass, are built up consuming energy. Most of these chemical reactions do not take place spontaneously, but they are catalyzed by enzymes, special proteins synthesized (expressed) from genes written in the cell's DNA. Cells have mechanisms to control the activity of some of these enzymes [18], therefore adapting the rate at which the corresponding catalyzed reactions occur as function of the input chemical compounds available in the environment, or other environmental parameters, such as temperature or pH. Among different adaptation mechanisms, in this paper we focus on the regulation of enzyme expression from their corresponding DNA genes as function of the input chemical compounds.

The processes underlying the regulation of enzyme expression can be mathematically formalized as follows. Transcription factors are special proteins that change their state between inactive (TF) and active (TF^*) depending on environmental conditions, such as the quantity of determinate chemical compounds in the cell metabolism at a specific moment. When active, transcription factors interact with DNA genes that encode metabolic enzymes to alter their expression. The rate at which each enzyme is expressed directly influences the rate of the corresponding catalyzed reaction. According to a commonly accepted model in biology, given a determinate concentration of active transcription factors $[TF^*]$, the rate R_e at which an enzyme is expressed is given by one of the following two sigmoidal expressions, called Hill's functions [3], [28]:

$$R_e = \frac{\beta[TF^*]^n}{K_d^n + [TF^*]^n} \text{ if activation}, \quad (1)$$

$$R_e = \frac{\beta}{1 + \left(\frac{[TF^*]}{K_d}\right)^n} \text{ if repression}; \quad (2)$$

where β is the maximum expression level of the enzyme, n is the Hill's coefficient having values between 1 and 4 depending on how many transcription factors cooperatively interact with the DNA gene, and K_d is the equilibrium constant [28]. Equation (1) models the situation where a higher concentration of transcription factors increases the enzyme expression from zero to β (activation), while (2) models the opposite (repression). These expressions show a sigmoidal behavior [28] with respect to the active transcription factor concentration $[TF^*]$, with most of the curve values close to two distinct states, namely, zero and maximum expression
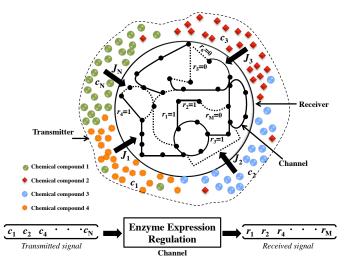

Fig. 1: Sketch of the proposed molecular communication system based on cell metabolism.

level β. As a consequence, the commonly accepted logical approximation of the Hill's function expresses the relation between the concentration of active transcription factors $[TF^*]$ and the enzyme expression rate R_e as follows [3]:

$$R_e \simeq \beta H([TF^*] - K_d) \text{ if activation}, \quad (3)$$
$$R_e \simeq \beta H(K_d - [TF^*]) \text{ if repression};$$

where $H(.)$ is the Heaviside step function, equal to 1 when the argument is positive, and 0 *vice versa*. According to the logical approximation in (3), the enzyme expression, and the activity of the corresponding chemical reaction within the cell metabolism, can be either ON (= maxima enzyme expression rate and corresponding rate of the reaction) or OFF (= no expressed enzyme and absence of the corresponding chemical reaction in the cell metabolism) depending on the quantity of specific chemical compounds in the cell metabolism.

B. Molecular Communication System Abstraction

In this paper, as shown in Fig 1, we propose to abstract cell metabolism, and its aforementioned adaptation mechanisms, as a molecular communication system. In this abstraction, the **Transmitter** is represented by the environment surrounding the cell, where the *Transmitted Signal* is the set of chemical compounds present in this environment that are input of the pathways that compose the cell metabolic network. The **Channel** is given by the mechanisms that regulate the expression of determinate enzymes in function of the chemical compounds in input, and the **Receiver** is the cell itself, where the *Received Signal* is the resulting aforementioned activity (ON/OFF) of the chemical reactions catalyzed by these enzymes. This abstraction is more formally expressed as

$$\{c_1, c_2, ...c_N\} \xrightarrow[\text{Regulation}]{\text{Enzyme Expression}} \{r_1, r_2, ...r_M\}, \quad (4)$$

where c_i is the concentration (number of molecules per unit volume) of the chemical compound i, N is the number of chemical compounds present in the environment surrounding

the cell and input of the metabolic pathway network, r_i is a binary value equal to 1 if the enzyme-expression-regulated reaction i is ON, and equal to 0 if the same reaction is OFF, M is the number of enzyme-expression-regulated reactions that change their state upon variations in the concentrations of input chemical compounds c_i. According to this abstraction, the enzyme expression regulation can be viewed as a binary encoding of the information contained in the chemical composition of the environment surrounding the cell.

In the following, we propose a method to obtain a preliminary analysis of the steady-state mutual information of the proposed molecular communication system through the definition of a theoretical upper bound and a well-known and computationally efficient simulation technique based on the cell DNA code (genome) information.

III. STEADY-STATE MUTUAL INFORMATION OF THE PROPOSED MOLECULAR COMMUNICATION SYSTEM

In this paper, we define the steady-state mutual information I of the aforementioned molecular communication system as the amount of information about the chemical composition of the surrounding environment measured in bits that a cell is able to represent in the binary state of its enzyme-expression-regulated metabolic reactions at steady state, after any evolution of the enzyme-expression regulation channel. According to information theory [6], this mutual information is defined as follows:

$$I(\{c_i\}_{i=1}^{N}; \{r_i\}_{i=1}^{M}) = H(\{c_i\}_{i=1}^{N}) - H(\{c_i\}_{i=1}^{N} \,|\, \{r_i\}_{i=1}^{M}),$$
(5)

where the input entropy $H(\{c_i\}_{i=1}^{N})$ is defined as

$$H(\{c_i\}_{i=1}^{N}) = -\int P\left(\{c_i\}_{i=1}^{N}\right) \log_2 P\left(\{c_i\}_{i=1}^{N}\right) d\{c_i\}_{i=1}^{N},$$
(6)

where the integration $\int$ is performed throughout the possible values that the set of chemical compound concentrations $\{c_i\}_{i=1}^{N}$ can assume. The conditional entropy of the input given the output $H(\{c_i\}_{i=1}^{N} \,|\, \{r_i\}_{i=1}^{M})$ is then defined as follows:

$$H(\{c_i\}_{i=1}^{N} \,|\, \{r_i\}_{i=1}^{M}) = -\sum_{k=1}^{K} P\left(\left[\{r_i\}_{i=1}^{M}\right]_k\right)$$
(7)
$$\int P\left(\{c_i\}_{i=1}^{N} \,\Big|\, \left[\{r_i\}_{i=1}^{M}\right]_k\right)$$
$$\log_2 P\left(\{c_i\}_{i=1}^{N} \,\Big|\, \left[\{r_i\}_{i=1}^{M}\right]_k\right) d\{c_i\}_{i=1}^{N};$$

respectively, where K is equal to the total number of different sets of binary values at the output of the system $\left[\{r_i\}_{i=1}^{M}\right]_k$, resulting from the all the possible values that the input chemical compound concentrations $\{c_i\}_{i=1}^{N}$ can assume, and $P(.)$ is the probability distribution of the argument random variable/s. In the aforementioned definition of mutual information, we are ignoring possible memory in the system, *i.e.*, the values in $\{r_i\}_{i=1}^{M}$ could depend on the past trajectory of the values of the input concentrations $\{c_i\}_{i=1}^{N}$. This might be the effect of hysteretic behaviors in the gene regulatory functions, which we

currently ignore given assumptions stated in Sec. II, with the justification that many of these mechanisms are even poorly understood in biology [7].

A. Estimation of Optimal Enzyme Expression Regulation through Flux Balance Analysis

Flux Balance Analysis (FBA) is a well-known and computationally efficient mathematical method that allows to simulate cell metabolism by computing an optimal solution given determinate constraints [22]. As detailed in the following, by stemming from the knowledge of the cell's genetic code, or genome, through FBA we are able to obtain an estimate of the state $\{r_i^*\}_{i=1}^{M}$ of the aforementioned enzyme-expression-regulated chemical reactions that results into an overall maximum production of biomass in the cell. The FBA-estimated chemical reaction states $\{r_i^*\}_{i=1}^{M}$ are those that maximize the growth of the cell given a chemical composition of the surrounding environment $\{c_i\}_{i=1}^{N}$, and represent the best regulation of these chemical reactions that the cell might ever achieve. The aforementioned mechanisms of activation or repression that might be in place for the regulation of enzyme expressions have been most probably acquired through evolution, and tend to reach this optimal solution, although they might just realize a subset of the needed reaction state adaptations [7]. The estimation of the optimal enzyme expression regulation through FBA can be formalized as follows:

$$\{c_1, c_2, ...c_N\} \xrightarrow[\text{Analysis}]{\text{Flux Balance}} \{r_1^*, r_2^*, ...r_M^*\},$$
(8)

In particular, FBA is based on the knowledge of all the possible chemical reactions that might be active within the cell metabolic network in any possible environmental condition, termed GEnome scale Model (GEM) [16]. Briefly, a GEM is computed from the knowledge of the DNA code, or genome, of the particular cell, and by searching for known genes that encode metabolic enzymes, and therefore possibly activating metabolic chemical reactions, which are described in extensively curated online catalogs. Subsequently, further chemical reactions are included in the GEM through comparisons with the genomes and the corresponding known metabolic pathways of other similar organisms that have been already extensively studied and annotated. Once a GEM is obtained, the set of possible metabolic reactions is expressed through the stoichiometric matrix S, where each row represents a chemical compound possibly participating in metabolic reactions, and each column represents a metabolic reaction possibly active in the cell metabolism. Each entry of the stoichiometric matrix S is the stoichiometric coefficient that indicates how many molecules of a chemical compound are consumed (coefficient < 0) or produced (coefficient > 0) in one of the possible reactions. The FBA solution in terms of $\mathbf{v}^*$, which is a column vector that contains the optimal flux of each reaction, defined as number of molecules per unit volume and unit time that are consumed/produced by that reaction, is obtained through

a Linear Program (LP) formalized as follows [22]:

$$\text{maximize} \quad a'\mathbf{v}$$
$$\text{subject to} \quad S\mathbf{v} = 0$$
$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \,,$$

where a is a column vector that contains the weight coefficients of the fluxes that the FBA optimizes. In our case, the entries of a are equal to 1 only at the indexes corresponding to the chemical compounds that are considered part of the aforementioned biomass produced by the cell and responsible for cell growth, and other entries as 0. The column vectors $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$ constrain the minimum and maximum flux, respectively, of each corresponding reaction considered in the FBA, and define the space where the LP searches for the optimal solution. The values of $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$ are set to reasonable biological limiting values [22], with the exception of the reaction corresponding to the uptake of the input chemical compounds present in the surrounding environment $\{c_i\}_{i=1}^{N}$ for which we are estimating the chemical reaction states $\{r_i^*\}_{i=1}^{M}$. This is expressed as follows:

$$v_{min,i} = v_{max,i} = J_i\left(\{c_i\}_{i=1}^{N}\right), \quad i = 1, ..., N \quad (9)$$

where J_i is in general a function of all the input concentrations $\{c_i\}_{i=1}^{N}$ that returns the flux of input chemical compound i, and depends on the particular method employed by cells to uptake this chemical compound, e.g., facilitated diffusion or active transport through the cell membrane. The expression of $J_i(.)$ is in general known from biochemistry literature. As an example, the expressions for the input glucose J_{gl} and lactose J_{lac} fluxes considered in the numerical example of this paper are as follows [26]:

$$J_{gl}(c_{gl}) = \frac{J_{gl}^{max} c_{gl}}{\Phi_{gl} + c_{gl}}, \quad (10)$$

$$J_{lac}(c_{gl}, c_{lac}) = \frac{J_{lac}^{max} c_{lac}}{k_{lac} + c_{lac}} \left(1 - \phi_{gl} \frac{J_{gl}^{max} c_{gl}}{k_{gl} + c_{gl}}\right),$$

where the parameter values can be found in Table 3 of [26].

The estimates of the chemical reaction states $\{r_i^*\}_{i=1}^{M}$ are finally computed from the optimal flux vector $\mathbf{v}^*$, which is obtained by the FBA given the chemical composition of the surrounding environment $\{c_i\}_{i=1}^{N}$, as follows:

$$r_i^* = \begin{cases} 0, & \text{if } v_i^* = 0 \\ 1, & \text{otherwise} \end{cases}, \quad (11)$$

B. An Upper Bound to the Steady-state Mutual Information

Given the optimal estimates of the chemical reaction states $\{r_i^*\}_{i=1}^{M}$ obtained through the FBA from the knowledge of the cell's genome for all the values that our input set of chemical compound concentrations $\{c_i\}_{i=1}^{N}$ can assume, we can compute the following steady-state mutual information:

$$I(\{c_i\}_{i=1}^{N}; \{r_i^*\}_{i=1}^{M}) = H(\{c_i\}_{i=1}^{N}) - H(\{c_i\}_{i=1}^{N} \,|\, \{r_i^*\}_{i=1}^{M}), \quad (12)$$

where $H(\{c_i\}_{i=1}^{N} \,|\, \{r_i^*\}_{i=1}^{M})$ is computed through (7) by substituting the chemical reaction states $\{r_i\}_{i=1}^{M}$ resulting from

the real regulation of the enzyme expression with the FBA-estimated chemical reaction states $\{r_i^*\}_{i=1}^{M}$.

In this paper, we consider the mutual information (12) computed with the results of the FBA as an upper bound to the real steady-state mutual information (5) that we would obtain in reality as a result of the enzyme expression regulation. This is formalized as follows:

$$I(\{c_i\}_{i=1}^{N}; \{r_i^*\}_{i=1}^{M}) \geq I(\{c_i\}_{i=1}^{N}; \{r_i\}_{i=1}^{M}). \quad (13)$$

The expression in (13) can be proven through the Data Processing Inequality from information theory [6], which states that the aforementioned inequality holds true if the steady-state chemical reaction states $\{\hat{r}_i\}_{i=1}^{M}$ given a set $\{\hat{c}_i\}_{i=1}^{N}$ of values for the input concentrations can be probabilistically determined from the sole knowledge of the chemical reaction states $\{\hat{r}_i^*\}_{i=1}^{M}$, without the need of having knowledge of the input concentrations. This is expressed as follows [6]:

$$P\left(\{r_i\}_{i=1}^{M} \,\Big|\, \{c_i\}_{i=1}^{N}, \{r_i^*\}_{i=1}^{M}\right) = P\left(\{r_i\}_{i=1}^{M} \,\Big|\, \{r_i^*\}_{i=1}^{M}\right). \quad (14)$$

Equation (14) can be explained by considering that the chemical reaction states $\{\hat{r}_i^*\}_{i=1}^{M}$ are those that underlie the optimally regulated cell metabolism that maximizes the cell growth rate (or biomass production) given a set of values for the input concentrations $\{\hat{c}_i\}_{i=1}^{N}$. In reality, when subject to the same input concentrations $\{\hat{c}_i\}_{i=1}^{N}$, a cell reaches the steady-state chemical reaction states $\{\hat{r}_i\}_{i=1}^{M}$, which might be in general different from the aforementioned optimal states. If these states are indeed not optimal, the cell will not grow (produce biomass) and reproduce at the maximum rate possible given the input concentrations $\{\hat{c}_i\}_{i=1}^{N}$. When considering multiple cells in a population subject to the same input concentrations, if different cells show different steady-state chemical reaction states (because of cell-cell variability), those that have states closer to the optimal states will grow faster, and ultimately outnumber other cells. As a consequence, cells have evolved gene expression regulation mechanisms, such as those described in Sec. II-A, through which they adapt their steady-state chemical reaction states as close as possible to optimality given a set of input concentrations [7]. Given the optimal chemical reaction states $\{\hat{r}_i^*\}_{i=1}^{M}$ and the knowledge of the gene expression regulation mechanisms in place in a particular cell species, we are theoretically able to estimate the probability distribution of the steady-state chemical reaction states $\{\hat{r}_i\}_{i=1}^{M}$, which are those that best approximate the optimal states. As a consequence, under the aforementioned assumptions, the conditional probability of the steady-state chemical reaction states $\{\hat{r}_i\}_{i=1}^{M}$ given the input concentrations $\{\hat{c}_i\}_{i=1}^{N}$ and chemical reaction states $\{\hat{r}_i^*\}_{i=1}^{M}$ is equal to the same probability but only conditioned to the chemical reaction states $\{\hat{r}_i^*\}_{i=1}^{M}$, as expressed in (14).

IV. NUMERICAL RESULTS

In this section, we present a proof-of-concept numerical example of the abstraction and analysis method proposed in this paper. In particular, we focused on the *E. coli* bacterium
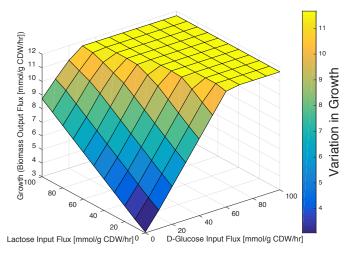
Fig. 2: Optimal E.Coli K12 MG1655 growth as a function of the input flux of D-Glucose and Lactose in the environment.

K-12 MG1655, which is considered one of the golden standard model systems in synthetic biology labs, and whose genome is completely known [12]. By stemming from this genome, we built the corresponding GEM, and subsequently performed the FBA by using the KBase (Department of Energy Systems Biology Knowledgebase) software application suite [14]. To obtain our numerical results, we developed a model of the external environment containing the known minimal set of chemical compounds necessary for this *E. coli* strain to grow, i.e., produce biomass, and at the same time allowing the variation of the concentrations of key compounds that result in changes in the optimal FBA-computed states of metabolic reactions. For this, we based our environment on the K-12 MG1655 minimal media [9], enriched with metal tracers common to other two standard media, namely, the Lysogeny Broth (LB) and the Carbon-D-Glucose media [5]. All compound fluxes contained in $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$ were set -100 and 100 [mmol per gram cell dry weight per hour] ([mmol/g CDW/hr]), respectively. On top of the defined media environment, we introduced two other input compounds, namely, D-Glucose and Lactose, for which we simulated a variation in their concentration, and consequent corresponding values in $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$ according to (9) and (10). In particular, we obtained our numerical results by performing FBA on every combination of input fluxes of D-Glucose and Lactose ranging from 0 to 100 [mmol/g CDW/hr] with increments of 10 [mmol/g CDW/hr], for a total of 121 different combinations of input fluxes.

In Fig. 2 we show the results of the FBA in terms of growth rate, or equivalently, output flux of produced biomass, as defined in Sec. III-A, computed for different combinations of values of the input fluxes of D-Glucose and Lactose. In these results, the variation in optimal growth rate, which is dependent on the optimal metabolic reaction states, as discussed below, varies from a minimum value of 3.061 [mmol/g CDW/hr] when the fluxes of D-Glucose and Lactose are absent from the environment, to a maximum value of 11.67 [mmol/g CDW/hr]. These curves show also a saturation in the optimal growth rate for D-Glucose fluxes on the higher end of the
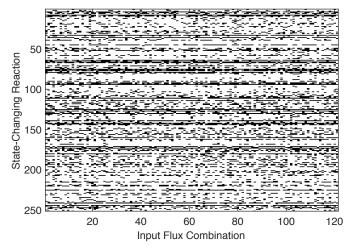

Fig. 3: FBA-estimated binary chemical reaction states $\{r_i^*\}_{i=1}^M$ for each combination of D-Glucose and Lactose input fluxes, where white = ON state; black = OFF state.

range, and the minimal value of D-Glucose flux to obtain this saturation varies as function of the lactose flux, from a minimum of 30 [mmol/g CDW/hr] to a maximum of 70 [mmol/g CDW/hr].

In Fig. 3, for each of the 121 tested combinations of the input fluxes of D-Glucose and Lactose, one for each column of the matrix, we show the binary values of FBA-estimated chemical reaction states $\{r_i^*\}_{i=1}^M$ as defined in Sec. III-A, one for each column, where the number of metabolic reactions M that show a state change within the considered combination of input fluxes of D-Glucose and Lactose is equal to 251.

The computation of the upper bound of the steady-state mutual information is finally realized by applying the expressions in (5), (6), and (7), taking into account that the possible combinations of input fluxes of D-Glucose and Lactose are drawn from a discrete set. For these preliminary results, we make the assumption that these combinations are equiprobable. As a consequence, the corresponding combinations $\{c_i\}_{i=1}^N = \{c_{gl}, c_{lac}\}$ computed through (10) can be as well considered equiprobable with probability density $P(\{c_{gl}, c_{lac}\}) = 1/(\#\text{of input combinations}) = 1/121$. The resulting input entropy $H(\{c_i\}_{i=1}^N)$, where $N = 2$, is then computed through (6) by substituting the integral with a summation over the number of input combinations, which results into $\log_2(121) = 6.92$ bits. To compute the conditional entropy of the input given the output $H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M)$, where $N = 2$ and $M = 251$, we translated (7) into the following formula:

$$H\left(\{c_{gl}, c_{lac}\} | \{r_i^*\}_{i=1}^M\right) = -\sum_{y=1}^{Y} P_Y \sum_{x=1}^{X_y} P(x|y) \log_2 P(x|y),$$
(15)

where Y and P_Y correspond to the number of times and the probability, respectively, that a reaction state combination is found more than once in the data shown in Fig 3, X_y is the number of times the reaction state combination y is found in

the data, and $P(x|y)$ is the probability of having a combination of input concentrations $x = \{c_{gl,x}, c_{lac,x}\}$ given the reaction state y, which we consider as a uniform distribution in the number of different combinations of input concentrations that have been found resulting into the same the reaction state y. In our data we found a total of 4 different reaction combinations that are repeated twice, which results in a conditional entropy of the input given the output $H\left(\{c_{gl}, c_{lac}\}|\{r_i^*\}_{i=1}^{M}\right) = 0.53$ bits. Finally, the following value is found for the upper bound of the steady-state mutual information:

$$I\left(\{c_{gl}, c_{lac}\}\right); \{r_i^*\}_{i=1}^{115}) = 6.39\,\text{bits} \qquad (16)$$

V. Conclusion

In this paper, we have introduced a method to obtain an upper bound to the steady-state mutual information of a communication system based on cell metabolism and its regulation. This upper bound stands a theoretical limit of the ability of a particular cell to internally represent the information contained in the chemical composition of the external environment. In particular, we presented a method to evaluate this upper bound by modeling the regulation of chemical reactions activity in cell metabolism as a binary encoder of the external concentration of chemical compounds, and subsequently applying the Flux Balance Analysis method to simulate the outcome of this model in optimal conditions.

The abstraction and analysis method developed in this paper will potentially help the design of techniques to control functionalities in cells engineered through genetic circuits [20]. Future work will be focused on a thorough modeling and evaluation of this molecular communication system, including models of the noise source and the dynamic behavior of metabolic regulation, and the investigation of its information theoretical capacity.

References

[1] I. F. Akyildiz, J. M. Jornet, and M. Pierobon, "Nanonetworks: A new frontier in communications," *Communications of the ACMs*, vol. 54, no. 11, pp. 84–89, November 2011.

[2] I. F. Akyildiz, M. Pierobon, S. Balasubramaniam, and Y. Koucheryavy, "The internet of bio-nano things," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 32–40, March 2015.

[3] U. Alon, *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2006.

[4] G. Aminian, H. Arjmandi, A. Gohari, M. Kenari, and U. Mitra, "Capacity of lti-poisson channel for diffusion based molecular communication," in *In Proc. of 2015 IEEE International Conference on Communications (ICC)*, June 2015.

[5] G. Bertani, "Lysogeny at mid-twentieth century: P1, P2, and other experimental systems," *Journal of Bacteriology*, vol. 186, no. 3, pp. 595–6008, 2004.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Wiley, 2006.

[7] E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, Elsevier, 2006.

[8] J. Dobson, "Remote control of cellular behaviour with magnetic nanoparticles," *Nature Nanotechnology*, vol. 3, pp. 139–143, 2008.

[9] EcoCyc, "Escherichia coli K-12 substr. MG1655 Growth Medium: M9 medium with 2% glycerol," http://biocyc.org/ECOLI/NEW-IMAGE?type=Growth-Media&object=MIX0-59, [Online; accessed 13-March-2016].

[10] A. Einolghozati, M. Sardari, and F. Fekri, "Capacity of diffusion-based molecular communication with ligand receptors," in *IEEE Information Theory Workshop (ITW)*, October 2011.

[11] L. Felicetti, M. Femminella, G. Reali, T. Nakano, and A. Vasilakos, "Tcp-like molecular communications," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 12, pp. 2354–2367, 2014.

[12] K. Hayashi, N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, B. L. Wanner, H. Mor, and T. Horiuchi, "Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110," *Mol Syst Biol.*, vol. 2, February 2006.

[13] L. J. Kahl and D. Endy, "A survey of enabling technologies in synthetic biology," *Journal of Biological Engineering*, vol. 7, no. 1, p. 13, May 2013.

[14] KBase, "Department of Energy Systems Biology Knowledgebase (KBase)," http://kbase.us, [Online; accessed 13-March-2016].

[15] W. A. Lim, "The promise of optogenetics in cell biology: interrogating molecular circuits in space and time," *Nature Reviews*, vol. 11, pp. 393–403, 2010.

[16] D. McCloskey, B. Ø. Palsson, and A. Feist, "Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli." *Mol Syst Biol.*, vol. 9, p. 661, 2013.

[17] D. B. Menendez, V. R. Senthivel, and M. Isalan, "Sender-receiver systems and applying information theory for quantitative synthetic biology," *Curr Opin Biotechnol*, vol. 31C, pp. 101–107, March 2015.

[18] C. M. Metallo and M. G. V. Heiden, "Understanding metabolic regulation and its influence on cell physiology," *Molecular Cell*, vol. 49, pp. 388–398, February 2013.

[19] R. Mosayebi, H. Arjmandi, A. Gohari, M. Nasiri-Kenari, and U. Mitra, "Receivers for diffusion-based molecular communication: Exploiting memory and sampling rate," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 12, pp. 2368–2380, December 2014.

[20] C. J. Myers, *Engineering genetic circuits*. Chapman & Hall/CRC, Mathematical and Computational Biology Series, 2009.

[21] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*. W. H. Freeman, 2005, ch. 12.2, pp. 425–429.

[22] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nat Biotechnol.*, vol. 28, no. 3, pp. 245–248, March 2010.

[23] S. Payne and L. You, "Engineered cell-cell communication and its applications," *Adv Biochem Eng Biotechnol*, vol. 146, pp. 97–121, 2014.

[24] M. Pierobon and I. F. Akyildiz, "Capacity of a diffusion-based molecular communication system with channel memory and molecular noise," *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 942–954, February 2013.

[25] A. Rhee, R. Cheong, and A. Levchenko, "The application of information theory to biochemical signaling systems," *Phys Biol.*, vol. 9, no. 4, p. 045011, August 2012.

[26] M. Santillán, "Bistable behavior in a model of the lac operon in escherichia coli with variable growth rate," *Biophysical Journal*, vol. 94, pp. 2065–2081, 2008.

[27] K. Srinivas, A. Eckford, and R. Adve, "Molecular communication in fluid media: The additive inverse gaussian noise channel," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4678–4692, 2012.

[28] G. Tkačik, C. G. Callan Jr., and W. Bialek, "Information capacity of genetic regulatory elements," *PHYSICAL REVIEW E*, vol. 78, p. 011910(17), 2008.

[29] J. E. Toettcher, C. A. Voigt, O. D. Weiner, and W. A. Lim, "The promise of optogenetics in cell biology: interrogating molecular circuits in space and time," *Nature Methods*, vol. 8, pp. 35–38, 2011.