NON-EQUILIBRIUM MODEL APPROACHES FOR GENETIC SYSTEMS

Report presented by
Lucas Goiriz Beltrán

as a Master's thesis for the program
Master in Mathematical Research

carried out under the supervision of
Dr. J. Alberto Conejero
Dr. Guillermo Rodrigo Tárrega
Dr. David Fuente Herraiz

Aknowledgements

Firstly, I would like to thank Guillermo Rodrigo for everything he has done for me since we met, way back in 2018: from supervising me on my TFG to accepting me into his research group as a predoctoral student. I will always look up to you. Thank you.

Second, I would like to thank Alberto Conejero for introducing me to the Investmat Masters program and giving me extra support when things weren't working out as I planned. Without Alberto's first approach in my 4^{th} year of Biotechnology, I wouldn't have dived into the passionate world of mathematics. Thank you.

Third, I would like to thank David Fuente for bearing with me during the most uncertain times during the pandemic and making the scripts run properly.

Next, I would like to thank everyone that has worked with me during this last year, especially my friends Nicolas Firbas, Aleixandre Muñoz, and Héctor Ariza. It has been a pleasure to have met you, and I hope that someday our professional paths cross once again.

Finally, I would thank the two most important people I met during this academic year: my dear friends Antoni López Martínez and Christian Cobollo. These two brilliant mathematicians approached me back in September 2019 and told me the following: *"We've noticed that you've never learnt the background that this level of mathematics demands. But we think you've got the spirit. We are going to make sure to fill most of your knowledge gaps"*. And so they did: real and complex analysis, topology, and even algebra (a field both of them did not enjoy much) while we had lunch at the faculty of Pharmacy's cafeteria, some pizza at Plaza Xuquer or even while having a couple of drinks after dinner. What these gentlement didn't know is that they ignited a hunger for knowledge that was buried deep inside me. I will make sure to keep on learning mathematics so we can chat about all the things you learned *"en la ehcuela"*. Thank you.

Index

1 Introduction **1**

1.1 Systems biology 1

1.2 Genetic network and regulation 2

1.3 Mathematical modelling in Biology 4

1.4 Objectives 7

2 RNA-RNA interactions in gene expression regulation **8**

2.1 Thermodynamic Models 8

2.2 Energy Landscape 10

2.3 Kinetic Modeling 14

2.4 Experimental Validation 16

3 Gene expression regulation role in cellular Metabolism **18**

3.1 Metabolic models and Flux Balance Analysis 18

3.2 Multi Objective optimization 21

3.2.1 sp-MODE algorithm 24

3.3 Resource allocation model 26

3.4 Simulations 29

4 Conclusions **30**

References **35**

1 Introduction

1.1 Systems biology

The developments in molecular biology that started in the 1960s with the central dogma of molecular biology[1] altogether with the techniques developed during the 1970s that allowed DNA manipulation, transfer and cloning, were the first footsteps of the current genomic era of biology. These developments in molecular biology and genomic understanding were possible due to the parallel increase of computing power and instrumental technology, such as sequencing technologies, during the following decades until reaching its most important landmark in the 2000s with the completion of the international scientific effort of constructing the first human genome. Currently, there are sequence and map data from whole genomes of over 1000 organisms in all three main domains of life (Bacteria, archaea, and eukaryota) as well as "non-living" entities such as viruses, phages, viroids, plasmids, and organelles [National Center for Biotechnology Information, 2020].

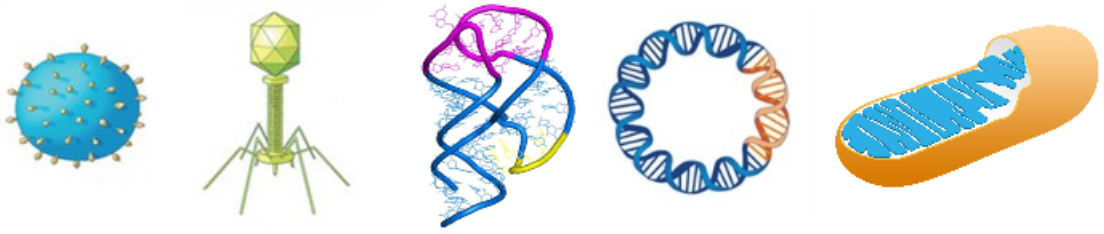

Figure 1: Examples of non-living biological entities. From left to right: virus, bacteriophage, viroid, plasmid and an organelle (the mytochondria).

This means that there is free world-wide access to lists of basic components of living organisms encoded at the genetic level. In other words, massive amounts of biological data ready to be employed for any purpose. This massive data availability and its continuous growth thanks to high throughput technologies raised the issue of addressing an ever-growing gap between raw data and interpretation. It is here where the field of Systems Biology comes into the spotlight.

Systems Biology is a field whose objective is to mathematically and systematically analyze and integrate huge biological datasets to infer and predict how cells respond to environmental stimuli and how these responses originate from its molecular components' natural interactions. This new approach towards biology has been quickly incorporated in a variety of fields with satisfactory results. Some examples are drug discovery [Germain, 2017, Zou et al., 2013, Oh et al., 2018], cancer research and [Archer et al., 2016, Zhou, 2016, Cherubini et al., 2017], neurological diseases research [Villoslada et al., 2009, Faundez et al., 2019] among others.

The systematic analysis of huge data sets by Systems Biology is possible thanks to the growth of computational power and the design in silico mechanistic models through mathematical software, leading to the "dissection" of a cell's natural behaviour. The use of

[1]The central dogma of molecular biology states that DNA encodes RNA which, in turn, encodes protein macromolecules.

mathematics in the description of cellular processes by this field contributes to developing other novel applied fields such as Synthetic Biology, where models originating from Systems Biology models are applied in experimental designs.

A Systems Biology way of understanding a system involves four key stages [Kitano, 2001, Kitano, 2002]:

1. The structures and interactions which compose the system have to be identified.
2. The behavior of the system under various conditions has to be analyzed.
3. Once the system is properly characterized, mechanisms that systematically control the cell state can be modulated to avoid pathological behavior or to enhance a desirable cellular function.
4. Finally, protocols to construct the biological systems with the desired properties can be designed.

During these steps, the system is frequently perturbed (either by biological, chemical, or genetic modifications), and its responses are closely monitored, yielding the information of interest, which is integrated into the form of mathematical models that describe the structure of the biological system and its responses towards perturbations [Ideker et al., 2001]. Therefore, mathematical models are the *rosetta stone* in Systems Biology, bringing together organization, understanding, and exploitation of information from a biological system.

1.2 Genetic network and regulation

One of the most interesting biological systems that is suitable for a Systems Biology approach is a cell's genetic network. This is due to a recent discovery that revealed that a cell's phenotypes are the product of gene-gene and protein-protein interactions plus biomolecular networks, rather than the accumulation of effects of a single gene or molecule at a time. These interactions and their effects occur during the complex process of gene expression regulation. We can approach them at several levels: at the epigenome, at the transcriptional level (at the moment of gene transcription into messenger RNA [mRNA] by RNA polymerases), the post-transcriptional level, and the post-translational level.

In order to understand the importance of regulation, there is a need to review how the internal mechanisms of gene expression function [2]. It is of general knowledge that the genetic information in eukaryotic cells is stored in the nucleous in the form of a macromolecule composed by a deoxyrribose phosphate backbone plus the nitrougenous bases Adenine [A] with Thymine [T], Guanine [G] with Cytosine [C]. This macromolecule, known as DNA, interacts using hydrogen bonds between the complementary nitrogenous bases (A with T, G with C) with another DNA to form the known double helix structure. From this point on, the consecutive combination of nitrogenous bases will be dubbed as *sequence*. Particular sequences in the DNA exist for different purposes: some are coding sequences (known as genes), while some other are non-coding sequences (they have no proteic product).

[2]For a more detailed review seek the most recent edition of [Alberts, 2017].

Transcription occurs when a particular enzyme called RNA polymerase (with or without the aid of DNA binding proteins known as transcription factors) attaches to a non-coding stretch of DNA called promoter, which precedes a coding sequence (a gene or group of genes). The polymerase then synthethyses mRNA through the nucleotides (a ribose phosphate attached to a nitrogenous base) found in the cell's cytosol (aqueous interior of the cell). Note that the difference between DNA and RNA mainly lies in the macromolecule's backbone (deoxyribose vs. ribose) and the change of one of its bases (Thymine by Uracil [U]). The mRNA synthesis finishes at the stop site and once released, it may suffer post-transcriptional modifications (common in eukaryotes and uncommon in prokaryotes). It may be degraded, or it may bind to other macromolecules that promote or prevent its movement or translation.

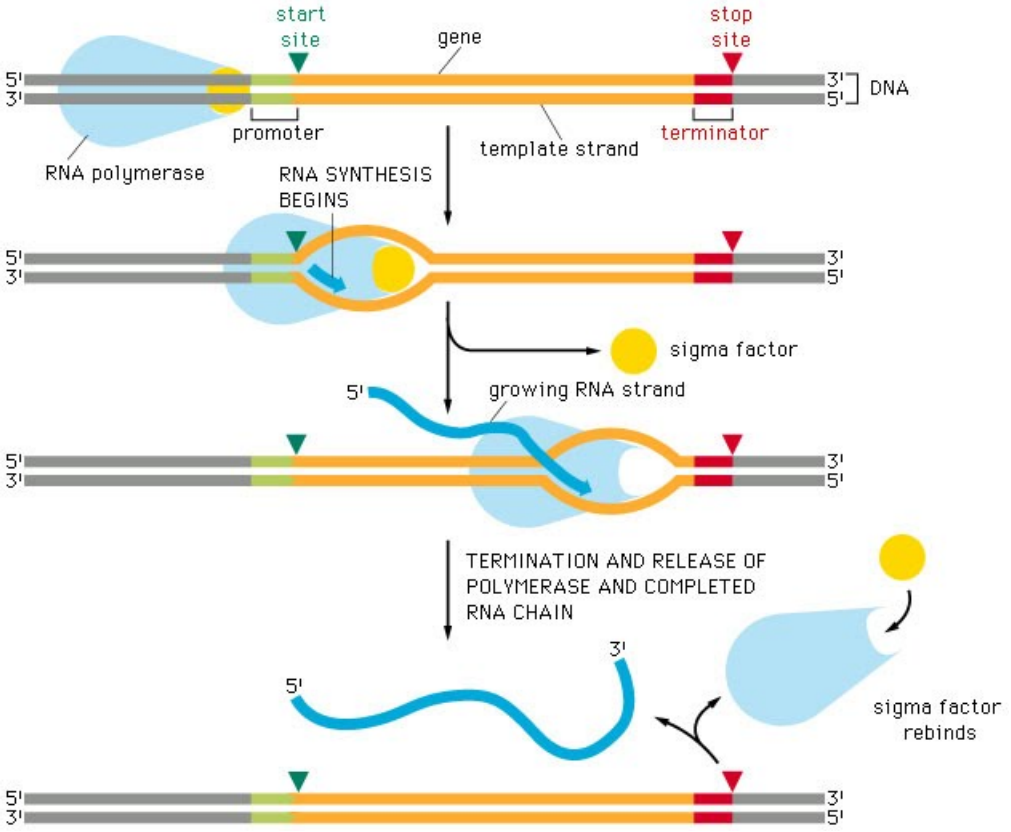

Figure 2: Overview of mRNA synthesis. Note that sigma factor is a transcription factor.

If the mRNA integrity is not compromised, it makes its way to the cytoplasm and recruits the ribosome to begin translation. The ribosome complex moves through the sequence until fiding a particular combination of bases (AUG) known as the start codon, where it begins building the polypeptide chain (protein) through the aminoacids provided by the specific complementary transporter RNA (tRNA) to the triplet (3 bases) where the ribosome stands. Step by step, the protein is synthesized, and it automatically starts adapting a particular spatial conformation (this is because the polypeptide seeks a particular structure that exposes the hydrophilic aminoacids while hiding the hydrophobic ones from the cells liquid media, in order to increase entropy). The translation process ends when the ribosome meets a terminator codon (UAG, UGA, or UAA). The complete process from the moment in which an mRNA is transcribed from a gene or set of genes until the protein product is synthesized

has many key points, where additional machinery of the cell may halt, enhance or repress the process by different mechanisms (as mentioned previously in this work). Formally, the cell's capacity of fine-tuning the moment, the amount, and the speed at which a gene is expressed is known as gene expression regulation.

1.3 Mathematical modelling in Biology

Complex processes, such as the one explained in the previous section, are the target of modelling from the moment they are discovered. The main law that has been employed to build dynamic systems that describe each component of a biological system's behavior has its origins in the fields of chemistry and thermodynamics. It is known as the law of mass action.

Definition 1 (Law of mass action)**.** *Consider the chemical reaction at constant temperature and volume given by the expression*

$$\mathrm{aA} + \mathrm{bB} \underset{\tau}{\overset{\sigma}{\rightleftharpoons}} \mathrm{cC} + \mathrm{dD}$$

where A and B are two reactants, C and D are two products, σ the rate constant of the forward reaction, τ is the rate constant of the reverse reaction and the undercase letters represent the coefficients (number of molecules needed for one instance of the reaction) of its respective uppercase letter chemical species; then the forward reaction rate r is

$$r_{\text{forward}} = \sigma \left[A\right]^a \left[B\right]^b \tag{1}$$

where $[\cdot]$ denotes the concentration of a chemical species.

Similarly, the reverse rate corresponds to $r_{reverse} = \sigma \left[C\right]^b \left[D\right]^d$. Thus, if the chemical species of interest is A, let $A(t)$ be the concentration of said species over time, then the law of mass action allows to define the rate at which the concentration of A changes over time:

$$\frac{dA(t)}{dt} = -\sigma A^a(t) B^b(t) + \tau C^c(t) D^d(t) \tag{2}$$

The other species rate can be expressed similarly. However, this cumbersome expression is simplified in biological modelling as interactions are most of the time bimolecular (i. e. they only need of two molecules to interact in order for the reaction to occur). Mathematically this means that for a single direction in a reaction, coefficients usually sum up to 2 (in this example, $a + b = 2$ and $c + d = 2$). Furthermore, as biological interactions are usually integrated into a pathway (also known as reaction cascade), the species participate in various reactions simultaneously (both as reactants and products), so the expression can be elongated in a number of terms.

There has been invested much interest in these systems from a mathematical point of view since their interesting properties, described by classical ordinary differential equations tools, can be directly observed in experiments. One remarkable example is [Gardner et al., 2000]'s genetic toggle switch in *Escherichia coli*, where a synthetic system in bacteria which

starts in an unstable steady state and by blocking gene expression by a repressor substance, the system moves towards a stable equilibrium, which can only be altered by adding another repressor. The highlight is that the bacteria can produce the repressor substances by itself, each repressor control its counterpart's expression and that the change produced pervades during generations. However, this complex system is simple in mathematical terms and beautifully obeys the following system of equations:

$$\frac{du}{dt} = \frac{\alpha_1}{1+v^\beta} - u \tag{3a}$$

$$\frac{dv}{dt} = \frac{\alpha_2}{1+u^\gamma} - v \tag{3b}$$

where u and v are the concentration of the repressors, α_i are the effective synthesis rate of the repressors, β, and γ are the strength of repressors against its counterpart's expression.

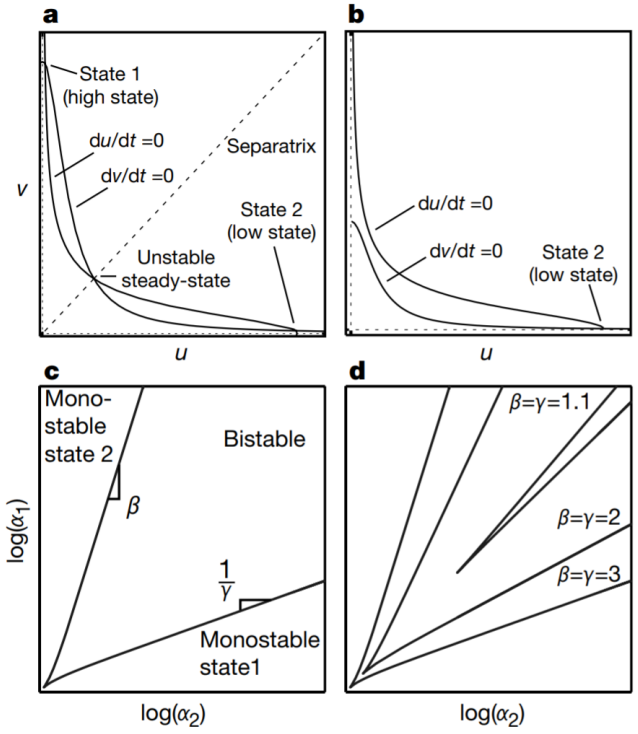

Figure 3: Geometric structure of the toggle equations. **a)** Bistable (2 equilibirum points) scenario.promoter strengths. **b)** A monostable toggle network (1 equilibrium point). **c)** Parameter characterization that show bifurcation between bistability and monostability. **d)** Reducing the strength of repression reduces the size of the bistable region. Courtesy of [Gardner et al., 2000].

.

This analysis brings huge biological information, as it allows to select the parts to construct the switch in order for it to be stable and behave as planned. This means that, as shown in **Figure 3**, parameters β and γ determine if the system is either monostable or bistable, thus for a successful construction of the toggle switch, adequate repressors have to

be chosen. Also, the smaller the bistable region (due to promoter strength), the easier the switch is triggered by random fluctuations.

Another scenario where the same principles were applied and results were more than satisfactory was with [Elowitz and Leibler, 2000]'s repressilator. A procedure similar to that of [Gardner et al., 2000] was performed, but this time with 3 repressors acting in a chain, meaning that the final result was some kind of clock, where repressors concentration increased and progressively decreased in a cyclical manner.

Again, the design was possible thanks to mathematical modelling. The equations describing the system are the following:

$$\frac{dm_i}{dt} = -m_i + \frac{\alpha}{1 + p_j^n} + \alpha_0 \tag{4a}$$

$$\frac{dp_i}{dt} = -\beta\,(p_i - m_i) \tag{4b}$$

where $\begin{pmatrix} i = lacI,\ tetR,\ cI \\ j = cI,\ lacI,\ tetR \end{pmatrix}$ are the repressors (drawn by index pairs), α_0 is the number of protein copies per cell produced from a given promoter type during continuous growth in the presence of saturating amounts of repressor, $\alpha + \alpha_0$ in its absence, β denotes the ratio of the protein decay rate to the mRNA decay rate and n is a Hill coefficient[3]. Again, the model proved to be successful for an optimal design of the system.

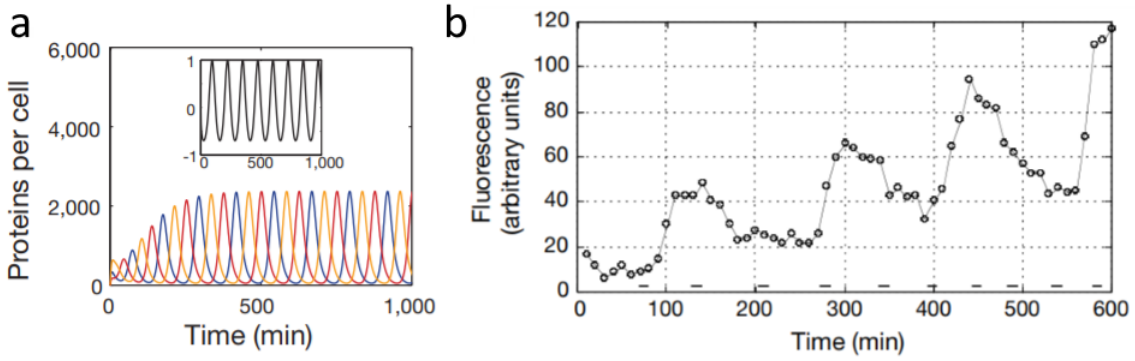

Figure 4: **a)** Numerical integration results of the total systems of equations resulting from **Equations 4a** and **4b**. Note the oscillatory nature of all three repressors. **b)** Experimental measure of the fluorescence of the fusion protein resulting from a repressor and GFP[4], indicating the oscillatory nature of the repressor. Figure modified from [Elowitz and Leibler, 2000].

Although these systems result simple from a mathematical point of view, there are multiple efforts from mathematicians in properly characterizing them to obtain generalized models and discover interesting properties. For the case of the repressilator, lot of work on that matter has been already done, see for instance [Müller et al., 2006, Buzzi and Llibre, 2014].

[3]The Hill coefficient measures how sensitive the effect of the repressor is.
[4]Green Fluorescent Protein. Routinely used in molecular biology to track other proteins.

Similarly, the toggle switch has also permitted to develop interesting mathematical results from its initial equations [Perez-Carrasco et al., 2018]. We also refer to [Fuente et al., 2020], where a genetic four-state machine is analyzed using deterministic and stochastic modeling and a spatial approach via a cellular automaton.

1.4 Objectives

Coming back to the matter of gene expression regulation, all levels of regulation mechanisms for a target gene can be studied as a network of interactions of the said gene with DNA binding proteins, enzymes, and non-coding nucleic acid sequences. Since they suffer from the effect of gene regulation mechanisms, increasing the network's size drastically. As shown through various examples in the previous section, this allows the formulation of dynamic systems with more or less complexity (depending on the scope taken on the system of interest). This means that all elements participating in the network are not in thermodynamic equilibrium. This implies that their abundancy inside the cell changes or can be triggered to change over time, and are subject to flux of matter and energy by chemical reactions (see **Definition 1**), which are in turn influenced by the metabolic needs of the cell, in a continuous and discontinuous manner.

The aim of this work is to study how a biological system at the genetic level behaves under the conditions of thermodynamic non-equilibrium, as to date, most approaches attempting to answer this question are build upon the assumption of thermodynamic equilibrium. This task will be performed at two different scopes:

- At local scope, where post-transcriptional regulation mediated by RNA-RNA interactions will be studied (**Section 2**).
- At global scope, where a cell metabolism network is restricted by the availability of cellular resources will be studied (**Section 3**).

This approach is inspired by the two strategies followed in the field of biology: "bottom up", i. e. studying every biological component individually and then describing the behaviour of the biological system as a sum of its parts, and "top down", namely studying a biological system as a whole and later describing each component.

2 RNA-RNA interactions in gene expression regulation

While most gene expression regulation occurs at the transcriptional level, other regulation levels, such as the post-transcriptional level, exist to complement the main transcriptional control to fine-tune expression or allow faster responses [Dahan et al., 2011]. In particular, the cases of regulation mechanisms based on RNA-RNA interactions are pervasive as they can be found in several differentiated biological contexts [Waters and Storz, 2009, Guil and Esteller, 2015, Paillart et al., 2004]. One such type of regulation consists of a small RNA (sRNA), naked or together with a protein, to target a given mRNA to regulate its translation or stability [Waters and Storz, 2009]. Even though the quantitative description of all these regulations that occur in vivo from simple, accurate, and formal mechanistic models is still a challenge.

2.1 Thermodynamic Models

Models based on statistical mechanics, also known as thermodynamic models, have been already exploited to quantitatively understand transcription regulations mediated by protein-DNA interactions [Bintu et al., 2005, Brewster et al., 2014, Phillips, 2015]. We assume that a system is in thermodynamic equilibrium if no further macroscopic change takes place. The most relevant characterization of thermodynamic equilibrium for the subject matter is defined through Gibbs free energy:

Definition 2 (Gibbs free Energy)**.** *Gibbs free Energy is a thermodynamic potential that characterizes the maximum amount of non-expansion work that can be extracted from a thermodynamically closed system*[5] [Adkins, 1997]*. It is formulated as*

$$G = U - TS + PV$$

where U is the system's internal energy, T is the absolute thermodynamic temperature, S is entropy, P is pressure, and V is volume. At thermodynamic equilibrium, given that T and P are fixed, G is minimum.

Since macromolecules in a cell are not in saturation, in order for them to interact, there is a dependence of random diffusion and colision. There is a need to predict how these macromolecules will distribute themselves inside the cell. Given that the Second Law of thermodynamics states that in an isolated system, the entropy never decreases, it is simple to point out that such distribution will tend to maximize the entropy of the system, this distribution is the Boltzmann Distribution.

[5] A system capable of exchanging heat and work with its surroundings, but not matter.

Definition 3 (Boltzmann Distribution)**.** *The Boltzmann distribution is a probability distribution that indicates the chances that a system will be at a certain state given said state's energy and the system's temperature. Its formulation is*

$$p_n = \frac{e^{\frac{-\varepsilon_n}{k_B T}}}{\sum_{j=1}^{M} e^{\frac{-\varepsilon_j}{k_B T}}}$$

where p_n is the probability of state n, ε_n is the energy of state n, k_B is the Boltzmann constant (which relates the average kinetic energy of gas particles with the temperature of the gas), and T is the thermodynamic temperature.

For the sake of simplicity, the term $\frac{1}{k_B T}$ will be referred to as β from now on. Note that $\sum_{j=1}^{M} e^{\frac{-\varepsilon_j}{k_B T}}$ is the canonical partition function, which results from the constraint that all the probabilities of all possible states must sum up to 1. Given this probability distribution, the Boltzmann factor can be defined:

Definition 4 (Boltzmann factor)**.** *The Boltzmann factor is the ratio of probabilities of two states of a system, which indicates the spontaneity of transitioning from one state to another. Its formulation is*

$$\frac{p_i}{p_j} = e^{\beta(\varepsilon_j - \varepsilon_i)}$$

where p_i, p_j are the probabilities of states i and j, k_B is the Boltzmann constant, and T is the thermodynamic temperature.

The Boltzmann factor can be easily derived by directly computing the ratio of probabilities of two states by means of **Definition 3** and simplifying the terms. Under the conditions of thermodynamic equilibrium and assuming that macromolecules are not in saturation, consider a simple system of DNA and DNA-binding proteins, which has two possible states: free DNA with free DNA-binding proteins and DNA bound with DNA-binding proteins. Therfore, for this particular case the probability of each state can be computed as follows:

$$p_{\text{free}} = \frac{e^{-\beta G_{\text{free}}}}{e^{-\beta G_{\text{free}}} + e^{-\beta G_{\text{bound}}}} \tag{5a}$$

$$p_{\text{bound}} = \frac{e^{-\beta G_{bound}}}{e^{-\beta G_{\text{free}}} + e^{-\beta G_{\text{bound}}}} \tag{5b}$$

Note that the energy employed is the Gibbs free Energy (see **Definition 2**). Consequently, computing the Boltzmann factor for this state change is simple:

$$\frac{p_{\text{free}}}{p_{\text{bound}}} = e^{\beta(G_{\text{bound}} - G_{\text{free}})} \tag{6}$$

Furthermore, $G_{\text{bound}} - G_{\text{free}}$ represents the change in Gibbs free Energy between unbound macromolecules and bound macromolecules, namely, it is the net free energy release associated with the intermolecular interaction, which is noted as ΔG_{inter}.

As DNA-protein interaction mediated post-transcriptional level expression control depends on the intermolecular interaction, the dynamic regulatory range, also known as fold change in expression r of a particular gene is proportional to **Equation 6**, and thus $r \propto e^{\beta \Delta G_{\text{inter}}}$.

However, in the case of post-transcriptional regulations based on RNA-RNA interactions this assumption of thermodynamic equilibrium appears to be inappropriate, as previous results have pointed out that only ΔG_{inter} is not sufficient to predict r [Mutalik et al., 2012, Rodrigo et al., 2013, Peterman et al., 2014]. Instead, and mostly from heuristic approaches, other Boltzmann factors relative to additional free energies characterizing the interaction, such as the free energy of seed[6] pairing (denoted by ΔG_{seed}), need to be considered in a combinatorial way to achieve relatively good estimates [Mutalik et al., 2012, Rodrigo et al., 2013]. This certainly challenges the conventional wisdom and forces to find a novel thermodynamic picture from which to derive a formal relationship. Interestingly, two molecular features mainly mark the difference between transcriptional and post-transcriptional regulations. First, RNAs are typically short-lived macromolecules, whilst DNA and proteins are not [Schwanhäusser et al., 2011]. Second, RNA-RNA interactions are weaker (with less binding affinity) than protein-DNA interactions [Rodrigo et al., 2017].

A non-equilibrium thermodynamic scheme to explain RNA-RNA interactions will be adopted in this work. As a working example, we considered the IS10 transposition system[7] that occurs in bacteria. There, an sRNA (dubbed RNA-OUT) interacts with the 5' untranslated region (dubbed RNA-IN) of a given mRNA, encoding a transposase, to repress the protein synthesis [Kittle et al., 1989]. RNA-OUT is a structured sRNA with a stem-loop conformation (with an external loop of 6 nucleotides and a long stem with 3 small internal loops, as already proposed [Kittle et al., 1989] and here assumed), whilst RNA-IN is a non-structured element in which the Shine-Dalgarno box[8] and the start codon are exposed to the solvent. Next, the energy landscape associated with the interaction at the base pair resolution will be described. In turn, a simple but general statistical model that emerges from a regulatory scenario in a steady state but out of the equilibrium will be derived [Gnesotto et al., 2018].

2.2 Energy Landscape

Definition 5 (Energy Landscape)**.** *An Energy Landscape is a continuous function* $f : X \to \mathbb{R}$ *associating each physical state with an energy, where* X *is a topological space.*

A less formal definition, which is used routinely by biologists, would be a mapping of possible states of a system. In this work, it is employed to describe the number of interacting base pairs of the system and their corresponding Gibbs free energy contribution to the system. Thus, it can be geometrically interpreted as the graph of the energy function across the target system's configuration space. The energy landscapes for individual RNA folding,

[6] Short stretches of nucleotides involved in the initial binding of RNAs [Gorski et al., 2017].

[7] Sequence of DNA that is capable of moving from one position in the DNA of the host organism to another.

[8] A ribosomal binding site in bacterial and archaeal mRNA key for protein synthesis initiation [Shine and Dalgarno, 1975].

namely computation of possible spacial conformations of single RNA molecules, can be represented under $f : \mathbb{R}^9 \to \mathbb{R}$. The nine dimensions represent that each RNA building block (nucleic acid) has 7 degrees of freedom corresponding to the sugar backbone torsions, and there are 2 more degrees of freedom for the nitrogenous base torsions [Chen and Dill, 2000, Schlick, 2018]. However, in this work's scope, the landscape describes RNA-RNA interaction, and thus it only depends on the number of interacting base pairs between RNA strands, which is a discrete value. Thus, there is a need to properly define the nature of the energy landscape employed by RNA-RNA interactions.

Definition 6 (RNA-RNA Interaction Energy Landscape)**.** *Consider the discrete space* $X = \{0, \ldots, n\}$*, where* n *is the maximum number of interacting base pairs between two RNA strands, predefined by their nucleic acid sequence. Then, the energy landscape is the function* $f : X \to \mathbb{R}$*, associating each intermolecular RNA-RNA base pairing with an energy contribution to the system.*

Clearly, such a function is continuous since its domain is in a discrete space. To resolve the energy landscape associated with the sRNA-mRNA interaction with precision, the different gains and losses of free energy that occur as the reaction coordinate progresses need to be computed, which is shown in **Figure 5**.

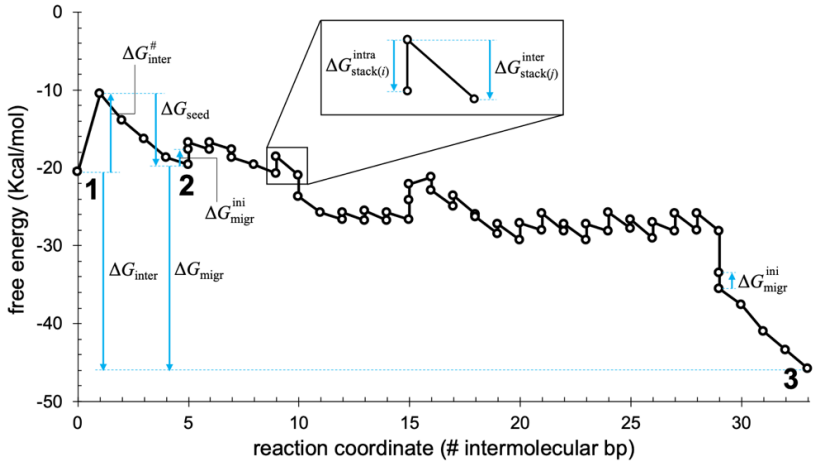

Figure 5: Energy landscape of the sRNA-mRNA interaction in the IS10 system at the base pair (bp) resolution. The intramolecular folding state is labelled as 1, the just-the-seed-paired intermolecular folding state as 2, and the final intermolecular folding state as 3. Blue vertical arrows mark the different free energies that characterize the interaction.

In the particular case of the IS10 system, this ranges from 0 (no interaction; state 1) to 33 base pairs (final complex; state 3). Notably, the different secondary structures that are progressively formed in that interaction range (base pair by base pair) can be evaluated by hand according to a simplified physicochemical model with the stacking and looping free energies previously determined at 37°C [Xia et al., 1998, Mathews et al., 1999]. Initially, the species are just folded intramolecularly (see **Figure 6a**), as this process occurs much faster than the eventual interaction (in the order of seconds or even minutes) [Franch et al., 1999]. The free energy of RNA-IN is directly $G_{\text{IN}} = 0$ and the free energy of RNA-OUT

$G_{\text{OUT}} = -20.5$ Kcal/mol (from 20 stacks and 4 loops). Then, the free energy of the state 1 reads $G_1 = G_{\text{IN}} + G_{\text{OUT}} = G_{\text{OUT}}$.

The sRNA-mRNA interaction can be divided into three different stages. In first place, the two species have to meet in order to start the interaction. This association clearly incurs in a thermodynamic penalty (denoted by $\Delta G^{\#}_{\text{inter}}$) due to entropic considerations [Xia et al., 1998]. Here, we considered $\Delta G^{\#}_{\text{inter}} \approx 10$ Kcal/mol without loss of generality (see **Figure 5**). This value can be assumed, to some extent, independent of the sequence, and it is expected to vary from in vitro to in vivo due to molecular crowding effects or spurious interactions that can take place in the cell [Schoen et al., 2009]. In second place, the seed regions of both RNAs interact to form a partly-stabilized complex (state 2; see **Figure 6b**). Here, the seed region consists of 5 nucleotides, from which $\Delta G_{\text{seed}} = -9.1$ Kcal/mol is calculated, and in the case of RNA-OUT they are located in the external loop of its structure (scenario of kissing loop). Then, the free energy of the state 2 reads $G_2 = G_1 + \Delta G^{\#}_{\text{inter}} + \Delta G_{\text{seed}}$. Note that G_2 might even be lesser than G_1 if the seed region were long enough [Salim et al., 2012]. In third place, the intermediate complex starts a process of branch migration to break the intramolecular base pairs and form the intermolecular ones, ending in a stable hybridization state between the two RNAs (see **Figure 6c**).

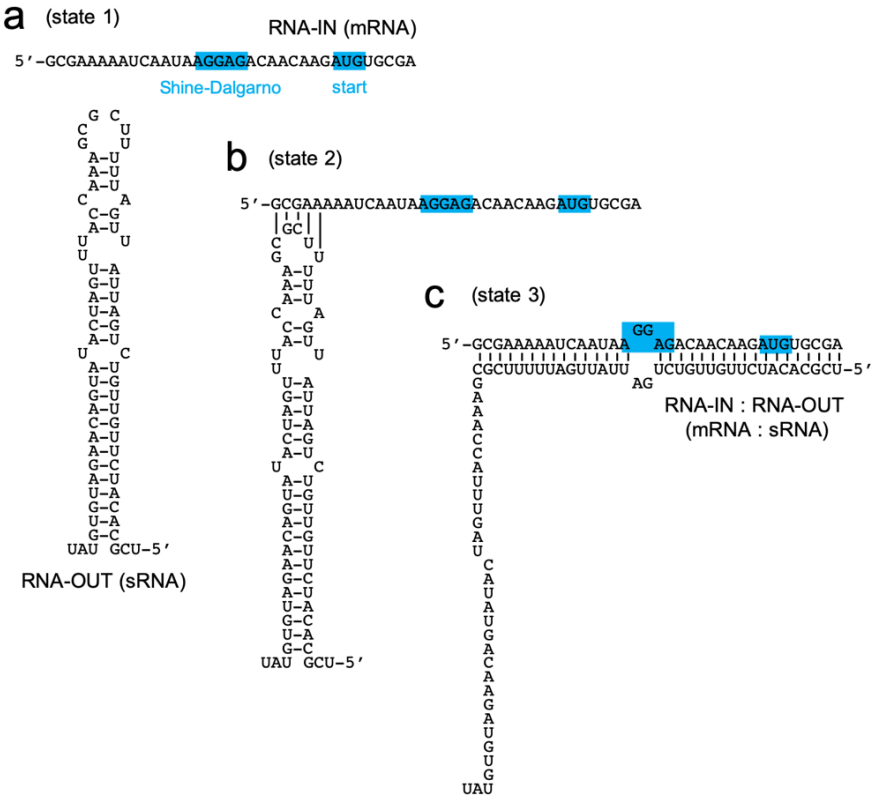

Figure 6: **a)** Intramolecular RNA secondary structures in the state 1. **b)** RNA secondary structure in the state 2 with intra- and intermolecular contacts. **c)** Intermolecular RNA secondary structure in the state 3. The Shine-Dalgarno box and the start codon are highlighted in blue.

By aggregating the free energies of all 27 stacks and 1 loop of this final complex, $G_{\text{IN:OUT}} =$

-55.8 Kcal/mol was obtained; the free energy of the state 3 is then $G_3 = \Delta G^{\#}_{\text{inter}} + G_{\text{IN:OUT}}$. For branch migration to start, nonetheless, the system incurs in a thermodynamic penalty (denoted by $\Delta G^{\text{ini}}_{\text{migr}}$), as the resulting secondary structures that need to be formed are suboptimal. We considered $\Delta G^{\text{ini}}_{\text{migr}} \approx 2$ Kcal/mol, which roughly corresponds to the gain in free energy if the external loop of RNA-OUT is enlarged. This small gain is released once the intramolecular structure is completely destroyed (at a reaction coordinate of 29 base pairs; see **Figure 5**). For simplicity, we neglected potential intramolecular interactions occurring in the 3' end of RNA-OUT upon hybridization (see **Figure 5c**). Denoting by ΔG_{migr} the free energy gap between the states 2 and 3 ($G_3 = G_2 + \Delta G_{\text{migr}}$), it is possible to compute $\Delta G_{\text{inter}} = \Delta G^{\#}_{\text{inter}} + \Delta G_{\text{seed}} + \Delta G_{\text{migr}} = -25.3$ Kcal/mol. Clearly, these calculations agree with a decomposition of the intermolecular interaction into incremental steps, i.e., $\Delta G_{\text{migr}} + \Delta G_{\text{seed}} = G_{\text{IN:OUT}} - G_{\text{OUT}}$ is satisfied.

As proposed by [Srinivas et al., 2013], this detailed energy landscape can be abstracted into a simple figure if considering the whole branch migration as a single process and that its different elementary steps are fast and reversible so that detailed balance can be applied. This means that there exists a continuous function $g : \mathbb{R} \to \mathbb{R}$ that explains the same phenomena as a function f.

Consequently, an effective free energy barrier is defined as $\Delta G^{\#}_{\text{migr}} = \Delta G^{\text{ini}}_{\text{migr}} - G_1$, noting that $|G_{\text{OUT}}|$ is indeed the gross gain of free energy during the interaction, as the intramolecular structure of RNA-OUT is destroyed. This way, the resulting energy landscape has three stable states and two energy barriers (see **Figure 7**), and a practical mathematical formulation can then be developed. According to the transition state theory, a given kinetic rate is proportional to the Boltzmann factor of the free energy barrier faced [Laidler and King, 1983].

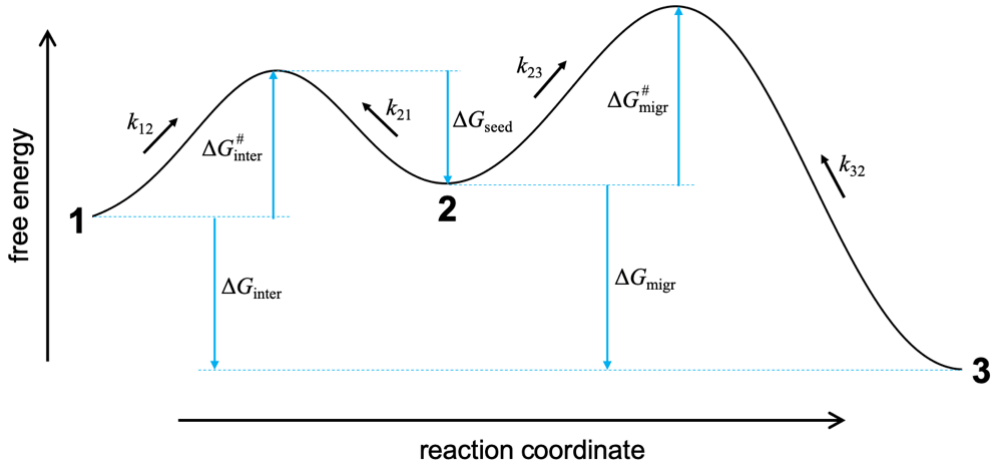

Figure 7: Abstracted energy landscape of the sRNA-mRNA interaction in the IS10 system. There are two transition states and three stable states (labelled as 1, 2, and 3). Blue vertical arrows mark the different free energies that characterize the interaction. Black arrows over the landscape designate the kinetic constants.

Thus, $k_{12} \propto e^{-\beta \Delta G^{\#}_{\text{inter}}}$, $k_{21} \propto e^{\beta \Delta G_{\text{seed}}}$ and $k_{23} \propto e^{\beta\left(\Delta G_{\text{migr}} - \Delta G^{\#}_{\text{migr}}\right)}$, where k_{ij} is the kinetic constant for the change of state i into state j.

2.3 Kinetic Modeling

Recalling the law of mass action (**Definition 1**) from **Subsection 1.3**, let x_i be the concentration of the RNA species in the folding state i, x_i denotes mRNA and x'_i denotes sRNA, then the system of nonlinear ordinary differential equations that govern the sRNA-mRNA interaction is defined as

$$\begin{aligned}
\dot{x_1} &= \alpha - k_{12} x_1 x'_1 - \delta x_1 + k_{21} x_2 \\
\dot{x'_1} &= n\alpha - k_{12} x_1 x'_1 - \delta x'_1 + k_{21} x_2 \\
\dot{x_2} &= k_{12} x_1 x'_1 - \left(k_{21} + k_{23} + \delta\right) x_2 + k_{32} x_3 \\
\dot{x_3} &= k_{23} x_2 - \left(k_{32} + \delta\right) x_3
\end{aligned} \tag{7}$$

where α is the transcription rate of the mRNA, n is the relative transcription rate of the sRNA with respect to the mRNA (e.g., due to different gene copy numbers or promoter strengths), and δ the RNA degradation rate (for simplicity, assumed equal for all species, single or complex). Of note, if we consider the total amount of mRNA, given by $y = x_1 + x_2 + x_3$, it turns out $\dot{y} = \alpha - \delta y$, which entails mass conservation.

The kinetic parameters employed are $n = 1$, $\alpha = 10$ nM/min, $\delta = 0.1$ min^{-1}, $k_{12} = 0.06$ nM^{-1}min^{-1}, $k_{21} = 6$ min^{-1}, $k_{23} = 1$ min^{-1} and $k_{32} = 0.01$ min^{-1}, corresponding to the IS10 transposition system and the corresponding boltzmann probabilities of duplexes.

The dynamic regulatory range (ratio of protein concentrations with and without sRNA) can be defined as $r = 1 - \frac{x_3}{y}$, assuming that the translation rate per mRNA is equal and substantial in the states 1 and 2, but 0 in the state 3.

In order to apply the Hartman-Grobman theorem, which allows to asses the stability of an equilibrium point by means of the linearisation of the system on that point, firstly we solve for the equilibrium points ($\dot{x_i} = 0$), and two distinct sets of solutions are obtained: $x_1 = -40.3015$, $x'_1 = -40.3015$, $x_2 = 13.9038$, $x_3 = 126.3980$ and $x_1 = 28.7249$, $x'_1 = 28.7249$, $x_2 = 7.0633$, $x_3 = 64.2118$. As negative concentrations make no sense in this context, the second set is considered for the equilibrium point study. The Jacobian matrix of the system is

$$J = \begin{pmatrix}
-0.1 - 0.06x'_1 & -0.06x_1 & 6 & 0 \\
-0.06x'_1 & -0.1 - 0.06x_1 & 6 & 0 \\
0.06x'_1 & 0.06x_1 & -7.1 & 0.01 \\
0 & 0 & 01 & -0.11
\end{pmatrix} \tag{8}$$

and the corresponding eigenvalues to the solution of interest are

$$Eig(J) = \begin{bmatrix} -10.2066 \\ -0.450411 \\ -0.1 \\ -0.1 \end{bmatrix} \tag{9}$$

As all the eigenvalues have a negative real part, the point of interest is stable, thus meaning that the RNA species evolve to a stationary picture in which all folding states are represented.The stationary picture of folding states has also been confirmed by numerical simulations (see **Figure 8**).

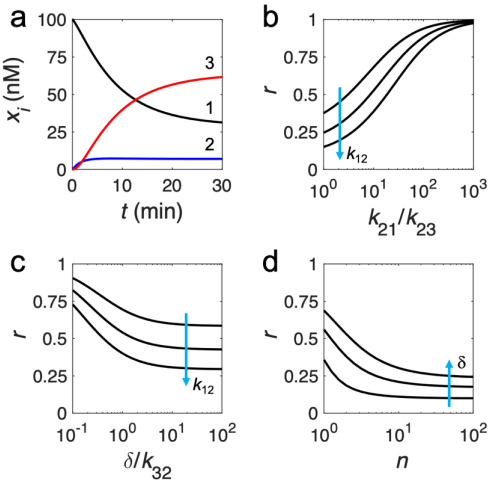

Figure 8: Numerical simulations of the RNA-based regulated response with a kinetic model. **a)** Time-dependent concentrations (x_i) of the different RNA species that correspond to the three stable states (labelled as 1, 2, and 3). The initial condition is $x_1 = 100$ nM and $x_2 = x_3 = 0$. **b**, **c**, **d)** Dynamic regulatory range (r) as a function of distinct kinetic parameters. k_{12} varies from 0.01 to 0.1 $\text{nM}^{-1}\text{min}^{-1}$ in **b**, **c)**, and δ varies from 0.1 to 0.3 min^{-1} in **d)**.

According to kinetic parameter values inferred from experimental data (in studies of kissing loop interactions) [Salim et al., 2012], this process takes about half an hour, the system is far from being completely in the final intermolecular folding state (about two thirds of mRNAs are paired to sRNAs), and the just-the-seed-paired intermolecular folding state is the least represented. Our simulations confirm that r improves by promoting the formation of the final intermolecular complex (i.e., increasing k_{23} and reducing k_{21}; see **Figure 8b)**, by enforcing the irreversibility of this process.

(i.e., increasing δ and reducing k_{32}; **Figure 8c)**, and by working with an excess of sRNA in the cell (i.e., increasing n; **Figure 8d)**. However, the improvement is limited due to the nonlinear nature of the system, consequence of accounting for the intermediate state 2. Indeed, this acts as a buffer to balance sRNA release (to go back to state 1) and branch migration (to go forward to state 3). The system can be analytically solved in steady-state if we consider that the sRNA is much more expressed than the mRNA ($n >> 1$). In this

regard, the mRNA amount limits the reaction and $x_1' \approx \frac{n\alpha}{\delta}$ is a good approximation. Thus, we obtain

$$r \approx \frac{1}{1+\frac{\frac{n\alpha}{\delta}k_{12}\left(\frac{k_{23}}{k_{32}+\delta}-1\right)}{\frac{n\alpha}{\delta}k_{12}+k_{21}+\frac{k_{23}\delta}{k_{32}+\delta}}} \approx \frac{1}{1+\left[e^{\beta(\Delta G_{\text{inter}}-\mu)}+e^{\beta(\Delta G_{\text{trans}}-\mu')}\right]^{-1}} \tag{10}$$

where $\Delta G_{\text{trans}} = \max\left(\Delta G_{\text{seed}}, \Delta G_{\text{seed}}^{\text{wild-type}}\right) + \Delta G_{\text{migr}}^{\#}$ and μ (also μ') represents a kind of chemical potential of the regulator ($e^{\beta\mu} \propto e^{\beta\mu'} \propto x_1'$. To obtain a compact expression in energetic terms, we considered k_{21} as the fastest kinetics in the system. Certainly, in the case of structured sRNAs, $\Delta G_{\text{migr}}^{\#} > -\Delta G_{\text{seed}}\}$ typically holds, meaning that once the just-the-seed-paired intermolecular folding state is reached, the reaction tends to preferentially go back (i.e., $k_{21} > k_{23}$) [Salim et al., 2012]. In addition, if n is moderate, $k_{21} > \frac{n\alpha}{\delta}k_{12}$ also holds. The contribution of ΔG_{seed} is saturated by $\Delta G_{\text{seed}}^{\text{wild-type}}$ to account for cases in which $k_{21} < \frac{n\alpha}{\delta}k_{12}$ (this saturation effect is in tune with what follows from DNA reactions in vitro [Srinivas et al., 2013]). Interestingly, ΔG_{trans} is in essence the free energy gap between the two transition states in Fig. 2, which makes **Equation 10** a satisfying combination of a standard Gibbs formulation with the Curtin-Hammett principle [Seeman, 1986].

It is worth to note that for systems in which there is a sufficient free energy gap between the states 2 and 3 (e.g., $\Delta G_{\text{migr}} < -20$ Kcal/mol), once the final intermolecular complex is formed, the probability of going back is negligible because the RNA will be degraded before this process can take place (i.e., $\delta >> k_{32}$). Hence, $x_3 \approx \frac{k_{23}}{\delta}x_2$, which already breaks the conventional scheme of thermodynamic equilibrium. In this case, the dynamic range is independent of ΔG_{inter} and the model predicts a maximal performance of $r \approx e^{\beta(\Delta G_{\text{trans}}-\mu')}$. Consequently, a stable seed pairing (e.g., enriched in GC content) and weak intramolecular structures (e.g., with internal loops and low GC content) are requirements to achieve a high dynamic range. By contrast, if we consider that the sRNA-mRNA interaction is a process much faster than the transcription and degradation of these RNA species, we recover a scenario of thermodynamic equilibrium. Indeed, $k_{12}x_1'x_1 = k_{21}x_2$ and $k_{23}x_2 = k_{32}x_3$, which means detailed balance. Consequently, it turns out

$$r \approx \frac{1}{1+\frac{n\alpha}{\delta}\frac{k_{12}k_{23}}{k_{21}k_{32}}} = \frac{1}{1+e^{-\beta(\Delta G_{\text{inter}}-\mu)}} \tag{11}$$

the conventional Fermi-Dirac statistics [Phillips, 2015] that was here considered as the null model. Note that by evaluating $lim_{\mu'\to\infty}$ on **10**, we get **11**.

2.4 Experimental Validation

To validate our theory, we considered the dynamic ranges reported for a large series of mutants of the IS10 system (mutations introduced in both the sRNA and the mRNA), which were experimentally measured by changes in fluorescence [Mutalik et al., 2012]. For each system, we calculated the values of ΔG_{inter} with the RNAcofold routine of the ViennaRNA package [Hofacker et al., 1994]. We obtained values that differ a bit from our previous manual calculations as a consequence of a different value of $\Delta G_{\text{inter}^{\#}}$ and the consideration of

additional physicochemical features [Mathews et al., 1999]. Moreover, we calculated the corresponding values of ΔG_{seed} by following the manual procedure aforementioned, considering the sequence motifs previously described in the species RNA-IN and RNA-OUT [Mutalik et al., 2012]. Only the most stable consecutive stackings were considered in the case of an intermolecular structure between the seed regions with a loop or bulge. To estimate $\Delta G_{\text{migr}^{\#}}$, we calculated $|G_{\text{OUT}}|$ with the RNAfold routine [Hofacker et al., 1994] by imposing the seed region to be unpaired, and then we added the value of $\Delta G^{\text{ini}}_{\text{migr}}$. For RNA-OUT mutants in which the external loop is very large, we set $\Delta G^{\text{ini}}_{\text{migr}} \approx 0$, as the base pairing that follows the seeding is not a process of branch migration but between unpaired regions. We then represented the experimental values of r vs. the computed free energies. For representation purposes, we defined $\phi = \left[e^{\beta(\Delta G_{\text{inter}}-\mu)} + e^{\beta(\Delta G_{\text{trans}}-\mu')}\right]^{-1}$, revealing a good agreement between theory and experiments with our novel scaling law from non-linear thermodynamics (see **Figure 9a**).

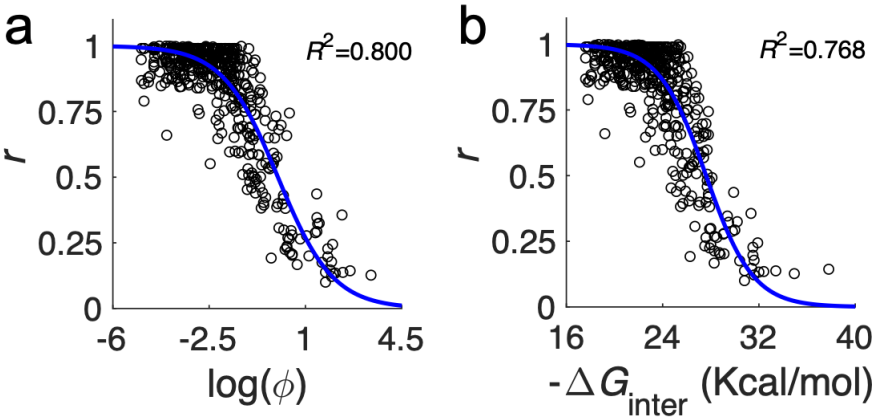

Figure 9: Correlation of experimentally measured dynamic range (r) against a computationally predicted free energy for different mutants of the IS10 system (529 data points). **a)** Non-equilibrium thermodynamic model (r vs. ϕ), fitted with $\beta = 0.545$ mol/Kcal, $\mu = -26.6$ Kcal/mol, and $\mu' = 20.7$ Kcal/mol. **b)** Equilibrium thermodynamic model (r vs. ΔG_{inter}), fitted with $\beta = 0.514$ mol/Kcal and $\mu' = -27.6$ Kcal/mol.

ΔG_{inter} still exhibits substantial prediction ability (see **Figure 9b**), as $e^{\beta(\Delta G_{\text{inter}}-\mu)}$ is the dominant term in ϕ for most of the mutant systems. Nonetheless, an F test showed a statistically significant better prediction ability in the former case ($F_{1,526} = 81.7$, $P < 10^{-6}$). Therefore, for a single gene, expression based on RNA-RNA interactions is more suited to be modelled under a non-thermodynamic equilibrium approach.

3 Gene expression regulation role in cellular Metabolism

In the previous section, a local non-equilibrium thermodynamic approach regarding RNA-RNA interaction-based gene expression regulation was addressed. However, a cell's gene expression profile cannot be appropriately predicted exclusively using this approach as there are many more mechanisms that play their roles in regulation, and consequently, the resulting phenotype can differ from the predicted one. Intuitively, this issue can be solved by considering each gene, its regulation mechanisms, and their interactions and then adding up their individual effects to predict a phenotype. Nevertheless, the latent incompleteness of the data required in terms of unknown gene functions, gene products, and interactions, makes this task impossible. Thus, a *"top-down"* approach seems more suitable as it allows us to conclude even though some mechanisms and interactions occurring inside the cell remain unknown.

From a thermodynamic point of view, a cell is an open system; namely, it is in constant matter and energy exchange with its surroundings. Moreover, cells are organized structures which have a guided information flux, as stated by the central dogma of molecular biology. Clearly, for a cell to proliferate and survive, it needs to maintain its internal order, which is entropically unfavorable at the cost of requiring a constant supply of energy. This ceases when a cell dies, leading the cell's internal composition towards thermodynamic equilibrium. So, a cell is, by natural definition, a thermodynamical non-equilibrium system [Nelson et al., 2017]. The cell obtains energy by means of its internal metabolism, which in turn depends on its gene expression profile, as biochemical reactions are possible thanks to proteic catalyzers known as enzymes. In addition, as a cell's phenotype can be characterized by the metabolites (substances), it consumes and produces, it is worth analyzing and modelling a cell's metabolism as it provides indirect information about expression profiles in a thermodynamic non-equilibrium context. The branch of Systems Biology devoted to studying and predicting a cell's metabolism employs metabolic models for these purposes.

3.1 Metabolic models and Flux Balance Analysis

Metabolic models are a mathematical representation of a cell's metabolism as a network of metabolites transformed into other metabolites through biochemical reactions. These models allow for a holistic view of the cell's biochemical functioning. The target variables of study in these models are the metabolic fluxes, namely, the rates of the metabolic reactions defined as the number of substrate molecules traversing each metabolic reaction per unit time, which describe how mass and energy flow through the metabolic network.

The elaboration of one of these models is tedious. It involves exploration of diverse databases, such as KEGG, BioCyc, ENZYME, BRENDA or BiGG among others, and is performed in a semi-automatic manner, needing a final manual curation for error pruning. Next, this network has to be analyzed in some manner, being Flux Balance Analysis (FBA), the chosen tool. FBA is employed to compute the flow of metabolites through the reconstructed network, which allows making predictions about the cell's growth rate or the production rate of a certain compound of interest. To perform FBA on a reconstructed net-

work, the system needs to be defined in such a way that reactions products are computable. The representation employed is a particular kind of adjacency matrix of a weighted graph, known as the Stoichiometric matrix.

Definition 7 (Stoichiometric matrix)**.** *A stoichiometric matrix $S \in \mathbb{R}^m \times \mathbb{R}^n$ is a sparse matrix representing the metabolic network of an organism, where the rows $i \in \{1, \ldots, m\}$ correspond to the metabolites and the columns $j \in \{1, \ldots, n\}$ to the reactions. Each value $s_{i,j}$ corresponds to the stoichiometric coefficient of the metabolite i participating in the reaction j. Metabolites which fulfill $s_{i,j} < 0$ are substrates of the reaction j, metabolites which fulfill $s_{i,j} > 0$ are products of the reaction j while metabolites with $s_{i,j} = 0$ do not participate in the reaction j.*

In simpler terms, the stoichiometric matrix is a representation of the topology of the constructed metabolic network, which incorporates the stoichiometry of the biochemical reactions of said network. The sparsity of S is due to the small number of metabolites that simultaneously participate in a single reaction. It is common that $n > m$ as the number of reactions is greater than the number of metabolites. The stoichiometric matrix allows defining the mass balance of the cell's metabolism, where the concentration change of each metabolite over time equals the difference between the rates at which it is produced and consumed. Formally, it can be formulated as

$$\frac{d\boldsymbol{X}}{dt} = S\boldsymbol{\nu} \tag{12}$$

where $\boldsymbol{X} \in \mathbb{R}^m$ is the concentration of metabolites, S is the stoichiometric matrix (**Definition 7**) and $\boldsymbol{\nu} \in \mathbb{R}^n$ are the fluxes of all reactions composing the metabolic network of interest. Besides, as metabolic regulation achieves a balance between the rate of input of a substrate and the rate of substance degradation or consumption plus the fact that metabolic transients are faster than cellular growth rates and changes in the cell's surroundings. It is said that metabolic fluxes can be considered to be in steady-state [Nelson et al., 2017, Varma and Palsson, 1994].

Definition 8 (Steady state)**.** *A system is in steady state if its state variables are unchanging in time, yielding*

$$\frac{\partial x}{\partial t} = 0$$

where x is a state variable of the system of interest.

Although a metabolic flux is variable and subject to metabolic demands, steady-state is maintained by balancing the rate of the substrate provided by a preceding reaction in the metabolic pathway and the rate that the substrate is converted into a product, which keeps substrate concentration constant. Steady states are not unique. In fact, cells adjust their internal composition through gene expression machinery to reach a new steady-state, which may be more convenient for survival. Applying **Definition 8** on metabolite concentration in **Equation 12**, it yields

$$S\nu = \mathbf{0} \tag{13}$$

There is a negligible accumulation or depletion of intracellular metabolites. Their concentrations are invariant with time and, thus, balanced. Additionally, it assumes a steady-state that unlinks the metabolic network of kinetic and gene expression regulation data of individual enzymes. Any ν that satisfies **Equation 13** is in the null space of S. As mentioned formerly, there are usually more reactions than there compounds ($n > m$); namely, there are more unknown variables than equations being thus an underdetermined the system, i. e. there are multiple possible solutions. In order to delimit the functional states of the system, i. e. getting rid of those solutions that biologically infeasible [Price et al., 2004], system constraints have to be defined. These constraints are related to reaction directionality, enzyme capacity and/or environmental conditions.

When defining constraints mathematically, a particular distinction between balances and bounds can be made [Price et al., 2004]. Mass conservation and experimental information about a certain flux are examples of balances, ultimately resulting in equality constraints. On the other hand, directionality and reversibility of biochemical reactions for example are defined as bounds.

Definition 9 (Reaction reversiblity)**.** *A reaction is reversible when the products can react and yield the reactants. A certain reaction i is reversible if its flux $\nu_i \in (-\infty, +\infty)$. Trivially, a reaction i is irreversible if its flux $\nu_i \in [0, +\infty)$.*

Additionally, these intervals can be customized using experimental data, i. e. flux bounds can be set to real values to find biologically feasible solutions relevant to the user's interest. Now that the main building blocks have been introduced, the system can be presented as the following optimization problem.

$$\max_{\nu} \quad Z(\nu) = c^T \nu \tag{14a}$$

$$\text{subject to:} \quad S\nu = \mathbf{0} \tag{14b}$$

$$a_j \leq \nu_j \leq b_j \qquad j \in \{1, \ldots, n\} \tag{14c}$$

where $c \in \mathbb{R}^n$ is a vector that contains the weights of each reaction on the objective function, a_j, and b_j are the lower and upper bounds for ν_j. For a reversible flux, a_j is set arbitrarily as a big negative number (for example, -1000) while b_j is set as a big positive number (for example, 1000). In the case of an irreversible flux, $a_j = 0$ while the definition of b_j is equivalent to the case of reversible fluxes. These values can be restricted up to such point that the system is forced to a solution in which, for example, a certain flux $\nu_j = \alpha, \quad \alpha \in \mathbb{R}$. This is accomplished by setting $a_j = \alpha = b_j$. The solution yielded by FBA is a set of fluxes corresponding to a particular optimal reaction network state. **Figure 10** is a summary of the aforementioned FBA process.

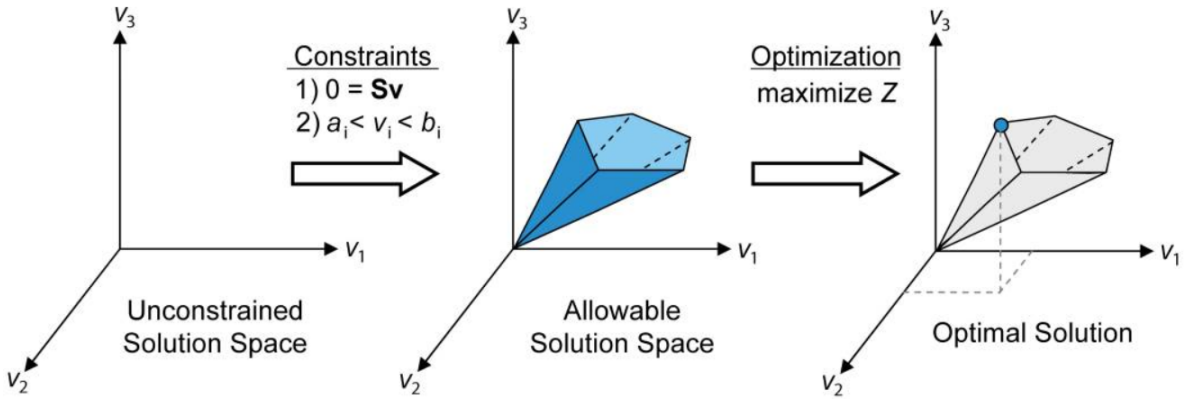

Figure 10: Constraining, politope construction and optimal solution choice by FBA, courtesy of [Orth et al., 2010].

3.2 Multi Objective optimization

One of the main disadvantages of employing FBA is that optimization is performed in such a way that only one objective is satisfied, which doesn't agree with a cell's natural behavior as cells perform various activities depending on their needs (growth, secondary metabolite production, stress responses...), and for that the different cellular resources are accordingly distributed. In addition, in FBA restrictions have to be linear. It is therefore interesting to tackle the matter from a Multi Objective Problem (MOP) perspective. The classical definition of a multi-objective problem is the following:

$$\min_{x \in X} \quad \boldsymbol{f}(x) = \quad [f_1(x), \ldots, f_k(x)] \tag{15a}$$

$$\text{subject to:} \quad \boldsymbol{g}(x) \leq \boldsymbol{0} \tag{15b}$$

$$\boldsymbol{h}(x) = \boldsymbol{0} \tag{15c}$$

$$a_j \leq x_j \leq b_j, \qquad j \in \{1, \ldots, n\} \tag{15d}$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, $\boldsymbol{f}(x)$ is the objective vector, with $k \geq 2$ components, $\boldsymbol{g}(x)$ are the inequality constraints, $\boldsymbol{h}(x)$ are the equality constraints, a_j and b_j are the lower and upper bounds, respectively, of x_j in the decision space. Note that if an objective is to be maximized, it can be rewritten in terms of minimization since $\max_{x \in X} f_q(x) = -\min_{x \in X} -f_q(x)$ with $q \in \{1, \ldots, k\}$. In general, it is not possible to find a unique solution that is optimal for all objectives in a simultaneous manner. However, there exists a space of sub-optimal solutions, that are more optimal for certain objectives in detriment of other objectives. [Miettinen, 1999] introduced some concepts that allow the comparison between different sub-optimal solutions in order to select the best candidate solution.

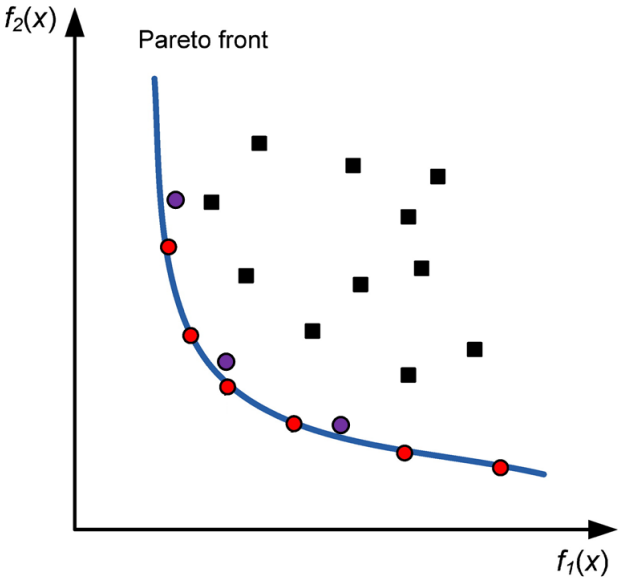

Figure 11: The objective vector solution space of a bi-objective optimization problem, where $f_1(x)$ and $f_2(x)$ are said objectives. The blue curve represents the pareto front, which consists of infinite Pareto optimal solutions. The Pareto front's epigraph denotes the feasible solution space. The red dots are particular Pareto optimal solutions. The purple dots are non-dominated solutions, while the black squares represent dominated solutions. Figure modified from [Huang et al., 2015].

Definition 10 (Dominance)**.** *An objective vector* $\boldsymbol{f}(x^a)$ *dominates another objective vector* $\boldsymbol{f}(x^b)$ *if* $\boldsymbol{f}_k(x^a) \leq \boldsymbol{f}_k(x^b)$*, for all* $k \in \{1, \ldots, q\}$ *and* $\boldsymbol{f}_k(x^a) < \boldsymbol{f}_k(x^b)$ *for at least one* k*. Dominance will be denoted with* $\preccurlyeq$*, where* $\boldsymbol{f}(x^a) \preccurlyeq \boldsymbol{f}(x^b)$ *implies that* $\boldsymbol{f}(x^a)$ *dominates* $\boldsymbol{f}(x^b)$*.*

It has strict dominance if $\boldsymbol{f}_k(x^a) < \boldsymbol{f}_k(x^b)$*, for all* $k \in \{1, \ldots, q\}$*.*

It has weak dominance if $\boldsymbol{f}_k(x^a) \leq \boldsymbol{f}_k(x^b)$*, for all* $k \in \{1, \ldots, q\}$*.*

Definition 11 (Pareto optimality)**.** *An objective vector* $\boldsymbol{f}(x^*)$ *is Pareto optimal if a* $\boldsymbol{f}(x)$ *such that* $\boldsymbol{f}(x) \preccurlyeq \boldsymbol{f}(x^*)$ *does not exist.*

A less formal definition would be that none of a Pareto optimal objective vector components cannot be improved without deteriorating at least one of its other components. That is to say, a Pareto optimal solution is not dominated. The set of Pareto optimal solutions is known as Pareto set (see **Definition 12**).

Definition 12 (Pareto set)**.** *The set* $V_P := \{x^* : \boldsymbol{f}(x^*) \text{ is Pareto optimal}\}$*, where* $\boldsymbol{f}(x^*)$ *is the objective vector of the solution* x^**, is known as Pareto set. Complementary, the set* $Z_P := \{\boldsymbol{f}(x^*) : x^* \in V_P\}$ *is known as Pareto front.*

An ilustrative example is provided in **Figure 11**.

The Pareto front, however, is unknown and has to be approximated with non-dominated solutions (see **Figure 11**). The set of non-dominated objective vectors conform the Pareto front approximation (denoted as Z_P^*) and their corresponding decision variables conform the

Pareto set approximation (denoted as V_P^*). Furthermore, in order to narrow down the Pareto front, [Miettinen, 1999] defined the following concepts:

Definition 13 (Utopian objective vector)**.** *The utopian objective vector* $\boldsymbol{f}^{\text{utopia}}$ *is an objective vector whose components are obtained by solving*

$$\begin{aligned} \min_{x \in X} \quad & \boldsymbol{f}_q(x) && \text{for all} \quad q \in \{1, \dots, k\} && \text{(16a)} \\ \text{subject to:} \quad & \boldsymbol{g}(x) \leq \boldsymbol{0} && && \text{(16b)} \\ & \boldsymbol{h}(x) = \boldsymbol{0} && && \text{(16c)} \\ & a_j \leq x_j \leq b_j && j \in \{1, \dots, n\} && \text{(16d)} \end{aligned}$$

thus it is a solution which is minimal for all objectives simultaneously.

In rare cases where this vector exists, it is the solution to the multi-objective problem. However, as mentioned previously in this work, the usual is that there is no unique solution that is optimal for all objectives at the same time. Nevertheless, it is a valuable being since it contains the lower bounds of the Pareto optimal solutions, although its components cannot be directly computed as the Pareto optimal set is unknown. That is why this vector has to be approximated.

Definition 14 (Ideal objective vector)**.** *The ideal objective vector* $\boldsymbol{f}^{\text{ideal}}$ *is an unfeasible objective vector whose components are defined by*

$$\boldsymbol{f}^{\text{ideal}} = \boldsymbol{f}^{\text{utopia}} + \varepsilon_q \qquad \text{for all} \quad q \in \{1, \dots, k\}$$

where $\varepsilon_q > 0$.

Definition 15 (Nadir objective vector)**.** *The nadir objective vector* $\boldsymbol{f}^{\text{nadir}}$ *is an objective vector whose components are obtained by solving*

$$\begin{aligned} \max_{x \in X} \quad & \boldsymbol{f}_q(x) && \text{for all} \quad q \in \{1, \dots, k\} && \text{(17a)} \\ \text{subject to:} \quad & \boldsymbol{g}(x) \leq \boldsymbol{0} && && \text{(17b)} \\ & \boldsymbol{h}(x) = \boldsymbol{0} && && \text{(17c)} \\ & a_j \leq x_j \leq b_j && j \in \{1, \dots, n\} && \text{(17d)} \end{aligned}$$

Thus, parallel to $\boldsymbol{f}^{\text{utopia}}$, $\boldsymbol{f}^{\text{nadir}}$ contains the upper bounds of the Pareto optimal solutions, and again it is not possible to compute as the Pareto optimal set is unknown and has to be approximated. However, this approximation is taken as the true nadir objective vector.

In this work, following the steps of [Siurana, 2017] to solve a multi-objective flux optimization problem, the *Generate-First Choose Later*[9] (GFCL) approach based on differential evolution embedded with a spherical pruning mechanism, dubbed sp-MODE (*spherical pruning Multi-Objective Differential Evolution*) designed by [Reynoso-Meza et al., 2012] will be employed. This strategy has several advantages: first, there is no information loss from the

[9]Generating many potential Pareto optimal solutions and then selecting the most preferable.

Pareto front (such as trade-off levels between solutions), which are not available in *Aggregate Objective Function* (AOF) derived approaches. The variables of interest, just like in FBA, will be metabolic fluxes, which are dependent on the stoichiometric model of the metabolic network. The objectives and constraints will be once more balances and bounds, although in this scenario, unlike in FBA, metabolic redundancies can be taken into account. Due to these reasons, the problem to be solved is of the large scale multi-modal type. In the next section, this algorithm is explained in detail.

3.2.1 sp-MODE algorithm

Like other multi-objective evolutionary algorithms, the first step consists of generating an initial population of solutions based on some initial distribution. Then, the fitness with respect to each individual's objective functions coming from the initial population is evaluated. This allows the selection of non-dominated solutions from the current population in order to form a first approximation of the Pareto set. Next, the evolutionary process begins: every generation, a new population of solutions is generated by means of the mutation and crossover operators given by the differential evolution mechanism.

The mutation operator generates a mutant vector $\boldsymbol{y}^i|_G$ for each parent vector $\boldsymbol{v^i}|_{\boldsymbol{G}}$ at generation G following **Equation 18**:

$$\boldsymbol{y}^i|_G = \boldsymbol{v}^{r_1}|_G + \lambda\left(\boldsymbol{v}^{r_2}|_G - \boldsymbol{v}^{r_3}|_G\right) \tag{18}$$

where $r_1 \neq r_2 \neq r_3 \neq i$ are randomly selected and λ is known as the scaling factor.

The crossover operator generates a child vector $\boldsymbol{x}^i|_G = [x_1^i|_G, \ldots, x_n^i|_G,]$ for each target vector $\boldsymbol{v^i}|_{\boldsymbol{G}}$ and its mutant vector $\boldsymbol{y}^i|_G$ following **Equation 19**:

$$x_j^i|_G = \begin{cases} y_j^i|_G & if\ rand(0,1) \leq Cr \\ v_j^i|_G & otherwise \end{cases} \tag{19}$$

where $j \in \{1, \ldots, n\}$ and Cr is the crossover probability rate.

Then, the individuals of the new population (*children*) are evaluated with respect to the objective functions. Then, these individuals are selected if they strictly dominate their respective parents. Thus the Pareto set approximation is updated with preferable solutions. In addition, to avoid the accumulation of similar solutions and to prevent diversity decrease, a pruning mechanism based on spherical coordinates, which are normalized with respect to a reference, is applied. This mechanism divides the objective space in a number of spherical sectors, and then it selects a single solution according to a certain index), which will populate each sector (see **Figure 12**).

Before explaining with detail the procedure, it is required to introduce certain definitions.

Definition 16 (Normalised spherical coordinates)**.** *Let v^1 and $\boldsymbol{f}(v^1)$, then the normalised spherical coordinates from a reference solution $\boldsymbol{f}^{ref}$ are*

$$S\left(\boldsymbol{f}\left(v^1\right)\right) = \left[\left\|\boldsymbol{f}\left(v^1\right)\right\|_2, \boldsymbol{\beta}\left(\boldsymbol{f}\left(v^1\right)\right)\right]$$

where D is the dimension of the objective space, $\boldsymbol{\beta}(\boldsymbol{f}(v^1)) = [\beta_1(\boldsymbol{f}(v^1)), \ldots, \beta_{D-1}(\boldsymbol{f}(v^1))]$ is the arc vector and $\|\boldsymbol{f}(v^1)\|_2$ is the Euclidean distance with the reference solution.

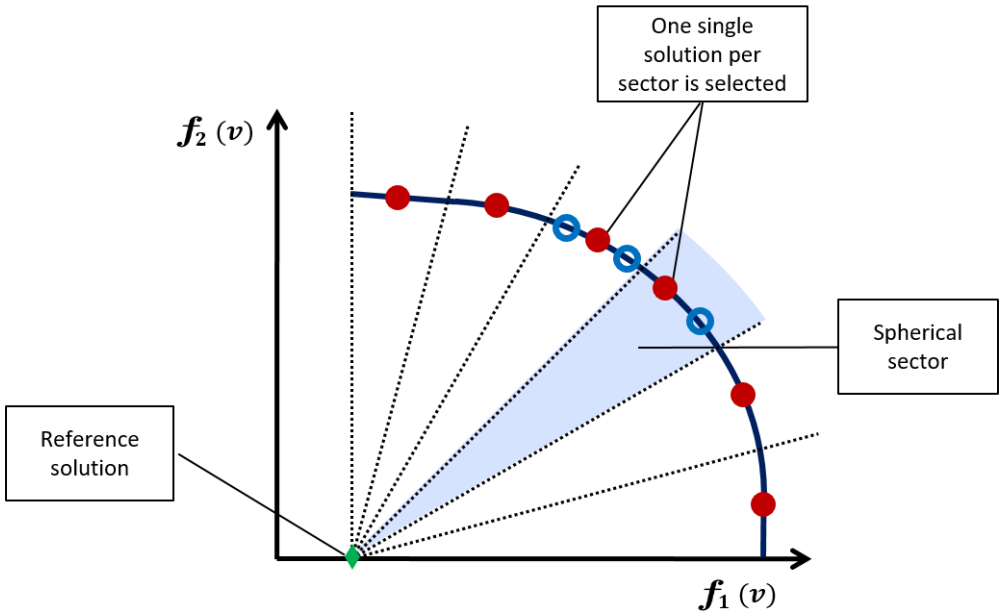

Figure 12: Spherical prunning mechanism on a bi-objective space, courtesy of [Siurana, 2017].

Note that $\boldsymbol{f}^{\text{ref}}$ must dominate existing solutions. In order to ensure that this condition is fulfilled for any possible population, $\boldsymbol{f}^{\text{utopia}}$ (see **Definition 13**) or its approximation $\boldsymbol{f}^{\text{ideal}}$ (see **Definition 14**) are employed as the reference. Since during the evolutionary process, the population composition is constantly changing, the components of $\boldsymbol{f}^{\text{ideal}}$ and $\boldsymbol{f}^{\text{nadir}}$ (see **Definition 15**) have to be recomputed for each generation.

Definition 17 (Sight range)**.** *The sight range from the reference solution $\boldsymbol{f}^{ref}$ to the Pareto fron approximation $\boldsymbol{f}_P^*$ is bounded by $\boldsymbol{\beta}^U$ and $\boldsymbol{\beta}^L$:*

$$\boldsymbol{\beta}^U = \left[\max \beta_1 \left(\boldsymbol{f}\left(v^i\right)\right), \ldots, \max \beta_{D-1} \left(\boldsymbol{f}\left(v^i\right)\right)\right] \quad \textit{for all} \quad \boldsymbol{f}\left(v^i\right) \in \hat{A}|_G \tag{20a}$$

$$\boldsymbol{\beta}^U = \left[\min \beta_1 \left(\boldsymbol{f}\left(v^i\right)\right), \ldots, \min \beta_{D-1} \left(\boldsymbol{f}\left(v^i\right)\right)\right] \quad \textit{for all} \quad \boldsymbol{f}\left(v^i\right) \in \hat{A}|_G \tag{20b}$$

where A is the set of the best solutions found so far in the iterative process, $\hat{A}|_G$ is the set of solutions found at generation G prior to prunning and $A|_G$ is the resulting set after prunning.

In the case of $\boldsymbol{f}^{\text{ref}} = \boldsymbol{f}^{\text{utopia}}$, it is trivial to prove that $\boldsymbol{\beta}^U = \left[\frac{\pi}{2}, \ldots, \frac{\pi}{2}\right]$ and $\boldsymbol{\beta}_L = [0, \ldots, 0]$.

Definition 18 (Spherical grid)**.** *Given a set of solutions in the objective space, the spherical grid on the $D-$ dimensional is*

$$\boldsymbol{\Lambda}^{\boldsymbol{f}_P^*} = \left[\frac{\beta_1^U - \beta_1^L}{\beta_1^\varepsilon}, \ldots, \frac{\beta_{D-1}^U - \beta_{D-1}^L}{\beta_{D-1}^\varepsilon}\right]$$

where $\boldsymbol{\beta}_\varepsilon = \left[\beta_1^\varepsilon, \ldots, \beta_{D-1}^\varepsilon\right]$ are arc increments.

Definition 19 (Spherical sector)**.** *The normalised spherical sector of a solution* v^1 *is defined as*

$$\mathbf{\Lambda}_\varepsilon\left(v^1\right) = \left[\left\lceil \frac{\beta_1\left(\boldsymbol{f}\left(v^1\right)\right)}{\Lambda_1^{\boldsymbol{f}_P^*}} \right\rceil, \ldots, \left\lceil \frac{\beta_{D-1}\left(\boldsymbol{f}\left(v^1\right)\right)}{\Lambda_{D-1}^{\boldsymbol{f}_P^*}} \right\rceil\right]$$

Definition 20 (Spherical pruning)**.** *Gven two solutions* v^1 *and* v^2 *from a set,* v^1 *has preference over* v^2 *in the spherical sector iff*

$$\left[\mathbf{\Lambda}_\varepsilon\left(v^1\right) = \mathbf{\Lambda}_\varepsilon\left(v^2\right)\right] \wedge \left[\left\|\boldsymbol{f}\left(v^1\right)\right\|_p < \left\|\boldsymbol{f}\left(v^2\right)\right\|_p\right]$$

where $\left\|\boldsymbol{f}\left(v\right)\right\|_p = \left(\sum_{a=1}^{D}\left|\boldsymbol{f}_a\left(v\right)\right|^p\right)^{1/p}$ *is a suitable p-norm.*

The spherical pruning procedure occurs after applying the dominance criterion (**Definition 10**) over the new computed solutions in a generation. What it does is that it calculates the normalized spherical coordinates for each member of the set of solutions (**Definition 16**). Next, the spherical grid is built (**Definition 17** and **18**) and the spherical sector (**Definition 19**) for each member of the set of solutions is computed. Finally, for each solution, it perseveres if no other solution has the same spherical sector (i. e. it is alone in a spherical sector) or if it has the lowest norm amongst solutions that share a spherical sector. Once the whole optimization process has arrived at completion, a set of solutions is obtained (i. e. the Pareto set approximation), which exhibits different levels of the trade-off between objectives. Although the spherical pruning mechanism has gotten rid of solutions that were thought to be non-pertinent, it is up to the researcher to evaluate the pertinency of the solutions one by one through specific metrics or criteria.

3.3 Resource allocation model

To model how a cell distributes their proteic resources, [Goelzer et al., 2011] modified the traditional FBA approach by considering that the set of proteins of a cell can be divided into 3 subsets: metabolic-related proteins, ribosomal proteins, and growth-related proteins. To describe the FBA modifications applied to the traditional FBA, some notation is needed: for a given integer N_x, I_x denotes the set $\{1, \ldots, N_x\}$, N_m denotes the number of enzymes, N_i the number of internal metabolites, N_p the number metabolic precursors, N_r the recycled metabolites, N_G the proteins on other cell subsystems needed for growth, N_c the macro components (i. e. DNA, the cell lipid membrane or the cell wall), μ the cell growth rate, R the ribosome concentration, ν metabolic fluxes and P_G protein concentration.

Besides the global constraint of steady state (see **Equation 13**)The 3 protein subsets have to satisfy certain constraints to ensure a proprer coordination at $\mu > 0$.

a) **Metabolic capability** constraint. The network has

   1) to be large enough to produce all metabolic precursors required for growth. Thus, the fluxes of the metabolic precursors produced by the network have to be greater than

or equal to the ones consumed during the synthesis of the cell components. For all $i \in I_p$

$$-\sum_{j=1}^{N_m} S_{p_{ij}}\nu_j + \mu\left(\sum_{j=1}^{N_m} C_{M_{ij}}^{M_p}|\nu_j| + C_{R_i}^{M_p} + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_p}\right) - \nu_Y \leq 0 \tag{21}$$

where S_{ij} is a sub-part of the stoichiometric matrix S relative to a metabolite of interest, ν_Y denotes the exchange flux of said metabolite with the cell's environment. $C_{M_{ij}}^{M_p}$, $C_{R_i}^{M_p}$ and $C_{G_{ij}}^{M_p}$ denote the amount of said metabolite needed for ribosome synthesis, enzyme synthesis and protein synthesis.

2) to be large enough to maintain the concentration of all macrocomponents (X_c) equal to its average $\left(\hat{X}_c\right)$. Let S_c be the sub-part of the stoichiometric matrix related to macrocomponents, then for all $i \in I_c$

$$-\sum_{j=1}^{N_m} S_{c_{ij}}\nu_j + \mu\bar{X}_{c_i} \leq 0 \tag{22}$$

3) to be large enough to recycle the metabolites produced during the synthesis of cell components. For all $i \in I_r$

$$\sum_{j=1}^{N_m} S_{r_{ij}}\nu_j + \mu\left(\sum_{j=1}^{N_m} C_{M_{ij}}^{M_r}|\nu_j| + C_{R_i}^{M_r}R + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_r}P_{G_j}\right) \leq 0 \tag{23}$$

where S_r is a sub-part of the stoichiometric matrix S relative to a recycled metabolite (X_r). Similarly to **Equation 21**, $C_{M_{ij}}^{M_r}$, $C_{R_i}^{M_r}$ and $C_{G_{ij}}^{M_r}$ denote the amount of said recycled metabolite needed for ribosome synthesis, enzyme synthesis and protein synthesis.

b) **Translation capability** constraint. The cellular translation machinery capability has to suffice to maintain a constant concentration of all cell proteins at a given growth rate μ:

$$\mu\left(\sum_{j=1}^{N_m} C_{M_j}^{R}|\nu_j| + C_R^R R + \sum_{j=1}^{N_G} C_{G_j}^{R}P_{G_j}\right) - k_T R \leq 0 \tag{24}$$

where k_T is the translation machinery's efficiency. Once again, similar to **Equation 21**, $C_R^R > 0$, $C_{M_j}^R > 0$ and $C_{G_j}^R > 0$ are the total amount of aminoacids required for ribosome synthesis, enzyme synthesis and protein synthesis respectively.

c) **Density** constraint. In order for the cell components (proteins, metabolites, macrocomponents...) to efficiently diffuse along the cytosol, the cell must maintain a constant internal density [citar a nabos aqui]:

$$\sum_{j=1}^{N_m} C_{M_j}^{D}|\nu_j| + C_R^D R + \sum_{j=1}^{N_G} C_{G_j}^{D}P_{G_j} - \bar{D} \leq 0 \tag{25}$$

where $\bar{D}$ is the mean density measured in aminoacid residues of cell components. $C_{M_j}^D > 0$, $C_{G_j}^D > 0$ and $C_R^D > 0$ are the density values in aminoacid residues for enzymes, proteins and ribosomes respectively.

The optimization problem of maximizing growth rate μ from a metabolic network with a traditional stoichiometric matrix S extended with protein synthesis reactions, subject to the the aforementioned restrictions, has been expressed in matrix form by [Groot et al., 2019] as follows:

$$A\left(x,\mu\right)\cdot\begin{bmatrix}\nu\\ \nu_{\text{synth}}\\ \nu_{\text{synth,rib}}\\ \mu\end{bmatrix}=\begin{bmatrix}S & -M_{\text{enz}} & -M_{\text{rib}} & -x\\ \left[\frac{\mu}{k_{\text{cat},j}}\right]_{r\times r} & -\mathbb{I}_{r\times r} & \mathbf{0}_{r\times 1} & 0\\ \mathbf{0}_{1\times r} & \mathbf{1}_{1\times r} & 1-\frac{k_{\text{cat,rib}}}{\mu} & 0\\ \mathbf{0}_{1\times r} & \left[\rho_j\right]_{1\times r} & \rho_{\text{rib}} & -1\end{bmatrix}\cdot\begin{bmatrix}\nu\\ \nu_{\text{synth}}\\ \nu_{\text{synth,rib}}\\ \mu\end{bmatrix}=0 \quad (26)$$

where $A(x,\mu)$ is an extended matrix whose terms depend on cellular growth rate μ and the concentrations of metabolites x; ν are metabolite fluxes; S is the stoichiometric matrix, M are stoichiometric matrices denoting metabolites consumed for a specified entity synthesis (speficied by the underscore), $\left[\frac{\mu}{k_{\text{cat},j}}\right]_{r\times r}$ is the diagonal matrix in which the j^{th} diagonal element is the cell's growth rate divided by the catalytic capacity of the enzyme j, $\mathbb{I}_{r\times r}$ is the identiity matrix of dimensions $r\times r$, the row vector $\left[\rho_j\right]_{1\times r}$ contains the volumetric information about enzymes j and ρ_{rib} contains the volumetric information about the ribosome.

3.4 Simulations

By means of [Bulović et al., 2019]'s python routine that automatically constructs a resource allocation model described by **Equation 26**, we constructed said model for *Escherichia coli* K12 corresponding to a fixed growth rate μ and developed a framework for it to be exploited by [Reynoso-Meza et al., 2012]'s spMODE algorithm. The objectives employed were solving for **Equation 26** without modifications, maximizing growth, maximizing ribosome synthesis (i. e. simply maximizing a subset of reactions of the system) and parsimony (i. e. force the system to yield flux solutions that are close to experimental fluxes provided to the algorithm). In total 10 models (corresponding to increasing percentages of $\mu_{\max} \approx 0.6808 \mathrm{h}^{-1}$) were tested (see **Figure 13**).

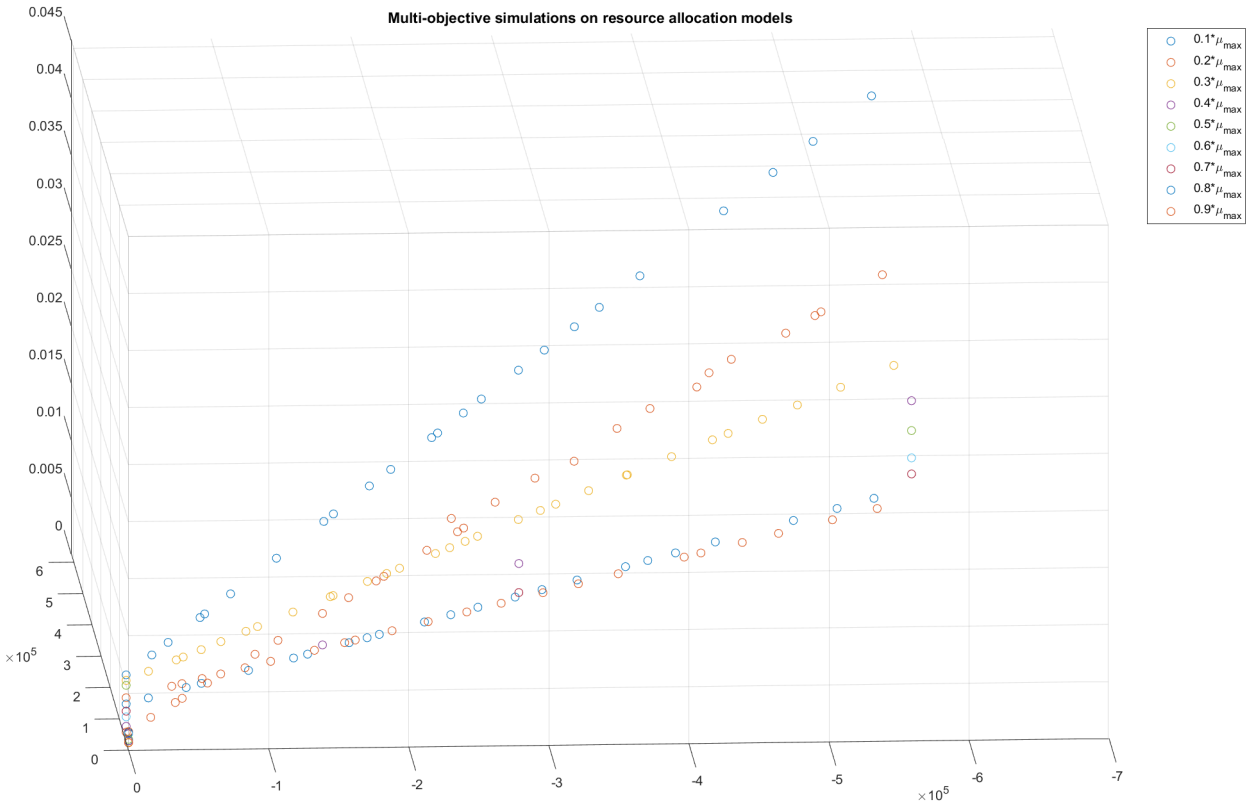

Figure 13: Solution space of simulated models, corresponding to distinct values of growth rate μ.

As shown in the figure above, the system has distinct feasible solutions. However there is an absence of a traditional Pareto front, as what can be seen is that for a fixed μ, the solutions appear to be linear combinations of extreme cases. No interesting tradeoff between objectives can be observed, and thus, not much information about gene expression can be inferred.

4 Conclusions

This work focused on studying how a biological system at genetic level behaves under the conditions of non thermodynamic equilibrium under both local and global scopes. Regarding the local approach by means of RNA-RNA interaction mediated gene expression regulation, a significant step forward in our basic understanding of this process has been reached. A novel scaling law for expression prediction from computational energetic and structural calculations is derived by following a steady state's non-equilibrium thermodynamic scheme. This law is dependent on different free energies, not only the free energy of formation, as a result of RNA degradation. The more efficient this enzymatic process (as in the case of bacteria), the further from a scenario of equilibrium we are. ΔG_{inter} will be a suitable predictor of regulatory activity when the final complex is not stable enough, as degradation will not be the dominant process at that moment, whilst ΔG_{trans} will govern the regulation when the global reaction is very favorable. Importantly, our model can be exploited to study natural RNA-based regulation or even to guide the design of synthetic systems [Rodrigo et al., 2017, Green et al., 2014].

This is developed for the IS10 system, in which the sRNA acts as a repressor of translation, but it may be adapted straightforwardly to deal with an activator of a cis-repressed mRNA. Of note, our results contribute to appreciate the importance of the seed region, disclosing from ab initio calculations its impact on regulatory activity; an effect only captured empirically until now [Mutalik et al., 2012]. In addition, our results reveal a trade-off in gene regulatory systems based on RNA-RNA interactions. On the one hand, the sRNA needs to be structured to enhance its stability in a cellular context, avoiding a premature degradation by the action of ribonucleases [Saramago et al., 2014]. On the other hand, that structure contributes to the difficult branch migration process by imposing an effective free energy barrier [Srinivas et al., 2013]. In sum, as we realize that enzymatic activity manifests into breaking detailed balance over some biochemical processes, we expect a wider application of non-equilibrium thermodynamics to predictably map genotype and phenotype, especially if this leads to amenable mathematical expressions connecting microscopic descriptors of the molecular world with mesoscopic parameters [Gnesotto et al., 2018].

Regarding the global approach, by means of the multi-objective optimisation of resource allocation metabolic models, not much can be inferred about gene expression regulation in a non-equilibrium scope. The linear combinations of solutions yielded by [Reynoso-Meza et al., 2012]'s algorithm on [Goelzer et al., 2011] metabolic model approach suggests that the system was much too rigid and had little freedom to allocate proteins in a way that biological conclusions could be drawn. We hypothesize that the rigidness was inherited from the nature of the restrictions imposed by [Goelzer et al., 2011]. Given that scenario, it seems that equivalent results could be obtained by addressing the multi-objective problem through the an *Aggregate Objective Funciton* approach. Therefore, by formulating an adequate aggregate objective function, a simple FBA would have sufficed.

Other works, such as [Siurana, 2017]'s adaptation of [Reynoso-Meza et al., 2012] algorithm for *Synechocystis* adaptability to different growth media have had very satisfactory results by adding restrictions on biochemical reactions. Following a parellel line, we worked to conduct such a work with adding protein allocation restrinctions to a *Escherichia coli* metabolic model. We managed to replicate results that can originate from a FBA approach. We are

currently formulating a new set of restrictions based on [Mori et al., 2016]'s approach towards of protein sectors and once we have the awaited experimental data, the possible benefits of [Reynoso-Meza et al., 2012]'s multicriterium algorithm against FBA can be evaluated.

References

[Adkins, 1997] Adkins, C. J. (1997). *Equilibrium thermodynamics*. Cambridge Univ. Press.

[Alberts, 2017] Alberts, B. (2017). *Molecular Biology of the Cell*. Garland Science.

[Archer et al., 2016] Archer, T. C., Fertig, E. J., Gosline, S. J., Hafner, M., Hughes, S. K., Joughin, B. A., Meyer, A. S., Piccolo, S. R., and Shajahan-Haq, A. N. (2016). Systems approaches to cancer biology. *Cancer Research*, 76(23):6774–6777.

[Bintu et al., 2005] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124.

[Brewster et al., 2014] Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323.

[Bulović et al., 2019] Bulović, A., Fischer, S., Dinh, M., Golib, F., Liebermeister, W., Poirier, C., Tournier, L., Klipp, E., Fromion, V., Goelzer, A., and et al. (2019). Automated generation of bacterial resource allocation models. *Metabolic Engineering*, 55:12–22.

[Buzzi and Llibre, 2014] Buzzi, C. A. and Llibre, J. (2014). Hopf bifurcation in the full repressilator equations. *Mathematical Methods in the Applied Sciences*, 38(7):1428–1436.

[Chen and Dill, 2000] Chen, S.-J. and Dill, K. A. (2000). Rna folding energy landscapes. *Proceedings of the National Academy of Sciences*, 97(2):646–651.

[Cherubini et al., 2017] Cherubini, C., Filippi, S., and Loppini, A. (2017). Systems biology modeling of nonlinear cancer dynamics. *Methods in Molecular Biology Systems Biology*, pages 203–213.

[Dahan et al., 2011] Dahan, O., Gingold, H., and Pilpel, Y. (2011). Regulatory mechanisms and networks couple the different phases of gene expression. *Trends in Genetics*, 27(8):316–322.

[Elowitz and Leibler, 2000] Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338.

[Faundez et al., 2019] Faundez, V., Wynne, M., Crocker, A., and Tarquinio, D. (2019). Molecular systems biology of neurodevelopmental disorders, rett syndrome as an archetype. *Frontiers in Integrative Neuroscience*, 13.

[Franch et al., 1999] Franch, T., Petersen, M., Wagner, E. H., Jacobsen, J. P., and Gerdes, K. (1999). Antisense rna regulation in prokaryotes: rapid rna/rna interaction facilitated by a general u-turn loop structure. *Journal of Molecular Biology*, 294(5):1115–1125.

[Fuente et al., 2020] Fuente, D., Garibo-i Orts, Ó., Conejero, J., and Urchueguía, J. (2020). Rational design of a genetic fnite state machine: Combining biology, engineering, and mathematics for bio-computer research. *Mathematics*, 8(8):1362.

[Gardner et al., 2000] Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342.

[Germain, 2017] Germain, R. N. (2017). Will systems biology deliver its promise and contribute to the development of new or improved vaccines? *Cold Spring Harbor Perspectives in Biology*, 10(8).

[Gnesotto et al., 2018] Gnesotto, F. S., Mura, F., Gladrow, J., and Broedersz, C. P. (2018). Broken detailed balance and non-equilibrium dynamics in living systems: a review. *Reports on Progress in Physics*, 81(6):066601.

[Goelzer et al., 2011] Goelzer, A., Fromion, V., and Scorletti, G. (2011). Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6):1210–1218.

[Gorski et al., 2017] Gorski, S. A., Vogel, J., and Doudna, J. A. (2017). Rna-based recognition and targeting: sowing the seeds of specificity. *Nature Reviews Molecular Cell Biology*, 18(4):215–228.

[Green et al., 2014] Green, A. A., Silver, P. A., Collins, J. J., and Yin, P. (2014). Toehold switches: De-novo-designed regulators of gene expression. *Cell*, 159(4):925–939.

[Groot et al., 2019] Groot, D. H. D., Lischke, J., Muolo, R., Planqué, R., Bruggeman, F. J., and Teusink, B. (2019). The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints. *Cellular and Molecular Life Sciences*, 77(3):441–453.

[Guil and Esteller, 2015] Guil, S. and Esteller, M. (2015). Rna–rna interactions in gene regulation: the coding and noncoding players. *Trends in Biochemical Sciences*, 40(5):248–256.

[Hofacker et al., 1994] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie Chemical / Monthly*, 125(2):167–188.

[Huang et al., 2015] Huang, Z. H., Zhang, L. L., Cheng, S. Y., Zhang, J., and Xia, X. H. (2015). Back-analysis and parameter identification for deep excavation based on pareto multiobjective optimization. *Journal of Aerospace Engineering*, 28(6):A4014007.

[Ideker et al., 2001] Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372.

[Kitano, 2001] Kitano, H. (2001). *Foundations of systems biology*. MIT Press.

[Kitano, 2002] Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560):1662–1664.

[Kittle et al., 1989] Kittle, J., Simons, R., Lee, J., and Kleckner, N. (1989). Insertion sequence is10 anti-sense pairing initiates by an interaction between the 5' end of the target rna and a loop in the anti-sense rna. *Journal of Molecular Biology*, 210(3):561–572.

[Laidler and King, 1983] Laidler, K. J. and King, M. C. (1983). Development of transition-state theory. *The Journal of Physical Chemistry*, 87(15):2657–2664.

[Mathews et al., 1999] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of Molecular Biology*, 288(5):911–940.

[Miettinen, 1999] Miettinen, K. (1999). *Nonlinear multiobjective optimization*. Kluwer Academic.

[Mori et al., 2016] Mori, M., Hwa, T., Martin, O. C., Martino, A. D., and Marinari, E. (2016). Constrained allocation flux balance analysis. *PLOS Computational Biology*, 12(6).

[Müller et al., 2006] Müller, S., Hofbauer, J., Endler, L., Flamm, C., Widder, S., and Schuster, P. (2006). A generalized model of the repressilator. *Journal of Mathematical Biology*, 53(6):905–937.

[Mutalik et al., 2012] Mutalik, V. K., Qi, L., Guimaraes, J. C., Lucks, J. B., and Arkin, A. P. (2012). Rationally designed families of orthogonal rna regulators of translation. *Nature Chemical Biology*, 8(5):447–454.

[National Center for Biotechnology Information, 2020] National Center for Biotechnology Information (2020). Home - genome - ncbi. https://www.ncbi.nlm.nih.gov/genome/. Accessed: 01-08-2020.

[Nelson et al., 2017] Nelson, D. L., Cox, M. M., and Lehninger, A. L. (2017). *Lehninger principles of biochemistry*. W.H. Freeman and Company.

[Oh et al., 2018] Oh, S.-J., Choi, Y.-K., and Shin, O. S. (2018). Systems biology-based platforms to accelerate research of emerging infectious diseases. *Yonsei Medical Journal*, 59(2):176.

[Orth et al., 2010] Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248.

[Paillart et al., 2004] Paillart, J.-C., Shehu-Xhilaga, M., Marquet, R., and Mak, J. (2004). Dimerization of retroviral rna genomes: an inseparable pair. *Nature Reviews Microbiology*, 2(6):461–472.

[Perez-Carrasco et al., 2018] Perez-Carrasco, R., Barnes, C. P., Schaerli, Y., Isalan, M., Briscoe, J., and Page, K. M. (2018). Combining a toggle switch and a repressilator within the ac-dc circuit generates distinct dynamical behaviors. *Cell Systems*, 6(4).

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

[Phillips, 2015] Phillips, R. (2015). Napoleon is in equilibrium. *Annual Review of Condensed Matter Physics*, 6(1):85–111.

[Price et al., 2004] Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897.

[Reynoso-Meza et al., 2012] Reynoso-Meza, G., Sanchis, J., Blasco, X., and Herrero, J. M. (2012). Multiobjective evolutionary algortihms for multivariable pi controller tuning. *Expert Systems with Applications*, 39(9):7895–7907.

[Rodrigo et al., 2013] Rodrigo, G., Landrain, T. E., Majer, E., Daròs, J.-A., and Jaramillo, A. (2013). Full design automation of multi-state rna devices to program gene expression using energy-based optimization. *PLoS Computational Biology*, 9(8).

[Rodrigo et al., 2017] Rodrigo, G., Prakash, S., Shen, S., Majer, E., Daròs, J.-A., and Jaramillo, A. (2017). Model-based design of rna hybridization networks implemented in living cells. *Nucleic Acids Research*, 45(16):9797–9808.

[Salim et al., 2012] Salim, N., Lamichhane, R., Zhao, R., Banerjee, T., Philip, J., Rueda, D., and Feig, A. L. (2012). Thermodynamic and kinetic analysis of an rna kissing interaction and its resolution into an extended duplex. *Biophysical Journal*, 102(5):1097–1107.

[Saramago et al., 2014] Saramago, M., Bárria, C., Santos, R. F. D., Silva, I. J., Pobre, V., Domingues, S., Andrade, J. M., Viegas, S. C., and Arraiano, C. M. (2014). The role of rnases in the regulation of small rnas. *Current Opinion in Microbiology*, 18:105–115.

[Schlick, 2018] Schlick, T. (2018). Adventures with rna graphs. *Methods*, 143:16–33.

[Schoen et al., 2009] Schoen, I., Krammer, H., and Braun, D. (2009). Hybridization kinetics is different inside cells. *Proceedings of the National Academy of Sciences*, 106(51):21649–21654.

[Schwanhäusser et al., 2011] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

[Seeman, 1986] Seeman, J. I. (1986). The curtin-hammett principle and the winstein-holness equation: new definition and recent extensions to classical concepts. *Journal of Chemical Education*, 63(1):42.

[Shine and Dalgarno, 1975] Shine, J. and Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature*, 254(5495):34–38.

[Siurana, 2017] Siurana, M. P. (2017). Modelling and multiobjective optimization for simulation of cyanobacterial metabolism. *[Unpublished Doctoral Thesis] Universitat Politècnica de València.*

[Srinivas et al., 2013] Srinivas, N., Ouldridge, T. E., Šulc, P., Schaeffer, J. M., Yurke, B., Louis, A. A., Doye, J. P. K., and Winfree, E. (2013). On the biophysics and kinetics of toehold-mediated dna strand displacement. *Nucleic Acids Research*, 41(22):10641–10658.

[Varma and Palsson, 1994] Varma, A. and Palsson, B. O. (1994). Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12(10):994–998.

[Villoslada et al., 2009] Villoslada, P., Steinman, L., and Baranzini, S. E. (2009). Systems biology and its application to the understanding of neurological diseases. *Annals of Neurology*, 65(2):124–139.

[Waters and Storz, 2009] Waters, L. S. and Storz, G. (2009). Regulatory rnas in bacteria. *Cell*, 136(4):615–628.

[Xia et al., 1998] Xia, T., Santalucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of rna duplexes with watson-crick base pairs. *Biochemistry*, 37(42):14719–14735.

[Zhou, 2016] Zhou, X. (2016). Computational systems biology in cancer brain metastasis. *Frontiers in Bioscience*, 8(1):169–186.

[Zou et al., 2013] Zou, J., Zheng, M.-W., Li, G., and Su, Z.-G. (2013). Advanced systems biology methods in drug discovery and translational biomedicine. *BioMed Research International*, 2013:1–8.