Regulatory systems of relevance for bioproduction

Ana Bulovic

To cite this version:

Ana Bulovic. Regulatory systems of relevance for bioproduction. Life Sciences [q-bio]. humboldt university, berlin, Germany, 2023. English. NNT : . tel-04379726

HAL Id: tel-04379726

https://hal.inrae.fr/tel-04379726

Submitted on 8 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Regulatory systems of relevance for bioproduction

DISSERTATION
zur Erlangung des akademischen Grades
Doctor of Philosophy
(Ph.D.)

im Fach
Biophysik

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

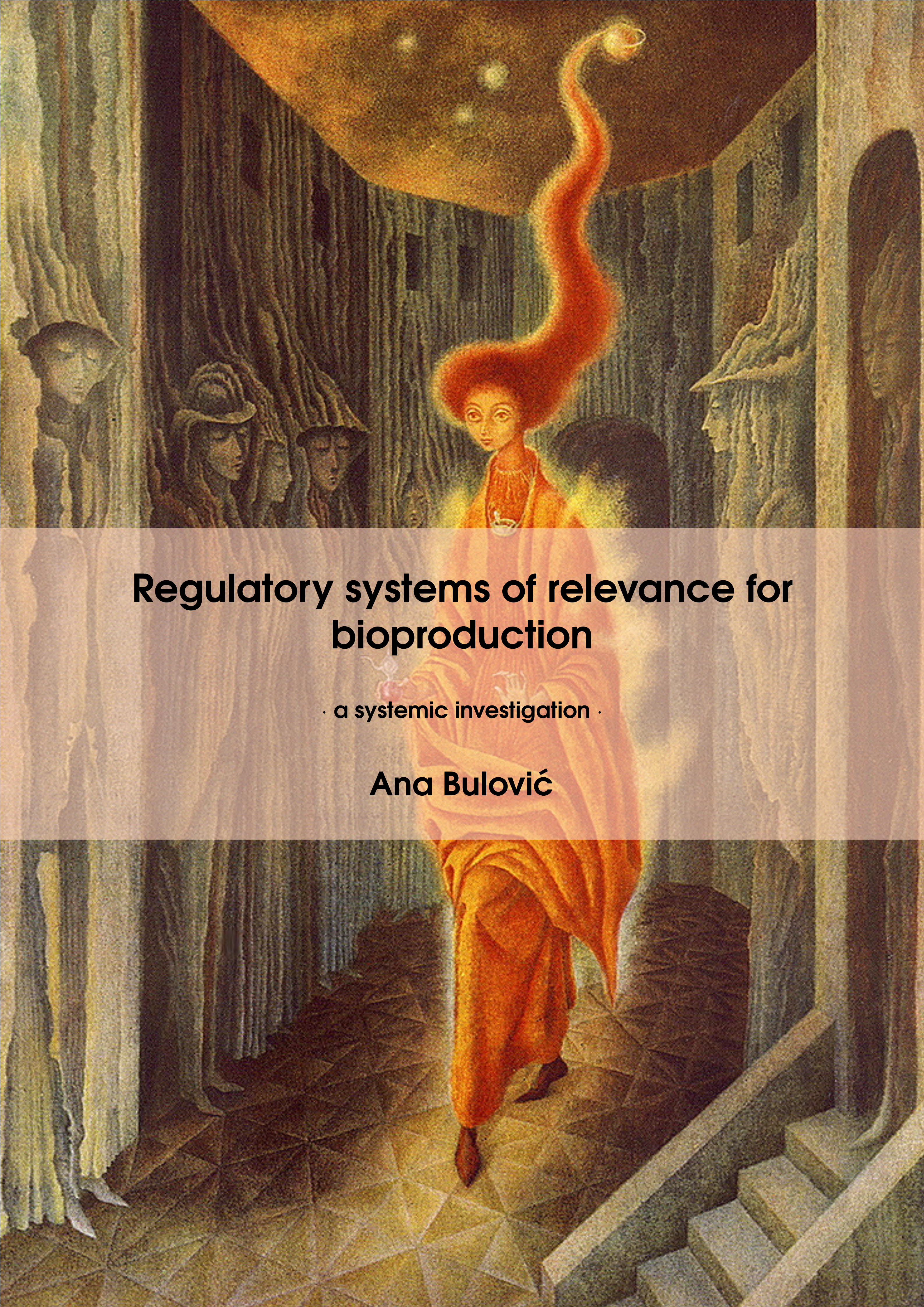
von
M.Sc. Ana Bulović
27.08.1989. Split, Kroatien.

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. Christian Ulrichs

Gutachter/innen:
1.
2.
3.

Tag der mündlichen Prüfung:



Regulatory systems of relevance for bioproduction

· a systemic investigation ·

Ana Bulović

But it is no longer a question of either maps or territories. Something has disappeared: the sovereign difference, between one and the other, that constituted the charm of abstraction. Because it is difference which constitutes the poetry of the map and the charm of the territory, the magic of the concept and the charm of the real.

Jean Baudrillard, Simulacra and Simulation (1981)

First of all, whatever we say is words, and what we want to talk about is generally not words. Second, whatever we *mean* by what we say is not what the thing actually *is*, though it may be similar. For the thing is always *more* than what we mean and is never exhausted by our concepts. And the thing is also *different* from what we mean, if only because no thought can be absolutely correct if extended indefinitely. The fact that the thing has qualities going beyond whatever we think and say about it is behind our notion of objective reality. Clearly, if reality were ever to cease to show new aspects that are not in our thought, then we could hardly say that it had an objective existence independent of us.

David Bohm and F. David Peat, Science, Order and Creativity (1987)

LATEXtemplate used for this work is **The Legrand Orange Book**, available under Creative Commongs BY-NC-SA 3.0 lincence.

Cover image: Remedios Varo, La Llamada (available for informational and educational purposes under fair use policy).

Abstract

Bacterial hosts, and *Escherichia coli* in particular, are used extensively for the production of industrial recombinant protein. The stress induced in the cells by this procedure is systemic - it introduces radical changes in the finely tuned system of mRNA and protein expression. Due to the complex and interwoven nature of the bacterial cell, it is no simple thing to understand the type and extent of these changes. This thesis deals with the problem of understanding and modeling such stress conditions, in which the entire cellular state is grossly affected.

I have attempted to tackle this problem in a number of ways. I first model and analyze the regulatory mechanisms involved in the cellular response to the stress provoked by recombinant protein expression, and show that, despite its apparent complexity, it has some unexpected and "simple" properties. Afterwards I shift the emphasis from regulation to cellular investment of resources. Since bioproduction is resource-wise very costly, it is reasonable to expect that many stress effects are due to the shifts in resource investment brought on by the genetic modification of the bacterium. For this purpose, I develop and calibrate a steady-state whole-cell model of *E. coli*. It is implemented in Resource Balance Analysis, a modeling framework able to realistically represent the cost of cellular events and account for a number of constraints under which cells operate - those of energy, efficiency and space - which lead to resource-related cellular decisions. This models shows good predictive power and because of its scope, level of detail and ease of manipulation, it can be used to assist experimental design in bioproduction. Lastly, I create a model whose purpose is to test whether the regulation of the bioproduction-induced stress responses can be explained by the tendency of the cell to implement resource strategies optimal for growth. For this purpose, I develop a simple time-resolved model of the heat shock response which takes into account the cellular constraints of energy, efficiency and space. I show that the obtained response to stress under the assumption of parsimonious resource allocation closely resembles one determined by experiment. The conclusions drawn from the three modeling approaches show that integrating the idea of resource allocation into cell models can help shed light on many regulatory events and adaptations taking place during bioproduction, and the tools developed in this thesis can help optimize the process of recombinant protein expression in *Escherichia coli*.

Zusammenfassung

Bakterielle Wirte, und insbesondere Escherichia coli, werden in großem Umfang für die Produktion industrieller rekombinanter Proteine verwendet. Der durch dieses Verfahren in den Zellen induzierte Stress ist systemisch - er führt zu radikalen Veränderungen in dem fein abgestimmten System der mRNA- und Proteinexpression. Aufgrund der komplexen und verwobenen Natur der Bakterienzelle ist es nicht einfach, Art und Ausmaß dieser Veränderungen zu verstehen. Diese Arbeit beschäftigt sich mit dem Problem des Verständnisses und der Modellierung solcher Stressbedingungen, bei denen der gesamte zelluläre Zustand grob beeinträchtigt wird.

Ich habe versucht, dieses Problem auf verschiedene Weise anzugehen. Zunächst modelliere und analysiere ich die Regulationsmechanismen, die an der zellulären Antwort auf den durch rekombinante Proteinexpression provozierten Stress beteiligt sind, und zeige, dass sie trotz ihrer scheinbaren Komplexität einige unerwartete und "einfache" Eigenschaften hat. Danach verlagere ich den Schwerpunkt von der Regulation auf die zelluläre Investition von Ressourcen. Da die Bioproduktion ressourcenmäßig sehr kostspielig ist, liegt die Vermutung nahe, dass viele Stresseffekte auf die Verschiebungen in der Ressourceninvestition zurückzuführen sind, die durch die genetische Modifikation des Bakteriums hervorgerufen werden. Zu diesem Zweck entwickle und kalibriere ich ein Steady-State-Ganzzellmodell von *Escherichia coli*. Es ist in Resource Balance Analysis implementiert, einem Modellierungsrahmen, der in der Lage ist, die Kosten von zellulären Ereignissen realistisch darzustellen und eine Reihe von Einschränkungen zu berücksichtigen, unter denen Zellen operieren, die der Energie, der Effizienz und des Platzes, die zu ressourcenbezogenen zellulären Entscheidungen führen. Dieses Modell zeigt eine gute Vorhersagekraft und kann aufgrund seines Umfangs, seines Detaillierungsgrads und seiner einfachen Manipulierbarkeit zur Unterstützung der Versuchsplanung in der Bioproduktion verwendet werden. Schließlich erstelle ich ein Modell, dessen Zweck es ist, zu testen, ob die Regulierung der Bioproduktions-induzierten Stressreaktionen durch die Tendenz der Zelle, für das Wachstum optimale Ressourcenstrategien zu implementieren, erklärt werden kann. Zu diesem Zweck entwickle ich ein einfaches zeitaufgelöstes Modell der Hitzeschockreaktion, das die zellulären Beschränkungen von Energie, Effizienz und Raum berücksichtigt. Ich zeige, dass die erhaltene Reaktion auf Stress unter der Annahme einer sparsamen Ressourcenallokation der experimentell ermittelten Reaktion sehr ähnlich ist. Die Schlussfolgerungen aus den drei Modellierungsansätzen zeigen, dass die Integration der Idee der Ressourcenallokation in Zellmodelle helfen kann, Licht in viele regulatorische Ereignisse und Anpassungen zu bringen, die während der Bioproduktion stattfinden, und die in dieser Arbeit entwickelten Werkzeuge können helfen, den Prozess der rekombinanten Proteinexpression in *Escherichia coli* zu optimieren.

Contents

Abstract 5

Zusammenfassung 7

1 Introduction 13

1.1 *Escherichia coli* in science and industry 14

1.1.1 Resources required for bacterial growth 15

1.1.2 Growth of bacterial cultures 16

1.2 Stress effects of protein overexpression 17

1.3 Control mechanisms in bacteria 19

1.4 Modeling in systems biology 20

1.4.1 Dynamical modeling 20

1.4.2 Constraint-based modeling 21

1.4.3 Dynamical modeling with constraints 22

1.5 A coherent cell state 23

1.6 Experimental techniques 24

1.6.1 Growth curve measurements 24

1.6.2 Proteomics 24

1.6.3 Total amino acid concentration measurements 25

1.6.4 Fluxomics 26

1.7 Layout of this thesis 26

I Proteome quality control during bioproduction

2 Recombinant protein production in *E. coli* 33

2.1 How proteins assume and maintain functional states 33

2.1.1 Protein folding 34

2.1.2 Chaperoning systems in *E. coli* 36

2.1.3 Proteolysis and protein degradation 39

2.2 Production of recombinant protein in *E. coli* 40

2.2.1 RP production induced stress 40

2.2.2 Process optimization 41

2.3 Unfolded protein and heat shock response 42

2.3.1 σ^{32}-mediated unfolded protein response 42

3 Model of the unfolded protein stress response 47
3.1 Mathematical preliminaries 48
3.1.1 Mathematical modeling of chemical reactions 48
3.1.2 Equilibria of dynamical systems 50
3.2 Dynamical model of HSR 53
3.3 Model simplification 56
3.4 Existence of equilibria 57
3.5 System behavior at the equilibrium point 61
3.6 Discussion 62
3.6.1 New understanding of heat shock regulation 62
3.6.2 Growth condition 63
3.6.3 ODEs, equilibrium points and biological systems 63
3.6.4 Modeling protein production 64

II whole-cell model of *Escherichia coli*

4 Resource Balance Analysis 71
4.1 Conceptual ideas behind RBA 71
4.1.1 Resource allocation and the growth rate 72
4.1.2 Concepts in RBA setting 73
4.2 Formulation of the RBA problem 74
4.2.1 RBA constraints 75
4.2.2 The linear programming problem 78
4.2.3 Enzyme and process machine efficiencies 79
4.3 RBA exemplified on a *toy* model 80
4.3.1 Model simulation 84
4.4 Related modeling paradigms 86
4.4.1 FBA with molecular crowding 86
4.4.2 MOMENT 87
4.4.3 ME-models 88
4.5 RBApy software 88
4.5.1 Model format 90
5 An RBA model of *E. coli* 93
5.1 Parameterization of the *E. coli* RBA model 95
5.1.1 Percentage of protein per compartment 95
5.1.2 Percentage of *housekeeping* protein per compartment 97
5.1.3 Cytosolic density 97
5.1.4 Total amino acid concentration 99
5.1.5 Default apparent catalytic rate of enzymes 99
5.1.6 Enzyme specific catalytic rates 99
5.1.7 Efficiencies of process machines 102

5.2 Results 103
5.2.1 Growth rate prediction . . . 103
5.2.2 Predictions for growth on glucose . . . 103
5.2.3 The case of dehydrogenase substitution . . . 104
5.2.4 Simulating the engineered CO_2-fixing *E. coli* strain . . . 106
5.3 Discussion 107
5.3.1 Gratuitous protein production in RBA . . . 107
5.3.2 Inferring regulation by exploring the RBA model . . . 108

III Dynamics under constraints

6 Dynamics under constraints . . . 111
6.1 The optimal control problem 111
6.2 Model of adaptation to change in temperature 113
6.2.1 Compact model representation . . . 121
6.3 Model parameterization 124
6.3.1 Kinetic rate of enzymes . . . 124
6.3.2 Ribosome translation rate . . . 125
6.3.3 Folding and unfolding rate . . . 126
6.3.4 Parameters without estimates . . . 126
6.4 Model simulation 126
6.4.1 Receding horizon control . . . 128
6.4.2 Numerical simulation . . . 129
6.4.3 Basic model functionality . . . 129
6.4.4 Adaptation to change in temperature . . . 130
6.4.5 Comparison to the whole-cell RBA model . . . 131
6.5 Different model versions 132
6.6 Discussion 134
6.6.1 Concentrations or molecule counts . . . 134
6.6.2 Parameter sensitivity analysis . . . 134
6.6.3 "Overshoot" in protein expression . . . 134
6.6.4 Comparison to other simple cell models . . . 135
6.6.5 Possible model uses . . . 135

IV Conclusions and outlook

7 Conclusions . . . 139
7.1 Tools and practices in systems biology 139
7.2 Elegance and utility of the parsimonious resource allocation paradigm 140

8 Outlook 143
8.1 Dynamical whole-cell models 143
8.2 Exploring regulation through parsimonious resource allocation paradigm 143
8.3 Towards predictive whole-cell models 144
8.4 Understanding temperature effect on cells 144

9 Further thoughts 145
9.1 On modeling in biology 145
9.2 A personal note 145

V .

Acronyms 149

Bibliography 165

VI .

A Appendix 169
A.1 Miscellaneous parameter values 170
A.2 σ^{32} regulon 171
A.3 Full matrix of the toy RBA model 175
A.4 Percentages of individual nucleic and amino acid in *E. coli* 176
A.5 Flux distribution (growth on glucose, no PQQ) 177
A.6 Flux distribution (growth on glucose, with PQQ) 178
A.7 Proteomaps of growth on glucose (experiment - prediction comparison) 179
A.8 Redistribution of resources for the simulated CO_2 engineered strain 180
A.9 Optimal control model steady state for a range of temperatures 181
A.10 JSON Bocop-compatible format 182

Acknowledgments 187

Declaration of authorship 189

Final throughs 191

1. Introduction

> Anyone who wants to analyze the properties of matter in a real problem might want to start by writing down the fundamental equations and then try to solve them mathematically. Although there are people who try to use such an approach, these people are the failures in this field; the real successes come to those who start from a physical point of view, people who have a rough idea where they are going and then begin by making the right kind of approximations, knowing what is big and what is small in a given complicated situation. These problems are so complicated that even an elementary understanding, although inaccurate and incomplete, is worth while having, and so the subject will be one that we shall go over and over again, each time with more and more accuracy, as we go through our course in physics.
>
> Richard Feynman, Lectures on Physics, Chapter 39

This thesis belongs to the expanding field of systems biology [1] and aims to study certain aspects of living beings by finding and analyzing suitable mathematical representations, a praxis that is commonly referred to in the field simply as *modeling*. My original interest in the topic came from a somewhat naïve belief in the power of representation combined with an amazement that an organism as complicated as a human being could function reliably, even for an instant. Because of their extreme complexity[1], the systems of interest in systems biology can be represented in innumerable ways, each of which highlights certain aspects and ignores others. This plethora of choice which the modeler always faces makes modeling an art of sorts.

The main research question of this thesis is how to appropriately model and understand the systemic stress responses in bacterial cells, especially under conditions relevant for bioproduction of recombinant protein (RP). RPs inherit their name from *recombinant* DNA, which is DNA produced by merging genetic material of more than one species. Recombinant DNA can be *designed* in a laboratory so as to encourage the production of foreign protein once introduced into a host organism[2]. As it interferes with one of the central processes in all cells (that of protein production), the expression of RP can have diverse effects on the cell. Some of these effects are protein specific and are caused by the biochemical properties of the RP which introduce a certain disbalance in the cell either via toxicity or via shifts of its metabolic or compositional balance. However, some of the effects are more general and get provoked with a large set of recombinantly expressed proteins. These are: (i) the change in the space available for functional cellular components, (ii) the change in the distribution of energy, precursors, and other resources among cellular processes, and (iii) the changes in maintenance requirements of the cellular proteome. All these have a profound influence on the entire cellular state. In simpler organisms,

---

[1] The complexity of biological organisms is encountered on multiple levels: in terms of physical aspects relevant for their understanding, in terms of their complex structure composed of a vast number of different chemical entities, and especially in terms of the coordinated interaction of those parts which ensures their smooth functioning.

[2] The process of expressing a foreign protein in a host organism is often difficult and not necessarily always successful.

such as bacteria or yeast, this can be easily noticed in the reduction of their *growth rate* - the rate of conversion of foodstuffs to new cells.

Bacteria, due to their ancient origins, relatively small size[3] and fast reproduction have had a lot of opportunity for improvement. Faced with the condition of limited and irregularly available resources, energetic efficiency becomes one very desirable trait. Therefore, it is reasonable to assume that most organisms present today, through competition and *survival of the fittest*, have become as energetically efficient as possible and necessary for the ecological niches which they occupy [4]. The logical extension of this thought is to assume that much of the regulation present in these organisms is there to ensure this efficiency. When we introduce processes into the cell which tamper with this efficiency and its implementation through regulation, we introduce changes that greatly affect the state of the cell.

In this introductory chapter, I will first give a short description of the organism of interest - Gram-negative bacterium *Escherichia coli*, and will explain some concepts regarding the growth of bacterial cultures and bacterial metabolism, followed by the effects of the stress caused by overexpression of gratuitous protein in bacterial hosts. Afterwards, I will give a description of what is assumed under a *coherent cell state*. I conclude with a short description of mathematical modeling approaches of relevance, after which I describe the experimental techniques whose results were used in this work. The last section contains a layout of this thesis.

The three parts of the thesis that follow address the issues of (1) structural understanding of the regulatory response activated by overexpression of protein, (2) representation of the coherent cellular state on the level of the whole cell of *E. coli* and (3) application of the coherent cellular state to the study of dynamic cellular response to heat shock.

Please note I have tried to make this thesis self-contained, and have therefore included some information which might seem banal to the informed reader. I made this decision because of my personal preference to study from books, and specifically from self-contained ones. Moreover, I have often found that the devil lies in the banal: whenever knowledge is presented as obvious, it can suppress the capacity to question it and lead to overlooking some very important assumptions inherent in the explanation. I hope I have succeeded without making the text too burdensome for those "in the know". Good reading!

1.1 *Escherichia coli* in science and industry

Genetically modified versions of *E. coli* are used in scientific and industrial laboratories throughout the world [5]. This bacterium is one of the main workhorses of the pharmacological industry, producing different metabolites and proteins of interest. The reasons for this were initially its fast growth and ease with which it grows in chemically defined media, as well as the simplicity of genetic modifications. Today, its usefulness is augmented by the abundance of knowledge accumulated over the years and the versatility of tools that enable its manipulation. Many dedicated databases summarize the metabolic, genetic, proteomic and regulatory information gathered thus far on this organism [6].

Its structure is that typical of Gram-negative bacteria. Its cytosol, which houses its DNA and most of its protein, is shielded from the environment by two membranes - the inner plasma membrane and the outer membrane (made out of lipopolysaccharide and protein) between which is a layer of peptidoglycan. The entire space between the plasma membrane the outer membrane is called periplasm. Many RPs are expressed in the cytosol. However, as the cytosol of *E. coli* is crowded

---

[3] Small size is meant here as a comparison to the minimum size which guarantees the necessary stability of a self-replicating unit [2], such as the proposed RNA world protocell [3].

with macromolecules [7], the purification of the RPs can be difficult. In the periplasm the density of macromolecules is much lower, rendering it an attractive place for the expression of RPs. Protein purification from the periplasm is also much more efficient [8].

Under all growth conditions, *E. coli* cells precisely monitor the energetic richness (in terms of foodstuffs) and the chemical composition of their surrounding medium and adapt to it. The capacity of the medium to support growth will influence the internal organization of the cell. This adaptation will often be observable in the change of the growth rate of the cells. In this thesis I study how the introduction of a gratuitous RP influences the cellular protein production capacity, which is one of the most costly cellular processes. Therefore, it is interesting and relevant for this thesis to consider which cellular processes depend most heavily on the availability of resources or - to put it in different words - which cost most resources to maintain.

1.1.1 Resources required for bacterial growth

The Christian notion of the Garden of Eden is that of a place of abundance of nature, where the necessities of man are modest in comparison. Even if the relative abundance in which we live today makes us somewhat blind to it, the myth carries an important message: energetic needs are very important, not trivial to satisfy, and guide a great part of the development of all species, making abundance seem close to the idea of ultimate fulfilment. It is thus logical when considering the organization of an organism[4] to identify the processes that require most resource investment to be maintained.

First, I would like to explain what I mean by cellular investment of resources (for an attempt at a systemic definition which goes along the lines of what I propose here see [9]). A cell can loosely (and somewhat poorly) be defined as a physical implementation of a set of interdependent functions coordinated by the needs of survival. All of these functions are implemented as complex series of chemical reactions. In general, this set of reactions requires three types of resources: (*i*) an implementational structure which facilitates the reactions, (*ii*) input of energy and (*iii*) chemical reaction substrates. Most cellular functions rely on an implementational structure composed of protein and RNA. The macromolecular complexes which aid in the implementation of metabolic functions are called enzymes and comprise a great part of the cellular protein content. Other functions implemented by protein and RNA are, for example, the duplication of DNA, transcription into mRNA, and protein production. Energy necessary to fuel all of these processes is most often available through high-energy bonds of certain metabolites, such as adenosine triphosphate (ATP) and guanosine triphosphate (GTP). The building blocks for cellular construction are either directly taken up from the growth medium or synthesized through metabolic modification of other substrates.

Under relatively stable environmental conditions, cellular growth will mean the faithful duplication of all cellular components, most of which are macromolecules. The main macromolecular constituents of the cell are: protein, RNA, DNA, lipids, and carbohydrates. Because of their often complicated and to a certain degree sensitive structures, macromolecules can take up forms which prevent them from fulfilling their biological function. Therefore, apart from the creation of new material, the cell constantly needs to invest resources into maintenance of its existing structures. All of the classes of macromolecules in cells require production, modification and maintenance[5], and the cost of their existence in the cell can be quantified through the cost of these three categories. The measurements of macromolecular structure of an *E. coli* cell [10] show that of all

---

[4] The notions of organization and organism are closely related: organism, from Greek ὀργανισμός (*organismos*), is derived from the word ὄργανον (*organon*), also being the origin of the word organization.

[5] Maintenance can be assumed to comprise degradation as its final step.

the classes of macromolecules, proteins are by far the most abundant, and account for 50% of the dry cell weight [11]. This is because almost all of the processes of production, modification and maintenance of all classes of macromolecules are, at least in part, performed by proteins in their enzymatic form. Their production requires ribosomes, very large and complex structures composed of protein and RNA, amino acids and a high amount of energy. Many proteins need further assistance to assume their functional form, in form of folding, re-folding, and post-translational modification. Chaperones and proteases, which provide this kind of assistance, are large protein complexes and often extremely costly.

Of all the other macromolecules, RNAs are the most abundant. In cells, RNAs serve as ribosome constituents (rRNAs), as transfer RNAs (tRNAs), as messenger RNAs (mRNAs) and as small RNAs (sRNAs). As almost all of these functions are directly related to protein production, their synthesis cost can be assumed to be part of protein production cost. DNA in *E. coli* comprises only about 3% of total dry cell weight. Even if its maintenance and duplication is a complex process, it can be considered minor in terms of cellular resource investment compared to protein production. Apart from metabolites present in the cytosol, the rest of the dry cell weight is composed of the constituents of cell membranes: lipids, lipopolysaccharides, peptidoglycan, etc. However, protein investment in the production of these components is only a minor part of the total proteome [11].

The conclusion we can draw from this short survey is that protein production and maintenance are what uses up most of the cellular resources. This was important to emphasize in the light of the topic of this thesis, which is the production of RPs. Since it is so costly, it is to be expected that this process will be under tight regulatory control and that all attempts at interfering with it will have extensive consequences on the cellular state. Next, I shortly describe how bacterial cultures are grown in a laboratory and give a rough outline of experiments for expression of RPs.

1.1.2 Growth of bacterial cultures

The late 19^{th} century saw the rise of the development of experimental techniques for growth of microorganisms in laboratory environments. In that period the first chemical (and minimal) medium was defined [12] and first enrichment cultures got cultivated, in which the conditions for growth of a specific microorganism were optimized [13]. Microbial growth in a laboratory is performed in a device called *bioreactor*. In the simplest case, a bioreactor is a vessel containing the growth-supporting liquid medium, usually shaken or stirred. If all the medium is provided at the beginning of an experiment, the bioreactor can be classified as a *batch* reactor. If the nutrients are supplied during culture growth, one talks of a *fed batch* reactor. In a *continuous* reactor, nutrients are constantly supplied and products are continuously taken up from the culture (with chemostat being the most common example). The work of this thesis assumes the growth of cultures in a batch reactor.

Monitoring the growth of microbial cultures in batch reactors in a laboratory often showed qualitatively very similar results. The curve displaying these results became known as the famous *growth curve*, which tracks the number of viable cells over time (see Figure 1.1.1). At first, when the cells are added to a fresh medium, they need to adapt to the new growth condition, during the so-called *lag phase*. Once the cells modify their internal configuration to better suit a new growth condition, they start growing faster. In this period, the death rate is negligible compared to the growth rate. This phase is known as the *exponential phase* or *log phase*, and the increase in the number of cells can be described as:

$$\frac{dN(t)}{dt} = \alpha(t)N(t) \tag{1.1.1}$$

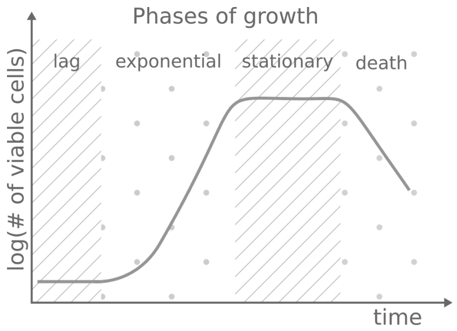


Figure 1.1.1: Stylistic representation of a typical "growth curve" (the number of viable cells over time) when microbial cultures are grown in a batch reactor, comprised of the adaptation dominated *lag* phase, *exponential* growth phase, followed by the *stationary* phase brought on by the depletion of nutrients or accumulation of toxic byproducts, finally ending in the *death* phase.

as was first introduced by Malthus [14] to describe the growth of populations. $\alpha(t)$ is known as the *growth rate* and in the exponential phase it can be assumed to be more or less constant $\alpha(t) = \alpha_{ep}$. After most of the foodstuffs has been consumed or after a toxic byproduct has accumulated in the medium to a sufficient degree, cells enter what is known as *stationary phase* [15]. The name of the phase implies that there is no change in the number of viable cells during this time. This does not specify what exactly happens to the cells, but in fact many important changes do happen: cells change their state from one compatible with growth to another one compatible with survival. They become smaller in size, different in shape, their membranes become more protective of external influence, and their internal composition more apt to shield them from different potential sources of stress. Finally, when the resources are completely depleted, the culture enters into the *death phase* (for a good overview, see [16]). In this phase, a large number of cells undergo *programmed cell death*. In this way, the bacterial population sacrifices many and provides the foodstuffs for the survival of a small part of the original population. If the experiment is left to run for an extended period of time, the death phase is followed by the so-called *long-term stationary phase*, in which there is a periodic increase and decrease in the number of viable cells, and which can last for a very long time [17].

A great community of scientists has been studying the growth of bacterial cultures (for a historical account see [18]). From being denounced as a field of study by one of its pioneers [6], it soon became a field in its own right, and quite a fervent one. Under the apparent simplicity of "growth curves", there lurk many questions regarding the regulation of synthesis of cellular material for a broad range of growth rates [20, 21]. The work of this thesis focuses on the *log* phase of bacterial growth.

1.2 Stress effects of protein overexpression

Bacteria are often used to overexpress protein of industrial or pharmacological use. The genetic code of these bacteria is altered to express the Protein of Interest (PoI). Sometimes further modifications are made to improve the capacity of a species to produce it. The optimization of a species for some industrial use is called *chassis design*. Such PoI is generally harvested from batch-grown cultures. The expression of PoI normally does not start from the beginning of the culture growth, but is generally induced at some later stage. The reason for this is that the

---

[6] "The study of the growth of bacterial cultures does not constitute a specialized subject or branch of research: it is the basic method of Microbiology.", Jacques Monod [19]

expression of a high yield gratuitous protein hinders cellular growth and would result in slow or even non-growing cultures.

One common type of problem in such experiments is that it leads to a significant perturbation of the energy balance in the cell causing cells to undergo a type of *starvation*. Another (also very common) kind of a problem is related to the way in which the PoI assumes its native, properly folded and functional state. All proteins require a specific biochemical environment to fold properly, but some proteins are more "troublesome" than others. They can require precise pH, presence of folding assistants (chaperones) and other modification mechanisms (phosphorylation, glycosylation). Some fold quickly, some very slowly, in iterative steps. Some proteins, which are produced in one, but need to assume their functional state in another compartment, often need to remain unfolded until they reach their destination. When such "demanding" proteins are expressed in high amounts, they can interfere with the native proteome's capacity to assume its functional state. It is precisely these kinds of problems that are of interest in this thesis, so I continue to describe them in more detail.

The native cellular proteome, energetically costly and abundant as it is, is under constant cellular supervision and quality control. In order to perform their functions, the enzymes of the cell need to maintain a specific shape, but with a certain degree of "elasticity" allowing them to undergo conformational changes. This shape is a result of properties of the original amino acid chain and of its interaction with the cellular environment through a notoriously complex process of folding. To achieve this functional shape, certain enzymes need assistance. Such assistance is provided by *molecular chaperones*, which can help the folding process in a number of ways: by *holding* the protein in a partially unfolded form and thus slowing down its folding, by providing an isolated environment more favorable for folding than the cytosol, or by disrupting parts of the secondary or tertiary structure which folds in a wrong way.

Holdases bind a protein and slow down further folding, thus giving the protein higher chances of folding in the right order. *Foldases* directly assist the folding of a protein. If the proteins are not in their functional shape, they can be either unfolded, misfolded, partially or completely folded. *Unfolded* proteins are those that still have not assumed their functional shape, but without any part of them being wrongly folded. *Misfolded* proteins are those that have gone through certain errors in folding which need to be corrected before the successful folding can continue. In unfolded and misfolded form, proteins have parts of their hydrophobic "insides" exposed. These hydrophobic patches bind equally well to other hydrophobic patches of the same protein as to those of other proteins. In the latter case, different proteins start binding each others hydrophobic patches and start forming *aggregates*, bound mixtures of misfolded proteins. These aggregates quickly become insoluble and present a great problem for the cell as they occupy space and are a pool of unusable resources. The cell tries to avoid their formation by refolding or degrading misfolded proteins. To degrade protein, the cell has enzymes that are able to cleave peptide bonds - *proteases*.

Chaperones and proteases are the main actors in the cellular response to the accumulation of unfolded protein - the so-called Unfolded Protein Response (UPR). Because of its ubiquity in RP expression experiments, the UPR will be the focus of this thesis, together with another stress response most related to it which happens outside of the laboratory: the Heat Shock Response (HSR). Their relation and their differences will be described in section 2.3.

1.3 Control mechanisms in bacteria

When a certain relevant condition of their environment changes (be it positive or negative), bacteria need to adapt in order to align their internal state in a best way possible to the environmental condition [7]. In order to achieve such adaptations, which can range from slight modifications of the metabolism to large-scale changes of the proteomic makeup, bacteria (and cells in general) have mechanisms in place which *regulate* their internal state. Because of their role in regulating the internal cellular state to achieve a certain goal (survival, reproduction, etc.), these mechanisms are called *regulatory mechanisms*. These can be incredibly diverse, and a big part of our progress in understanding cells over the last decades has been due to our ever increasing appreciation of the diversity of the possible ways in which cells regulate themselves. It would be difficult to pinpoint any cellular function which is not regulated on multiple levels and through a number of diverse mechanisms, as will become obvious in the next chapter, when describing the regulation of the UPR.

It is first instructional to ask what kinds of stimuli can an *E. coli* cell perceive. As a complex chemical system, it can react to the changes of its environment and internal state by sensing (among other things) (i) temperature, which influences the structure and stability of macromolecules and diffusion and binding of molecules in general, (ii) concentrations of ions and metabolites, the binding of which can modulate almost all cellular functions, and (iii) concentrations of macromolecules which serve to influence the functions of other macromolecular machines by modulating their transcription, translation or activity. The detection of such signals can result in a plethora of changes on all levels of cellular organization - they can influence transcription, translation, efficiencies of particular metabolic processes, degradation of cellular macromolecules etc. Such changes usually happen in a carefully coordinated regulatory cascade in order to bring about the necessary systemic changes in the cellular state, such as entrance into stationary state, change in the cell wall permeability, change in the the growth rate, etc.

This regulation is complex and layered, and here I would like to mention some of the mechanisms of control one has to have in mind when analyzing regulation of any process. I will focus on bacteria, and because of the great multitude of ways in which cells regulate their own states, I will list only a subset sufficient to illustrate the vastness of such mechanisms. These can be:

- structure of DNA [22],
- methylation of DNA
- proximity of genes on DNA
- organization of multiple genes into single transcription units - operons
- distance of promoter to gene
- affinity of promoter to RNA polymerase and transcription factors
- tertiary structure of mRNA as a regulator of translation
- utilization of rare molecules in composition of biological polymers which slow down and regulate production
- stability, conformational and functional changes of macromolecules induced by binding to single molecules, other macromolecules, or by changes in temperature
- concentration of different metabolites and ions in the cell.

Depending on which part of the cellular adaptation it affects, regulation can be transcriptional, post-transcriptional, translational or post-translational. Transcription is performed by a protein

---

[7] This statement can generally be considered to be true, if one refrains from specifying precisely what it means to align the internal state in the best way possible to the environmental condition. It could be said to mean a cellular state which has shown the highest probability of survival over time in similar conditions, but that would again be to say little to nothing.

complex called RNA polymerase, which binds the promoter gene regions. One type of transcriptional control relevant for this thesis is implemented by the so-called σ factors, proteins which bind and modify the affinity of the RNA polymerase DNA binding region to a specific set of promoters [23].

Each σ factor exerts its influence through an increased transcription initiation of a set of promoters, thus influencing the composition of the cellular mRNA and consequently, the protein pool. *E. coli* has seven σ factors. One of these factors, σ^{70}, aids the transcription of a great number of *E. coli* genes under conditions of normal growth and is thus considered the housekeeping σ factor, and is the product of the *rpoD* gene. The other six factors are preferentially used in specific stress situations *E. coli* encounters. These are:

- σ^{32} - the "heat shock" σ factor, regulating the cellular adaptation to increased temperatures (heat shock),
- σ^{24} - the extreme heat shock σ factor,
- σ^{38} - the σ factor regulating the entry into the stationary phase,
- σ^{28} - σ factor responsible for motility and flagellar synthesis,
- σ^{54} - σ factor active in regulation of nitrogen-related genes and in conditions of nitrogen limitation, and
- σ^{19} - involved in the regulation of transcription of ferric citrate transport genes.

1.4 Modeling in systems biology

Since the span of potential systems of interest is great in systems biology - going from small network motifs [24] all the way to entire cellular models [25], the span of mathematical modeling approaches is equally broad. Here I present the modeling approaches relevant for this thesis.

1.4.1 Dynamical modeling

As the knowledge of the internal workings of the living beings accumulated, it became clear that they are full of complex dynamical phenomena. Interest arose to understand how these dynamical systems can exhibit certain behaviors, while functioning reliably within a wider cellular context. The science of dynamic systems in physics and engineering has had a long and successful history in representation, analysis and synthesis. One of the most common formalisms for modeling dynamical systems are Ordinary Differential Equations (ODEs): equations which involve not only different functions of the variables, but also their instantaneous change over a certain independent variable. The independent variable in all our applications will be time and is designated here as t:

$$\frac{dx^i(t)}{dt} = f^i(x^1,...,x^n,u^1,...,u^m,t) \tag{1.4.1}$$

The usage of ODEs in modeling of biological phenomena has a long history since it became evident that cells are full of precisely controlled chemical reactions and that those can be described through the relations of substrate and product concentrations in the form of ODEs (see section 3.1 on chemical kinetics).

The way in which the instantaneous changes of variables depend on the variables will determine type and properties of a system. One such important property is whether this dependence is linear or not. If it is, it is possible to find closed-form expressions for all the variables as functions of the independent variable (time), usually as the sum of constants and time exponentials:

$$x^i = x^i(t) = C_0 + C_1 e^{k_1 t} + \ldots + C_n e^{k_n t} \tag{1.4.2}$$

where k_j can be a real or a complex number. It is also possible to characterize the solutions to such systems and determine whether they exhibit stable, unstable or oscillating behavior. If, however, the dependence on the set of variables is nonlinear, it is often impossible to find closed-form expressions, such as in Equation 1.4.2 and it can be difficult or impossible to determine the stability properties of the system. Such systems might not even have existing solutions, or the solutions might exist only within a limited period of time. The nonlinear systems generally have a richer set of possible behaviors than the linear ones, but unlike the linear ones, which can always be analyzed using the same set of techniques, nonlinear ODE systems require case-specific treatment, which often is mathematically advanced.

As it turns out, most systems apt for describing biological phenomena of interest are nonlinear. As it is often impossible to find the closed-form solutions of such systems, their behavior over time is usually approximated numerically. The basic idea in numerical solving of ODEs is to use the value of the tangent at one or more timepoints to approximate the future behavior of the system. There is a number of algorithms available for such approximations and their suitability depends on the characteristics of the particular application.

Even if it is possible to find an approximation of the behavior of a nonlinear system over time, there are certain things this approach cannot tell us. We cannot know that the behavior we see is "typical" as $t \to \infty$. We also cannot know if the observed behavior is stable for small changes in parameter values. Such things need to be established through detailed system-specific analysis.

1.4.2 Constraint-based modeling

What are possible solutions given a number of constraints? Which is the best one according to a certain criteria? These are the kinds of questions that one can tackle with constraint-based modeling approaches. Researchers working in production of metabolic compounds through bacterial fermentation were asking quite similar questions: given the restrictions imposed by the structure of the bacterial metabolism and necessities of growth, what is the maximum yield of a particular metabolite one can produce? In an early modeling paper [26], Papoutsakis gives a clear and concise description of the goals of modeling:

> "The establishment of thermodynamic and biochemical constraints which determine the theoretically highest yield for each product and the calculation of these maximal yields would be of both fundamental and practical importance. They would allow us to establish rationally the upper bounds for the productivity of the fermentations, which in turn can be used as a guide in feasibility studies, and experimentation for genetic and bioreactor-productivity improvements."

As the models got more refined and manual manipulation of equations became hard, methods based on flux balancing [27] offered an automatized solution to the problem. The method known today as Flux Balance Analysis (FBA) introduced the use of linear optimization to find the optimal flux distribution for a certain condition of growth, defined by imposing limits on a number of import or export fluxes. This modeling paradigm assumes the metabolism of a cell to be in steady state. What follows from that assumption is that the sum of all production and consumption fluxes for all metabolites must be zero. This can be written in matrix form as:

$$Sv = 0 \tag{1.4.3}$$

where S is the stoichiometric matrix and v is the vector of fluxes. Typically, flux through the so-called *biomass reaction* is optimized. This reaction commonly includes a great number of compounds in a stoichiometry representing their content in the cell. These stoichiometric coefficients are estimated from experiments which determine the chemical cellular composition (see [10] for an example).

Since within the FBA only the metabolism is considered in detail, while the rest of the cell is lumped in the biomass equation, the applicability of the method to study different phenomena is limited[8]. During the years, many tried circumventing this restriction by taking into account other important cellular constraints, such as limited space [28]. Recently, the cell has successfully been described as a set of linear constraints which take into account, among other things: metabolism, availability of enzymes, different cellular processes (such as protein production and folding) and limited space in all compartments [29]. This is the cellular formulation used in this thesis.

1.4.3 Dynamical modeling with constraints

If the cell is represented by a dynamical framework, it is possible to study its adaptive mechanisms over time. The constraint-based steady-state description of the cell allows for the study of optimal cellular configurations under given conditions, and for the analysis of the investment of resources in different cellular processes under the assumpton of parsimonious resource allocation. The combination of theset two approaches provides a basis for investigating the adaptation of the cell under the set of resource allocation constraints. Such an approach enables us to situate the dynamical system of interest into an appropriate cellular context.

One way in which the two approaches can be combined is through *optimal control*. Optimal control answers the question of what is the best way to control a dynamical system to achieve a certain goal under a set of linear equality and inequality constraints. The first question of this nature regarded a mechanical system. It came as a challenge from Johann Bernoulli to the "most brilliant mathematicians in the world":

> "Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time."

The field of application for optimal control today is very broad. Within its scope one can ask a question such as: What is the best way the cell should distribute its resources to adapt itself to a new condition under the assumption of growth rate maximization?

Let us assume that some portion of interest of the cell is described as a set of ODEs:

$$\frac{dx^i(t)}{dt} = f^i(x^1,...,x^n,u^1,...,u^m,t) \tag{1.4.4}$$

where x is a vector of system states, and u the vector of controls which act on the system. It is the goal of optimal control to determine the time course of the control vector u

$$u^j = u^j(t), j = 1,...,m \tag{1.4.5}$$

such that a certain criterion is maximized:

$$J = \phi(x^1(t_f),...,x^n(t_f),t_f) + \int_{t_0}^{t_f} L(x^1,...,x^n,u^1,...,u^m,t)dt \tag{1.4.6}$$

The controls in a biological system can be, for example, the rate of mRNA or protein production. The states of the system can be the concentrations of respective macromolecular species. Such an approach was taken and further explored in chapter 6.

---

[8] For example, all available resources are always used, which is not the case in real organisms.

1.5 A coherent cell state

What our pre-scientific ancestors lacked in technology, they amply made up in imagination. As the atomic theory traces its origins at least some 2300 years into the past [30], so the first idea of microscopic living beings greatly predates the microscope [31]. However, technological advances did bring out a new important aspect of living beings to the forefront - they are incredibly complex. A single bacterial cell is able to sense its chemical surroundings, move in the direction of better foodstuffs, use a very limited set of nutrients to produce the great metabolic variety that constitutes its cell, duplicate itself with great fidelity, produce injection like formations through which it will secrete toxins into another organism, exchange its genetic code in a modular fashion with other bacteria, regulate its pH, form tissue-like formations with other bacteria - the list of fascinating phenomena is indeed very long.

All of the phenomena mentioned can rightfully be studied and formalized within the domain of systems biology. How one chooses to formally describe a biological system will, of course, greatly depend on its nature. When studying the information processing capacities of a signaling pathway, the cellular context within which this system operates can well be ignored to a significant degree without seriously affecting the analysis and conclusions thereby obtained. However, when dealing with phenomena of systemic stress, in which one very important aspects of the cell is perturbed - such as its capacity to produce protein - then it is exactly the *state of the cell* that becomes the topic of modeling. The model of the cellular state will not always necessarily be the same - the biological condition of interest will guide the design of a relevant *cell state model*. However, to model the cellular state means to account for energetic and spatial requirements for achieving a particular cellular configuration. The set of states which best represent the limiting energetic and spatial requirements then become a matter of informed choice.

In this thesis, the *state of the cell* means the following: a mathematical formalization in which the cell is represented with a number of states across a range of granularity, but such that these states adhere to certain consistency restrictions. All of these restrictions can be viewed as particular instances of one general restriction: operational capacity encoded in the cellular configuration must be sufficient for the requirements imposed by that same configuration. An example might help to clarify this general idea: the production flux of a certain amino acid needed for protein synthesis puts a requirement on the minimal amount of enzyme in its synthesis pathway.

Even if the idea is quite intuitive, I think it is important to state it clearly. I have mentioned both cellular configuration and cellular operational capacity, i.e., the amount of enzyme and the reaction flux, or the amount of ribosome and the protein production flux. But how does protein production flux relate to the amount of protein? In a simplistic way, one can say that the protein production flux acts to either maintain or change the current protein make-up of the cell. When focusing on the first possibility - that of maintenance - it is clear that the operational capacity of the cell serves to maintain its configuration over *time*.

When looking at a bacterial cellular state in a coarse-grained manner, where states are lumped in broad categories such as ribosomes and metabolic enzymes, then a good rough measure of the cellular state is the rate of its growth [32][9]. For a specific growth rate, and on a certain level of modeling granularity, the cellular states are similar, regardless of the different conditions which brought about that growth rate. To illustrate: to maintain the same rate of growth in different conditions, the cell needs to maintain more or less the same flux of protein production, regardless of the precise proteome composition. This will imply a similar amount of ribosome, which will in turn imply that the level of RNA in both cultures will more or less be the same.

---

[9] Growth rate of bacteria is computed as $\ln(2)/T_d$, where T_d is the time it takes for a growing population of bacteria to double its size.

Under different growth conditions, the growth rate is a type of measure of how efficiently the foodstuffs can be converted into new cells. This approximation is grounded in the expectation that, due to competition for limited resources, organisms have become as energetically efficient as possible throughout the course of evolution.

1.6 Experimental techniques

The diversity, precision and ingenuity of experimental techniques available today which have allowed for measurement of a great number of cellular features are truly impressive. Here, I mention those relevant for this thesis.

1.6.1 Growth curve measurements

As mentioned in subsection 1.1.2, the so-called growth curve describes the number of viable cells over time with its y-axis normally represented in log scale. The most common way to determine the growth of the culture is to relate its density to an optical measurement, called the Optical Density (OD). OD measurements are performed through the use of spectrophotometer, in which a light of a particular wavelength is used to pass through an aqueous sample. The reduction in the light intensity due to scattering and absorption is proportional to the density of the cell culture, within a certain range. For microorganisms which do not express pigments, most reduction of intensity is not due to absorption, but to scattering [33], meaning that in these cases OD measurement can be considered a type of turbidity measurement. For a particular wavelength of light used ($600nm$ for example), the OD measurement will be marked as OD_{600}. As the increase in density can lead to incorrect measurements, after a certain threshold, the samples first have to be diluted. If we want to compute the growth rate of the culture, then the OD measurements suffice. The doubling time can be taken as the time needed for the OD to double in the exponential phase of culture growth. Often, however, OD measurements are used not solely to compute the growth rate, but as a way to measure the number of cells in a culture. There is no general way to relate the number of cells to an OD measurement, since too many parameters would need to be taken into account. Also, this relationship will be influenced by the cellular shape (which influences the angle of scattered light) and culture density, by the shape of the spectrophotometer and by the size of the detection sensor. Therefore, the only way to establish a reliable relationship is by sample-specific calibration [34].

1.6.2 Proteomics

Proteomics is a name given to a variety of experimental techniques through which one can determine presence, or the relative or absolute amount of (a subset of) proteins present in a cell or in a culture. The most common usage of the term proteomics today involves the techniques based on mass spectrometry (MS) [35] (even if it can mean other experimental techniques as well). As all the experiments used for the completion of this work were of MS type, I will provide a short description of this technique alone. In further text, the name *proteomics* will be assumed to mean *MS based proteomics*.

Analysis of complex protein samples is a very difficult and still not fully resolved problem [36]. Due to inherent limitations of our current state-of-the-art in proteomics, all of the steps, from sample preparation to data interpretation, are usually specifically optimized to suit the needs of the research question at stake. Yet, certain steps are common to most proteomics experiments and these will be outlined here. In order to be analyzed, the proteome first needs to be separated from the rest of the cellular content. If only a part of the cell is to be analyzed, or the sample complexity should be reduced, the cells first need to be fractionated and necessary compartments isolated. After protein extraction, depending on the research question, it might be necessary to treat the sample to alter its dynamic range (high abundance protein depletion, low abundance protein enrichment, etc.). After sample treatment, the proteins need to be "digested". Digestion is performed by enzymes that are able to cut proteins at specific places, particular to each digestion enzyme. The peptide mixture can be subjected to some form of sample simplification, such as separation by mass or isoelectric point. As the mass spectrometer is a device that measures mass to charge ratio of ions, the peptides need to be ionized before their placement into the measurement device. The great break-throughs of the 1980s in ionisation technologies enabled today's common usage of MS for complex protein mixture analysis: Matrix Assisted Laser Desorption Ionization (MALDI) [37] and ElectroSpray Ionization (ESI) [38]. Mass spectrometer will then record the ionized macromolecules as a ratio of mass and charge - m/z ratio. The intensity of any m/z point in the final spectrum will be the sum of intensities of all measurements that recorded this particular m/z ratio. Finally, it is necessary to "make sense" of the resulting spectrum. For the purpose of identification of the peptides from an m/z spectrum, databases of sequenced and annotated genomes of the species under investigation are used. Often, MS proteomics are designed to give either relative or absolute quantitative measures of individual proteins in the mixture. The intensity of the MS spectrum does not correspond in any simple way to the abundance of the protein, and two spectra cannot directly be compared to establish relative abundance. In relative proteomics, the fact is used that the difference in peak intensity does provide a good measure of relative abundance if the samples are measured in the same analyte. In order to be able to distinguish one sample from the other, they are usually grown on a medium containing a certain isotope which causes a predictable shift in protein mass. If the quantification is to be absolute, one common solution is to spike the analyte with a protein of known concentration or a number of proteins over a range of concentrations and thus establish a calibration curve to transform the spectrum information into protein abundance.

1.6.3 Total amino acid concentration measurements

To understand the overall cellular state, it is necessary to know the total protein content of the cell. Proteomics experiments cannot be used for this purpose, as they are known to have detection issues with some of the most abundant protein groups (membrane proteins, ribosomes, etc.). Therefore, an independent total protein concentration measurement is necessary. First, to note, a method to measure *total protein concentration*, which would imply measuring the concentration of fully formed proteins, does not exist to my best knowledge. What is actually measured is the weight of protein per unit of volume, which can be converted into protein concentration only if the protein sample contains a single known protein or a number of known proteins in known ratios.

As previously mentioned, most methods used for measurement of total protein concentration are based on measuring a certain physical, chemical or biochemical property of amino acids or peptide bonds that can, within a certain range, be extrapolated to a measure of total amino acid concentration (for a review of existing methods see [39]). One of the methods commonly used is the famous Bradford essay [40], due to its simplicity, reproducibility, and relative insensibility to some of the reagents typically used in proteomics. It relies on the property of the a dye (Coomassie Brilliant Blue G-250) to exist in two different colors (blue and red), and to change color from red to blue upon binding with protein. Absorbance at both of the wavelengths indicates the amount of bound and unbound dye - a measure which can be taken to indicate the total protein concentration within a certain range. The relation between the color and total protein concentration is established through calibration, for which a known concentration of a purified protein is used. The result of the measurement is given in units of g/l, with a typical significant range being in $\mu g/ml$.

This measurement can be converted into a measure of total amino acid concentration, under a set of assumptions. One must assume a certain distribution of amino acids in the sample, which allows to compute the weight of an *average* amino acid. Its weight and the total weight of protein in unit volume can then be used to compute the approximate total amino acid concentration, as was done in [32] (see Table 2, notes *b* and *h*).

1.6.4 Fluxomics

Fluxomics measures in a relative or absolute way the fluxes through a set of metabolic reactions. Even if there is a number of techniques available for determination of metabolic fluxes, here I will focus on just one - Metabolic Flux Analysis (MFA). The reason to focus on this technique is its recent widespread usage and the fact that it was applied to obtain the data for the calibration procedures in this thesis. For a good early review, see [41], where many historically significant references can be found.

To perform MFA, certain assumptions need to be fulfilled. First, the cellular culture needs to be either in steady-state - meaning that all of the metabolite concentrations and reaction fluxes need to be constant, within the bounds of "cellular" noise - or in pseudo steady-state - meaning that the rate of change should be significantly lower than the rate of measurement. Second, the part of the metabolic network studied needs to be represented in a stoichiometric model, which is used for interpretation of raw measurements. The measurement of metabolic fluxes is usually given in the unit of $mmol/(h \times gCDW)$, where CDW stands for cell dry weight.

1.7 Layout of this thesis

This thesis is organized in three parts. Part I deals with modeling and analysis of the UPR and HSR in bacterium *E. coli*. It consists of two chapters: chapter 2 focuses on the dynamical properties of UPR. It lays out the most important biological actors in this response, the proteome quality control mechanisms (chaperones and proteases) as well as the regulatory structure in place to protect the cell from this kind of stress. In chapter 3, I introduce the mathematical framework used for modeling of this stress condition. This is followed by the development of the full model, model simplification and an analysis of the properties of the simplified regulatory scheme.

Part II is dedicated to the whole-cell steady-state model of *E. coli*. First, in chapter 4, the RBA framework is intuitively and then formally introduced. I then offer one practical example using a small toy model to illustrate what kind of computations are necessary to make and simulate an RBA model. I shortly discuss related modeling paradigms, their similarities and differences. The last part of the chapter deals with `RBApy` - Python software for creating and simulating whole-cell bacterial RBA models, with special attention given to the XML format in which these models are encoded. chapter 5 deals specifically with the development, validation and exploration of the whole-cell RBA model of *E. coli*. Special attention is given to the parameterization of the model. I provide a number of simulations validating the applicability of the model to a range of biological situations. I explore the different potential applications of the model in understanding cellular regulation and discuss its potential in analyzing the costs in gratuitous protein production.

In Part III, I again explore the UPR, but this time embedded within a coherent cellular context through a set of linear equality and inequality constraints, similar to those described in chapter 4. This formulation allows the study dynamical cellular stress response under resource constraints, a point very important in production of RPs. I show that the predicted response matches quite well the experimentally observed one.

Proteome quality control during bioproduction

2 Recombinant protein production in *E. coli* 33
2.1 How proteins assume and maintain functional states
2.2 Production of recombinant protein in *E. coli*
2.3 Unfolded protein and heat shock response

3 Model of the unfolded protein stress response 47
3.1 Mathematical preliminaries
3.2 Dynamical model of HSR
3.3 Model simplification
3.4 Existence of equilibria
3.5 System behavior at the equilibrium point
3.6 Discussion

> If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat.
>
> Douglas Adams

Bacteria, and *Escherichia coli* in particular, have been used for decades as protein production hosts for industrial and research purposes. More often than not, such proteins come from species other than the host, and are therefore called recombinant proteins. The term *recombinant* derives from recombinant DNA, which has genetic material of two or more different organisms. Despite its wide usage and great research efforts, recombinant protein technology is plagued with problems and most of them center around producing a well folded, functional variant of the recombinant protein [42, 43]. This problem is so common because proteins depend heavily on a precise biochemical environment for their proper maturation, which host strains are not always able to provide. Because the issue of proper folding is of great importance, even for native proteins, all cells have dedicated systems for assisting the newly produced proteins to reach their destined compartments, as well as native and functional states. When the host cells are modified to express gratuitous recombinant protein, such proteome quality control systems get additionally activated. Because of the central importance of the quality state of the proteome, this system is involved in regulating many stress responses, and functions as a mediator of many recombinant protein production related stress effects. To lay the groundwork for understanding these stress responses, in chapter 2 I expound upon the biological background necessary for understanding these issues: starting from the process of protein folding itself, followed by the description of a part of the proteome quality control network in *E. coli*. I shortly describe the process of recombinant protein production in *E. coli* as well as the stress effects related to it. Finally, I characterize in relative detail the regulation of Heat Shock Response (HSR) and Unfolded Protein Response (UPR), the latter being the most common stress response in recombinant protein production. However, a complete survey of these topics is not within the scope of this thesis. Instead, the descriptions offered here serve to provide context and scope for the model presented in chapter 3. In chapter 3, I shortly outline the mathematical preliminaries and modeling decisions necessary for understanding of the proposed model, and then continue to detail the dynamical model for the onset of the unfolded protein response. After introducing a number of simplifications, we show that the model structure allows for a single equilibrium point, regardless of the choice of parameters. I discuss this property, quite atypical for a complex nonlinear dynamical system, yet found often in biological systems. Finally, I examine the limitations stemming from both the choices made in model formulation and the selection of the modeling framework itself. I pinpoint to some important aspects of adaptation which cannot be modeled with the chosen framework. This discussion is a prelude to the rest of the thesis, as it introduces the necessity for a different modeling framework for understanding of this systemic stress response.

2. Recombinant protein production in *E. coli*

To fully appreciate the problems related to overexpression of foreign protein in bacterial cells, it is first necessary to understand the subtleties and complexities involved in folding and maintenance of a functional proteome. Proteins, which carry out most functions in the cell, are polypeptides - or chains of amino acids - that can vary greatly in length and composition. In order to function, they need to assume their native state characterized by relative stability, which allows them to maintain the proper state under cellular conditions for long periods of time, and by flexibility which is needed to undergo proper conformational changes necessary for their function.

Both this stability and the flexibility should be maintained across a number of conditions an organism might encounter. One of the most important aspects of their surroundings, especially for unicellular organisms, is temperature with its direct effect on the properties of all the cellular components, proteins included.

On Earth, there are not so many environments which life has not populated. Among those, some are close to our living temperatures, some are extremely hot, some extremely cold. Notably, we find that, for example, in two bacteria inhabiting very different environments in terms of temperature, still most of the functions implemented by their cellular proteins are the same. However, the proteins themselves do differ, as the ones in the hot-environment bacteria will need to have higher stability in order to function. This shows us that stability is a protein property that can be modulated according to need [44]. Particular stability of a protein might be a result of a compromise between a result stable enough to maintain its functional shape, but still flexible enough to perform function related conformational changes [45]. Combining these two requirements with the fact that protein production is expensive, that all the proteins in the cell are different and have unique folding, cofactor and compartment requirements, the proteome maintenance becomes a necessary and complex system, one that exists in every living cell, and traces its origin back the very root of the tree of life [46, 47]. Stability (and conformational flexibility) of enzymes is directly related to the availability of this proteome maintenance machinery. Having enzymes that are too stable is costly, because they are slow in performing their function, but having enzymes that are less stable costs the cell in terms of the machinery in place to ensure proteome quality [48]. In this sense, stability can be seen as a trade off between different cellular costs. I continue by describing the process of folding and the proteome quality control systems in *E. coli*.

2.1 How proteins assume and maintain functional states

The so-called *central dogma of molecular biology* [49] explains how the sequence information is passed from coding DNA (cDNA) to messenger RNA (mRNA), and finally to protein. All protein information is considered to be encoded in its amino-acid sequence [50]. Under the term *protein information* we can assume a set of structures a protein can reach in which it is capable of implementing its function within a cell or an organism. The full functional structure of the protein does not necessarily include only the amino-acid chain encoded by its gene - it can include other chemical partners as well. These can be other proteins, RNA, DNA or ions, for example. However, the capacity of binding all of these is encoded in the original amino-acid sequence.

The functions that proteins perform in cells are exclusively related to their capacity to bind and potentially alter other molecules or macromolecules. Protein function relates to its structure, and this can be altered by the cell by introducing mutations into the gene coding for it if need arises. It is important to note that the structure of a protein is not a concept as simple as structure of objects closer to our experience. While objects we are familiar with have a well defined structure which is usually resilient to many mechanical and chemical stresses, protein structure is a more fluid notion. First of all, it greatly depends on the chemical or biochemical environment which surrounds it. This requirement of a particular biochemical environment often does not imply only a particular organism, but a particular state of that organism. For example, some proteins required solely at a range of temperatures will achieve a functional state only within that range. Even if in its appropriate environment, some proteins have parts or are entirely intrinsically unstructured or disordered [51].

The process through which nascent polypeptide chains assume their functional states is called *folding* and it can generally be a multi-step process, depending on the specificities of a particular protein. For cytosolic proteins, which need not be translocated or integrated into a membrane, this process could require some or all of the following steps: (i) co-translational folding, (ii) spontaneous folding, (iii) chaperone-assisted folding, (iv) post-translational modification and (v) assembly into multi-protein complexes. Since post-translational modification and complex assembly are protein-specific, I focus on the first three steps - co-translational, spontaneous and chaperone-assisted folding. Protein folding, as stated, is sensitive to intracellular conditions, and does not always end successfully. Proteins, instead of assuming their native state, can assume other states of relative stability, but in which they are unable to perform their cellular function. This happens when a portion or an entire protein *misfolds*, a process which often leaves hydrophobic patches exposed on the surface of the protein. These hydrophobic patches, which are normally buried in the interior of the protein and are one of the important factors in achieving protein stability, once exposed, easily bind to hydrophobic patches of nascent proteins which still had not had a chance to fold, or to those of other misfolded proteins. This can cause creation of potentially large insoluble protein complexes - so called *aggregates* or *inclusion bodies*. Such deleterious effects need to be controlled, and the cell has an elaborate proteome quality control system of chaperones and proteases in place for that purpose.

2.1.1 Protein folding

For a number of decades now, the major hypothesis as to how proteins assume their native state from a chain of amino acids owes its formulation to Christian Anfinsen and his colleagues [52, 53], for which he got awarded a Nobel Prize in chemistry in 1972. What he and his colleagues called the "thermodynamic hypothesis" states that [53]:

> the three-dimensional structure of a native protein in its physiological milieu (solvent, *p*H, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment.

This describes one characteristic of a final *native* state of a protein - the one with lowest Gibbs free energy - but it still leaves an open question as to how this is achieved. Does each protein fold into a completely unique structure? Yes, and no. Even if all proteins are unique, there is some kind of orderliness to their structure. Most proteins do have common structural elements, out of which larger, functional structural parts are formed. For example, most proteins have common elements of what is known as *secondary structure* - such as α-helices and β-sheets. These (and other) elements are then assembled into so-called *protein domains*, functional units of protein structure which are able to fold independently of the rest of the protein. A protein can have one or more such domains.

Today it is assumed that one of the most important ways in which proteins obtain and maintain their native state is by "hiding" of hydrophobic parts of their structure from the aqueous environment of the cell through a process known as *hydrophobic collapse* [1] [54].

As both the formation of secondary structures and the hydrophobic collapse are potentially very fast events ($< 1\mu s$ for secondary structure [55] and $< 100ns$ for hydrophobic collapse[56] [2]), and as these events are notoriously difficult to detect, the discussion is still ongoing as to the order of events leading to a formation of a native protein state [58]. Formation of tertiary structure, through which functional domains of the protein are formed, usually involves the formation of non-local interactions, by which distant parts of the amino acid chain become physically close.

Co-translational folding. This complicated and still unsolved question of acquiring of native states by proteins as is presented here (and as is often studied *in vitro*) is still only a part of what takes place in a complex cellular environment and is embedded in other cellular processes. For example, one important thing to keep in mind is that the protein begins its interaction with the cellular environment as it is being assembled by the ribosome. The speed of translation in *E. coli* varies in the range of 15 - 20 $\frac{aa}{s}$[32], while the secondary structure formation (α-helices and β-sheets) can happen at the order of less than a microsecond [55]. This gives an indication that for many proteins the first structural elements are formed co-translationally. When eukaryotic proteins are expressed in bacterial hosts, they are prone to misfold. This tendency can be reduced by reducing the speed at which they are translated [59]. It was also noted that *E. coli* populations growing in rich nutrient environments do have a higher propensity for protein aggregation [60], a fact that might be related to the speed of translation which is higher at higher growth rates, but could also be explained through cellular distribution of resources. The translation speed in some cases is purposefully slowed down. This slowing down of translation is called *ribosome stalling* and is implemented in a variety of ways through specific characteristics of the mRNA or amino acid sequence of a protein [61]. mRNA might purposefully exhibit a secondary structure which allows for translation only under certain conditions [62]. In other cases, certain regions of an mRNA might be enriched in rare codons [63]. Due to low availability of charged tRNAs corresponding to those codons, the translation of that region will be slowed down, indicating that it is beneficial (or necessary) for those proteins to achieve their native state. A similar purpose is assumed for the stretches of positively charged amino acids in eukaryotes [64], and for polyproline stretches in the amino acid chain [65] due to their bond geometry which is incompatible with the passage through the ribosome [66].

Spontaneous folding. After exiting the ribosome, proteins can spontaneously fold. In fact, many cellular proteins are capable of spontaneous folding, not requiring any additional assistance. The process of spontaneous folding still is an ongoing mystery in the field of biological research. Due to the speed at which certain folding events happen and the notorious difficulty of experimentally keeping track of such events, it has not been easy to elucidate the events leading to a spontaneous assuming of a native state by proteins. There are three major theories that propose the order of events by which the native state is achieved [67]:

- Framework model [68, 69] which suggests that the native state is achieved through a

---

[1] Hydrophobic collapse is a working theory at least for a class of globular proteins.

[2] The duration of hydrophobic collapse can vary substantially depending on the protein in question [57].

Table 2.1.1: Size, stoichiometry and abundance of the most important and most abundant *E. coli* chaperones. ^{a}: data taken from [72]

	Length (AA)	Stoichiometry	Copy per cell^a
TF	432	1	13128-69639
DnaJ	376	2	938-5688
DnaK	638	1	10567-45951
GrpE	197	2	4731-23174
GroEL	548	14	11204-57895
GroES	97	14	13733-59001

hierarchy of intermediate states, whereby the elements of secondary structure form first, and are later assembled into tertiary structure.

- Hydrophobic collapse [54] in which hiding of the hydrophobic protein parts results in a much smaller protein volume (and an intermediary state referred to as a *molten globule*), in which further folding is more easily achieved.
- Nucleation - condensation mechanism [70] which proposes that the formation of secondary and tertiary structure begins at a nucleation site, which then serves as a folding nucleus.

Folding assistance. However, some proteins require the formation of certain bonds or presence of binding partners to function. One very typical example is the formation of disulfide bonds, or binding of metabolites or ions. This is usually achieved by a number of proteins which assist other proteins to assume their natural state. In 1987, John Ellis proposed the name of *molecular chaperone* for a class of proteins "whose function is to ensure that the folding of certain other polypeptide chains and their assembly into oligomeric structures occur correctly"[71].

Apart from such specific assistance, such as facilitating a formation of a particular bondf or delivering a specific ion, certain chaperones offer a more generic kind of folding assistance. They bind a large class of nascent proteins and help in their folding through actions of *holding* and *folding* of unfolded protein, *refolding* of misfolded protein and *disaggregation* of aggregated protein. This assistance begins right at the exit of the ribosome tunnel, and is continued after the full synthesis of the protein through a coordinated action of a chaperone network. However, all of this does not ensure that all proteins will end up in their native state. When folding errors occur beyond the corrective capacity of the chaperone network, such proteins are degraded by a class of proteins known as proteases. The chaperoning and proteolytic activities in the cell are the two most important aspects of proteome quality control and maintenance.

2.1.2 Chaperoning systems in *E. coli*

Chaperone is a name for any macromolecular species which assists macromolecules assume their functional state. Bacteria, archaea and eukaryotes, while possessing different chaperones, are all equipped with the same five core chaperone families. These have been originally named by their molecular mass, and are: HSP20, HSP60, HSP70, HSP90 and HSP100 (HSP standing for heat shock protein). The reason I mention these rather unintuitive names is that they are still used frequently in the literature. Perhaps the two most studied families of chaperones in *E. coli* are the chaperonin HSP60 (GroEL/S) and the chaperone HSP70 (DnaK).

The **HSP70** family of chaperones has been reported to have a vast number of functions in *de novo* folding of protein, refolding of misfolded protein, prevention of aggregation and re-solubilization of aggregated protein. Experimental studies of DnaK, the most studied HSP70 *E. coli* chaperone, have shown it can exhibit all of these functions under different conditions, and that it functions as a "central hub" for the chaperone network [73]. Apart from DnaK, *E. coli* has two other, more specific chaperones belonging to the HSP70 family - HscA and HscC [74]. Much like the HSP70, the **HSP60** family chaperonins have been evolutionarily conserved [75]. The chaperonin GroEL/S - member of the HSP60 family - is involved in folding of newly synthesized proteins [76]. The **HSP100**/Clp class of chaperones, and its *E. coli* member ClpB, is reported to have a role in disaggregation and re-solubilization of aggregated protein in concert with other chaperones [77]. The most prominent member of the **HSP90** family in *E. coli* is HtpG, and aids in reactivation of inactive proteins. Chaperones of the **HSP20** are small heat shock proteins which assist the disaggregation of protein by interlacing the inclusion bodies and facilitating their dissolution. IbpA and IbpB are most studied members of this family in *E. coli* [78].

In addition, *E. coli* has a number of chaperones which do not belong to any of the afore named families, but perform vital roles in proteome quality maintenance. One such chaperone, and one of the most important chaperones in *E. coli*, is the Trigger Factor (TF). It associates to the ribosome and offers the first folding assistance to the proteins being synthesized [79].

Apart from the chaperones providing general assistance in folding, refolding and disaggregation, *E. coli* has a number of specific chaperones in charge of delivery of metals (NarJ, CopA, IscA, PaoD), integration of cofactors (HemW) or formation of particular bonds (DsbA/B/C/D). Here, I focus on the two most important and well-characterized general chaperoning systems which assist proteome quality maintenance through folding, refolding and preventing of aggregation - DnaJK-GrpE and GroEL/S.

DnaJK-GrpE

Even if today they are mostly known as protein-specific chaperones, dnaJ and dnaK genes were given their names for their relation to DNA replication of bacteriophage lambda [80]. Chaperone DnaK and its co-chaperone DnaJ are transcribed from the same operon under the regulation of σ^{32} dependent promoter [3]. GrpE is transcribed in a single gene transcription unit, also under the regulation of σ^{32}.

DnaK performs the ATP-dependent chaperone activity of folding of nascent proteins, refolding of misfolded protein, prevention of misfolding and aggregation and assembly of protein complexes [73]. It consists of an ATPase domain and a substrate binding domain, capped by a so-called lid domain. Its activity is regulated by its co-chaperone DnaJ, a nucleotide exchange factor, and the binding of the substrate protein. Without the action of the two other proteins, DnaK binds ATP very tightly (with a dissociation constant $K_D \approx 1nM$), hydrolyzes it very slowly (with a rate of hydrolization $k_{hyd} \approx 0.02min^{-1}$), and releases ADP at a similar rate of ATP hydrolysis [81]. DnaJ acts to modulate the speed at which DnaK hydrolyzes ATP, while GrpE modulates the rate at which DnaK releases bound ADP. In fact, DnaK ATPase activity alone correlates poorly with its foldase activity, showing the importance of its complex partners for its regulation [82].

In simple terms, the proposed mechanism of DnaK-DnaJ-GrpE coordinated activity is this [83]: ATP-bound DnaK quickly binds and dissociates substrate proteins. DnaJ exhibits greater substrate specificity and delivers the substrates to DnaK through simultaneous ATP hydrolysis [84]. This results in a more stable complex of ADP-bound DnaK and the substrate. GrpE finishes the cycle of folding/holding by speeding up the release of the peptide through the exchange of ADP for ATP, thus reducing DnaK substrate affinity [83].

---

[3] σ factors in bacteria are a special kind of transcription factors which bind the RNA polymerase and change its affinity towards a certain set of promoters. σ^{32} is the so-called heat-shock sigma factor because of its activation upon heat shock. This σ factor is discussed in detail in section 2.3

E. coli DnaK null mutants grow slowly at intermediate growth temperatures (between 30 to 37°C), and result in unviable cells at higher temperatures ($\geq 42°C$) which do not grow even after the subsequent lowering of temperature [85]. DnaK and TF have overlapping function in folding of *de novo* synthesized proteins - neither null mutant exhibits serious growth defects for an intermediate range of temperatures, while the double null mutant causes lethality [86]. A proteomic study has shown that DnaK has a broad range of substrates and was identified in complex with around 700 proteins [73]. The more aggregation-prone and heterooligomer forming proteins are statistically enriched in the set of DnaK substrates compared to the *E. coli* proteome [87].

GrpE also acts as a thermosensor - at higher temperatures, it undergoes a conformational change which again increases the affinity of DnaK to ADP and allows the bound substrates (prevent denaturation) to remain shielded from the potential toxic effect of the cellular environment at high temperatures. GrpE seems to be required for growth at all temperatures [88]. As this is not true for DnaK, this might imply a function other than the nucleotide exchange factor of DnaK.

GroEL/S

GroEL-GroES complex performs an ATP-dependent chaperonin function in *E. coli*, and is required in all growth conditions [89, 90, 91]. Both groEL and and groES are transcribed from the same operon (groE) under control of a promoter transcribed by both σ^{32} and σ^{70} bound RNA polymerase. A curious and complex nature of GroEL/S assisted folding made it a target of extensive study. GroES forms two heptameric ring complexes that are joined back to back (see Figure 2.1.1 A). These rings (called *cis* and *trans*) form cavities for substrate proteins which go through cycles of occupancy and vacancy coordinated by the binding and hydrolysis of ATP and release of ADP [92]. The interior surface of the GroEL cavity is hydrophobic and binds to the exposed hydrophobic patches of the non-native protein [93]. Two rings exhibit allosteric effects on one another. The bound state of one ring to 7 ATP molecules leaves the other ring in an unbound state at physiological ATP concentrations [94]. The binding of ATP increases the GroEL affinity for GroES, and binding of both initiates a series of conformational changes: the hydrophobic surface of the empty cavity is oriented toward GroES, the size of the cavity increases, and the non-native protein becomes exposed to a hydrophilic surface that promotes folding [95].

The activity of GroEL/S is additionally regulated by temperature [96].

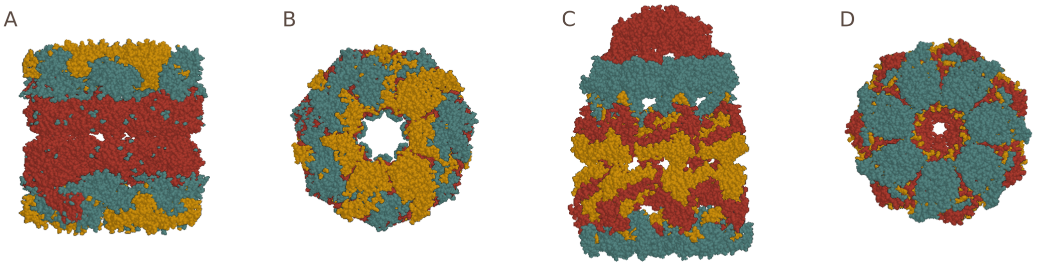


Figure 2.1.1: 3D structure of the GroEL-GroES complex. **(A)** Side view of GroEL (original source here). GroEL forms a cylinder with two cavities (top and bottom). **(B)** Top view of GroEL (original source here). **(C)** Side view of the GroEL-GroES complex (original source here). GroES acts as a lid of GroEL. Its binding on one side causes a conformational change in GroES and a release of the peptide, as well as the GroES and ADP on the opposite side. **(D)** Top view of GroEL-GroES complex (original source here). *(All images available in the public domain)*

2.1.3 Proteolysis and protein degradation

Proteolysis, or hydrolysis of peptide bonds, is necessary for maintaining the functional state of the proteome. On one side, specific proteolysis is necessary for a certain class of proteins, which need to be modified after translation by removing a peptide region, as is the case for the signal peptide in secretion-destined proteins. On the other side, proteolysis is used to degrade aberrant proteins which attenuates their potentially toxic effects to the rest of the cell and releases valuable amino acids back into the metabolism. Proteases fulfill one other very important function - they are involved in many regulatory functions, often by rapidly degrading transcription factors, thus enabling their high production flux and fast release in case of need. Since soluble and membrane-embedded proteins have very different structures and compositions, dedicated proteases exist for both. Different mechanisms for peptide bond cleavage have evolved, two of which are present in *E. coli*: (i) serine proteases [97] (ClpP, Lon) and (ii) metalloproteases (FtsH, RseP). [98, 99]. Here, I will mention only the most common general purpose cytosolic protease, Lon, and a membrane protease important in the regulation of the HSR, FtsH.

Lon

Lon (also known as protease La) is an ATP-dependent cytosolic serine protease [100]. It belongs to the class of AAA+ proteolytic machines. This class of proteases has ATPase domains which provide energy from ATP hydrolysis for unfolding the substrate protein into an enclosed chamber where it is subsequently degraded through the activity of the proteolytic domain. In its functional form, it is a homohexamer that makes a ring formation [101]. Lon is transcribed as a single gene in σ^{70} and σ^{32} dependent manner, and in an operon with ClpX in σ^{24} dependent manner. Lon protease has several regulatory functions in cell division and capsule synthesis, possibly linking proteome quality with important cellular decisions. It performs a general proteolytic function in removal of aberrant proteins [102, 103].

FtsH

In vivo, FtsH is assembled into a hexameric ring structure [104]. Each subunit has a smaller trans-membrane region and a larger cytoplasmic region, composed of ATPase and protease domains. The protease domain is dependent on Zn^{2+} for proteolytic function. ATPase functionality seems to facilitate the introduction of proteins into the proteolytic chamber by unfolding the substrate proteins. The protease is found in complex with two proteins, bound on the side of the inner membrane, HflK and HflC.

FtsH is under transcriptional regulation of two sigma factors, σ^{70} and σ^{32}, and itself is an important regulatory actor in the cell. It is known to have an essential role in lipopolysaccharide biosynthesis, enacted by degradation of LpxC and KdtA. It fact, this seems to be the regulatory function making this gene essential, since the FtsH null mutant suppressor mutations (sfhCs) modulate this cellular function [105]. It is involved in the regulation of the HSR by degrading σ^{32}, the sigma factor responsible for transcription of many of the heat shock genes. This function seems not to be essential, as σ^{32} activity can be regulated by DnaJ and DnaK [106]. Additional FtsH substrates have been determined by a trapping approach in [107] and revealed to include IscS (sulphur delivery protein), DadA (D-amino acid dehydrogenase), FdoH (formate dehydrogenase subunit) and an uncharacterized protein YfgM. As a proteome quality control actor, it is reported to remove the uncomplexed form of SecY and ATP synthase subunits [108].

2.2 Production of recombinant protein in *E. coli*

E. coli has a long history of being used as an "expression system" for recombinant proteins (RP). This history begins with the discoveries of how to cut up DNA at specific target sites with the use of bacterial restriction enzymes [109], and how to reassemble it later by the use of DNA ligases [110]. First experiments of the kind were done on bacterial plasmids - small circular fragments of DNA that bacteria can excrete and take up from the environment, and which often convey nonessential but useful genetic information (such as antibiotic resistance). Having confirmed that a single plasmid constructed by "cutting and pasting" parts of different plasmids is biologically functional [111], researchers soon after incorporated eukaryotic genetic material into a bacterial plasmid [112]. This began the era of molecular cloning - introducing foreign DNA into an organism which then continues to replicate it. In the context of RP expression, the DNA strain carrying the protein coding sequence (cDNA) is inserted into an organism-specific expression vector (most commonly a plasmid). The vector is introduced into the bacterial cell through transformation [113], a natural capacity of bacteria (and *E. coli* among them) to take up plasmids under certain biological conditions. Since there is no guarantee that all the cells will uptake the plasmid, researchers developed ways to detect the transformed cells, usually by supplementing the plasmid with a gene for antibiotic resistance. By introducing the appropriate antibiotic in the medium, non-transformed cells are killed, ensuring the survival of the plasmid carrying cells only. In order to make sure that the plasmid they carry is the one into which a recombinant gene has been successfully inserted, the placement of the recombinant DNA is chosen so that it disrupts a *tag* gene. One typical principle used is the disruption of the encoding region of a color producing gene, with the lack of color indicating the presence of recombinant DNA. This allows for visual detection of colonies carrying the target gene.

Once the culture carrying the proper recombinant DNA is isolated, it is possible to begin with protein expression and purification. Protein expression involves growing large high-density populations of transformed bacteria in bioreactors under optimized growth conditions. Bioreactors most commonly used for this purpose are batch or fed-batch. Density of the culture is monitored through OD measurements. The gene of interest is usually not expressed from the start, but is under the control of an *inducible* promoter. The inducer can be a chemical introduced into the medium, which activates the transcription of the recombinant gene, or a temperature change. After induction, cells start producing the PoI. Once the cells have produced the PoI in sufficient quantity, it needs to be extracted and purified from the rest of the cellular protein. For this purpose, PoI is often equipped with a tag, a peptide *extension* with an affinity to a particular biochemical or chemical substance. The tags can also serve a dual role and improve the solubility of the protein, and can be removed chemically or enzymatically from the PoI.

2.2.1 RP production induced stress

Production of gratuitous and foreign protein in bacterial cells can have stress causing effects upon the host. These effects can be *metabolic*, *spatial*, *folding*, *toxicity* or *population* related. Metabolic stress effects are caused by an increase in energy and precursor demands imposed by the production of the RP. One of the most common metabolic issues in RP production is that the codons used in the recombinant gene are *rare* codons in the host organism. Because of its central importance, the cell is continuously monitoring its protein producing capacity. If this capacity becomes impaired, it is detected by the presence of uncharged tRNAs at the ribosome binding site, and eventually leads to activation of the stringent response which completely alters the cellular make up, as it prepares for survival, instead of growth. Apart from the usage of rare codons, the increase in energy and precursor demands cause rearrangements in catabolism and anabolism. When the rate of production impairs the cell's ability to adapt, it will lead to a stress response.

Spatial disturbances related to RP production are caused in part by the fact that the cell is occupied by protein not useful to its functioning. Therefore, less space is available for the cellular configuration necessary to maintain a particular growth rate. As the cell size is tightly controlled in *E. coli* [114] and depends on concentrations of a number of regulatory proteins, the occupancy of the cytosol by the RP alters this fine-tuned regulation and disturbs the relation between the cell size and the growth rate [115]. This can read to growth cessation or filamentous growth, during which the *E. coli* cells continue growing without dividing [116].

Folding related issues are caused by an increase in the amount of newly synthesized protein compared to the availability of the proteome quality maintenance machinery. In fact, chaperones and proteases are often upregulated in *E. coli* RP production experiments. Additionally, some RPs might depend on a very different biochemical environment for their successful folding, and might therefore be unable to fold in *E. coli*, even once the proteome maintenance machinery has been upregulated. This can cause an accumulation of misfolded and aggregated protein in the host cells, leading to what is called the *unfolded protein response*, a cellular stress response in many ways similar to the HSR. Consequently, many host strain optimizations involve co-expression of necessary chaperones [117, 118].

Some recombinantly produced proteins are enzymes with a particular metabolic, ribonuclease or protease activity. These proteins, even if outside their native environments, can cause toxic effects by interacting with the host cell in the way of metabolic imbalances or degradation of native cellular components.

When bacterial cultures are grown to high density, as is the case in the RP production processes, they can excrete metabolic products which can additionally deter their growth. One common example of this is the accumulation of acetate in the growth medium which impairs growth [119].

Growth related effects. All the aforementioned burdens and stress effects of RP production have as an unavoidable consequence the decrease of the growth rate of the host population [120]. While initially the protein production experiments were performed in the *log* growth phase of the bacterial culture, it later became clear that the relation between the successful high yield production and the growth rate is not at all trivial, as different experiments have shown that it is beneficial to decouple the growth from RP production [121], to induce at low growth rate [122, 123] or even in stationary phase [124].

2.2.2 Process optimization

As outlined, it is not easy to intuitively determine the best strategy for obtaining optimal yield. This, in fact, is almost an obvious fact. The highly complex and optimized self-replicating bacterial cell is fundamentally altered by interfering with its capacity to control its resource investment, its rate of procreation and complete fine-tuning of its internal organisation. Additionally, a number of stress responses may be (and often are) triggered which alter the cellular state. The codons should be optimized in a way not to stall ribosomes to a degree which might trigger a stringent response, but to still ensure that translation is not too fast for the proper folding of the protein. The aspect of time adds to the complexity of the problem - is fast and intense production better than a slow and less intense one? It would also be beneficial, of course, to use the best medium composition and its optimal consumption (one in which the greatest investment in medium is directed towards the production of PoI), a thing not at all trivially related to the speed of cellular growth. The search for the best medium is typically a large part of the optimization of RP production. Very fast we find ourselves in front of a complicated optimization problem. I will touch upon this perspective in the second part of my thesis.

2.3 Unfolded protein and heat shock response

Each cellular function is "surrounded" by a regulatory network ensuring its smooth "functioning". In this section I describe the regulatory mechanisms in place to maintain the cellular proteome in a functional state. This machinery was discovered mostly through the study of the cellular response to heat shock and to λ-phage infection [125]. The greatest part of the disturbance introduced into the cell by the production of RP is due to the inability of RPs to successfully fold in *E. coli* environment, and is called the Unfolded Protein Response (UPR). This response corresponds to a great degree to a response the cell mounts to an increase in temperature, through what is called the Heat Shock Response (HSR). Because UPR can be in a way considered a subset of HSR, and because the literature on HSR is much more extensive, we will focus our study on it. Within the scope of the modeling relevant for this thesis, the two stresses can be assumed equal, and their difference is noted later.

The regulatory scheme of proteome quality maintenance was first characterized and studied under the laboratory conditions of the so-called "heat shock". In these experiments, the cell culture would be moved from "room temperature" to a higher temperature environment. This would provoke, among other things, expression of a number of proteins (then termed the "heat shock proteins"), whose roles were soon shown to be involved in the maintenance of the proteome quality which is particularly sensitive to temperature. Most common roles attributed to these proteins are of assistance in folding and the maintenance of the folded state of the proteome and the degradation of misfolded and damaged proteins. For detailed description of this system in *E. coli* see section 2.1. It was later understood that the unfolding of the proteome caused by other growth conditions will provoke the synthesis of the same proteostasis proteins [126]. The cellular reaction to any condition which causes the unfolding of its proteome is termed UPR.

2.3.1 σ^{32}-mediated unfolded protein response

The importance of σ factors in regulating bacterial stress responses has already been noted in section 1.3. The stress σ factor σ^{32} (product of the *rpoH* gene) - also known as the *heat shock* σ factor - performs an important regulatory role when cells are exposed to heat, but also when the cell is faced with the unfolding of its proteome. In such circumstances, σ^{32} in great part replaces the "housekeeping" σ factor σ^{70}, and becomes the predominant σ factor in the cell. This happens as the chaperones and proteases which constantly sequester and degrade this σ factor under normal growth conditions become occupied by unfolded, misfolded and aggregated proteins. This serves to indicate to the cell that it is not equipped with enough of proteostasis machinery, and the released σ^{32} factor binds RNA polymerase and initiates the transcription of, among other things, chaperones and proteases - the same ones that contribute to its sequestration and degradation.

This is illustrated in Figure 2.3.1, where it is shown how the unfolding of the proteome is linked to the activation of the σ^{32}-mediated HSR and UPR.

Mechanism of response activation. It was in the beginning of the 80s that the researchers first started assembling an image of how the heat shock was regulated in *E. coli*. In 1981, it was discovered that a mutation in the then recently discovered *htpR* gene led to the inability of the cell to induce synthesis of the then-known heat shock proteins, and thus, a putative positive regulatory role was assigned to it [127]. A set of proteins affected by this change was identified, but the roles of most of these proteins were not clear at the time.

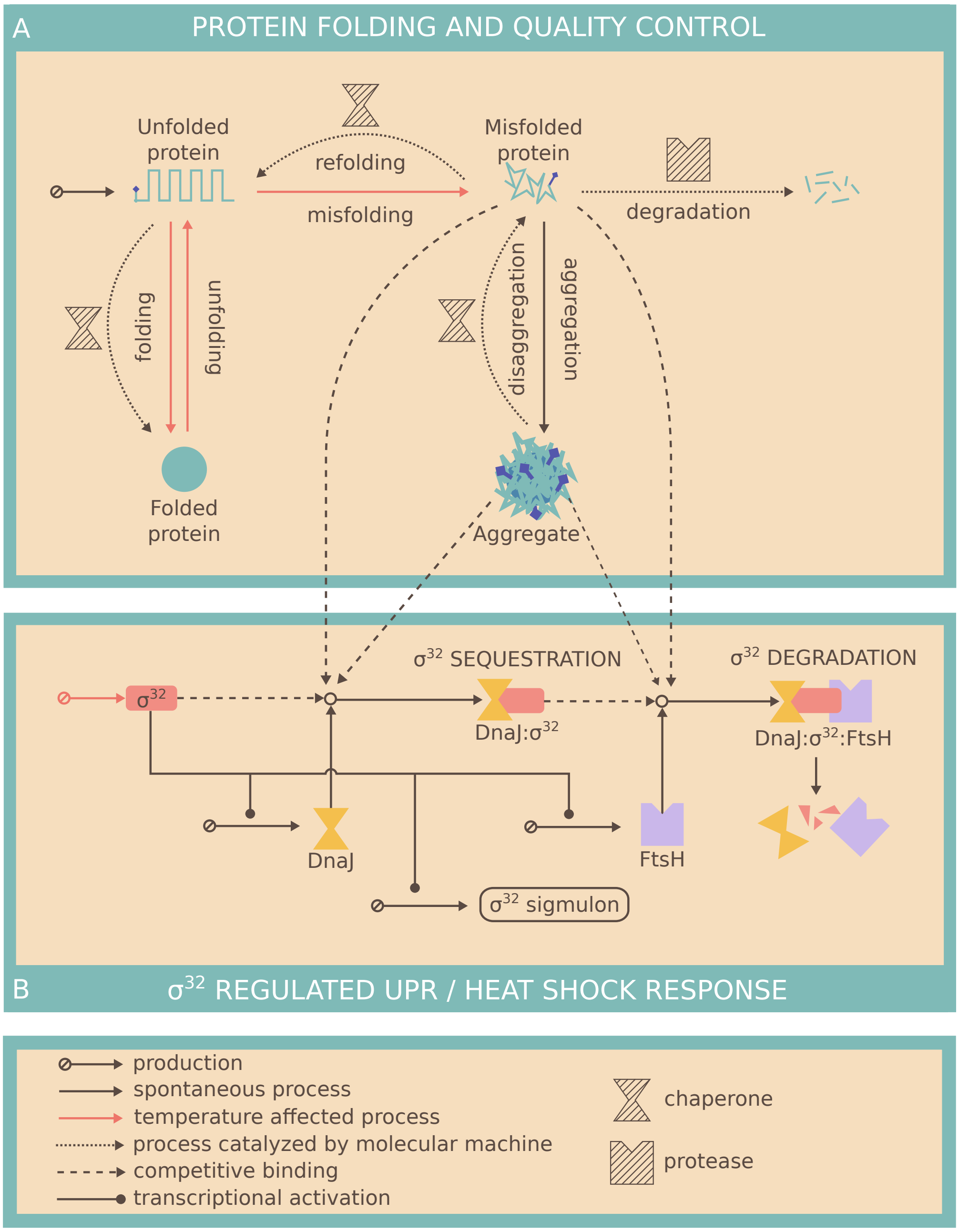


Figure 2.3.1: Relation of HSR and protein quality control. **(A)** Nascent proteins can assume a number of states: unfolded, folded, misfold and aggregated (this is of course not a comprehensive list, but serves rather the purpose of illustration). Chaperones are protein complexes in charge of holding, folding, refolding and disaggregation of protein. Proteases are in charge of degrading misfolded or aggregated proteins. **(B)** HSR, as well as UPR are regulated primarily through the action of the σ factor σ^{32}, which, when attached to the RNA polymerase, helps transcribe a number of genes related to proteome maintenance. σ^{32} is constantly sequestered and degraded by the very same chaperones and proteases it helps transcribe. When misfolded and aggregated proteins accumulate, they bind the proteome quality control machinery, resulting in the rise of cellular σ^{32} levels.

Parallel to these discoveries, in 1983, it was shown that *dnaK* gene, which was by that time already identified as encoding for one of the heat shock proteins, was a regulator of HSR. The strain carrying a mutation in this gene was unable to shut off the heat shock response after a temperature increase, and kept on synthesizing heat shock proteins 2 hours into the adaptation. Conversely, the strain overproducing *dnaK* had a comparatively mild response to heat shock [128]. It was thus showed that at least one heat shock protein exhibits a negative regulatory function in the process of adaptation.

The *htpR* gene was later proven to encode for σ^{32} which, complexed with the RNA polymerases, induces the transcription of heat-shock promoters, and thus enacts a positive transcriptional regulatory role in the heat shock adaptation [129]. A further link in the regulatory mechanism was uncovered when it was shown that the regulatory effect of *dnaK* is related to the synthesis and stability of σ^{32}, as it was shown that it is the mutation in the *dnaK* gene which prevents the shut-off of σ^{32} synthesis in the post-adaptation stage. However, the exact molecular mechanism by which this happened was not clear [129].

The σ^{32} factor is a short-lived protein under normal growth conditions, with a half-life of about 1 minute. By establishing a correlation between the relative σ^{32} concentration and the *dnaK-dnaJ* mRNA synthesis, researchers postulated that changes in concentration of σ^{32} regulate the HSR [130]. The same group established the effect of the *dnaK*, *dnaJ* and *grpE* null mutants on the synthesis and stability of σ^{32}, demonstrating a difference in half-life in each of the null mutants [131]. Today, with a better understanding of the functioning of the chaperone complex DnaJK-GrpE, this does not come as a surprise (see section 2.1.2).

DnaK is soon shown to be a protein belonging to the then newly discovered class of molecular chaperones [132]. Further experiments show that it is the binding, and possibly the chaperone function of DnaK which acts as a regulator of σ^{32} stability [133]. A theory is proposed that it is the sequestering of DnaK by unfolded proteins that initiates HSR [134]. While being sequestered by unfolded protein, DnaK is not able to bind and destabilize σ^{32}.

Role of chaperones in σ^{32} inactivation. With the understanding of the chaperone function of the DnaJK-GrpE complex, it became obvious that the short half life of the σ^{32} factor cannot be explained through its action alone, but that there should be another, quite possibly proteolytic mechanism in place (such as was already discovered for other processes regulated by constant production and degradation of transcription factors). In 1995, *FtsH* protease is implicated in the degradation of σ^{32} [135, 136]. As already explained in section 2.1.3, FtsH is a membrane-bound metalloprotease with distinct roles in membrane protein integration and protein secretion. By isolating an *ftsH* null mutant, researchers were able to show that its role in regulating heat shock was not essential, as the strain lacking FtsH eventually seized the production of the heat shock proteins, albeit more slowly, and with higher steady-state levels of heat shock proteins than in the wild type [106]. Considering the binding affinities of σ^{32} to RNAP and to DnaJK / GroELS, the pure sequestration model becomes quite unlikely [137].

Temperature sensing. A further nuance in the regulation of HSR was added by the discovery that the structure of mRNA encoding for σ^{32} is sensitive to temperature, and allows for higher rates of translation upon temperature increase [138]. This regulatory effect marks a difference between the HSR and UPR. While increase in temperature will result in a higher level of translation of the σ^{32} protein, accumulation of unfolded and aggregated protein does not exhibit any known influence upon this rate.

DnaJK-GrpE and FtsH interaction. A link was postulated between the role of chaperones and FtsH, because σ^{32} is the only known substrate of FtsH which requires to be delivered to it by a chaperone [139], thus indicating that it is a regulatory mechanism, not a biological necessity for this protease. The actual mechanism of interaction is as of today still not clear, and researchers have not been able to reconstruct it *in vitro*.

The first proteases suspected to degrade σ^{32} were the cytosolic proteases of more general scope. The discovery that it was in fact FtsH, the only essential and membrane bound protease, brought to light the possibility that this part of the regulation serves to monitor the state of the inner membrane proteome.

Regulation by and of σ^{32}. σ^{32} is the protein product of the *rpoH* gene, which is under extensive transcriptional control. This allows the cell to precisely tune its production rate in response to many cellular signals (the significance of which is still not fully understood). It is transcribed from at least four different promoters [140], two of them under control of the housekeeping sigma factor σ^{70}, one under the extreme heat shock σ factor σ^{24} [4] and one under the control of the nitrogen limitation factor σ^{54} [142], with usage of the promoters changing greatly with temperature [141]. Additionally, it is regulated by the cAMP-CRP-CytR nucleoprotein complex [143] and DnaA [144].

There are 99 known operons which are under the control of the σ^{32} promoter, with 152 genes encoding for protein and RNA products (see Table A.2.1) [5]. As can be seen from Table A.2.1, there is a number of central cellular functions under significant regulation by σ^{32}:

1. Post-translational modification, protein turn-over and chaperones
2. Translation, ribosomal structure and biogenesis
3. RNA processing and modification
4. Replication, recombination and repair
5. Defense mechanisms
6. Cell wall / membrane biogenesis
7. Metabolism, and in particular inorganic ion, carbohydrate and nucleotide transport and metabolism
8. Signal transduction mechanisms

Of course, the complexity of the response triggered cannot be fully appreciated by such a list, as almost all of the genes listed are also under other types of regulatory control which enables for very precise regulation of their production. Still, such a list gives an overview of the changes the cell undergoes when faced with heat shock / unfolded protein shock. Particularly, its obvious coordination with the central cellular functions, such as translation, implies complex and systemic adaptation [145].

---

[4] The extreme heat shock σ factor was discovered in the attempt to understand the transcriptional activation of σ^{32} [141].

[5] Data taken from EcoCyc database [6]

3. Model of the unfolded protein stress response

> Complete freedom from stress is death.
>
> Hans Selye, Stress without distress, 1974

In this chapter I mathematically describe and analyze the regulation in *E. coli* responsible for the maintenance of the quality of its proteome, and show how we can further our understanding of this system through the exercise of modeling. In particular, we show that the structure of this system (as represented in our model) guarantees the existence of a sole equilibrium for all parameter values.

Several aspects of proteome quality maintenance in *E. coli* have already been modeled. *FoldEco*[146] is a modeling study focusing on detailed reconstruction of the proteostasis network in *E. coli* in a dynamic model which allows for inference of effects of parameters on protein destiny, and can be used as a database of parameters related to proteostasis. On the other hand, a smaller model of the proteostatic regulatory network has been studied in [147] and its follow-ups [148, 149]. This study proposes a link between the design of the biological regulatory 'circuit' of the HSR and the control systems as implemented through the science of automatic control, and it discusses the reasoning behind such a design.

The work in this thesis, in turn, focuses on the analysis of the mathematical representation of the system by a set of ODEs and the considerations as to the validity of such a representation. Such nonlinear models of intracellular interactions are notoriously difficult to parameterize due to the difficulty of associating the mathematical parameters to precise *in vivo* measurements and the poor correspondence to *in vitro* ones. Moreover, any behavior that the system is shown to exhibit with one set of parameters is not necessarily the behavior it would exhibit in a different point of parameter space. In light of great difficulty in obtaining relevant parameter values for cellular systems [150], I focus my analysis on the *structure* of the system and discuss the properties evident in it.

In order to perform this analysis, I first propose a relatively detailed model of production and assumption of different stability states of protein (unfolded, partially folded, folded, misfolded, aggregated) and the regulatory mechanisms in place to react to the changes in the distribution of cellular protein between these states. I then propose a number of simplifications to this model and offer justifications as to why these simplifications can be made. The analysis of the simplified model brings to light a property inherent in its structure - a single equilibrium point. I discuss the potential implications of such an unexpected simplicity found in a complex nonlinear dynamical system. After the demonstration of this property, I list a number of possible limitations of the model, and examine the applicability of the ODE framework to represent regulatory systems with feedback in biological systems. This discussion serves as an introduction to the rest of the thesis, as it pinpoints the possible need for different modeling paradigms for understanding systemic stress responses. With its focus on structural properties, this work goes along with the growing interest in the structural analysis of biological networks (work focusing on characterization of ODE models of metabolic networks [151, 152, 153], network motifs [154] and generic structural considerations of chemical reaction networks in terms of stability [155, 156], just to name a few). The chapter begins by a short description of the mathematical framework used for the construction of this model and its analysis.

3.1 Mathematical preliminaries

3.1.1 Mathematical modeling of chemical reactions

When studying events taking place inside of a cell, one is most often faced with interactions of different chemical components. In this case, the cell (or its portion of interest to the researcher) is the chemical system under study. The changes in the state of a chemical system over time happen as a result of elementary chemical reactions. Elementary reactions are such so that the conversion from substrates to products occurs in a single step, with no detectable chemical intermediates. One class of elementary reactions is the *unimolecular* elementary reactions, in which one chemical component goes through spontaneous alteration, either by dissotiating into smaller chemicals, by isomerization or by radioactive decay. If the reaction involves more than one substrate, the prerequisite of it happening is the collision of all the substrates. The simultaneous collision of three chemical components in such a way that allows for the reaction to take place, although not impossible, is already highly unlikely, while there are no elementary reactions involving four or more substrates. Therefore, the only typical elementary reaction, apart from the *unimolecular* one, is the bimolecular one, which involves the collision and subsequent transformation of two substrates.

Apart from the *elementary* chemical reactions, a chemical system can be described by complex chemical reactions. These might be described as a "lump-sum" of multiple elementary reactions. The study of the temporal evolution of the state of the chemical system governed by a set of chemical reactions falls under the domain of *chemical kinetics*.

A chemical system at a time t is described by a set of chemical components, a vector of their quantities at time t and a set of reactions describing evolution of the system over time. Additionally, a chemical system may incorporate a description of space in which the reactions are taking place. This space is either modeled (such as in reaction-diffusion systems), or is represented by a set of assumptions as to its size, mixedness and openness (or closedness) to its surroundings.

A description of each of the chemical reactions needs to include (i) a list of substrate and product chemical components, (ii) a stoichiometric coefficient associated to each of the chemical components - a number of units of that component used in or produced by the reaction, (iii) a vector of quantities of each of the substrates and products and (iv) a function describing the speed at which the reaction takes place, also known as the *reaction rate*. While the first two requirements can be stated unambiguously, the third one can be measured (at least theoretically speaking), the last requirement is subject to choice. The choice of the formulation of reaction rates indeed makes up for a large portion of what can be called the *modeler's choice*. Indeed, the choice, even at a superficial inspection, gives eight possible ways to go: the time can be either continuous or discrete, as well as the chemical component quantities, and the change in these quantities can be described either deterministically or stochastically. All of the listed options can be valid for particular chemical systems. Often, the modeler is governed not only by the estimated suitability of a particular modeling choice, but also by her own expertise and the availability of mathematical tools for the study of the system (given a certain choice). Even after the choice of continuous vs. discrete and deterministic vs. stochastic has been made, the formulation of the rate expression for a given chemical reaction does not simply follow as a consequence of the underlying physical events. Our understanding and characterization of elementary reactions has greatly progressed, both in terms of experimental techniques and *in silico* molecular dynamics simulation methods. Even so, a direct correspondence between the physical events taking place and a chemical reaction rate has been established only in a very limited set of circumstances [1]. Assuming one had a good framework for description of elementary reactions, one would still be left with the problem that most reactions one wishes to describe are complex, and the sequence of events which leads to their taking course is often full of intermediates which are very difficult to detect and characterize. In a complex reaction, the number of possible paths which a reaction can take can be very great indeed.

Mass action kinetic law

One of the most common ways used to model elementary reactions has for years been the so-called *mass action kinetic law* [157]. This law assumes continuous time, continuous concentrations of the chemical components and a deterministic change in the chemical system. Although the derivation of this law, which took form in the first part of the 19^{th} century, was initially made on the basis of empirical observation [158], it was later shown that such a principle can be derived from classical thermodynamics under certain assumptions (see [159] for one such derivation). The kinetic law of mass action states that the rate of reaction is proportional to the product of reacting species' *active masses* (concentrations) and what they called the *chemical affinity* [160]. For a reaction of the form:

$$A + B \xrightarrow{k} C \tag{3.1.1}$$

kinetic law of mass action describes the changes in concentrations:

$$\frac{dC_A}{dt} = \frac{dC_B}{dt} = -\frac{dC_C}{dt} = -kC_A C_B \tag{3.1.2}$$

where C_X denotes the concentration of species X. The constant k corresponds to the original notion of *chemical affinity*. While the product of concentrations relates to the probability of collision of two possibly interacting chemical components, the interpretation of k is more subtle: it includes the dependence of the reaction rate on temperature and the probability that the collision will result in a reaction taking place.

While the expression in Equation 3.1.2 is often generalized to the following form:

$$n_A A + n_B B \xrightarrow{k} n_C C + n_D D \tag{3.1.3}$$

in which the reaction rate becomes:

$$\frac{dC_A}{dt} = \frac{dC_B}{dt} = -kC_A^{n_A} C_B^{n_B} \tag{3.1.4}$$

there is no evidence that this formulation can be extended to non-elementary reactions, and it is safe to assume that any reaction involving a simultaneous collision of more than three chemical components cannot be classified as an elementary reaction. In fact, the law was formulated as a consequence of study of elementary reactions. The mass action law assumes that all the chemical components are mixed homogeneously.

From the point of view of the type of differential equations used in the formulation of the reaction rates according to the mass action kinetics law, we can see that all the ODEs are *ordinary* (they

---

[1] Expressions derived from thermodynamics, such as the transition state theory or the Eyring–Polanyi equation, explain only what happens at a chemical equilibrium.

describe change over a single variable - time), *autonomous* (they do not depend explicitly on time) and in a general case *nonlinear*.

The mass action law has been the basis for the derivation of many other rate laws. For example, in a simple enzymatic reaction system

$$E+S \leftrightarrows ES \rightarrow E+P \tag{3.1.5}$$

where E is the enzyme, S the substrate and P the product, it might be safe to assume that the rate limiting step is the catalytic activity of the enzyme (therefore the conversion of ES to $E+P$), and that the reversible binding of the substrate to the enzyme achieves equilibrium within a time frame relevant for the catalysis. Such kind of assumptions have been used to derive different enzyme kinetic rate laws, such as Michaelis Menten or Hill equation. These formulations offer a single equation describing the rate at which the enzyme performs its catalytic activity, and thus provides a "simple" description of a potentially very complex underlying process.

Rate formulation in this thesis

Even if many details are left out in this short exposition, it does serve to begin to appreciate the subtlety involved in modeling chemical systems. Modeling cells provides additional challenges, for a number of reasons: (i) they are not exactly well mixed containers, (ii) system volume is subject to change, (iii) many of their species are present in such low amounts that they cannot be considered concentrations, (iv) they are not isolated systems, but instead exchange chemical material with their surroundings and (v) reactions happening on a membrane cannot be considered as reactions taking place in a three dimensional space, just to name a few.

The other way in which the chemical experiment set up in laboratory conditions and those taking place inside a cell differ is the latter are purposeful. This purpose is implemented through a complex chemical environment of the rest of the cell which monitors and interferes with each chemical reaction. Even if this is an argument difficult to take into account, I believe it is very important, and can help guide our modeling efforts towards more sensible, and thus useful, models. The reason for this is that when we model a chemical reaction taking place in a container set up by a chemist, we don't need to ask ourselves the question such as "What purpose does this reaction serve?", whereas when the modeled reaction takes place within a living entity, this question might be the single most important question one can ponder about.

In this thesis, I have chosen to model all the interactions using the afore described *mass action kinetic law*. Since our system of interest is well studied, with decades of experimental effort into elucidating the workings of each reaction, we could identify the elementary reactions of our chemical system with a reasonable degree of confidence. Formulating the time evolution of our chemical system in such a way allows us to use the existing analytical tools for the analysis of nonlinear ODEs.

3.1.2 Equilibria of dynamical systems

In subsection 1.4.1, I briefly describe the kind of dynamical systems used for the implementation of the model in this chapter. There, I also mention how complex the behavior of such dynamical systems can be when they are nonlinear. As we see in the formulation of the "mass action law", the dynamic description of chemical interactions often includes nonlinear terms. We have also mentioned that for such systems it is often impossible to obtain analytical expressions for their trajectories. However, even if we were equipped with an analytical "solution" of a nonlinear dynamical system, that might not necessarily be the best way in which we could understand the *behavior* and characteristics of the system. If, following the example from [161], we take a system:

$$\dot{x} = \sin x \tag{3.1.6}$$

and obtain an analytical expression defining its evolution in time:

$$t = \left| \frac{\csc x_0 + \cot x_0}{\csc x + \cot x} \right| \tag{3.1.7}$$

where csc stands for cosecant, and cot for cotangent. This expression, while being exact, does not contribute much to our understanding of the system - in fact - we would again need to analyze this expression - instead of the original system - so as to gain some intuition of what possible behaviors this system allows for.

However, we know that, chaotic systems aside [2] , dynamical systems can exhibit a number of *typical* behaviors. These typical behaviors would become obvious if we would initialize the system at a number of points $\boldsymbol{x_0}$ and observe what happens to the system after a long time [3]. They are related to the number and type of *equilibrium points* and *limit cycles* of the system. Given a system whose change over time is described by:

$$\frac{d\boldsymbol{x}}{dt} = \dot{\mathrm{x}}(t) = f(\boldsymbol{x}) \tag{3.1.8}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$, equilibrium points are points $\boldsymbol{x}^*$ such that:

$$f(\boldsymbol{x}^*) = 0 \tag{3.1.9}$$

Since this means that the change of the system over time is zero ($\dot{\boldsymbol{x}}(t) = 0$, the system that starts out in $\boldsymbol{x}(t_0) = \boldsymbol{x}^*$ remains in it forever:

$$\boldsymbol{x}(t) = \boldsymbol{x}^*, \qquad t \in [t_0, \infty) \tag{3.1.10}$$

Periodic (closed) orbits of period T are such that

$$\boldsymbol{x}(t) = \boldsymbol{x}(t+T), \qquad t \in [t_0, +\infty) \tag{3.1.11}$$

While the detailed study of nonlinear dynamical systems is in terms of required mathematical level rather difficult, in certain cases some intuition can be gained through the use of graphical methods. As is the case with the appeal of Feynman diagrams [162], dynamic systems can sometimes be characterized graphically by using *phase portraits* in a way which allows for a good engineering-level understanding of their behavior.

This is particularly suitable for analysis of two-dimensional systems. We will take this system as an example:

$$\begin{aligned} \dot{x} &= x(x-y) \\ \dot{y} &= y(2x-y) \end{aligned} \tag{3.1.12}$$

---

[2] We can indeed leave chaotic systems aside in this consideration, because the regulated and predictable behavior necessary for the functioning of an organism has excluded chaotic systems as potential solutions to any of the regulatory problems faced by the cell and organism.

[3] The number of these points could theoretically be infinite, and the observation time as well.

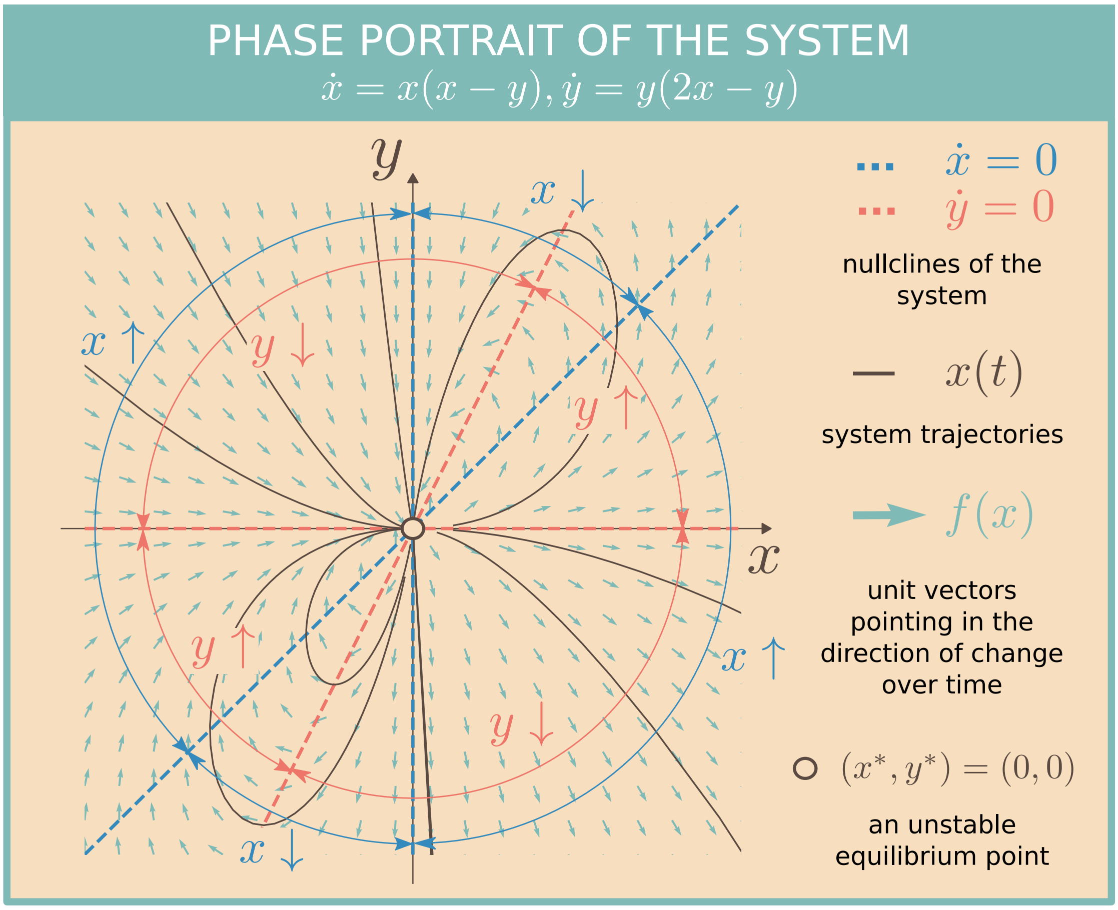


Figure 3.1.1: Phase portrait of a dynamical system described by Equation 3.1.12. The red and blue dashed lines represent nullclines associated with the variables x and y, correspondingly. The dark brown lines are trajectories of the system initiated at different points in the phase space. The green arrows are unit vectors pointing in the direction of change over time for x and y. The system has one equilibrium point $(x^*, y^*) = (0,0)$. Since there are trajectories which, if initiated in close proximity of lead away from it and remain distant at $t \to \infty$, the equilibrium point is unstable.

By setting $\dot{x} = 0$ we obtain a set of points in phase space at which x does not change over time. This set of points (usually curves) is called a *nullcline*. In our case, the nullclines of x are:

$$x = 0 \tag{3.1.13}$$

$$x = y \tag{3.1.14}$$

While the y nullclines are:

$$y = 0 \tag{3.1.15}$$

$$y = 2x \tag{3.1.16}$$

The intersection of all the nullclines defines the equilibrium point of the system, which is $(x = 0, y = 0)$. We can graphically represent this information in a phase plot, as in Figure 3.1.1. We notice that the nullclines for a particular variable separate the regions in which that variable increases (positive derivative) or decreases. These regions are marked by arcs in blue and red for x and y. Determining these regions by knowing that there can be no change in variables along their nullclines, we are able to sketch the trajectories of the system. This allows us to conclude that the one equilibrium point of our system is unstable, as there are trajectories which start arbitrarily close to it, but lead away from it as $t \to \infty$.

This idea can in some cases be extended to systems of dimension higher than two, if it is possible to simplify the system of nullclines until we reduce it to two expressions with two variables. In that case, if the shape of the curves are such so that only a single intersection is possible, we can show that the system has a single equilibrium point.

3.2 Dynamical model of HSR

With the aim to understand the behavior of the stress response in *E. coli* during the change in cellular proteome quality due to unfolding, misfolding and aggregation of protein, I first propose a model of the state of the cellular proteome (in terms of its foldedness) and the regulatory mechanism in place to monitor and react to the changes in its quality.

In this section, I present the full model in ODE form, as graphically represented in Figure 3.2.1 and Figure 2.3.1.

As already mentioned, UPR can be considered as a subset of HSR. Therefore, the model presented here, even if describing HSR, in fact describes both. The full heat shock regulation model integrates the regulatory loop as represented in Figure 2.3.1 and Figure 3.2.1. It assumes constant production of cellular protein C_p. The protein pool is divided into proteins that can spontaneously fold P_{sf} which constitute a portion α of all cellular protein (and are thus produced at a rate equal to αC_p) and obligatory GroELS substrates P_G (produced at rate $(1-\alpha)C_p$). We include this difference since a part of the proteome of *E. coli* requires GroELS under all growth conditions. The obligate GroELS substrates are not many (~ 85), but include 13 proteins essential under all growth conditions [163]. Both types of protein can assume a variety of states: unfolded uP, native nP, misfolded mP and aggregated aP. The obligate GroELS substrates can assume one additional state: partially folded pfP. Apart from $GroEL/S$, this model features another chaperone complex - the $DnaJK-GrpE$. This complex was chosen because of its central role in almost all the aspects of protein *de novo* folding, refolding, prevention of aggregation, disaggregation and re-solubilization of aggregated protein [73] .

While only P_{sf} can spontaneously fold, both P_{sf} and P_G can, once in native state, spontaneously unfold. Both types can be folded back into native state through binding of the chaperone. The

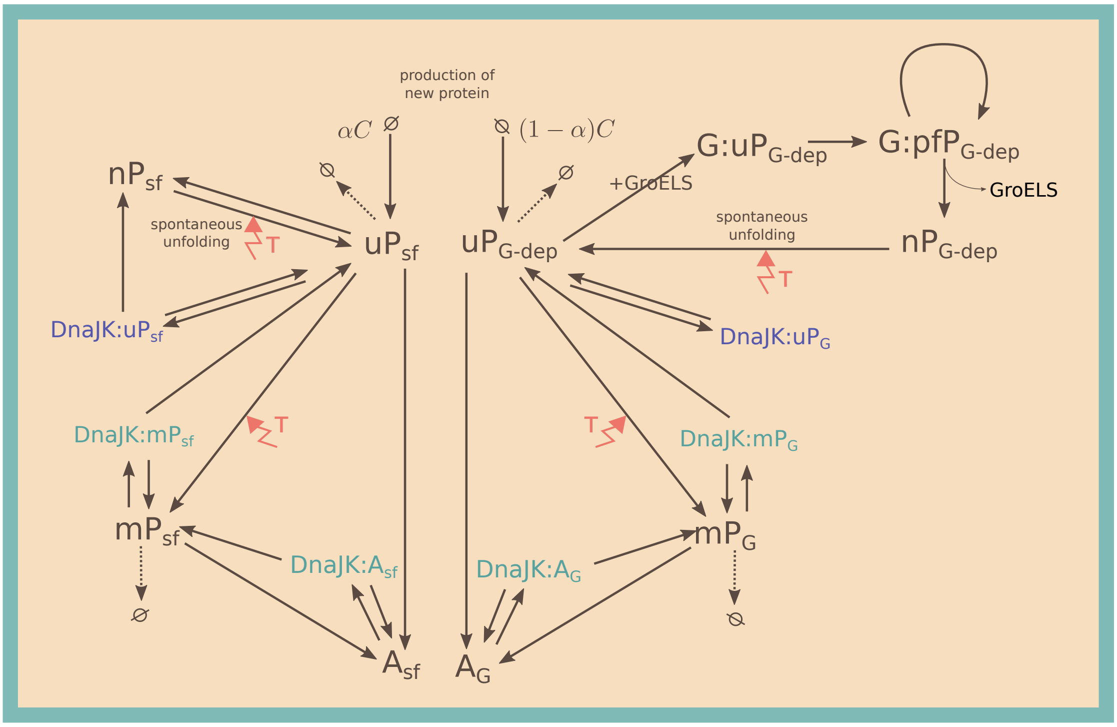


Figure 3.2.1: Simplified model of the protein folding dynamics in bacteria. The cellular protein pool is divided into two parts - proteins which require the assistance of *GroELS* chaperonin complex folding P_G (right side of the figure), and those that do not, and can fold spontaneously P_{sf} (left). Both of these types of protein can exist in multiple states: unfolded uP, native nP, misfolded mP and aggregated aP. The GroELS dependent protein can exist in one additional state: partially folded pfP. Proteins are produced in their unfolded form and can achieve native form either through spontaneous (for the P_{sf} class) or chaperone-assisted folding. By unfolding, the native state can change to the unfolded state. Unfolded proteins can misfold and aggregate. Misfolded proteins can revert to the unfolded state by the action of the chaperone. Aggregated proteins can be converted back to misfolded proteins by the action of the chaperone.

unfolded protein can misfold, and both unfolded and misfolded protein can aggregate. Misfolded protein can be reverted to unfolded state by the chaperones. As misfolded proteins are relatively stable, they cannot be immediately refolded to native state, but need chaperone assistance to revert back to the more unstable, unfolded state [164]. Once aggregated, protein can be reverted to misfolded state by the action of the chaperone. Unfolded and misfolded protein can be subject to degradation by the protease *Lon* whose concentration is assumed constant. Unfolded and misfolded protein can be degraded by the action of the protease *FtsH*. All reactions are modeled through mass action kinetics.

The mass balance equations of the model are the following:

$$\begin{aligned}
[P_{tot_{sf}}] &= [uP_{sf}] + [nP_{sf}] + [mP_{sf}] + [aP_{sf}] \\
&\quad + [DJK : uP_{sf}] + [DJK : mP_{sf}] + [DJK : aP_{sf}] \qquad (3.2.1) \\
[P_{tot_G}] &= [uP_G] + [nP_G] + [mP_G] + [aP_G] \\
&\quad + [DJK : uP_G] + [DJK : mP_G] + [DJK : aP_G] \\
&\quad + [G : uP_G] + [G : pfP_G] \qquad (3.2.2) \\
[\sigma^{32}_{tot}] &= [\sigma^{32}] + [DJK : \sigma^{32}] + [DJK : \sigma^{32} : FtsH] \qquad (3.2.3) \\
[DnaJK_{tot}] &= [DnaJK] \\
&\quad + [DJK : \sigma^{32}] + [DJK : \sigma^{32} : FtsH] \\
&\quad + [DJK : uP_{sf}] + [DJK : mP_{sf}] + [DJK : aP_{sf}] \\
&\quad + [DJK : uP_G] + [DJK : mP_G] + [DJK : aP_G] \qquad (3.2.4) \\
[GroELS_{tot}] &= [GroELS] + [G : uP_G] + [G : pfP_G] \qquad (3.2.5) \\
[FtsH_{tot}] &= [FtsH] + [DJK : \sigma^{32} : FtsH] \qquad (3.2.6)
\end{aligned}$$

Relation between protein in the aggregated state and aggregates (which are composed of n_P protiens):

$$\begin{aligned}
[A_{sf}] &= n_P [aP_{sf}] \qquad (3.2.7) \\
[A_G] &= n_P [aP_G] \qquad (3.2.8)
\end{aligned}$$

Change in total quantities:

$$\begin{aligned}
[\dot{P}_{tot_{sf}}] &= (1-\alpha)C_p - \mu [P_{tot_{sf}}] - k_{deg_1}[uP_{sf}] - k_{deg_2}[mP_{sf}] \qquad (3.2.9) \\
[\dot{P}_{tot_G}] &= \alpha C_p - \mu [P_{tot_G}] - k_{deg_1}[uP_G] - k_{deg_2}[mP_G] \qquad (3.2.10) \\
[\dot{\sigma}^{32}_{tot}] &= f_P(T, \sigma^{70,38,24,54}) - \mu [\sigma^{32}_{tot}] - k_{deg_3}[DJK : \sigma^{32} : FtsH] \qquad (3.2.11) \\
[DJ\dot{K}_{tot}] &= f_P(\sigma^{32}) - \mu [DJK_{tot}] \qquad (3.2.12) \\
[Gro\dot{E}LS_{tot}] &= f_P(\sigma^{70,32}) - \mu [GroELS_{tot}] \qquad (3.2.13) \\
[Fts\dot{H}_{tot}] &= f_P(\sigma^{70,32}) - \mu [FtsH_{tot}] \qquad (3.2.14)
\end{aligned}$$

All protein states of the spontaneous folding protein type:

$$\begin{aligned}
[u\dot{P}_{sf}] &= (1-\alpha)C_p - \mu [uP_{sf}] - k_{deg}[Lon][uP_{sf}] \\
&\quad - k_F(T)[uP_{sf}] + k_U(T)[nP_{sf}] \\
&\quad - k_B[DJK][uP_{sf}] + k_D[DJK : uP_{sf}] \\
&\quad - k_M(T)[uP_{sf}] + k_R[DJK : mP_{sf}] \\
&\quad - k_A(T)[uP_{sf}] \qquad (3.2.15) \\
[n\dot{P}_{sf}] &= k_F(T)[uP_{sf}] - k_U(T)[nP_{sf}] + k_{D_2}[DJK : uP_{sf}] \qquad (3.2.16) \\
[m\dot{P}_{sf}] &= k_M(T)[uP_{sf}] - \mu [mP_{sf}] - k_D[Lon][mP_{sf}] \\
&\quad - k_B[DJK][mP_{sf}] + k_D[DJK : mP_{sf}] - k_R[DJK : mP_{sf}] \\
&\quad - k_A(T)[mP_{sf}] + k_{DA}[DJK : aP_{sf}] \qquad (3.2.17) \\
[a\dot{P}_{sf}] &= k_A(T)[uP_{sf}] + k_A(T)[mP_{sf}] - \mu [aP_{sf}] \\
&\quad - k_B[DJK][aP_{sf}] + k_D[DJK : aP_{sf}] \qquad (3.2.18)
\end{aligned}$$

All protein states of the proteins that require GroELS for folding:

$$\begin{aligned}
\dot{[uP_G]} = {} & \alpha C_p - \mu[uP_G] - k_{deg}[Lon][uP_G] \\
& + k_U(T)[nP_G] \\
& - k_B[DJK][uP_G] + k_D[DJK:uP_G] \\
& - k_M(T)[uP_G] + k_R[DJK:mP_G] \\
& - k_A(T)[uP_G] && (3.2.19) \\
\dot{[nP_G]} = {} & k_{f2}[G:pfP_G] - \mu[nP_G] && (3.2.20) \\
\dot{[mP_G]} = {} & k_M(T)[uP_G] - \mu[mP_G] - k_D[Lon][mP_G] \\
& - k_B[DJK][mP_G] - k_D[DJK:mP_G] \\
& - k_A(T)[mP_G] + k_{DA}[DJK:aP_G] && (3.2.21) \\
\dot{[aP_G]} = {} & k_A(T)[uP_G] + k_A(T)[mP_G] - \mu[aP_G] \\
& - k_B[DJK][aP_G] + k_D[DJK:aP_G] && (3.2.22)
\end{aligned}$$

Dynamics of the complexes are:

$$\begin{aligned}
\dot{[DJK:uP_{sf}]} &= k_B[DJK][uP_{sf}] - (k_{D_1} + k_{D_2})[DJK:uP_{sf}] && (3.2.23) \\
\dot{[DJK:mP_{sf}]} &= k_B[DJK][mP_{sf}] - k_D[DJK:mP_{sf}] && (3.2.24) \\
\dot{[DJK:aP_{sf}]} &= k_B[DJK][aP_{sf}] - k_D[DJK:aP_{sf}] && (3.2.25) \\
\dot{[DJK:uP_G]} &= k_B[DJK][uP_G] - k_D[DJK:uP_G] && (3.2.26) \\
\dot{[DJK:mP_G]} &= k_B[DJK][mP_G] - k_D[DJK:mP_G] && (3.2.27) \\
\dot{[DJK:aP_G]} &= k_B[DJK][aP_G] - k_D[DJK:aP_G] && (3.2.28) \\
\dot{[G:uP_G]} &= k_B[G][uP_G] - k_{f1}[G:uP_G] && (3.2.29) \\
\dot{[G:pfP_G]} &= k_{f1}[G:uP_G] - k_{f2}[G:pfP_G] && (3.2.30) \\
\dot{[DJK:\sigma^{32}]} &= k_B[DJK][\sigma^{32}] - k_D[DJK:\sigma^{32}] && (3.2.31) \\
\dot{[DJK:\sigma^{32}:FtsH]} &= k_B[DJK:\sigma^{32}][FtsH] - k_{deg}[DJK:\sigma^{32}:FtsH] && (3.2.32)
\end{aligned}$$

3.3 Model simplification

The model presented above has 24 ODEs and 6 algebraic equations. The sheer size of the model prohibits anything but numeric simulation. In order to be able to analyze it, I first introduce a number of assumptions and simplifications which result in a more comprehensible model.

***Assumption 1* - FtsH interaction.** Because there is no biological evidence that the free quantity of FtsH is a signal relevant in regulating the σ^{32}-mediated HSR or UPR (unlike the case of chaperones) and the degradation proceeds fast in *in vivo* conditions (half-life of σ^{32} shorter than 1 minute), and because its precise role in the coordination of the response is still not well understood, I propose to make the following assumption: FtsH does not need to undergo complexing with $[DJK:\sigma^{32}]$ in order to degrade the σ factor, but instead, degradation proceeds at a pace proportional to $[FtsH_{tot}][DJK:\sigma^{32}]$. Since in my model FtsH is used solely to degrade σ^{32}, which *in vivo* should occupy a minimal portion of the protease, it is reasonable to approximate the available FtsH portion with the total amount [4]. Furthermore, since the transcription of the *ftsH* gene depends on both the housekeeping σ^{70} and the heat shock σ^{32} sigma factors, it is not trivial

---

[4] The simulation of the parameterized model yielded the portion of FtsH occupied by σ^{32} less than one percent ([147], Supplementary table 2).

to represent its transcription as a function of σ^{32}. The reason for this is that the increase in σ^{32} is accompanied by a decrease in σ^{70}, making the consequences for the transcription of *ftsH* unclear without further experimental investigation. Therefore, I assume that the production of FtsH is not dependent on the quantity of free σ^{32} but rather constant, making $[FtsH_{tot}]$ a parameter of the system.

Simplification 1 **- Removal of GroELS.** The first simplification introduced in the model is the removal of the difference between the obligate GroELS substrates and the other proteins. While both DnaJK-GrpE and GroELS take part in controlling for the activity of σ^{32}, they seem to perform the same sequestration role [165]. Due to the similarity of their regulatory role, I remove the GroELS, and keep only the DnaJK chaperone, and thereby obtain a model simple enough to be analyzed.

Simplification 2 **- Bundling of non-native protein states.** As an additional simplification, I chose to bundle the non-native protein states (unfolded, misfolded, aggregated) into a single non-native protein state - *unfolded*. Since the chaperone DnaJK through its folding, refolding and disaggregation function interacts with all of the named non-native protein states, this simplification still preserves the basic relation of DnaJK to non-native protein.

With these simplifications, I obtain the following model. Mass balance equations are:

$$[P_{tot_{sf}}] = [uP_{sf}] + [nP_{sf}] + [DJK : uP_{sf}] \tag{3.3.1}$$
$$[\sigma^{32}_{tot}] = [\sigma^{32}] + [DJK : \sigma^{32}] \tag{3.3.2}$$
$$[DJK_{tot}] = [DJK] + [DJK : \sigma^{32}] + [DJK : uP_{sf}] \tag{3.3.3}$$

Changes in total quantities are:

$$\frac{d[P_{tot_{sf}}]}{dt} = (1-\alpha)C_p - \mu[P_{tot_{sf}}] - k_{deg_{uP}}[Lon][uP_{sf}] \tag{3.3.4}$$
$$\frac{d[\sigma^{32}_{tot}]}{dt} = f_{P:32}(T, \sigma^{70,38,24,54}) - \mu[\sigma^{32}_{tot}] - k_{deg_{32}}[FtsH_{tot}][DJK : \sigma^{32}] \tag{3.3.5}$$
$$\frac{d[DJK_{tot}]}{dt} = f_{P:DJK}(\sigma^{32}) - \mu[DJK_{tot}] \tag{3.3.6}$$

All protein states of the spontaneous folding protein type:

$$\frac{d[uP_{sf}]}{dt} = (1-\alpha)C_p - (k_{deg_{uP}}[Lon] + \mu)[uP_{sf}] - k_F(T)[uP_{sf}] + k_U(T)[nP_{sf}] \cdots$$
$$- k_{B_{D:uP}}[DJK][uP_{sf}] + (k_{D_{D:uP}} + k_{F_{DJK}})[DJK : uP_{sf}] \tag{3.3.7}$$
$$\frac{d[nP_{sf}]}{dt} = k_F(T)[uP_{sf}] - (k_U(T) + \mu)[nP_{sf}] + k_{F_{DJK}}[DJK : uP_{sf}] \tag{3.3.8}$$

Complexes:

$$\frac{d[DJK : uP_{sf}]}{dt} = k_{B_{D:uP}}[DJK][uP_{sf}] - (k_{D_{D:uP}} + k_{F_{DJK}} + \mu)[DJK : uP_{sf}] \tag{3.3.9}$$
$$\frac{[DJK : \sigma^{32}]}{dt} = k_{B_{D:32}}[DJK][\sigma^{32}] - (k_{D_{D:32}} + \mu)[DJK : \sigma^{32}] \tag{3.3.10}$$

3.4 Existence of equilibria

In order to better comprehend the properties inherent in the structure of this system, we attempt to determine the number of its equilibrium points.

I resolve the complexes:

$$
\begin{aligned}
[DJK:uP_{sf}] &= \frac{k_{B_{D:uP}}}{k_{D_{D:uP}} + k_{F_{DJK}} + \mu}[DJK][uP_{sf}] = A[DJK][uP_{sf}],\\
[DJK:\sigma^{32}] &= \frac{k_{B_{D:32}}}{k_{D_{D:32}} + \mu}[DJK][\sigma^{32}] = B[DJK][\sigma^{32}].
\end{aligned}
\tag{3.4.1}
$$

Next, I solve for total quantities at an equilibrium: From the ODE describing the change in the total protein quantity 3.3.4, we deduce:

$$
[P_{tot_{sf}}] = \frac{(1-\alpha)C_p}{\mu} - \frac{k_{deg_{uP}}}{\mu}[uP_{sf}] = C - D[uP_{sf}].
\tag{3.4.2}
$$

The equilibrium point associated to ODE defined by 3.3.5 is given by

$$
[\sigma^{32}_{tot}] = \frac{f_{P:32}(T, \sigma^{70,38,24,54})}{\mu} - \frac{k_{deg_{32}}}{\mu}[FtsH_{tot}][DJK:\sigma^{32}].
\tag{3.4.3}
$$

It remains to replace $[DJK:\sigma^{32}]$ by its expression at the equilibrium, i.e. Equation 3.4.1, in order to deduce that

$$
[\sigma^{32}_{tot}] = \tilde{f}_{P:32}(T, \sigma^{70,38,24,54}, \mu) - E[DJK][\sigma^{32}]
\tag{3.4.4}
$$

with $\tilde{f}_{P:32}(T, \sigma^{70,38,24,54}, \mu) = f_{P:32}(T, \sigma^{70,38,24,54})/\mu$.

Finally, the equilibrium point of ODE defined by 3.3.6 is easily obtained and given by

$$
[DJK_{tot}] = \frac{f_{P:DJK}(\sigma^{32})}{\mu} = \tilde{f}_{P:DJK}(\sigma^{32}, \mu).
\tag{3.4.5}
$$

At the equilibrium, by using Equation 3.3.8, the native protein is expressed as a function of free unfolded protein and DnaJK concentrations:

$$
\begin{aligned}
[nP_{sf}] &= \frac{k_F(T)}{k_U(T)+\mu}[uP_{sf}] + \frac{k_{F_{DJK}}}{k_U(T)+\mu}[DJK:uP_{sf}]\\
&= \frac{k_F(T)}{k_U(T)+\mu}[uP_{sf}] + A\frac{k_{F_{DJK}}}{k_U(T)+\mu}[DJK][uP_{sf}]\\
&= F(T)[uP_{sf}] + G(T)[DJK][uP_{sf}]
\end{aligned}
$$

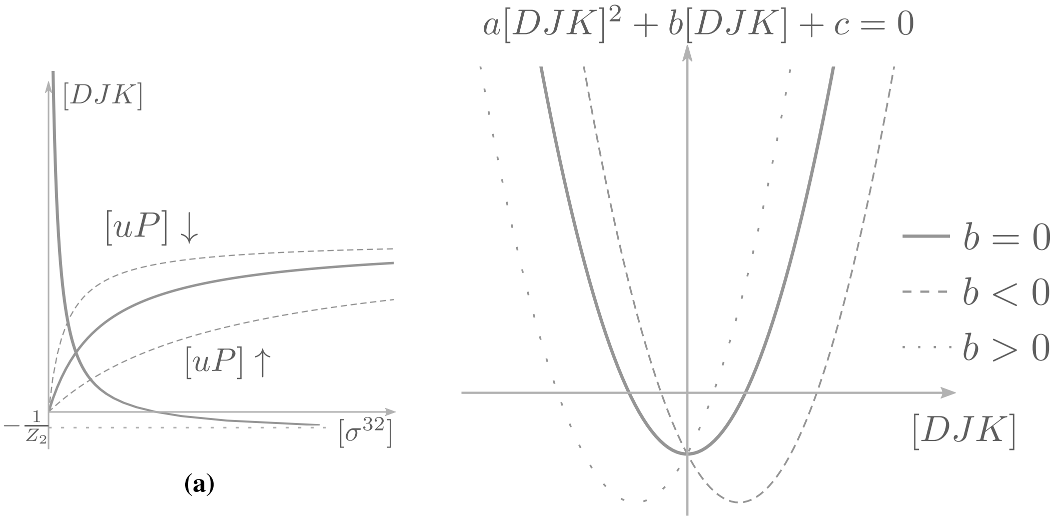
All the introduced constants are listed in Table 3.4.1.
By introducing the obtained expressions into the mass balance constraints, I obtain:

$$
\begin{aligned}
[uP_{sf}] &= \frac{C}{1+D+F(T)+(A+G(T))[DJK]},\\
[\sigma^{32}] &= \frac{\tilde{f}_{P:32}(T, \sigma^{70,38,24,54}, \mu)}{1+(B+E)[DJK]},\\
[DJK] &= \frac{\tilde{f}_{P:DJK}(\sigma^{32}, \mu)}{1+A[uP_{sf}]+B[\sigma^{32}]}.
\end{aligned}
\tag{3.4.6}
$$

Table 3.4.1: Definition of various constants.

$$
\begin{aligned}
A &= \frac{k_{B_{D:uP}}}{k_{D_{D:uP}} + k_{F_{DJK}} + \mu}, & B &= \frac{k_{B_{D:32}}}{k_{D_{D:32}} + \mu} \\
C &= \frac{(1-\alpha)C_p}{\mu}, & D &= \frac{k_{deg_{uP}}}{\mu} \\
E &= B\frac{k_{deg_{32}}}{\mu}[FtsH_{tot}], & F(T) &= \frac{k_F(T)}{k_U(T) + \mu},
\end{aligned}
$$

$$
G(T) = \frac{k_{F_{DJK}}}{k_U(T) + \mu} \times A = \frac{k_{F_{DJK}}}{k_U(T) + \mu} \times \frac{k_{B_{D:uP}}}{k_{D_{D:uP}} + k_{F_{DJK}} + \mu}.
$$



(a)

(b)

Figure 3.4.1: **(a)**: Intersection of nullclines defined by 3.4.8 and 3.4.9 under the assumption that $[uP]$ is constant. Qualitative inspection of the curves shows that they can intersect at a single point, thus ensuring the existence of a single equilibrium point. **(b)**: Qualitative illustrations of the possible curves defined by Equation 3.4.10, where the intersections of the curve with the x axis define a value of $[DJK]$ at the equilibrium point. One can see that regardless of the value of the parameter b (3.4.12), for all values of $[\sigma^{32}] > 0$, we show that $[DJK]$ will have a single positive value, thus defining a single equilibrium point.

Function $f_{P:DJK}([\sigma^{32}])$ is assumed to be a monotonously growing function of σ^{32}, and for the simplicity of calculation, it is further assumed to be a linear function $\hat{Y}_1(\sigma^{32}) = Y_1\sigma^{32}$. I introduce new constants for the purpose of simplification, and write:

$$[uP] = \frac{X_1}{X_2 + X_3[DJK]} \tag{3.4.7}$$

$$[DJK] = \frac{Y_1[\sigma^{32}]}{1 + Y_2[uP] + Y_3[\sigma^{32}]} \tag{3.4.8}$$

$$[\sigma^{32}] = \frac{Z_1}{1 + Z_2[DJK]} \tag{3.4.9}$$

The expression for $[DJK]$ seen as a function of $[\sigma^{32}]$ is of Michaelis-Menten type, and is thus a monotonously growing function of $[\sigma^{32}]$. The expression for $[\sigma^{32}]$ is monotonously decreasing with $[DJK]$. In such a case, for a fixed $[uP]$, there is a single crossing that marks the equilibrium point of the system, as illustrated in Figure 3.4.1a.

If we input the expression for $[uP]$ into the expression for $[DJK]$, we get a second degree polynomial of $[DJK]$ of a shape:

$$a[DJK]^2 + b[DJK] + c = 0 \tag{3.4.10}$$

where

$$a = X_3(1 + Y_3)[\sigma^{32}] = \theta_1[\sigma^{32}] \tag{3.4.11}$$

$$b = X_2 + X_1Y_2 + (X_2Y_3 - X_3Y_1)[\sigma^{32}] = \theta_2 + \theta_3[\sigma^{32}] \tag{3.4.12}$$

$$c = -X_2Y_1[\sigma^{32}] = -\theta_4[\sigma^{32}] \tag{3.4.13}$$

with all constants $X_1, X_2, X_3, Y_1, Y_2, Y_3, Z_1, Z_2 > 0$, and $\theta_1, \theta_2, \theta_4 > 0$.
This equation has two solutions:

$$[DJK] = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3.4.14}$$

or, in the expanded form:

$$[DJK] = \frac{-(\theta_2 + \theta_3[\sigma^{32}]) \pm \sqrt{(\theta_2 + \theta_3[\sigma^{32}])^2 + 4\theta_1\theta_4[\sigma^{32}]^2}}{2\theta_1[\sigma^{32}]} \tag{3.4.15}$$

To determine the type of solutions possible for Equation 3.4.15, we first look at the discriminant. The constants θ_1, θ_2 and θ_4 are always positive. θ_3 can assume any value, but it is present only under a square, and thus the expression under the root is always positive and greater than b. Since $[DJK]$ is necessarily non-negative, only one of the two solutions is acceptable (the positive one). This is illustrated in Figure 3.4.1.

By showing that this solution is a strictly non decreasing function of $[\sigma^{32}]$, we know that the two nullclines described by the expressions of $[DJK]$ and $[\sigma^{32}]$ can intersect only at a single point, which would define the equilibrium of the system. This is more difficult to show graphically because $[\sigma^{32}]$ appears in all the parameters of the quadratic equation (3.4.15). Therefore, to show that the $[DJK]$ is a monotonous function of $[\sigma^{32}]$, I show that the derivative $\partial[DJK]/\partial[\sigma^{32}]$ is non negative for $[\sigma^{32}] \in \mathbb{R}_{\geq 0}$

$$\frac{\partial[DJK]([\sigma^{32}])}{\partial[\sigma^{32}]} = \frac{-\theta_2\left(\theta_2 + \theta_3[\sigma^{32}] - \sqrt{(\theta_2 + \theta_3[\sigma^{32}])^2 + 4\theta_1\theta_4[\sigma^{32}]^2}\right)}{2\theta_1[\sigma^{32}]^2\sqrt{(\theta_2 + \theta_3[\sigma^{32}])^2 + 4\theta_1\theta_4[\sigma^{32}]^2}} \tag{3.4.16}$$

Since I have already shown that $\sqrt{(\theta_2+\theta_3[\sigma^{32}])^2+4\theta_1\theta_4[\sigma^{32}]^2}$ is positive and larger than $b=\theta_2+\theta_3[\sigma^{32}]$ Therefore, the value $\left(\theta_2+\theta_3[\sigma^{32}]-\sqrt{(\theta_2+\theta_3[\sigma^{32}])^2+4\theta_1\theta_4[\sigma^{32}]^2}\right)$ will always be negative. Since θ_2 is positive, the expression in the numerator will always be positive. I have shown that for $[\sigma^{32}] \in \mathbb{R}^+$, $[DJK]$ is a strictly non decreasing function of $[\sigma^{32}]$, and therefore the expressions Equation 3.4.15 and Equation 3.4.9 can intersect at a single point which is the equilibrium point of the system.

I thus show that the system described by equations (Equation 3.3.1-3.3.3) under the set of assumptions listed in the beginning of section 3.3 has a unique equilibrium point. What is important to note is that the simplified system for which we have shown this property numbers 7 (in general) nonlinear ODEs and 3 algebraic equations. Keeping in mind that even the existence of solutions cannot be assumed for any nonlinear dynamical system, a single equilibrium point for the entire parameter space is indeed a very strong property.

3.5 System behavior at the equilibrium point

In order to analyze the behavior of the system at the equilibrium points by observing the influence of variables and parameters, I refer to Equation 3.4.6 and rewrite it expanding the parameter expressions:

$$\begin{aligned}
[uP_{sf}] &= \frac{1}{\mu}\frac{(1-\alpha)C_p}{1+\frac{k_{deg_{uP}}}{\mu}+\frac{k_F(T)}{k_U(T)+\mu}+\frac{k_{B_{D:uP}}}{k_{D_{D:uP}}+k_{F_{DJK}}+\mu}(1+\frac{k_{F_{DJK}}}{k_U(T)+\mu})[DJK]} \\
[DJK] &= \frac{\tilde{f}_{P:DJK}(\sigma^{32},\mu)}{1+\frac{k_{B_{D:uP}}}{k_{D_{D:uP}}+k_{F_{DJK}}+\mu}[uP_{sf}]+\frac{k_{B_{D:32}}}{k_{D_{D:32}}+\mu}[\sigma^{32}]} \\
[\sigma^{32}] &= \frac{\tilde{f}_{P:32}(T,\sigma^{70,38,24,54},\mu)}{1+\frac{k_{B_{D:32}}}{k_{D_{D:32}}+\mu}(1+\frac{k_{deg_{32}}}{\mu}[FtsH_{tot}])[DJK]}
\end{aligned} \tag{3.5.1}$$

The effect of chaperones. Let us first remember that the inactivation of σ^{32} happens due to chaperoning by both DnaJK-GrpE and GroELS systems. Therefore, what is represented by $[DJK]$ variable in our system is in fact the action of all the chaperones that act on the sigma factor. If we imagine a null mutant of one of these chaperones, that would correspond to the low amount of $[DJK]$ in our model and could be integrated into the model by reducing the Y_1 parameter (3.4.8). Such a change would result in a new equilibrium point with a higher level of unfolded protein $[uP]$ and $[\sigma^{32}]$, indicating a higher level of heat shock protein synthesis, as is the case in cells [131]. Unfortunately, the situation in which all the chaperones which regulate σ^{32} levels are knocked out is impossible to accomplish due to essentiality of GroELS in all growth situations. Our model predicts that in this case, an equilibrium point would exist, determined solely by (apart from folding and unfolding dynamics for the unfolded protein) the rate of dilution:

$$[uP_{sf}] = \frac{\frac{(1-\alpha)C_p}{\mu}}{1+\frac{k_{deg_{uP}}}{\mu}+\frac{k_F(T)}{k_U(T)+\mu}} \tag{3.5.2}$$

$$[DJK] = 0 \tag{3.5.3}$$

$$[\sigma^{32}] = \frac{f_{P:32}(T,\sigma^{70,38,24,54})}{\mu} \tag{3.5.4}$$

under the assumption that the production of σ^{32} is constant. An excessive production of the chaperones, on the other hand, would result in lower amounts of the unfolded protein and lower levels of σ^{32} at the equilibrium point, as has been shown experimentally [131].

rpoH null mutant. I further analyze the result of eliminating σ^{32} from our system by assuming that $f_P(\sigma^{70,38,24,54}, T) = 0$, and thus simulating the situation of an *rpoH* null mutant. At the equilibrium point of the system, as it is described by Equation 3.5.1, the effect of setting $[\sigma^{32}] = 0$ would result in zero levels of chaperone, and therefore increased level of unfolded protein. In the real heat shock regulation system, some chaperones are produced under the influence of the housekeeping sigma factor, so the actual chaperone level would not be zero. However, the *rpoH* null mutants do result in low levels of chaperones and impossibility of growth at temperatures over $20°C$ [166].

Degradation by FtsH. The parameter $[FtsH_{tot}]$ appears solely in the expression for σ^{32}, and is bundled in the constant Z_2 (Equation 3.4.9). Since I have shown that under the assumption of positivity for all constants the system will always have a single equilibrium point, and since this parameter will be positive even if $[FtsH_{tot}] = 0$, I conclude that the system will maintain this property in the absence of FtsH. This has been experimentally confirmed in [106], as well as shown in simulations of a heat shock model of similar structure in [147].

Protein folding. Assuming that the folding of proteins is completely prevented, both spontaneously and through the action of the chaperone ($k_F(T) = 0$, $k_{F_{DJK}} = 0$), the expression for the unfolded protein reduces to:

$$[uP_{sf}] = \frac{\frac{(1-\alpha)C_p}{\mu}}{1 + \frac{k_{deg_{uP}}}{\mu} + \frac{k_{B_{D:uP}}}{k_{D_{D:uP}}+\mu}[DJK]} \tag{3.5.5}$$

while the expressions for $[DJK]$ and $[\sigma^{32}]$ do not change significantly. One can see that because of the binding of uP to DJK, the structure of the system can still be represented as in Equation 3.4.7. Consequently, as the positivity for all constants still holds, the system retains the same property of a single equilibrium point.

Growth rate. The equilibrium point of the system under analysis exists only for the positive growth rate $\mu > 0$. At zero growth rate, this system no longer has an equilibrium point (remembering that the expressions for $\tilde{f}_{P:DJK}(\sigma^{32})$ and $\tilde{f}_{P:32}(T, \sigma^{70,38,24,54})$ have μ in the denominator (see Equation 3.4.5 and 3.4.4). Change in the growth rate has no simple effect on the system, but instead depends on the exact parameter values.

3.6 Discussion

Whenever a model of a certain phenomenon is presented, it will have its limits, and its benefits. The benefits need to fit the purpose at hand, while making sure that the limits are not important enough for the situation of interest so as to limit the usefulness of the model.

3.6.1 New understanding of heat shock regulation

The regulation of HSR in *E. coli* has been studied for half a century, and still, many details remain unclear, such as the mechanism of inactivation of σ^{32} by chaperones, the role of the chaperones in its degradation, or the reason behind its degradation depending on the only essential and membrane-anchored protease.

There have been new discoveries in the regulation mechanism that have not been included in this thesis. Here, I shortly review the new nuances of the heat shock regulation that have been uncovered in the last 10 years.

In 2013, Lim et al. [167] have shown that σ^{32} is targeted to the membrane by the co-translational SRP (Signal Recognition Particle) mechanism. This discovery added weight to the assumption that σ^{32} regulatory mechanism monitors the state of the membrane proteome via its degradative regulation by FtsH. Additionally, *rpoH* gene is exactly downstream of the operon from which *ftsY* - signal recognition particle receptor - is transcribed. As σ^{32} is a transcriptional regulator of the *ftsX-ftsE-ftsY* operon, its physical proximity might be significant [168, 169].
FtsH has poor unfoldase activity, and degrades σ^{32} very slowly *in vitro* [135]. It was initially assumed that the action of the chaperones introduces a conformational change in σ^{32} which makes it more susceptible to degradation. Researchers have been unable to confirm this, as the addition of chaperones did not speed up the *FtsH*-mediated degradation of σ^{32} *in vitro* [170]. It was recently suggested that other ubiquitin-like modifier proteins might play a role in σ^{32} modification prior to its degradation [5] [172].

3.6.2 Growth condition

The regulatory mechanism in place to maintain a properly folded proteome in *E. coli* that has been studied here is the one that gets activated when *E. coli* is grown aerobically in a laboratory. The question can rightly be asked if the same mechanism is responsible for the proteome maintenance task in the anaerobic environment, more typical for *E. coli*, in which the long periods of stationary phase existence are interspersed with periods of slow growth.
The question has been posed whether heat shock as manifesting in the aerobic laboratory conditions is not to a great extent caused by oxidative stress. In fact, it has been shown that in anaerobic conditions, the *rpoH* null mutant is not impaired in dealing with heat stress [173], meaning that a regulatory mechanism other than the one governed by σ^{32} is in control.
Be it so, the growth condition of aerobic exponential growth is still relevant for this study, because the UPR (as a subset of the UPR) during recombinant protein production happens mostly exactly under such laboratory growth conditions.

3.6.3 ODEs, equilibrium points and biological systems

Use of ODEs to model the reactions inside living cells arose as a result of understanding that cells are 'chemical factories' bereft of any *elan vital*, whose inner workings can be understood in great part as chemical reactions. This understanding brought about the possibility to model the cellular systems with already existing empirical mathematical representations for describing chemical reactions, also known as the 'mass action law' (see section 3.1). The mass action law was derived for chemical reactions happening in well mixed containers with both substrates and products present in great numbers. Although such a chemical reaction can be considered 'noisy' [6] , the mixedness of the container and the great number of particles involved allows the system to be well represented by a number of continuous variables - *concentrations*, whose change can also be considered continuous, and which can be represented by ODEs. The cell, however, often does not resemble a well mixed container in which all the interacting species are numerous enough to be considered concentrations. For example, a single mRNA can support production of hundreds of proteins. Whether 5 or 9 copies of a particular mRNA exist will have great influence on the protein copy number. Such noise present at the very core of cellular functioning (protein production) makes it doubtful whether the cellular species can be represented well by ODE models.

---

[5] It has been suggested that there are proteins in bacteria which have a role in 'tagging' damaged proteins for degradation, as do the ubiquitin proteins in eukaryotes, and are called Pup - prokaryotic ubiquitin-like proteins [171].

[6] From the point of view of observing the occurrences of single reactions.

Given this situation, we could limit our modeling attempts to cellular populations. Since most observations and measurements for bacteria have been done at population level, at which this kind of noise becomes negligible due to averaging, it might be proper to say that our models describe entire populations. This indeed might be a good argument for open-loop systems [7] in which there is a cascade of effect in one direction. However, the system modeled here uses feedback to implement control. Since feedback present in our system can be said to act only on the level of a single cell, and in no way on the population scale, we again seem to be at an impasse.

However, numerous studies in systems biology have seen that ODE models often do recover important behavior inherent in the systems under study (keeping in mind that these systems greatly differ from one another and encompass a great variety of characteristics). This empirical appropriateness in face of theoretical unsuitability makes for an interesting puzzle, one which poses an interesting question: What does *representability* mean in the case of cellular systems and ODEs? ODE systems are rich mathematical objects. Can all the conclusions reached by their analysis and simulation be applied to the cellular system at hand?

One possible answer (on an informal level) could be that ODE models represent well what happens in single cells not at the level of concentration, but at the level of concentration means as averaged over a population. If this were the case, the transient behavior of the ODE system might not necessarily carry such a great significance, while the behavior at the equilibria could be meaningful for understanding of the biological system. Following this line of reasoning, the presented work focused exactly on this part of system analysis. Moreover, since the actual system in the cell is highly stochastic, the stability analysis of our ODE representation is not very meaningful. In a deterministic setting, the multiple feedback systems often exhibit oscillations. It may be, for example, that the cellular noise has a *dithering* function to augment stability of the system [174]. Even if this last assumption is not really the case, the relation between the actual stochastic system in the cell and an ODE representation is by no means a trivial one, and there is no simple and general way to see if the stability of one indicates anything about the stability of the other.

3.6.4 Modeling protein production

When modeling HSR, or UPR, we are modeling a systemic stress response which influences some of the core cellular processes - such as production of protein. The rate at which proteins are produced in a cell is tightly controlled, balancing the costly process with the energetic and metabolic output capacity of the cell. When stress is induced, this rate will almost certainly change. As the temperature increases, a cell undergoes several important changes: diffusion of molecules in the cell changes, efficiency of chemical reactions increases, and the stability of proteins and other macromolecules decreases. The first two changes have positive impact on the growth capacity of the cell, while the other two represent a burden to cellular growth. The balance of such effects will determine the rate of growth of the cell, and thus the rate of protein production.

I have, however, represented the growth rate, which is also the rate of dilution of all the cellular components - μ - as a constant. Considering all that has been stated above, that means to misrepresent the actual state of affairs. It is possible to circumvent this issue by assigning an algebraic expression for the growth rate, tying it to some other rate in the cell, such as the rate of protein production, as was done in [175]. In the aforementioned model, the change in growth was due to change in available nutrient, which influences the rate of precursor and energy production, which in turn influences the rate of protein production.

---

[7] Open-loop systems are control systems which use no feedback, unlike the closed-loop systems.

This is of course possible, but would introduce a number of problems to our analysis. First, I would need to extend the model greatly, introducing further functional divisions into the protein pool so as to make protein production a function of the cellular state, for which I would need to include ribosomes, metabolic enzymes and potentially mRNAs of all the protein species. Then, a decision would be necessary as to how to make the model sensitive to the cellular state. In [175], the cell model adapts to the metabolic state, which is a direct consequence of the external metabolite concentration and transporter and metabolic enzyme efficiency. Here, the changes in the cell state and the growth rate should be a consequence of changes in metabolic enzyme efficiency and protein stability. While it is possible to make the necessary modifications in the line of those proposed in [175], the resulting model could only be simulated, and not analyzed, which would defeat its purpose for this study.

However, the line of reasoning highlights an important consideration - the growth rate is a consequence of a cellular state. How the growth rate is determined in the cell, and how this can be modeled and simulated will be the focus of the rest of this thesis.

II

whole-cell model of *Escherichia coli*

4 Resource Balance Analysis 71
4.1 Conceptual ideas behind RBA
4.2 Formulation of the RBA problem
4.3 RBA exemplified on a *toy* model
4.4 Related modeling paradigms
4.5 **RBApy** software

5 An RBA model of *E. coli* 93
5.1 Parameterization of the *E. coli* RBA model
5.2 Results
5.3 Discussion

In the discussion of the previous chapter, I have mentioned some difficulties in modeling systemic stress responses with the ODE formalism. If aiming at a good representation of a *coherent cellular state* (see section 1.5), one needs to introduce additional complexities and nonlinearities into an ODE model, which (almost certainly) exclude the possibility of model analysis. The advantage of such a model extension is to make the behavior of the modeled cell more realistic. However, for bioengineering and bioproduction needs, the possible level of detail of an ODE model is not sufficient. Even if there were no theoretical limit to the size of an ODE model, the practical issues of model parameterization and simulation soon come to the foreground as the model increases in size and complexity. For this reason, most ODE models are limited in scope to the analysis of a relatively small regulatory system in the cell. Exceptions do exist [176], but they are not common. Also, due to the mentioned limitations, ODE models have not had great success in bioengineering and bioproduction industries. These industries, however, have more readily accepted the simpler (in terms of parameterization and simulation) constraint-based genome-scale models, which allow for some useful *in silico* exploration of the metabolism.

In contrast, genome-scale metabolic models could not satisfy the research needs of this work, since they leave out the aspect most important to this work: the production, folding and assembly of proteins. However, in the last 15 years, a number of research groups has worked to bridge this gap, introducing modeling frameworks which allow for representation of the entire cellular metabolism while taking into account the macromolecular composition of the cell in some ways [28, 29, 177, 178]. Of the paradigms listed, Resource Balance Analysis (RBA) is the paradigm most suited for taking into account all the cellular processes relevant for protein production. RBA approach is based on the idea of *coherency* of the cellular state, an issue of great importance for this thesis. Therefore, in the next two chapters I explain the ideas behind and the formulation of the RBA problem, and I develop, parameterize and validate a whole-cell RBA model of *Escherichia coli*.

4. Resource Balance Analysis

> "What is it that you've learned, what you're able to do?"
> "I can think. I can wait. I can fast."
> "That's everything?"
> "I believe, that's everything!"
> "And what's the use of that? For example, the fasting– what is it good for?"
> "It is very good, sir. When a person has nothing to eat, fasting is the smartest thing he could do.""
>
> Hermann Hesse, Siddhartha

4.1 Conceptual ideas behind RBA

Allocation of resources could well be the crowning concept of modernity. It is a concept used over an incredibly broad range of human activities: ranging from energy distribution, to business planning all the way to self-help literature advising individuals on how to treat their most precious resource: time (and/or money). Hand in hand with the allocation of resources comes the idea of *optimizing* the allocation of resources. RBA deals with this idea on a scale of bacterial growth. I would now like to provide an image useful for understanding the reasoning behind RBA. This image is not accurate as it implies bacteria having a will which guides them to acquire certain properties, it is simply useful as an illustration. Let us for a moment imagine a life of a fast growing bacteria in a rich medium. Each cell is busy reproducing itself with a certain doubling time [1]. The food is plenty, the space is ample. But the bacteria has learnt throughout millennia that this kind of situation will not last forever. If it does not manage to grow fast, the food will be eaten by others. The bacteria can choose either to influence the environment to prevent such occurrence, or to modify itself to better be able to utilize the available resources. Here we will consider the second possibility. If they choose to direct their attention inwards, they will modify their internal regulation to help them attain the best possible internal state to outcompete other bacteria. This will be the state in which they attain the highest possible growth rate, which at the same time will be the highest rate at which they are able to secure the food resources for themselves. They can perform internal reorganization - allocate more resources to ribosomes, for example, but then some metabolic precursors become scarce. They could allocate more resources to metabolism, but then there are not enough ribosomes to produce all the necessary metabolic enzymes. Finally, they develop a complex and robust regulation scheme which allows them to achieve the "sweet spot" of resource allocation, in which each cellular process is allocated just the right amount of resources to help the cell achieve higher growth rate. Of course, the cells also know that if some kind of stress comes their way, they need to be ready to change their internal configuration into a state that is more adapted to survival than to growth. This means that cells will never be fully optimized for fast growth, but I will show that this can be a reasonable

---

[1] The doubling time is the time a single cell requires to grow to the double of its size right after the last doubling.

assumption in certain situations, which already allow us to study the cellular configuration in detail.

4.1.1 Resource allocation and the growth rate

Monod's discovery on how the growth rate of bacteria depends on the external concentration of the sole carbon source [179] lead to posing of many questions such as: what limits the maximum growth rate of bacteria and do bacteria in fact aim to achieve a maximum possible growth rate or not.

The appealing idea of optimality was used very early on to explain the growth of bacteria. The model of cell growth as an optimal process proposed in 1966 managed to predict the lag phase during batch growth of a bacterial culture by assuming that bacteria aim to optimize the accumulation of biomass at the moment of depletion of the growth substrate [180].

The phenomenon of *diauxie* characterized by Monod [179] led to additional questions. *Diauxie* occurs in some batch growth situations when cells are presented with two carbon sources in the medium, but do not use them simultaneously. The bacteria first deplete the medium of the one *preferred* carbon source, after which they switch to the second one. In the work that eventually got them the Nobel Prize in Physiology, Monod and Jacob showed how diauxie is achieved in the case of glucose and lactose, with the glucose being the preferred carbon source [181] through a regulatory mechanism known today as *catabolite repression*. They showed that glucose acts as a repressor on the synthesis of the proteins responsible for uptake and catabolysis of lactose. While this showed how diauxie is implemented in the cell, the question of *why* this happens still remained. First model to attempt at an explanation of *diauxie* was proposed in 1984 [182]. This work proposed that such diauxic behavior is a consequence of "judicious investment of cellular resources in synthesizing different key proteins according to an optimal regulatory strategy" and that the need for optimal investment of resources lies in their limited availability. The proposed criterion of optimization is the maximization of cell mass productivity for the time period in which at least one growth substrate is present in the medium. The problem with such a hypothesis is that the cell has no way of predicting when exactly the medium will be depleted, and cannot possibly adjust its behavior in the beginning of log phase according to what would be optimal for the entire duration of growth. In their follow-up work, the same group explores the consequences of assuming that instead of the previously used "long-term perspective", the cell optimizes according to a "short-term perspective" [183].

The first, "long-term perspective" criterion used - that of maximization of biomass during the entire period of growth - would be to assume that bacteria maximize their *yield*. Although different definitions appear in the literature, here I define yield as the amount of biomass produced per substrate consumed. The second, "short-term perspective" criterion of maximization of the instantaneous rate of biomass accumulation is what is known today as *growth rate* maximization.

In the following years, the growth rate was shown to have an unexpected relation to the internal state of the bacterial cell. Namely, in their work on the composition of the bacterium *Salmonella typhimurium* under different growth conditions, Schaechter, Maaløe and Kjeldgaard showed that cellular mass, RNA and DNA content can be described as a function of the growth rate [184]. Already in 1928, Henrici reported that the size of the bacterial cell changes with the rate of its growth [185]. Bremer and Dennis [186] performed a detailed study over a wide range of growth rates and found that certain ratios of macromolecules in the cell (like the ratio of RNA to protein) show linear dependency on the growth rate. This and other related works suggested that such characteristics of cells are not directly related to the medium on which they are growing, but to the growth rate which the cells obtain in a certain medium.

These results touched upon an ongoing debate in the field: What determines the rate of growth of bacterial cultures given a certain medium and a growth condition - the rate of ATP synthesis, the efficiency of the metabolism to produce precursors and energy, or the rate of synthesis of biopolymers? Obviously, there exists a mechanism to attune the biopolymer synthesis rate to metabolic efficiency, as the rate of translation is attuned to the availability of charged tRNAs. One of the first attempts to model the rate of growth as a result of interaction of a choice of cellular processes was done by Marr in 1991 [187].

RBA is a cell modeling framework that captures the basic cellular resource limitations and under the assumption of growth maximization, predicts the metabolic fluxes, enzyme, ribosome, chaperone and other process machinery concentrations for a particular growth medium. This assumption of growth rate optimization, although not always correct, can be assumed valid in a number of growth situations [188]. Often the assumption of optimal growth is not true if the bacteria have not been evolutionarily pressured to optimize growth on a certain substrate. If such pressure is exerted in a laboratory, they can adapt and reach growth rates very close to those predicted experimentally under the assumption of growth rate maximization [189].

4.1.2 Concepts in RBA setting

RBA [29] is a modeling paradigm that enables the modeler to systematically take into account the cost of different cellular processes and analyze their optimal balance under the assumption of growth rate maximization. It models the growth of an average cell in a batch culture population that is exponential growth phase. The modeling formalism assumes that the cell is at steady state, growing at the rate μ. The rate of growth μ is the rate of expansion of the cellular volume:

$$\frac{dV}{dt} = \mu V \tag{4.1.1}$$

which can also be expressed as $\ln 2/T_d$, where T_d is the average doubling time within the population [2]. RBA models metabolism and a set of cellular processes all taking place within the confines of cellular compartments of limited space (see Figure 4.2.1). Metabolism can be modeled at the desired level of detail (from simple to full genome-scale metabolic reconstructions). The modeled cellular processes will always include protein translation due to its central importance in cellular resource distribution, but can also include other processes such as protein folding, secretion, or anything that is of interest to the modeler. All of these processes are facilitated by the so-called "molecular machines" - for metabolism the enzymes, for translation the ribosomes, for protein folding the chaperones, etc. In order to be built, the molecular machines require cellular resources (in terms of energy, precursors and process molecular machines), and once they are built, they occupy cellular space. To be able to compute precisely the amount and type of cellular resources required for their construction and the amount of space they will take up, RBA uses the information about the exact macromolecular composition of all molecular machines. The abundance and capacity of these molecular machines limit the fluxes in the cell: in the sense in which the number of available ribosomes limits the flux of production of protein, and the abundance of a metabolic enzyme limits its respective metabolic flux. As already stated, RBA allows the model designer to decide the level of detail of the model. However, at our present state of knowledge, even the most detailed model will leave something out - a portion of the cellular proteome will be unrepresented. Those proteins can be assigned to the so-called non-enzymatic "housekeeping protein" pool which the cell needs to produce.

---

[2] This can be easily computed from the respective ODE, assuming that the time it takes for the volume to grow from V_0 to $2V_0$ is T_d.

This modeling approach falls into the family of constraint-based models from the point of view of mathematical formulation and into the family of cell models regarding its scope. On a black box level, it requires (i) a metabolic reconstruction annotated with proteins associated to each reaction, (ii) amino acid sequences of the associated proteins, (iii) a total concentration of protein the cell has at each growth rate and (iv) specification of modeled cellular processes and their associated process machines. It predicts the metabolic fluxes and concentrations of all molecular and process machines at the maximum obtainable growth rate (which is also a prediction of the model). Let us now see in more detail how such an idea can be translated into a mathematical object.

4.2 Formulation of the RBA problem

The full conceptual and mathematical formulation of the RBA is presented in [29], while the conceptual aspects are further elaborated in [190]. RBA describes a cell at steady state as a set of related reaction fluxes, enzyme and cellular machine concentrations and the rate of growth. This description is put forth by the flexible establishing of relations between the general constraints of the cell and those imposed by the rate at which the cell is growing. General constraints are those that the cell "needs to live with", regardless of the rate at which it is growing. These are: (i) stoichiometry of the metabolic network, (ii) composition of metabolic enzymes and process machines (ribosomes, chaperones), (iii) cost of cellular processes and (iv) limitation of available cellular space. The constraints related to the growth rate concern the need to produce intracellular species at a flux which counteracts their dilution. The bacterial colony is said to grow at a growth rate μ if its population dynamics can be described by the following equation:

$$\frac{dV(t)}{dt} = \mu(t)V(t) \tag{4.2.1}$$

where $V(t)$ is the total volume of the cells in the colony. The steady-state assumption corresponds to the biological situation of the so-called *balanced growth* regime during which the population grows exponentially and the cellular internal composition is constant, within the limits of cellular noise [184, 186]. The change in concentration of species x in volume V can be described as:

$$\frac{dC_x(t)}{dt} = \frac{d}{dt}\frac{N_x(t)}{V(t)} = \frac{1}{V(t)}\frac{dN_x}{dt} - \frac{N_x(t)}{V(t)}\frac{1}{V(t)}\frac{dV(t)}{dt} = 0 \tag{4.2.2}$$

At steady state, the change in concentration is zero. Combining this with the expression for growth rate in Equation 4.2.1, one obtains the following expression:

$$\frac{1}{V(t)}\frac{dN_x(t)}{dt} = \mu\frac{N_x(t)}{V(t)} = \mu C_x(t) \tag{4.2.3}$$

Therefore, the production flux required for maintaining the steady-state concentration C_x of intracellular species x is μC_x. This simple statement has three types of consequences within RBA. First, all macromolecular machines (enzymes, process machines) need to be produced at a rate that equals their dilution due to growth. Secondly, this, in turn, imposes a constraint on the metabolism which needs to provide precursors and energy and absorb the byproducts released in the process of their production. And thirdly, as is obvious from Equation 4.2.3, the flux of production of macromolecules depends on the growth rate. This flux of production is generally not a spontaneous process that requires solely precursors and energy, but also the related process machine (or a number of them) is required in sufficient amount so as to facilitate the necessary flux. For translation this would mean:

$$k_T R \geq \mu P_{tot} \tag{4.2.4}$$

where k_T is the translation rate, R is the concentration of the ribosome and P is the total concentration of protein in the model. In RBA, the only species represented by concentration are macromolecular species - protein, RNA and possibly DNA. Individual metabolites are not represented by a concentration, but instead only have associated stoichiometries and fluxes in all the reactions in which they take part.
As can be seen on Figure 4.2.1, an RBA model consists of the metabolism (represented by metabolic fluxes), macromolecular machines associated to all non-spontaneous reactions, a set of cellular processes and their corresponding molecular machines P, and a representation of cellular space by a notion of density D and protein concentration P_{tot}. The metabolism is represented by its stoichiometric matrix S and an associated set of metabolic enzymes E. There are some metabolites and macromolecules that are not specifically represented in the model but still need to be produced - such as mRNA, for example. For some species, one might want to impose that a certain concentration must be present an a certain growth rate, or that they need to be produced or degraded with a specific flux. Such concentrations and fluxes are called *target* concentrations and fluxes.

4.2.1 RBA constraints

Now I proceed to show how such constraints can be formalized. The type of the constraints will later lead us to the type of constraint-based optimization problem to be solved.
(C_1) **Metabolic capability constraint**
The steady state assumption imposes a constraint on the metabolism - it needs to be able to produce all the precursors and energy required for all of the synthesis fluxes. It also needs to be able to absorb all the metabolites released during the synthesis of cellular components. These constraints can be formulated as:

$$S\vec{v} + \mu(C_e\vec{e} + C_p\vec{p} + C_{t_c}\vec{t_c}) + C_{t_f}\vec{t_f} = 0 \tag{4.2.5}$$

where S is the stoichiometric matrix, v the vector of metabolic fluxes, C_e, C_p, C_{t_c} and C_{t_f} are the matrices which, for each enzyme, process machine, target species and target flux, give the stoichiometry of substrate metabolites required for and product metabolites generated by their synthesis. One typical target species is housekeeping protein. That leaves us with the unknowns $\vec{v}$, $\vec{e}$ and $\vec{p}$, which are exactly the constituents of the vector x estimated in the optimization procedure. This constraint can be written in matrix form:

$$\begin{bmatrix} S & \mu C_e & \mu C_e & \mu C_{t_c} & C_{t_f} \end{bmatrix} \begin{bmatrix} \vec{v} \\ \vec{e} \\ \vec{p} \\ \vec{t_c} \\ \vec{t_f} \end{bmatrix} = \vec{0} \tag{4.2.6}$$

(C_2) **Capacity constraints**
Capacity constraints link the abundance of a molecular machine to the flux of a reaction it can facilitate. These kinds of constraints are relevant for enzymes and process machineries.
(C_{2a}) **Process capacity constraints**

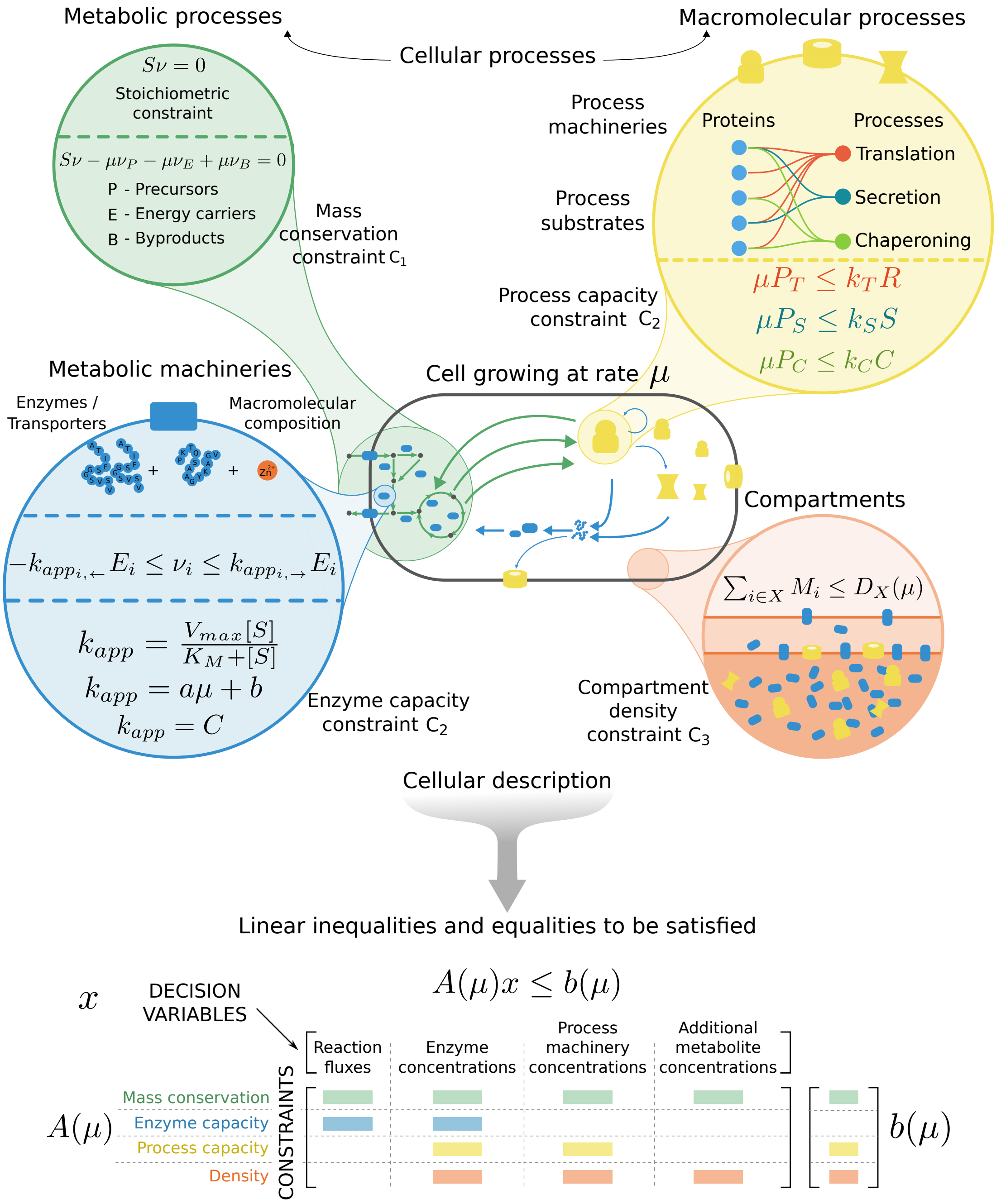


Figure 4.2.1: Resource Balance Analysis model represents the cell growing at rate μ as a set of metabolic fluxes and concentrations of macromolecular species - enzymes and process machines. The relation between them is established through a number of constraints (C_1 to C_3). These constraints formulate a linear feasibility problem for a particular value of μ. A series of LP feasibility problems are solved to find the highest μ for which the problem is still feasible.

Macromolecular processes in the RBA context are all the processes which take part in the construction of macromolecules (translation, chaperoning, transcription, etc.), and whose cost can be defined on the level of the individual constituent of the macromolecule (amino acid, RNA, etc.). The most important macromolecular process (from the resource allocation perspective) is translation, and all RBA models need to include it. As seen in Figure 4.2.1, the list of macromolecular processes can be extended to include all such processes that are of interest to the modeler. Each process definition needs to specify the process machinery p - such as ribosome, chaperone or RNA polymerase) in terms of its exact macromolecular composition. Additionally, one needs to specify the capacity of the process machinery k_P in terms of the number of individual macromolecule constituents it is able to process over time. For example, in case of the ribosome, this is the number of amino acids over time. Since not all macromolecules need all the processes in order to be produced, each process has an associated set of macromolecules. This is mathematically represented by matrices $M_x, x \in \{e, p, t_c, t_f\}$ whose entries represent the cost in terms of macromolecular machinery for all the intracellular species (enzymes, process machineries, target species) and target fluxes. As an example, each amino acid in a protein has a machinery cost of 1 for the process of translation.
The process capacity constraints can be written in the following way:

$$\mu(M_e\vec{e} + M_p\vec{p} + M_{t_c}\vec{t_c}) + M_{t_f}\vec{t_f} \leq \mathrm{diag}(k_P)\vec{p} \tag{4.2.7}$$

The corresponding matrix notation is:

$$\begin{bmatrix} 0 & \mu M_e & \mu M_p - \mathrm{diag}(k_P) & \mu M_{t_c} & M_{t_f} \end{bmatrix} \begin{bmatrix} \vec{v} \\ \vec{e} \\ \vec{p} \\ \vec{t_c} \\ \vec{t_f} \end{bmatrix} \leq \vec{0} \tag{4.2.8}$$

(C_{2b}) **Enzyme capacity constraints**
In general, in RBA one can assume that enzymes can facilitate both the forward and backward reaction. The flux of the reaction is limited by the enzyme concentration and its capacity (reactions per unit of time). Therefore, one can write:

$$-\mathrm{diag}(k_E^b)\vec{e} \leq \vec{v} \leq \mathrm{diag}(k_E^f)\vec{e} \tag{4.2.9}$$

or:

$$\begin{aligned} \vec{v} - \mathrm{diag}(k_E^f)\vec{e} \leq 0 \\ -\vec{v} - \mathrm{diag}(k_E^b)\vec{e} \leq 0 \end{aligned} \tag{4.2.10}$$

In a matrix form the expression becomes:

$$\begin{bmatrix} \Psi & -\mathrm{diag}(k_E^f) & 0 & 0 & 0 \\ -\Psi & -\mathrm{diag}(k_E^b) & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{v} \\ \vec{e} \\ \vec{p} \\ \vec{t_c} \\ \vec{t_f} \end{bmatrix} \leq \vec{0} \tag{4.2.11}$$

where Ψ is a matrix mapping reactions to enzymes.

(C_3) **Density constraint** Compartments in RBA are not only voluminous parts of the cell enclosed by membranes, but include the membranes as well. The density constraint imposes an upper bound on the macromolecular occupancy of all the compartments. Due to difficulties in computing the volume and surface occupied by macromolecules and macromolecular processes, in RBA these occupancies are expressed as concentrations of amino acids per gram of cell dry weight. This choice of density representation is guided by the fact that the most important class of macromolecules modeled in RBA are proteins. When computing the contribution of other macromolecules to the density (such as RNA), they need to be scaled so as to account for the different volume their constituent molecules (ribonucleic acids, for example) take up compared with proteins.

These contributions of individual macromolecules to the density constraints are mathematically represented by matrices W_E, W_P and W_{T_C}. Since an RBA model can have multiple compartments, a density constraint needs to be defined for all of them. Therefore, the number of rows of the W matrices will correspond to the number of compartments:

$$W_e\vec{e}+W_p\vec{p}+W_{t_c}\vec{t_c}\leq\vec{D} \tag{4.2.12}$$

or in matrix notation:

$$\begin{bmatrix}0 & W_e & W_p & W_{t_c} & 0\end{bmatrix}\begin{bmatrix}\vec{v}\\ \vec{e}\\ \vec{p}\\ \vec{t_c}\\ \vec{t_f}\end{bmatrix}\leq\vec{D} \tag{4.2.13}$$

4.2.2 The linear programming problem

All the aforementioned constraints in matrix form have the same vector of unknowns. This makes it is easy to "stack up" the matrices that multiply them from the left side into the complete matrix of constraints that formalizes the RBA problem.

$$\begin{array}{c} \\ C_1\\ C_{2a}\\ C_{2b}^f\\ C_{2b}^b\\ C_3\end{array}\begin{array}{c}\begin{array}{ccccc}\vec{v} & \vec{e} & \vec{p} & \vec{t_c} & \vec{t_f}\end{array}\\ \begin{bmatrix}S & -\mu C_e & -\mu C_p & -\mu C_{t_c} & -C_{t_f}\\ 0 & \mu M_e & \mu M_p-\operatorname{diag}(k_P) & \mu M_{t_c} & M_{t_f}\\ \Psi & -\operatorname{diag}(k_E) & 0 & 0 & 0\\ -\Psi & -\operatorname{diag}(k_E) & 0 & 0 & 0\\ 0 & W_e & W_p & W_{t_c} & 0\end{bmatrix}\end{array}\begin{bmatrix}\vec{v}\\ \vec{e}\\ \vec{p}\\ \vec{t_c}\\ \vec{t_f}\end{bmatrix}\begin{array}{c}=\\ \leq\\ \leq\\ \leq\\ \leq\end{array}\begin{bmatrix}\vec{0}\\ \vec{0}\\ \vec{0}\\ \vec{0}\\ \vec{D}\end{bmatrix} \tag{4.2.14}$$

Since $\vec{t_c}$ and $\vec{t_f}$ are not actually vectors of unknowns, but actually parameters of the model that need to be known at the time of simulation, they can be moved to the right-hand side of the equation:

$$\begin{array}{c} \\ C_1(n_m)\\ C_{2a}(n_p)\\ C_{2b}^f(n_e)\\ C_{2b}^b(n_e)\\ C_3(n_c)\end{array}\begin{array}{c}\begin{array}{ccc}\vec{v}(2n_r) & \vec{e}(n_e) & \vec{p}(n_p)\end{array}\\ \begin{bmatrix}S & \mu C_e & \mu C_p\\ 0 & \mu M_E & \mu M_P-\operatorname{diag}(k_P)\\ \Psi & -\operatorname{diag}(k_E) & 0\\ -\Psi & -\operatorname{diag}(k_E) & 0\\ 0 & W_e & W_p\end{bmatrix}\end{array}\begin{bmatrix}\vec{v}\\ \vec{e}\\ \vec{p}\end{bmatrix}\begin{array}{c}=\\ \leq\\ \leq\\ \leq\\ \leq\end{array}\begin{bmatrix}-\mu C_{t_c}\vec{t_c}+C_{t_f}\vec{t_f}\\ -\mu M_{t_c}\vec{t_c}-M_{t_f}\vec{t_f}\\ \vec{0}\\ \vec{0}\\ \vec{D}-W_{t_f}\vec{t_c}\end{bmatrix} \tag{4.2.15}$$

where n_m is the number of metabolites, and n_r the number of reactions (with $2n_r$ being the number of fluxes, taking into account the reversibility of reactions), n_e number of enzymes, n_p number of cellular processes and n_c number of compartments.
The variables of the problem are the metabolic fluxes $\vec{v}$, the concentrations of enzymes and processing machineries $\vec{E}$ and $\vec{P}$ and the growth rate value μ. The goal now is to find the maximum growth rate μ for which the problem presented in Equation 4.2.15 is still feasible. In this formulation, it is a nonlinear programming problem. The nonlinear nature of the problem would limit the usability of the framework and prevent the development of whole-cell models. However, it is possible to pose it as a linear programming (LP) feasibility problem for any value of $\mu = \mu^* \geq 0$.

$$
\begin{aligned}
&\text{find } (\vec{v} \in \mathscr{R}^{2n_r}, \vec{e} \in \mathscr{R}_+^{n_e}, \vec{p} \in \mathscr{R}_+^{n_p}) \\
&\text{subject to:} \\
&\quad \mu = \mu^*, \mu^* \geq 0 \\
(C_1):\ & -\underset{(n_m \times 2n_r)}{S}\ \vec{v} + \mu\Big(\underset{(n_m \times n_e)}{C_e}\ \vec{e} + \underset{(n_m \times n_p)}{C_p}\ \vec{p} + \underset{(n_m \times n_{t_c})}{C_{t_c}}\ \vec{t_c}\Big) + \underset{(n_m \times n_{t_f})}{C_{t_f}}\ \vec{t_f} = \underset{(n_m \times 1)}{\vec{0}} \\
(C_{2a}):\ & \mu\Big(\underset{(n_p \times n_e)}{M_e}\ \vec{e} + \underset{(n_p \times n_p)}{M_p}\ \vec{p} + \underset{(n_p \times n_{t_c})}{M_{t_c}}\ \vec{t_c}\Big) + \underset{(n_p \times n_{t_f})}{M_{t_f}}\ \vec{t_f} \leq \underset{(n_p \times n_p)}{\operatorname{diag}(k_P)}\vec{p} \\
(C_{2b}):\ & -\underset{(n_e \times n_e)}{\operatorname{diag}(k_E^b)}\vec{e} \leq \vec{v} \leq \underset{(n_e \times n_e)}{\operatorname{diag}(k_E^f)}\vec{e} \\
(C_3):\ & \underset{(n_c \times n_e)}{W_e}\ \vec{e} + \underset{(n_c \times n_p)}{W_p}\ \vec{p} + \underset{(n_c \times n_{t_c})}{W_{t_c}}\ \vec{t_c} \leq \underset{(n_c \times 1)}{\vec{D}}
\end{aligned}
\tag{4.2.16}
$$

Written this way, it is possible to "scan" the range of μ for which the problem 4.2.16 is feasible. For example, by starting from an arbitrarily large growth rate μ^* for which the problem is infeasible, it is possible to find the maximum feasible growth rate μ_{max} by binary search.

4.2.3 Enzyme and process machine efficiencies

The attainable fluxes through enzymes, transporters and process machines are related to their abundance by parameters describing their efficiencies. These efficiencies can depend on the growth rate or on the concentration of the substrate in the medium. The functions readily available in the current RBA implementation are constant, linear, Michalis-Menten and multiplication. Multiplication can be used to combine other function types.
Enzymes. In RBA, the efficiency associated to an enzyme is called an *apparent catalytic rate*. While the catalytic rate normally describes the velocity of conversion of a bound substrate to product, the apparent catalytic rate is a broader term. It takes into account all the effects which can change the rate of the conversion reaction (such as temperature, pH, concentrations of substrates), except the enzyme concentration.

$$
k_{app}^{E_i} = f(T, pH, [s_1], \ldots, [s_n]) \tag{4.2.17}
$$

In RBA, apparent catalytic rates are parameters of the model. They can be expressed either as constants or as functions of the growth rate. In [191] it was shown that the apparent catalytic rate of many enzymes shows a linear dependence of the growth rate, with most enzymes showing an increase in k_{app} with the increase in μ. The reason behind this is that as the growth rate increases, so do the substrate pools, making the apparent enzyme efficiency higher.
Transporters. In order to account for the exact composition of the medium, the efficiency of a transporter depends on the concentration of its substrate(s) in the medium. Usually, this

dependency is expressed as a Michaelis-Menten function of the external substrate concentration. For a transporter T_i involved in an exhange reaction of a single external substrate s_j, this relation would be:

$$k_{app}^{T_i} = k_{max}^{T_i} \frac{[s_j]}{K_M^{T_i} + [s_j]} \tag{4.2.18}$$

In case multiple substrates are transported, the overall apparent catalytic rate is a multiplication of the corresponding Michaelis-Menten terms.

Process machines. The efficiencies of process machines is most often described as a function of the growth rate. Ribosome efficiency was best fit to a Michaelis-Menten function of the growth rate [191], while the efficiencies of other process machines were shown to have a linear relation to the growth rate [192].

4.3 RBA exemplified on a *toy* model

When starting to use complex mathematical and computational frameworks, it does not help when the problem at hand is high-dimensional. It becomes hard to separate errors due to misunderstanding of conceptual ideas from errors due to indexing and similar practical issues. Therefore, I decided to build a simple RBA model of a "toy cell", the main purpose of which is to illustrate all the steps of translating an RBA problem description into an actual linear programming feasibility problem. Hopefully, this model will help others to better understand the construction of the RBA matrices, and consequently to explore their model, to find "bugs" and to fix issues more easily.

The little cell which I build for this purpose is blatantly simple (see Figure 4.3.1) - it uses a single type of energy molecule E and a single type of amino acid AA to assemble its scarce pool of protein, composed of:

- Four transporter species, catalyzing the following transport reactions:
    1. $E_e^{P1} \xrightarrow{\text{T1}} E_c^{P1}$ (import of type I energy precursor)
    2. $E_e^{P2} \xrightarrow{\text{T2}} E_c^{P2}$ (import of type II energy precursor)
    3. $AA_e^{P} \xrightarrow{\text{T3}} AA_c^{P}$ (import of amino acid precursor)
    4. $AA_e \xrightarrow{\text{T4}} AA_c$ (direct import of amino acid)
- Three metabolic enzyme species, catalyzing the following conversions
    1. $E^{P1} \xrightarrow{\text{E1}} E^{P2}$ (conversion of type I into type II energy precursor)
    2. $E^{P2} \xrightarrow{\text{E2}} E$ (conversion of type II energy precuror into energy species)
    3. $AA^{P} \xrightarrow{\text{E3}} AA$ (conversion of amino acid precursor into amino acid)
- ribosomes R and
- two target species concentrations, namely the cytosolic and membrane housekeeping proteins (HP_c, HP_m).

The concentration of the target species is a parameter of the model to be computed from data. The metabolism of the toy cell is given by the following stoichiometric matrix:

$$S = \begin{array}{c|ccccccc} & T_1 & T_2 & T_3 & T_4 & E_1 & E_2 & E_3 \\ \hline E^{P1} & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ E^{P2} & 0 & 1 & 0 & 0 & 1 & -1 & 0 \\ E & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ AA^{P} & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ AA & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array} \tag{4.3.1}$$

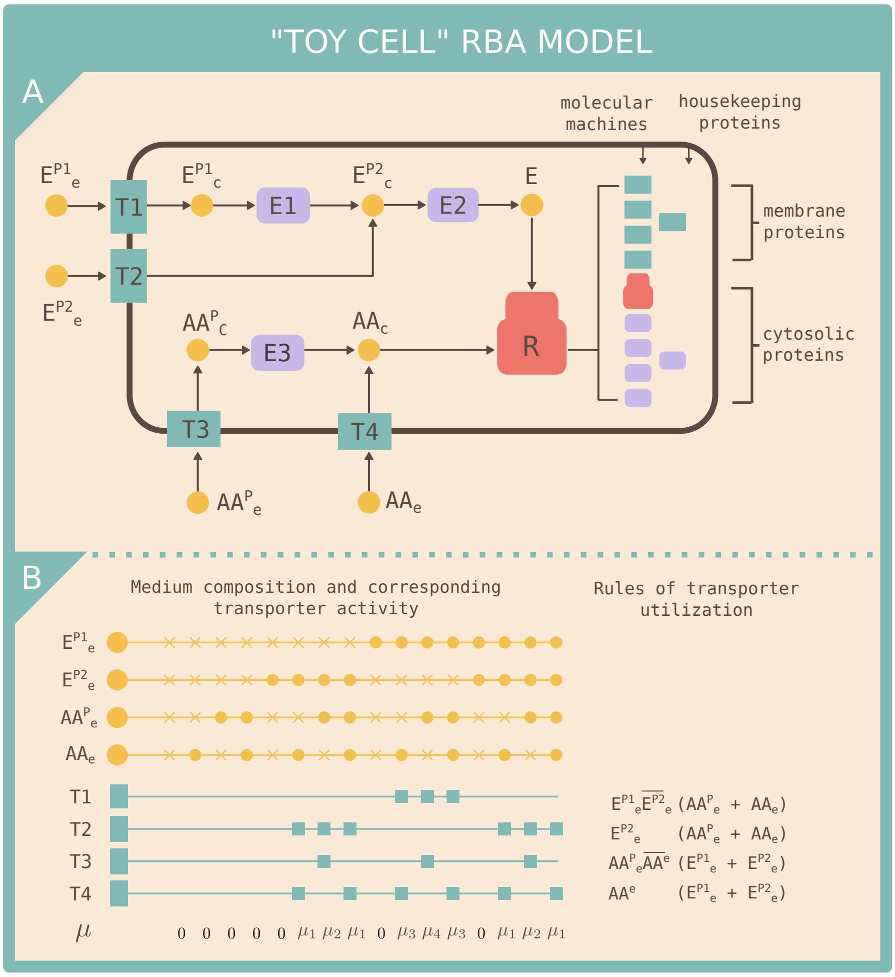


Figure 4.3.1: **(A)** A simple (toy) cell model developed to simplify the understanding of the construction of an RBA problem. The metabolism of the cell has two different energy precursors for which the synthesis price differs in terms of enzyme cost. The same is true for two amino acid precursors. Transporters, enzymes, ribosomes and housekeeping proteins need precursors, energy and process machinery to be produced. **(B)** Illustration of the "decision making" capacity of an RBA model. When a "cheaper" substrate is available in the medium (one requiring fewer resources to utilize), it is preferred, the transporters for the more expensive one are not expressed, and a higher growth rate is obtained.

It was chosen to capture, albeit in a limited sense, the 'decision making process' of the cell in terms of the most efficient utilization of resources for a defined purpose - in our case, the maximization of growth rate. These trade-offs are realized by providing the cell two different 'pathways' for energy production: (1) by importing the Type I precursor $\left[E_e^{P1} \xrightarrow{T_1} E_c^{P1} \xrightarrow{E_1} E_c^{P2} \underset{a}{\xrightarrow{E_2}} E_c\right]$ and (2) by importing the Type II precursor $\left[E_e^{P2} \xrightarrow{T_2} E_c^{P2} \xrightarrow{E_2} E_c\right]$, and two ways of amino acid production: (1) by importing the precursor $\left[AA_e^P \xrightarrow{T_3} AA_c^P \xrightarrow{E_3} AA_c\right]$ and (2) by direct import of amino acids $\left[AA_e \xrightarrow{T_4} AA_c\ \right]$.

The parameterization of the toy model in many ways closely resembles that of the genome-scale model. This allows me to illustrate not only the principles of RBA, but also the practical aspects of the process of translating biological notions into mathematical constraints, and finally into

Table 4.3.1: RBA toy example parameters.

Parameter	Description	Value
P_{tot}	Total cellular protein	$6-0.2\mu$
F_c	Fraction of cytosolic protein	0.8
F_m	Fraction of membrane protein	0.2
$F_{c(ne)}$	Nonenzymatic protein fraction (cytosol)	0.15
$F_{m(ne)}$	Nonenzymatic protein fraction (membrane)	0.2
k_{app}^{E}	Metabolic enzyme efficiency	$10[s^{-1}]$
k_{app}^{T}	Transporter efficiency (substrate S)	$\frac{40[S]}{0.8+[S]}[s^{-1}]$
k_T	Ribosome efficiency	$20[s^{-1}]$
n_T	Amino acids in transporter	2000
n_E	Amino acids in enzyme	1000
n_{HP_c}	Amino acids in cytosolic houskeeping protein	300
n_{HP_m}	Amino acids in membranous houskeeping protein	300
n_R	Amino acids in ribosome	10000

matrices supplied to the LP solver[3]. Since RBA is formulated as a linear programming problem:

$$\begin{aligned}
\underset{x}{\text{maximize}} \quad & c^T x \\
\text{subject to} \quad & Ax = b \\
& Cx \leq d \\
& x \geq 0
\end{aligned}$$

with x representing the vector of decision variables to be identified, the problem formulation needs to fit this schematic.
The vector of decision variables is:

$$\vec{x} = \begin{bmatrix} v_{T_1} & v_{T_2} & v_{T_3} & v_{T_4} & v_{E_1} & v_{E_2} & v_{E_3} & T_1 & T_2 & T_3 & T_4 & E_1 & E_2 & E_3 & R \end{bmatrix} \tag{4.3.2}$$

where $\vec{v}$ is the vector of fluxes through metabolic reactions, $\vec{E}$ a vector of enzyme concentrations (including transporters) and $\vec{P}$ a vector of cellular machinery concentrations (in our case only the ribosomes). The last two entries are not actually decision variables - they are temporarily placed in the vector for the ease of understanding of the matrix manipulations that follow. I now go through the constraints named in section 4.2 and explain them in full detail.
(C_1) **Mass conservation constraint.** This constraint describes the cost of maintaining the concentrations of macromolecular and target species at their steady-state concentrations:

$$S\vec{v} + \mu(C_e\vec{e} + C_p\vec{p} + C_{t_c}\vec{t_c}) = 0 \tag{4.3.3}$$

Since the target concentrations $\vec{t_c} = [HP_c \;\; HP_m]$ are not decision variables of the system, it is possible to transfer them to the right hand side:

$$S\vec{v} + \mu(C_e\vec{e} + C_p\vec{p}) = -\mu C_{t_c}\vec{t_c} \tag{4.3.4}$$

---

[3] The LP solver used for this example is IBM's optimization package CPLEX (and its corresponding Python wrapper).

Each of our macromolecular species requires a certain number of units of amino acid species AA (consult Table 4.3.1), and three times that amount of energy species E. Thus, C_E matrix is:

$$C_E = \begin{array}{c} \\ E^{P1} \\ E^{P2} \\ E \\ AA^P \\ AA \end{array} \begin{array}{c} \begin{array}{ccccccc} T_1 & T_2 & T_3 & T_4 & E_1 & E_2 & E_3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3n_T & -3n_T & -3n_T & -3n_T & -3n_E & -3n_E & -3n_E \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -n_T & -n_T & -n_T & -n_T & -n_E & -n_E & -n_E \end{bmatrix} \end{array} \tag{4.3.5}$$

C_p and C_{t_c} matrices are constructed in the same fashion. Negative values indicate the metabolites are used, and the positive values, if there were any, would indicate metabolites released through the process of construction of macromolecular species. It is now possible to construct the mass conservation constraints, which are 5 in total - one for each metabolite (excluding the external ones). For the target species, the expressions that would be the result of the $C_{t_c}\vec{t_c}$ can be considered the portion of nonenzymatic protein in each compartment and computed in the following way:

$$\begin{aligned} n_{HP_c}HP_c &= P^{ne}_{cyt}(\mu) = F_{c(ne)} \times F_c \times P_{tot}(\mu) \\ n_{HP_m}HP_m &= P^{ne}_{mem}(\mu) = F_{m(ne)} \times F_m \times P_{tot}(\mu) \end{aligned} \tag{4.3.6}$$

(C_2) **Capacity constraint.** I first relate the concentration and translation rate of the ribosome to the flux of new protein that needs to be produced:

$$k_T R - \mu\left(\vec{M_e}\vec{e} + M_p\vec{p} + \vec{M_{t_c}}\vec{t_c}\right) = 0 \tag{4.3.7}$$

M_e is a vector containing the corresponding number of amino acids for each metabolic enzyme and the same is the case for M_p and M_{t_c}:

$$\begin{aligned} M_e &= \begin{bmatrix} n_{T_1} & n_{T_2} & n_{T_3} & n_{T_4} & n_{E_1} & n_{E_2} & n_{E_3} \end{bmatrix} \\ M_p &= \begin{bmatrix} n_R \end{bmatrix} \\ M_{t_c} &= \begin{bmatrix} n_{HP_c} & n_{HP_m} \end{bmatrix} \end{aligned} \tag{4.3.8}$$

Next, I write the constraints on the capacities of individual enzymes and transporters. Since all of the metabolic enzymes can facilitate reactions just in the forward direction, the following constraints apply:

$$v_\theta \leq k^+_\theta \theta, \qquad \theta \in \{T_1, T_2, T_3, T_4, E_1, E_2, E_3\} \tag{4.3.9}$$

The matrix Ψ mapping metabolic fluxes to corresponding metabolic enzymes (see Equation 4.2.10) is an identity matrix of dimension 7. This is due to the fact that the order between metabolic enzymes and corresponding fluxes is maintained (see Equation 4.3.2), there are no reactions catalyzed by more than one enzyme, and no backward reactions.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{T_1} \\ v_{T_2} \\ v_{T_3} \\ v_{T_4} \\ v_{E_1} \\ v_{E_2} \\ v_{E_3} \end{bmatrix} - \begin{bmatrix} k^+_{T_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & k^+_{T_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & k^+_{T_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k^+_{T_4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k^+_{E_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & k^+_{E_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k^+_{E_3} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ E_1 \\ E_2 \\ E_3 \end{bmatrix} \leq \vec{0} \tag{4.3.10}$$

(C_3) **Density constraint.** The cell is assumed to be able to fit a certain total amount of protein $P_{tot}(\mu)$, which is distributed in the cytosol and membrane compartments (see Table 4.3.1). The cytosolic fraction is $P_{cyt}(\mu) = F_c P_{tot}(\mu)$ and the membrane fraction is $P_{mem}(\mu) = F_c P_{mem}(\mu)$. In this simple model, I assume that all the macromolecules take up space that is directly proportional to the number of amino acids that constitute them. In this case, the model has two density constraints, one for the cytosol, and one for the membrane:

$$n_{E_1}E_1 + n_{E_2}E_2 + n_{E_3}E_3 + n_R R \leq P^e_{cyt}(\mu) \tag{4.3.11}$$

$$n_{T_1}T_1 + n_{T_2}T_2 + n_{T_3}T_3 + n_{T_4}T_4 \leq P^e_{mem}(\mu) \tag{4.3.12}$$

where $P^e_{cyt}(\mu)$ represents the enzymatic portion of the cytosolic protein and is computed as $(1 - F_{c(ne)}) \times F_c \times P_{tot}(\mu)$. The same holds for $P^e_{mem} = (1 - F_{m(ne)}) \times F_m \times P_{tot}(\mu)$. This constraint can be expressed in matrix form:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & n_{E_1} & n_{E_2} & nE_3 & n_R \\ n_{T_1} & n_{T_2} & n_{T_3} & n_{T_4} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ E_1 \\ E_2 \\ E_3 \\ R \end{bmatrix} \leq \begin{bmatrix} P_{cyt}(\mu) - P^e_{cyt}(\mu) \\ P_{mem}(\mu) - P^e_{mem}(\mu) \end{bmatrix} \tag{4.3.13}$$

The expression $P_{cyt}(\mu) - P^e_{cyt}(\mu)$ equals the nonenzymatic portion of the protein, as stated in Equation 4.3.6.

With this, all of the model constraints have been written in matrix form. The last thing left to do is to stack up the constraints in one matrix. The full matrix of the toy RBA model is given in section A.3.

4.3.1 Model simulation

The final toy model can now be simulated to assess its ability to qualitatively reproduce cellular ’decision making’ behavior, based on optimal allocation of resources for the maximization of growth rate. The Python implementation of this model is available as a public Gist. In order to run it, one needs to have the CPLEX optimizer by IBM installed, and its Python wrapper [194]. The simulation results are indicated in the bottom part of Figure 4.3.1. The yellow dots represent all possible combinations of metabolite presence in the medium. For a number of media the toy cell does not grow at all. This happens in cases when there is either the energy precursor or amino acid precursor missing in the medium. All the other 11 cases will result in four different cellular configurations, because the metabolites E^{P2} and AA are preferred over E^{P1} and AA^P since they require smaller investment in cellular resources. Therefore, when either of them is present in the medium, it will be used and its alternative will not. As expected, when growing on the preferred sources, the *toy* cell obtains the highest growth rate (see Figure 4.3.2), while the lowest growth rate is obtained when only the non-preferred substrates are present in the medium. It is already visible from Figure 4.3.2 how this extremely simple RBA cell shows correspondence with experimentally determined bacterial growth laws [193] - such as the increase in total protein quantity and fraction of total protein allotted to ribosomes with the increase in growth rate.

In RBA, the efficiency of transporters depends on the concentration of their respective substrates in the medium. In the toy model in particular (but also often in genome-scale RBA models) this

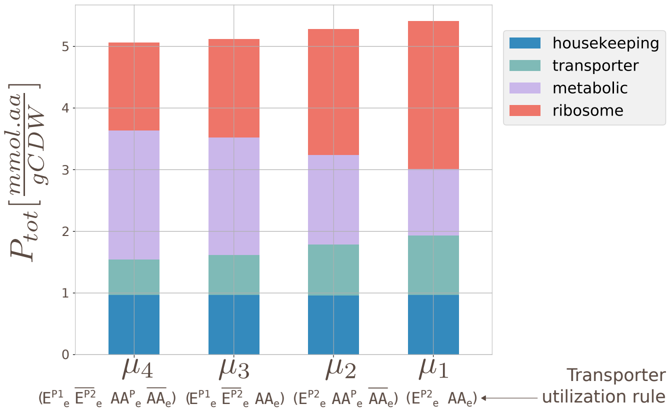


Figure 4.3.2: Given the four possible nutrients available in the medium (E^{P1}, E^{P2}, AA and AA^P) and the $2^4 = 16$ combinations thereof, the *toy* cell grows on 11. In those 11 growth situations, four different growth rates appear, depending on the presence of preferred nutrients in the medium. The cell grows fastest (and achieves the biggest size) when growing on AA and E^{P2} (μ_1), regardless of the availability of other substrates, and the slowest when only AA^P and E^{P1} are present in the medium (μ_4). Below the growth rates, there are logical expressions describing the occurrence of that particular growth rate in terms of precursor availability in the medium. We can see how this simple model recaptures the so-called "bacterial growth laws" [193], such as an increase in total protein and ribosomal fraction in the total protein pool with the increase in growth rate.

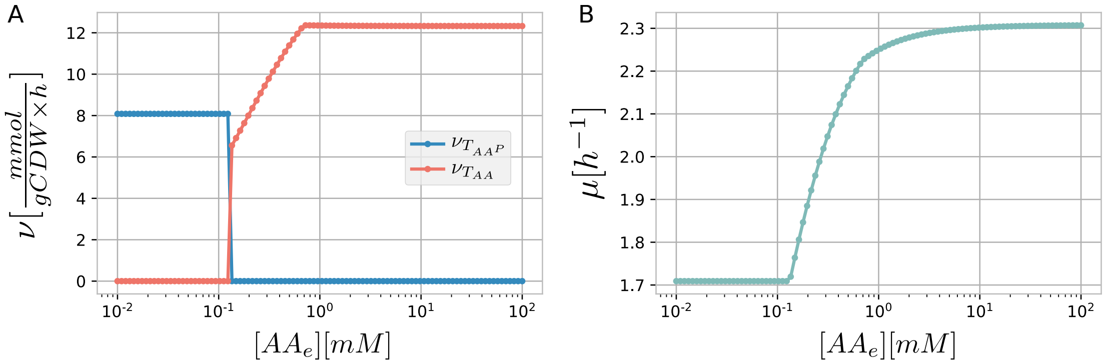


Figure 4.3.3: **(A)** Substitution of transporter usage under a changing concentration of substrate AA when the concentration of AA^P is held constant and high. Transporter efficiency is determined by the concentration of the substrate in the medium. While the concentration of the preferred substrate AA is low, the cell uses the substrate AA^P which is more costly in terms of resources, but its high concentration makes the transporter T_{AA^P} more efficient. When the concentration of AA becomes high enough so that the T_{AA} efficiency increases until making it favorable for the cell, the cell switches to the more resource-efficient substrate AA. **(B)** Cellular growth rate under changing concentration of AA. The switch to AA from AA^P results in the increase in growth rate, and therefore, more efficient utilization of cellular resources.

efficiency is expressed as a Michaelis-Menten function of the substrate concentration - $V_{max}\frac{[S]}{K_M+[S]}$. I have already shown that when all the substrates are present in high concentrations, the AA substrate is preferred over AA^P, as it results in higher growth rate. However, if AA were available at a low concentration, that would render its respective transporter quite inefficient, causing the cell to utilize AA^P. This situation is illustrated in Figure 4.3.2. The cell grows utilizes AA^P and grows at the same growth rate until the concentration of AA becomes high enough to render its transporter more efficient. After that point, the cell preferentially uses AA. As the concentration of AA continues to increase, making its transporter more efficient, the growth rate increases as well, reflecting the fact that the cell needs to invest less resources in order to obtain the same flux. This effect evetually comes to a saturation point, because of the Michaelis-Menten type efficiency of the transporter.

4.4 Related modeling paradigms

RBA was first proposed in 2009 by the first author and coworkers [195], who described the cell as a convex optimization problem. At that time, the idea of bacterial growth as a consequence of optimization of the cellular self-replicating process in terms of "cellular economics" and "allocation of resources" is explored for single proteins [196], on the cellular level either phenomenologically [197] or with small dynamical models [198].

4.4.1 FBA with molecular crowding

As briefly mentioned in subsection 1.4.2, FBA is a constraint-based modeling paradigm which allows for predictions of metabolic fluxes based on the stoichiometry of the metabolic network and experimentally determined quantification of certain fluxes under the assumption of maximization of biomass production [27].

$$\begin{aligned} \max_{v\in\mathscr{R}^N_{\geq 0}} \quad & c^T v \\ \text{subject to} \quad & Sv = 0 \\ & v \leq b \end{aligned}$$

where v is the vector of metabolic fluxes, S is the stoichiometric matrix and $b \in \mathscr{R}^N_{>0}$ is the vector by which the fluxes are constrained either to an experimentally determined value, or to a high number (thus limiting the feasibility region of the problem). $c \in \mathscr{R}^N$ is a vector of coefficients which determines which reaction will be maximized. Normally, there are one or more "biomass" reactions in the model and vector c contains all zeros expect for a single one at a position corresponding to the biomass reaction to be maximized. A biomass reaction typically takes into account all the precursors, energy and ions necessary for the construction of a new cell and the production of metabolites released by the cell during growth.

In FBA, the only limit on cellular growth is imposed by limiting the fluxes. The higher the bounds on the fluxes, the higher will be the predicted growth rate, with no upper limit. Additionally, when the simulated medium has a number of carbon sources, the model cell will utilize them all and have a correspondingly higher growth rate. Therefore, the phenomenon of *diauxie* cannot be captured in such a model. To circumvent such problems, a new constraint was introduced in the FBA formulation, attempting to capture the consequences of limited cellular space [28], giving rise to the modeling paradigm known as FBA with molecular crowding (FBAwMC). It states that cellular volume is limited and sets the upper bound on the sum of voluminous contributions of enzymes:

$$\sum_{i=1}^{N} v_i n_{E_i} \leq V_{max} \tag{4.4.1}$$

which can be expressed in terms of enzyme concentrations by dividing the expression by cellular mass M:

$$\sum_{i=1}^{N} v_i [E_i] \leq \frac{V_{max}}{M} = \frac{1}{C} \tag{4.4.2}$$

where $[E_i]$ is the concentration of enzyme E_i and C is the cytosolic density of the bacterial cell. By assuming that the flux through a metabolic reaction is directly proportionate to the concentration of enzyme $\nu_i = k_i E_i$, the constraint in Equation 4.4.2 can be expressed in terms of variables ν:

$$\sum_{i=1}^{N} a_i \nu_i \leq 1 \tag{4.4.3}$$

where the coefficient a_i corresponds to $a_i = \frac{C v_i}{k_i}$.
As also noted in their paper, the FBAwMC does not take into account the volume taken up by ribosomes - either by their protein or their RNA part [28]. Additionally, in comparison with RBA, the proteins are not built by necessary substrates, energy and molecular machines, but are just represented voluminously.

4.4.2 MOMENT

MOMENT stands for MetabOlic Modeling with ENzyme kineTics [178]. Like the two methods described above, it extends the metabolic network with macromolecular expression. In particular, it introduces a new set of variables representing gene products g_i which are estimated during the optimization procedure. The constraints which supplement the original FBA formulation have to do with limiting fluxes through reactions based on the concentration of a particular gene product:

$$\nu_i \leq k_{cat}[g_i] \tag{4.4.4}$$

and with limiting the amount of protein in the cell by accounting for their voluminous contribution:

$$\sum_{i} g_i MW_i \leq C \tag{4.4.5}$$

where MW_i is the molar mass of a gene product g_i and C_i is the parameter of the model denoting the total protein weight. Additionally, MOMENT does not assume that instantaneous accumulation of the biomass (growth rate) is the optimization criteria, but instead optimizes for maximal ATP yield at minimal enzymatic usage [199]:

$$\frac{\nu_{ATP}}{\nu_{glc}} - \varepsilon \sum \nu_i^2 \tag{4.4.6}$$

Therefore, MOMENT can predict metabolic fluxes and gene product concentrations, but does not take into account additional cellular processes nor the cellular cost of producing gene products.

4.4.3 ME-models

ME-models take their name from the so-called *M-models* - genome scale models of metabolism. In ME-models the acronym stands for modeling of metabolism and macromolecular expression (ME). ME-models can represent a variety of cellular processes, such as the production of mRNA and protein, protein modification and assembly [177, 200]. This is achieved by introducing a number of "coupling constraints" to the original FBA formulation. For example, the flux of translation of an mRNA is limited by its rate of degradation and the flux of a reaction through an enzyme is limited by its dilution. The formation flux of the ribosome is related to the total flux of translation, as is the formation flux of RNA polymerase to the total flux of transcription. For example, for an enzyme, its formation flux and the flux of the reaction it facilitates are thus related:

$$v_{E_i}^{f} - \sum_{r}^{r \in \mathscr{R}_{E_i}} \frac{\mu}{k_{i,r}^{eff}} v_r = 0 \tag{4.4.7}$$

where $\mathscr{R}_{E_i}$ is the set of reactions catalyzed by enzyme E_i, $v_{E_i}^{f}$ the formation flux of enzyme E_i and v_r the flux through reaction r. The predictions of such a model are the metabolic fluxes, mRNA and protein abundances.

From its scope and predicted quantities, RBA and ME-Models have a similar scope. However, they are two problems quite differently posed and differently implemented. For example, the RBA problem is formulated to impose a single constraint on a metabolite for the production flux of all the molecular machines it is a part of. This makes model simulation fast and model extension quite simple. While the constraints given in [177, 200] outline certain conceptual differences between the two modeling paradigms, to my knowledge, no full problem formulation of the ME-models has yet been published. This clearly makes a detailed comparison somewhat difficult.

4.5 RBApy software

RBApy is a free Python software which helps automate the process of creating, modifying and simulating bacterial resource allocation models encoded in an RBApy-XML format [192] (see Figure 4.5.1).

The `preRBA` package of the software provides tools for the creation and modification of such models, and it can be used as a standalone tool. The creation of a new model requires the modeler to supply an annotated metabolic reconstruction (in Systems Biology Markup Language (SBML) format) and a Uniprot ID [203] of the organism of interest. Using the gene associations of all the annotated reactions, `preRBA` will proceed to download the available information on the corresponding gene products from Uniprot. The information that `RBApy` will attempt to gather is the following:

- Sequences of all gene products
- Cofactors required by individual proteins and their corresponding identifiers in the metabolic reconstruction
- Protein subcellular localization
- Stoichiometries of individual proteins in enzyme complexes
- Additional chemical components the cell needs to produce (from the biomass equation).

There are a number of reasons why it is highly unlikely that `preRBA` will be completely successful in gathering this information. Uniprot, for example, encodes the enzyme stoichiometry

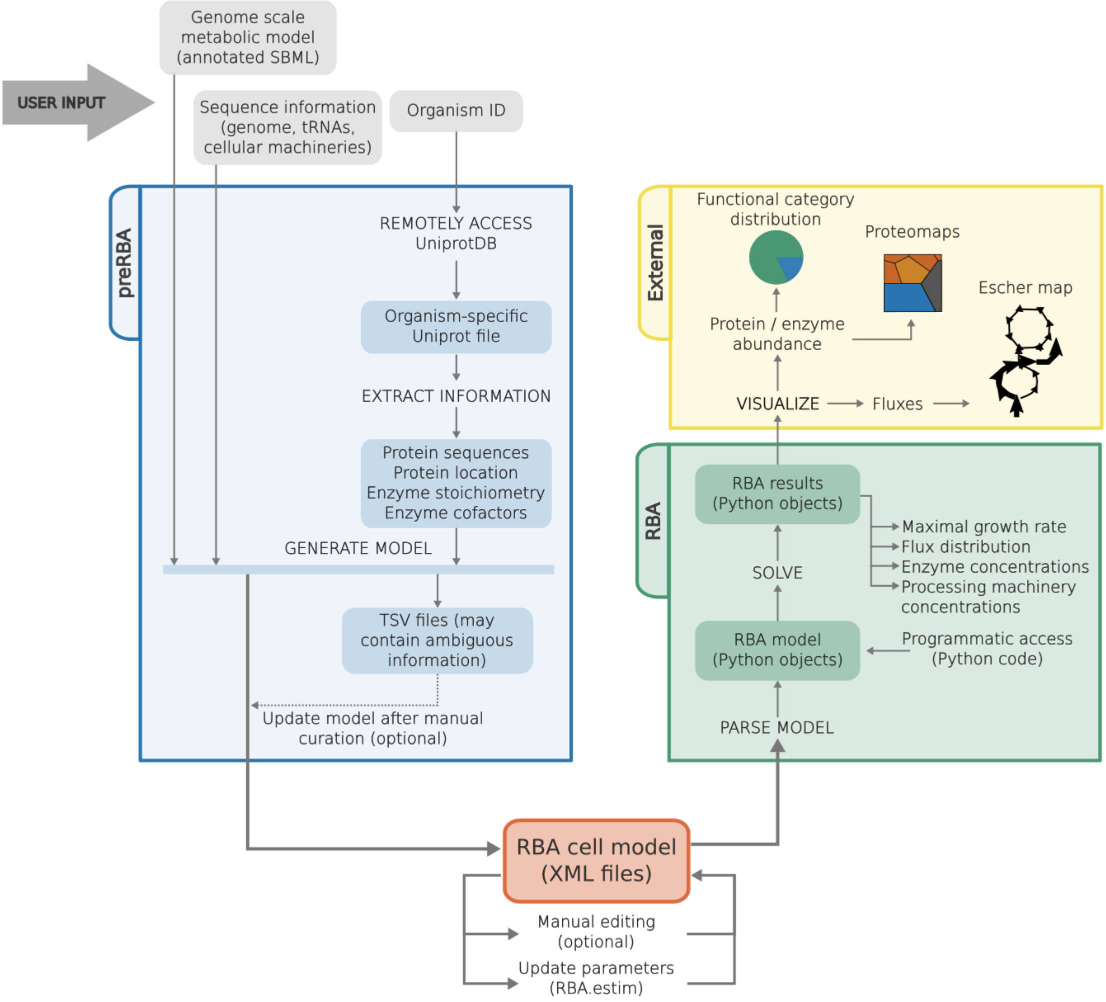

USER INPUT
Genome scale metabolic model (annotated SBML)
Sequence information (genome, tRNAs, cellular machineries)
Organism ID
preRBA
REMOTELY ACCESS UniprotDB
Organism-specific Uniprot file
EXTRACT INFORMATION
Protein sequences
Protein location
Enzyme stoichiometry
Enzyme cofactors
GENERATE MODEL
TSV files (may contain ambiguous information)
Update model after manual curation (optional)
External
Functional category distribution
Proteomaps
Escher map
Protein / enzyme abundance
VISUALIZE
Fluxes
RBA
RBA results (Python objects)
Maximal growth rate
Flux distribution
Enzyme concentrations
Processing machinery concentrations
SOLVE
RBA model (Python objects)
Programmatic access (Python code)
PARSE MODEL
RBA cell model (XML files)
Manual editing (optional)
Update parameters (RBA.estim)

Figure 4.5.1: Process of creating a bacterial RBA model by using RBApy. User needs to provide an annotated metabolic reconstruction in SBML format, Uniprot ID of the modeled organism, and a file containing sequences of macromolecular process machines which are to be included in the model (i.e. ribosomes, chaperones). RBApy retrieves Uniprot data on enzymes (stoichiometry) and proteins (localization, amino acid sequence, cofactors) and creates an RBA model. This model can further be refined with the use of (a) available calibration methods or manual editing of "helper" files which list all ambiguous information. The simulation results can be interfaced to Escher maps [201] for flux or to Proteomaps [202] for protein abundance visualization.

information in plain text, which is often difficult to parse without ambiguity, and gene IDs get replaced or become obsolete. In such cases, `preRBA` generates the so-called *helper* files, which the modeler can fill in with the missing information. After modifying the helper files, `preRBA` should be called again, and a new model will be generated, updated in the places corresponding to the newly given information. The process can be repeated until a satisfactory level of detail and accuracy is achieved. The final model is encoded in a set of XML files which contain an RBA model of the organism of interest (see subsection 4.5.1).

`RBApy.RBA`, unlike the `preRBA`, is better used as an API, even if a standalone script is offered. It offers the user a programmatic access to the `RBA` object, enabling model modification and parameterization. It also offers a default simulation scenario, in which growth rate is maximized. The simulation results can then be exported into various formats compatible with online visualization tools (such as Proteomaps [202] for protein or Escher maps [201] for flux visualization).

4.5.1 Model format

`RBApy` utilizes a novel Extensible Markup Language (XML) format to represent an RBA cellular model. One might rightly ask if it was necessary to develop a new format in face of existing formats. SBML [204] is a standardly used format to represent biological models and supports constraint-based models through its Flux Balance Constraints Package [205]. This format is well adapted for encoding of FBA models, in which most reactions are metabolic reactions, and in which macromolecular species facilitating them are not present. However, RBA introduces the problem of modeling macromolecular species (proteins, RNA, macromolecular complexes), each of which needs to be produced through an action of one or more process machines. In a reaction-centered modeling format, such as SBML, this would require addition of numerous new reactions. For each individual macromolecule, one would need to add a specific synthesis reaction, listing all the necessary substrates in energy and precursors and the products released. In order to ensure the proper mass conservation, all of these terms for all the individual macromolecules would need to be added to all the reactions of the associated substrates and products. But not only do metabolites take part in the synthesis of macromolecules - process machines are needed as well, as are ribosomes for translation. Would it be an acceptable solution to add ribosome "consumption" and "release" in each of the macromolecule synthesis reactions? The situation is further complicated by the fact that functional proteins often are formed not through the action of one, but many cellular processes. Apart from translation, a protein might require folding and post-translational modification. In that case, one would need to modify all the synthesis reactions to add the necessary substrates and products required by the new process, and continue to update all the metabolite reactions to reflect this.

These remarks suffice to make it clear that this kind of encoding is not suited for simple addition and removal of cellular processes, nor is it suited for changing the processing requirements of individual proteins. Additionally, the above-described reaction-centric solution adds to the numerical complexity of the problem and could slow down or even impede the process of finding an optimal solution of an RBA linear programming problem. RBA-xml format was developed to resolve these issues.

For this purpose, an RBA model is split into a number of XML files. Their entire description can be found in the official documentation [206], available on the GitHub pages of the `RBApy` project. Here, I explain the functional links between them and design ideas that favored such an organization. Metabolism is described in the `metabolism.xml` file in a standard way. Each reaction specifies whether it is reversible or not, it has an associated list of substrates, products and their stoichiometries. Additionally, each reaction by default has an enzyme associated to it, which is listed in the `enzymes.xml` file, under the ID `reaction_ID_enzyme`. Each enzyme entry lists the identifiers and stoichiometry of proteins that constitute it, as well as an identifier of a function which describes its efficiency. These (and other) functions can be found in the `parameters.xml` file. The composition of proteins in terms of amino acids and cofactors is given in the `proteins.xml` file, as is the composition of DNA and RNA given in `dna.xml` and `rnas.xml` files correspondingly. Cellular processes in charge of producing these proteins and other macromolecules are described in the `processes.xml` file. The representation of cellular processes is the most important novelty introduced by the RBA-xml format and is illustrated in Figure 4.5.2. Each process is described by:

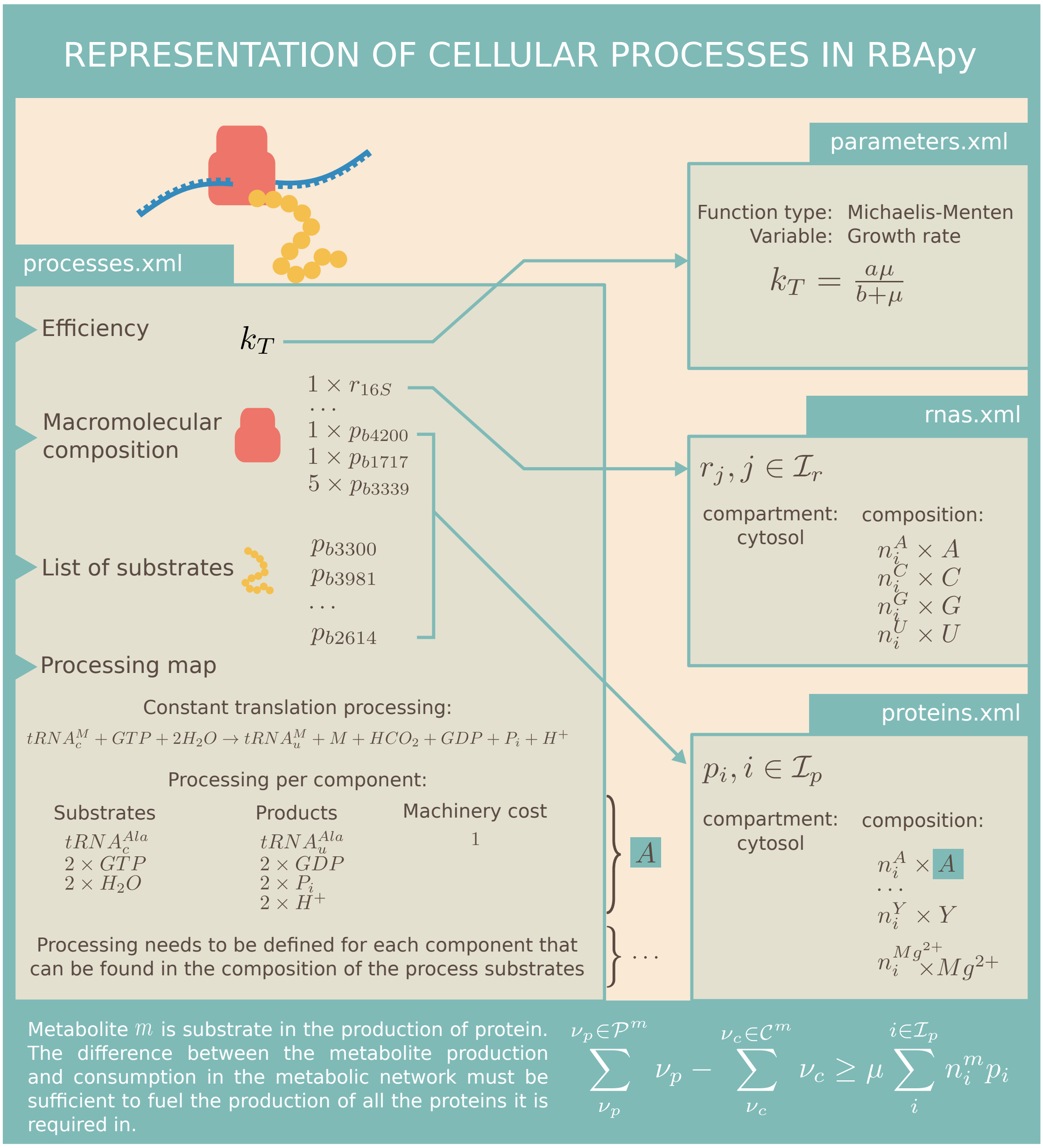

REPRESENTATION OF CELLULAR PROCESSES IN RBApy
parameters.xml
Function type: Michaelis-Menten
Variable: Growth rate
$k_T = \frac{a\mu}{b+\mu}$
processes.xml
Efficiency
k_T
Macromolecular composition
$1 \times r_{16S}$
$\dots$
$1 \times p_{b4200}$
$1 \times p_{b1717}$
$5 \times p_{b3339}$
List of substrates
p_{b3300}
p_{b3981}
$\dots$
p_{b2614}
Processing map
Constant translation processing:
$tRNA_c^M + GTP + 2H_2O \rightarrow tRNA_u^M + M + HCO_2 + GDP + P_i + H^+$
Processing per component:
Substrates
$tRNA_c^{Ala}$
$2 \times GTP$
$2 \times H_2O$
Products
$tRNA_u^{Ala}$
$2 \times GDP$
$2 \times P_i$
$2 \times H^+$
Machinery cost
1
A
$\dots$
Processing needs to be defined for each component that can be found in the composition of the process substrates
rnas.xml
$r_j, j \in \mathcal{I}_r$
compartment: cytosol
composition:
$n_i^A \times A$
$n_i^C \times C$
$n_i^G \times G$
$n_i^U \times U$
proteins.xml
$p_i, i \in \mathcal{I}_p$
compartment: cytosol
composition:
$n_i^A \times A$
$\dots$
$n_i^Y \times Y$
$n_i^{Mg^{2+}} \times Mg^{2+}$
Metabolite m is substrate in the production of protein. The difference between the metabolite production and consumption in the metabolic network must be sufficient to fuel the production of all the proteins it is required in.
$\sum_{\nu_p}^{\nu_p \in \mathcal{P}^m} \nu_p - \sum_{\nu_c}^{\nu_c \in \mathcal{C}^m} \nu_c \geq \mu \sum_{i}^{i \in \mathcal{I}_p} n_i^m p_i$

Figure 4.5.2: Cellular processes in RBApy are listed in the `processes.xml` file, where they are represented by their efficiency, the macromolecular composition of their process machinery, list of substrate macromolecules they process and the processing map. The efficiency of the process machinery is listed in the `parameters.xml` file. The composition of macromolecules is listed in `proteins.xml` and `rnas.xml` files, where for each protein or RNA, where each macromolecule has a list of components and their associated stoichiometries. The processing map indicates if there is a fixed cost associated with the process (such as the cost of translation initiation) and defines the processing for each metabolite that can be found in the composition of substrate macromolecules. Such representation of processes allows the `RBApy` software to easily formulate a single constraint for each of the metabolites involved in cellular processes.

- its efficiency,
- the composition of its macromolecular machine (as ribosomes are for translation),
- list of substrate macromolecules (all the proteins in case of translation) and
- a processing map listing a constant processing cost (such as that of translation initiation) and a processing rule for each component that can be found in the composition of its substrate macromolecules.

A requirement for any particular metabolite is computed as a total sum of all thus described requirements, resulting in a single constraint per metabolite (constraint C_1). The capacity constraint (constraint C_2) for a particular cellular process is formalized from the list of all the substrate macromolecules and their concentrations. This kind of encoding allows a much cleaner model construction, modification and maintenance, and was a major reason behind introducing a new XML model encoding format.

5. An RBA model of *E. coli*

> Beyond a critical point within a finite space, freedom diminishes as numbers increase. This is as true of humans in the finite space of a planetary ecosystem as it is of gas molecules in a sealed flask. The human question is not how many can possibly survive within the system, but what kind of existence is possible for those who do survive.
>
> Frank Herbert, Dune

In this chapter I present all the information that was necessary for the creation, calibration and verification of the *Escherichia coli* RBA model. I would first like to explain my motivation for building this model. The topic in this study is in great part the production and secretion of recombinant protein in *Escherichia coli*. Modeling of such a process is far from trivial, as it (a) is bound to one of the cells most important functions - the production of protein, (b) is energetically costly, (c) it introduces big changes in the cellular state due to reduction of space and other resources available for the "normal" cellular functions, (d) it changes the rate of growth of the population and (e) it can influence the stability and state of foldedness of the rest of the cellular proteome. The effects (a)-(d) are basically the problems of resource allocation, while (e) can be considered to fall in the domain of stress responses.

With most modeling paradigms often used in the field of systems biology, one would need to resort to modeling and analyzing the dynamics of the interaction of a relatively few reactant species [175] or to the stoichiometric analysis of the species' metabolism. Due to the nature of the problem, neither is particularly adapted to a detailed study of the stated problem, especially taking into account that this work is motivated by the optimization of the industrial processes, in which generic phenomenological conclusions are of little use. Resource Balance Analysis was, to my knowledge, the modeling approach most suited to the study of a protein production - a process central to cellular allocation of resources. It allows for a detailed, cell-level description and an easy integration of an entire metabolism of the modeled organism. Apart from that, as I will discuss in chapter 6, it provides a framework upon which dynamic models can be built, still taking into account all the constraints which form the basis of RBA.

The RBA model of *Escherichia coli* was constructed through the use of `RBApy` software [192]. It was based on the most current genome-scale metabolic reconstruction at the time of model creation [207] and calibrated using physiological data measurements [32] and a comprehensive proteomics dataset [208]. The entire model along with the calibration is available at the Github pages of SysBioIntra group: https://github.com/SysBioInra/Bacterial-RBA-models.

The metabolic reconstruction used for the creation of the *E. coli* RBA model is `iJO1366` [207], available on the BiGG Models database [209]. It is a compartment-specific reconstruction, with metabolites assigned to either cytoplasm, periplasm or external compartment. It is composed of 1805 metabolites and 2583 reactions, most of them annotated with a gene association rule. The information on the exact sequences of all the proteins involved in the model, as well as their stoichiometry in the enzymatic complex, necessary cofactors and localization was downloaded from Uniprot [203] through `RBApy`. In the developed RBA *E. coli* model, apart from translation, I

Table 5.0.1: Values and parameters needed for parameter estimates of the RBA model of *E. coli*. Parameters whose value isn't specified are dimensionless. $/^a$ - the data sources used are the rRNA sequences.

Parameter	Symbol	Data sources	Formula	Value	Unit	Section
Fractions of ind. NAs in rRNA	$\vec{p_{NA}}$	rRNA sequence				A.4
Fractions of ind. of AAs in proteome	$\vec{p_{AA}}$	[210]				A.4
Molar masses of nucleic acids	$M\vec{W}_{NA}$				$\frac{g}{mol}$	A.4
Molar masses of amino acids	$M\vec{W}_{AA}$				$\frac{g}{mol}$	A.4
Weighted average nucleic acid	$M\bar{W}_{NA}$		$\vec{p_{NA}}^T M\vec{W}_{NA}$	340.19	$\frac{g}{mol}$	
Weighted average amino acid	$M\bar{W}_{AA}$		$\vec{p_{AA}}^T M\vec{W}_{AA}$	108.28	$\frac{g}{mol}$	
Fraction of stable RNA that is tRNA	p_{tRNA}	[32]		0.14	/	
Fraction of mRNA in total RNA	p_{mRNA}	[211]		0.05	/	
Fraction of RNA that is rRNA	p_{rRNA}		$1 - p_{tRNA} - p_{mRNA}$	0.81	/	
Number of nucleotides per ribosome	$N_{na/rib}$	rRNA sequence		4593	$\frac{na}{rib}$	
Ribosome scaling factor	$d_{r/R}$		$N_{na/rib}\frac{M\bar{W}_{NA}}{M\bar{W}_{AA}}$	14430	/	5.1.4
Ribosome maturation time	$T_{mat/R}$	[29, 212]		5	min	
Ribosome efficiency	$k_T(\mu)$	[29, 32]	$k_T = \frac{27\mu}{0.5+\mu}$		$\frac{aa}{s}$	
Fraction of protein in CDW	$p_{p/CDW}(\mu)$	[213]	$-0.28\mu + 0.64$			
Fraction of RNA in CDW	$p_{R/CDW}(\mu)$	[213]	$0.14\mu + 0.05$			
Fraction of cytoplasmic protein	$p_{cyt/P}(\mu)$	[208]	$0.73 + 0.04\mu$			5.1.1
Fraction of inner membrane protein	$p_{im/P}$	[208]	0.08			5.1.1
Fraction of outer membrane protein	$p_{om/P}(\mu)$	[208]	$0.09 - 0.007\mu$			5.1.1
Fraction of periplasmic protein	$p_{p/P}(\mu)$	[208]	$0.10 - 0.04\mu$			5.1.1
Fraction of secreted protein	$p_{s/P}$	[208]	0.0009		/	5.1.1
Cytosolic density	D_{cyt}		see Equation 5.1.18	4.89	$\frac{mmol.aa}{gCDW}$	5.1.3
Total protein concentration	$P_{tot}(\mu)$		$5.91 - 1.04\mu$		$\frac{mmol.aa}{gCDW}$	5.1.4

have included the process of protein folding (chaperoning) and secretion.

5.1 Parameterization of the *E. coli* RBA model

In this section I describe all the steps necessary for basic parameterization of a bacterial RBA model. However, due to the lack of experimental data suitable for a complete parameterization, some steps could not have been performed - as is the estimation of growth rate dependent molecular machine efficiencies. In this section, I will first discuss the estimation of the so-called *physiological* parameters: (i): percentage of protein per compartment, (ii) percentage of housekeeping protein per compartment, (iii) cytosolic density and (iv) the total amino acid concentration. Next, the estimation of the molecular-machine-specific parameters is discussed: (i) enzyme-specific catalytic rates and (ii) efficiencies of process machines. Finally, I provide some information on experimental methods for the generation of the data used in parameterization of the *E. coli* RBA model.

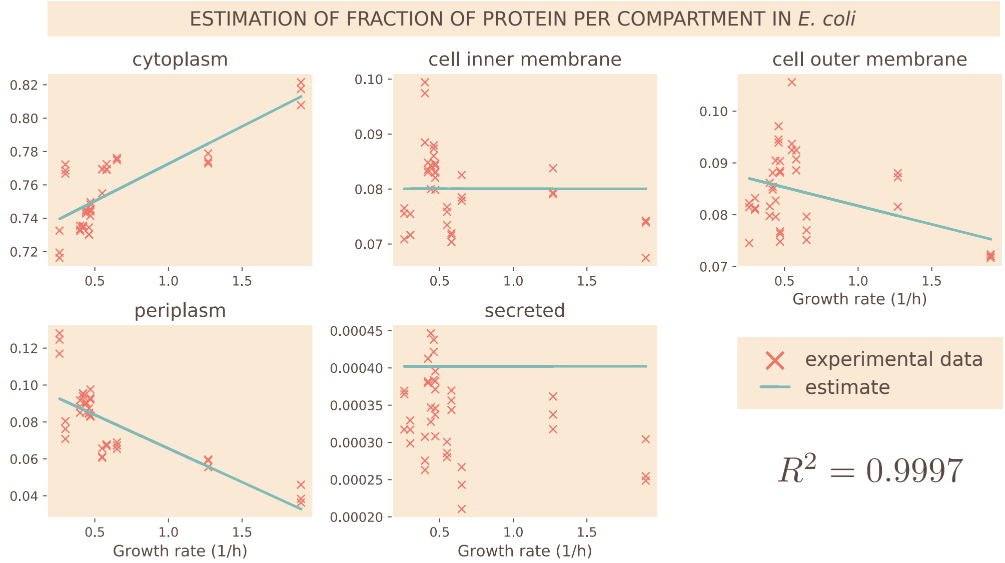


Figure 5.1.1: Linear and constant fits for the dependence of percentage of protein per compartment as the function of the growth rate, using the data from [208] and Uniprot subcellular localization annotations.

5.1.1 Percentage of protein per compartment

In RBA, the amount of protein per compartment is represented by the concentration of amino acids in unit of $mmol.aa/gCDW$. This amount is later used in the optimization procedure as an inequality constraint limiting how much of protein produced for metabolic and process requirements can fit into each compartment. With the change in growth rate, bacteria change their size, and with it the amount of protein. Therefore, the protein concentration per compartment is modeled as a function of the growth rate. The amount of protein per cell has been shown to vary linearly with the growth rate [32]. The estimation procedure assumes that the protein concentration per compartment is either constant or a linear function of the growth rate.

The estimates of percentages of protein per compartment necessarily all need to be non-negative and need to sum to one.

I have used the subcellular localization information from the proteomics data from [208] (Supplementary Table S13) to estimate this parameter in *E. coli*. For each experiment $e, e \in \mathscr{E}$, there is an associated growth rate μ_e, and the percentages of protein allocated to different compartments $y_e^c, c \in \mathscr{C}$ computed from the proteomics data. It is necessary to find coefficients $a_e^c, b_e^c, e \in \mathscr{E}, c \in \mathscr{C}$ such that minimize the sum of residuals

$$S = \sum_{e \in \mathscr{E}, c \in \mathscr{C}} \left(y_e^c - (a^c \mu_e + b^c) \right)^2 \tag{5.1.1}$$

Additionally, for all the growth rates μ_e, the goal is to minimize the difference of the sum of percentages of protein in all compartments from 1:

$$S_{tot} = \sum_{e \in \mathscr{E}} \left(1 - \sum_{c \in \mathscr{C}} a^c \mu_e + b^c \right)^2 \tag{5.1.2}$$

The problem can be formulated as follows:

$$\begin{array}{ll} \min\limits_{x} & \|Ax - y\|_2 \\ \text{s.t.} & Ax \geq 0 \end{array} \tag{5.1.3}$$

This formulation takes into account both the expressons in Equation 5.1.1 and Equation 5.1.2. The vector x of unknowns contains the linear coefficients $a^c, b^c, c \in \mathscr{C}$, and matrix A contains the experimentally determined growth rates $\mu_e, e \in \mathscr{E}$ and ones:

$$\begin{array}{c|cccccc|}
 & \multicolumn{2}{c}{\overbrace{\hphantom{\mu_1 \quad 1}}^{c_1}} & & & \multicolumn{2}{c}{\overbrace{\hphantom{\mu_{N_{\mathscr{E}}} \quad 1}}^{c_{N_{\mathscr{C}}}}} \\
\cline{2-7}
 & \mu_1 & 1 & 0 & \cdots & 0 & 0 \\
N_{\mathscr{E}} & \vdots & & 0 & & & \vdots \\
 & \mu_1 & 1 & 0 & \cdots & 0 & 0 \\
 & \vdots & & \ddots & & & \\
 & 0 & 0 & \cdots & 0 & \mu_{N_{\mathscr{E}}} & 1 \\
N_{\mathscr{E}} & \vdots & & & & & \vdots \\
 & 0 & 0 & \cdots & & \mu_{N_{\mathscr{E}}} & 1 \\
 & \mu_1 & 1 & \cdots & & \mu_{N_{\mathscr{E}}} & 1 \\
N_{\mathscr{E}} & \vdots & & & & & \vdots \\
 & \mu_1 & 1 & \cdots & & \mu_{N_{\mathscr{E}}} & 1 \\
\cline{2-7}
\end{array}
\begin{bmatrix} a^{c_1} \\ b^{c_1} \\ \vdots \\ a^{c_{N_{\mathscr{C}}}} \\ b^{c_{N_{\mathscr{C}}}} \end{bmatrix}
=
\begin{bmatrix} y_{e_1}^{c_1} \\ \vdots \\ y_{e_{N_{\mathscr{E}}}}^{c_1} \\ \vdots \\ y_{e_1}^{c_{N_{\mathscr{C}}}} \\ \vdots \\ y_{e_{N_{\mathscr{E}}}}^{c_{N_{\mathscr{C}}}} \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{5.1.4}$$

where $N_{\mathscr{E}}$ is the number of experiments and $N_{\mathscr{C}}$ the number of compartments. This can be efficiently computed if the problem is reformulated as a quadratic programming problem:

$$\begin{array}{lll} \min\limits_{x} & \frac{1}{2} x^T Q x + c^T x & \\ \text{s.t.} & Gx \quad \leq \quad h & \end{array} \tag{5.1.5}$$

where $Q = A^T A$, $c = -A^T y$, $G = -A$, $h = \vec{0}$.

The result of this parameter estimation procedure for *E. coli* using the data available in [208] can be seen on Figure 5.1.1. The estimated values are:

$$p_{cyt/P}(\mu) = 0.04\mu + 0.73 \tag{5.1.6}$$

$$p_{im/P} = 0.08 \tag{5.1.7}$$

$$p_{om/P}(\mu) = -0.01\mu + 0.09, \qquad \mu \in [0.26, 1.9] \tag{5.1.8}$$

$$p_{p/P}(\mu) = -0.04\mu + 0.10 \tag{5.1.9}$$

$$p_{s/P} = 0.0004 \tag{5.1.10}$$

with Pearson correlation coefficient of $R^2 = 0.9997$.

5.1.2 Percentage of *housekeeping* protein per compartment

An RBA model will never consider *all* the cellular processes. Proteins inolved in the processes that are not represented in the model are designated as *housekeeping* proteins. In order to take into account the fact that the cell needs to produce them, in RBA there is a constraint ensuring that a specific percentage of protein in all compartments will be allocated to these *housekeeping* proteins. They are represented by a single protein of a length and composition which reflect the average length and the average amino acid composition of protens in *E. coli*. Amino acid composition of an average protein can be estimated from experiments determining organism's residue composition, such as was done in [10]. Also, the cellular description in RBA terms might have different levels of detail for different compartments, so it is necessary do describe the percentage of housekeeping protein for each individual compartment. The data that one can use to determine these percentages is a quantitative proteomics experiment, in which proteins have been assigned subcellular localization and a function. In this way, one can "count" the amount of protein (or amino acids, to be more exact) that correspond to functions represented in the model for each compartment. Much like in the subsection 5.1.1, I had the data to estimate this percentage for a span of growth rates in *E. coli*. As the percentage of housekeeping protein in different compartments does not relate in any way (unlike the case of total protein per compartment), this fit can be obtained through a simple least squares procedure for each compartment independently. The percentages of housekeeping protein per compartment were computed as:

$$p^{ne}_{cyt}(\mu) = 0.011\mu + 0.149 \tag{5.1.11}$$

$$p^{ne}_{im}(\mu) = -0.041\mu + 0.243 \tag{5.1.12}$$

$$p^{ne}_{om}(\mu) = 0.017, \qquad \mu \in [0.26, 1.9] \tag{5.1.13}$$

$$p^{ne}_{p}(\mu) = 0.047\mu + 0.022 \tag{5.1.14}$$

$$p^{ne}_{s} = 1 \tag{5.1.15}$$

with Pearson correlation coefficient of $R^2 = 0.954$.

5.1.3 Cytosolic density

It has been experimentally shown that *E. coli* cells have constant buoyant cell density independent of the growth rate [214, 215]. It had been a surprising fact because of the great changes in cell volume and chemical composition that *E. coli* undergoes as the growth rate changes. In RBA this is represented as an assumption on the constant cytosolic density $D_c[\frac{mmol}{gCDW}]$, which stands for a constant volume occupied by cytosolic macromolecules per gram of cell dry weight [187]. The cytosolic macromolecues are metabolic enzymes and ribosomes, thus taking into account the protein and rRNA content. Since proteins make up most part of the cellular space, the cytosolic density is expressed in terms of the concentration of an average amino acid in gram of cellular dry weight - $\frac{mmol.aa}{gCDW}$. Ribosomes take up a significant portion of the cellular volume and are composed not only of protein, but of RNA as well. In order to express how many "average" amino acids are contained in the RNA content of one ribosome, I define a scaling constant $d_{r/R}$. "Average" amino acid (or nucleic acid) is computed as a weighted mean of molar masses of different amino acids (nucleic acids), weighted by their frequency in the *E. coli* cell.

$$\overline{MW}_{NA} = \vec{p_{NA}}^T \vec{MW}_{NA} \qquad \overline{MW}_{AA} = \vec{p_{AA}}^T \vec{MW}_{AA} \tag{5.1.16}$$

where the $\vec{MW}_x$ is a vector containing the molar masses of amino acids (nucleic acids) in g/mol. Therefore, the number of average amino acids in the RNA content of one ribosome - $d_{r/R}$ - relates to the number of nucleic acids in one ribosome $N_{aa/rib}$ and the ratio of $\overline{MW}_{NA}$ and $\overline{MW}_{AA}$:

$$d_{r/R} = N_{aa/rib}\frac{\overline{MW}_{NA}}{\overline{MW}_{AA}} = 4593\frac{340.19}{108.28} = 14430 \tag{5.1.17}$$

To compute the cytosolic density, I have consolidated three relevant datasets [32, 187, 210] and have thus chosen a growth rate which is present in all three datasets, that of $1[h^{-1}]$. Cytosolic density can be expressed as a sum of all the cytosolic protein and of the RNA content of ribosomes converted to average amino acid:

$$D_{cyt} = p_{cyt}(\mu)P_{tot}(\mu) + d_{r/R}R(\mu) \qquad [\frac{mmol.aa}{gCDW}] \tag{5.1.18}$$

Fraction of protein assigned to cytosol $p_{cyt}(\mu)$ is estimated using the package `RBApy.estim` from proteomics data [208] as a linear function of the growth rate. The process of estimation is described in the supplementary text S6, to be found here (link).

$$p_{cyt} = 0.04\mu + 0.73 \tag{5.1.19}$$

What remains to be computed are the total amino acid and ribosome concentration. To compute the total amino acid concentration in $[\frac{mmol}{gCDW}]$, I have used the comprehensive dataset of [32], and express it as:

$$C_{aa/CDW} = N_{aa/CDW}[\frac{\#aa}{\mu gCDW}] \times \frac{1}{N_a}[mol] \times 10^6 \times 10^3 \qquad [\frac{mmol.aa}{gCDW}] \tag{5.1.20}$$

N_a is the Avogadro constant, and $N_{AA/CDW}$ is the number of amino acids in a μg of cell dry weight. The two factors 10^6 and 10^3 serve to scale the $[\mu g]$ to $[g]$ and $[mol]$ to $[mmol]$. For ease of comparison, the abbreviations used here to express the formula for total amino acid count are the same as used in the [32]: P_M stands for protein/mass expressed in units of $10^{17} aa/OD_{460}$, $M_C(\mu g)$ stands for μg of cell dry weight per 10^9 cells, and M_C stands for OD_{460} units per 10^9 cells. I compute $N_{AA/CDW}$ in the following way:

$$N_{aa/CDW} = \frac{P_M}{\frac{M_C(\mu g)}{M_C}} \qquad [\frac{\#AA}{\mu gCDW}] \tag{5.1.21}$$

By inputing Equation 5.1.21 into Equation 5.1.20 I finally obtain a value for total concentration of amino acids:

$$C_{aa/CDW} = \frac{5.2 \times 10^{17}[\frac{\#AA}{OD_{460}}]}{\frac{433[\mu g/10^9\text{cells}]}{2.5[OD_{460}/10^9\text{cells}]}} \times \frac{1}{6.022 \times 10^{23}[mol^{-1}]} \times 10^6 \times 10^3 = 5\frac{mmol.aa}{\mu CDW} \tag{5.1.22}$$

for the growth rate of $\mu = 1[h^{-1}]$.
I compute the ribosome concentration for $\mu = 1[h^{-1}]$ by assuming that the cell has as many active ribosomes R_a as needed to translate the flux of total protein at steady state:

$$k_T R_a = \mu C_{aa/CDW} \tag{5.1.23}$$

where the active ribosomes depend on the growth rate μ and the maturation time $T_{mat/R}$: $[R_a] = e^{-\mu T_{mat/R}}[R] - p_{R_a}[R]$. By taking into account the expression for active ribosomes in Equation 5.1.23, I obtain the following expression for the ribosome concentration:

$$R = \frac{\mu C_{AA/CDW}}{k_T \times p_{R_a}} \tag{5.1.24}$$

The maturation time of ribosomes is assumed to be 5 minutes [212]. The ribosome concentration thus computed is:

$$R = 7.24 \times 10^{-5} \qquad [\frac{mmol}{gCDW}] \tag{5.1.25}$$

By inputing the ribosomal and the total amino acid concentration in Equation 5.1.18 I finally obtain the cytosolic density for $\mu = 1[h^{-1}]$ from to be:

$$D_{cyt} = 4.89 \qquad [\frac{mmol.aa}{gCDW}] \tag{5.1.26}$$

5.1.4 Total amino acid concentration

Total protein content is determined as a linear function of μ, by solving a system of two linear equations with two unknowns for a set of different growth rates. The equations are:

$$\begin{gathered} \mu P_{tot} - k_T R_a = 0 \\ p_{cyt} P_{tot} + d_{r/P} R = D_{cyt} \end{gathered} \tag{5.1.27}$$

The system for each μ is solved as:

$$\begin{bmatrix} \mu & -k_T e^{-\mu t_{mat}} \\ p_{cyt} & d_{r/P} \end{bmatrix} \begin{bmatrix} P_{tot} \\ R \end{bmatrix} = \begin{bmatrix} 0 \\ D_{cyt} \end{bmatrix}$$

When looking at the list of data needed for the computation of P_{tot}, one can see that what is needed are the percentages of cytosolic proteins for each growth rate and the cytosolic density (which is also a function of the percentage of cytosolic protein). The type of data that could be used for such a computation is the proteomics data, such as [208].
After solving this system for a range of values for $\mu = (0.4..0.1..1.9)$, the final linear fit of total protein with respect to the growth rate is:

$$P_{tot} = 5.91 - 1.04\mu$$

5.1.5 Default apparent catalytic rate of enzymes

Default value chosen for enzymatic efficiency is $k_{app} = 12.5s^{-1}$, and was obtained as a best fit for predicted growth rates to growth rates determined experimentally for cells grown in batch cultures for 12 different media [208]. On Figure 5.1.2, I show the differences in growth rate prediction as a consequence of change in the default enzyme efficiency value.

5.1.6 Enzyme specific catalytic rates

Even if the apparent catalytic rates of enzymes are in reality complex functions of, among other things, substrate and product concentrations, temperature, regulation and cofacor availability, all these effects are difficult to measure, especially systematically, for all active enzymes in a particular *in vivo* situation. Due to the lack of suitable experimental data on genome scale, RBA utilizes a simplification for the estimates of k_{app} values [191]. Ideally, apparent catalytic rates of individual enzymes would be estimated either as constants or as linear functions of the growth rate. This requires a series of comaparable proteomics and fluxomics experiments done for a range of growth rates, as done in [191]. For the calibration of the *E. coli* RBA model, no such

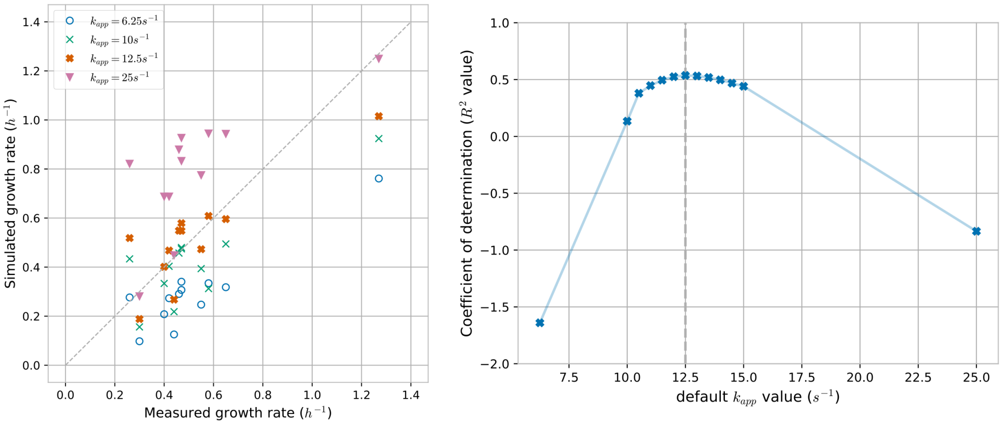


Figure 5.1.2: Left: predictions of growth rate for four different values of default enzyme efficiency. Right: Change in the goodness of fit as a function of the default enzyme efficiency. The best fit is obtained for default $k_{app} = 12.5s^{-1}$.

range of growth rates was available. I were able to identify matching [1] proteomics and fluxomics experiments for a single condition - batch growth on glucose. Proteomics data used was the one measured in [208], while the fluxomics data used was one as measured in [216].

To compute the enzyme specific apparent catalytic rates for the batch growth on glucose, I first used the flux values measured in [216] to constrain an FBA model of the metabolic reconstruction used for the creation of the *E. coli* RNA model. Names of reactions and constraints on fluxes are given in Table A.1.2. Thereby I obtain the values for the metabolic fluxes for which no measurement was available.

Since I had only one proteomics experiment and a single value for the abundance of individual proteins (no data available on biological replicates, for example), our k_{app} estimation results in a very simple formula:

$$k_{app}^{E_i} = \frac{\tilde{v}_i}{[\tilde{E}_i]} \tag{5.1.28}$$

where $\tilde{v}_i$ is the estimated flux catalyzed by enzyme E_i, and $[\tilde{E}_i]$ is the estimated concentration of the enzyme. The proteomics datasets used in the calibration give protein measures in counts per cell. In order to transform this measure in a concentration in $mmol/gCDW$, we needed to know the dry cellular weight for growth on glucose, which I took to be 417.64 $[fg]$ (data taken from Supplementary of [208]). Therefore, the concentration was computed as:

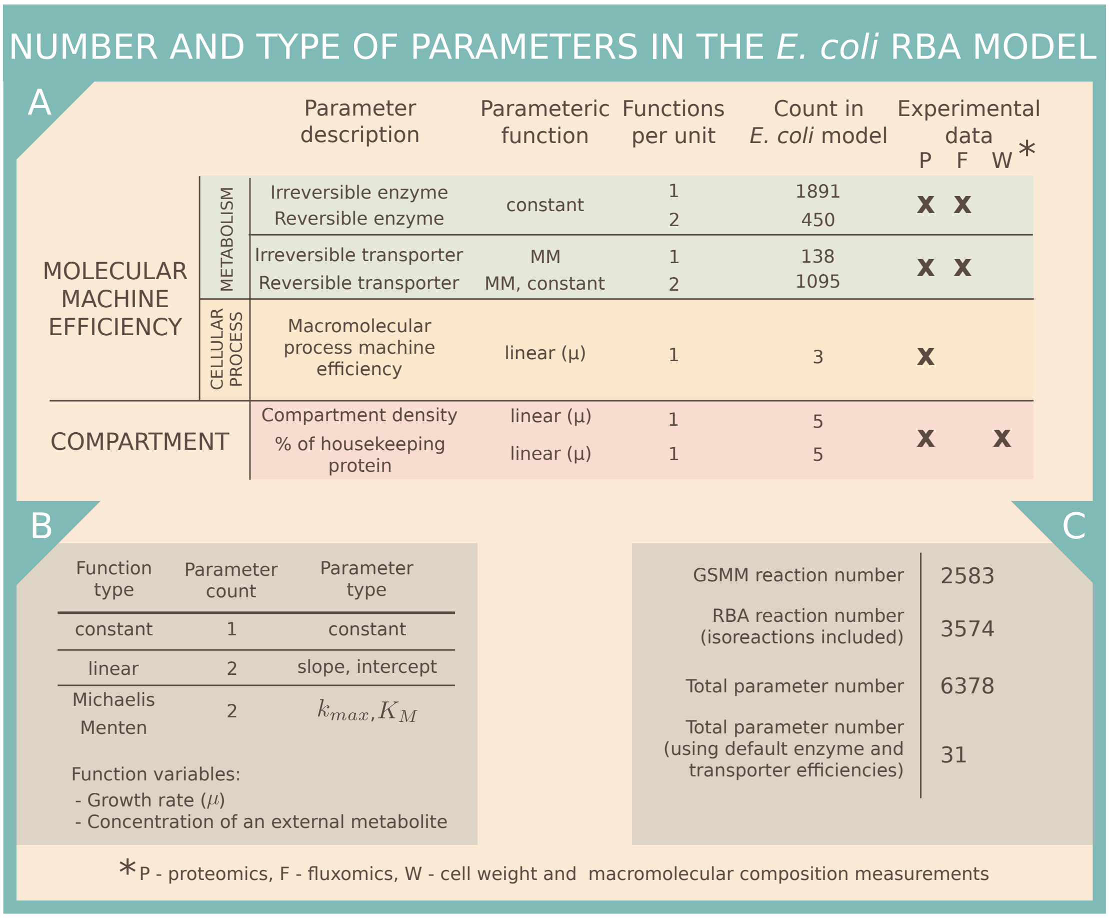
$$[\tilde{E}_i] = \frac{\tilde{E}_i}{CDW \times R} \times 10^3 \qquad [\frac{mmol}{gCDW}] \tag{5.1.29}$$

where $[E_i]$ represents the concentration of the i^{th} enzyme, and E_i the count per cell, and R the Avogadro constant. Since enzymes are often composed not of a single, but of a number of proteins with their corresponding stoichiometry, the enzyme count also needs to be estimated. Assuming that the protein abundances are log-normally distributed, I used the geometric mean of the individual protein abundances (corrected for their stoichiometry in the complex) to estimate the enzyme abundance.

By using this procedure, I obtain the individual k_{app} estimates for 406 enzymes.

---

[1] By *matching* in this context I assume the same strain, same medium and a comparable growth rate.


NUMBER AND TYPE OF PARAMETERS IN THE *E. coli* RBA MODEL
A
B
C

		Parameter description	Parametric function	Functions per unit	Count in *E. coli* model	Experimental data* P	F	W
MOLECULAR MACHINE EFFICIENCY	METABOLISM	Irreversible enzyme	constant	1	1891	X	X	
		Reversible enzyme		2	450			
		Irreversible transporter	MM	1	138	X	X	
		Reversible transporter	MM, constant	2	1095			
	CELLULAR PROCESS	Macromolecular process machine efficiency	linear (μ)	1	3	X		
COMPARTMENT		Compartment density	linear (μ)	1	5	X		X
		% of housekeeping protein	linear (μ)	1	5			

Function type	Parameter count	Parameter type
constant	1	constant
linear	2	slope, intercept
Michaelis Menten	2	k_{max}, K_M

Function variables:
- Growth rate (μ)
- Concentration of an external metabolite

GSMM reaction number	2583
RBA reaction number (isoreactions included)	3574
Total parameter number	6378
Total parameter number (using default enzyme and transporter efficiencies)	31

*P - proteomics, F - fluxomics, W - cell weight and macromolecular composition measurements

Figure 5.1.3: Number and type of parameters in the *E. coli* RBA model. **(A)** The two basic types of parameters in RBA models are the molecular machinery efficiency parameters and the parameters related to the protein occupancy of different compartments. The efficiency parameters for enzymes can be expressed as a constant or as a linear function of the growth rate, while the efficiency of transporters can also be represented as a Michaelis-Menten type function of the concentration of the substrate in the medium. Proteomics and fluxomics data is generally needed for the estimation of these parameters. The efficiencies of process machines are generally expressed as functions of the growth rate and can be estimated from a comprehensive proteomics dataset. The compartment-related parameters are the total amino acid concentration and the concentration of housekeeping protein per compartment. Both are described either as constants or as linear functions of the growth rate. **(B)** RBA supports different function types in describing molecular machinery efficiencies. These can be either constant, linear with respect to growth rate, Michaelis-Menten with respect to the growth rate or substrate concentration. Users can define additional functions if necessary. **(C)** Numbers relating to the *E. coli* model. Compartment-related parameters are always necessary, as well as the process machinery efficiencies. In case all the enzyme apparent catalytic rates are estimated, the *E. coli* RBA model requires 6378 parameters. When there is no suitable fluxomics and proteomics data available, the parameterization of the *E. coli* model can be done so as to estimate the default apparent catalytic rate for all enzymes. In this case, the model requires only 31 parameters.

5.1.7 Efficiencies of process machines

The rate of protein translation was taken directly from the calibration done in [191], since in that paper it was done with data for *E. coli*. The rate is taken to be:

$$k_T = \frac{27\mu}{0.5+\mu} \tag{5.1.30}$$

The efficiencies of other process machineries were estimated using proteomics datasets from [208] for growth on 12 different carbon sources and 1 supplemented with 20 amino acids. I estimate the total amino acid concentration flux that needs to be processed by a particular machinery, and divide it by the abundance of the machinery, obtaining process machinery efficiency in $[\frac{\#AA}{h}]$.

Folding. I consider two major chaperoning systems (GroEL/S and DnaJK) in exponential growing cells and describe them as one cellular process having single machinery composed of the two chaperoning systems in their right stoichiometries. I assume that this process machinery needs to fold 10% of all protein [217]. The total concentration of amino acids that needs to be folded per unit time can be estimated by using the value for total amino acid concentration obtained in subsection 5.1.4.

$$v_{P_{fold}} = \mu \times 0.1 \times P_{tot}(\mu) = 0.31 \qquad [\frac{mmol.aa}{gCDW}] \tag{5.1.31}$$

The total number of amino acids in the chaperone complex consisting of all subunits in their correct stoichiometries (tig, dnaJ, dnaK, groL, groS, grpE) is $N_{AA/ch} = 10829$. Number of measured amino acids of the same complex is $N_{AA/ch/mes} = 3.8 \times 10^7$. Efficiency of the chaperone complex becomes:

$$k_{CH} = \frac{v_{P_{fold}}}{\frac{N_{AA/ch/mes}}{N_{AA/ch}}} \tag{5.1.32}$$

The folding efficiency as a linear function of the growth rate is

$$k_{CH}(\mu) = 7.2\mu + 1.59 \qquad [s^{-1}] \tag{5.1.33}$$

with the coefficient of determination being $R^2 = 0.97$.

Secretion. I model the general secretory *sec* pathway of *Escherichia coli*, since most non-cytosolic proteins are translocated to their compartments via this pathway [218]. The concetration of amino acids to be secreted per unit time will be:

$$v_{P_{sec}} = \mu \times (1 - p_{cyt}(\mu)) \times P_{tot}(\mu) \tag{5.1.34}$$

The rest of the procedure is the same as in the case of folding, and the final linear relation between the growth rate and secretion efficiency is

$$k_{SEC}(\mu) = 118.23\mu - 6.94 \qquad [s^{-1}] \tag{5.1.35}$$

with the coefficient of determination being $R^2 = 0.98$.

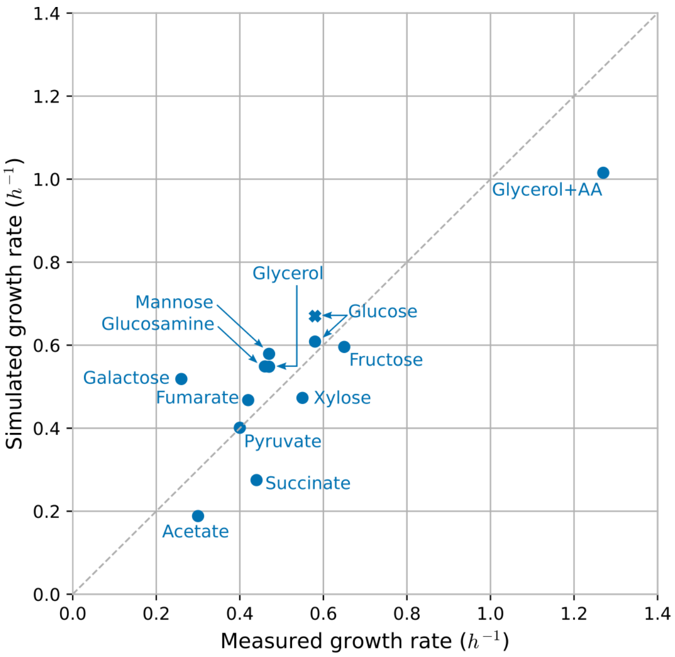


Figure 5.2.1: Prediction of growth rates for 12 different media: 11 minimal media and one medium with glycerol and 20 amino acids. The experimental measurements for the growth rates were taken from [208]. ’X’ marks the prediction for growth on glucose using the estimated apparent catalytic rates for individual enzymes.

5.2 Results

5.2.1 Growth rate prediction

Using the *Escherichia coli* RBA model in which I have calibrated the physiological parameters, but have used the same default apparent catalytic rate for all enzymes of $12.5\frac{1}{s}$ (see subsection 5.1.5), I simulated the growth on twelve different media used in [208].

Eleven of the media are minimal media with a single carbon source, given in the order of increasing growth rate (galactose, acetate, pyruvate, fumarate, succinate, glucosamine, glycerol, mannose, xylose, glucose, fructose), and the twelfth is is a medium with glycerol supplemented with twenty amino acids. As can be seen on the Figure 5.2.1, predictions of the growth rate are good ($R^2 = 0.58$) even without enzyme-specific model calibration.

5.2.2 Predictions for growth on glucose

As explained in subsection 5.1.6, for growth on glucose it was possible to estimate the individual enzyme catalytic rates of 417 enzymes due to the availability of appropriate proteomics [208] and fluxomics [216] datasets in growth conditions in which the growth rate was similar enough to indicate a similar internal cellular organization. Figure 5.2.2 shows the comparison of the predictions obtained here to those obtained by [219] for 183 enzymes present in both datasets. Also, in the Figure 5.2.3, section (A), it is possible to see the comparison of the cummulative histogram of the catalytic rates obtained here with those available in the Brenda [220] database. These catalytic rate values were used to obtain predictions for the flux distribution and the abundances of enzymes and molecular machines. The Figure 5.2.3, section (B) shows the comparison of the measured [216] and predicted flux values for a subset of the central carbon metabolism fluxes and the fluxes of import of glucose and export of acetate. The exchange

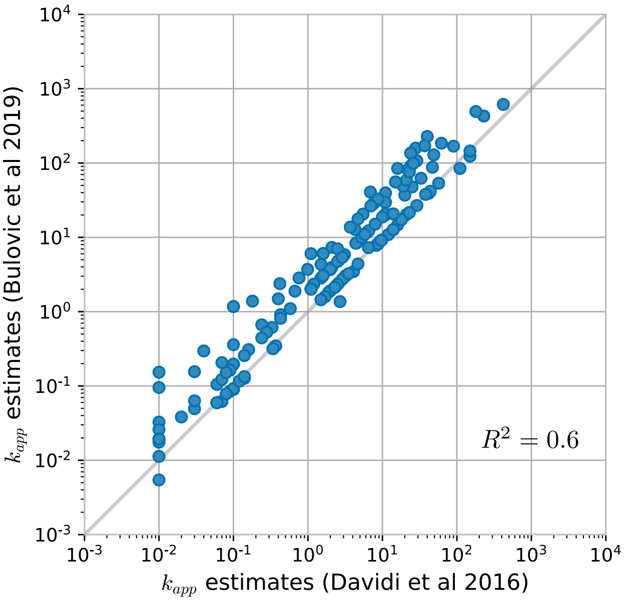


Figure 5.2.2: Comparison of the estimated catalytic rates for 183 enzymes for which [219] also offer a prediction. The coefficient of determination between the two sets of predictions is $R^2 = 0.6$

fluxes are predicted almost exactly. The Figure 5.2.3, sections (C) and (D) show the comparison of the experimentally measured [208] and predicted enzyme and macromolecular machine abundances by using the default and enzyme-specific catalytic rates respectively. There is an obvious improvement in the quality of the predictions once the enzyme-specific catalytic rates are available.

5.2.3 The case of dehydrogenase substitution

The respiratory chain of *Escherichia coli* is highly modular, with a set of molecular species functioning as electron donors (NADH, formate, glucose, hydrogen, pyruvate etc.) and as electron acceptors (oxygen, fumarate, nitrate, nitrite, etc) [221], with dedicated enzymes serving as dehydrogenases and reductases for different substrates. The electrons are passed from dehydrogenases to the reductases via a quinone pool. *E. coli* can not only choose the substrate to use, but for certain substrates it can use different enzymes to achieve different flux of hydrogen ions to the periplasm per molecule of substrate [221]. This modularity enables *E. coli* to adapt to different environmental situations. For example, glucose dehydrogenase, a relatively small enzyme (796 amino acids) could be preferentially used in case of growth on glucose instead of the expensive NADH dehydrogenase (4878 amino acids) if it were not for the cofactor that this enzyme requires for functioning and for which *E. coli* has no biosynthetic pathway - pyrroloquinoline quinone (PQQ). However, if this cofactor is externaly supplied in glucose minimal media, *E. coli* will exhibit chemotaxis towards it, incorporate it, switch to using the cheaper enzyme and grow faster [222, 223]. The enzyme is easily activated, since the PQQ needs to be incorporated on the side of the enzyme facing periplasm [224].

The original metabolic reconstruction used for the creation of the *E. coli* RBA model has no mechanism to import pyrroloquinoline quinone. I have added the necessary metabolic species (external and periplasmic PQQ), as well as the import reactions from the external to

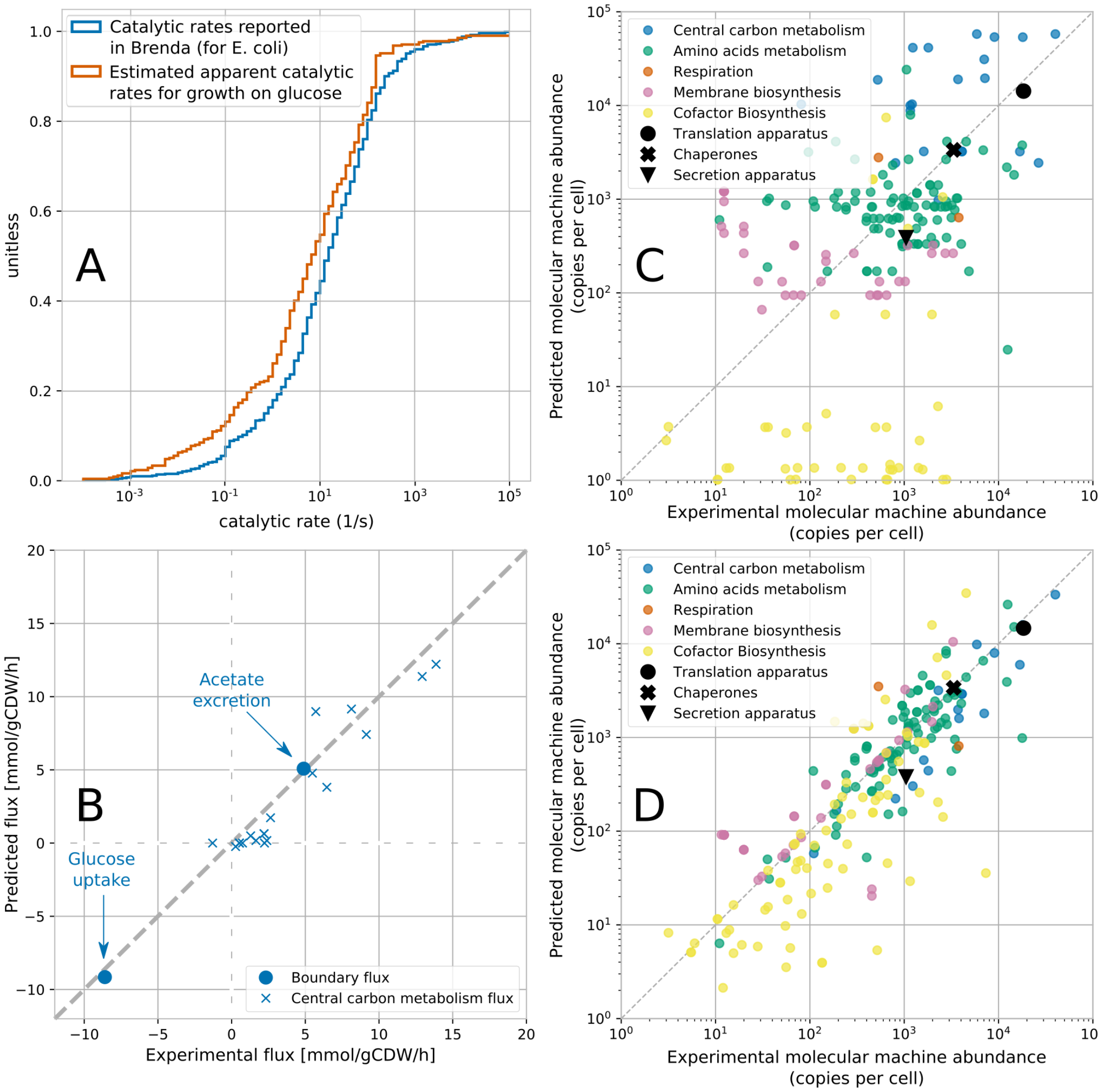


Figure 5.2.3: (A) Cumulative histogram of catalytic rates taken from the Brenda database [220] and the ones obtained by the RBA model calibration in *E. coli*. (B) Comparison of experimentally measured and predicted central carbon metabolism fluxes, as well as glucose import and acetate export fluxes. (C) Comparison of experimentally measured and predicted enzyme abundances using the same default apparent catalytic rate for all enzymes of $12.5s^{-1}$. (D) Comparison of experimentally measured and predicted enzyme abundances using the enzyme-specific apparent catalytic rates obtained through model calibration.

the periplasmic space, one for each of the general outer membrane porins. I have performed simulations for growth on glucose minimal media without and with the PQQ present in the medium. In the absense of PQQ, the model predicts the usage of NADH dehydrogenase and NADPH quinone reductase, coupled with the cytochrome oxidase *bo3*, growing at the growth rate of $\mu = 0.61h^{-1}$. With PQQ present in the medium, the model predicts that the cell will use the combination of glucose and NADH dehydrogenase and grow at an increased growth rate of $\mu = 0.64h^{-1}$.

One can assume that the chemotaxis towards PQQ of *E. coli* could be a consequence of the more efficient utilization of resources for growth followings its uptake. This demonstrates the predictive

Table 5.2.1: Changes introduced into the wild type *E. coli* RBA model to mimic the CO_2-fixing strain [225]. (All metabolite identifiers except $r15b_D$ are available in the BIGG database [227])

Reaction	Enzyme	Modification	Organism
$r15b_D + CO_2 + H_2O \longleftrightarrow 2 \times 3pg + 2 \times H^+$	Rubisco	addition	*R. rubrum*
$r15b + O_2 \longleftrightarrow 3pg + 2 \times pglyc$	Rubisco	addition	*R. rubrum*
$ru5p_D + ATP \longleftrightarrow r15b_D + ADP$	phosphoribulokinase	addition	*S. elongatus*
$HCO_3 + H^+ \longleftrightarrow CO_2 + H_2O$	carbonic anhydrase	addition	*R. rubrum*
$accoa + glx + H_2O \longrightarrow coa + H^+ + mal_L$	malate synthase	removal	/
$icit \longrightarrow glx + succ$	isocytrate lyase	removal	/
$ATP + f6p \longrightarrow ADP + fdp + H^+$	Phosphofructokinase	removal	/
$2pg \longleftrightarrow 3pg$	Phosphoglyc. mutase	removal	/
$g6p + NADP \longleftrightarrow 6pgl + NADHP + H^+$	Glucose 6p dehydr.	removal	/
$glyc_R + ATP \longrightarrow 3pg + ADP + H^+$	Glycerate kinase	removal	/

power of the resource allocation paradigm, as well as the advantage of having a cofactor-specific genome-scale model in which testing of such scenarios is simple.

5.2.4 Simulating the engineered CO_2-fixing *E. coli* strain

This passage serves to illustrate how RBA models can be used to model engineered strains by mimicking the genetic modifications done to the wild type. For this purpose, I have modeled the egineered CO_2 fixing strain developed by [225]. The process of adjusting the model required introducing four new reactions: two catalyzed by the type II Rubisco enzyme (from *Rhodospirillum rubrum ATCC 11170*), one by a phosphoribulokinase (from *Synechococcus elongatus PCC 7942*) and one by a carbonic anhydrase (from *Rhodospirillum rubrum*). To model the deletions reported in [225], I removed two reactions of the glyoxylate shunt: MALS and ICL, two of glycolysis: PFK and PGM and one reaction of the pentose phosphate pathway: G6PDH2r. I additionally removed one reaction of the glyoxylate metabolism (GLYCK) which is not disabled in the engineered strain, but which is reported to be active only during the growth on glycolate as carbon source [226]. The list of all modifications is reported in Table 5.2.1.

I have used the same default apparent catalytic rate $k_{app} = 12.5s^{-1}$ as for the growth-rate simulations (see subsection 5.2.1), except for the enzymes of the carbon fixation. Carbonic anhydrase is known to be among the fastest enzymes, operating close to the diffusion limit, so I set its efficiency to $10000s^{-1}$. Due to lack of specific information of the catalytic rate of phosphoribulokinase, I set it to the default apparent catalytic rate of $12.5s^{-1}$.

$$k^R_{MM}([CO_2]) = \frac{k^{MM}_{max}[CO_2]}{K^{MM}_M + [CO_2]} \tag{5.2.1}$$

$$k^R_{CI}([CO_2],[O_2]) = \frac{k^{CI}_{max}[CO_2]}{K^{CI}_M(1 + [O_2]/K^{CI}_I) + [CO_2]} \tag{5.2.2}$$

Rubisco activity is modeled either as a Michaelis-Menten function of CO_2 concentration (see Equation 5.2.1), either as competitive inhibition by oxygen (see Equation 5.2.2). I assume the carbon dioxide is dissolved under $0.1atm$ and oxygen is present under normal atmospheric conditions. Parameters for the Rubisco activity were taken as minimum, maximum and median values reported in [228]. The values of all the used Rubisco parameters can be found in Table 5.2.2. The updated model is available on the Github pages of the `SysBioInra` group - https://github.com/SysBioInra/Bacterial-RBA-models/tree/master/Escherichia-coli-CO2-fixing.

Table 5.2.2: Predictions of growth rate and percentage of Rubisco in the cytosol for two different enzyme kinetics - Michaelis-Menten function of CO_2 - k^R_{MM} and competitive inhibition by oxygen - k^R_{CI}. ^{a} - Percentage of dry weight of Rubisco in all cytosolic proteins.

Enzyme kinetics			$\mu(h^{-1})$	$p^{cyt}_{Rubisco}(\%)^a$	Comment
$k^R_{MM}([CO_2])$					
k^{MM}_{max}	K^{MM}_M				
1.31	446		0.1827	4.98	Median values taken from Brenda for *R. rubrum*
1.31	14		0.183	4.44	K_M set to median K_C value of [228]
0.32	14		0.1685	17.7	k_{max} set to lowest $k_{cat,C}$ value of [228]
12.6	14		0.1866	0.54	k_{max} set to highest $k_{cat,C}$ value of [228]
$k^R_{CI}([CO_2],[O_2])$					
k^{CI}_{max}	K^{CI}_M	K^{CI}_I			
3.16	14	446	0.1854	1.88	Median of values of [228] for $k_{cat,C}$, K_C and K_O
0.32	14	446	0.17	16.11	k_{max} set to lowest $k_{cat,C}$ value of [228]
12.6	14	446	0.1867	0.48	k_{max} set to highest $k_{cat,C}$ value of [228]

The engineered strain grows at the growth rate of $0.12h^{-1}$ [225], much slower than its wild-type equivalent grows on glucose ($0.65h^{-1}$). The modified *E. coli* model predicts the growth rate between 0.16 to $0.18h^{-1}$, with Rubisco taking up from ~ 1 to 18% of cellular protein, depending on the parameters chosen to model Rubisco activity. When CO_2 is removed from the *in silico* medium, the modified *E. coli* model does not support growth, which shows that CO_2 is indeed a necessary carbon source. In section A.8 I show the changes in the resource allocation between the unmodified cell, and the two versions of the CO_2-fixing *E. coli* - one with high and one with low Rubisco efficiency.

With the example of this modified strain, and the ease with which I obtain realistic predictions of growth rate and the percentage of total protein that is Rubisco without any additional parameter estimation, I show the power of RBA models, and this calbrated *E. coli* model in particular. This model can be thought of a first step towards developing a tool for *in silico* assisted bioengineering experiment design.

5.3 Discussion

In this chapter, I have shown the process of development and parameterization of a genome-scale steady-state cell model of *Escherichia coli* in RBA framework. With a number of different simulations, I have demonstrated the usefulness of such a model in predicting realistic cellular states on one hand, and regulatory events based on parsimonous resource allocation on the other. However, the presented results only begin to cover the ways in which such a model can be used. In this discussion I would like to suggest some future research directions which can be aided by the use of the developed model.

5.3.1 Gratuitous protein production in RBA

Even if this thesis has its original motivation in understanding the issues in recombinant protein production, due to time limitations, I have not been able to fully explore this issue with the *E. coli* RBA model. I can imagine several ways in using such a model for bioproduction. First goes in the direction of analyzing the expression experiment in terms of culture growth and induction. Since the developed RBA model is a steady-state model, in its unmodified form it cannot be used to directly simulate a growth and induction experiment. However, it can be used to estimate the growth rate at which the capacity of the cell to produce gratuitous proten is the highest. In the model, the recombinant target protein can be represented by a *target concentration*. It is possible to find the maximum attainable concentration of the target protein and the corresponding growth rate. This can give us a reasonable approximation to the theoretical upper limit on the yield and some indication on which phase of the growth is most adapted to achieving the highest product concentration.

Secondly, RBA can help in understanding the type and amount of cellular burden imposed on the cell by expression of recombinant protein. One can assume that that burden can be the (i) energetic or (ii) precursor burden, (iii) burden on the process machines necessary for the production of protein (translation, chaperoning, possibly secretion) and (iv) the burden in terms of the occupied cellular space. Such "disection" of the burden can help in understanding what causes the greatest growth defect and therefore has the highest impact on the cells. This can in turn help to make informed decisions on how to alleviate the effects of a particular kind of burden. For example, one can study the changes in the simulated metabolism caused by the overexpression of a protein. The changes in the simulated metabolism can serve to indicate potential overexpression or knockout targets, which could help in adapting the metabolism to the task of producing recombinant protein [229] (instead of to that of growth and proliferation).

5.3.2 Inferring regulation by exploring the RBA model

Resource Balance Analysis models can be used, as showed in subsection 5.2.3, to infer certain types of regulation, which are in place to provide the cell with a more resource efficient solution under specific environmental conditions. Maybe the most famous such example of regulation due to resource allocation which had caused quite some polemic in the field is the so-called "overflow" metabolism, also known as the Warburg effect in cancer cells. The overflow metabolism is a name for a metabolic strategy in which certain microbes, when growing fast, do not completely oxidize the growth substrate through respiration, but through seemingly inefficient substrate utilization excrete a number of "overflow" metabolites. Recently it has been shown that this strategy is in fact more efficient in terms of proteome allocation [198, 230].

RBA model is the perfect *in silico* tool to aid in the understanding of such cellular decisions. One way in which this capacity of the RBA model can be used is to predict the preferece of carbon sources and the underlying regulatory structure of catabolite repression. Under the assumption that the phenomenon of catabolite repression is driven by resource utilization efficiency, one can simulate the growth on all the combinations of a choice of carbon sources. This was done in [231]. By detecting the utilization of transporters for growth on all the combinations of 9 carbon sources ($2^9 = 512$ growth conditions), they were able to predict their utilization hierarchy for *B. subtilis* with almost a perfect match to the experimentally determined one. Additionally, while RBA cannot model the toxicity effects caused by a high presence of certain metabolites or inorganic ions in the medium, it can serve to understand the effects on the cell when certain chemicals necessary for growth are present in low amounts. This can help in understanding the cell-level adaptations *E. coli* goes through in such situations.

III

Dynamics under constraints

6 Dynamics under constraints 111
6.1 The optimal control problem
6.2 Model of adaptation to change in temperature
6.3 Model parameterization
6.4 Model simulation
6.5 Different model versions
6.6 Discussion

6. Dynamics under constraints

In the first part of the thesis, I have described, modeled and analyzed the regulatory network in charge of proteome maintenance in conditions when its quality is compromised, as in the case of heat shock. I have presented my reasoning for thinking why it is necessary to incorporate a *coherent cellular state* when modeling events that bring about big rearrangements in the cell. In the second part, I have presented RBA, a modeling framework capable of representing and simulating coherent cellular states on genome scale. In this final part of the thesis, I will apply the RBA principle of representing coherent cellular states to the study of HSR.

The time evolution of the cellular state will described by a system of ODEs, and the coherency of that state through time will be ensured through a number of linear RBA-like constraints. As in Part II, I assume that the cellular configuration reflects the goal of maximizing the growth rate. In this way, I will study what is the cellular response to the change in temperature under the assumption of parsimonious allocation of resources. By comparing that response to the one obtained by studying the known HSR regulatory network, and to the one determined by experiment, I can address the question of whether the known regulation is in place in order to ensure (near) optimal allocation of resources for growth. A mathematical framework that allows the posing of optimization problems for systems described by ODEs and a number of linear constraints is *optimal control*. I will use this framework for the study of the optimal response of the cell to the change in temperature in terms of parsimonious resource allocation.

Temperature change is a systemic change that perturbs the cell on many levels, the effects of which accumulate from the most basic to the most complex: (a) the rate of diffusion and the osmotic pressure, (b) membrane fluidity and state and stability of macromolecules, which directly relates to (c) their enzymatic activity. Enzymatic activity of central cellular processes (such as the metabolism and translation) then, in turn, influences the production of the rest of the cell. The proposed model will not take into account the changes in diffusion, osmotic pressure, or membrane fluidity, but will instead focus on enzymatic activity and stability of proteins.

In this chapter, I will shortly present the mathematical framework of optimal control, followed by the model of the cellular adaptation to change in temperature. I then present model parameterization and several simulations demonstrating how the optimal allocation of resources seems to be one of the strong guiding principles behind the regulatory network organization. Lastly, I present the simulation software and the different model versions I have developed for this study, ranging from the simplest cellular representation involving only the ribosome and a single metabolic enzyme, to the final model.

6.1 The optimal control problem

Let us imagine a system endowed with specific internal dynamics (a car), with controls that can be externally operated and which influence its behavior (gas and brake pedal). If the problem of interest is how to achieve optimum performance of such as system (get from A to B as fast as possible), then one method for formalizing such a problem can be optimal control. If the system can be described as an optimal control problem, it is possible to obtain the synthesis of optimal controls over time. This problem falls under the domain of optimization problems for continuous dynamical systems.

Dynamic system. Let us imagine a system that can be described as a set of n system states $x_i(t), i = 1, 2, ..., n$, and their time evolution as a set of n ordinary differential equations:

$$\frac{dx_i}{dt} = f(x_1, ..., x_n, u_1, ..., u_m), \qquad i = 1, 2, ..., n \tag{6.1.1}$$

where the variables $u_i, i = 1, 2, ..., r$ are what are called controls, which can vary over time:

$$u_i = u_i(t) \tag{6.1.2}$$

As the name says, controls are used to control the behavior of the system. An example of a control variable is the angle of the gas pedal which can (after a certain transformation) be related to the acceleration of the car (system state).

Admissible controls. In the definition of an optimal control problem, one can impose certain limitations on the controls. This can be done, for example, by describing a set in $\mathbf{R}^m$ to which they are limited, or by describing the type of change over time they can exhibit. This defines a set of *admissible controls* $u(t) \in U$.

Boundary constraints. Within the optimal control problem definition, it is possible to define the boundary constraints on system states, for initial and for terminal time: $x(t_0)$ and $x(t_f)$.

$$\psi(x(t_0), x(t_f)) = 0 \tag{6.1.3}$$

Path constraints. It is also possible to constrain the states variables at time points inside the time interval $[t_0, t_f]$, or over the entire time interval. They can be expressed in terms of equality or inequality:

$$g(x(t), u(t)) \leq 0 \tag{6.1.4}$$

Performance index. A part of the definition of an optimal control problem is the objective (or performance index) to be minimized. This objective can take on a number of forms, depending on the problem definition. The general form of the index is

$$J = \phi[x(t_f), t_f] + \int_0^{t_f} L[x(t), u(t), t]dt \tag{6.1.5}$$

In the case of no integral performance index ($L = 0$), the problem is a *Mayer* optimal control problem, and in the case of no terminal performance index ($\phi = 0$), it is a *Lagrange* optimal control problem. If both ϕ and L are non-zero, the problem is called a *Bolza* problem. Therefore, in a general term, the optimal control problem can be described by the following optimization formulation:

$$\begin{aligned} \min \quad & J(x,u) = \phi[x(t_f), t_f] + \int_0^{t_f} L[x(t), u(t), t]dt \\ s.t. \quad & \dot{x}(t) = f(x(t), u(t), t) \\ & u(t) \in U \\ & g(x(t), u(t)) \leq 0 \\ & \psi(x(t_0), x(t_f)) = 0 \end{aligned} \tag{6.1.6}$$

The synthesis of optimal control can be obtained analytically for some relatively simple systems. For an understanding of how this synthesis is derived, see [232]. For the original derivation of the optimal control problem, see [233]. In this work, we used a numerical simulator Bocop [234] to find the solutions to the optimization problem defined in the next section.

6.2 Model of adaptation to change in temperature

In this section, I describe HSR within a cellular context using the dynamical RBA approach described in [235]. This is a lumped description of the cell designed to model the adaptation to heat shock in bacteria under a set of important cellular constraints. It includes metabolism, protein production and proteome quality control in terms of chaperones (folding) and proteases (degradation). As shown in Figure 6.2.2, the cell imports the single necessary nutrient S_{ext} and converts it to energy, a generic precursor species S and the "biomass" metabolite B. S is then used to produce the rest of the cellular constituent species - the mRNAs and proteins. Metabolism is represented by three enzymes. The so-called *spontaneously folding* enzyme E_{sf} is produced directly in its folded form. The state of this enzyme is assumed not to depend on temperature, as it cannot unfold or aggregate. The *chaperone-assisted* enzyme E_{ca} cannot fold spontaneously but instead depends on chaperones for its folding. It can unfold and when unfolded, can also aggregate. The third enzyme E_{ts} is the *temperature-sensitive* enzyme. It can fold spontaneously, but its folding and unfolding depend on temperature. It can also be assisted in folding by chaperones. The aggregates of both the chaperone-assisted and the temperature-sensitive enzyme can be digested by the protease P and converted back to the precursor species S at the expense of some energy, proportional to the length of the enzyme. The rest of the proteome which is not functionally represented in the model is assumed to take a certain percentage p_{HP} of the cellular protein and is termed housekeeping protein - HP. All of the mentioned macromolecular machines - the ribosome R, three metabolic enzymes - E_{sf}, E_{ca} and E_{ts}, the chaperone C, the protease P and the housekeeping protein HP are produced in proportion to their relative mRNA abundance in the total mRNA pool. All the mRNA species are synthesized directly from the metabolite S without the assistance of specialized machinery, while the proteins require ribosome R to be produced. The set of all protein components produced by the ribosome are:

$$\mathbb{C}_R = \{E_{sf}, E^u_{ca}, E^u_{ts}, R, C, P, HP\} \tag{6.2.1}$$

The corresponding mRNA production fluxes are decision variables denoted as $v_P^{^m\zeta}(t) : \zeta \in \mathbb{C}_R$. The set of all protein components assisted in folding by the chaperone is:

$$\mathbb{C}_C = \{E^u_{ca}, E^u_{ts}\} \tag{6.2.2}$$

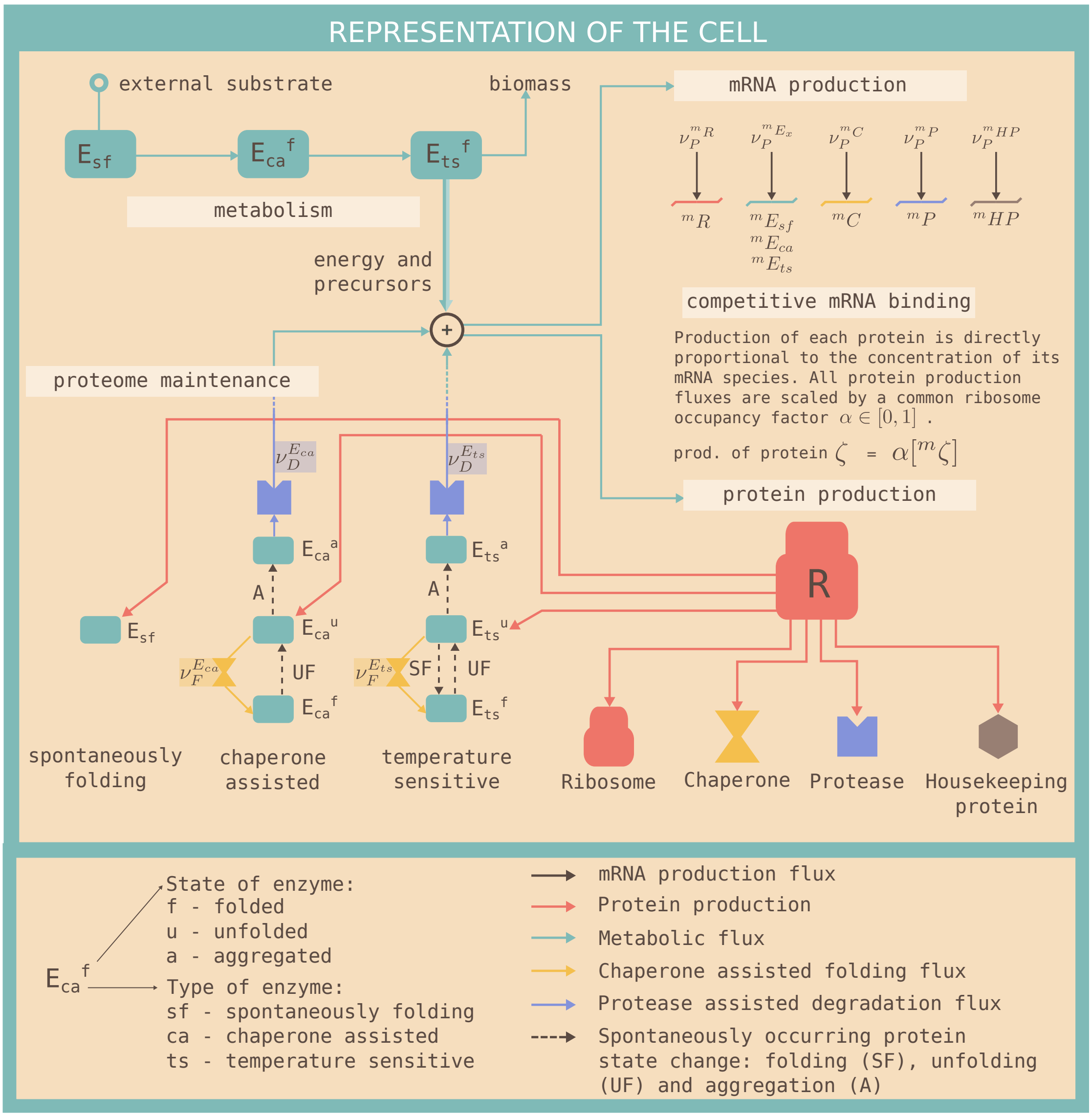
The corresponding folding fluxes are decision variables $v_F^{\zeta}(t) : \zeta \in \mathbb{C}_C$ The set of all protein components that can be degraded by the protease is:

$$\mathbb{C}_P = \{E^a_{ca}, E^a_{ts}\} \tag{6.2.3}$$

The corresponding degradation fluxes are decision variables $v_D^{\zeta}(t) : \zeta \in \mathbb{C}_P$. The concentration of a protein species of a component ζ is designated simply as $[\zeta]$, while its mRNA species is $[^m\zeta]$. The model contains 23 parameters, represented either by constants or by sigmoidal or exponential functions of the temperature T.

This model is graphically depicted in Figure 6.2.1. Stated in this way, it allows us to combine the dynamic changes to the cellular configuration under the influence of change in temperature and resource allocation constraints. As in the steady-state RBA problem, the working assumption is that the cell maximizes its rate of growth μ.

Protein synthesis. Proteins are synthesized from their corresponding mRNA molecules and precursors by the macromolecular machine R - ribosome. The total flux of protein production



Production of each protein is directly proportional to the concentration of its mRNA species. All protein production fluxes are scaled by a common ribosome occupancy factor $\alpha \in [0, 1]$.

prod. of protein ζ = $\alpha[^{m}\zeta]$

Figure 6.2.1: Graphical depiction of the optimal control heat shock model. The cell is made of ribosomes, proteome maintenance machinery (chaperones and proteases), metabolic enzymes and housekeeping protein. Metabolism is composed of three enzyme types, which represent three broad categories of protein with respect to their folding needs: sf - spontaneously folding, ca - chaperone-assisted and cannot fold spontaneously and ts - temperature-sensitive, which are more prone to unfolding as the temperature increases. Metabolic enzymes facilitate the flux of external substrate uptake. This flux is needed to fuel the production of "biomass" species B and protein production substrate species S. Enzymes can spontaneously fold (except the chaperone-assisted one), unfold and aggregate. They can be assisted by chaperones in their folding and disaggregation and can be degraded by proteases in the aggregated state. Degradation produces substrate flux which can again be used in protein production.

is limited by the availability and the efficiency of the ribosome. The flux of the production of a protein ζ will depend on the proportion of its mRNA ${}^m\zeta$ in the total pool of mRNA. This is modeled by making all production fluxes of protein linearly dependent on their corresponding mRNA concentrations, and scaled by a decision variable α. This scaling factor $\alpha : 0 \leq \alpha \leq 1$ is in place to model the possibility that not all mRNAs are being transcribed (in case there is not enough ribosome). But, as the scaling factor is the same for all protein species, all of them will still be transcribed proportionally to the portion of their mRNA in the total mRNA pool. The mRNA species are not produced through the action of any particular molecular machine, but directly from the precursors. They are modeled as voluminous species and take up space in the cell. The flux of production of mRNA ${}^m\zeta$ is a decision variable $\nu_P^{{}^m\zeta}$. The time evolution of mRNA species is given by:

$$[\dot{{}^m\zeta}](t) = \nu_P^{{}^m\zeta}(t) - (\mu(t) + k^m_{deg})[{}^m\zeta](t), \qquad \zeta \in \mathbb{C}_R \tag{6.2.4}$$

where k^m_{deg} represents a faster-than-dilution degradation of mRNA that serves to achieve their short half-lives. The time evolution of all protein species except for the chaperone-assisted enzyme is given by:

$$[\dot{\zeta}](t) = \alpha(t)[{}^m\zeta](t) - \mu(t)[\zeta](t), \qquad \zeta \in \mathbb{C}_R \setminus \{E^u_{ca}, E^u_{ts}\} \tag{6.2.5}$$

As E_{ca} and E_{ts} can undergo unfolding and aggregation, their dynamics are different and are given in the next paragraph. The total production flux of protein is limited by the efficiency and the availability of the translation apparatus R.

$$\alpha(t) \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta [{}^m\zeta](t) \leq k_R(T(t))[R](t) \tag{6.2.6}$$

Proteome maintenance. One of the enzymes, E^u_{ca}, requires chaperones for folding, while E^u_{ts} is not their obligatory substrate but can be assisted in its folding by chaperones. Both of these enzymes are first produced in their unfolded form: E^u_{ca} and E^u_{ts}. Their folded forms are denoted as E^f_{ca} and E^f_{ts}. Once folded, they can unfold with an unfolding rate $k_u(T(t))$ that depends on temperature. When in unfolded form, they can spontaneously assume an aggregated state E^a_{ca} and E^a_{ts}. Aggregation is modeled as if each protein is individually converted into the *aggregated* state, without taking into account the cumulative effect of this phenomenon. This choice was made for the sake of simplicity. A more exact but numerically infeasible approach would be something along the lines of Smochulowski's coagulation [236]. The aggregated proteins can be degraded by the protease into the generic precursor metabolite S. The corresponding dynamics for the different forms of the chaperone-assisted enzyme is:

$$[\dot{E}^u_{ca}](t) = \alpha(t)[{}^mE_{ca}](t) - \nu_F^{E_{ca}}(t) + k_u(T(t))[E^f_{ca}](t) - (\mu(t) + k_{agg})[E^u_{ca}](t) \tag{6.2.7}$$

$$[\dot{E}^f_{ca}](t) = \nu_F^{E_{ca}}(t) - (\mu(t) + k_u(T(t)))[E^f_{ca}](t) \tag{6.2.8}$$

$$[\dot{E}^a_{ca}](t) = k_{agg}[E^u_{ca}](t) - \nu_D^{E_{ca}}(t) - \mu(t)[E^a_{ca}](t) \tag{6.2.9}$$

The dynamics for the temperature-sensitive enzyme is different in that it can spontaneously fold with a temperature-dependent folding rate $k_f(T(t))$, and in its aggregated form it cannot be rescued by the chaperone, but only degraded by the protease.

$$[\dot{E_{ts}^{u}}](t) = \alpha(t)[^{m}E_{ts}](t) - v_F^{E_{ts}}(t) + k_u(T(t))[E_{ts}^{f}](t) - (\mu(t) + k_{agg} + k_f(T(t)))[E_{ts}^{u}](t) \tag{6.2.10}$$

$$[\dot{E_{ts}^{f}}](t) = k_f(T(t))[E_{ts}^{u}](t) + v_F^{E_{ts}}(t) - (\mu(t) + k_u(T(t)))[E_{ts}^{f}](t) \tag{6.2.11}$$

$$[\dot{E_{ts}^{a}}](t) = k_{agg}[E_{ts}^{u}](t) - v_D^{E_{ts}}(t) - \mu(t)[E_{ts}^{a}](t) \tag{6.2.12}$$

The folding fluxes are limited by the availability of the chaperone:

$$\sum_{\zeta}^{\zeta \in \mathbb{C}_C} n_\zeta v_F^{\zeta}(t) \leq k_C[C](t) \tag{6.2.13}$$

Degradation of aggregates is limited by the availability of the protease:

$$\sum_{\zeta}^{\zeta \in \mathbb{C}_P} n_\zeta v_D^{\zeta}(t) \leq k_P[P](t) \tag{6.2.14}$$

Metabolism. The cell is taking up external nutrient S_{ext} whose concentration is assumed to be constant. The flux of conversion of the external nutrient (v_M) is facilitated by three metabolic enzymes: E_{sf} and E_{ca}^{f} and E_{ts}^{f}, and is limited by their availability:

$$v_M(t) \leq k_{M_{sf}}(T(t))[E_{sf}](t) \tag{6.2.15}$$

$$v_M(t) \leq k_{M_{ca}}(T(t))[E_{ca}^{f}](t) \tag{6.2.16}$$

$$v_M(t) \leq k_{M_{ts}}(T(t))[E_{ts}^{f}](t) \tag{6.2.17}$$

The uptake flux is converted into three metabolic fluxes - one of energy production $v_M^E(t)$, one of precursor production $v_M^P(t)$, and one of "biomass" production $v_M^B(t)$, which produces the biomass species B. The three metabolic fluxes $(v_M^E(t), v_M^P(t), v_M^B(t))$ sum up to the flux of uptake of the external metabolite, under appropriate stoichiometries:

$$v_M(t) = n_B v_M^B(t) + n_{M_p} v_M^P(t) + n_{M_e} v_M^E(t) \tag{6.2.18}$$

The time evolution of the biomass species B is described as:

$$[\dot{B}](t) = n_B v_M^B(t) - (\mu(t) + C_B)[B](t) \tag{6.2.19}$$

where C_B is a positive constant that ensures a requirement for biomass even at zero growth rate. I assume that the cell needs to maintain a constant concentration of $B = B_0$. The resulting flux expression becomes:

$$v_M^B(t) = \frac{B_0}{n_B}(\mu(t) + C_B) = \tilde{B_0}(\mu(t) + C_B) \tag{6.2.20}$$

Since it is completely determined by the growth rate, $v_M^B(t)$ is not a control variable of the system. Each of the other two metabolic fluxes, $v_M^P(t)$ and $v_M^E(t)$, takes up a predetermined portion of the remaining of the metabolic flux when the biomass production flux has been deducted:

$$\begin{aligned} v_M^E(t) &= p_E\Big(v_M(t) - n_B v_M^B(t)\Big) \\ &= p_E\Big(v_M(t) - B_0(\mu(t) + C_B)\Big) \end{aligned} \tag{6.2.21}$$

$$v_M^P(t) = (1 - p_E)\Big(v_M(t) - B_0(\mu(t) + C_B)\Big) \qquad p_E \in (0,1) \tag{6.2.22}$$

The energy flux v_M^E is used for both production and degradation of macromolecular species[1]. This leads to the following constraint:

$$\begin{aligned} v_M^E(t) = &\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta\left(\gamma_p^m v_P^{^m\zeta}(t) + \gamma_p^p\alpha(t)[^m\zeta](t)\right) \\ &+\gamma_d^m k_{deg}^m \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[^m\zeta](t) + \gamma_d^p \sum_{\eta}^{\eta\in\mathbb{C}_P} n_\eta v_D^\eta(t) \end{aligned} \tag{6.2.23}$$

where γ_p^m, γ_p^p, γ_d^m and γ_d^p are nonnegative constants that scale the energy cost of mRNA and protein production and degradation according to the length of the corresponding macromolecules.

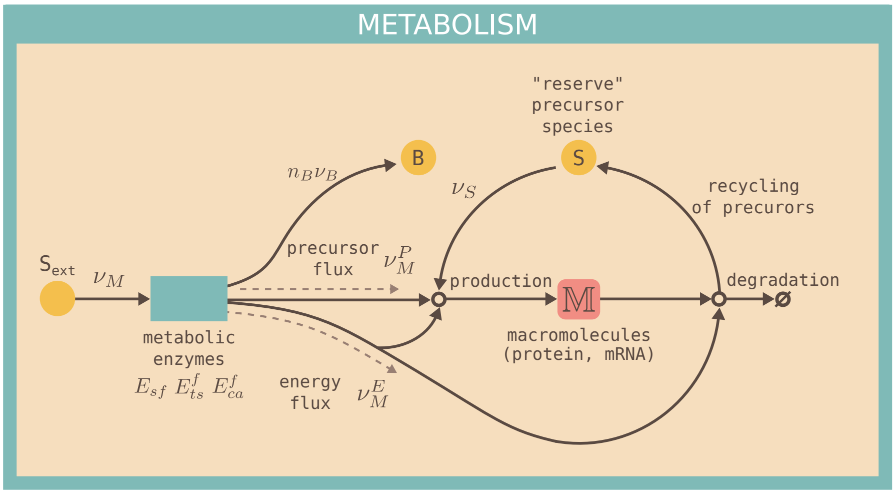


Figure 6.2.2: Metabolic part of the temperature response model. External substrate uptake flux v_M is split into three fluxes (under appropriate stoichiometric relations): (i) v_B - the flux of production of biomass metabolite B, v_M^E - energy flux and v_M^P - precursor flux. Energy is required for both the production and degradation of macromolecular species, while the precursors are consumed by the production, and released by degradation. The precursors released through degradation are accumulated in the precursor storage metabolite S. The flux v_S from metabolite S is used for the production of macromolecular species.

The precursor flux v_M^P is used in the production of new macromolecular species. The degradation of those species feeds into the pool of a *reserve* voluminous precursor species S. This species participates in the macromolecular density of the cell and can be utilized in the creation of new macromolecular species. The dynamics of the species S is described by:

$$\begin{aligned} [\dot{S}](t) = &-v_S(t) - \mu(t)[S](t) \\ &+ \underbrace{k_{deg}^m \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[^m\zeta](t)}_{\text{mRNA degradation}} + \underbrace{\sum_{\eta}^{\eta\in\mathbb{C}_P} n_\eta v_D^\eta(t)}_{\text{protein degradation}}, \qquad [S](t)\geq 0, t\in[t_0,t_f] \end{aligned} \tag{6.2.24}$$

where v_S is the flux of usage of metabolite S for the creation of new macromolecular species. This flux, together with the metabolic precursor flux v_M^P must equal the production of all new

---

[1] If necessary, this energy flux can be required in other cellular processes as well, such as protein folding or degradaton.

macromolecules:

$$v_S(t) + v_M^P(t) = \underbrace{\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta \left(v_P^{{}^m\zeta}(t) + \alpha(t)[{}^m\zeta](t) \right)}_{\text{mRNA and protein production}} \tag{6.2.25}$$

The schematic of the metabolism is available in Figure 6.2.2.

Temperature dependence. The temperature change is introduced in the model through temperature dependence of a number of model parameters. This is the case for the peptide elongation rate of ribosomes ($k_R(T(t))$), protein folding and the unfolding rates ($k_f(T(t))$, $k_u(T(t))$) and metabolic rates of enzymes ($k_{M_{ca}}(T(t))$, $k_{M_{sf}}(T(t))$). These parameters were made dependent on temperature because of their importance in the HSR, and because of the availability of data which allowed me to estimate them. Some parameters, such as the rate of protein aggregation (k_{agg}), rate of folding by the chaperone (k_C) and the rate of degradation by the protease (k_P), even if they most probably are temperature dependent, were described by constants since there was no data available which would allow me to estimate them. Temperature is represented in the model by a system variable $T(t)$, described by the first-order response system around the external temperature value T_{ext}:

$$\dot{T}(t) = -s_T(T(t) - T_{ext}) \tag{6.2.26}$$

where s_T is a scaling factor determining the speed at which the internal temperature reaches a new steady state upon achange in external temperature T_{ext}.

Growth rate. I derive the expression for the growth rate from the rate of change of volume and concentrations of voluminous species. Voluminous species, those that take up volume in the cell, are assumed to be all the macromolecular species and the precursor metabolite S. The set of all macromolecular species are $\mathbb{M}$:

$$\mathbb{M} = \{\zeta : \zeta \in \mathbb{C}_R\} \cup \{{}^m\zeta : \zeta \in \mathbb{C}_R\} \cup \{E_{ca}^f, E_{ca}^a, E_{ts}^f, E_{ts}^a\} \tag{6.2.27}$$

and the set of all voluminous species:

$$\mathbb{V} = \{\zeta : \zeta \in \mathbb{M} \cup \{S\}\} \tag{6.2.28}$$

It is reasonable to assume that the volume is proportional to the weighted sum of voluminous species

$$V(t) = \beta \sum_{x}^{i\in\mathbb{V}} n_x N_x(t) \tag{6.2.29}$$

where n_x is the stoichiometric coefficient denoting the number of units of precursor S stored in the species x (for species S the stoichiometric coefficient is $n_S = 1$), N_x is the amount of the voluminous species $x, x \in \mathbb{V}$ and β is a constant that converts the amount of the precursor S into units of volume. In this model, however, there is no explicit mention of species amounts, but their concentrations. Therefore, instead of relating the volume to species amounts, it is more convenient to relate the rate of volume change to species concentrations. The rate of change of the total volume taken up by a bacterial culture can be described as:

$$\frac{dV(t)}{dt} = \mu(t)V(t) \tag{6.2.30}$$

where $\mu(t)$ is what is commonly referred to as the *growth rate*. By taking a time derivative of both sides of Equation 6.2.29, we get:

$$\frac{dV(t)}{dt} = \mu(t)V(t) = \beta \sum_{x}^{x\in\mathbb{V}} n_x \frac{dN_x(t)}{dt} \tag{6.2.31}$$

or simply:

$$\mu(t) = \beta \sum_{x}^{x\in\mathbb{V}} n_x \frac{1}{V(t)} \frac{dN_x(t)}{dt} \tag{6.2.32}$$

The expression $dN_x(t)/dt$ can be substituted with appropriate terms from the expression of a time derivative of the concentration of species x:

$$\frac{dC_x(t)}{dt} = \frac{1}{V(t)}\frac{dN_x(t)}{dt} - \frac{N_x(t)}{V(t)}\frac{dV(t)/dt}{V(t)} = \frac{1}{V(t)}\frac{dN_x(t)}{dt} - \mu(t)C_x(t) \tag{6.2.33}$$

where $C_x(t)$ is the concentration of species x. I next substitute $\frac{1}{V(t)}\frac{dN_x}{dt}$ with the expression from Equation 6.2.33:

$$\mu(t) = \beta \sum_{x}^{i\in\mathbb{V}} n_x\left(\frac{dC_x(t)}{dt} + \mu(t)C_x(t)\right) \tag{6.2.34}$$

The dynamics of all the voluminous species can generally be described by the terms of production, conversion, degradation and dilution:

$$\frac{dC_x(t)}{dt} = \nu_{P_x}(t) + \nu_{C_x}(t) - \nu_{D_x}(t) - \mu(t)C_x(t), \qquad \nu_{P_x}, \nu_{D_x} \geq 0 \tag{6.2.35}$$

where $\nu_{P_x}(t)$ is the flux of production of species x, $\nu_{C_x}(t)$ is the flux of conversion of species x to or from another voluminous species and $\nu_{D_x}(t)$ is the flux of degradation of species x. Degradation here assumes the removal of a voluminous species. One example of this would be an export to a different compartment or the external medium. By introducing this expression into Equation 6.2.34, the expression becomes:

$$\begin{aligned}\frac{dV(t)/dt}{V(t)} &= \mu(t) = \beta \sum_{x}^{x\in\mathbb{V}} n_x(\nu_{P_x}(t) + \nu_{C_x}(t) - \nu_{D_x}(t) - \mu(t)C_x(t) + \mu(t)C_x(t)) \\ &= \beta \sum_{x}^{x\in\mathbb{V}} n_x(\nu_{P_x}(t) + \nu_{C_x}(t) - \nu_{D_x}(t))\end{aligned} \tag{6.2.36}$$

In this model, all the voluminous species have zero production and degradation flux. Production of macromolecules from precursors, and their degradation back into the precursors are both conversion fluxes, since no macromolecular volume is thereby lost (assuming that the stoichiometry of the two steps is equal). The only flux that can be classified as a production flux is the precursor uptake flux ν_M^P, since it contributes to the accumulation of voluminous species in the cell. The sum of this flux and v_S, the flux that determines the rate of conversion of accumulated internal substrate S into macromolecules (see Equation 6.2.24 and Figure 6.2.2), is then distributed for the production of all the macromolecular voluminous species.

$$
\begin{align}
\sum_{i}^{i\in\mathbb{V}} v_{C_x}(t) = &- \underbrace{v_S(t)}_{\text{conversion into macromolecules}} \underbrace{+k^m_{deg}\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[{}^m\zeta](t)}_{\text{mRNA degradation}} + \underbrace{\sum_{\eta}^{\eta\in\mathbb{C}_P} n_\eta v_D^\eta(t)}_{\text{protein degradation}} \tag{S}\\
&+ \underbrace{\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta v_P^{{}^m\zeta}(t)}_{\text{production}} - \underbrace{k^m_{deg}\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[{}^m\zeta](t)}_{\text{degradation}} \tag{mRNA}\\
&+ \underbrace{\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta \alpha(t)[{}^m\zeta](t)}_{\text{production}} - \underbrace{\sum_{\eta}^{\eta\in\mathbb{C}_P} n_\eta v_D^\eta(t)}_{\text{degradation}} \tag{protein}\\
&- v_F^{E_{ca}}(t) + k_u(T(t))[E^f_{ca}](t) - k_{agg}[E^u_{ca}](t) \tag{E-ca-u}\\
&+ v_F^{E_{ca}}(t) - k_u(T(t))[E^f_{ca}](t) \tag{E-ca-f}\\
&+ k_{agg}[E^u_{ca}](t) \tag{E-ca-a}\\
&- v_F^{E_{ts}}(t) + k_u(T(t))[E^f_{ts}](t) - k_{agg}[E^u_{ts}](t) \tag{E-ts-u}\\
&+ v_F^{E_{ts}}(t) - k_u(T(t))[E^f_{ts}](t) \tag{E-ts-f}\\
&+ k_{agg}[E^u_{ca}](t) \tag{E-ts-a}\\
&= -v_S(t) + \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta v_P^{{}^m\zeta}(t) + \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta \alpha(t)[{}^m\zeta](t) \tag{6.2.37}
\end{align}
$$

By replacing $v_S(t)$ with the expression from Equation 6.2.25, whereby $v_S(t)$ is:

$$
v_S(t) = \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta \Big(v_P^{{}^m\zeta}(t) + \alpha(t)[{}^m\zeta](t) \Big) - v_M^P(t) \tag{6.2.38}
$$

I obtain the final expression for the sum of conversion fluxes:

$$
\sum_{x}^{x\in\mathbb{V}} v_{C_x}(t) = v_M^P(t) \tag{6.2.39}
$$

The precursor uptake flux $v_M^P(t)$ is the final result of the sum of all the conversion fluxes, and thereby the only flux contributing to the expression of the growth rate. The expression for the growth rate $\mu(t)$ becomes:

$$
\mu(t) = \beta v_M^P(t) = \frac{1}{D_c} v_M^P(t) \tag{6.2.40}
$$

where $D_c[\frac{mmol.AA}{gCDW}]$ is the so-called cytosolic density, described in detail in subsection 5.1.3. This quantity is not the only one that could be used to convert concentrations of chemical components into volume, but it is a convenient one. One great benefit is that D_c has already been estimated for the whole-cell *E. coli* RBA model. Additionally, using the same quantity allows for an easier comparison of the two models.

Density constraint. At the initial time t_0, the sum of all voluminous species must not exceed the maximal cytosolic density D_c:

$$\sum_{\zeta}^{\zeta \in \mathbb{V}} n_{\zeta}[\zeta](t_0) \leq D_c \tag{6.2.41}$$

Performance index. The interest of this model is to study cellular resource rearrangement under conditions of changing temperature and the assumption of growth rate maximization. Due to the formulation of the Mayer type optimal control problem, the performance index can be stated as a function of system states at the final time of the simulation:

$$J = \phi[x(t_f), t_f] \tag{6.2.42}$$

Since the growth rate μ is not a system variable (see Equation 6.2.40), it cannot be used directly in the formulation of the performance index. Additionally, maximizing the maximum growth rate at final time t_f would not ensure the maximization of the growth rate throughout the simulation. This can be circumvented by introducing the population size X as a system variable:

$$\frac{dX(t)}{dt} = \mu(t)X(t) \tag{6.2.43}$$

The maximization of the population size at final time would then correspond to the maximization of the growth rate over the entire simulation time interval. This, however, would introduce an exponentially growing variable into the system, which could lead to numerical difficulties during simulation. I solve this by introducing a *proxy* variable, which does not directly correspond to the population size, but ensures that the objective of the optimization corresponds to the maximization of growth rate over the entire simulation time interval:

$$\frac{dX(t)}{dt} = \mu(t) \tag{6.2.44}$$

The performance index then becomes

$$J = X(t_f) \tag{6.2.45}$$

6.2.1 Compact model representation

Our system is described by a system of differential equations:

$$\dot{x}(t) = f(x(t), u(t)) \tag{6.2.46}$$

where $x(t)$ are system variables, and $u(t)$ system controls. Changes in system variables over time are described by the following ODEs:

$$\begin{aligned}
[{}^{m}\dot{\zeta}](t) =& \nu_P^{{}^{m}\zeta}(t) - (\mu(t) + k_{deg}^{m})[{}^{m}\zeta](t), && \zeta \in \mathbb{C}_R && (6.2.47)\\
[\dot{\zeta}](t) =& \alpha(t)[{}^{m}\zeta](t) - \mu(t)[\zeta](t), && \zeta \in \mathbb{C}_R \setminus \{E_{ca}^{u}, E_{ts}^{u}\} && (6.2.48)\\
[\dot{E}_{ca}^{u}](t) =& \alpha(t)[{}^{m}E_{ca}](t) - \nu_F^{E_{ca}}(t) + k_u(T(t))[E_{ca}^{f}](t) && &&\\
& - (\mu(t) + k_{agg})[E_{ca}^{u}](t) && && (6.2.49)\\
[\dot{E}_{ca}^{f}](t) =& \nu_F^{E_{ca}}(t) - (\mu(t) + k_u(T(t)))[E_{ca}^{f}](t) && && (6.2.50)\\
[\dot{E}_{ca}^{a}](t) =& k_{agg}[E_{ca}^{u}](t) - \nu_D^{E_{ca}}(t) - \mu(t)[E_{ca}^{a}](t) && && (6.2.51)\\
[\dot{E}_{ts}^{u}](t) =& \alpha(t)[{}^{m}E_{ts}](t) - \nu_F^{E_{ts}}(t) + k_u(T(t))[E_{ts}^{f}](t) && && (6.2.52)\\
& - (\mu(t) + k_{agg} + k_f(T(t)))[E_{ts}^{u}](t) && &&\\
[\dot{E}_{ts}^{f}](t) =& k_f(T(t))[E_{ts}^{u}](t) + \nu_F^{E_{ts}}(t) - (\mu(t) + k_u(T(t)))[E_{ts}^{f}](t) && && (6.2.53)\\
[\dot{E}_{ts}^{a}](t) =& k_{agg}[E_{ts}^{u}](t) - \nu_D^{E_{ts}}(t) - \mu(t)[E_{ts}^{a}](t) && && (6.2.54)\\
[\dot{S}](t) =& -\nu_S(t) - \mu(t)[S](t) + && &&\\
& + k_{deg}^{m} \sum_{\zeta}^{\zeta \in \mathbb{C}_R} n_{\zeta}[{}^{m}\zeta](t) + \sum_{\eta}^{\eta \in \mathbb{C}_P} n_{\eta}\nu_D^{\eta}(t) && && (6.2.55)\\
\dot{X}(t) =& \mu(t) && && (6.2.56)
\end{aligned}$$

The system controls $u(t)$ are:

- $\nu_M(t)$ - flux of uptake of external substrate
- $\nu_S(t)$ - use of internal metabolite S as a precursor for the creation of new macromolecules
- $\nu_P^{{}^{m}\zeta}(t)$, $\zeta \in \mathbb{C}_R$ - the production fluxes of all mRNA species
- $\nu_F^{E_{ts}}(t)$, $\nu_F^{E_{ca}}(t)$ - folding fluxes of the temperature sensitive and chaperone-assisted enzyme
- $\nu_D^{E_{ts}}(t)$, $\nu_D^{E_{ca}}(t)$ - degradation fluxes of the aggregated form of E_{ts} and E_{ca}
- $\alpha(t) \in [0,1]$ - ribosome pool occupancy scaling factor

The boundary conditions of this system $\Phi(x(t_0), t_0) = 0$ are:

$$\sum_{\zeta}^{\zeta \in \mathbb{C}_R} n_{\zeta}\Big([\zeta](t_0) + [{}^{m}\zeta](t_0)\Big) + n_{E_{ca}}([E_{ca}^{f}](t_0) + [E_{ca}^{a}](t_0)) - D_c \leq 0 \tag{6.2.57}$$

$$X(t_0) - X_0 = 0 \tag{6.2.58}$$

$$T(t_0) - T_0 = 0 \tag{6.2.59}$$

The equality and inequality path constraints, $C_{eq} = 0$ and $C_{in} \leq 0$, of this system are:

$$
\begin{aligned}
&\texttt{(g.r.)} && D_c\mu(t) - n_S\nu_M^P(t) = 0 && \text{(6.2.60)}\\
&\texttt{(met-sf)} && \nu_M(t) \leq k_{M_{sf}}(T(t))[E_{sf}](t) && \text{(6.2.61)}\\
&\texttt{(met-ca)} && \nu_M(t) \leq k_{M_{ca}}(T(t))[E_{ca}^f](t) && \text{(6.2.62)}\\
&\texttt{(met-ts)} && \nu_M(t) \leq k_{M_{ts}}(T(t))[E_{ts}^f](t) && \text{(6.2.63)}\\
&\texttt{(P-flux)} && \nu_M^P(t) - (1-p_E)\Big(\nu_M(t) - B_0(\mu(t)+C_B)\Big) = 0 && \text{(6.2.64)}\\
&\texttt{(E-flux)} && \nu_M^E(t) - p_E\Big(\nu_M(t) - B_0(\mu(t)+C_B)\Big) = 0 && \text{(6.2.65)}\\
&\texttt{(P-met)} && \nu_S(t) + \nu_M^P(t) - \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta\Big(\nu_P^{^m\zeta}(t) - \alpha(t)[^m\zeta](t)\Big) = 0 && \text{(6.2.66)}\\
&\texttt{(E-met)} && \nu_M^E(t) - \sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta\Big(\gamma_p^m\nu_P^{^m\zeta}(t) + \gamma_p^p\alpha(t)[^m\zeta](t)\Big) && \\
& && -\gamma_d^m k_{deg}^m\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[^m\zeta](t) + \gamma_d^p\sum_{\eta}^{\eta\in\mathbb{C}_P} n_\eta\nu_D^\eta(t) = 0 && \text{(6.2.67)}\\
&\texttt{(ribo)} && \alpha(t)\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta[^m\zeta](t) \leq k_R(T(t))[R](t) && \text{(6.2.68)}\\
&\texttt{(chap)} && \sum_{\zeta}^{\zeta\in\mathbb{C}_C} n_\zeta\nu_F^\zeta(t) \leq k_C[C](t) && \text{(6.2.69)}\\
&\texttt{(prot)} && \sum_{\zeta}^{\zeta\in\mathbb{C}_P} n_\zeta\nu_D^\zeta(t) \leq k_P[P](t) && \text{(6.2.70)}\\
&\texttt{(HP-ss)} && \alpha(t)[^mHP](t) - \mu(t)[HP_0] = 0 && \text{(6.2.71)}\\
& && && \text{(6.2.72)}
\end{aligned}
$$

The optimization problem is:

$$
\begin{aligned}
\min_{\{\nu_P^{^m\zeta}(t):\zeta\in\mathbb{P}_R\},\{\nu_F^\zeta(t):\zeta\in\mathbb{P}_C\},\{\nu_D^\zeta(t):\zeta\in\mathbb{P}_P\},\nu_M(t),\nu_S(t),\alpha(t)} \quad & -X(t_f)\\
\text{s.t.} \quad & \dot{x}(t) = f(x(t),u(t))\\
& \Phi(x(t_0),t_0) = 0\\
& C_{eq}(x(t),u(t)) = 0\\
& C_{in}(x(t),u(t)) \leq 0 \qquad \text{(6.2.73)}
\end{aligned}
$$

6.3 Model parameterization

In this work, I take into account the fact that temperature affects the rate of translation, the catalytic rates of enzymes, and the rates of folding and unfolding of proteins. Ribosomes are assumed to be produced in a properly folded form, regardless of the temperatures, as are chaperones, proteases and the spontaneously folding enzyme E_{sf}. The unfolding of the two other enzymes, E_{ca} and E_{ts}, depends on temperatures, as does folding of the temperature-sensitive enzyme E_{ts}. All the parameters (temperature dependent and otherwise) of the model are given in Table 6.4.1.

6.3.1 Kinetic rate of enzymes

In choosing parameter values for the kinetic rates of all the enzymes $k_{app}(T)$ as a function of temperature, I have not used a single set of measurements. Instead, I have combined a number of published observations about the quantity and the type of change in the catalytic rate with temperature. The reason behind this is that there is no genome-wide estimate of temperature dependence on the catalytic rates of enzymes. The individual enzyme studies, due to the very specific and unique nature of protein molecules, can hardly be extrapolated onto genome-scale. For this reason, I have chosen to assume a catalytic rate of $k_{app}(T = 37°C) = 12.5s^{-1}$. This rate has been estimated as the best fit for a range of growth rates of the genome-scale RBA model of *E. coli* in subsection 5.1.5 (see also Figure 5.1.2). I then take into account the observation made in [237] that the catalytic rates of enzymes roughly doubles from every 10°C, further assuming that $k_{app}(T = 27°C) = 6.25s^{-1}$ and $k_{app}(T = 47°C) = 25s^{-1}$. For $T = 0°C$, I take the apparent catalytic rate to be 0. Arrhenius equation describes the exponential increase in the rate of reaction with temperature. For enzymes, however, it has been shown that this exponential relation breaks down after a certain temperature [238]. The potential reason behind it is that at

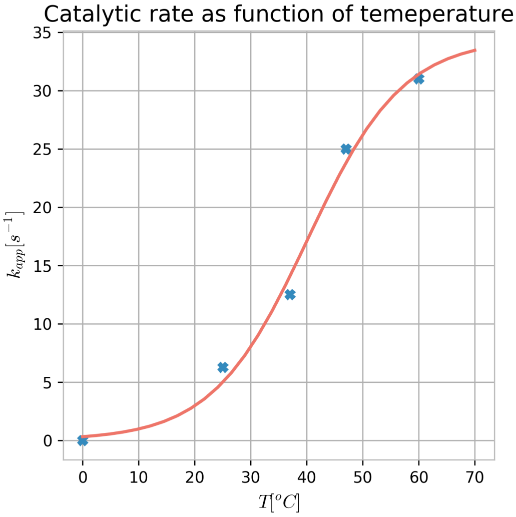


Figure 6.3.1: Apparent catalytic rate of enzymes as a function of temperature. Data points are estimated as follows: for $T = 37^oC$, we take the value provided computed in [192] as the best fit for a number of growth conditions $k_{app} = 12.5s^{-1}$. As noted in [237] that the catalytic rate doubles each 10^oC, we assume the following values: $k_{app}(27^oC) = 6.25$, and $k_{app}(47^oC) = 25$. Finally, we assume that the increase doesn't continue exponentially after a range of temperatures optimal for growth of *E. coli* (non-Arrhenius dependence), as noted in [238], and set $k_{app}(60^oC) = 31$. Additionally, we assume that enzymes do not perform any catalytic function at $T = 0^oC$. These data points are fit to a sigmoidal curve $k_{app}(T) = \frac{a}{b+ce^{-dT}}$.

higher temperatures the enzyme populates a greater variety of stable states, not all of which are capable of catalyzing the reaction, which effectively lowers its overall catalytic rate. With this reasoning in mind, I choose a value of $k_{app}(T=60°C)=31s^{-1}$. The functional dependence between the catalytic rate and temperature I assume to be of sigmoidal type:

$$k_{app}(T)=\frac{a}{b+ce^{-dT}} \tag{6.3.1}$$

By fitting the sigmoidal curve the the chosen data points, I obtain the values of the constants to be $a=2.4$, $b=0.07$, $c=7.82$, $d=0.12$. The chosen data points and the obtained functional dependence of the apparent catalytic rate on temperature are shown in Figure 6.3.1. The fit was obtained by using `scipy.optimize.curve_fit` function, which fits the data points to a user-defined function, using a non-linear least squares method.

6.3.2 Ribosome translation rate

In 1998, Farewell and Neidhardt analyzed what happens to the peptide chain elongation rate in *E. coli* with increase in temperature [239]. They have determined that the peptide chain elongation rate increases in the normal range of temperatures (25 to 37°C) in the same manner as the growth rate. At higher temperatures, the rate of growth decreases, but the peptide chain elongation rate continues to increase. This indicates that the decrease in the growth rate is not due to the capacity of ribosomes to produce protein, but that the reason lies elsewhere. In the low range of temperatures, the growth rate decreased faster than the peptide elongation rate, showing that it is also not limiting for growth at low temperatures. I use this data for the parameterization of the peptide chain elongation rate. I add one additional point at $T\approx 50°C$ of $k_T=18.6s^{-1}$. This point is added to accommodate for a comment made in [239] that the peptide chain elongation rate decreases after the range of temperatures for which they have published the data. As they do not

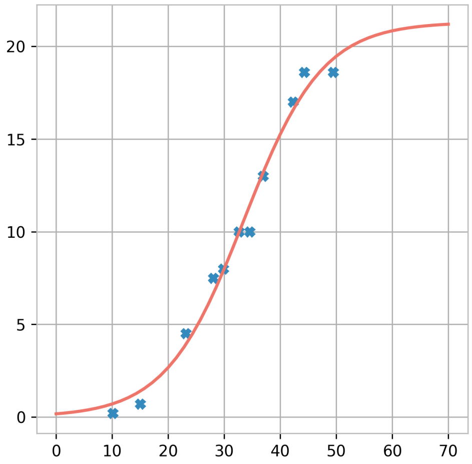


Figure 6.3.2: Peptide chain elongation rate of the ribosome as a function of temperature. All the data points (except the last one) are taken from [239]. The last data point is assumed based on the comment in the same article claiming that after the measured range of temperatures the elongation rate begins to decrease. As we had no numerical values for this point, we assumed that the rate stays the same after the last measured point. The data is fit to a sigmoidal curve $k_T(T)=\frac{a}{b+ce^{-dT}}$.

offer a number for this observation, I have just assumed that the rate saturates after the one at the highest measured temperature. When fitting the data to a sigmoidal curve (as in Equation 6.3.1), the following values of constants are obtained: $a = 5.83$, $b = 0.27$, $c = 33.56$ and $d = 0.14$. The fitting was performed as in subsection 6.3.1. The data points and the results of the fit are shown in Figure 6.3.2.

6.3.3 Folding and unfolding rate

The folding and unfolding rates are chosen so as to exhibit the Arrhenius dependence on temperature, with both k_f and k_u increasing with temperature. Again, as for the catalytic rate, it is very difficult to obtain a value of the two rates which would be meaningful for the entire proteome. Because of this, I use a rough estimate of the two values, similar to the way they are described in [240].

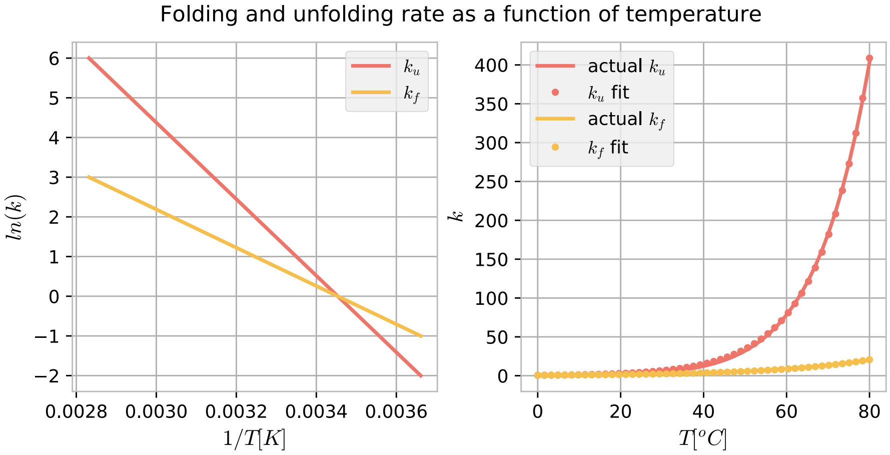


Figure 6.3.3: Protein folding and unfolding rate as a function of temperature. I assume that the unfolding rate is dependent on temperature as described by the Arrhenius equation. Left: Arrhenius plot. Logarithm of k_f and k_u depends linearly on $1/T$. The data is not taken for any particular protein, but is a rough estimate as given in [240]. Right: Transformation of the plot on the left to linear coordinates. This curves are fit to an exponential: $k = ae^{bT-c}$.

This choice is depicted in Figure 6.3.3. The dependence of the logarithm of k_u and k_f to the inverse of temperature is described by an exponential dependence of k_u and k_f on temperature. The values thus obtained are:

$$k_f(T) = 1.84e^{0.04T-1.09} \qquad k_u(T) = 2.18e^{0.08T-1.38} \tag{6.3.2}$$

6.3.4 Parameters without estimates

There are certain parameters in the model which has not been computed from existing data (see Table 6.4.1). This is true for the aggregation rate of proteins and all the scaling factors relating the distribution of total metabolic flux into precursor and energy, and the ones relating the length of macromolecule to the cost of production and degradation in terms of energy. For these parameters I was not able to find sufficient data which would allow me to estimate them within the time available for the completion of my thesis.

6.4 Model simulation

Table 6.4.1: Parameters of the optimal control model of HSR. Some parameters have been taken directly from data (such as ribosome length) or from published literature, some estimated from published data, and some chosen without reference to data.

Parameter name	Symbol	Value	Unit	Comment
Number of S in ribosome	n_R	20120		From *E. coli* ribosome
Number of S in chaperone	n_C	5000		Chosen
Number of S in protease	n_P	5000		Chosen
Number of S in E_{sf}	$n_{E_{sf}}$	16600		Most of metabolic pool
Number of S in E_{ca}	$n_{E_{ca}}$	400		2% of metabolic pool
Number of S in E_{ts}	$n_{E_{ts}}$	3000		15% of metabolic pool
Number of S in HP	n_{HP}	300		Average *E. coli* protein
Translation efficiency	$k_R(T)$	$5.83/(0.27+33.56e^{-0.14T})$	s^{-1}	Estimated
Catalytic rate of enzymes	$k_M(T)$	$2.4/(0.07+7.82e^{-0.12T})$	s^{-1}	Estimated
Chaperone efficiency	k_C	25	s^{-1}	Estimated
Protease efficiency	k_P	25	s^{-1}	Estimated
Folding rate	$k_f(T)$	$1.84e^{(0.04T-1.09)}$	s^{-1}	Estimated
Unfolding rate	$k_u(T)$	$2.18e^{(0.08T-1.38)}$	s^{-1}	Estimated
Aggregation rate	k_{agg}	0.1	min^{-1}	Chosen
mRNA degradation rate	k^m_{deg}	0.1	min^{-1}	Typical in *E. coli*
% of v_M diverted to energy	p_E	0.15		Chosen
Energy cost of mRNA prod. (scaling factor)	γ^m_p	0		Chosen
Energy cost of protein prod. (scaling factor)	γ^p_p	0.1		Chosen
Energy cost of mRNA deg. (scaling factor)	γ^m_d	0		Chosen
Energy cost of protein deg. (scaling factor)	γ^p_d	0.1		Chosen
Cytosolic density	D_c	4.89	$mmol/gCDW$	Taken from [192]
Biomass steady-state concentration	B_0	0.387	$mmol/gCDW$	Taken from [235]
HP steady-state concentration	HP_0	0.00163	$mmol/gCDW$	Taken from [235]

6.4.1 Receding horizon control

The goal of the above-described model is to predict the proteome rearrangement under the conditions of changing temperature. Temperature is a state of the model, described as a first-order system adjusting to a value set by a system parameter T_{ext}. The temperature change can therefore be introduced by changing T_{ext}, after which the system temperature will reach a new steady state, equal to T_{ext}.

In the optimal control problem, the solutions are optimized for the entire duration of the simulation time. This means that if within a single optimal control problem one would introduce a temperature change, the optimal solution would be the one in which the adaptation takes place even before the temperature changes. While this is acceptable and even desired in a system where control can be implemented externally (like a rocket or a car), it is unrealistic within a cellular setting. Therefore, a single simulation cannot suffice for the type of investigation required for this thesis. For this reason, I implemented a receding horizon control simulation strategy. The optimization problem is repeatedly solved, and the solutions for step n act as constraints on the initial values of state variables for step $n+1$. While in the first simulation step the optimal control problem is described by Equation 6.2.73, every next (n^{th}) simulation step has additional boundary constraints which ensure continuity with the previous ($(n-1)^{th}$) step. The boundary conditions are the following:

$$\sum_{\zeta}^{\zeta\in\mathbb{C}_R} n_\zeta\Big([\zeta](t_0)+[{}^m\zeta](t_0)\Big)+n_{E_{ca}}([E^f_{ca}](t_0)+[E^a_{ca}](t_0))+S(t_0)-D_c\leq 0 \tag{6.4.1}$$

$$X(t_0)-X_0=0 \tag{6.4.2}$$

$$T(t_0)-T_0=0 \tag{6.4.3}$$

$${}^mE_{sf}(t_0)-{}^mE_{sf0}=0 \tag{6.4.4}$$

$$E_{sf}(t_0)-E_{sf0}=0 \tag{6.4.5}$$

$${}^mE_{ca}(t_0)-{}^mE_{ca0}=0 \tag{6.4.6}$$

$$E^u_{ca}(t_0)-E^u_{ca0}=0 \tag{6.4.7}$$

$$E^f_{ca}(t_0)-E^f_{ca0}=0 \tag{6.4.8}$$

$$E^a_{ca}(t_0)-E^a_{ca0}=0 \tag{6.4.9}$$

$${}^mE_{ts}(t_0)-{}^mE_{ts0}=0 \tag{6.4.10}$$

$$E^u_{ts}(t_0)-E^u_{ts0}=0 \tag{6.4.11}$$

$$E^f_{ts}(t_0)-E^f_{ts0}=0 \tag{6.4.12}$$

$$E^a_{ts}(t_0)-E^a_{ts0}=0 \tag{6.4.13}$$

$${}^mC(t_0)-{}^mC_0=0 \tag{6.4.14}$$

$$C(t_0)-C_0=0 \tag{6.4.15}$$

$${}^mP(t_0)-{}^mP_0=0 \tag{6.4.16}$$

$$P(t_0)-P_0=0 \tag{6.4.17}$$

$$\tag{6.4.18}$$

The simulation script I provide with the thesis allows for the definition of the initial temperature and (optionally) a number of other temperatures of the system and times at which they occur.

6.4.2 Numerical simulation

As already mentioned, each simulation step in the receding horizon simulation was performed by the Bocop solver. Bocop is an open-source toolbox for solving optimal control problems. The optimal control problem, as defined in Equation 6.2.73, is infinite-dimensional. What needs to be estimated, namely the controls $u(t)$ of the system, are continuous in time and are therefore defined by an infinite number of points over a finite interval. To transform this problem into a finite-dimensional optimization problem (nonlinear programming (NLP) problem), Bocop uses the so-called direct method, through which the states and the controls of the system are discretized in time [2]. Such a discretized problem is then passed to the IpOpt solver [241], which provides the solution to the NLP by the interior-point line search filter method. The problem defined in Equation 6.2.73 is solved for a predefined final time through implicit Euler numerical integration algorithm.

6.4.3 Basic model functionality

Balanced growth. It is instructional to first make sure that the model exhibits expected behavior when growing on a single temperature. Since the external substrate is constant, it is expected that the simulated cells will grow in the balanced regime typical of the exponential growth phase [242], the one in which all the intracellular species are constant[3]. This expectation is based on the knowledge that there is (in terms of an RBA model) just one cellular configuration that maximizes growth [29]. To ensure the basic functionality of the receding horizon implementation, I first compare the single simulation performed by Bocop for a fixed final time to a receding horizon simulation performed as described in subsection 6.4.1.

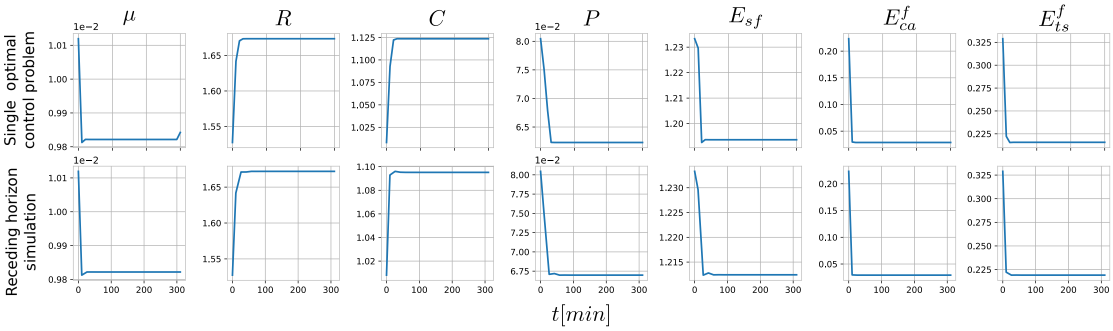


Figure 6.4.1: Simulation of growth on constant temperature of $T = 37°C$. The variables shown are the growth rate μ (in $[min^{-1}]$) and a selection of system states (measured in $mmol$ of substrate S per $gCDW$). The upper row shows the results of a single optimal control problem simulation with fixed final time $t_f = 300[min]$. The lower row shows the results of the receding horizon type simulation for the same time duration. The results differ only slightly, to a degree that can be expected in a complex optimization problem.

In Figure 6.4.1, I show that the results obtained from the two simulation methods are almost identical, save for the numerical effects, most probably due to different time quantization. The simulation was done for $T = 37°C$. The capacity of the model to predict balanced growth is the first and basic test of its validity. Even if I show the simulation results for a single temperature, the balanced growth is predicted for a whole range of biologically relevant temperatures. In

---

[2] The other method by which an optimal control problem can be optimized is the so-called *indirect* method, by which the optimality criterion is first obtained analytically, and is afterwards discretized to obtain a solution.

[3] The constancy of intracellular species is of course just a useful abstraction of the actual noisy cellular state.

section A.9 I show how the model reaches a steady state for temperatures from $20°C$ to $60°C$. One important remark about all the simulation results is that under no condition does the system behavior start at a steady state, but instead features initial and final "adaptation". This initial adaptation is typical of problems that exhibit the so-called "turnpike" behavior[4].

Growth rates for a range of temperatures. It is known that *E. coli* (like all ectotherms) has a range of temperatures most suited for its survival and reproduction. Bellow and above this range of temperature, growth is seriously impaired. When using the model of this chapter to simulate growth for a range of temperatures, this feature is recovered, if only qualitatively.

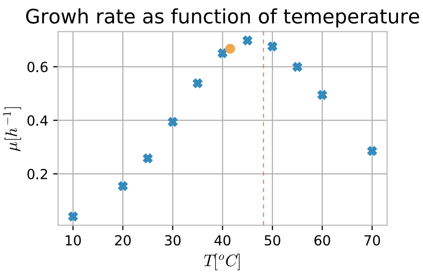


Figure 6.4.2: Dependence of growth rate on temperature. The optimal predicted temperature for growth is around $45°C$, close to the experimentally determined one of $T_{opt} \approx 42°C$ for growth on glucose. A single point at $T = 42°C$ depicts the experimentally determined value [208]. The dashed vertical line shows the temperature at which most *E. coli* strains stop growing [243].

In Figure 6.4.2, one can see that the predicted optimal temperature for growth is around $T = 45°C$, which is not too far from the experimentally determined value of $T = 42°C$. The predicted range of growth rates is reasonable for *E. coli* (the experimentally determined growth rate of *E. coli BW25113* on glucose for $T = 37°C$ is $0.58h^{-1}$ and for $T = 42°C$ is $0.66h^{-1}$ [208]).

6.4.4 Adaptation to change in temperature

The next step in model exploration is to predict the adaptation to an increase in temperature, keeping in mind that this behavior is a reflection of a single principle, that of optimization of resources for growth and that no regulatory effects have been accounted for. I first illustrate the adaptation to the most common type of heat shock performed in the laboratory - the transfer from 27 to $42°C$. Figure 6.4.3 shows the change in the growth rate, the most abundant cellular components, and their corresponding mRNA species. As is known from experimental data, the amount of chaperones and proteases increases upon heat shock. Further on, after a stage of adaptation, the system enters a new steady state, as is also known of heat shock in *E. coli*, at least for a range of temperatures in which the cells do not enter the stationary phase [244].

To better illustrate how the cellular investment of resources changes with temperature, in Figure 6.4.5 I show the distribution of protein in functional categories for a range of temperatures from 20 to $60°C$. As the temperature increases, the ribosomes and metabolic enzymes become more efficient. This is reflected in the decrease in the cellular investment in these components. Because the increase in temperature also increases the rate of misfolding of proteins, the cellular requirement for proteome maintenance machinery (chaperones and proteases) increases, as shown in Figure 6.4.5.

---

[4] I have not proven that the problem falls in the category of turnpike problems, I just observe that under all simulation conditions the problem exhibits the typical turnpike behavior, in which most of the time is spent at steady state, preceeded and followed by a deviation from steady state.

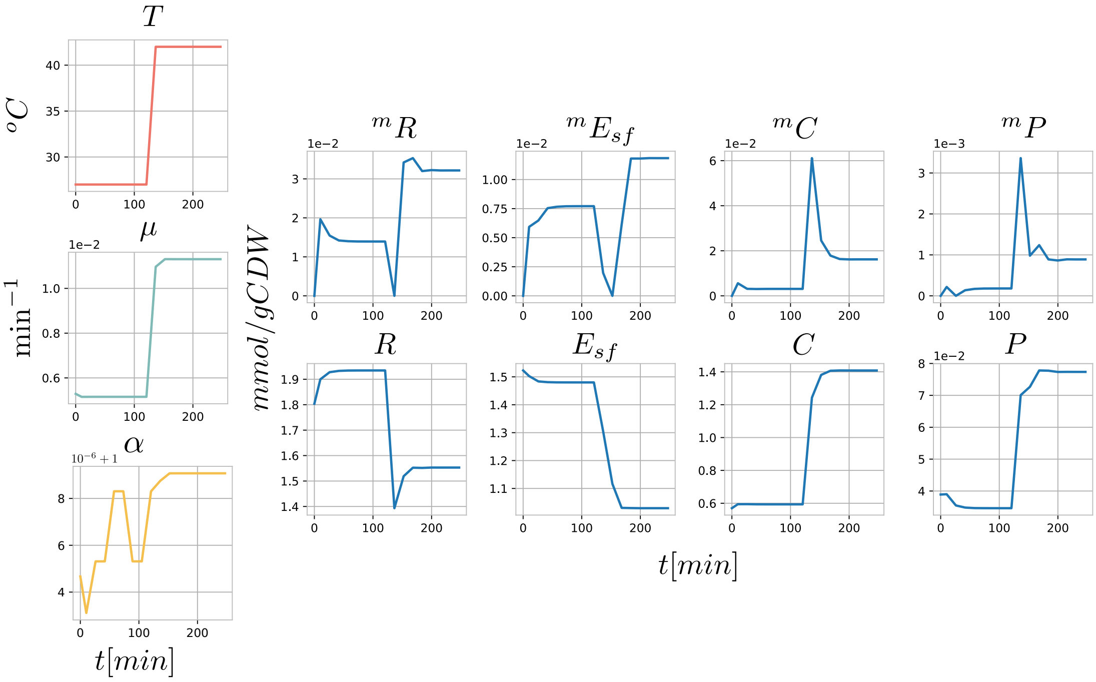


Figure 6.4.3: Response of the optimal control model to the change in temperature from 27 to $42°C$. The left column shows the change in temperature, growth rate μ and the mRNA pool translation scaling factor α, while the right part of the plot shows the predicted adaptation of major macromolecular species of the model.

6.4.5 Comparison to the whole-cell RBA model

Even if the whole-cell RBA model developed in Part III and the simple model developed in this chapter find themselves on opposite sides of the detail spectrum, they are based on the same principles of parsimonious resource allocation and the constraints developed from those principles. The parameterization used for the whole-cell model RBA of *E. coli* was performed by using the data obtained with cells grown on $37°C$. It is therefore instructive to make the comparison by simulating the growth on that temperature. For the simulation of the whole-cell model, I use the glucose minimal medium. Apart from the fact that the model presented in this chapter is a very simplified cellular model, one additional difference from the whole-cell model is that in this model there is only one compartment - the cytosol. The values that I compare are the growth rate and the percentage of ribosomes in the cytosolic fraction.

One of the very important constraints on the model - the density constraint (see subsection 4.2.1) - is implemented in both models using the same value of the $D_c = 4.89$ parameter. However, another very important parameter, namely the efficiency of the ribosome, is quite different. In the whole-cell RBA model, this efficiency for the growth on glucose and the thus obtained growth rate of $\mu = 0.65s^{-1}$ is $k_T = 23.52s^{-1}$. The value obtained in subsection 6.3.2, whereby the ribosome efficiency is the function of temperature, is $k_T(T = 37°C) = 13s^{-1}$.

It is due to this difference that the growth rate of the two models differs significantly for the temperature of $T = 37°C$ and is $\mu = 0.65h^{-1}$ for the whole-cell RBA, and $\mu = 0.41h^{-1}$ for the HSR model. The ribosome fractions are $f_R = 33.14\%$ for the whole-cell model and $f_R = 39.29\%$ for the HSR model. However, if the ribosome efficiency is taken from the whole-cell model and set constant to that value in the HSR model, the growth rate of the HSR model becomes $\mu = 0.6h^{-1}$, and the ribosome fraction $f_R = 33.59\%$, which is quite similar to the whole-cell

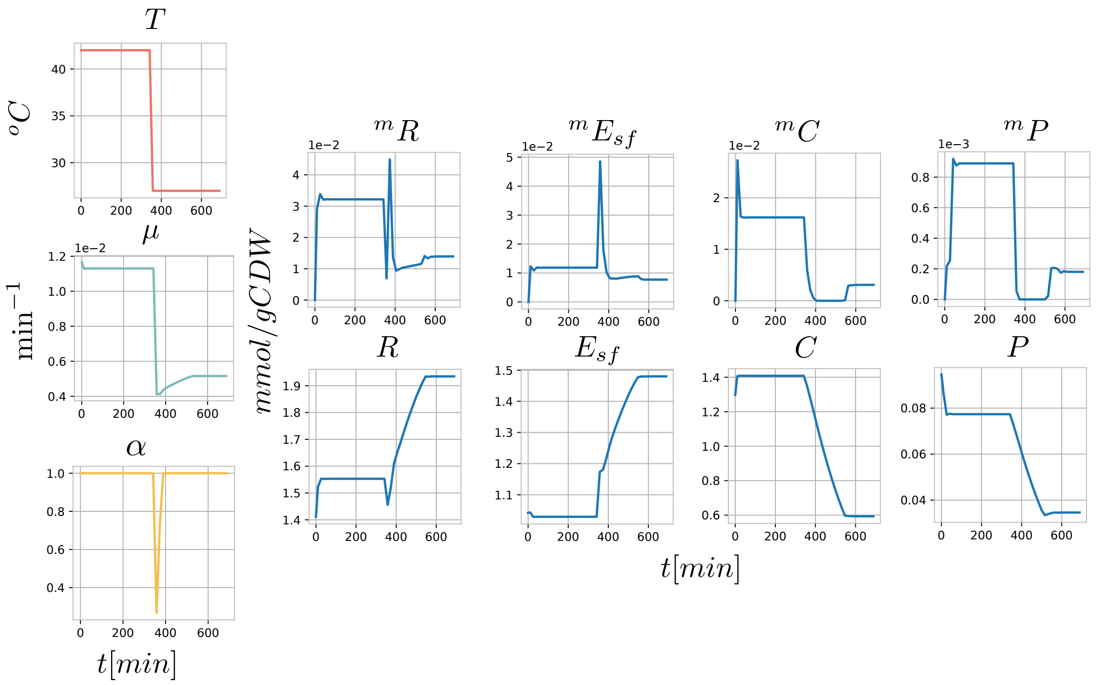


Figure 6.4.4: Response of the optimal control model to the change in temperature from 42 to 27°C. The left column shows the change in temperature, growth rate μ and the mRNA pool translation scaling factor α, while the right part of the plot shows the predicted adaptation of major macromolecular species of the model. This change in temperature leads to a decrease in the growth rate. Due to this, the cellular adaptation to the change is slower (notice the different time span in comparison to Figure 6.4.3). Also, this causes the mRNA pool to temporarily not be fully translated and results in a drop in α at the time of the temperature change, which allows for a faster adaptation on the protein level.

RBA model. This shows that, on a global scale, the two models correspond well to one another.

6.5 Different model versions

During the process of development of the final optimal control model, I have made several simpler models. These were implemented to better understand the effects of different constraints on the numerical solubility of the optimal control problem and for the purposes of debugging. I have developed two types of models. They differ in how the proteins are produced - with or without the mRNA intermediate. The first model type (without mRNA production) has not been detailed in the thesis because it has proven not to be suitable for this study. However, I can imagine that it can be useful for educational purposes (due to its simplicity), or maybe for the study of different biological phenomena. In this model, the synthesis flux of the protein x is under the direct control of a control variable v_P^x, making the dynamics of that protein correspond to:

$$\frac{d[x](t)}{dt} = v_P^x(t) - \mu(t)x(t) + \dots \tag{6.5.1}$$

where … stands for protein conversion (such as from unfolded to folded). In comparison to that, in the second model type, the control is exerted on mRNA production

$$\frac{d[{}^m x](t)}{dt} = v_P^{{}^m x}(t) - (\mu(t) + k_{deg}^m)[{}^m x](t) \tag{6.5.2}$$

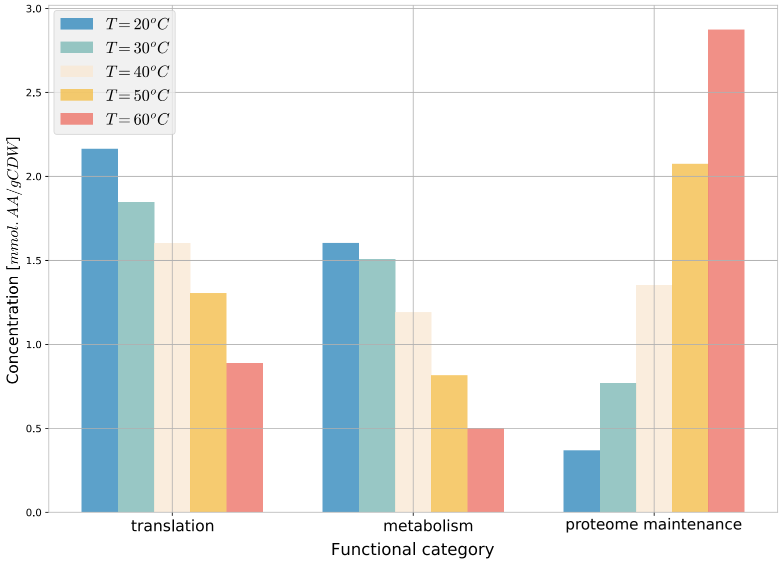


Figure 6.4.5: Change in cellular composition for growth at different temperatures. Translation includes ribosomes, metabolism three types of enzymes, and proteome maintenance chaperones and proteases. At lower temperatures, the cell needs to invest more resources into translation and metabolism due to their slower rates, but the requirements on proteome maintenance are low, as the unfolding rate is negligible. At higher temperatures, most cellular resources are invested in proteome maintenance.

while the protein synthesis is regulated by the relative abundances of mRNAs competing for the ribosome:

$$\frac{d[x](t)}{dt} = \alpha(t)[^m x](t) - \mu(t)x(t) + \dots \qquad 0 \leq \alpha(t) \leq 1 \tag{6.5.3}$$

Both model types are available in varying degrees of granularity, starting from the simplest - featuring just a single enzyme and a ribosome - to the full heat shock model. All of the models are available in two formulations necessary for the simulation as receding horizon problems (as described in subsection 6.4.1).

For the encoding of optimal control problems, I have developed a simple and intuitive JSON format. While Bocop does offer a format for encoding optimal control problems, I found it somewhat difficult to edit. The reason for that is that, in Bocop, the model information is spread out in multiple files, all of which need to be updated upon making a single change. For example, to add a constant, one has to add it to the `constants.def` file, change the number of constants in the `problem.def` file, and edit all the `tpp` files which reference constants. While this might be feasible in a process by which a model is directly transferred from paper to the computer, it can be burdensome and error-prone in the process of model development. By introducing a new compact JSON format, I believe I have made model editing simpler. With the format, I offer `Python` scripts with which the JSON problem formulation can be converted to the Bocop problem formulation. The example of the simplest model (without mRNA production) encoded in this format is given in section A.10.

All the models with and without mRNA production, as well as the scripts necessary to convert them to Bocop format or to simulate them, can be found in a public Git repository - https://github.com/abulovic/opt-ctrl-cell-models. The repository also features instructions on what is necessary to simulate the models.

6.6 Discussion

The developments in this chapter complement the analysis performed in chapter 3. The phenomenon studied (HSR) is the same, while the methodology is different: in chapter 3 I analyze the behavior of the system by integrating the detailed knowledge on the macromolecular interaction leading to the regulation of the HSR, while in this chapter I assume a parsimonious allocation of resources favoring growth and through it predict the adaptation to a change in temperature.

6.6.1 Concentrations or molecule counts

A reformulation of the problem, in which we don't deal with concentrations but with molecule counts, would allow for simplification of constraints. If the problem were to be formulated in terms of molecule counts, it would not be necessary to consider dilution. The dilution term includes the growth rate, which is described by a linear relation of control variables. Because of this, all equations which describe a concentration of any molecule involve both system and control variables, adding to the complexity of the problem. One drawback of the formulation using molecule counts would be that all the species would exhibit exponential growth, which would lead to the need of using ever-smaller integration times to correctly assess the solutions.

6.6.2 Parameter sensitivity analysis

As mentioned in subsection 6.3.4, the model features several parameters that were chosen *ad hoc*, as I was not able to find the appropriate data. If such data is not available in the literature, the next step in estimating the impact of these parameters on the outcomes of the model would be to perform the parameter sensitivity analysis. Analytical estimates of parameter sensitivities are not easy to obtain in the case of an optimal control problem with mixed constraints. However, to get a first understanding of how important certain parameters are for the outcome of model simulation, one could increase and decrease each parameter for a certain percentage, while keeping the other parameter values fixed. The resulting change in the simulation results would be due to that parameter change. This procedure would, in the least, allow assessing which of the non-estimated parameters influence the simulation outcome the most, and would therefore be best to tackle with data.

6.6.3 "Overshoot" in protein expression

In none of the simulations of temperature change was there a noticeable "overshoot" in the protein expression levels, regardless of the increase in temperature. To my knowledge, all of the dynamic models which tackle to describe the HSR have obtained simulations (for physiological values of parameters) in which the levels of both chaperones and proteases "overshoot" before settling to a new steady-state level [147, 149]. Additionally, it has been noted in [147] that the overshooting type of response is beneficial for the cell, as it allows for faster response and a lower steady-state level of necessary chaperones and proteases. There are some potential issues with this conclusion. The first has been discussed in some detail already (see subsection 3.6.3) and relates to the possibility of making conclusions about the regulation of noisy biological systems by analyzing their deterministic dynamic representations. The second issue might be the fact that these models do not take into account the increase of the peptide elongation rate with temperature. If they would, the overshoot might be much more pronounced, leading to the next issue. Would the cellular regulation be organized in a way so as to greatly overproduce the number of necessary chaperones? Chaperones and proteases are extremely expensive in terms of resources required for their production and the space they occupy. Their overproduction would lead to a significant decrease in the growth rate, which is another factor not taken into account by the aforementioned dynamical models. Since there is no evidence that chaperones are actively degraded in *E. coli*, that means that the decrease in their concentration is due solely to dilution. The decrease in the growth rate could make the effects of dilution very slow, and it could take the cell potentially a long time to reach the new steady-state level of the required chaperones. Also, to my knowledge, there is no experimental data supporting the protein expression overshoot. The only time-resolved chaperone level measurements I know of are presented in [245], and they show a steady increase in the chaperone level over 30 minutes. This goes more in the line of results obtained in this thesis, as well as in the work presented in [245].

6.6.4 Comparison to other simple cell models

In recent years, researchers have developed many simple cell models through which they have attempted to analyze and explain different aspects of bacterial growth-related phenomena. The first of these, to the best of my knowledge, was the model by Molenaar *et al.* [198], which describes the cell by a set of ODEs and linear constraints under the assumption of growth rate maximization. The problem formulation closely resembles the one proposed in this thesis, and in general, the one proposed in RBA [195]. They formulate the problem within the framework of nonlinear optimization, and I within the framework of optimal control. However, optimal control problems are nonlinear optimization problems of special structure. Therefore, from a theoretical standpoint, the difference is minimal.

The other class of small cell models can be represented by the model developed by Weisse *et al.* [175], in which the cell is fully described by a set of ODE equations. Certain assumptions (such as the constant protein content) are not represented by constraints but are directly integrated into the ODE formulation. The benefit of such a model formulation is that it is easier to analyze with mathematical tools and that the numerical simulation might be simpler. The scope of this model is different than that of the one proposed in this thesis, as with this model it is not possible to integrate assumptions about a certain objective being optimized.

6.6.5 Possible model uses

The type of model proposed in this chapter has already been implemented for bioreactor optimization [235]. Because of its dynamical nature, the model is adapted for simulations of bioreactor culture growth and can be used to design an optimal strategy for, for example, maximizing the yield of biomass or wan industrial product. Because it explicitly takes into account the production of protein and inherent growth limitations, it is suited for analysis of RP production problems.

While RP production can be triggered by introducing a certain chemical into the growth medium, it is also possible to induce their production by increasing the temperature of the culture in a process called heat induction. Since many problems related to RP production are related to the capacity of the protein to fold, the increased expression of chaperones and proteases followed by an increase in temperature might prove to be beneficial. The exact temperature at which the effects of increased chaperone expression and increased metabolic and translation rates overweigh the effects of higher proteome maintenance requirements could be analyzed by such a model.

IV Conclusions and outlook

7 Conclusions 139
7.1 Tools and practices in systems biology
7.2 Elegance and utility of the parsimonious resource allocation paradigm

8 Outlook 143
8.1 Dynamical whole-cell models
8.2 Exploring regulation through parsimonious resource allocation paradigm
8.3 Towards predictive whole-cell models
8.4 Understanding temperature effect on cells

9 Further thoughts 145
9.1 On modeling in biology
9.2 A personal note

7. Conclusions

Stress is a change in the living (internal of external) condition of an organism such that it renders its current state suboptimal. This often provokes a series of responses that aim to better align the state of the organism to the new condition. I have studied the HSR of *E. coli* through the development of a detailed regulatory ODE model and its subsequent analysis, and through the application of a parsimonious resource allocation paradigm on a small cell model. I have additionally developed a whole-cell RBA model of *E. coli*. Here I will present the conclusions I have reached regarding the usage of modeling tools in general, and ODEs in particular, and about the application of the parsimonious resource allocation paradigm in understanding cellular states and regulation.

7.1 Tools and practices in systems biology

> For these ideas are not the foundation of science, upon which everything rests: that foundation is observation alone. They are not the bottom but the top of the whole structure, and they can be replaced and discarded without damaging it.
>
> Sigmund Freud, On Narcissism

If you query Wolfram Alpha for the square root of -1, it will readily provide you with an answer, i. This answer is correct under a certain set of assumptions, but those assumptions are to a great degree hidden from the user and not stated as a part of the answer. Indeed, under different assumptions, the provided answer would be incorrect. This was the lesson taught to us at the first lecture of our first mathematics class at the university. The lesson is that tools can be useful, but it can be dangerous to derive conclusions based on their results if we are not completely aware of the assumptions that underlie their functioning. Abstracted from this simple example, this lesson of the first day's lecture remains one of the most valuable during my whole education.

Lying at the intersection of many sciences, systems biology requires solid knowledge of many different fields: biology, mathematics, physics and computer science. It was my impression during my thesis that the diversity of the backgrounds from which the researchers in systems biology come from, combined with the complexity of problems addressed, contributes to seeing modeling methodologies as tools that can be used almost as black boxes. It seemed to me only natural that this could happen (and indeed it had happened to me) since a single researcher in systems biology can and often is simultaneously facing problems that require such a broad spectrum of knowledge that it can easily become a daunting task to have the proper theoretical understanding needed for their successful implementation. The researcher should have a solid enough grasp on experimental methods to be able to judge which published data can safely be used for purposes of modeling and to be able to clearly state the data and metadata requirements during (collaborative) experiment design. He or she should know enough about the biological question at hand so as to be able to judge what is essential, what could be important, and what could be disregarded in the model. He or she should have enough background in probability and statistics so as to know how to perform parameter estimation. It is necessary to know which modeling tools are available so as to make an educated choice for the most appropriate one, and then be mathematically literate enough so as to manipulate the tool, and not be manipulated by it. Additionally, the usage of complex tools, especially when applied to large models, can easily divert attention away from its formulation (assumptions and limitations) and focus it on its (technically) successful application. As most models are implemented on a computer, there should also exist an awareness of the potential numerical issues surrounding model simulation.

Some of the tools often used in systems biology (such as ODEs and FBA) have been around for quite a long time now, enough for the new generations of modelers to not remember the original usage, together with the assumptions and simplifications originally made in order to "tame" the biological problem so that it can be mathematically described. However, I have found that looking into the original publications describing the modeling methodologies offers indispensable insight into their assumptions and their possible application, as well as helps one appreciate their intricacy. I have also found that finding scientific literature written for other fields can help better understand the common and diverging points of their application and therefore better appreciate the limitations and potential usages of the modeling method. For me, one of the most important conclusions of my thesis is that all the time devoted to the proper understanding of the modeling techniques is time well spent, allowing me to slowly develop away from the tinkerer and towards the hacker.

Within my thesis, I have looked into more detail into the application of ODEs and mass action kinetics for modeling the chemical interactions of living systems through my study and modeling of the HSR in *E. coli*. These considerations have led me to believe that it is prudent to be more careful with the conclusions reached through the usage of these tools. Many models describing chemical systems in cells involve feedback. Experimental analysis of these systems is often based on measurements involving populations, not single cells. Any conclusions thereby reached have to be trimmed by the fact that there can be no intracellular feedback directly detected on the population level. Furthermore, the noise in macromolecular expression levels inside cells directly affects the stability of these feedback systems, and conclusions reached by analyzing deterministic models might not necessarily be meaningful.

7.2 Elegance and utility of the parsimonious resource allocation paradigm

Parsimonious resource allocation is a lens through which one can look at phenomena concerning living beings by investigating whether an observable trait can be explained through the need of the organism to be efficient in the use of its resources available to it. One consequence of the study and application of this principle has been that it made clear that the understanding of any particular regulatory system in the cell requires a good overall knowledge of the organization of its most important functions (such as metabolism and protein production). In bacteria, almost every type of stress will cause great changes in the entire cell because it will affect its growth, and therefore its production of macromolecules and metabolic activity. Therefore, the study of each type of stress will require the modeler to be aware of its global effects, even if only to make an educated decision as to which of those can be disregarded in the modeling process.

Resource Balance Analysis is the modeling tool that allows for an integrated description of the cell (by functionally linking its resource-relevant processes) and the prediction of cellular states through the assumption of parsimonious resource allocation. The cell state prediction obtainable through RBA involves the growth rate, concentrations of enzyme and macromolecular process machines (such as ribosomes), as well as metabolic reaction fluxes. I have shown how RBA can be used to successfully and without detailed parameterization predict the steady-state growth of *E. coli* populations and have also demonstrated its utility in explaining certain regulatory events (as in the case of dehydrogenase substitution).

In order to study whether the regulation of the HSR can be understood through the lens of parsimonious resource allocation, I have developed a dynamic simple cell model which includes RBA-like constraints and maximization of growth. The model was implemented in the optimal control framework in which one can define the dynamic behavior of the system through ODEs and impose linear equality and inequality constraints on the system (RBA-like constraints) and maximize a certain objective (maximization of growth). The model includes the processes most relevant for this response: protein production, chaperoning and degradation, but doesn't involve any description of its regulation. The predictions of the model qualitatively correspond to what has been observed experimentally in HSR. This shows that RBA-like constraints coupled with a dynamic description of an adaptive cellular system and assumption of growth maximization can successfully describe adaptation to change in temperature in *E. coli*, and presumably other stress responses as well.

8. Outlook

In this chapter, I offer a few thoughts on what would be the continuation of the work done in my thesis.

8.1 Dynamical whole-cell models

While metabolic genome-scale models have been around for quite a while, the last decade has seen the development of genome-scale cell models. As the *E. coli* model developed in this thesis, they are becoming quite accurate in predicting cellular states even when described with a remarkably low number of parameters, illustrating the predictive power of the underlying paradigms. As was the case with genome-scale metabolic models, these models will soon find their way to everyday usage by researchers in both science and industry.

While all the whole-cell genome-scale modeling paradigms known to me describe the cells in steady-state, one natural development in this field will be their dynamical counterpart. This will allow such models to further the understanding of adaptive cellular processes (such as growth and stress, for example), and aid in experimental design when time is of critical issue. One example of a case when such models would be of great use is in designing experiments for the expression of recombinant protein. Since they can account for growth defects caused by RP expression and can reflect in detail the cellular configuration on a specific medium, they will help in optimizing the timing of induction and duration of the culture growth. Incorporation of effects such as temperature in these models can further help in finding the optimal conditions for such experiments.

8.2 Exploring regulation through parsimonious resource allocation paradigm

Bacteria (especially ones that cannot form spores) are almost under constant pressure to acquire nutrients to survive. This has led to a type of internal organization which favors parsimonious resource allocation. Many regulatory mechanisms that exist in bacteria are in place to ensure that the resources are not idly wasted. Parsimonious allocation of resources was shown to be the reason behind the highly debated overflow metabolism and the *diauxie* shift.

It was in fact modeling which contributed to the understanding of these phenomena. Integration of the parsimonious resource allocation paradigm, first with simple cell models, and then later with genome-scale cell models, allowed to test *in silico* whether this principle can explain the phenomena observed *in vivo*.

Apart from these larger rearrangements of metabolism, I have shown that with the genome-scale cell model of *E. coli* this principle can be used to explain regulation involved in local metabolic rearrangements, such as is the case in the substitution of the NADH dehydrogenase by glucose dehydrogenase. By showing that it is possible to predict the cellular rearrangement in face of changing temperature by the sole assumption of maximization of growth (the equivalent of efficient resource utilization), I show that this may be the reason behind the actual cellular regulatory mechanism.

There is a great amount of work to be done in our understanding of how far the idea of parsimonious resource allocation can be used to understand cellular states and adaptations. Adaptations of metabolism to the medium are the first and most obvious example which has been under some scrutiny but is still far from fully explored. Cellular adaptations in conditions of scarcity might also be interesting to study under this paradigm.

8.3 Towards predictive whole-cell models

Even if *E. coli* is such a well-studied organism, there is a serious lack of data needed for the full parameterization of a whole-cell RBA model. Some of the most important cellular features have not been measured (such as the number of ribosomes), or have been indirectly estimated 70 years ago. There are no systematic proteomics and fluxomics experiments performed in the same conditions which would allow for full parameterization of the *E. coli* RBA model. These experiments would allow for a creation of a well-parameterized cell model of one of the most used model organisms in research and industry. The lack of suitable omics data is often the case because the data available is not obtained in close collaboration of experimental and theoretical researches. It is enough that certain metadata is missing, and the whole expensive and time cumbersome experiment can become useless for modeling purposes. The development of predictive, well parameterized genome-scale whole-cell models will require a joint effort of the theoretical and experimental communities.

8.4 Understanding temperature effect on cells

Temperature and the stress caused by a change in temperature is one condition that can affect all the cells on Earth. Regardless of whether an ectotherm or an endotherm, the temperature of an organism remains crucial: either the organism adapts to a range of temperatures within which it can still function, or it needs to ensure through its own metabolic activity the stable temperature at which it can live. In each of these organisms, temperature affects all of its components - the density of its membrane, the speed at which reactions occur, the stability of its macromolecules. There has been a lot of work done in understanding how individual cellular components react to temperature (metabolic rate of individual enzymes and stability of proteins for example), but (to my knowledge) not much work done in understanding how cells adapt to such radical and simultaneous changes in all of their components. Therefore, what is known under the term of heat shock (the unfolding of proteins and the cellular adaptive reaction to it) is only the most obvious and well-studied systemic consequence of temperature change.

Since it is one of the basic guiding and limiting conditions for all life, I believe that a better understanding of adaptation to temperature will provide great insight into cellular regulation and flexibility.

9. Further thoughts

9.1 On modeling in biology

Modeling in biology might seems a new activity on the first glance, but that is true only if one considers mathematical modeling. I think, however, modeling is a very old activity. Modeling (for me) is an attempt to represent one's knowledge by a system larger than the body of evidence one is basing it on. The purpose of modeling should be extension of knowledge, projection of the gathered facts into the unknown, whereby the unknown is transformed into a map which can then be explored. Each map is necessarily incomplete, and the exploration of the actual territory leads to creation of new maps. The elegance of a map lies in its simplicity and its clarity. Map allows the mind to travel far without the burder of all the detail of the actual territory. This travel is necessary because the keen observation of things such as is necessary in science can lead to short sightedness and lack of enthusiasm, at least in my case. Maps are created out of need.

As I have entered the field of systems biology, I was full of questions. My questions were mostly related to: (i) how organisms work (I list these questions in detail in section 9.2), (ii) the internal organization of mathematical systems typically used for representation of living cells in systems biology and (iii) the limits and range of validity of these systems when used to model living systems. While the questions of type (i) were the driving motive for the research, and the questions of type (ii) were interesting, fun and very rewarding, the questions of type (iii) seemed to me the proper domain of systems biology. I have attempted to pose some of these questions, especially in the Part I of the thesis.

9.2 A personal note

Many have pointed out that there are parts of my thesis, certain reflections or thoughts, which seem inappropriate for a thesis work. I should be more on point, less poetic and less philosophical. While I certainly know that many people are inclined to be sceptical about such thoughts, for me they form the core of my scientific interest. Everything else, for me, is just roadwork that needs to be done to chart a way to that mysterious place. The fact that I find this construction work pleasurable and with its own awards has made me be able to persist in this often quite challenging task. But as much as I like my construction work, and as much as it is an enjoyable exercise in intellect, it is not at the center of what drove me to spend six years of my life in reading pages after pages of obscure texts, full of acronyms and technical jargon. I was after something. And that something needs to have a place in this thesis, because it has been the source of it, the path of it, and the destination of it.

There are certain moments of clarity which impale us with their force so much that they have the power to alter the course of our lives. I believe often the hard and cold realization of how much our soft bodies and minds are at a mercy of external forces has been a profound source of interest for scientists of all ages. The will to predict and control natural forces with the force of the intellect has been a strong motive at all times. But even if the forces of the older times were more frightening, when our control over our surroundings was less far-reaching, the forces of the world seemed to work more on the outside, then on our inside. Today, we are plagued by this uncomfortable feeling that we are controlled on the inside, by our own chemistry and ancestry. In a sense, we are no longer free. There is no plausible *elan vital* in the 21^{st} century.

My core of interest came out of a number of questions which plagued me at the time when I was 22 or 23 years old. First was, why is humanity doing so poorly? With all the beauty I could imagine myself, with all the deep understanding that sometimes shone through arts and sciences, why is world full of mysery and violence? Is the depth and mystery I felt in my own consciousness really limited to creating such a world? As I learned more about biology, I started wondering at the incredible, well regulated complexity that lies underneath our behaviour. Even if the world of human creation seems complex and large, still I felt that the magnificent complexity of the organisation of a cell or an organism is larger still, and more impressive still. I wondered about the relation of these two worlds. One that seems almost automatic, without will, a series of perfectly orchestrated events. The other that must be a direct consequence of it, and yet seems independent of it - the world of social behavior, experience, memory, emotion... The interface of these two worlds was what interested me the most.

I will list my questions here, as naive as they were. I do not want to censor them with the knowledge that came afterwards. I still find their naivety beautiful.

How can systems as complex as living beings (even made of a single cell) function reliably? Millions of interacting components, organised in cells, organs, organisms...

Why does this complexity result in a behaviour such as one can observe among more complex organisms?

Why is not harmonious organisation such as seems possible at a cellular level not extended to the next level (individual perception, social organization)?

What kind of light does that shine on suffering? Is it just an extension of the underlying harmonious organisation, but we, as unsuspecting vessels of evolution, are just oblivious to it?

These were my questions. This was the reason and motivation to spend years looking into how cells work. I must say I have not managed to completely answer any of these questions.

V

Acronyms 149

Bibliography 165

Acronyms

ATP adenosine triphosphate.

cDNA coding DNA.
CDW cell dry weight.

ESI ElectroSpray Ionization.

FBA Flux Balance Analysis.
FBAwMC FBA with molecular crowding.

GTP guanosine triphosphate.

HSR Heat Shock Response.

LP linear programming.

MALDI Matrix Assisted Laser Desorption Ionization.
ME macromolecular expression.
MFA Metabolic Flux Analysis.
MOMENT MetabOlic Modeling with ENzyme kineTics.
mRNA messenger RNA.
MS mass spectrometry.

NLP nonlinear programming.

OD Optical Density.
ODE Ordinary Differential Equation.
ODEs Ordinary Differential Equations.

PoI Protein of Interest.
PQQ pyrroloquinoline quinone.

RBA Resource Balance Analysis.
RNA riboucleic acid.
RP recombinant protein.

SBML Systems Biology Markup Language.

TF Trigger Factor.

UPR Unfolded Protein Response.

XML Extensible Markup Language.

Bibliography

[1] Hiroaki Kitano. "Systems biology: a brief overview". In: *science* 295.5560 (2002), pages 1662–1664.

[2] Erwin Schrödinger. *What Is Life? the physical aspect of the living cell and mind*. Cambridge University Press Cambridge, 1944.

[3] Jack W Szostak, David P Bartel, and P Luigi Luisi. "Synthesizing life". In: *Nature* 409.6818 (2001), pages 387–390.

[4] Peter A Parsons. "Environments and evolution: interactions between stress, resource inadequacy and energetic efficiency". In: *Biological Reviews* 80.4 (2005), pages 589–610.

[5] Vargas-Maya Naurú Idalia and Franco Bernardo. "Escherichia coli as a model organism and its application in biotechnology". In: *Escherichia coli-Recent Advances on Physiology, Pathogenesis and Biotechnological Applications*. IntechOpen, 2017.

[6] Ingrid M Keseler et al. "The EcoCyc database: reflecting new knowledge about Escherichia coli K-12". In: *Nucleic acids research* 45.D1 (2016), pages D543–D550.

[7] Steven B Zimmerman and Stefan O Trach. "Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli". In: *Journal of molecular biology* 222.3 (1991), pages 599–620.

[8] FJM Mergulhão, David K Summers, and Gabriel A Monteiro. "Recombinant protein secretion in Escherichia coli". In: *Biotechnology advances* 23.3 (2005), pages 177–202.

[9] Gita Mahmoudabadi et al. "Defining the Energetic Costs of Cellular Structures". In: *bioRxiv* (2019), page 666040.

[10] Frederick C. Neidhardt and H. Edwin Umbarger. "Chemical composition of Escherichia coli". In: *Escherichia coli and Salmonella* (1996), pages 13–16.

[11] Karl Peebo et al. "Proteome reallocation in Escherichia coli with increasing specific growth rate". In: *Molecular BioSystems* 11.4 (2015), pages 1184–1193.

[12] J Raulin. "Chemical studies on growth". In: *Ann. Sci. Nat. Bot* 11 (1869), pages 93–299.

[13] Martinus W Beijerinck. "Anhaufungsversuche mit ureumbakterien". In: *Zentralbl. Bakterol. Parasitenkd. Infektionskr. Hyg. Abt. II.* 7 (1901), pages 33–61.

[14] Thomas Robert Malthus. *An Essay on the Principle of Population*. 1798.

[15] Roberto Kolter, Deborah A Siegele, and Antonio Tormo. "The stationary phase of the bacterial life cycle". In: *Annual review of microbiology* 47.1 (1993), pages 855–874.

[16] P Pletnev et al. "Survival guide: Escherichia coli in the stationary phase". In: *Acta Naturae* 7.4 (27) (2015).

[17] Steven E Finkel. "Long-term survival during stationary phase: evolution and the GASP phenotype". In: *Nature Reviews Microbiology* 4.2 (2006), pages 113–120.

[18] Moselio Schaechter. “A brief history of bacterial growth physiology”. In: *Frontiers in microbiology* 6 (2015), page 289.

[19] Jacques Monod. “The growth of bacterial cultures”. In: *Annual review of microbiology* 3.1 (1949), pages 371–394.

[20] Frederick C Neidhardt and Boris Magasanik. “Studies on the role of ribonucleic acid in the growth of bacteria”. In: *Biochimica et biophysica acta* 42 (1960), pages 99–116.

[21] Ole Maaløe and Niels Ole Kjeldgaard. “Control of macromolecular synthesis: a study of DNA, RNA, and protein synthesis in bacteria”. In: (1966).

[22] RD Wells et al. “DNA structure and gene regulation”. In: *Progress in nucleic acid research and molecular biology*. Volume 24. Elsevier, 1980, pages 167–267.

[23] CA Gross et al. “The functional and regulatory roles of sigma factors in transcription”. In: *Cold Spring Harbor symposia on quantitative biology*. Volume 63. Cold Spring Harbor Laboratory Press. 1998, pages 141–156.

[24] Uri Alon. “Network motifs: theory and experimental approaches”. In: *Nature Reviews Genetics* 8.6 (2007), pages 450–461.

[25] Jonathan R Karr et al. “A whole-cell computational model predicts phenotype from genotype”. In: *Cell* 150.2 (2012), pages 389–401.

[26] Eleftherios Terry Papoutsakis. “Equations and calculations for fermentations of butyric acid bacteria”. In: *Biotechnology and bioengineering* 26.2 (1984), pages 174–187.

[27] Amit Varma and Bernhard O Palsson. “Metabolic flux balancing: basic concepts, scientific and practical use”. In: *Bio/technology* 12.10 (1994), pages 994–998.

[28] Qasim K Beg et al. “Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity”. In: *Proceedings of the National Academy of Sciences* 104.31 (2007), pages 12663–12668.

[29] Anne Goelzer, Vincent Fromion, and Gérard Scorletti. “Cell design in bacteria as a convex optimization problem”. In: *Automatica* 47.6 (2011), pages 1210–1218.

[30] Geoffrey Stephen Kirk, John Earle Raven, Malcolm Schofield, et al. *The presocratic philosophers: a critical history with a selection of texts*. Cambridge University Press, 1983.

[31] Marcus Terentius Varro. *The Three Books of M. Terentius Varro Concerning Agriculture*. University Press, 1800.

[32] Hans Bremer and Patrick P Dennis. “Modulation of chemical composition and other parameters of the cell at different exponential growth rates”. In: *EcoSal Plus* 3.1 (2008).

[33] John B Bateman, Jack Wagman, and Edwin L Carstensen. “Refraction and absorption of light in bacterial suspensions”. In: *Kolloid-Zeitschrift und Zeitschrift für Polymere* 208.1 (1966), pages 44–58.

[34] Keiran Stevenson et al. “General calibration of microbial growth in microplate readers”. In: *Scientific reports* 6.1 (2016), pages 1–7.

[35] Ruedi Aebersold and Matthias Mann. “Mass spectrometry-based proteomics”. In: *Nature* 422.6928 (2003), pages 198–207.

[36] Kondethimmanahalli Chandramouli and Pei-Yuan Qian. “Proteomics: challenges, techniques and possibilities to overcome biological sample complexity”. In: *Human genomics and proteomics: HGP* 2009 (2009).

[37] Michael Karas, Doris Bachmann, and Franz Hillenkamp. “Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules”. In: *Analytical chemistry* 57.14 (1985), pages 2935–2939.

[38] Masamichi Yamashita and John B Fenn. “Electrospray ion source. Another variation on the free-jet theme”. In: *The Journal of Physical Chemistry* 88.20 (1984), pages 4451–4459.

[39] Christine V Sapan and Roger L Lundblad. “Review of methods for determination of total protein and peptide concentration in biological samples”. In: *PROTEOMICS–Clinical Applications* 9.3-4 (2015), pages 268–276.

[40] Marion M Bradford. “A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding”. In: *Analytical biochemistry* 72.1-2 (1976), pages 248–254.

[41] Christoph Wittmann and Elmar Heinzle. “Mass spectrometry for metabolic flux analysis”. In: *Biotechnology and Bioengineering* 62.6 (1999), pages 739–750.

[42] Gopal Jee Gopal and Awanish Kumar. “Strategies for the production of recombinant protein in Escherichia coli”. In: *The protein journal* 32.6 (2013), pages 419–425.

[43] Hans Peter Sørensen and Kim Kusk Mortensen. “Advanced genetic strategies for recombinant protein expression in Escherichia coli”. In: *Journal of biotechnology* 115.2 (2005), pages 113–128.

[44] Peter A Fields et al. “Adaptations of protein structure and function to temperature: there is more than one way to ‘skin a cat’”. In: *Journal of Experimental Biology* 218.12 (2015), pages 1801–1811.

[45] Peter A Fields. “Protein function at thermal extremes: balancing stability and flexibility”. In: *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 129.2-3 (2001), pages 417–431.

[46] William R Boorstein, Thomas Ziegelhoffer, and Elizabeth A Craig. “Molecular evolution of the HSP70 multigene family”. In: *Journal of molecular evolution* 38.1 (1994), pages 1–17.

[47] Mathieu Rebeaud et al. “On the evolution of chaperones and co-chaperones and the expansion of proteomes across the Tree of Life”. In: *bioRxiv* (2020).

[48] Nobuhiko Tokuriki and Dan S Tawfik. “Chaperonin overexpression promotes genetic variation and enzyme evolution”. In: *Nature* 459.7247 (2009), pages 668–673.

[49] Francis HC Crick. “On protein synthesis”. In: *Symp Soc Exp Biol*. Volume 12. 138-63. 1958, page 8.

[50] CB Anfinsen and HA Scheraga. “Experimental and theoretical aspects of protein folding”. In: *Advances in protein chemistry*. Volume 29. Elsevier, 1975, pages 205–300.

[51] Peter E Wright and H Jane Dyson. “Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm”. In: *Journal of molecular biology* 293.2 (1999), pages 321–331.

[52] Christian B Anfinsen et al. “Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease”. In: *Journal of Biological Chemistry* 207.1 (1954), pages 201–210.

[53] Christian B Anfinsen. “Principles that govern the folding of protein chains”. In: *Science* 181.4096 (1973), pages 223–230.

[54] Noel T Southall, Ken A Dill, and ADJ Haymet. “A view of the hydrophobic effect”. In: *The Journal of Physical Chemistry B* 106.3 (2002), pages 521–533.

[55] Jan Kubelka, James Hofrichter, and William A Eaton. “The protein folding ‘speed limit’”. In: *Current opinion in structural biology* 14.1 (2004), pages 76–88.

[56] Mourad Sadqi, Lisa J Lapidus, and Victor Muñoz. “How fast is protein hydrophobic collapse?” In: *Proceedings of the National Academy of Sciences* 100.21 (2003), pages 12117–12122.

[57] Neil Ferguson and Alan R Fersht. “Early events in protein folding”. In: *Current opinion in structural biology* 13.1 (2003), pages 75–81.

[58] Daria V Fedyukina and Silvia Cavagnero. “Protein folding at the exit tunnel”. In: *Annual review of biophysics* 40 (2011), pages 337–359.

[59] Efrain Siller et al. “Slowing bacterial translation speed enhances eukaryotic protein folding efficiency”. In: *Journal of molecular biology* 396.5 (2010), pages 1310–1318.

[60] Ulfat I Baig et al. “Protein aggregation in E. coli: short term and long term effects of nutrient density”. In: *PloS one* 9.9 (2014), e107445.

[61] Daniel N Wilson, Stefan Arenz, and Roland Beckmann. “Translation regulation via nascent polypeptide-mediated ribosome stalling”. In: *Current Opinion in Structural Biology* 37 (2016), pages 123–133.

[62] Donald Oliver, Jessica Norman, and Shameema Sarker. “Regulation of Escherichia coli secA by cellular protein secretion proficiency requires an intact gene X signal sequence and an active translocon”. In: *Journal of bacteriology* 180.19 (1998), pages 5240–5242.

[63] Thomas F Clarke IV and Patricia L Clark. “Rare codons cluster”. In: *PloS one* 3.10 (2008), e3412.

[64] Jianli Lu and Carol Deutsch. “Electrostatics in the ribosomal tunnel modulate chain elongation rates”. In: *Journal of molecular biology* 384.1 (2008), pages 73–86.

[65] Christopher J Woolstenhulme et al. “Nascent peptides that block protein synthesis in bacteria”. In: *Proceedings of the National Academy of Sciences* 110.10 (2013), E878–E887.

[66] Paul Huter et al. “Structural basis for polyproline-mediated ribosome stalling and rescue by the translation elongation factor EF-P”. In: *Molecular cell* 68.3 (2017), pages 515–527.

[67] Valerie Daggett and Alan R Fersht. “Is there a unifying mechanism for protein folding?” In: *Trends in biochemical sciences* 28.1 (2003), pages 18–25.

[68] Peter S Kim and Robert L Baldwin. “Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding”. In: *Annual review of biochemistry* 51.1 (1982), pages 459–489.

[69] OB Ptitsyn et al. “Evidence for a molten globule state as a general intermediate in protein folding”. In: *FEBS letters* 262.1 (1990), pages 20–24.

[70] Alan R Fersht. “Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications”. In: *Proceedings of the National Academy of Sciences* 92.24 (1995), pages 10869–10873.

[71] John Ellis. “Proteins as molecular chaperones”. In: *Nature* 328.6129 (1987), pages 378–379.

[72] Gene-Wei Li et al. “Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources”. In: *Cell* 157.3 (2014), pages 624–635.

[73] Giulia Calloni et al. “DnaK functions as a central hub in the E. coli chaperone network”. In: *Cell reports* 1.3 (2012), pages 251–264.

[74] Pierre Genevaux, Costa Georgopoulos, and William L Kelley. “The Hsp70 chaperone machines of Escherichia coli: a paradigm for the repartition of chaperone functions”. In: *Molecular microbiology* 66.4 (2007), pages 840–857.

[75] Costa Georgopoulos and WJ Welch. “Role of the major heat shock proteins as molecular chaperones”. In: *Annual review of cell biology* 9.1 (1993), pages 601–634.

[76] Arthur L Horwich et al. “Folding in vivo of bacterial cytoplasmic proteins: role of GroEL”. In: *Cell* 74.5 (1993), pages 909–917.

[77] Axel Mogk et al. “Identification of thermolabile Escherichia coli proteins: prevention and reversion of aggregation by DnaK and ClpB”. In: *The EMBO journal* 18.24 (1999), pages 6934–6949.

[78] H Nakamoto and L Vigh. “The small heat shock proteins and their clients”. In: *Cellular and Molecular Life Sciences* 64.3 (2007), pages 294–306.

[79] Anja Hoffmann, Bernd Bukau, and Günter Kramer. “Structure and function of the molecular chaperone Trigger Factor”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1803.6 (2010), pages 650–661.

[80] Haruo Saito and Hisao Uchida. “Initiation of the DNA replication of bacteriophage lambda in Escherichia coli K12”. In: *Journal of molecular biology* 113.1 (1977), pages 1–25.

[81] Rick Russell, Robert Jordan, and Roger McMacken. “Kinetic characterization of the ATPase cycle of the DnaK molecular chaperone”. In: *Biochemistry* 37.2 (1998), pages 596–607.

[82] Lyra Chang et al. “Mutagenesis reveals the complex relationships between ATPase rate and the chaperone activities of Escherichia coli heat shock protein 70 (Hsp70/DnaK)”. In: *Journal of Biological Chemistry* 285.28 (2010), pages 21282–21291.

[83] Ezra V Pierpaoli et al. “The power stroke of the DnaK/DnaJ/GrpE molecular chaperone system”. In: *Journal of molecular biology* 269.5 (1997), pages 757–768.

[84] Stefan Rüdiger, Jens Schneider-Mergener, and Bernd Bukau. “Its substrate specificity characterizes the DnaJ co-chaperone as a scanning factor for the DnaK chaperone”. In: *The EMBO journal* 20.5 (2001), pages 1042–1050.

[85] KH Paek and GRAHAM C Walker. “Escherichia coli dnaK null mutants are inviable at high temperature.” In: *Journal of bacteriology* 169.1 (1987), pages 283–290.

[86] Elke Deuerling et al. “Trigger factor and DnaK cooperate in folding of newly synthesized proteins”. In: *Nature* 400.6745 (1999), pages 693–696.

[87] Tatsuya Niwa et al. “Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins”. In: *Proceedings of the National Academy of Sciences* 106.11 (2009), pages 4201–4206.

[88] DEBBIE Ang and COSTA Georgopoulos. “The heat-shock-regulated grpE gene of Escherichia coli is required for bacterial growth at all temperatures but is dispensable in certain mutant backgrounds.” In: *Journal of bacteriology* 171.5 (1989), pages 2748–2755.

[89] Tomoya Baba et al. “Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection”. In: *Molecular systems biology* 2.1 (2006).

[90] Emily CA Goodall et al. “The essential genome of Escherichia coli K-12”. In: *MBio* 9.1 (2018), e02096–17.

[91] OLIVIER Fayet, T Ziegelhoffer, and C Georgopoulos. “The groES and groEL heat shock gene products of Escherichia coli are essential for bacterial growth at all temperatures.” In: *Journal of bacteriology* 171.3 (1989), pages 1379–1385.

[92] Alan M Roseman et al. “The chaperonin ATPase cycle: mechanism of allosteric switching and movements of substrate-binding domains in GroEL”. In: *Cell* 87.2 (1996), pages 241–251.

[93] Zhanglin Lin, Frederick P Schwarz, and Edward Eisenstein. “The hydrophobic nature of GroEL-substrate binding”. In: *Journal of Biological Chemistry* 270.3 (1995), pages 1011–1014.

[94] Shubhasis Haldar et al. “Chaperonin-assisted protein folding: relative population of asymmetric and symmetric GroEL: GroES complexes”. In: *Journal of molecular biology* 427.12 (2015), pages 2244–2255.

[95] Daniel K Clare et al. “ATP-triggered conformational changes delineate substrate-binding and-folding mechanics of the GroEL chaperonin”. In: *Cell* 149.1 (2012), pages 113–123.

[96] Pierre Goloubinoff et al. “GroES binding regulates GroEL chaperonin activity under heat shock”. In: *FEBS letters* 407.2 (1997), pages 215–219.

[97] Lizbeth Hedstrom. “Serine protease mechanism and specificity”. In: *Chemical reviews* 102.12 (2002), pages 4501–4524.

[98] Andreas Martin, Tania A Baker, and Robert T Sauer. “Distinct static and dynamic interactions control ATPase-peptidase communication in a AAA+ protease”. In: *Molecular cell* 27.1 (2007), pages 41–52.

[99] Sina Langklotz, Ulrich Baumann, and Franz Narberhaus. “Structure and function of the bacterial AAA protease FtsH”. In: *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research* 1823.1 (2012), pages 40–48.

[100] Istvan Botos et al. “The catalytic domain of Escherichia coli Lon protease has a unique fold and a Ser-Lys dyad in the active site”. In: *Journal of Biological Chemistry* 279.9 (2004), pages 8140–8148.

[101] Seong-Cheol Park et al. “Oligomeric structure of the ATP-dependent protease La (Lon) of Escherichia coli.” In: *Molecules & Cells (Springer Science & Business Media BV)* 21.1 (2006).

[102] Ewa Laskowska et al. “Degradation by proteases Lon, Clp and HtrA, of Escherichia coli proteins aggregated in vivo by heat shock; HtrA protease action in vivo and in vitro”. In: *Molecular microbiology* 22.3 (1996), pages 555–571.

[103] Ran Rosen et al. “Protein aggregation in Escherichia coli: role of proteases”. In: *FEMS microbiology letters* 207.1 (2002), pages 9–12.

[104] Hajime Niwa et al. “Hexameric ring structure of the ATPase domain of the membrane-integrated metalloprotease FtsH from Thermus thermophilus HB8”. In: *Structure* 10.10 (2002), pages 1415–1424.

[105] Teru Ogura et al. “Balanced biosynthesis of major membrane components through regulated degradation of the committed enzyme of lipid A biosynthesis by the AAA protease FtsH (HflB) in Escherichia coli”. In: *Molecular microbiology* 31.3 (1999), pages 833–844.

[106] Takashi Tatsuta et al. “Heat shock regulation in the ftsH null mutant of Escherichia coli: dissection of stability and activity control mechanisms of σ32in vivo”. In: *Molecular microbiology* 30.3 (1998), pages 583–593.

[107] Kai Westphal et al. “A trapping approach reveals novel substrates and physiological functions of the essential protease FtsH in Escherichia coli”. In: *Journal of Biological Chemistry* 287.51 (2012), pages 42962–42971.

[108] Akio Kihara, Yoshinori Akiyama, and Koreaki Ito. “FtsH is required for proteolytic elimination of uncomplexed forms of SecY, an essential protein translocase subunit”. In: *Proceedings of the National Academy of Sciences* 92.10 (1995), pages 4532–4536.

[109] Herbert W Boyer and Daisy Roulland-Dussoix. “A complementation analysis of the restriction and modification of DNA in Escherichia coli”. In: *Journal of molecular biology* 41.3 (1969), pages 459–472.

[110] David A Jackson, Robert H Symons, and Paul Berg. “Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli”. In: *Proceedings of the National Academy of Sciences* 69.10 (1972), pages 2904–2909.

[111] Stanley N Cohen et al. “Construction of biologically functional bacterial plasmids in vitro”. In: *Proceedings of the National Academy of Sciences* 70.11 (1973), pages 3240–3244.

[112] John F Morrow et al. “Replication and Transcription of Eukaryotic DNA in Esherichia coli”. In: *Proceedings of the National Academy of Sciences* 71.5 (1974), pages 1743–1747.

[113] Stanley N Cohen, Annie CY Chang, and Leslie Hsu. “Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA”. In: *Proceedings of the National Academy of Sciences* 69.8 (1972), pages 2110–2114.

[114] An-Chun Chien, Norbert S Hill, and Petra Anne Levin. “Cell size control in bacteria”. In: *Current biology* 22.9 (2012), R340–R349.

[115] Sattar Taheri-Araghi et al. “Cell-size control and homeostasis in bacteria”. In: *Current Biology* 25.3 (2015), pages 385–391.

[116] Ki Jun Jeong and Sang Yup Lee. “Enhanced production of recombinant proteins in Escherichia coli by filamentation suppression”. In: *Applied and environmental microbiology* 69.2 (2003), pages 1295–1298.

[117] Ario de Marco et al. “Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli”. In: *BMC biotechnology* 7.1 (2007), pages 1–9.

[118] Ario De Marco. “Protocol for preparing proteins with improved solubility by co-expressing with molecular chaperones in Escherichia coli”. In: *Nature protocols* 2.10 (2007), page 2632.

[119] Stéphane Pinhal et al. “Acetate metabolism and the inhibition of bacterial growth by acetate”. In: *Journal of bacteriology* 201.13 (2019).

[120] Moshe Kafri et al. “The cost of protein production”. In: *Cell reports* 14.1 (2016), pages 22–31.

[121] Martin Lemmerer et al. “Decoupling of recombinant protein production from Escherichia coli cell growth enhances functional expression of plant Leloir glycosyltransferases”. In: *Biotechnology and bioengineering* 116.6 (2019), pages 1259–1268.

[122] Joseph R Tunner et al. “Use of glucose starvation to limit growth and induce protein production in Escherichia coli”. In: *Biotechnology and bioengineering* 40.2 (1992), pages 271–279.

[123] Claire Turner, Malcolm E Gregory, and Michael K Turner. “A study of the effect of specific growth rate and acetate on recombinant protein production of Escherichia coli JM107”. In: *Biotechnology letters* 16.9 (1994), pages 891–896.

[124] Jingxing Ou et al. “Stationary phase protein overproduction is a fundamental capability of Escherichia coli”. In: *Biochemical and biophysical research communications* 314.1 (2004), pages 174–180.

[125] Costa Georgopoulos. “Toothpicks, serendipity and the emergence of the Escherichia coli DnaK (Hsp70) and GroEL (Hsp60) chaperone machines”. In: *Genetics* 174.4 (2006), pages 1699–1707.

[126] Dawn A Parsell and Robert T Sauer. “Induction of a heat shock-like response by unfolded protein in Escherichia coli: dependence on protein level not protein degradation.” In: *Genes & Development* 3.8 (1989), pages 1226–1232.

[127] Frederick C Neidhardt and Ruth A VanBogelen. “Positive regulatory gene for temperature-controlled proteins in Escherichia coli”. In: *Biochemical and biophysical research communications* 100.2 (1981), pages 894–900.

[128] Kit Tilly et al. “The dnaK protein modulates the heat-shock response of Escherichia coli”. In: *Cell* 34.2 (1983), pages 641–646.

[129] Alan D Grossman, James W Erickson, and Carol A Gross. “The htpR gene product of E. coli is a sigma factor for heat-shock promoters”. In: *Cell* 38.2 (1984), pages 383–390.

[130] David B Straus, William A Walter, and Carol A Gross. “The heat shock response of E. coli is regulated by changes in the concentration of σ32”. In: *Nature* 329.6137 (1987), pages 348–351.

[131] David Straus, William Walter, and Carol A Gross. “DnaK, DnaJ, and GrpE heat shock proteins negatively regulate heat shock gene expression by controlling the synthesis and stability of sigma 32.” In: *Genes & Development* 4.12a (1990), pages 2202–2209.

[132] Dorota Skowyra, Costa Georgopoulos, and Maclej Zylicz. “The E. coli dnaK gene product, the hsp70 homolog, can reactivate heat-inactivated RNA polymerase in an ATP hydrolysis-dependent manner”. In: *Cell* 62.5 (1990), pages 939–944.

[133] Krzysztof Liberek et al. “The DnaK chaperone modulates the heat shock response of Escherichia coli by binding to the sigma 32 transcription factor.” In: *Proceedings of the National Academy of Sciences* 89.8 (1992), pages 3516–3520.

[134] Elizabeth A Craig and Carol A Gross. “Is hsp70 the cellular thermometer?” In: *Trends in biochemical sciences* 16 (1991), pages 135–140.

[135] Toshifumi Tomoyasu et al. “Escherichia coli FtsH is a membrane-bound, ATP-dependent protease which degrades the heat-shock transcription factor sigma 32.” In: *The EMBO Journal* 14.11 (1995), pages 2551–2560.

[136] Christophe Herman et al. “Degradation of sigma 32, the heat shock regulator in Escherichia coli, is governed by HflB.” In: *Proceedings of the National Academy of Sciences* 92.8 (1995), pages 3516–3520.

[137] Eric Guisbert et al. “Convergence of molecular, modeling, and systems approaches for an understanding of the Escherichia coli heat shock response”. In: *Microbiology and Molecular Biology Reviews* 72.3 (2008), pages 545–554.

[138] Hiroki Nagai, Harumi Yuzawa, and Takashi Yura. “Interplay of two cis-acting mRNA regions in translational control of sigma 32 synthesis during the heat shock response of Escherichia coli”. In: *Proceedings of the National Academy of Sciences* 88.23 (1991), pages 10515–10519.

[139] Lisa-Marie Bittner, Jan Arends, and Franz Narberhaus. “When, how and why? Regulated proteolysis by the essential FtsH protease in Escherichia coli”. In: *Biological chemistry* 398.5-6 (2017), pages 625–635.

[140] JW Erickson et al. “Regulation of the promoters and transcripts of rpoH, the Escherichia coli heat shock regulatory gene.” In: *Genes & development* 1.5 (1987), pages 419–432.

[141] James W Erickson and Carol A Gross. “Identification of the sigma E subunit of Escherichia coli RNA polymerase: a second alternate sigma factor involved in high-temperature gene expression.” In: *Genes & development* 3.9 (1989), pages 1462–1471.

[142] Mark Pallen. “RpoN-dependent transcription of rpoH?” In: *Molecular microbiology* 31.1 (1999), pages 393–393.

[143] Birgitte H Kallipolitis and Poul Valentin-Hansen. “Transcription of rpoH, encoding the Escherichia coli heat-shock regulator σ32, is negatively controlled by the cAMP-CRP/CytR nucleoprotein complex”. In: *Molecular microbiology* 29.4 (1998), pages 1091–1099.

[144] QP Wang and JM Kaguni. “dnaA protein regulates transcriptions of the rpoH gene of Escherichia coli.” In: *Journal of Biological Chemistry* 264.13 (1989), pages 7338–7344.

[145] Gen Nonaka et al. “Regulon and promoter analysis of the E. coli heat-shock factor, σ32, reveals a multifaceted cellular response to heat stress”. In: *Genes & development* 20.13 (2006), pages 1776–1789.

[146] Evan T Powers, David L Powers, and Lila M Gierasch. “FoldEco: a model for proteostasis in E. coli”. In: *Cell reports* 1.3 (2012), pages 265–276.

[147] HJCM El-Samad et al. “Surviving heat shock: control strategies for robustness and performance”. In: *Proceedings of the National Academy of Sciences* 102.8 (2005), pages 2736–2741.

[148] Hana El Samad et al. “Optimal performance of the heat-shock gene regulatory network”. In: *IFAC Proceedings Volumes* 38.1 (2005), pages 19–24.

[149] Hiroyuki Kurata et al. “Module-based analysis of robustness tradeoffs in the heat shock response system”. In: *PLoS computational biology* 2.7 (2006), e59.

[150] Jeremy Gunawardena. “Models in systems biology: the parameter problem and the meanings of robustness”. In: *Elements of Computational Systems Biology. New Jersey: John Wiley and Sons* (2010), pages 21–48.

[151] Frédéric Grognard, Yacine Chitour, and Georges Bastin. “Equilibria and stability analysis of a branched metabolic network with feedback inhibition”. In: *IFAC Proceedings Volumes* 37.3 (2004), pages 171–176.

[152] Ismail Belgacem and Jean-Luc Gouzé. “Global stability of enzymatic chains of full reversible Michaelis-Menten reactions”. In: *Acta biotheoretica* 61.3 (2013), pages 425–436.

[153] David H Anderson and Towanna Roller. “Equilibrium points for nonlinear compartmental models”. In: *Mathematical biosciences* 103.2 (1991), pages 159–201.

[154] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2019.

[155] Martin Feinberg. “Chemical reaction network structure and the stability of complex isothermal reactors—I. The deficiency zero and deficiency one theorems”. In: *Chemical engineering science* 42.10 (1987), pages 2229–2268.

[156] Martin Feinberg. “The existence and uniqueness of steady states for a class of chemical reaction networks”. In: *Archive for Rational Mechanics and Analysis* 132.4 (1995), pages 311–370.

[157] Eberhard O Voit, Harald A Martens, and Stig W Omholt. “150 years of the mass action law”. In: *PLoS Comput Biol* 11.1 (2015), e1004012.

[158] Claude-Louis Berthollet. *Essai de statique chimique*. Volume 1. Didot, 1803.

[159] Andrei B Koudriavtsev, Reginald F Jameson, and Wolfgang Linert. *The law of mass action*. Springer Science & Business Media, 2001.

[160] Cato M Guldberg and P Waage. “Etudes sur l’Affinité”. In: *Forhandlinger: Videnskabs-Selskabet i Christiana* 35 (1864).

[161] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[162] Frank Wilczek. *Why Feynman diagrams almost saved space*. 2016. URL: `https://www.quantamagazine.org/why-feynman-diagrams-are-so-important-20160705/` (visited on 01/15/2021).

[163] Michael J Kerner et al. “Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli”. In: *Cell* 122.2 (2005), pages 209–220.

[164] Sandeep K Sharma et al. “The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase”. In: *Nature chemical biology* 6.12 (2010), page 914.

[165] Eric Guisbert et al. “A chaperone network controls the heat shock response in E. coli”. In: *Genes & development* 18.22 (2004), pages 2812–2821.

[166] Yan Ning Zhou et al. “Isolation and characterization of Escherichia coli mutants that lack the heat shock sigma factor sigma 32.” In: *Journal of Bacteriology* 170.8 (1988), pages 3640–3649.

[167] Bentley Lim et al. “Heat shock transcription factor σ 32 co-opts the signal recognition particle to regulate protein homeostasis in E. coli”. In: *PLoS Biol* 11.12 (2013), e1001735.

[168] Grigory Kolesov et al. “How gene order is influenced by the biophysics of transcription regulation”. In: *Proceedings of the National Academy of Sciences* 104.35 (2007), pages 13948–13953.

[169] Otto Pulkkinen and Ralf Metzler. “Distance matters: the impact of gene proximity in bacterial gene regulation”. In: *Physical review letters* 110.19 (2013), page 198101.

[170] Adam Blaszczak, Costa Georgopoulos, and Krzysztof Liberek. “On the mechanism of FtsH-dependent degradation of the σ32 transcriptional regulator of Escherichia coli and the role of the DnaK chaperone machine”. In: *Molecular microbiology* 31.1 (1999), pages 157–166.

[171] Michael J Pearce et al. “Ubiquitin-like protein involved in the proteasome pathway of Mycobacterium tuberculosis”. In: *Science* 322.5904 (2008), pages 1104–1107.

[172] Xibing Xu et al. “Heat shock transcription factor δ32 is targeted for degradation via an ubiquitin-like protein ThiS in Escherichia coli”. In: *Biochemical and biophysical research communications* 459.2 (2015), pages 240–245.

[173] Alondra Diaz-Acosta et al. “Effect of anaerobic and stationary phase growth conditions on the heat shock and oxidative stress responses in Escherichia coli K-12”. In: *Archives of microbiology* 185.6 (2006), pages 429–438.

[174] G Zames and N Shneydor. “Dither in nonlinear systems”. In: *IEEE Transactions on Automatic Control* 21.5 (1976), pages 660–667.

[175] Andrea Y Weiße et al. “Mechanistic links between cellular trade-offs, gene expression, and growth”. In: *Proceedings of the National Academy of Sciences* 112.9 (2015), E1038–E1047.

[176] Jens Hahn. “From Parts to the Whole”. In: (2020).

[177] Ines Thiele et al. “Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage”. In: *PloS one* 7.9 (2012), e45635.

[178] Roi Adadi et al. “Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters”. In: *PLoS computational biology* 8.7 (2012), e1002575.

[179] Jacques Monod. “Recherches sur la croissance des cultures bacteriennes”. In: (1942).

[180] CH Swanson et al. “Bacterial growth as an optimal process”. In: *Journal of theoretical biology* 12.2 (1966), pages 228–250.

[181] François Jacob and Jacques Monod. “On the regulation of gene activity”. In: *Cold Spring Harbor symposia on quantitative biology*. Volume 26. Cold Spring Harbor Laboratory Press. 1961, pages 193–211.

[182] P Dhurjati et al. “A cybernetic view of microbial growth: modeling of cells as optimal strategists”. In: *Biotechnology and bioengineering* 27.1 (1984), pages 1–9.

[183] Dhinakar S Kompala, Doraiswami Ramkrishna, and George T Tsao. “Cybernetic modeling of microbial growth on multiple substrates”. In: *Biotechnology and bioengineering* 26.11 (1984), pages 1272–1281.

[184] Moselio Schaechter, Ole Maaløe, and Niels O Kjeldgaard. “Dependency on medium and temperature of cell size and chemical composition during balanced growth of Salmonella typhimurium”. In: *Microbiology* 19.3 (1958), pages 592–606.

[185] Arthur Trautwein Henrici et al. “Morphologic variation and the rate of growth of bacteria”. In: (1928).

[186] Patrick P Dennis and Hans Bremer. “Macromolecular composition during steady-state growth of Escherichia coli B/r”. In: *Journal of bacteriology* 119.1 (1974), pages 270–281.

[187] Allen G Marr. “Growth rate of Escherichia coli.” In: *Microbiological reviews* 55.2 (1991), pages 316–333.

[188] Benjamin D Towbin et al. “Optimality and sub-optimality in a bacterial growth law”. In: *Nature communications* 8.1 (2017), pages 1–8.

[189] Rafael U Ibarra, Jeremy S Edwards, and Bernhard O Palsson. “Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth”. In: *Nature* 420.6912 (2002), pages 186–189.

[190] Anne Goelzer and Vincent Fromion. “Bacterial growth rate reflects a bottleneck in resource allocation”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1810.10 (2011), pages 978–988.

[191] Anne Goelzer et al. “Quantitative prediction of genome-wide resource allocation in bacteria”. In: *Metabolic engineering* 32 (2015), pages 232–243.

[192] Ana Bulović et al. “Automated generation of bacterial resource allocation models”. In: *Metabolic engineering* (2019).

[193] Matthew Scott and Terence Hwa. “Bacterial growth laws and their applications”. In: *Current opinion in biotechnology* 22.4 (2011), pages 559–565.

[194] IBM ILOG Cplex. “V12. 1: User’s Manual for CPLEX”. In: *International Business Machines Corporation* 46.53 (2009), page 157.

[195] Anne Goelzer, Vincent Fromion, and Gérard Scorletti. “Cell design in bacteria as a convex optimization problem controller”. In: *CDC*. 2009.

[196] Erez Dekel and Uri Alon. “Optimality and evolutionary tuning of the expression level of a protein”. In: *Nature* 436.7050 (2005), page 588.

[197] Matthew Scott et al. “Interdependence of cell growth and gene expression: origins and consequences”. In: *Science* 330.6007 (2010), pages 1099–1102.

[198] Douwe Molenaar et al. “Shifts in growth strategies reflect tradeoffs in cellular economics”. In: *Molecular systems biology* 5.1 (2009).

[199] Robert Schuetz, Lars Kuepfer, and Uwe Sauer. “Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli”. In: *Molecular systems biology* 3.1 (2007), page 119.

[200] Joshua A Lerman et al. “In silico method for modelling metabolism and gene product expression at genome scale”. In: *Nature communications* 3 (2012), page 929.

[201] Zachary A King et al. “Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways”. In: *PLoS computational biology* 11.8 (2015), e1004321.

[202] Wolfram Liebermeister et al. “Visual account of protein investment in cellular functions”. In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pages 8488–8493.

[203] UniProt Consortium. “UniProt: a hub for protein information”. In: *Nucleic acids research* 43.D1 (2014), pages D204–D212.

[204] Michael Hucka et al. “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”. In: *Bioinformatics* 19.4 (2003), pages 524–531.

[205] Brett G Olivier and Frank T Bergmann. “The systems biology markup language (SBML) level 3 package: flux balance constraints”. In: *Journal of integrative bioinformatics* 12.2 (2015), pages 660–690.

[206] BioSys group. *XML format for RBA models,version 1*. INRA Jouy, France, Dec. 2018.

[207] Jeffrey D Orth et al. “A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011”. In: *Molecular systems biology* 7.1 (2011), page 535.

[208] Alexander Schmidt et al. “The quantitative and condition-dependent Escherichia coli proteome”. In: *Nature biotechnology* 34.1 (2016), pages 104–110.

[209] Zachary A King et al. “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic acids research* 44.D1 (2016), pages D515–D522.

[210] Frederick Carl Neidhardt, John L Ingraham, and Moselio Schaechter. *Physiology of the bacterial cell: a molecular approach*. Volume 20. Sinauer Sunderland, 1990.

[211] Atsuko Shinhara et al. “Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli”. In: *BMC genomics* 12.1 (2011), page 428.

[212] L Lindahl. *Intermediates and time kinetics of the in vivo assembly of*. 1975.

[213] Kaspar Valgepea et al. “Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins”. In: *Molecular BioSystems* 9.9 (2013), pages 2344–2358.

[214] Herbert E Kubitschek et al. “Independence of buoyant cell density and growth rate in Escherichia coli.” In: *Journal of bacteriology* 158.1 (1984), pages 296–299.

[215] HE Kubitschek, WW Baldwin, and R Graetzer. “Buoyant density constancy during the cell cycle of Escherichia coli.” In: *Journal of bacteriology* 155.3 (1983), pages 1027–1032.

[216] Bart RB Haverkorn Van Rijsewijk et al. “Large-scale 13 C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli”. In: *Molecular systems biology* 7.1 (2011), page 477.

[217] Elke Deuerling et al. “Trigger factor and DnaK possess overlapping substrate pools and binding specificities”. In: *Molecular microbiology* 47.5 (2003), pages 1317–1328.

[218] Anthony P Pugsley. “The complete general secretory pathway in gram-negative bacteria.” In: *Microbiological reviews* 57.1 (1993), pages 50–108.

[219] Dan Davidi et al. “Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements”. In: *Proceedings of the National Academy of Sciences* 113.12 (2016), pages 3401–3406.

[220] Lisa Jeske et al. “BRENDA in 2019: a European ELIXIR core data resource”. In: *Nucleic acids research* 47.D1 (2018), pages D542–D549.

[221] Robert B. Gennis and Valley Stewart. “Respiration”. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Edited by Frederick C. Neidhardt. ASM Press, 1996. Chapter 1.

[222] BART J van Schie et al. “Energy transduction by electron transfer via a pyrrolo-quinoline quinone-dependent glucose dehydrogenase in Escherichia coli, Pseudomonas aeruginosa, and Acinetobacter calcoaceticus (var. lwoffi).” In: *Journal of bacteriology* 163.2 (1985), pages 493–499.

[223] Rob De Jonge et al. “Pyrroloquinoline quinone, a chemotactic attractant for Escherichia coli.” In: *Journal of bacteriology* 178.4 (1996), pages 1224–1226.

[224] MD Elias et al. “C-terminal Periplasmic Domain of Escherichia coliQuinoprotein Glucose Dehydrogenase Transfers Electrons to Ubiquinone”. In: *Journal of Biological Chemistry* 276.51 (2001), pages 48356–48361.

[225] Niv Antonovsky et al. “Sugar synthesis from CO2 in Escherichia coli”. In: *Cell* 166.1 (2016), pages 115–125.

[226] MK Ornston and LN Ornston. “Two forms of D-glycerate kinase in Escherichia coli”. In: *Journal of bacteriology* 97.3 (1969), pages 1227–1233.

[227] Zachary A King et al. “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic acids research* 44.D1 (2015), pages D515–D522.

[228] Avi I Flamholz et al. “Revisiting Trade-offs between Rubisco Kinetic Parameters”. In: *Biochemistry* 58.31 (2019), pages 3365–3376.

[229] Justyna Nocon et al. “Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production”. In: *Metabolic Engineering* 24 (2014), pages 129–138.

[230] Markus Basan et al. “Overflow metabolism in Escherichia coli results from efficient proteome allocation”. In: *Nature* 528.7580 (2015), pages 99–104.

[231] Laurent Tournier, Anne Goelzer, and Vincent Fromion. “Optimal resource allocation enables mathematical exploration of microbial metabolic configurations”. In: *Journal of mathematical biology* 75.6-7 (2017), pages 1349–1380.

[232] Jr. Arthur E. Bryson and Yu-Chi Ho. *Applied Optimal Control*. 270 Madison Avenue, New York: Taylor & Francis Group, 1975.

[233] R. V. Gamkrelidze L. S. Pontryagin V. G. Boltyanskii and E. F. Mishchenko. *Mathematical theory of optimal processes*. Interscience Publishers, 1962.

[234] Inria Saclay Team Commands. *BOCOP: an open source toolbox for optimal control*. `http://bocop.org`. 2017.

[235] Guillaume Jeanne et al. “Dynamical resource allocation models for bioreactor optimization”. In: *IFAC-PapersOnLine* 51.19 (2018), pages 20–23.

[236] M von Smoluchowski. “Drei vortrage uber diffusion, brownsche bewegung und koagulation von kolloidteilchen”. In: *Zeitschrift fur Physik* 17 (1916), pages 557–585.

[237] Arren Bar-Even et al. “The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters”. In: *Biochemistry* 50.21 (2011), pages 4402–4410.

[238] Vickery L Arcus and Adrian J Mulholland. “Temperature, dynamics, and enzyme-catalyzed reaction rates”. In: *Annual review of biophysics* 49 (2020), pages 163–180.

[239] Anne Farewell and Frederick C. Neidhardt. “Effect of temperature on in vivo protein synthetic capacity in Escherichia coli.” In: *Journal of bacteriology* 180.17 (1998), pages 4704–4710.

[240] Michelle L Scalley and David Baker. “Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability”. In: *Proceedings of the National Academy of Sciences* 94.20 (1997), pages 10636–10640.

[241] Andreas Wächter and Lorenz T Biegler. “On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming”. In: *Mathematical programming* 106.1 (2006), pages 25–57.

[242] John L Ingraham, Ole Maaløe, Frederick Carl Neidhardt, et al. *Growth of the bacterial cell*. Sinauer Associates, 1983.

[243] Ian K Blaby et al. “Experimental evolution of a facultative thermophile from a mesophilic ancestor”. In: *Applied and environmental microbiology* 78.1 (2012), pages 144–155.

[244] FC Neidhardt, RA VanBogelen, and V Vaughn. “The genetics and regulation of heat-shock proteins”. In: *Annual review of genetics* 18.1 (1984), pages 295–329.

[245] R Srivastava, MS Peterson, and WE Bentley. “Stochastic kinetic analysis of the Escherichia coli stress circuit using σ32-targeted antisense”. In: *Biotechnology and Bioengineering* 75.1 (2001), pages 120–129.

[246] Roman L Tatusov et al. “The COG database: a tool for genome-scale analysis of protein functions and evolution”. In: *Nucleic acids research* 28.1 (2000), pages 33–36.

VI

A Appendix 169
A.1 Miscellaneous parameter values
A.2 σ^{32} regulon
A.3 Full matrix of the toy RBA model
A.4 Percentages of individual nucleic and amino acid in *E. coli*
A.5 Flux distribution (growth on glucose, no PQQ)
A.6 Flux distribution (growth on glucose, with PQQ)
A.7 Proteomaps of growth on glucose (experiment - prediction comparison)
A.8 Redistribution of resources for the simulated CO_2 engineered strain
A.9 Optimal control model steady state for a range of temperatures
A.10 JSON Bocop-compatible format

Acknowledgments 187

Declaration of authorship 189

Final throughs 191

A. Appendix

A.1 Miscellaneous parameter values

Table A.1.1: Efficiencies of the secretion and folding process machines as computed by a procedure described in subsection 5.1.7. All values are provided in units of $1/h$.

Growth rate	k_{sec}	k_{ch}
0.26	85692.63	12456.46
0.3	102719.05	13494.18
0.4	145285.10	16088.45
0.42	153798.31	16607.31
0.44	162311.52	17126.16
0.46	170824.73	17645.02
0.47	175081.33	17904.45
0.47	175081.33	17904.44
0.55	209134.17	19979.87
0.58	221903.98	20758.15
0.65	251700.22	22574.15
1.27	515609.71	38658.66
1.9	783775.81	55002.61

Table A.1.2: Upper and lower bounds of a set of central carbon metabolism reactions in *E. coli* obtained from fluxomics measurements [216]. These bounds were set on reactions of the *iJO1366* model [207] during FBA simulation for the estimation of individual apparent catalytic rates for growth on glucose, as explained in subsection 5.1.6. ^{a} - Reaction IDs correspond to the ones used in the *iJO1366* model.

Reaction ID^a	Lower bound $[\frac{mmol}{h \times gCDW}]$	Upper bound $[\frac{mmol}{h \times gCDW}]$
EX_glc__D_e	-8.26	/
EX_ac_e	4.89	/
GLCptspp	7.79	8.47
G6PDH2r	2.09	2.69
GND	1.21	2.09
PGI	5.32	6.1
EDD	0.09	1.39
PFK	5.84	7.08
TKT1	0.38	0.68
TKT2	0.12	0.42
TALA	0.38	0.68
GAPD	13.03	14.71
ENO	12.09	13.77
GLCptspp + PYK	9.01	10.97
PDH	8.5	9.78
CS	1.75	2.65
ICDHyr	1.75	2.65
SUCOAS	-1.73	-0.85
FUM	0.85	1.73

A.2 σ^{32} regulon

Table A.2.1: List of genes in the σ^{32} regulon (taken from EcoCyc [6]). Next to the gene IDs are listed the descriptions of their protein or RNA product, along with the COG category [246] to which they belong. For the category abbreviations used here, see this webiste. The genes listed without the horizontal separation line are transcribed from the same operon.

Gene name	Function	COG ID
topA	DNA topoisomerase I	L
yfbR	dCMP phosphohydrolase	F
mutL	DNA mismatch repair protein	L
miaA	tRNA dimethylallyltransferase	A
hfq	RNA-binding protein	A
hflX	ribosome rescue factor	J
hflK	regulator of FtsH protease	O
hflC	regulator of FtsH protease	O
slt	soluble lytic murein transglycosylase	M
rpmE	50S ribosomal subunit protein	J
yjhB	putative sialic acid tranporter	V
yjhC	putative oxidoreductase	V
yibA	putative lyase s	V
dnaK	chaperone	O
tpke11	putative small RNA	T
dnaJ	chaperone	O
xerD	site-specific recombinase	L
dsbC	protein disulfate isomerase	O
recJ	ssDNA specific exonuclase	L
yhdN	unknown	S
zntR	DNA binding transcriptional activator	R
htpG	chaperone	O
creA	unknown	S
creB	DNA-binding transcriptional regulator	K
creC	sensory histidine kinase	T
can	carbonic anhydrase 2	P
mhpT	3-hydroxyphenylpropionic acid transporter	G
ibpA	small heat shock protein	O
ibpB	small heat shock protein	O
pphA	phosphoprotein phosphatase 1	T
yjaZ	unknown	S
bssS	regulator of biofilm formation	V
hspQ	heat shock protein	T
yafU	putative IM protein	S
fxsA	/	S
yjhI	putative transcriptional regulator	R
yjhH	putative adolase	R
yjhG	D-xylonate dehydratase	G
rrsA	16S ribosomal RNA	J

Ctnd.

Table A.2.1 – *Continued from previous page*

Gene name	Function	Functional category
ileT	tRNA-Ile(GAU)	J
alaT	L-alanyl-tRNA^{alaT}	J
rrlA	23S ribosomal RNA	J
rrfA	5S ribosomal RNA	J
rrsB	16S ribosomal RNA	J
gltT	tRNA-Glu(UUC)	J
rrlB	23S ribosomal RNA	J
rrfB	5S ribosomal RNA	J
rrsC	16S ribosomal RNA	J
gltU	tRNA-Glu(UUC)	J
rrlC	23S ribosomal RNA	J
rrfC	5S ribosomal RNA	J
rrsD	16S ribosomal RNA	J
tilD	tRNA-Ile(GAU)	J
alaU	tRNA-Ala(UGC)	J
rrlD	23S ribosomal RNA	J
rrfD	5S ribosomal RNA	J
thrV	tRNA-Thr(GGU)	J
rrfF	5S ribosomal RNA	J
rrsE	16S ribosomal RNA	J
tgtE	tRNA-Glu(UUC)	J
rrlE	23S ribosomal RNA	J
rrfE	5S ribosomal RNA	J
rrsG	16S ribosomal RNA	J
gltW	tRNA-Glu(UUC)	J
rrlG	23S ribosomal RNA	J
rrfG	5S ribosomal RNA	J
rrsH	16S ribosomal RNA	J
ileV	tRNA-Ile(GAU)	J
alaV	tRNA-Ala(UGC)	J
rrlH	23S ribosomal RNA	J
rrfH	5S ribosomal RNA	J
glnS	glutamine - tRNA ligase	A
lapA	lipopolysaccharide assembly protein A	M
lapB	lipopolysaccharide assembly protein B	M
pyrF	orotidine-5' phosphate decarboxylase	F
yciH	putative translation factor	J
clpB	chaperone	O
cnoX	cheperedoxin (holdase)	O
yjiT	unknown	S
hslR	RNA chaperone (50S rRNA recycling)	A
hslO	molecular chaperone	O
ydhQ	putative adhesin-related protein	R
rdgB	dITP/XTP pyrophosphatase	F

Ctnd.

Table A.2.1 – *Continued from previous page*

Gene name	**Function**	**Functional category**
hemW	heme chaperone	O
lipB	lipoyl (octanoyl) transferase	I
metA	homoserine O-succynil transferase	E
hslV	peptidase component of HslVU protease	O
hslU	ATPase component of HslVU protease	O
trmA	tRNA methyltransferase	A
yrfG	purine nucleotidase	F
hslR	heat shock protein	A
ldhA	D-lactate dehydrogenase	G
rlmE	23S RNA methyltransferase	A
ftsH	ATP-dependent zinc metalloprotease	O
alaA	glutamate - pyruvate aminotransferase	E
ydeO	DNA-binding transcriptional regulator	K
htpX	zinc dependent endoprotease	O
holC	DNA polymerase III subunit	L
valS	valine tRNA ligase	A
ileS	isoleucine tRNA ligase	A
lspA	lipoprotein signal peptidase	O
fkpB	peptidyl-prolyl cis-trans isomerase	O
ispH	1-hydroxy-2-methyl-2-butenyl-4-diphosphate reductase	Q
ycjY	putative hydrolase	R
mpaA	murein tripeptide amidase	O
yafD	endonuclease/exonuclease/phosphatase domain	R
yafE	putative methyltransferase	R
cas2	CRIPR associated endoribonuclease	A
cra	DNA binding transcriptional regulator	K
lon	Lon protease	O
adiC	arginine:agmatine antiporter	V
grpE	nucleotide exchange factor	O
osmF	glycine betaine ABC transporter binding protein	E
yehY	subunit of glycine betain ABC transporter	E
yehX	subunit of glycine betain ABC transporter	E
yehW	subunit of glycine betain ABC transporter	E
raiA	stationary phase translation inhibitor	D
mngA	$2-O-\alpha$ mannosyl-D-glycerate specific PTS enzyme II	G
mngB	α mannosidase	G
pncC	NMN aminohydrolase	F
pgpC	phosphatidylglycerophosphatase C	M
tadA	tRNA adenosine^{3}4 deaminase	A
mlc	DNA-binding transcriptional repressor	K
ynfK	putative dethiobiotin synthetase	R
ackA	acetate kinase	C
ptsH	sugar non-specific of the PTS sugar system	G
ptsI	PTS enzyme I	G

Ctnd.

Table A.2.1 – *Continued from previous page*

Gene name	Function	Functional category
crr	subunit of glucose-specific PTS enzyme II	G
clpP	ATP-dependent Clp protease proteolytic subunit	O
clpX	ATP-dependent Clp protease subunit	O
prlC	oligopeptidase A	O
rsmJ	16S rRNA m^{2}G1516 methyltransferase	A
sdaA	L-serine deaminase I	E
nfuA	iron sulfur cluster carrier protein	O
ribE	6,7-dimethyl-8-ribityllumazine synthase	H
nusB	transcription antitermination protein	K
thiL	thiamine monophosphate kinase	H
pgpA	phosphatidylglycerophosphatase A	M
rfaD	ADP-L-glycero-D-mannoheptose 6-epimerase	G
waaF	ADP-heptose LPS-heptosyltransferase 2	M
waaC	ADP-heptose LPS-heptosyltransferase 1	M
waaL	O-antigen ligase	M
bssS	regulator of biofilm formation	V
gapA	glyceraldehyde-3-phosphate dehydrogenase A	G
yeaD	putative aldose 1-epimerase	R
yfjV	CP4-57 prophage, putative arsenite transporter	R
mutM	DNA formamidinopyrimidine glycosylase	L
rapA	RNAP binding ATPase and recycling factor	K
phoP	phosphorylated DNA-binding transcriptional dual regulator	K
phoQ	sensory histidine kinase	T
groS	cochaperonin GroES	O
groL	chaperonin GroEL	O
tyrR	DNA-binding transcriptional dual regulator	K
rpoD	σ^{70} factor	K
casD	type I-E CRISPR system cascade subunit	V
casE	pre-CRISPR RNA endonuclease	V
cas1	multifunctional nuclease	V
cas2	CRISPR-associated endoribonuclease	V
ybeZ	PhoH-like protein	S
ybeY	endoribonuclease	A
ybeX	CorC-HlyC family protein	S
lnt	apolipoprotein N-acyltransferase	O
rnlA	subunit of toxin-antitoxin system	V
narP	transcriptional dual regulator	K

A.3 Full matrix of the toy RBA model

$$A(\mu) = \begin{array}{c|ccccccccccccccc|} & v_{T_1} & v_{T_2} & v_{T_3} & v_{T_4} & v_{E_1} & v_{E_2} & v_{E_3} & T_1 & T_2 & T_3 & T_4 & E_1 & E_2 & E_3 & R \\ \hline e^{p1} & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ e^{p2} & 0 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ e & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -3\mu n_{T_1} & -3\mu n_{T_2} & -3\mu n_{T_3} & -3\mu n_{T_4} & -3\mu n_{E_1} & -3\mu n_{E_2} & -3\mu n_{E_3} & -3\mu n_R \\ aa^p & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ aa & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -\mu n_{T_1} & -\mu n_{T_2} & -\mu n_{T_3} & -\mu n_{T_4} & -\mu n_{E_1} & -\mu n_{E_2} & -\mu n_{E_3} & -\mu n_R \\ R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu n_{T_1} & \mu n_{T_2} & \mu n_{T_3} & \mu n_{T_4} & \mu n_{E_1} & \mu n_{E_2} & \mu n_{E_3} & \mu n_R - k_T \\ v_{T_1} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^T & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ v_{T_2} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^T & 0 & 0 & 0 & 0 & 0 & 0 \\ v_{T_3} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^T & 0 & 0 & 0 & 0 & 0 \\ v_{T_4} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^T & 0 & 0 & 0 & 0 \\ v_{E_1} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^E & 0 & 0 & 0 \\ v_{E_2} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^E & 0 & 0 \\ v_{E_3} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{app}^E & 0 \\ D_c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & n_{E_1} & n_{E_2} & n_{E_3} & n_R \\ D_m & 0 & 0 & 0 & 0 & 0 & 0 & 0 & n_{T_1} & n_{T_2} & n_{T_3} & n_{T_4} & 0 & 0 & 0 & 0 \end{array} \begin{bmatrix} v_{T_1} \\ v_{T_2} \\ v_{T_3} \\ v_{T_4} \\ v_{E_1} \\ v_{E_2} \\ v_{E_3} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ E_1 \\ E_2 \\ E_3 \\ R \end{bmatrix} \begin{matrix} = \\ = \\ = \\ = \\ = \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \\ \leq \end{matrix} \begin{bmatrix} 0 \\ 0 \\ 3\mu P_{tot}^{ne} \\ 0 \\ \mu P_{tot}^{ne} \\ -\mu P_{tot}^{ne} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ P_{cyt}^e \\ P_{mem}^e \end{bmatrix} \qquad (A.3.1)$$

where $P_{tot}^{ne} = P_{tot}(\mu)(p_c p_c^{ne} + p_m p_m^{ne})$ is the total nonenzymatic protein, and where $P_{cyt}^e = P_{tot} p_c (1 - p_c^{ne})$ and $P_{mem}^e = P_{tot} p_m (1 - p_m^{ne})$ are the enzymatic portion in cytosol and membrane correspondingly.

A.4 Percentages of individual nucleic and amino acid in *E. coli*

Table A.4.1: Percentage of nucleic acids in the ribosomal RNA of *Escherichia coli*. The sequences of 5S, 16S and 23S rRNA can be found on EcoliWiki. (5S, 16S, 23S)

Nucleic acid	Molar mass ($\frac{g}{mol}$)	Percentage (%)
A	345.2	25.8
G	361.2	31.5
C	321.2	22.9
U	322.2	19.8

Table A.4.2: Percentage of amino acids in the *E. coli* proteome. ^{a}Computed from chemical composition of amino acids without water. ^{b}Computed from values measured by [10]

Amino acid	Molar mass ($\frac{g}{mol}$)^a	Percentage (%)^b
Alanine	71.08	9.60
Arginine	156.19	5.53
Asparagine	114.10	4.51
Aspartate	115.09	4.51
Cysteine	103.14	1.71
Glutamate	128.13	4.92
Glutamine	129.16	4.92
Glycine	57.05	11.45
Histidine	137.14	1.77
Isoleucine	113.16	5.43
Leucine	113.16	8.42
Lysine	128.17	6.42
Methionine	131.20	2.87
Phenylalanine	147.18	3.46
Proline	97.12	4.13
Serine	87.08	4.03
Threonine	101.11	4.74
Tryptophan	186.21	1.06
Tyrosine	163.18	2.58
Valine	99.13	7.91

A.5 Flux distribution (growth on glucose, no PQQ)

$\mu = 0.609$

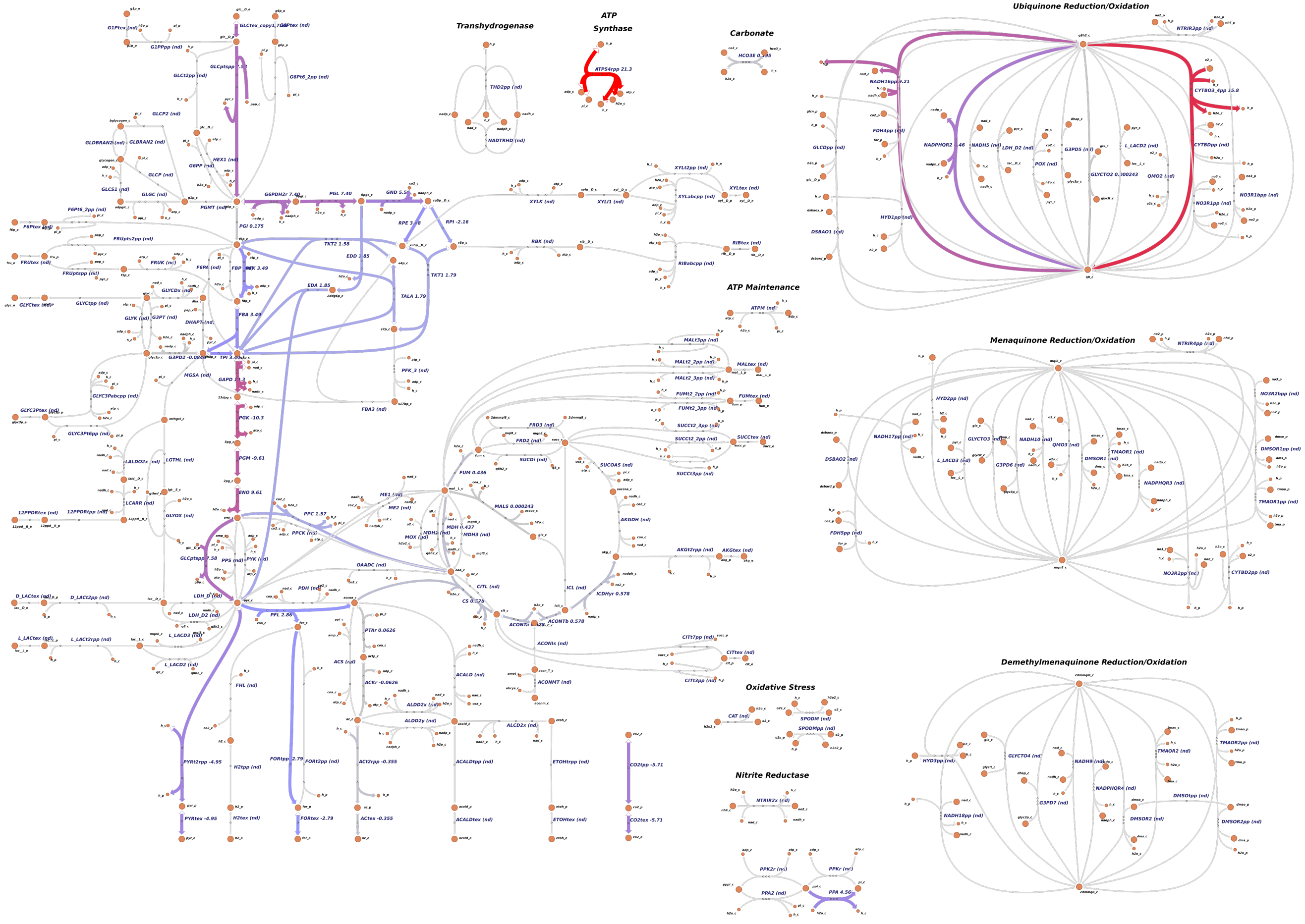


Figure A.5.1: Flux distribution for growth on glucose and no PQQ in the medium. Model predicts respiration through NADH dehydrogenase.

A.6 Flux distribution (growth on glucose, with PQQ)

$\mu = 0.640$

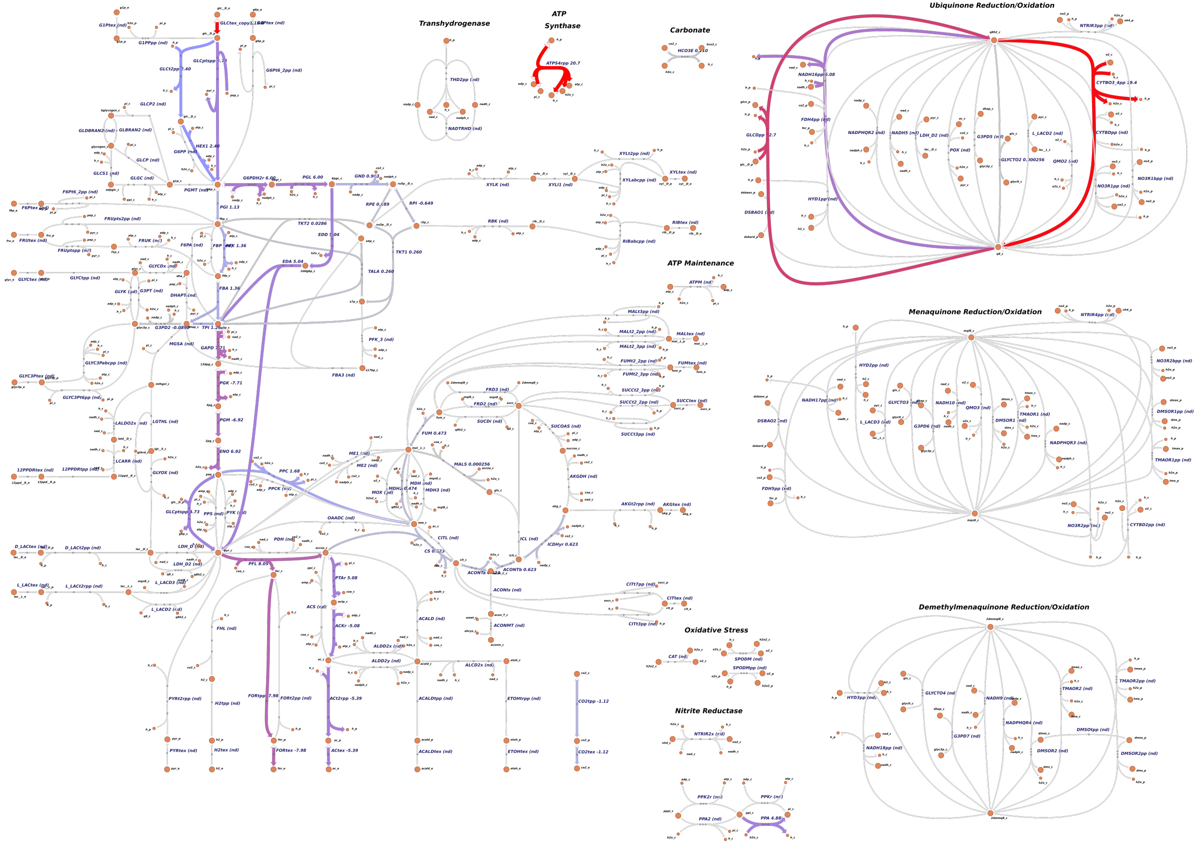
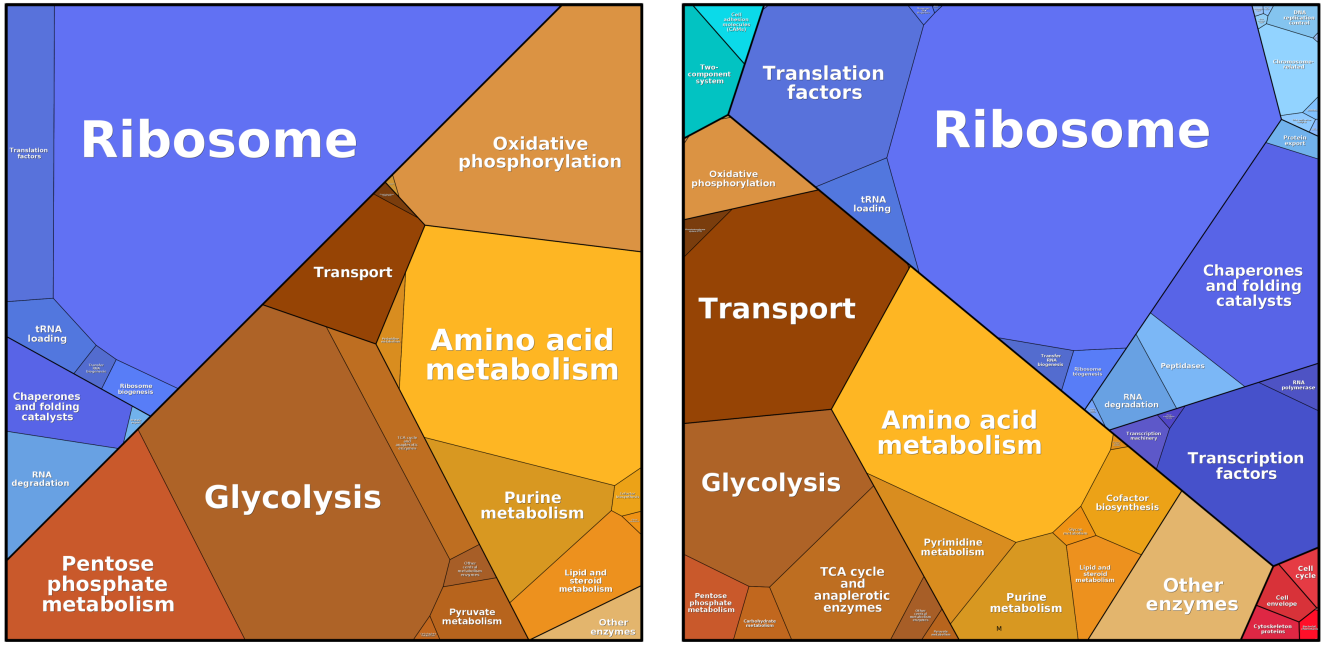

Transhydrogenase
ATP Synthase
Carbonate
Ubiquinone Reduction/Oxidation
ATP Maintenance
Menaquinone Reduction/Oxidation
Oxidative Stress
Nitrite Reductase
Demethylmenaquinone Reduction/Oxidation

Figure A.6.1: Flux distribution for growth on glucose and PQQ in the medium. Model predicts respiration through glucose dehydrogenase, which is cheaper for the cell to produce than the NADH dehydroganse, but requires PQQ as a cofactor. This results in a higher growth rate compared to when no PQQ is available in the medium.

A.7 Proteomaps of growth on glucose (experiment - prediction comparison)



Predicted protein abundances

Measured protein abundances

Figure A.7.1: Left: proteomaps showing the predicted protein abundances for *E. coli* growth on glucose. Right: proteomaps showing the measured protein abundances for *E. coli* growth on glucose [208].

A.8 Redistribution of resources for the simulated CO_2 engineered strain

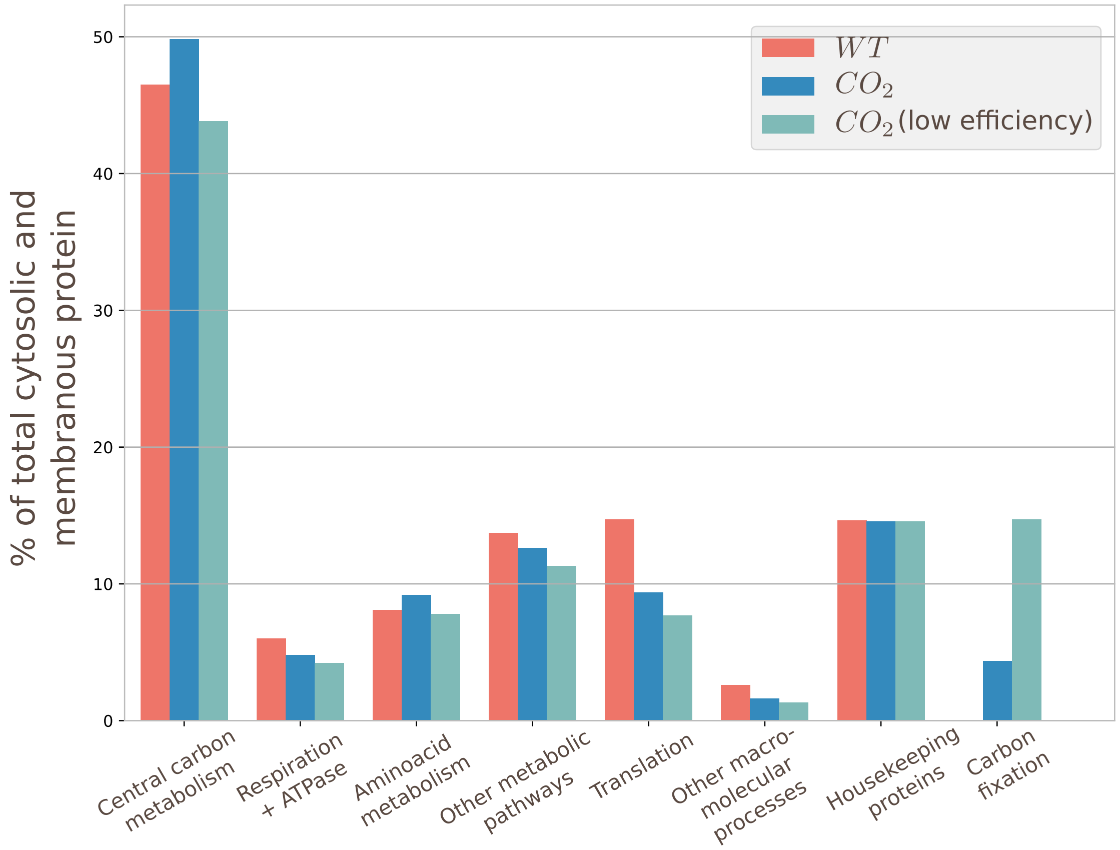


Figure A.8.1: Comparison of cellular resource allocation for the wild type strain, and for the two CO_2 fixing strains, one of low, and one of high Rubisco efficiency.

A.9 Optimal control model steady state for a range of temperatures

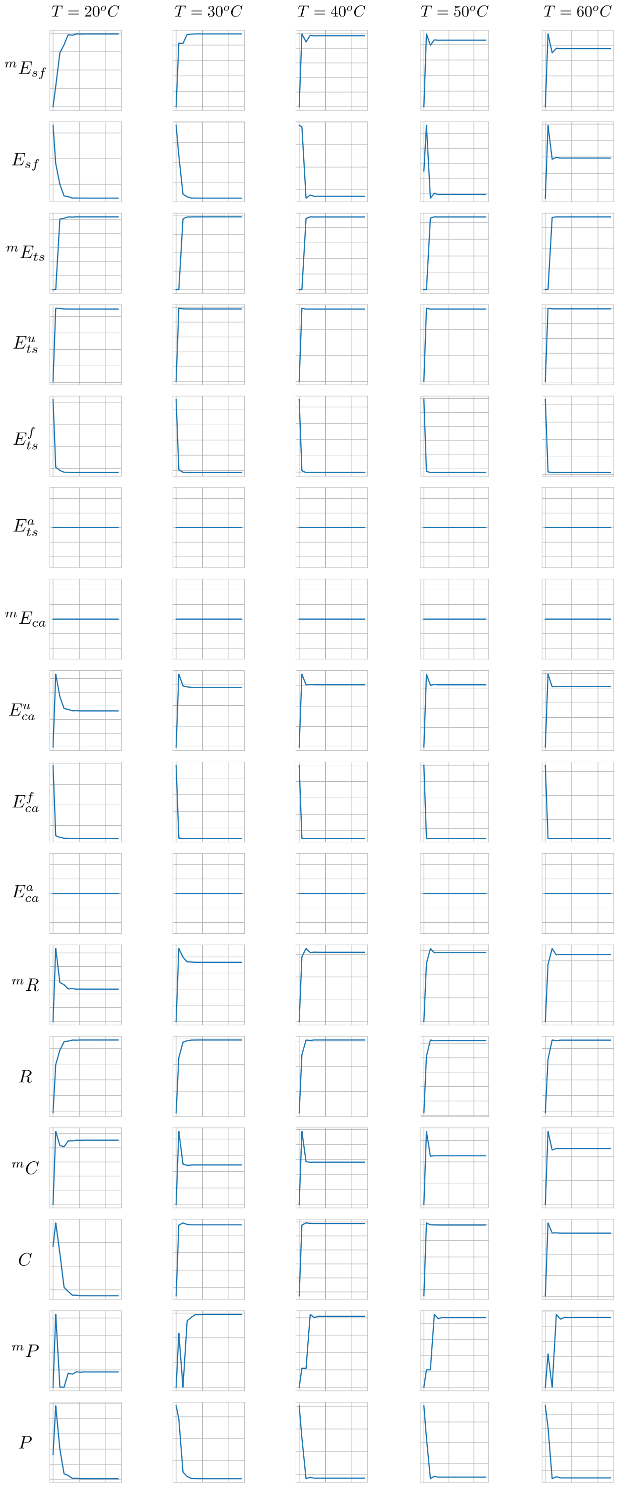


Figure A.9.1: The optimal control heat shock model described in section 6.2 reaches steady state for growth on a biologically relevant range of temperatures. The x-axis shows time evolution over 4 hours, and the y-axis the concentrations of individual macromolecular components (numbers were left out for sake of space).

A.10 JSON Bocop-compatible format

```
{
        "modelName": "Simple RBA problem",
        "states":
        [
                {
                        "name": "X",
                        "bound": {
                                "type": "lower",
                                "lb":  0,
                                "ub":  2e+20
                        },
                        "expression": "mu * X"
                 },
                 {
                        "name": "S_ext",
                        "bound": {
                                "type": "lower",
                                "lb":  0,
                                "ub":  2e+20
                        },
                        "expression": "-mu * X"
                 },
                 {
                        "name": "E",
                        "bound": {
                                "type": "lower",
                                "lb":  0,
                                "ub":  2e+20
                        },
                        "expression": "nu_P_E - mu * E"
                 },
                 {
                        "name": "R",
                        "bound": {
                                "type": "lower",
                                "lb":  0,
                                "ub":  2e+20
                        },
                        "expression": "nu_P_R - mu * R"
                 }
        ],
        "controls":
        [
                {
                        "name": "nu_P_E",
                        "bound": {
                                "type": "lower",
                                "lb":  0,
                                "ub":  2e+20
                        }
                },
                {
                        "name": "nu_P_R",
                        "bound": {
                                "type": "lower",
                                "lb":   0,
                                "ub":   2e+20
                        }
                }
        ],
        "parameters":
        [
        ],
        "algebraic":
        [
                {
                        "name": "mu",
                        "bound": {
                                "type": "free",
                                "lb":   0,
                                "ub":   0
                        }
                },
                {
                        "name": "nu_M",
                        "bound": {
                                "type": "free",
                                "lb":   0,
                                "ub":   0
                        }
                }
        ],
        "constants":
        [
                {
                        "name":         "n_R",
                        "value":        180000
                },
                {
                        "name":         "n_E",
                        "value":        20000
                },
                {
                        "name":         "D_c",
                        "value":        4.89
                },
                {
                        "name":         "k_R_slope",
                        "value":        84
                },
                {
                        "name":         "k_R_intercept",
                        "value":        -747.6
                },
                {
                        "name":         "k_M_slope",
                        "value":        120
                },
                {
                        "name":         "k_M_intercept",
                        "value":        -3000
                },
                {
                        "name":         "T",
                        "value":        37
                },
                {
                        "name":         "n_S",
                        "value":        0.5
                },
                {
                        "name":         "n_HP",
                        "value":        300
                },
                {
                        "name":         "X_0",
                        "value":        100
                },
                {
                        "name":         "S_ext_0",
                        "value":        1000
                }
        ],
        "boundarycond":
        [
                {
                        "name": "D_init",
                        "bound": {
                                "type": "equal",
                                "lb":  0,
                                "ub":  0
                        },
                        "expression": "(E_t0 * n_E + R_t0 * n_R ) - D_c"
                },
                {
                        "name": "X_init",
                        "bound": {
                                "type": "equal",
                                "lb":  0,
                                "ub":  0
                        },
                        "expression": "X_t0 - X_0"
                },
                {
                        "name": "S_ext_init",
                        "bound": {
                                "type": "equal",
                                "lb":  0,
                        "ub":  0
                },
                "expression": "S_ext_t0 - S_ext_0"
        }
],
"pathconstraints":
[
        {
                "name": "growth_rate",
                "bound": {
                        "type": "equal",
                        "lb":  0,
                        "ub":  0
                },
                "expression": "D_c * mu - (nu_P_R * n_R + nu_P_E * n_E)"
        },
        {
                "name": "capacity_met",
                "bound": {
                        "type": "upper",
                        "lb":  -2e+20,
                        "ub":  0
                },
                "expression": "nu_M - (k_M_slope * T + k_M_intercept) * E"
        },
        {
                "name": "capacity_ribo",
                "bound": {
                        "type": "upper",
                        "lb":  -2e+20,
                        "ub":  0
                },
                "expression": "(nu_P_R * n_R + nu_P_E * n_E) - (k_R_slope
                    * T + k_R_intercept) * R"
        },
        {
                "name": "metabolism_ss",
                "bound": {
                        "type": "equal",
                        "lb":  0,
                        "ub":  0
                },
                "expression": "n_S * nu_M - (n_E * nu_P_E + n_R * nu_P_R)"
        }
],
"criterion": "- X",
"time":
{
        "free": "none",
        "initial": 0,
        "final": 20
},
"discretization":
        {
                "steps": 9,
                "method": "euler_imp"
        },
        "fixed_part": "# Optimization :\noptimization.type string single\nbatch.
            type integer 1\nbatch.index integer -1\nbatch.nrange integer 3\
            nbatch.lowerbound double 100\nbatch.upperbound double 1000\nbatch.
            directory string time_step\n\n# Initialization :\ninitialization.
            type string from_init_file\ninitialization.file string none\n\n#
            Parameter identification :\nparamid.type string false\nparamid.
            separator string ,\nparamid.file string no_directory\nparamid.
            dimension integer 0",
        "solution": {
                "file": "problem.sol"
        }
}
```

Acknowledgements

Writing a PhD thesis can be a very selfish business, and those who bare with us in this (often mentally disturbing) process deserve many thanks.

I would first like to thank my supervisor Edda Klipp. Through these many years of my doctorate she has always been kind and supportive, and has allowed me a great freedom in the direction of my research. Also, she accepted me as a PhD candidate, even if I had no obvious previous knowledge of the topic.

I would then like to thank the TBP research group, for always providing a pleasent and relaxed work environment. I especially want to help Judith Wodtke and Katja Tummler for their help with reading and editing of the thesis.

I want to thank Anne Goelzer and Vincent Fromion, for their inexhaustive help, their humour and deep care for honesty in science and in life.

I want to thank Wolfram Liebermeister for many beautiful conversations and his unintrusive but sharp honesty.

I want to thank Paula Martinell Garcia, for our communal unraveling of the one true path which leads Onwards. Sharp as arrows, you and I!

I want to thank Roman Rainer, for being the founding member of our squad and the name-giver, and Olufemi Bolaji for all of this thoughts on this world and beyond.

I want to thank my mother and father, for all the yeses and nos they had chosen well and with care, and for their love for the world of human creation - music, literature, history, science and friendship.

I want to thank Ivan Dokmanić, who once saved me from a very dangerous road of actual work, shared his dark magic with me, and was a great inspiration for searching out the true understanding of things.

Finally, I wish to thank some people that I had never met. Thank you, Bertrand Russell, for the most kind and beautiful rational mind that I have ever come to know, and your honest quest for understanding. Your mental voyages gave wings to my mind, and showed how intellect can be at the same time piercing and gentle.

Thank you, David Bohm, for never keeping out of sight the original thirst for science, which comes from wonder when looking at the world. Your supple mind took this wonder to incredible places.

Declaration of authorship

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Berlin, May 1, 2021

Ana Bulović

Final thoughts

Among all the many misfortunes to which we are heir, it is only fair to admit that we are allowed the greatest degree of freedom of thought.

Andre Breton, Manifesto of Surrealism

Professionalism is environmental. Amateurism is anti-environmental. Professionalism merges the individual into patters of total environment. Amateurism seeks the development of the total awareness of the individual and the critical awareness of the groundrules of society. The amateur can afford to lose. The professional tends to classify and to specialize, to accept uncritically the groundrules of the environment. The groundrules provided by the mass response of his colleagues serve as a pervasive environment of which he is contentedly unaware. The “expert” is the man who stays put.

Marshall McLuhan, Medium is the Message

A civilized man judges and is judged according to his behavior, but even the term "civilized" leads to confusion: a cultivated "civilized" man is regarded as a person instructed in systems, a person who thinks in forms, signs, representations – a monster whose faculty of deriving thoughts from acts, instead of identifying acts with thoughts, is developed to an absurdity. If our life lacks brimstone, Le., a constant magic, it is because we choose to observe our acts and lose ourselves in considerations of their imagined form instead of being impelled by their force.

Antonin Artaud, Theater and its Double

Forgive me, father, I am not certain what my own wishes are. I shall always take pleasure in study, how could it be otherwise? But I do not believe that my life will be limited to study. A man's wishes may not always determine his destiny, his mission; perhaps there are other, predetermining, factors.

Hermann Hesse, Narziss und Goldmund

“I don't speak,” Bijaz said. “I operate the machine called language. It creaks and groans, but is mine own.”.

Frank Herbert, Dune

It was not growing up that slowly applied breaks to learning but an accumulation of “things I know”.

Frank Herbert, Dune

The theory is that my organism tends to actualize itself if I stand out of the way. It is an article of faith.

Paul Goodman