Proceedings of The Evry Spring School
on

advances in
Systems and
Synthetic
Biology

March 24th - 28th, 2014

Edited by

Patrick Amar, François Képès, Vic Norris

"But technology will ultimately and usefully be better served by following the spirit of Eddington, by attempting to provide enough time and intellectual space for those who want to invest themselves in exploration of levels beyond the genome independently of any quick promises for still quicker solutions to extremely complex problems."

Strohman RC (1977) Nature Biotech 15:199

FOREWORD

Systems Biology includes the study of interaction networks and, in particular, their dynamic and spatiotemporal aspects. It typically requires the import of concepts from across the disciplines and crosstalk between theory, benchwork, modelling and simulation. The quintessence of Systems Biology is the discovery of the design principles of Life. The logical next step is to apply these principles to synthesize biological systems. This engineering of biology is the ultimate goal of Synthetic Biology: the rational conception and construction of complex systems based on, or inspired by, biology, and endowed with functions that may be absent in Nature.

This annual School started in 2002. It was the first School dedicated to Systems Biology in France, and perhaps in Europe. Since 2005, Synthetic Biology has played an increasingly important role in the School. Generally, the topics covered by the School have changed from year to year to accompany and sometimes precede a rapidly evolving scientific landscape. Its title has evolved in 2004 and again in 2012 to reflect these changes. The first School was held near Grenoble after which the School has been held in various locations. It started under the auspices of Genopole®, and has been supported by the CNRS since 2003, as well as by several other sponsors over the years.

This book gathers overviews of the talks, original articles contributed by speakers, subgroups and students, tutorial material, and poster abstracts. We thank the sponsors of this conference for making it possible for all the participants to share their enthusiasm and ideas in such a constructive way.

Patrick Amar, Gilles Bernot, Marie Beurton-Aimar, Attila Csikasz-Nagy, Jürgen Jost, Ivan Junier, Marcelline Kaufman, François Képès, Pascale Le Gall, Jean-Pierre Mazat, Victor Norris, William Saurin, El Houssine Snoussi, Birgit Wiltschi.

ACKNOWLEDGEMENTS

We would like to thank the conference participants, who have contributed in a way or another to this book. It gathers overviews of the talks and discussions, original articles and tutorial material contributed by speakers, abstracts from attendees, posters and lectures proposed by the epigenesis groups to review or illustrate matters related to the scientific topic of the conference.

Of course the organisation team would like to express gratitude to all the staff of the *Ibis Evry Cathédrale* hotel for the very good conditions we have found during the conference.

Special thanks to the Epigenomics project for their assistance in preparing this book for publication. The cover photography shows a view of the *Place de l'Agora*, at Évry.

We would also like to express our thanks to the sponsors of this conference for their financial support allowing the participants to share their enthusiasm and ideas in such a constructive way.

They were:

- Centre National de la Recherche Scientifique (CNRS):
  http://www.cnrs.fr
- Genopole® Évry:
  http://www.genopole.fr
- GDRE CNRS 513 Biologie Systémique:
  http://www.mpi-magdeburg.mpg.de/CNRS_MPG
- Consortium BioIntelligence (OSEO)
- Institut National de Recherche en Informatique et en Automatique (INRIA):
  http://www.inria.fr/
- Institut des Systèmes Complexes - Paris Île de France:
  http://www.iscpif.fr/
- GDR CNRS 3003 Bioinformatique Moléculaire:
  http://www.gdr-bim.u-psud.fr

THE EDITORS

INVITED SPEAKERS

CÉCILE BONNARD	Sobios, Boulogne, FR
ANTOINE BRIL	Servier, Suresnes, FR
NICOLAS FROLOFF	Dassault systèmes, Vélizy, FR
ROMAN JERALA	National Inst. of Chemistry, Ljubljana, SI
JEAN PECCOUD	Virginia Tech, US
OLIVIER RIVOIRE	U. Joseph Fourier, Grenoble, FR
ALAN ROBINSON	MRC Mitochondrial Biology Unit, Cambridge, UK
HEIKE SIEBERT	FU Berlin, Berlin, DE
GUY-BART STAN	Imperial College London, UK
INES THIELE	U. Luxembourg, LU
PAUL VANOUSE	Department of Visual Studies, University at Buffalo, US
LINGCHONG YOU	Duke U., Durham NC, US

CONTENTS

SYNTHETIC BIOLOGY 2

PART I INVITED TALKS

bioPLM: a global collaborative platform for multidiscipline scientific innovation

Nicolas Froloff[1]

[1] Dassault systèmes, Vélizy, FR

Abstract

The purpose of the BioIntelligence Program is to develop an integrated software environment for the discovery and development of new biological entities and products (from molecules to biological pathways, cells, organs, including regulatory aspects) for life sciences industries and research institutes, and in particular for pharmaceuticals, cosmetics, and agrochemicals.

This digital environment for scientific collaboration and innovation is aimed at:

- Proposing a unified platform for exploring and analyzing biological information (which is intrinsically heterogeneous and extremely diverse), and for formulating scientific hypotheses to be tested in the lab;
- Building in silico models supported by this bioknowledge, that can be numerically simulated and confronted to experimental data;
- Managing all discovery and development activities of those industries by relying on the foundations necessary to all involved multidiscipline R&D teams (collaboration, industrial processes coverage, certification).

The software applications of the BioPLM platform will be presented, and will be demonstrated on a particular use case in a project of drug discovery in oncology.

Digital Clinic

Cécile Bonnard[1]

[1] Sobios, Boulogne, FR

Abstract

Digital Clinic is a software for modeling and simulation of clinical trials in oncology, integrated in the BioPLM platform. Using PK models, PD models and disease models, Digital Clinic can be used to test specific therapies with specific designs on specific populations.

The session will be focused on the use of Digital Clinic to optimize an administration schedule. During the development of an anticancer drug, real clinical trials have shown that patients encountered drug resistance, which reduce the efficacy of the drug, and therefore the tumor shrinkage. A stronger dose has been considered but the drug has a neutropenia effect that cannot be ignored. A neutropenia model was built linked to the PK model of the drug, and a tumor growth inhibition model was built in previous researches. A combined model can be built from these three models and then be used in Digital Clinic, in order to design an administration schedule that allows a better balance between efficacy and toxicity.

Data integration and knowledge management in therapeutic innovation

Antoine Bril[1]

[1] Institut de recherches Internationales Servier, Suresnes, FR

Abstract

The need to use the right drug at the right dose for the right patient requires evolution in the therapeutic innovation process and in the approaches used to understand pathologies. It is commonly said that the therapeutic innovation carried out in close collaboration between the pharmaceutical industry and academic laboratories, generates data and knowledge that together need to be integrated, shared and eventually exploited.

In this talk we will discuss the challenges facing both industry and academia in the discovery of new therapeutic solutions at a time where the biological complexity is investigated in a very sophisticated manner. Some of the questions that will be addressed are the following: How will it be possible to understand disease complexity? What is the level of information that will be shared between scientists to facilitate therapeutic innovation? Key factors of success will be introduced with particular attention to strategies aimed at developing integrative knowledge management.

Module-based Analysis of Complex Networks

Heike Siebert[1]

[1] Freie Universität Berlin, FB Mathematik und Informatik, Berlin, D

Abstract

The idea of modularization of a complex system appeals on several levels. A structuring of a network into modules and their interactions yields a more manageable representation and might clarify relations between system components. Identifying subnetworks that can be linked to specific functions may allow us to analyze them in isolation and still derive information valid for the original system, thus yielding reduction strategies. In general, an understanding of the interplay of subsystems responsible for distinct dynamical effects giving rise to the overall capabilities of the system opens the door for targeted control, e.g. in the context of drug design, or even network design in synthetic biology. Taking into account the different motivations for modularization, it is not surprising that a whole range of definitions for and related approaches to identification of modules have been developed. However, to make use of them one needs a clear understanding of the respective module properties in relation to the question the module-based analysis is supposed to address in the first place.

In this talk, we will be interested in exploiting modules to obtain a better understanding of the network organization in relation to its function and to reduce complexity for more efficient analysis of the system. Different module definitions will be introduced, starting with a purely static view based on the network topology and ranging to system modules related to specific dynamical aspects. We will look at methods for module identification as well as the related module-based approaches to network analysis. Lastly, we will critically discuss constraints and benefit of such methods.

Inference, organization and evolution of genomic features

Olivier Rivoire[1]

[1] Université Joseph Fourier, Grenoble, FR

Abstract

Genomes are being sequenced at an exponential pace, offering us an unprecedented opportunity to study quantitatively the evolutionary processes that shape their diversity. I will present a statistical framework to extract conserved genomic patterns from sequence data and investigate their evolutionary origins.

Computational modeling of the human metabolic reconstruction and the application to inheritable metabolic diseases

Ines Thiele[1]

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, LU

Abstract

Metabolism plays a central role in many human diseases, including diabetes, obesity and cardiovascular diseases. Shifts in metabolism and metabolite levels are often involved in human disease, either as cause or as consequence of pathogenic changes, and therefore offer great potential for diagnosis and elucidation of disease processes. Metabolic reconstructions describe the biological knowledge about a target organism in a structured manner and permit the conversion into a mathematical format and subsequent computation of physiological properties. As such, they facilitate the investigation of the mechanisms underlying genotype-phenotype relationships and indeed, inheritable metabolic diseases.

Here, we will present the most comprehensive metabolic reconstruction of human metabolism, which has been recently assembled in a community effort, and some promising biomedical applications. In particular, we will illustrate how this reconstruction and derived cell type specific models can be used to further our understanding of network wide, metabolic effects of single enzyme defects that are often associated with inheritable metabolic diseases.

Therefore, we created a detailed compendium of human inborn errors of metabolism captured by the human metabolic reconstruction. Using constraint-based analysis, we were able to predict novel biomarkers and to investigate systemic, metabolic effects of these rare metabolic diseases.

We will illustrate that the human metabolic reconstruction provides a deep insight into complex human metabolic phenotypes and disease states and represents a fundamental tool for the study of the systems biology of human metabolism.

Simulating the complexity of mitochondrial metabolism in health and disease by flux balance analysis

Alan Robinson[1]

[1] MRC Mitochondrial Biology Unit, Cambridge, UK

Abstract

Mitochondria not only produce the biochemical energy source ATP necessary for so many cellular functions, but have a major role in the creation, destruction and re-cycling of many other key metabolites of central metabolism. They are also now recognised as having an important role in homeosatis and apoptosis. Thus it is unsurprising that dysfunction of mitochondrial metabolism occurs not only in the rarer mitochondrial diseases, but also features in many complex and common diseases too, including cancer, heart failure, obesity and diabetes, and neurodegenerative diseases such as Parkinson's.

New discoveries about mitochondrial metabolism are also still continuing to surprise us. For example, the unexpected discovery of reductive carboxylation where parts of bioenergetic metabolism run in reverse to what is expected.

The interconnections of all these metabolic pathways and their control points makes for a truely complex system. Whilst deterministic methods have been used to study mitochondrial ATP production, these have often been limited to only 10's of reactions including glycolysis, the citric acid cycle and the electron transport chain.

In the first part of my talk, I will explain how we have built curated models of mitochondrial metabolism that incorporate 100's of reactions. To simulate and so understand mitochondrial metabolism with such large reaction networks, I will then explain how and why we use flux balance analysis. Whereas this technique makes many assumptions, including regarding the system being at steady-state, it is a powerful method of generating testable hypotheses about how mitochondrial metabolism changes under perturbations to the system that may arise during disease. It can also be use to suggest potential therapeutic options that ameliorate these perturbations.

In the second half of my talk, I will discuss the application of our models to understand the mechanisms of changes in mitochondrial metabolism in both rare mitochondrial diseases and in common conditions, such as the hypoxia experienced by cells during heart failure, stroke, and cancer.

Engineering bacteria to cooperate, sacrifice, and organize

Lingchong You[1]

[1] Duke University, Biomedical Engineering & IGSP, Durham NC, US

Abstract

A major focus of synthetic biology is the engineering of gene circuits to perform user-defined functions. In addition to generating systems of practical applications, such efforts have led to the identification and evaluation of design strategies that enable robust control of dynamics in single cells and in cell populations. On the other hand, there is an increasing emphasis on using engineered systems programmed by simple circuits to explore fundamental biological questions of broad significance. In this talk, I will discuss our efforts along this line of research, whereby we have used engineered gene circuits to examine the evolutionary dynamics of two common bacterial survival strategies and to program self-organized spatial pattern formation . I will also discuss the implications of these systems for medicine and materials fabrication.

References

1. A. Pai, Y. Tanouchi, and L. You. *Optimality and Robustness in quorum sensing (QS)-mediated regulation of a costly public good enzyme*. PNAS (2012).

2. Y. Tanouchi, A. Pai, N.E. Buchler, and L. You. *Programming stress-induced altruistic death in engineered bacteria*. Molecular Systems Biology 8:626 (2012).

3. S. Payne, B. Li, Y, Cao, D. Schaeffer, M.D. Ryser, and L. You. *Temporal control of self-organized pattern formation without morphogen gradients in bacteria*. Molecular Systems Biology. (2013).

Genetic Design Automation: engineering fantasy or scientific renewal?

Jean Peccoud[1]

[1] Virginia Bioinformatics Institute, Virginia Tech, US

Abstract

The hype surrounding the emerging field of synthetic biology has generated excitement and skepticism. The overarching vision of engineering living organisms using methods developed in other engineering fields often sounds naïve to experienced life scientists. However, the spectacular successes of synthetic biology are impossible to ignore. They challenge commonly accepted ideas and invite us to reassess the organization of research programs in biotechnology. In particular, they show that it is possible to organize the product development cycle in three stages focused on design, fabrication, and testing. Significant gains of productivity can be achieved by developing a generic infrastructure that can easily be customized to meet the needs of specific projects. In the foreseeable future, engineering living organisms will remain much slower and less predictable than the development of non-living products. Yet, adapting key engineering concepts to the biotech industry will make the product development cycle faster and cheaper.

Modularity as the engineering principle in synthetic biology

Roman Jerala[1]

[1] National institute of chemistry and Centre of excellence EN-FIST,
Ljubljana, Slovenia

Abstract

Modularity is extensively used in engineering to allow rapid and cost effective construction of different devices and structures. In biological systems functional modules are often obscured as the functional elements, such as proteins have been optimized for specific interactions. Construction of complex devices requires large sets of orthogonal elements, which may be difficult to harvest from nature. Nucleotide sequence provides a large and easily accessible combinatorial diversity that can underlay programming in biological systems. Knowledge of the recognition code of the sequence-specific DNA binding proteins allows us to prepare large number of orthogonal DNA binding proteins. Those domains can be used to encode the assembly of biosynthetic enzymes which enhances the yield of biosynthetic production. On the other hand designable DNA-binding TALE domains can be used to construct genetic logical NOR gates. Application of the single-layer designable NOR gate allowed us to prepare all 16 two-input functional logic gates and more complex information processing circuits. The same type of elements allows construction of dynamic bistable switches where the nonlinearity is introduced through feedback loops. In terms of the modular construction of structures designable orthogonal coiled-coil dimers provide the basic building blocks based on the specificity of interactions between segments of the polypeptide chain. This principle allows the design of completely new modular protein folds, composed of a single polypeptide chain where the final structure and potentially function can be encoded within the designed sequence and can be produced cost effectively in bacteria.

Taking a Systems and Control Engineering Approach in Synthetic Biology

Guy-Bart Stan[1]

[1] Department of Bioengineering, Imperial College London, UK

Abstract

In this talk I will give a brief overview of some of the research activities in my group, the ”Control Engineering Synthetic Biology” group at Imperial College London, where we focus our efforts on developing foundational forward-engineering methods to mathematically model, and rigorously analyse, design and control synthetic gene circuits and cellular metabolism so as to endow engineered cells with novel functionalities. The tools and approaches that we take rely on concepts and principles drawn from Robust Optimal Control and Dynamical Systems theory, applied to Synthetic Biology problems.

Some of the topics covered will include (if time allows): (a) Design of in vivo genetic feedback controllers for automatic robust regulation of branched and unbranched metabolic pathways, and (b) Exogenous data-based optimal feedback control of gene regulatory networks.

(a) Among Synthetic Biology’s most prominent applications is the manipulation of bacterial metabolism for the production of high-value chemicals in diverse sectors such as energy, biomedicine and food technology. In this regard, we are developing foundational tools for the analysis and design of feedback control synthetic biodevices that dynamically regulate bacterial metabolism according to pre-defined objectives such as stability and robustness. Because these feedback controllers are intracellular, they have a great potential for applications where cellular behaviour needs to be controlled without real-time human intervention.

(b) In the second part of the talk, I will present research results pertaining to the inference of (close-to) optimal feedback control strategies for exogenous control of biological systems (natural or synthetic) directly from input-output measurements, i.e., without the need for identifying a mathematical model of the system’s dynamics a priori. The examples discussed include data-based inference of optimal control strategies for regulation and reference trajectory tracking in synthetic gene regulatory networks such as the toggle-switch or the generalised repressilator.

References

1. http://www.bg.ic.ac.uk/research/g.stan/#Publications
2. http://www.bg.ic.ac.uk/research/g.stan/group/
3. http://rsif.royalsocietypublishing.org/content/10/78/20120671.long
4. http://www.bg.ic.ac.uk/research/g.stan/ECC_2013b.pdf
5. http://arxiv.org/abs/1303.3183
6. http://arxiv.org/abs/1303.2987
7. http://www.bg.ic.ac.uk/research/g.stan/#Lecture_Notes

PART II ARTICLES

GREAT: Genome REgulatory and Architecture analysis Tools

Costas Bouyioukos[1,2], Mohamed Elati[1] and François Képès[1,2]

[1] institute of Systems and Synthetic Biology, Genopole®, CNRS, University of Évry, France
[2] BioIntelligence Project, Genopole®, Évry, France

Abstract

Genome expression and layout are expected to be interdependent. Understanding this interdependence is key to whole-genome engineering. Evidence for non-random genome layout, defined as relative positioning of co-functional or co-regulated genes, stems from two main approaches. Firstly, the analysis of contiguous genome segments across species has highlighted the conservation of gene order (synteny) along chromosome regions. Secondly, the study of long-range regularities along chromosomes of one given species has emphasized periodic positioning of microbial genes that are co-regulated, co-expressed, evolutionarily correlated, or highly codon-biased. Tools to detect, visualise, systematically analyse, integrate and exploit gene position regularities along genomes have been developed. Here we report on recent advancements of these tools and discuss novel results.

1 Introduction

The way by which genomes are organised influences fundamental biological processes such as transcription and replication, and through evolution those fundamental biological processes are affecting genome organisation [5]. Ascertaining the interplay between genome organisation and transcription regulation will provide key insights into whole genome expression, transcription control and genome architecture. The latter can contribute into moving towards rational design and whole genome engineering.

2 Detection of regular positioning of co-regulated genes along chromosomes

The non-random arrangement of genes, either regulatory genes such as transcription factors (TFs), or TFs targets, has been observed and discussed for long, however predominantly under an evolutionary and genome synteny approach [13]. Numerous studies during the last decade have indicated that the positioning of various groups of genes demonstrates significant degrees of regularities. Genes that are co-functional, that is, either co-regulated or encoding parts of a functional complex, or members of the same pathway [8], co-expressed [1], evolutionary correlated [15] or highly codon-biased [2] have been found to be periodically positioned along the genomes in all eubacterial phyla. The approaches to detect periodicities are generally characterised by the application of a spectral method, either fast-Fourier transform (FFT), wavelet analysis or an autocorrelation function. However all these methods have a particular handicap when it comes to biological sequences where the spectral signal is weak and poor. Spectral methods cannot take into account the areas of a genome in which data are very sparse and no signal can be detected. Here, for all the reported periodicity analyses, we use a basic algorithm that is able to detect periodic patterns and is taking in to account both regions with strong periodic signal as well as genomic regions -and the length of these regions- with weak or no signal at all [7]. In the current report we consolidate and extend the applicability of this algorithm to detect periodicities, compute its significance, apply it to multiple chromosomes and improve the TF binding site predictions.

2.1 Periodic transcriptional organisation of the E. coli genome

The organisation of transcription within the prokaryotic nucleoid both depends on and determines the structure of the chromosomal fibre [10]. In order to study this relationship we obtained the TF regulatory network of *E. coli* from the Regulon database [14] and clear it from ambiguous TF-target gene predictions accepting only TF-target pairs that were obtained by at least two strong evidences or one strong and two weak evidences as they are classified in the RegulonDB main paper [14]. To detect only long range periodic signal proximal effects of genes need to be removed. We removed any consecutive set of genes where two gene start sites are closer than twice the average gene length in *E. coli*. The positions of the removed genes were replaced by a single coordinate located in the barycentre of the removed set. Next we applied the algorithm to detect periodic patterns of gene positioning in genomic sequences and it is described in [7]. The first step requires the examination for significant periods of a typical periodogram of the target genes of a major *E. coli* TF as it is illustrated in figure 1.

We apply the same analysis systematically to all the major TFs of *E. coli* and we identified (only by using the publicly available transcription start sites of genes into the genome) that there exist a level of significant periodic positioning for almost all the target gene sets of all the major regulators in *E. coli*.

In table 1, 10 out of 12 of *E. coli* TFs with the highest number of targets appear to have their target genes periodically organised along the genome with

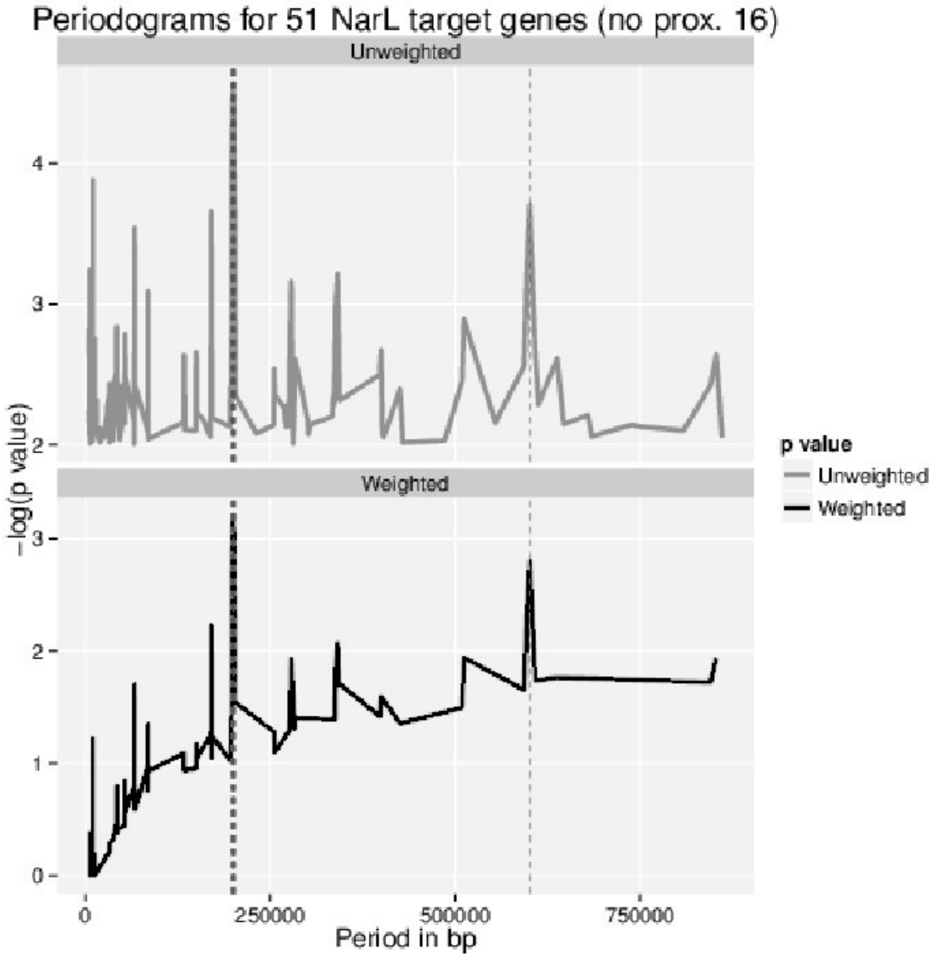


Figure 1: Periodograms of target genes of NarL -a major transcription factor of *E. coli*. The most significant periods are designated by the vertical dashed lines. Both the unweighted and the weighted -i.e corrected for multiple testing- p values are plotted. Two major period areas are observed: one set of periods that lay very close to 200Kbp and a single period at 600Kbp. (The actual periodicity calculations were conducted after the removal of proximal target genes and the replacement with their barycentre, on 16 positions -marked on the title of the graph-. More details on section 2.1.)

the period designated in the 2^{nd} column of the table. The table comprises TF genes that are designated by EcoCyc [12] as global transcription regulators, members of 2-components systems and nucleoid associated proteins.

TF	Period	p value
NarL	200,350	6.18E-4**
ArcA	180,888	1.4E-3**
H-NS	744,444	6.1E-3*
Fur	63,148	5.3E-3*
FNR	161,167	9.8E-3*
CpxR	30,915	1.6E-2*
CRP	16,004	1.3E-2*
Lrp	39,797	2.1E-2*
IHF	282,370	2.6E-2*
Fis	56,309	7.2E-2
LexA	17,115	7.5E-2

Table 1: Detected periods of TF targets genes in *E. coli* sorted by increasing p value (the period with the lowest p value per each TF is reported in the table). NarL is the TF whose targets are arranged in the most regular way. The periods were obtained from the same analysis that generated the periodicity spectra such the one depicting NarL regulating genes on Figure 1.

2.2 Integrated periodicity analysis for multiple TFs in multiple chromosomes.

Then we seek to extend this particular analysis for multiple TFs and in the genomes of organisms with multiple chromosomes. The first candidate was the yeast *Saccharomyces cerevisiae* where a previous study has detected potential periodic patterns of co-regulated genes [9]. Calculated periods for a single TF in a single chromosome were consolidated under the view of multiple TFs or under the view of the whole chromosome. Two novel algorithms were developed and used for this study (unpublished data) to complete the software suite for periodicity analysis. One is calculating periods by using a sliding window along the genome and is able to detect periodic patterns on genomic domains and regions instead of the whole genome. The second is collecting and extending different regions of periodicity and integrates them under the predominant period of a TF and/or under the predominant period of the whole chromosome. At a final step the algorithm checks and extends the result for potential harmonics of the predominant period in order to provide "global" solutions for all the chromosomes. In table 2 we report the results after running this algorithm to the full yeast TF interaction network and calculate the period for each TF with more than 4 targets.

The first reading of our results supports that major yeast regulators, such as the protein RNA polymerase I Enhancer Binding protein Reb1p, (which regulates

Chrom.	TF	Period	Begin	End	p value
2	Reb1p	356,048	84,259	568,634	0.03*
4	Dal82p	989,564	155,980	1,490,112	6E-4**
5	Rpn4p	19,091	82,603	504,582	7E-4**
7	Swi6p	21,463	23,935	1,494,578	2E-4**
8	Abf1p	28,557	48,364	498,751	0.01*
9	Cbf1p	20,929	117,818	292,150	6E-3**
10	Abf1p	236,098	36,584	640,819	0.02*
11	Hap1p	61,190	100,676	575,622	0.04*
12	Reb1p	8,291	35,653	976,964	7E-4**
13	Reb1p	81,496	30,209	880,695	0.03*
14	Abf1p	91,949	48,154	718,329	0.02*
15	Cin5p	142,666	52,942	1,049,509	0.02*
16	Mbp1p	522,429	169,337	888,837	0.03*

Table 2: Consolidated periods from all the yeast TFs organised per chromosome. The table illustrates the longest stretch (Begin – End columns for each particular chromosome), that a specific period, or its harmonics, can span. The longest possible interval of the period with the lowest p value per chromosome was selected. The spanning of each individual period covers almost all the chromosome length for many of the 16 yeast chromosomes.

more than 2500 genes according to the Saccharomyces Genome Database [3]), appear to arrange their targets periodically along a lengthy region of most of the yeast chromosomes. Additionally, another major regulator, the Autonomously replicating sequence binding factor 1 (Abf1p, alias Reb2p), a protein with multiple chromatin binding and gene activation/repression role, also appears to organise its targets periodically along a sorter interval of three yeast chromosomes. The results provide some initial preliminary evidence that global transcription regulators in the yeast genome have their targets regularly arranged.

3 The compromise between global gene position and local regulatory sequences

3.1 Boosting transcription factor binding sites prediction using genes positions

Current methods for the prediction of transcription factor binding sites (TFBS) are marginally successful in their ability to discriminate between many alternative variants of potential TFBSs. While the data on the consensus sequences

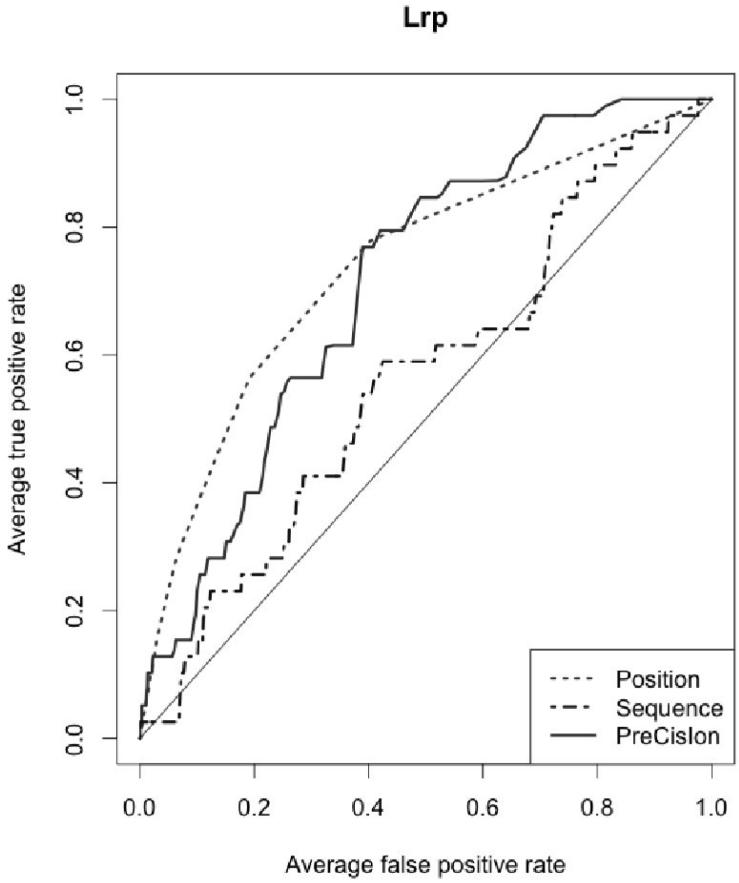


Figure 2: The receiver operating characteristic (ROC) curve -illustrates the relationship between specificity and sensitivity- for the prediction of TFBS of the *E. coli* TF Lrp. Combining the two predictors with the AdaBoost approach (PreCiSion curve) improves the area under curve (AUC).

for the corresponding regulatory sites is available, it often contains motifs with very low sequence conservation (for instance like TCRNNNNNNNACG, where N is R,A or G). This leads to predictions with high false-negative and false-positive rates. The difficulty lies in the specific nature of DNA-protein interactions and the promiscuity of the formation of the DNA-protein complex. Our method PreCisIon [6] address this issue by taking into account a dual view: a) a direct DNA sequence readout and b) a genome layout readout. All methods thus far have solely relied on local DNA sequence information, in addition to that, PreCiSion introduces gene TFBS position information. The underlying rationale is based on the observations that co-regulated genes are positioned at periodic intervals along chromosomes (see section 2.1). The combined classifier is then obtained with an iterative weight update scheme, between position and sequence information, using a modified version of the AdaBoost algorithm. PreCisIon consistently improves predictions from methods based on sequence information only. This is shown, by implementing a cross-validation analysis of the 20 major transcription factors from two phylogeneticaly remote model organisms. For *Bacillus subtilis* and *E. coli*, respectively, PreCisIon achieves on average an AUC (Area Under the ROC Curve) of 70% and 60%, with sensitivity of 80% and 70%, and with specificity of 60% and 56% [6]. A characteristic example is illustrated in figure 2.

3.2 Position sequence interdependence

PreCisIon (presented in section 3.1) operates a boosting algorithm to select the classifier that improves prediction in each iteration of the AdaBoost. As this feature utilises both the sequence score of a TFBS (based on the weight matrix of the site) and the position score (based on the genomic coordinate of the site) allow us to collect these two individual scores and check how they are correlated. The position score and the sequence score are unrelated and measured with two independent processes. However what we find interesting is that there exist cases where the position score and the sequence score are negatively correlated.

Period 149,652 bp	
Sequence ID	Spearman corr. coef.
Seq1	-0.007
Seq2	-0.111
Seq3	-0.132
Seq5	-0.518*
Seq6	-0.398*
Seq8	-0.233
Seq9	-0.153
Seq10	-0.314*

Table 3: Correlations table between position score and sequence score for a set of eight predicted Lrp binding sites and a period of 149652 bp. Negative correlation implies compromise between TFBS sequence quality and its position in the genome (the asterisk * designates significant correlation)

Table 3 summarises the Spearman correlation coefficients between the position score and the sequence score for several iterations of the boosting algorithm in PreCisIon. The observation of the negative correlation between these two scores provides some initial evidence that it is possible that the position of a TFBS in the genome affects the binding of the regulatory protein as well as the quality of the sequence motif. The negative correlation implies that it is possible for a "weak" TFBS sequence to be compensated by a very favourable position in the periodic arrangement and therefore become a real TFBS, or vice-versa. Not withstanding that this intriguing hypothesis of sequence vs. position interplay requires further study and corroboration with more data analysis as well as bench experiments.

4 Conclusions

We present a set of analytical tools and approaches to detect regularities in genomes of prokaryotes (*E. coli*) as well as multichromosome eukaryotes (*S. cerevisiae*). We report on the evidence that connects the periodic organisation of co-regulated genes with the fundamental processes of transcription and gene expression as it was firstly conceived in [9] and further supported by [5]. The area of detecting effective regular arrangements in genome architecture and connect them with transcription, as well as with the dynamic folding of the chromosome, together with the latest advances in experimental biology that are able to measure the contact probabilities of chromatin fibres [4] can reveal a more systematic view about the dynamic arrangement and organisation of the nucleoid and the nucleus. Understanding the epigenetic control of gene expression will pave the way for designing whole genomes in an informed rational way [11].

References

[1] Timothy E. Allen, Nathan D. Price, Andrew R. Joyce, and Bernhard Ø. Palsson. Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Computational Biology*, 2(1):e2, Jan 2006.

[2] Alessandra Carbone, Andrei Zinovyev, and François Képès. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16):2005–2015, Nov 2003.

[3] J Michael Cherry, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman,

Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue):D700–D705, Jan 2012.

[4] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data2 *Nature Reviews. Genetics*, 14(6):390–403, Jun 2013.

[5] Charles J. Dorman. Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nature Reviews. Microbiology*, 11(5):349–355, May 2013.

[6] Mohamed Elati, Rémy Nicolle, Ivan Junier, David Fernández, Rim Fekih, Julio Font, and François Képès. PreCisIon: Prediction of cis-regulatory elements improved by gene's position. *Nucleic Acids Research*, Dec 2012.

[7] Ivan Junier, Joan Hérisson, and François Képès. Periodic pattern detection in sparse boolean sequences. *Algorithms Mol Biol*, 5:31, 2010.

[8] Ivan Junier, Joan Hérisson, and François Képès. Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. *Journal of Molecular Biology*, 419(5):369–386, Jun 2012.

[9] François Képès. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *Journal of Molecular Biology*, 329(5):859–865, Jun 2003.

[10] François Képès. Periodic transcriptional organization of the e.coli genome. *Journal of Molecular Biology*, 340(5):957–964, Jul 2004.

[11] François Képès, Brian C. Jester, Thibaut Lepage, Nafiseh Rafiei, Bianca Rosu, and Ivan Junier. The layout of a bacterial genome. *FEBS Letters*, 586(15):2043–2048, Jul 2012.

[12] Ingrid M. Keseler, Amanda Mackie, Martin Peralta-Gil, Alberto Santos-Zavaleta, Socorro Gama-Castro, Céesar Bonavides-Martínez, Carol Fulcher, Araceli M. Huerta, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Imke Schröder, Alexander G. Shearer, Pallavi Subhraveti, Mike Travers, Deepika Weerasinghe, Verena Weiss, Julio Collado-Vides, Robert P.

Gunsalus, Ian Paulsen, and Peter D. Karp. Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Res*, 41(Database issue):D605–D612, Jan 2013.

[13] Eduardo P C. Rocha. The organization of the bacterial genome. *Annual Review of Genetics*, 42:211–233, 2008.

[14] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S. García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma MartÃnez-Flores, Alejandra Medina-Rivera, Gerardo Salgado-Osorio, Shirley Alquicira-Hernández, Kevin Alquicira-Hernández, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Araceli M. Huerta, César Bonavides-Martínez, Yalbi I. Balderas-Martínez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Verónica Jiménez-Jacinto, Leticia Vega-Alvarado, Victor Del Moral-Chávez, Alfredo Hernández-Alvarez, Enrique Morett, and Julio Collado-Vides. Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(Database issue):D203–D213, Jan 2013.

[15] Matthew A. Wright, Peter Kharchenko, George M. Church, and Daniel Segré. Chromosomal periodicity of evolutionarily conserved gene pairs. *Proceedings of the National Academy of Sciences of the United States of America*, 104(25):10559–10564, Jun 2007.

RetroPath: Retrosynthesis design of metabolic pathways

Pablo Carbonell[1]*, Tamás Fehér[1], Ioana Grigoras[1], Jean-Loup Faulon[1]

[1] iSSB Institute of Systems and Synthetic Biology, University of Évry
CNRS FRE3561, Genopole Campus 1, F-91030 Evry Cedex, France

Abstract

Engineering production cell factories to efficiently synthesize novel therapeutics and biofuels involves the development of strategies for constructively importing pathways into industrial strains. Here, we will learn the basics of RetroPath, a retrosynthesis-based pathway design framework, to enumerate and rank identified candidate pathways leading to the production of the target chemical. The retrosynthesis approach performs a backwards search for biosynthesis routes leading from a target compound to host metabolites through the iterative application of a defined set of biochemical transformation rules. To that end, RetroPath implements a method that codes reactions into molecular signatures, which are atomic subgraphs contained in molecular structures. Individual performances for the list of predicted pathways need to be characterized in order to prioritize the engineering of the most promising routes into the chassis organism. We introduce a ranking function of heterologous biosynthesis pathways based on several factors such as host compatibility, cytotoxicity, enzyme efficiency, and estimated steady-state fluxes. We discuss the practical aspects and challenges of implementing our constructive strategy into hosts organisms.

1 Introduction

Small molecules have become indispensable to our life. They can be used as therapeutics, fuels, building blocks for the chemical industry..., and for many other applications. Their production at the lowest cost and in environment friendly conditions, thus, is one of today's challenge. To address this issue, microbial fermentation is a key technology [6, 8, 10]. Bacteria and yeast were from the very beginning an obvious choice for large-scale production as early compounds of interest were endogenous metabolites of those microorganisms and, with the help of systems biology, the metabolic engineers were able to transform those natural producers into industrial strains. However a large number of value-added compounds are naturally synthetized by higher eukaryotes instead of microorganisms. Producing them through microbial fermentation might seem a straightforward method requiring "only" importing

*Corresponding author: pablo.carbonell@issb.genopole.fr

the pathway from the natural producer organism into the industrial strain. But host engineering is a challenging process requiring time and deep knowledge about the biochemical reactions and enzymes needed at each step and relying on the availability of part characterizations in the literature or in databases. In addition, for each enzymatic step, multiple protein sequences from various organisms may exist and some may be more efficient than others in a given cellular host chassis. This represents a source for pathway improvement, but also leads to a combinatorial explosion of possible pathway variants to be implemented (and tested) into the chosen strain. Moreover, different pathways (i.e. different intermediates and biochemical reactions) may exist for the synthesis of each compound, making the choice even more difficult.

To cope with all these limitations and in order to rationalize the strain engineering process, we developed RetroPath that is an integrated framework for automated pathway design in metabolic engineering projects [3]. In its full deployment version, the system is able to fulfill the engineering cycle going from modeling to design, construction and validation. Such type of system is much sought in the biotechnological sector, as it should accelerate the process of bringing a metabolic engineering project into reality from its conception to the cell factory.

2 RetroPath outline

Here, we will outline the basic steps that are implemented in a retrosynthesis-based pathway design. We start first by picking a target molecule of interest and a chassis organism for which RetroPath will generate the metabolic map containing heterologous biosynthesis routes to the host organism. We will analyze how to enumerate and rank them and, in particular, how to evaluate aspects influencing the efficiency of the pathway. Finally, we discuss practical aspects to be considered for construction of the selected pathways.

Requirements: a computer with internet connection in order to access the RetroPath server and online metabolic databases: KEGG, MetaCyc, BRENDA. Additional modeling tools: CellDesigner, OptFlux, EcoliTox, and Copasi.

3 Genome-scale metabolic models of the chassis organism

In order to start our design, we need a genome-scale model of the host or chassis organism. In biological databases, we can find basically two types of models that are widely-used for metabolic engineering, being both implemented in RetroPath. Such models can be classified depending on the way they were originally built:

1. **From genome-wide annotations**: This first model is mostly based on homology annotations and gap-filling analysis. Its objective is to have a full annotated genome of the organism's metabolism. MetaCyc and KEGG [1] are the main sources of information for these type of organism's models. Retropath uses this type of models in order to search and enumerate all potential pathways producing a desired compound.

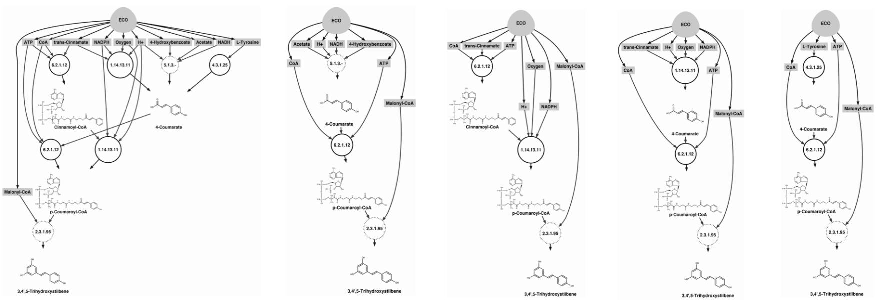
2. **From experimental data**: This second approach searches for models that can reproduce as much as possible experimentally observed phenotypes (measured fluxes). This second model is generally less detailed than the first one. BioModels database is a main source of information for this type of reconstructed models. Retropath uses this model in order to simulate the achievable steady-state fluxes of the engineered organism.

4 Pathway enumeration

In order to produce exogenous compounds, the corresponding metabolic routes containing heterologous enzymes that start from endogenous metabolites must be found. We should note that **not always the shortest pathway is the best solution** in terms of the cost associated with the pathway, as other factors such as enzyme efficiency or toxicity need to be taken into account. To evaluate all possibilities, the problem of performing a complete enumeration of the possible heterologous pathways leading to target production needs to be solved. This problem can be computationally approached by using a well-known technique in metabolic engineering, i.e. by computing the elementary modes present in a metabolic network [2]. By definition, any pathway producing a target compound can be formed by some positive linear combination of the elementary modes in the metabolic network. Here, such metabolic network is formed by the scope of reactions that can link endogenous precursors to the target compound.

In Figure 1, the output of the RetroPath interface is shown for the enumeration of pathways producing resveratrol (KEGG id: C03582), a natural product with reported health benefits that is found in plants, in particular in the skin of red grapes. RetroPath enumerated in total 4 pathways. For each pathway, gene constructs were determined by a machine-learning procedure that predicts performance of the reaction for the set of sequences in the KEGG database [3]. Pathways were ranked based on the scores of the best constructs, which consists in the net sum of the smallest costs in terms of scored gene sequences associated to each enzymatic step.

Map of reachable pathways



Ranked list of best pathways

id	cost	n_genes	putative	C00223	reaction	rcost	best_gene	gcost	C00540	reaction	rcost	best_gene	gcost	C00811	reaction	rcost	best_gene	gcost	C03582	reaction	rcost	best_gene	gcost
22672	2.78	3	2	R01616 sscan6 0	R01616	0.00	rpa:RPA4421	0.99	---	---	0.00	---	0.00	R01308 sscan6 0	R01308	0.00	eck:EC55989 2308	0.79	R01614 sscan6 0	R01614	0.00	bid:Bind 3897	1.00
22674	6.76	3	1	R08815 sscan6 0	R02253	0.00	osa:4338409	4.77	R02255 sscan6 0	R01616	0.00	rpa:RPA4421	0.99	---	---	0.00	---	0.00	R01614 sscan6 0	R01614	0.00	bid:Bind 3897	1.00
22673	6.76	3	1	R01616 sscan6 0	R01616	0.00	rpa:RPA4421	0.99	---	---	0.00	---	0.00	R02253 sscan6 0	R02253	0.00	osa:4338409	4.77	R01614 sscan6 0	R01614	0.00	bid:Bind 3897	1.00
22675	7.19	3	1	R01616 sscan6 0	R01616	0.00	rpa:RPA4421	0.99	---	---	0.00	---	0.00	R00737 sscan6 0	R00737	0.00	rsp:RSP 3574	5.20	R01614 sscan6 0	R01614	0.00	bid:Bind 3897	1.00

Figure 1: Enumerated pathways predicted by RetroPath for producing resveratrol in *E. coli*. Ranking is shown at the bottom and pathways are represented from left to right in ranking order.

5 Pathway steady-state simulation

After selection of the desired gene construct corresponding to a pathway, RetroPath provides the possibility of exporting the pathway into SBML format. Many software tools are available that accept SBML format for simulation. In particular, we are interested in inserting the pathway into a genome-scale metabolic model so that we can simulate the steady-state flux distribution of the engineered strain through flux balance analysis (FBA). To that end, we have to import the additional metabolites and reactions contained in the pathway into an *E. coli* model.

Here, we perform flux balance analysis simulations by using OptFlux [11]. Once the model of the engineered strain is loaded into the program, we select the objective function of the cell that we want to maximize. In our approach, we compare two types of competing objectives: a) **growth**, expressed as a linear combination of fluxes, which is the objective that naturally is found in a wild-type strain; b) **compound production**, which is our biotechnology objective. In most cases, we cannot expect the engineered cell to naturally optimize production as most likely it does not provide a selective advantage to the strain. Therefore, we consider that the first objective of growth is *the natural objective*, while the second is *the synthetic objective* that we want to implement. In a practical situation, the actual observed metabolic phenotype for our engineered strain would lie in between both objectives (see Figure 2). Several metabolic interventions are possible to shift the state of the cell in order to favor the biotechnology-oriented second objective of target production [7].

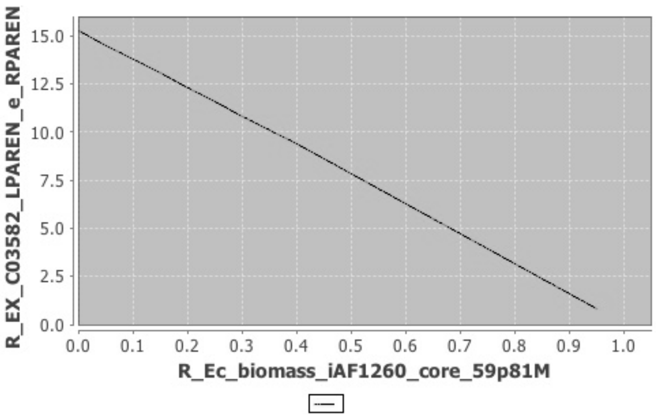


Figure 2: Flux variability analysis from OptFlux showing the tradeoff frontier lying between target production and biomass objectives.

6 Pathway toxicity

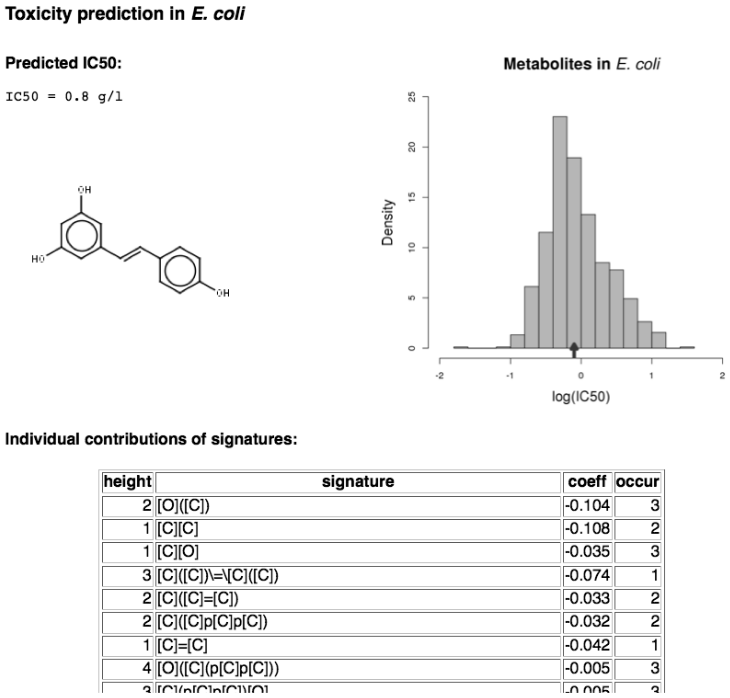
Another aspect to be considered on pathway design is toxicity of intermediates and by-products that might be accumulated in the cell. Our EcoliTox server [9] addresses this issue by providing an estimate of metabolites' IC50 (half maximal inhibitory concentration, i.e. minimum concentration of that specific metabolite that will inhibit growth by half). We can consider as a first approximate that pathway toxicity is the average toxicity of metabolites in the pathway.

7 Pathway ranking and selection

An overall score of the pathway may finally be obtained by combination of construct, flux and toxicity scores. The way each term is weighted in the global score is nontrivial. A first approach has been determined by considering the optimal case as the one that ranks at the top those scores that correspond to natural pathways [3]. In that way, we prioritize pathways that have been selected and optimized during natural evolution.

8 Pathway implementation

As described above, RetroPath provides a top-list of enzymes available for each step of the highest-ranking metabolic pathway. The genes of these enzymes can be PCR-amplified from the genome or cDNA of the natural host, or can be synthesized by a commercial service. The possibility of codon opti-



Toxicity prediction in *E. coli*

Predicted IC50:

```
IC50 = 0.8 g/l
```

Individual contributions of signatures:

height	signature	coeff	occur
2	[O]([C])	-0.104	3
1	[C][C]	-0.108	2
1	[C][O]	-0.035	3
3	[C]([C])\=\[C]([C])	-0.074	1
2	[C]([C]=[C])	-0.033	2
2	[C]([C]p[C]p[C])	-0.032	2
1	[C]=[C]	-0.042	1
4	[O]([C](p[C]p[C]))	-0.005	3
3	[C](p[C]p[C])[O]	-0.005	3

Figure 3: Metabolite toxicity prediction from the EcoliTox web server.

mization and the avoidance of restriction sites favor the latter choice, despite its currently higher price. The enzyme scores of the RetroPath output can be thought of as probabilities for a certain type of activity; therefore testing multiple enzyme candidates for each enzymatic step is highly recommended. Physically, this means building the chosen pathway using multiple gene combinations (referred to as constructs), which requires the use of rapid and efficient plasmid assembly techniques. We divide these tools into three large categories (for a more detailed comparison of assembly techniques, see current reviews written on this topic [4, 5]):

- The first group uses **restriction endonucleases and ligation** to stitch appropriate genes together. These classical methods have been much developed in the past decades to overcome the problems of internal restriction sites, repetitive use of restriction enzymes or sequence scars remaining in the joints. The most promising member of this category is GoldenGate cloning, which allows the high efficiency fusion of even ten DNA segments in a simple reaction.

- The second group of tools applies *in vitro* **enzymatic reactions on overlapping DNA fragments** for their assembly. These can be based on DNA-polymerase extension of the overlaps such as Circular Polymerase Extension Cloning, or can use the "chew-back and anneal" strategy, as in the case of Ligation Independent Cloning, or Gibson assembly.

- The third pool of construction techniques also utilizes overlapping DNA fragments, but uses **homologous recombination** to fuse them. The Red recombinases of phage Lambda, or the endogenous recombination machinery of yeast cells have both been applied successfully for this purpose.

Once all constructs are assembled, they are transformed into bacterial cells, induced, and production of the target compound, as well as some pathway intermediates are quantified. The most efficient constructs can go through further optimization to increase production titers with the help of FBA. Potential bottlenecks can be identified in the pathway, which can be eliminated by increasing the expression of the respective enzyme. Gene expression is usually modified by re-assembling the construct using altered regulatory sequences (e.g. promoters or ribosome binding sites), which further emphasizes the importance of using rapid and efficient plasmid construction techniques.

9 Conclusions

As metabolic engineering and synthetic biology progress into well-established biotechnology disciplines, more precise and rational protocols for designing and constructing metabolic pathways are required by the community. Here, we provided an introductory glimpse into RetroPath, a retrosynthesis-based approach that integrates state-of-the-art modeling techniques in order to streamline the otherwise challenging process of pathway design and construction. Our goal here was to showcase the promising capabilities for rational pathway design that are present in the proposed computational framework. As such, we believe that our retrosynthesis-based approach will progressively become in next years a first-stop reference for implementing advanced and innovative metabolic engineering and synthetic biology projects.

10 Acknowledgments

abSYNTH facilities; Genopole ATIGE; ANR Chair of Excellence; PRES UniverSud Paris; UPFellows; Marie Curie COFUND (EU FP7).

References

[1] Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**:112+.

[2] Carbonell, P., D. Fichera, S. Pandit and J.L. Faulon, (2012). Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst. Biol.*, **6**: 10+.

[3] Carbonell, P., A.G. Planson, D. Fichera and J.L Faulon, (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.* **5**: 122+.

[4] Cobb, R.E., Ning, J.C., Zhao, H., (2013). DNA assembly techniques for next-generation combinatorial biosynthesis of natural products. *J. Ind. Microbiol. Biotechnol.*

[5] Ellis, T., Adie, T., Baldwin, G.S., (2011). DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol.*, **3**: 109–18.

[6] Lee, W.J., Na, D., Park, J.M., Lee, J., Choi, D., Lee, S.Y.. (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.*, **8**: 536–546.

[7] Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.*, **14**: 2367–2376.

[8] Planson, A.G., P. Carbonell, I. Grigoras and J.L. Faulon, (2012). A retrosynthetic biology approach to therapeutics: from conception to delivery. *Curr. Opin. Biotech.*, **23**: 948–956.

[9] Planson, A.G., Carbonell, P., Paillard, E., Pollet, N., and Faulon, J.L. (2012). Compound toxicity screening and structure-activity relationship modeling in Escherichia coli. *Biotechnol. Bioeng.*, **109**: 846–850.

[10] Rabinovitch-Deere, C.A., Oliver, J.W.K., Rodriguez, G.M., Atsumi, S. (2013) Synthetic biology and metabolic engineering approaches To produce biofuels. *ACS Synth. Biol.*, **113**: 4611–4632.

[11] Rocha, I., Maia, P., Evangelista, P., Vilaca, P., Soares, S., Pinto, J., Nielsen, J., Patil, K., Ferreira, E., and Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.*, **4**: 45+.

Integrating time-series data on large-scale cell-based models: application to skin differentiation

Louis Fippo Fitime[1]*, Andrea Beica[1], Olivier Roux[1]
and Carito Guziolowski[1]

[1]LUNAM Université, École Centrale de Nantes, IRCCyN UMR CNRS 6597
(Institut de Recherche en Communications et Cybernétique de Nantes)
1 rue de la Noë – B.P. 92101 – 44321 Nantes Cedex 3, France.

Abstract

The way living organisms work and develop themselves is controlled by large and complex networks of genes, proteins, small molecules, and their interactions, called biological regulatory networks. Confronting time-series gene expression data with models may allow us to examine and characterize the dynamics of elements that compose such regulatory networks. In this work, we propose a way to model and simulate large-scale regulatory networks, by using the Process Hitting (PH) framework, in order to verify if the model can predict the experimental measures. The preliminary work presented here proposes: (1) a semi-automatic method to build a PH from a regulatory network of biochemical reactions, (2) a discretization scheme of the continuous time-series measurements, and (3) an approach to estimate the PH stochastic simulation parameters in an unbiased manner.

1 Introduction

The comprehension of the mechanisms involved in the regulation of a living cell is a fundamental issue. These mechanisms can be modeled as biological regulatory networks, which analysis requires to preliminary build a mathematical or computational model. By just considering qualitative regulatory effects between components, biological regulatory networks depict fairly well biological systems, and can be built upon public repositories such as the Pathways Interaction Database [8], and hiPathDB[10] for human regulatory knowledge.

This work aims to propose a dynamical model of large-scale systems based on the formal integration (complete validation/invalidation) of high-throughput experimental time-series data. So far this idea has been addressed separately by approaches that either: (a) focus first on modeling at small-scale the system and then on refining or improving it through the fitting with some data points, such as methods based on differential equations [9, 1, 6], or (b) integrate in an efficient and complete fashion large-scale models and high-throughput

*Corresponding author: `Louis.Fippo-Fitime@irccyn.ec-nantes.fr`

data regardless of the system dynamics [3, 5], or (c) fit dynamical data to middle-scale networks using stochastic approaches, and therefore without the guarantee of finding global optima [4]. Therefore, with this work we intend to fill the gaps between the previously cited methodologies and converge to a more realistic model of biological behavior.

For modeling and analyzing the biological system we rely on the Process Hitting (PH) framework[7], since it is especially useful for studying systems composed of biochemical interactions, and provides stochastic simulation as well as efficient static methods to model dynamical properties of the system. The PH framework uses qualitative and discrete information of the system, without requiring enormous parameter estimation tasks for its stochastic simulation. So far, this method has been successfully demonstrated only on very well-known systems and without exploiting high-throughput measures. We believe, however, that the use of high-throughput data has become unavoidable with the advent of massive, publicly available data sets in the form of well-standardized DNA microarray data and, more recently, in the form of phospho-proteomics data.

The main methodological and preliminar results of this work are: (i) semi-automatic PH generation from a biological system composed of biochemical reactions, extracted from public databases; (ii) discretization approach of time-series expression data, so we can reproduce these traces by using in a first attempt the PH stochastic simulation, and afterwards perform static reachability analyses to satisfy these data; and (iii) estimation of the the temporal and stochastic parameters of the simulation, based on statistical analyses of the full-compendium of time-series expression data. The biological system used as a case-study for this work is a cell-based model of skin differentiation, which is of key importance in wound healing.

2 Methods and data

2.1 The Process Hitting Framework

Process Hitting (PH) gathers a finite number of concurrent processes grouped into a finite set of sorts. A sort stands for a component of a biological system while a process, which belongs to a unique sort, corresponds to a unique state of the system components(sorts). At any time, exactly one process of each sort is present. A state of the PH corresponds to such a set of processes. We denote here a process by a_i where a is the sort and i is the process identifier within the sort a. The concurrent interactions between processes are defined by a set of *actions*. Actions describe the replacement of a process by another of the same sort conditioned by the presence of at most one other process in the current state. An action is denoted by $a_i \to b_j \upharpoonright b_k$, which is read as "$a_i$ *hits* b_j to make it bounce to b_k", where a_i, b_j, b_k are processes of sorts a and b, called respectively *hitter*, *target* and *bounce* of the action.

Definition 1 (Process Hitting) *A* Process Hitting *is a triple* $(\Sigma, L, \mathcal{H})$*, where:*

- $\Sigma = \{a, b, \dots\}$ *is the finite set of* sorts*;*
- $L = \prod_{a \in \Sigma} L_a$ *is the set of states with* $L_a = \{a_0, \dots, a_{l_a}\}$ *the finite set of* processes *of sort* $a \in \Sigma$ *and* l_a *a positive integer, with* $a \neq b \Rightarrow L_a \cap L_b = \emptyset$*;*
- $\mathcal{H} = \{a_i \to b_j \upharpoonright b_k \in L_a \times L_b \times L_b \mid (a, b) \in \Sigma^2 \wedge b_j \neq b_k \wedge a = b \Rightarrow a_i = b_j\}$ *is the finite set of* actions.

Given a state $s \in L$, the process of sort $a \in \Sigma$ present in s is denoted by $s[a]$. An action $h = a_i \to b_j \upharpoonright b_k \in \mathcal{H}$ is *playable* in $s \in L$ if and only if $s[a] = a_i$ and $s[b] = b_j$. In such a case, $(s \cdot h)$ stands for the state resulting from the play of the action h in s, with $(s \cdot h)[b] = b_k$ and $\forall c \in \Sigma, c \neq b, (s \cdot h)[c] = s[c]$.

Modeling cooperation. As described in [7], the cooperation between processes to make another process bounce can be expressed in PH by building a *cooperative sort*. Fig. 1 shows an example of a cooperative sort ab between sorts a and b, defined with 4 processes (one for each sub-state of the presence of processes a_1 and b_1). For the sake of clarity, processes of ab are indexed using the sub-state they represent. Hence, ab_{01} represents the sub-state $\langle a_0, b_1 \rangle$, and so on. Each process of sort a and b hit ab, which makes it bounce to the process reflecting the status of the sorts a and b (e.g., $a_1 \to ab_{00} \upharpoonright ab_{10}$ and $a_1 \to ab_{01} \upharpoonright ab_{11}$). Then, to represent the cooperation between processes a_1 and b_1, the process ab_{11} hits c_1 to make it bounce to c_2 instead of independent hits from a_1 and b_1. The same cooperative sort is used to make a_0 and b_0 cooperate to hit c_1 and make it bounce to c_0. Cooperation can be used to model protein-complex biochemical reaction. For instance a molecule a that cooperates with a molecule b to activate a molecule c, Fig. 1 (left), We model this interaction by four sorts Fig. 1 (right) a, b, c and ab. Sorts a, b and c represent components a, b and c. We introduce the cooperative sort ab to characterize constraints on components a and b. Cooperation can be a way to model protein-complex formation.

Example 1 *Fig. 1 represents a PH* $(\Sigma, L, \mathcal{H})$ *with* $\Sigma = \{a, b, c, ab\}$*, and:*

$$\begin{aligned} L_a &= \{a_0, a_1\}, & L_b &= \{b_0, b_1\}, \\ L_{ab} &= \{ab_{00}, ab_{01}, ab_{10}, ab_{11}\}, & L_c &= \{c_0, c_1, c_2\}. \end{aligned}$$

This example models a Biological Regulatory Network (BRN) where the component c *has three qualitative levels, components* a *and* b *are Boolean and* ab *is a cooperative sort. In this BRN,* ab *inhibits* c *at level* 2 *through the cooperative sort* ab *(e.g.*$ab_{00} \rightarrow c_2 \Rsh c_1$*,* $ab_{00} \rightarrow c_1 \Rsh c_0$*) while* a *and* b *activate* c *through the cooperative sort* ab *(e.g.* $ab_{11} \rightarrow c_0 \Rsh c_1$ $ab_{11} \rightarrow c_1 \Rsh c_2$ *). Indeed, the reachability of* c_2 *and* c_0 *is conditioned by a cooperation of* a *and* b*, as explained above.*

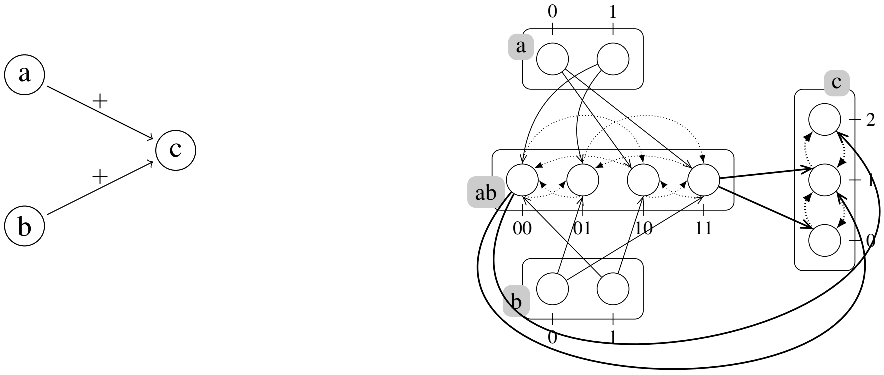


Figure 1: (left) Biological pattern example. Nodes are components and edges are interactions For instance, components a and b cooperate to activate c. (right) equivalent PH model. A PH example with four sorts: three components (a, b and c) and a cooperative sort (ab). Actions targeting processes of c are in thick lines.

2.2 Time-series microarray data

To illustrate our approach, we used the time series microarray data from calcium stimulated keratinocyte cells measured at 10 time-points. 200 transcripts were selected for their dynamic patterns, that is, their fold expression with respect to the non-stimulated cell was significant in at least one time point. We included in our model a subset of 12 of them: MKP3, MKP1, UPAR, HES5, ILB1, A20, SM22, IL8, ET1, TNF-a, TFR, DKK1. This subset was selected because we were able to automatically retrieve the regulatory mechanisms upstream of these 12 genes from public repositories of biochemical reactions. The full dataset (data not shown) was produced by the German Cancer Research Center (DKFZ) and is currently in the process of getting published.

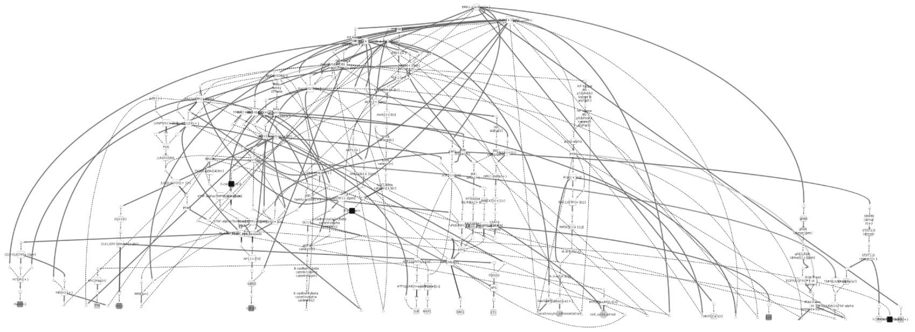


Figure 2: **RSTC network**

2.3 Interaction network

The interactions of the studied biological system were represented in a RSTC network, which stands for multi-layer receptor-signaling-transcription-cell state network, generated from the Pathway Interaction Database (PID). In order to build this network, we selected a set of seed nodes related to the biological process studied. The seed nodes for our case study were: (1) E-cadherin, which is a protein having Ca binding domains and which plays an important role in cell adhesion; (2) the 12 significantly differentially expressed genes accross the 10 time-points; and (3) the cell states of keratinocytes-differentiation and cell-cycle-arrest. The network was extracted automatically from the whole content of the NCI-PID database by using a subgraph algorithm to link the seed nodes[2]. Fig.2 shows the RSTC network obtained.

3 Results

3.1 Modeling the RSTC network as a PH model

In order to model the RSTC network with a PH model we selected known biological regulatory patterns (atomic set of biological components and their interacting roles), represented as biochemical reactions in the RSTC network, and proposed their PH representation. For instance a molecule a that cooperates with a molecule b to activate a molecule c, Fig. 1 (left), is a regulatory pattern because it is a protein-complex biochemical reaction that appears recurrent times. We model this pattern by four sorts Fig. 1 (right) a, b, c and ab. Sorts a, b and c represent components a, b and c. We introduce the cooperative sort ab to characterize constraints on components a and b. In our RSTC network, we found 11 regulatory patterns (see Appendix 4).

3.2 Integrating time-series gene expression data

3.2.1 Discretizing times-series data

Because PH simulation is discrete we need to discretize continuous experimental data, so we can compare our simulation outputs. The goal of this method was to better determine, according to the gene expression level, when a given molecule is activated or inhibited. To do this, we introduced the new analog concept of Significant Increase or Decrease to characterize the fact that a level of a molecule increases or decreases when crossing a threshold of significance; We limited the possible expression levels for a molecule to $\{0, 1, 2\}$. Algorithm 1 underlines the main steps of the proposed discretization method. For more details about the functions used in this algorithm see Appendix 4

Algorithm 1 Discretization of experimental data

```
Require: $X$ a table of experimental data
Ensure: $Y$ a table of discretized data
  for all gene $i$ in $X$ do
    $threshold \leftarrow computeThreshold(X[i,]);$
    $Y[i,0] \leftarrow initialState(threshold, X[i,]);$
    for all $j$ in $numberExpression$ do
      if $Increase(X[i,j], X[i,j+1])$ then
        $computeSignificativityOfIncrease(threshold, X[i,j], X[i,j+1]);$
        $fixSTATE(Y[i,j], Y[i,j+1]);$
      else
        $computeSignificativityOfDecrease(threshold, X[i,j], X[i,j+1]);$
        $fixSTATE(Y[i,j], Y[i,j+1]);$
      end if
    end for
  end for
```

To illustrate the result of the discretization algorithm 1 we plot in Fig. 3 the expression of the TFRC and IL8 genes from the times-series data with their respective discrete plots. On the discrete plot, one can clearly differentiate when a molecule is active or not, which is of extreme importance when modeling these steps in the PH framework since we want to have coherent simulation results.

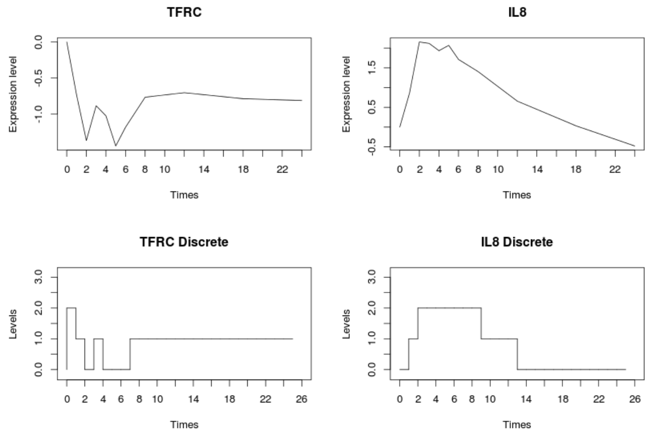


Figure 3: Illustration of discretisation of Experiment Data

3.2.2 Estimating the parameters for the PH-simulation

The simulation of the execution of the PH actions is done stochastically. Therefore, we need to relate each action with temporal and stochastic parameters, introduced into the PH framework to achieve dynamic refinement [7]. This is an important aspect of the modeling when taking into account the temporal and stochastic dimensions of biological reactions by performing simulations. On the one hand, we consider the probability of a reaction to occur, and on the other hand, we consider stochastic parameters in the aim at observing an expected behavior. In the PH framework, to play an action we need two essential parameters: the rate r or the temporal parameter because $t = r^{-1}$ and the stochasticity absorption sa. These two parameters will be estimated according to the expression profile of time series data of the experiment. To avoid overfitting in the estimation of these parameters, we propose that each component of the PH, representing a measured gene in the network, will take the estimated values of the parameters of its respective cluster in the experimental data.

1. The first step is to cluster the data set. The goal of the clustering process is to partition the genes into groups such that the profiles contained in the same group (cluster) are similar to each other and as different as possible of the profiles assigned to the other clusters. The particularity here is to choose the best clustering criteria.

2. For each cluster obtained in the previous step, estimate the value of r and sa associated to the cluster.

3. For each component of the PH model associated to the measured gene, determine its cluster, and assign it the previously estimated parameters, r and sa.

In our time-series data, the components of the PH which need to be associated specific parameters (step 3) are the 12 genes present in our RSTC network.

3.3 PH code generation

To simulate of the model, we generated a PINT code to be simulated by the PINT simulator[†]. For the PINT code generation we first list all the selected patterns in the biological reaction into a file. In this file, each line contains the name of the nodes belonging to the current reaction and the reaction type number. The list was then parsed, line by line and, after renaming the nodes using numbers (for readability and in conformity with the PINT language syntax) the corresponding PINT code for the PH process equivalent to each reaction was generated. This was implemented in the Java programming language.

4 Conclusions

This work describes the preliminary steps towards the integration of time-series data in large-scale cell-based models. We proposed a semi-automatic method to build a PH from a biological system composed of biochemical reactions, extracted automatically from public databases, relevant to keratinocyte stimulation induced by calcium. We then proposed a method to discretize time-series gene expression data, so they can be confronted to the PH simulations and logically explained by the PH static analysis. Finally we described a method to automatically estimate the temporal and stochastic parameters for the PH simulation, so this estimation process will not be biased by overfitting. As concrete perspectives of this work, we intend to *(i)* validate the RSTC network topology by confronting its *in-silico* simulation with real measurements of its components; *(ii)* compare the stochastic simulation results with reachability static analysis over the same PH components mapped to the 12 measured genes; and finally *(iii)* search for key-regulators up-stream the 12 genes which will control the dynamics of the system, to provide our biologist partners concrete hypotheses to test experimentally.

[†] Available at `http://process.hitting.free.fr`

References

[1] Grégory Batt, Delphine Ropers, Hidde De Jong, Johannes Geiselmann, Radu Mateescu, Michel Page, and Dominique Schneider. Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in escherichia coli. *Bioinformatics*, 21(suppl 1):i19–i28, 2005.

[2] Carito Guziolowski, Aristotelis Kittas, Florian Dittmann, and Niels Grabe. Automatic generation of causal networks linking growth factor stimuli to functional cell state changes. *FEBS Journal*, 279(18):3462–3474, 2012.

[3] Carito Guziolowski, Santiago Videla, Federica Eduati, Sven Thiele, Thomas Cokelaer, Anne Siegel, and Julio Saez-Rodriguez. Exhaustively characterizing feasible logic models of a signaling network using answer set programming. *Bioinformatics*, 29(18):2320–2326, 2013.

[4] Aidan MacNamara, Camille Terfve, David Henriques, Beatriz Peñalver Bernabé, and Julio Saez-Rodriguez. State–time spectrum of signal transduction logic models. *Physical Biology*, 9(4):045003, 2012.

[5] Alexander Mitsos, Ioannis N Melas, Paraskeuas Siminelakis, Aikaterini D Chairakaki, Julio Saez-Rodriguez, and Leonidas G Alexopoulos. Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS computational biology*, 5(12):e1000591, 2009.

[6] Mohammad Mobashir, Burkhart Schraven, and Tilo Beyer. Simulated evolution of signal transduction networks. *PloS one*, 7(12):e50905, 2012.

[7] Loïc Paulevé, Morgan Magnin, and Olivier Roux. Refining dynamics of gene regulatory networks in a stochastic π-calculus framework. In *Transactions on Computational Systems Biology XIII*, pages 171–191. Springer, 2011.

[8] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[9] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

[10] Namhee Yu, Jihae Seo, Kyoohyoung Rho, Yeongjun Jang, Jinah Park, Wan Kyu Kim, and Sanghyuk Lee. hipathdb: a human-integrated pathway database with facile visualization. *Nucleic acids research*, 40(D1):D797–D802, 2012.

Appendix A

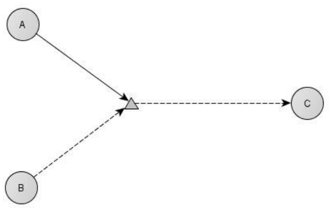
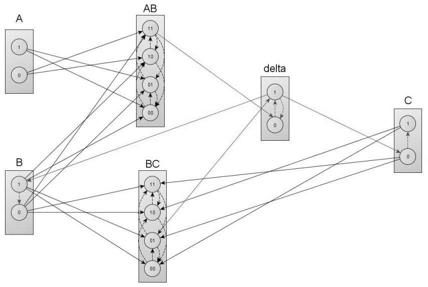
Biological Pattern no 1	PH Model
	 A AB delta C B BC 1 0 11 10 01 00

Figure 4: (left) Biological pattern: Molecules A and B cooperate to activate molecule C. After the activation of C, A remains active and B is desactivated. (right) equivalent PH model. AB and BC are regular sorts, while the sort delta models the reaction beginning or end.

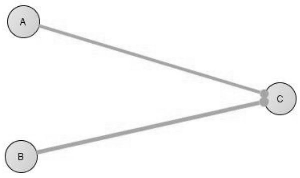
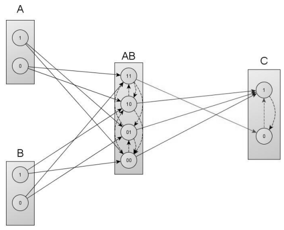
Biological Pattern no 2	PH Model
	

Figure 5: (left) Biological pattern: A and B cooperate to activate C. Both A and B remain active after end of reaction (right) equivalent PH model

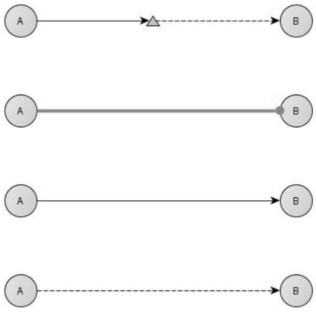
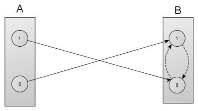
Biological Pattern no 3	PH Model
	

Figure 6: (left) Biological pattern: different types of activation. (right) equivalent PH model

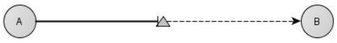
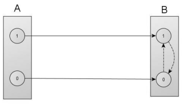
Biological Pattern no 4	PH Model
	

Figure 7: (left) Biological pattern of an inhibition reaction: the inhibitor presence leads to the desactivation of its target, while its absence leads to the activation of the target (right) equivalent PH model

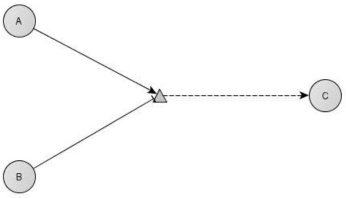
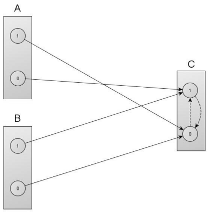
Biological Pattern no 5	PH Model
	

Figure 8: (left) Biological pattern. Molecule C is either activated by A, or inhibited by B; (right) equivalent PH model where A and B are not cooperating to modify C, each one has independent, opposite action on C.

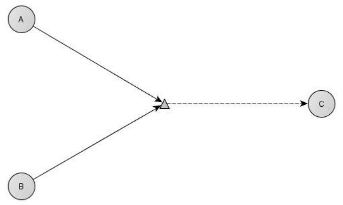
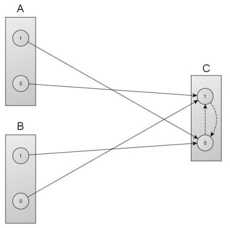
Biological Pattern no 6	PH Model
	

Figure 9: (left) Molecule C is activated by either A, or B, independantly one from other. (right) equivalent PH model

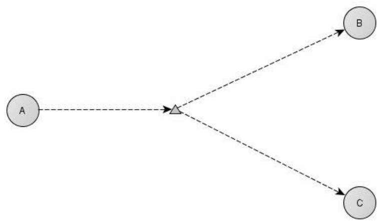
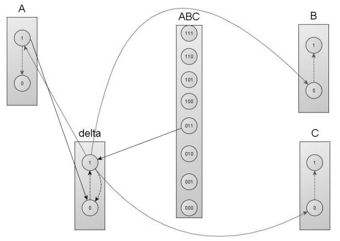
Biological Pattern no 7	PH Model
	 A 1 0 ABC 111 110 101 100 011 010 001 000 delta 1 0 B 1 0 C 1 0

Figure 10: (left) Complex A decomposes in components B and C. At the end of the reaction, A no longer exists/ is no longer active. (right) equivalent PH model. ABC is a regular cooperative sort and delta models the reaction, as explained in Pattern 1. For clarity purposes, the hits from A, B and C to the cooperative sort ABC have not been drawn.

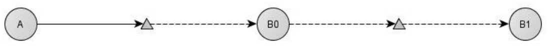
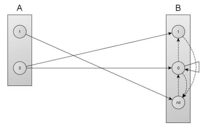
Biological Pattern no 8	PH Model
	

Figure 11: (left) $B0$ and $B1$ represent the same biological entity. (right) equivalent PH model, $B0$ and $B1$ are different process of the same sort; A create B, which then activates itself.

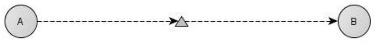
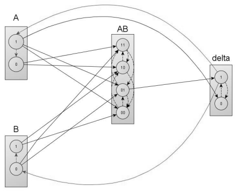
Biological Pattern no 9	PH Model
	 A 1 0 AB 11 10 01 00 delta 1 0 B 1 0

Figure 12: (left) A modification reaction: A activate B, then dissapears; The reaction begins when A is present, and ends when A has replaced by B. (right) equivalent PH model, AB is a cooperative sort and the delta sort models the reaction.

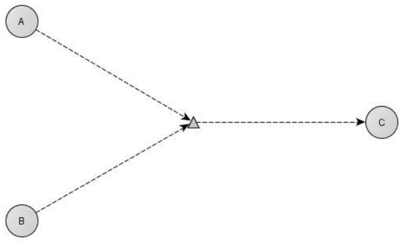
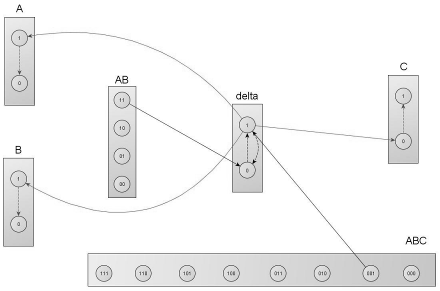
Biological Pattern no 10	PH Model
	 A 1 0 AB 11 10 01 00 delta 1 0 C 1 0 B 1 0 ABC 111 110 101 100 011 010 001 000

Figure 13: (left) A composite modification: A and B cooperate to create C, then disappear. (right) equivalent PH model. For clarity purposes, hits to cooperative sorts have not been drawn.

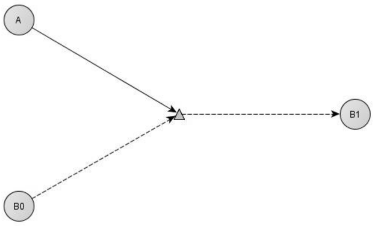
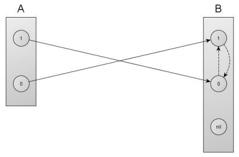
Biological Pattern no 11	PH Model
	

Figure 14: (left) Activation of non-binary sort: similar to Pattern 1, except for the non-binarity of the target source. $B0$ and $B1$ represent the same entity. Unlike pattern 8 (the other pattern dealing with non-binary sorts), entity B is already present, via the condition on $B0$, it just needs to be activates. (right) equivalent PH model.

Appendix B

Functions	Specifications
computeThreshold(X)	compute the threshold of the profile of expression represent by X
initialState(X)	fixe the initial state of the expression represent by X according to the initial value of X and the threshold
Increase(X,Y)	Test if the measure increases between the two times points X and Y
computeSignificativityOfIncrease(s,X1,X2)	compute the significance of the increase according to the threshold and X1 and X2
computeSignificativityOfDecrease(s,X1,X2)	compute the significance of the decrease according to the threshold and X1 and X2
fixSTATE(X1, X2)	fix the current state

Figure 15: Functions(first column) and Specifications(second column)

Appendix C

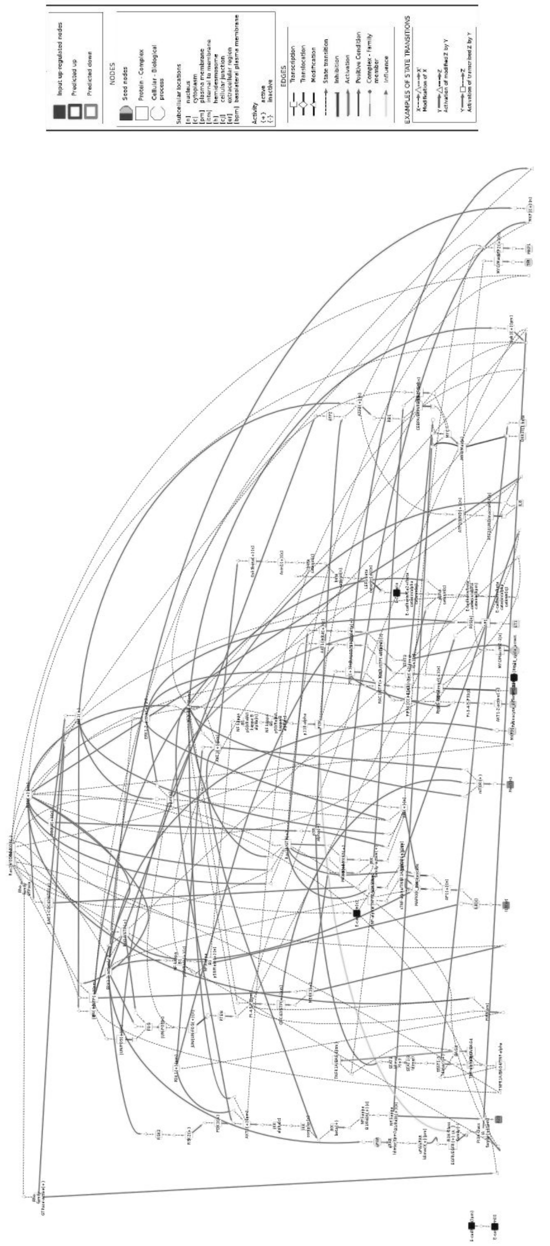


Figure 16: RSTC Network

Genetic networks modeling inspired from electronic design automation

Yves Gendrault[1], Morgan Madec[1], Christophe Lallement[1], Jacques Haiech[2]

[1] Laboratoire des Sciences de l'Ingénieur, de l'Informatique et de l'Imagerie (ICube), UMR 7357, Equipe SMH, 300 Boulevard Sébastien Brandt, F-67412 Illkirch Cedex 02.

[2] Laboratoire d'Innovation Thérapeutique (LIT), UMR 7200, 74 route du Rhin - F-67400 Illkirch.

Abstract

Synthetic biology is a new science, at the interface between biotechnology and engineering sciences. It tends to create new organisms by a rational combination of standardized biological components that are decoupled from their natural environment. The work presented in this paper focuses on aspects of in-silico design of these artificial biosystems and is a summary of the work done by our team. Biosystem design assistance has been identified as necessary for the development of the complexity of these systems but is rarely addressed in a general context by the scientific community. Unlike existing tools for synthetic biology, our work is based on the design methodology used in microelectronics and consists in adapting the tools of EDA (Electronic design automation). Their adaptation is based on three key elements: the structuring of the design through a design flow, reuse and standardization mechanisms through the use of biobricks and finally fast and accurate models understandable by electronic tools.

Design flow for synthetic biology

A design flow is a methodology for structuring stages of system design to ensure the effectiveness and reliability of the system achieved while reducing the time and cost of designing. The design flow used in the design of microelectronic circuits is a well-tried method, in particular the steps dedicated to digital parts. During 40 years, thanks to the combined evolution of technologies and design tools, processors have gone from just over 2,000 transistors integrated in 1970 to over 2 billion in 2010 [1].

Synthetic biology can be seen today at the same stage of development as microelectronics was in the 60's. Nevertheless, the outlook for this area is at least as promising [2] as the one experienced by microelectronics in recent years. Although the complexity of artificial biosystems developed today is modest (in terms of features), the development of CAD (Computer-aided design) tools is required straight away to follow the technological developments in this field. To carry out this task, we can find several tools, such as BioJADE, GenoCAD, SysBioSS and Tinkercell [3,4], which seem to be effective in some cases, but correspond only to some stages of a complete CAD tool. The assembly of these tools seems complex and several elements are missing in order to have an automated software suite. The original approach developed in this paper is to reuse existing tools from microelectronics, and to adapt them to biological material, rather than recreating the entire design environment.

The proposed design flow developed in [4] includes the following steps. Starting from the specifications of a biosystem, the first step is to achieve high-level system description. Then, a functional synthesis is made from this description. It provides a netlist of elementary functions of the system. During the next step, a search is made in a biobricks database (like the Registery of Standard Biological Parts, http://parts.igem.org) to transform this netlist into a concrete virtual biobricks assembly. Finally, the last steps intend to validate the biosystem through multi-abstractions simulations, thus requiring the development of associated models and post-design analysis.

These steps are based on the existence of a design library, called "design kit", which is effective and complete, including the biological functions used, their characteristics, the logical equivalent and the corresponding biobricks and models in several levels of abstraction. Models of biological mechanisms have been especially developed in two main levels of abstraction, high-level and low-level.

High-level modeling for design automation purpose

It has been shown [5] that most of the biological processes can be modeled as electronic functions and in particular as logic gates thanks to a digital abstraction (illustrated Figure 1.A). This property is very important because it can allow the reuse and adaptation of digital design tools (such as logic synthesizers, compilers, etc.) [6].

Two open-source tools used to perform the step of functional synthesis in the microelectronic design flow (composed of a register transfer level synthesis step and a step of mapping abstract functions and optimization) have been adapted to the constraints of biological material. We tested this step on two examples [6]: a state machine, which is an abstract machine that can switch between several defined states (the state machine is used to regulate a protein concentration) and a biological microprocessor that demonstrates the effectiveness of our approach on very complex systems.

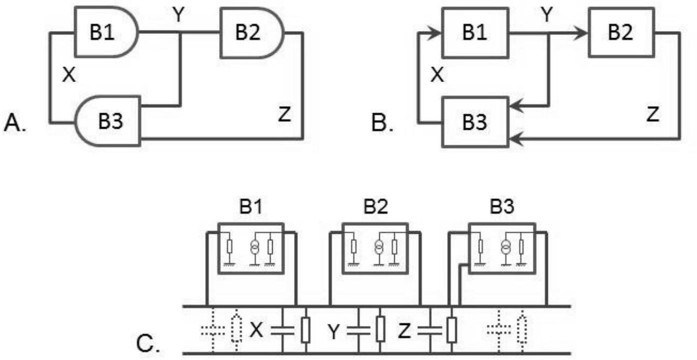


Figure 1: Representation of the various models of a set of biobricks. A. High-level model, B. Signal flow model, and C. Conservative model.

Low-level modeling for validation purpose

Biological mechanisms are usually modeled by mathematical equations linking the chemical species involved. These equations can be transformed into coupled ordinary differential equations (ODEs). This system can then be simulated on dedicated tools. If the different species involved don't interact, the systems can be described in the form of block diagrams and therefore simulated by flow signal type simulators (e.g. MATLAB-Simulink). This type of model (illustrated Figure 1.B) is easy to develop but has limitations, especially with the increase of the necessary number of couplings and feedbacks between blocks, which is the case when a mechanism consumes a species created by another mechanism.

To overcome this problem, we use an alternative modeling approach, called conservative (illustrated Figure 1.C), to solve all differential equations in parallel. To facilitate the formalization of this model, we converted the ODEs into an electrical network [7]. We started from the premise that all ODEs can be written as terms of synthesis/consumption and degradation. Then, we made the analogy with a parallel RC transient electronic circuit. The comparison between these equations enabled us to transform any set of ODEs into an electronic diagram, involving sources of positive current (for terms of synthesis), sources of negative current or variable resistors (for terms of consumption), parallel resistors (for natural degradation) and parallel capacitors (to store the species). This analogy allows us to have a conservative model much easier to instantiate in the case of a complex biosystem, and then to simulate the conservative model using an electronic simulator.

Both types of models have been implemented with two hardware description languages (HDLs) commonly used in electronics and more generally in systems engineering. The first one is the VHDL-AMS [8], highly prevalent, which shows the advantage of having a multi-nature type well suited for the modeling of living. It is very efficient to model hybrid systems but it is dependent on commercial compilers. The implementation using the second language, SystemC-AMS [9], which consists in a set of C + + open-source classes, was motivated by the existence of several types of model of computation corresponding to the two types of low-level models developed (TDF for signal flow models and ELN for conservative models). The advantage of SystemC-AMS is also that it includes a simulator, which is not the case for VHDL-AMS.

The various models presented have been validated on experimental results of the most advanced artificial biosystems published in the literature, which have confirmed the relevance of our models, as well as highlighted some limitations [4].

Fuzzy logic, a promising intermediate description level

Unlike for microelectronics, where the link between high and low-level models is straightforward, in biology, biobricks behavior may vary even if it corresponds to the same digital function. An intermediate level of description between the behavioral and the quantitative levels is required. This is the reason why new models, using the principles of fuzzy logic, have also been developed [10]. Fuzzy logic is a field of mathematics whose concept was developed by Zadeh in 1965, and which approximates the probability field.

Since its introduction, the concept of fuzzy logic has been used in many areas but remains undeveloped for the application on biological material to model biosystems. A computational core was developed specifically for this application in order to have flexible models, to tend either to a digital abstraction or to a nearby low-level model behavior. It is a quantitative model but it remains somehow discrete in that the system behavior is described with a finite number of rules.

The fuzzy logic models leads to accurate and predictive results with very fast time simulation, which was illustrated on complex biosystems [10]. Moreover, description of biobricks through fuzzy rules is more reliable than through digital ones. We can take advantage of this property to improve the classification of biobricks which can be very helpful for the optimization of the choice of biobricks used for targeted logic functions.

Conclusion and perspectives

Finally, this work represents an important contribution to the structuration and the automation of design steps for synthetic biosystems. It helps to outline a complete design flow, adapted from microelectronics, and to highlight its interests. From preliminary work, a number of obstacles have been identified. The first one is the development of a library of models with their associated components. At this level, the various works presented have allowed the development of models, their formalization, and their implementation with tools traditionally used in microelectronics. The library of these models has been developed using an open source language in order to reuse it in other projects without having to use commercial tools.

Improvement of these models is still in progress. We are developing a more accurate representation of the binding mechanism using a binding polynomial, unifying the various existing modeling approaches. The integration of biological noise in the models is also under investigation, as well as the development of stochastic models for the study of a single organism, rather than a population, as it is the case in the current low-level models.

References

[1] International Technology Roadmap for Semiconductors, *http://www.itrs.net/*.

[2] R. Carlson, “The changing economics of DNA synthesis.” *Nature biotechnology*, vol. 27, no. 12, pp. 1091–4, Dec. 2009.

[3] M.W. Lux, B.W. Bramlett, D.A. Ball and J. Peccoud, “Genetic design automation: engineering fantasy or scientific renewal?”, *Trends in Biotechnology*, vol. 30(2), pp. 120-126, 2012.

[4] Y. Gendrault, M. Madec, C. Lallement, J. Haiech, Modeling biology with HDL languages: a first step toward a Genetic Design Automation tool inspired from microelectronics, *IEEE Transactions on Biomedical Engineering*, 2013, accepted, under press.

[5] D. Gerber, M. Fussenegger, “Mammalian synthetic biology: Engineering of sophisticated gene network”, *Journal of Biotechnology*, vol. 130, 2007, pp.329-345.

[6] M. Madec, F. Pecheux, Y. Gendrault, L. Bauer, C. Lallement, EDA inspired Open-source Framework for Synthetic Biology, *BioCAS* 2013, Proc. pp. 374-377.

[7] Y. Gendrault, M. Madec, C. Lallement, F. Pecheux, J. Haiech, Synthetic biology methodology and model refinement based on microelectronic modeling tools and languages, *Biotechnol. J.,* vol 6, 2011, pp. 796-806

[8] G. Peterson, P. Ashenden, and D. Teegarden, The System Designer’s Guide to VHDL-AMS, 2nd ed. Morgan Kaufmann, 2002.

[9] SystemC-AMS language, http://www.systemc-ams.org.

[10] Y. Gendrault, M. Madec, V. Wlotzko, C. Lallement, J. Haiech, Fuzzy logic, an efficient intermediate abstraction level for synthetic biology, *BioCAS* 2013, Proc. pp. 370-373.

The Theatre Management Model of Plant Memory

Vic Norris[1], Camille Ripoll[1] and Michel Thellier[1]

[1] Department of Biology, University of Rouen, 76821 Mont Saint Aignan, France

Abstract

The existence of a memory in plants raises several fundamental questions. What might be the function of a plant memory? How might it work? Which molecular mechanisms might be responsible? Here, we sketch out the landscape of plant memory with particular reference to the concepts of functioning-dependent structures and competitive coherence. We illustrate how these concepts might be relevant with reference to the metaphor of a travelling, avant-garde theatre company and we suggest how using a program that simulates competitive coherence might help answer some of the questions about plant memory.

1 Introduction

The operation of a memory in plants was observed at the beginning of the 1980's [1]. Since then, several other examples of plant memory have been described. Moreover, two different aspects of plant memory have been distinguished [2]. After exposure to the first stimulus or stimuli, one aspect of memory is proposed to entail the plant modifying the pathway that transduces those stimuli and responding immediately; this modification affects the way the plant responds to exposure to the stimulus on subsequent occasions. In this aspect of memory the initiation of the response to the stimulus is immediate. The other aspect of memory is proposed to entail the plant storing information and recalling that information at a later time. In this aspect of memory the response to the stimulus using the memory is delayed.

The existence of a plant memory raises some fundamental questions. How and where does a plant encode information from the environment? How does the plant reconcile what appear to be very different aspects of memory? More specifically, what molecular mechanisms might a plant use to store information without immediately using it? And what mechanisms might it use to recall it? How does a plant integrate the environmental information it has stored along with its own capacities to respond in a final commitment to a growth strategy? In trying to answer these questions, two concepts that have been developed recently may be useful.

One is the concept of *competitive coherence* which has been implemented in an artificial learning program [3]. In competitive coherence, the overall operation of the memory depends on a competition between biological elements (such as genes and proteins) leading to their selection for membership of an active subset of elements from the vast set available to an organism [4,5]. This competition is based on the two patterns of discrete links possessed by each element, where such links include those between a transcription factor and its target genes, a protein kinase and its target, two enzymes in the same structure etc. One of the patterns of links, the *Now* links, connects those elements that are active at the same time as one another. The other pattern of links, the *Next* links, connects those elements that are active at one time with those that are active at the following time(s). Competitive selection between these patterns of elements for the inclusion of an element in the active subset might therefore help a plant find a coherent solution to (1) the need for a plant to have a phenotype that is consistent with its internal and external milieus at the present time with (2) its need for a phenotype that is consistent with its milieus at previous times.

The other concept is that of *functioning-dependent structures*, FDSs, which are assembled (or are disassembled) in response to their activity, such as the metabolizing of a sugar [6]. This activity reflects the environment, and hence the FDSs and their bigger relatives, the functioning-dependent hyperstructures, constitute a measure of the plant's response to the environment [7]. Given the activity-dependent interaction of enzymes with the cytoskeleton, the concept of FDSs can be taken even further to include the enzyme-decorated cytoskeleton as a metabolic sensor [8].

Here, we bring together ideas about FDSs and competitive coherence in a new approach to plant memory. We discuss both concepts in the framework of the metaphor of a theatre as previously used in the case of an influential model [9]. In our version of this metaphor, plant memory can be viewed as a travelling theatre in which the actors must be chosen and the play itself adapted in response to feedback from an audience representative of the different audiences who will pay to see the performance.

2 A unifying, theatre management model of plant memory

Our model of plant memory is the metaphor of a peripatic, avant-garde theatre. Management of this theatre entails casting to fill different roles in a play, rehearsing and adapting this play to the tastes of the audience, and, finally, performing the play. The casting process requires a cast of actors – a subset of genes, macromolecules and ions – to be chosen from a large number of unemployed candidates – unexpressed genes, unsynthesized or inactive macromolecules, and ions at ineffective concentrations. This casting continues through rehearsals. The play is dynamic and interactive: the plot, roles and cast change in response to feedback from a non-paying, invited audience, who represent the environment. The group of actors onstage at any one time changes in a meaningful sequence. An actor can be either onstage playing a role or waiting backstage (or in the wings) ready to come on.

2.1 Competitive coherence – choosing the cast

The casting process is described by competitive coherence, which is a learning strategy for choosing a subset of elements to determine the state of the system from a larger set of inactive elements [3, 5]. In the context of the theatrical model of plant memory, competitive coherence allows the management to select actors to be on onstage based on two sets of requirements. The first requirement is that the actors have a coherent relationship to the present scene; this corresponds to the requirement for memory to be *immediate*. The second requirement is that these actors have a coherent relationship to the actors present in the *preceding* scene; this corresponds to the requirement for memory to be *sequential* in order to take into account temporal changes. Taking the two requirements together, the metaphor translates into the developing plant storing a pattern of activity of elements (genes, macromolecules, ions etc.) that is coherent with both its present internal and external milieus and its preceding internal and external milieus.

2.2 The elements of memory – the actors

There are two classes of elements in our model. In the cytosol, functioning-dependent structures, FDSs (see below), mainly comprise the immediate aspect of memory. Cytosolic FDSs capture – indeed constitute – the state of the cell at any one time. Such FDSs include those involved in metabolism and signalling and may comprise both enzymes and cytoskeletal proteins. In the nucleus, the positions and modification status of certain genes reflect the history of the cell and hence the sequential aspect of memory. These genes may include those involved in sensing temperature, mechanical stimuli, signals from competitors etc.

2.3 The mechanisms for creating memories – the decisions of the director

The candidate mechanisms responsible for the immediate aspect of memory include the processes determining the dynamics of FDSs, which are assembled (or are disassembled) in response to their activity, such as the metabolizing of a sugar; this activity reflects the immediate response of the plant to the environment. These FDSs exist in the cytosol and inside organelles, and the laying down of the immediate aspect of memory may be based on the interactions of these FDS with one another, with the membrane and with the cytoskeleton. Note that FDSs can exist at several levels of organization and, for example, include not only structures within the organelles but also the structures made by organelles in association with the cytoskeleton. FDSs may be stabilised by the post-translational modification of their enzyme and cytoskeletal constituents.

The candidate mechanisms that may be responsible for the inclusion of an element such as a gene in the sequential aspect of memory include: (1) the methylation and demethylation patterns of the DNA and nucleosomal histones and (2) the binding of transcription factors to the gene. Certain of these mechanisms also favour the selection of a coherent set of STO genes because, for example, transcription factors may themselves interact with one another to increase the chances of the genes to which they bind becoming memory genes. The temporal coherence of the successive patterns of memory genes is achieved by coupling the selection of these genes to the expression of the genes controlling circadian and other rhythms.

Laying down the memory, that is, choosing the elements, is a learning process that involves reward and punishment. A part-time or unemployed actor who has just played a scene is more likely (than one who has never been called on) to hang around backstage and therefore be called on again to play another scene. A protein or gene that has been recruited temporarily to the memory to be expressed may undergo a modification that increases its chance of staying available in the ensemble of the memory (i.e., the equivalent of onstage and backstage); in general, modifications to elements, such as the methylation of genes or the phosphorylation of proteins, can increase or decrease the probability that these elements become part of the memory. An actor who has just come offstage from playing in one scene is likely to be followed by an actor who plays in the chronologically next scene in the play (and the first actor may phone this second actor to remind him that his scene is likely to be next). Here, the metaphor means that the genes that have just been expressed (or proteins that are recruited to an FDS) are modified to increase or decrease the probability that the *following* set of genes (or proteins) will be expressed (or recruited) again.

Competition for inclusion in the memory means that the plant can combine the coherence of the set of FDSs assembled and genes expressed in response to a specific environment at a particular time (i.e., immediate memory) with the temporal coherence of the sets of FDSs assembled and genes expressed as an environment changes over time (i.e., sequential memory). Candidate mechanisms for the competition between candidate elements in the immediate and sequential aspects of memory include: (1) the coupling of the chromatin – via the nucleoskeleton and the cytoskeleton – to the cytosolic FDSs and (2) the redistribution of calcium via the condensation and decondensation of calcium onto and from the chromatin and FDSs that constitute the memory [10–12].

3 Examples of the immediate aspect of plant memory

In *Arabidopsis thaliana*, cold shock and hyperosmosis result in a transient increase in cytosolic calcium; in the former case, cold pre-treatments reduce this increase [13] whilst a hyperosmotic-stress pre-treatment heightens it [14]. In *Nicotiana plumbaginifolia* seedlings, a wind stimulus also results in a transient increase in cytosolic calcium but repeated wind stimuli lead to a reduction in the amplitude of the corresponding calcium transients [15]. Other examples include the summing of electrical stimuli in the Venus flytrap [16], the interpretation of gravitropic stimulation in graminean coleoptiles [17] and the adaptation of the phosphate uptake system by the history of phosphate levels in *Anabaena variabilis* [18].

4 Examples of the sequential aspect of plant memory

Pricking the cotyledons of *Bidens pilosa* seedlings shortly after germination (Desbiez et al 1987), followed days later by transfer to a nutrient medium with a very low concentration of mineral ions results in a reduction of the hypocotyl growth, a reduction that does not occur in a standard nutrient medium [19]. Removal of the terminal bud (”decapitation”) of *Bidens pilosa* seedlings in certain conditions of light and mineral concentrations allows one of the cotyledonary buds to start growing before the other; the selection of the bud that grows first is random unless one of the cotyledons is pricked, in which case the other bud grows first; the information leading to the preferential growth of the bud at the axil of the non-pricked cotyledon is stored and is only revealed if the plants are subjected to a particular treatment, such as a change of temperature, at a time that can be days later [20]. Abiotic stimuli such as the manipulation of flax seedlings followed at a later time by a temporary depletion (e.g., for one day) of calcium in the nutrient medium leads to a production of epidermal meristems in the hypocotyl; without either the abiotic stimulus or the calcium depletion, the meristem production does not occur [21].

5 The implication of calcium in plant memory

In most cases, plants rapidly react to stimuli by raising transiently the concentration of free calcium in their cells [22]. This transient elevation of cytosolic calcium sets off a series of processes such as the opening of transmembrane ion channels, post-translational modifications of some proteins and the expression of certain genes. This chain of processes leads to the final response of the plant, which may involve leaf movements, morphogenetic changes and metabolic modifications. Moreover, the kinetics and magnitude of the transient increase of cytosolic calcium (which are different for the different stimuli) are proposed to orient the system towards a response appropriate to the particular stimulus that has been perceived [14, 23–25]. Consistent with this importance of calcium, plant memory is perturbed by treatment with pharmacological agents that are known to affect cytosolic calcium and hence any transient increase of cytosolic calcium that might follow stimulus perception [15, 26].

With its multiple bridging roles, calcium might increase the probability of colocalisation of specific proteins, nucleic acids and lipids and hence contribute to the formation and the stabilisation of FDSs. The condensation of calcium onto charged membranes and linear polymers and might also underpin the organisation of FDSs and chromatin and their putative action in memory. In many of our experiments, changes in the level of calcium are responsible for triggering what we have termed the recall of stored information and, significantly, calcium variations are central to diurnal rhythms in plants [27–31] and possibly to seasonal ones [32–34]. This is consistent with such variations playing a role in the sequential aspect of memory.

6 Implication of functioning-dependent structures or hyperstructures in memory

It is conceivable that the entire cytoskeleton with its associated enzymes acts as an FDS that can integrate intracellular and extracellular information [8]. Such metabolic sensing would put the enzyme-decorated cytoskeleton in a strong position to be a central player in memory. In higher plant cells, MAPs play a major role in the dynamics of the MT network [35] and the association of certain enzymes with the cytoskeleton contributes to these dynamics. The cytoskeleton binds some enzymes when they are active, that is, catalyzing their reactions, whilst it binds others when they are inactive. In plants, protein-protein interactions have been found between actin and enzymes that include cytosolic aldolase, three GAPDH isoforms and two enolase isoforms, as well as between tubulin and enzymes that include aldolase, GAPDH and sucrose synthase [36]. Dynamic interactions between microfilaments and MTs also occur in *A. thaliana* [37]. In line with the processing of information that is central to memory, in maize, the presence of sucrose is required for the association of sucrose synthase with microfilaments *in vitro* and probably *in vivo* [6].

FDSs can also include those found within mitochondria and chloroplasts and, at a higher level, those between the organelles and the cytoskeleton. In chloroplasts, association between glyceraldehyde-3-phosphate and phosphoribulokinase leads to the latter's activation which persists even after the enzymes separate [38]. An FDS comprising the Krebs Cycle enzymes might form via increased affinities of enzymes for one another in the presence of substrates and/or calcium and, in line with a role for calcium in FDS-mediated memory, the activities of three Krebs cycle dehydrogenases – pyruvate, isocitrate, and α-ketoglutarate dehydrogenase – are modulated by calcium (for references see [39]). The information concerning increasing demands for ATP, which is coded in calcium transients, is stored in the mitochondria of Hela cells and is proposed to involve changes in the activity of mitochondrial enzymes [39] which may in turn promote FDS assembly. Ilv5p, a mitochondrial protein that catalyses the synthesis of branched chain amino acids, has been implicated in the formation of mtDNA nucleoids in *S. cerevisiae* [40]. Also in this organism, a subunit of α-ketoglutarate dehydrogenase, Kgd2p, is one of twenty proteins found by cross-linking to be bound to mtDNA whilst double mutants affected in this protein and another DNA-binding protein, Abf2p, produce cells lacking mitochondria [41] and, for other references, see [42].

In our hypothesis, plant memory is also based on changes in the organisation of the chromatin during the learning period. Chromosome territories, for example, can be regarded as FDSs. If FDSs in the cytosol and nucleus do indeed constitute the molecular basis of memory, there should be systems that link them. This is the case: in animals and plants, the Sad1/UNC-84 (SUN) domain proteins are part of a complex that bridges the nuclear envelope to connect cytoskeletal elements to the nucleoskeleton and chromatin [43]. These complexes allow transmission of cellular signals to the nucleus and are essential in various cell functions such as the movement of the nucleus and movement within it. SUN domain proteins such as AtSUN1 and AtSUN2 are also found in plants [44].

7 Discussion

In our vision of plant development, the early stages of growth entail the plant receiving, processing and storing diverse information from the environment for use at later stages of development. This information comprises the temporal dynamics of factors that therefore include temperature, light, rainfall, wind, nutrients, competitors, predators and pests. To optimise its chances of survival, growth and reproduction, the plant must integrate these factors together with its own capacities. In other words, a plant must learn and a memory is therefore essential.

There are two opposed learning strategies that a plant might adopt based on the immediate and delayed aspects of memory. One strategy would be for the plant to grow immediately and to try to alter its particular pattern of growth during the acquisition of new information; this would have the advantage of the plant getting started but the disadvantage of the high probability of the selected growth pattern being far from optimal. The alternative strategy would be for the plant to learn about its environment for an entire year (or even longer!) before committing itself to a particular pattern of growth; this would have the advantage of a high probability of the pattern being right but the disadvantage that the plant would have left the field wide open to its competitors (which would change the environment that needed to be learnt). An attractive compromise solution would be for a plant to grow initially in a reversible pattern before committing itself to a growth pattern relatively early in the growing season.

In a computer simulation of the competitive coherence model of learning, competition between the equivalent of the processes responsible for the immediate and sequential aspects of memory (*Now* and *Next* processes, respectively) for inclusion in an active subset of elements is fundamental to learning. We have argued here that the *Now* and *Next* processes have their counterparts in plant memory. If so, it would mean that it is essential for the plant to respond to its environment in order for it to learn. It would therefore mean that the same molecular mechanisms would have to be involved in the immediate and sequential aspects of memory because the underlying processes would be competing for the same final mechanistic space. With this reasoning, if FDSs in the cytosol and nucleus are involved in memory, both types of FDS are involved in each of the processes.

An apparently different possibility would be if the immediate and sequential aspects of memory were to depend on, for example, cytosolic and nuclear FDSs, respectively. Here, the competition at each step of learning by the plant might entail finding coherent cellular solutions to the problem of integrating FDSs in the cytosol and nucleus. This is aesthetically attractive. Moreover, a shift in the importance given to structures in, say, the nucleus, might underpin the commitment step.

The feasibility of the above possibilities might be tested using a program that simulates competitive coherence. The role of cyclically activated proteins and cyclically expressed genes might be tested using elements in the program that are activated cyclically (i.e., selected cyclically for membership of the active subset of elements). The value of separating cytosolic and nuclear FDSs might be tested by dividing the elements of the program into two classes and by treating them differently. Such differential treatment could include a change in the relative weighting given to the elements such that, at the time representing commitment, the Next process (representing the sequential aspect of memory) takes on a greater importance in determining the behaviour of the system.

References

[1] Thellier, M., et al., Do memory processes also occur in plants? *Physiologia Plantarum*, 1982.**56**: p. 281-284.

[2] Trewavas, A., Aspects of plant intelligence. *Ann Bot*, 2003. **92**(1): p. 1-20.

[3] Norris, V., M. Engel, and M. Demarty, Modelling biological systems with competitive coherence. *Advances in Artificial Neural Systems*, 2012.**2012**: p. 1-20.

[4] Norris, V., Modelling E. coli: the concept of competitive coherence. *Comptes Rendus de l'Academie des Sciences*, 1998.**321**: p. 777-787.

[5] Norris, V., G.G. Nana, and J.N. Audinot, New approaches to the problem of generating coherent, reproducible phenotypes. *Theory Biosci*, 2013.

[6] Duncan, K.A. and S.C. Huber, Sucrose synthase oligomerization and F-actin association are regulated by sucrose concentration and phosphorylation. *Plant Cell Physiol*, 2007. **48**(11): p. 1612-23.

[7] Thellier, M., et al., Steady-state kinetic behaviour of functioning-dependent structures. *The FEBS journal*, 2006. **273**(18): p. 4287-99.

[8] Norris, V., et al., Sensor potency of the moonlighting enzyme-decorated cytoskeleton: the cytoskeleton as a metabolic sensor. *BMC Biochem*, 2013. **14**(1): p. 3.

[9] Baars, B.J., In the theatre of consciousness. *Journal of Consciousness Studies*, 1997.**4**: p. 292-309.

[10] Manning, G.S., Limiting laws and counterion condensation in polyelectrolyte solutions. I. Colligative properties. *J Chem Phys*, 1969.**51**: p. 924-933.

[11] Ripoll, C., V. Norris, and M. Thellier, Ion condensation and signal transduction. *BioEssays*, 2004.**26**: p. 549-557.

[12] Thellier, M., V. Norris, and C. Ripoll, Memory processes in the control of plant growth and morphogenesis. *Nova Acta Leopoldina*, 2013. **114**(391): p. 21- 42.

[13] Plieth, C., et al., Temperature sensing by plants: the primary characteristics of signal perception and calcium response. *Plant J*, 1999. **18**(5): p. 491-7.

[14] Knight, H., S. Brandt, and M.R. Knight, A history of stress alters drought calcium signalling pathways in Arabidopsis. *Plant J*, 1998. **16**(6): p. 681-7.

[15] Knight, M.R., S.M. Smith, and A.J. Trewavas, Wind-induced plant motion immediately increases cytosolic calcium. *Proc Natl Acad Sci U S A*, 1992. **89**(11): p. 4967-71.

[16] Volkov, A.G., et al., Plant electrical memory. *Plant Signal Behav*, 2008. **3**(7): p. 490-2.

[17] Nick, P. and E. Schafer, Spatial memory during the tropism of maize (Zea mays L.) coleoptiles. *Planta*, 1988. **175**(3): p. 380-8.

[18] Falkner, R. and G. Falkner, Distinct adaptability during phosphate uptake by the cyanobacterium Anabaena variabilis reflects information processing about preceding phosphate supply. *J Trace Microprobe Techn*, 2003.**21**: p. 363-375.

[19] Desbiez, M.O., et al., Effect of cotyledonary prickings on growth, ethylene metabolism and peroxidase activity in Bidens pilosus. *Plant Physiol Biochem*, 1987.**25**: p. 137-143.

[20] Desbiez, M.O., et al., Memorization and delayed expression of regulatory messages in plants. *Planta*, 1984. **160**(5): p. 392-9.

[21] Verdus, M.C., M. Thellier, and C. Ripoll, Storage of environmental signals in flax: their morphogenetic effect as enabled by a transient depletion of calcium. *Plant Journal*, 1997.**12**: p. 1399-1410.

[22] Knight, M.R., et al., Transgenic plant aequorin reports the effects of touch and cold-shock and elicitors on cytoplasmic calcium. *Nature*, 1991. **352**(6335): p. 524-6.

[23] Dolmetsch, R.E., et al., Differential activation of transcription factors induced by Ca2+ response amplitude and duration. *Nature*, 1997. **386**(6627): p. 855-8.

[24] McAinsh, M.R. and A.M. Hetherington, Encoding specificity in Ca2+ signalling systems. *Trends Plant Sci*, 1998.**3**: p. 32-36.

[25] Sanders, D., et al., Calcium at the crossroads of signaling. *Plant Cell*, 2002. 14 Suppl: p. S401-17.

[26] Verdus, M.C., et al., Pharmacological evidence for calcium involvement in the long-term processing of abiotic stimuli in plants. *Plant Signal Behav*, 2007. **2**(4): p. 212-20.

[27] Wood, N.T., et al., The calcium rhythms of different cell types oscillate with different circadian phases. *Plant Physiol*, 2001. **125**(2): p. 787-96.

[28] Love, J., A.N. Dodd, and A.A. Webb, Circadian and diurnal calcium oscillations encode photoperiodic information in Arabidopsis. *Plant Cell*, 2004. **16**(4): p. 956-66.

[29] Dodd, A.N., et al., Time of day modulates low-temperature Ca signals in Arabidopsis. *Plant J*, 2006. **48**(6): p. 962-73.

[30] Tang, R.H., et al., Coupling diurnal cytosolic Ca2+ oscillations to the CAS-IP3 pathway in Arabidopsis. *Science*, 2007. **315**(5817): p. 1423-6.

[31] Dalchau, N., et al., Correct biological timing in Arabidopsis requires multiple light-signaling pathways. *Proc Natl Acad Sci U S A*, 2010. **107**(29): p. 13171-6.

[32] Lazzaro, M.D. and W.W. Thomson, Seasonal variation in hydrochloric acid, malic acid, and calcium ions secreted by the trichomes of chickpea (Cicer arieteinum). *Physiologia Plantarum*, 1995.**94**: p. 291-297.

[33] DeHayes, D.H., et al., Physiological implications of seasonal variation in membrane-associated calcium in red spruce mesophyll cells. *Tree Physiol*, 1997. **17**(11): p. 687-95.

[34] Sinutok, S., S. Pongparadon, and S. Prathep, Seasonal variation in density, growth rate and calcium carbonate accumulation of Halimeda macroloba decaisne at Tangkhen bay, Phuket Province, Thailand. *Malaysian Journal of Science*, 2008.**27**: p. 1-8.

[35] Lloyd, C. and P. Hussey, Microtubule-associated proteins in plants–why we need a MAP. *Nat Rev Mol Cell Biol*, 2001. **2**(1): p. 40-7.

[36] Holtgrawe, D., et al., Cytoskeleton-associated, carbohydrate-metabolizing enzymes in maize identified by yeast two-hybrid screening. *Physiologia Plantarum*, 2005. **125**(2): p. 141-156.

[37] Sampathkumar, A., et al., Live cell imaging reveals structural associations between the actin and microtubule cytoskeleton in Arabidopsis. *Plant Cell*, 2011. **23**(6): p. 2302-13.

[38] Lebreton, S., et al., Information transfer in multienzyme complexes–1. Thermodynamics of conformational constraints and memory effects in the bienzyme glyceraldehyde-3-phosphate-dehydrogenase-phosphoribulokinase complex of Chlamydomonas reinhardtii chloroplasts. *European Journal of Biochemistry*, 1997.**250**: p. 286-295.

[39] Jouaville, L.S., et al., Regulation of mitochondrial ATP synthesis by calcium: evidence for a long-term metabolic priming. *Proceedings of the National Academy of Science U.S.A.*, 1999.**96**: p. 13807-13812.

[40] MacAlpine, D.M., P.S. Perlman, and R.A. Butow, The numbers of individual mitochondrial DNA molecules and mitochondrial DNA nucleoids in yeast are co-regulated by the general amino acid control pathway. *EMBO Journal*, 2000.**19**: p. 767-775.

[41] Kaufman, B.A., et al., In organello formaldehyde crosslinking of proteins to mtDNA: identification of bifunctional proteins. *Proceedings of the National Academy of Science U.S.A.*, 2000.**97**: p. 7772-7777.

[42] Trinei, M., et al., A hyperstructure approach to mitochondria. *Molecular Microbiology*, 2004.**53**: p. 41-53.

[43] Rothballer, A. and U. Kutay, The diverse functional LINCs of the nuclear envelope to the cytoskeleton and chromatin. *Chromosoma*, 2013. **122**(5): p. 415-29.

[44] Zhou, X., et al., Novel plant SUN-KASH bridges are involved in RanGAP anchoring and nuclear shape determination. *J Cell Biol*, 2012. **196**(2): p. 203-11.

Multi-way signal processing for multicolor fluorescent biosensing

Damien Parrello[1]*, Patrick Billard[1], Asfaw Zegeye [1], David Brie[2], Sebastian Miron[2] and Christian Mustin[1]

[1]Laboratoire Interdisciplinaire des Environnements Continentaux (LIEC)
UMR 7360 CNRS - Université de Lorraine, F-54500 Vandoeuvre-lès-Nancy
[2]Laboratoire en Automatique de Nancy (CRAN)
UMR 7039 CNRS - Université de Lorraine, F-54500 Vandoeuvre-lès-Nancy

Abstract

In this paper, we investigate the possibility of modelling the concomitant response of multiple bacterial genes to environmental factors by using fluorescent whole cell biosensors. An experimental procedure is proposed to obtain three-way data sets by synchronous fluorescence spectroscopy (SFS). A Candecomp/Parafac algorithm (CP) is then proposed to separate simultaneously all the fluorescence signals of reporter proteins and to interpret the response of the corresponding promoters. The joined estimation of several gene responses is the main original point of this blind identification procedure. The method proves to be more powerful than the traditional principal components analysis (PCA) or the singular value decomposition (SVD) and provides a promising multi-way strategy for in vivo monitoring of gene expression in system biology studies.

1 Introduction

One of the essential aspects to understand cellular systems is to identify the topology of gene regulatory networks from which the metabolic networks are organized and which reflect the functional and adaptive response of the system to its environment [1,2]. A popular method consists of coupling a reporter gene encoding a fluorescent protein with an environmentally responsive promoter [3]. A single wavelength measurement is then realized to detect and quantify the resulting fluorescence. However, monitoring simultaneously the expression of several genes using fluorescent reporters becomes increasingly challenging due to possible overlaps of their excitation/emission spectra. It is therefore necessary to choose compatible, spectrally distant reporter proteins to avoid fluorescence crosstalk. Consequently, only a limited number of reporters can be used despite the wide range of fluorescent proteins currently

*Corresponding author: damien.parrello@univ-lorraine.fr

available. Moreover auto?uorescence of culture media and bacteria cells may hamper the use of fluorescent reporters for *in vivo* or *in situ* approaches.

Traditional data analysis approaches just allow to study a system by analyzing its parts in pairs. For instance, data from a gene expression experiment will be organized in matrices such as *gene 1 x analyte 1, gene 2 x analyte 1, gene 1 x condition 2*, etc. Some properties of the system can depend on these two-way interactions but all properties depending on higher order interactions cannot be revealed and understood. Several studies show the limit of two-way models to detect the underlying structures in multi-way arrays, *i.e.* in complex systems [4,5]. The development of multi-way models was performed through the generalization of the standard two-way methods to higher order datasets. With increasing number of application areas, multi-way data analysis has become popular as an exploratory analysis tool but its potential is still largely unexploited in biology [6]. One of the most well-known and commonly applied multi-way models in literature is the Candecomp/Parafac (CP) model, a powerful multi-way data analysis which can also be used as a blind sources separation technique [7].

This paper proposes an original approach able to produce three-way models from CP decomposition of synchronous fluorescence spectra datasets. Models describe the dynamic expression of a set of genes of interest as the function of two crossed environmental parameters. Coupling synchronous fluorescence spectroscopy (SFS) and CP decomposition allows extending the potential of fluorescent whole-cell biosensors to the study of the relationship between several genes in a systemic and non-destructive fashion.

2 Materials and Methods

2.1 Bacterial strains

E. coli TOP10 strain was engineered to constitutively produce fluorescent proteins. Plasmid pPROBE-NT' [8] served as a backbone for the construction of a new promoter probe plasmid series where the *gfp* gene was switched by three others reporter genes encoding turboYFP, E2orange, and dsRed express2. The expression of each reporter gene is controlled by the *lac* promoter (not shown). These new plasmids were introduced in *E. coli* TOP10 by transformation to yield four fluorescent bioreporters strains:

- (gfp), *i.e.* *E. coli* TOP10- pPROBE-NT'lac constitutively producing green fluorescent protein GFP
- (yfp) *E. coli* TOP10-pPRlacY538 constitutively producing the green-yellow fluorescent protein turboYFP

- (E2or) *E. coli* TOP10 -pPRlacO561 constitutively producing the orange fluorescent protein E2orange
- (dsRx) *E. coli* TOP10-pPRlacX591 constitutively producing the red fluorescent protein dsRed express2.

These whole cell bioreporters covered a large domain of fluorescence ranging from blue-green (490 nm) to orange (over 560 nm) and from green (510 nm) to red (590 nm) in excitation and in emission respectively. On the contrary, the untransformed TOP10 strain displayed a very low basal fluorescence.

2.2 Culture conditions

E. coli TOP10 strains were cultured at 37°C overnight in Lysogeny broth (LB, Sigma, pH 7) supplemented with kanamycin (30 mg.L^{-1}) at 260 rpm. Before dispensing and patterning in microplate, fluorescent cell suspensions were rinsed twice in a saline solution (NaCl 9‰) and their optical density (OD$_{600\text{nm}}$) was adjusted to 2 ($\approx 4.10^8$cells.ml^{-1})

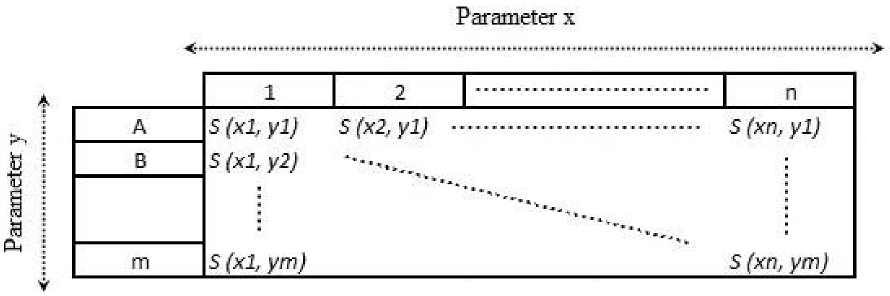
Cells suspensions were finally dispensed in black polypropylene 96-well microplates (Eppendorf Microplate 96/U) with an automated pipetting system (Eppendorf epMotion 5070) according to the pattern described below.

2.3 Microplate Patterning and Three-way dataset generation

We simulated the response of 4 genes following two arbitrary parameters by mixing several serial dilutions of cell suspensions of bioreporters in a black 96-well microplate according to the experimental matrix presented in figure1. This special patterning generates a three way dataset useable by signal processing method (CP). In a “real” experiment, the dispensing pattern must constraint experimental diversity, e.g. by varying gradually or randomly specific environmental factors, in order to identify and estimate the transcriptional response of the studied genes. It is quite similar to iterative refinement, a multi-way data production strategy used in systems biology.

Three-way experimental patterns (A, B and C) were developed to study the characteristics of CP separation profiles of known sources corresponding to general gene expressions patterns. Experimental Pattern A was developed to test CP decomposition efficiency on a trilinear dataset of four overlapping fluorescent signals, evolving independently. In addition, patterns B and C were designed to study the CP decomposition according to colinear or partial collinear response of reporter genes to the parameters of interest. For pattern B, TOP10- pPRlac-Y538 [yfp] and TOP10-pPRlac-O561 [E2or] strains were mixed to produce a trilinear dataset presenting a colinearity in one mode.

For pattern C, the same bioreporters [yfp] and [E2or] were mixed to design a trilinear dataset presenting a partial colinearity in mode 1 and independent behavior on mode 2 but similar to one in mode 1.

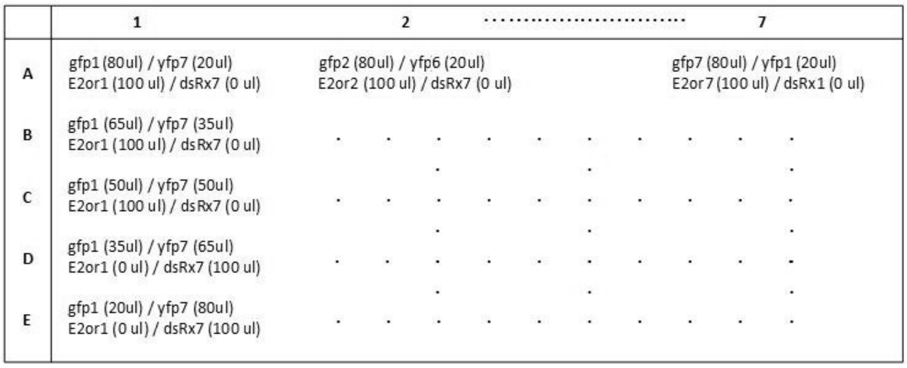


	1	2	....	n
A	S (x1, y1)	S (x2, y1)	....	S (xn, y1)
B	S (x1, y2)			⋮
	⋮			⋮
m	S (x1, ym)			S (xn, ym)

Figure 1: Generalized experimental matrix to produce three-way dataset for CP decomposition. S (x,y) represents spectra following the variation of parameters x and y.

In details, to generate the pattern A, [gfp] and [yfp] suspensions were diluted six times in a 2ml Eppendorf deepwell plate according to a dilution factor of 2/5 and [E2or] and [dsRx] suspensions according to a dilution factor of 1/3. The genuine TOP10 strain was used as a diluent to maintain a constant cell concentration in the wells. Cell suspensions were then distributed in the microplate as described in figure 2. From column 1 to column 7 and from line A to line E, the density of fluorescents reporters (gfp, yfp, E2or and dsRx) varied whereas the total cell density remain constant ($\mathrm{OD}_{600\mathrm{nm}} = 2$). Thus, the contributions of GFP, turbo-YFP, E2orange and dsRed express2 fluorescence signal decreased or increased as a function of the dilution rank (parameter x) or of the mixing ratios (parameter y).

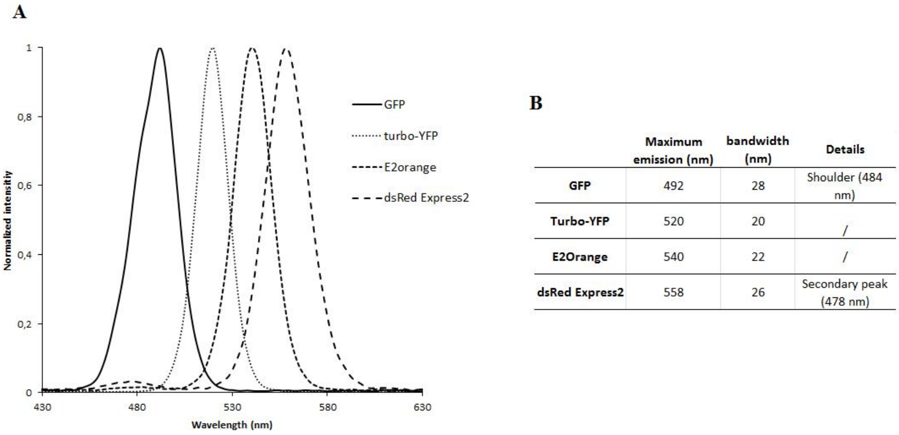
Before fluorescence reading, cell suspensions (36 wells) were shortly mixed with vortex mixer (Mixplate Eppendorf). Synchronous Fluorescence Spectra (SFS) were performed in 96 well microplate with two-grating monochromator spectrofluorometer *FLX-Xenius*® (SAFAS, Monaco), equipped with a 150 W Xenon lamp as excitation source. The Synchronous Fluorescence Spectra (SFS) analysis of biosensors suspensions was measured in the excitation wavelength range of 400-700 nm at constant offset $\Delta\lambda$ of 20 nm. Spectra were recorded with a spectral step of 2 nm, a mean scan speed of 600 nm.min^{-1}. The excitation and emission slits width were 10 nm and the photodiode detector was operating at a voltage of 750 V. Raw fluorescence data (without filtering or smoothing) were collected and exported for further signal processing with CP algorithms running under Matlab software.



	1	2	....	7
A	gfp1 (80ul) / yfp7 (20ul) E2or1 (100 ul) / dsRx7 (0 ul)	gfp2 (80ul) / yfp6 (20ul) E2or2 (100 ul) / dsRx7 (0 ul)		gfp7 (80ul) / yfp1 (20ul) E2or7 (100 ul) / dsRx1 (0 ul)
B	gfp1 (65ul) / yfp7 (35ul) E2or1 (100 ul) / dsRx7 (0 ul)	. . .	. . . . .	. .
C	gfp1 (50ul) / yfp7 (50ul) E2or1 (100 ul) / dsRx7 (0 ul)	. . .	. . . . .	. .
D	gfp1 (35ul) / yfp7 (65ul) E2or1 (0 ul) / dsRx7 (100 ul)	. . .	. . . . .	. .
E	gfp1 (20ul) / yfp7 (80ul) E2or1 (0 ul) / dsRx7 (100 ul)	. . .	. . . . .	. .

Figure 2: Microplate dispensing pattern (A) to simulate biosensors expression following two arbitrary parameters. Numbers 1 to 7 represent the rank of dilution of E.coli TOP10 bioreporters (GFP, turbo-YFP, E2Orange and dsRed Express2). Dispensed volume of each bioreporters (microliters) are given in parenthesis.

Figure 3 summarizes characteristics of GFP, turbo-YFP, E2orange and dsRed express2 SFS spectra expressed in TOP10 strains according to the spectral acquisition described above.



	Maximum emission (nm)	bandwidth (nm)	Details
GFP	492	28	Shoulder (484 nm)
Turbo-YFP	520	20	/
E2Orange	540	22	/
dsRed Express2	558	26	Secondary peak (478 nm)

Figure 3: Normalized SFS spectra of GFP, turbo-YFP, E2Orange and dsRed Express2 (A) and their spectral characteristics (B). Wavelength shift $\Delta\lambda = 20\text{nm}$.

2.4 Candecomp/Parafac (CP) decomposition

The estimation of CP decomposition of the three way SFS data is achieved by an optimized non- negative ALS algorithm. The code was developed by Bro and Anderson and is available in the Matlab N-way toolbox [9]. The R order CP decomposition of a three-way array X (three-order tensor) can be written as:

$$\chi = \sum_{i=1}^{R} a_i b_i c_i + E$$

where R is the number of factors (i.e. the number of fluorescence sources) $\mathbf{a_i}$, $\mathbf{b_i}$ and $\mathbf{c_i}$ the three component matrices and $\mathbf{E}$ a residual three-order tensor representing signal noise or behaviour that could not be included in the decomposition model. CP method approximates the three-way datasets as the sum of the tri-linear behavior of each identified factor.

With SFS data set presented herein, performing the CP decomposition yields the three factors representing both the spectra of each bioreporters and their responses to the cross parameters (x,y) (figure 4).

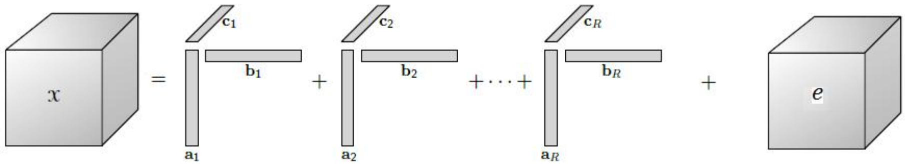


Figure 4: CP decomposition of a three-order tensor χ composed of R fluorescent sources ; Vectors a_i, b_i, c_i given respectively an estimation of the SFS and of responses versus the two modes, i.e. corresponding to parameters x,y.

3 Results and Discussion

3.1 Benefits of coupling synchronous fluorescence spectroscopy with CP decomposition

We combined the synchronous fluorescence spectroscopy with the CP decomposition in order to simultaneously analyze the expression of multiple genes in multi-parametric experiments.

On the one hand, SFS facilitates the data acquisition and improves the quality of the spectral information. The commonly used technique to collect the entire fluorescent signals is to perform an excitation-emission matrix. However this procedure is time consuming and bleaches fluorescent signals. Another alternative is to vary simultaneously (in a synchronous manner) the excitation (λ_{ex}) and emission (λ_{em}) wavelength while keeping a constant interval $\Delta\lambda$ between them. This procedure was proposed by Lioyd (1971) and leads to numerous advantages compared to classical fluorescent spectroscopies as demonstrated by Tuan Vo-Dinh (1978) [10,11]. SFS records the diversity of fluorescent signals (e.g. mix of biosensors) in a single multimodal spectrum, improves the sensibility and the spectral resolution, simplifies spectra, reduces interference signals, and gives the ability to select spectral information. This last property is particularly powerful when several genes are tagged or when biosensors are used simultaneously. Indeed the wavelength shift ($\Delta\lambda = \lambda_{emission} - \lambda_{excitation}$), can be tuned in order to amplify signals of interest (e.g close) or to limit interference signals [11,12]. A suitable approach consists to set this value close to Stoke's shift of fluorescence proteins.

On the other hand, CP decomposition solves the fluorescence overlapping issue without a priori about the source signals and reveals if any, the multilinear relationship between spectral components [7,13,14,15]. Generally, interferences and overlaps between emission bands (spontaneous ?uorescence from metabolites or medium) limit the interpretation of fluorescence data. Thus, the simultaneous monitoring of gene expressions remains complicated and lead to work with limited set of fluorescent proteins or reporters. The identification of spectral sources is a classical but difficult problem in signal processing, needing blind and powerful identification procedure. Traditional data analysis methods such as Partial Least Square projection (PLS) or Single Value Decomposition (SVD) require an extended knowledge of source spectra (Database) or are limited to bi-linear approaches. Two-way analysis methods do not take into account more than two parameters at the same time without disrupting the inner structures of datasets [16,17]. Moreover, the two-way identification procedures of unknown sources assume that the spectra are independent (orthogonal).These methods applied to a non-negative SFS data set would likely yield negative values in estimated spectra. Nevertheless, statistical approaches are especially useful for generating single dimensional solutions for multidimensional biological data.

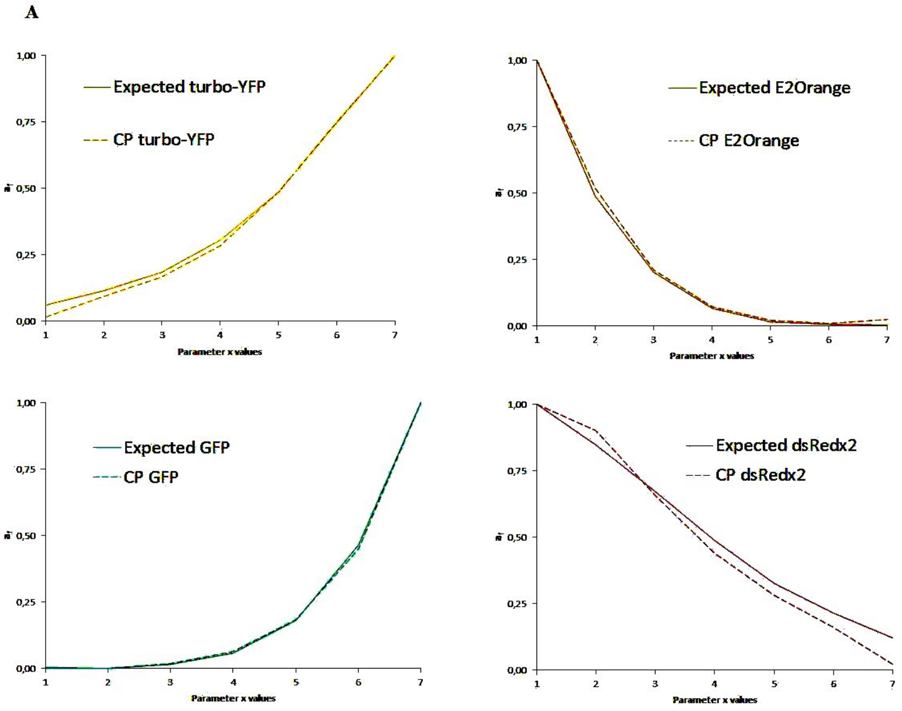
Performing the CP decomposition yields the three factors representing respectively the fluorescence spectra of each bioreporters and their responses to the studied environmental parameters. Multi-way CP decomposition is, thus, a promising tool to study complex systems such as microorganisms interacting with their surrounding and is able to integrate multiple variables in a criss-cross fashion.

3.2 CP decomposition of SFS datasets

CP decomposition is sensitive to the behavior of the identified sources. Characterizing the separation profiles of known sources corresponding to general cases (independence, colinearity, partial colinearity behaviors) is essential to correctly interpret gene relationships in real studies.

The results of CP decomposition on three experimental three-way datasets (A, B and C) are presented hereafter. They are compared to the expected behaviour of gene response (set by dispensing pattern) and to reference spectra of the fluorescent proteins (figure 3).

Figure 5 displays the results of the CP decomposition with 4 sources corresponding to the experimental pattern A (presented in Figure 2). Taking into consideration pipetting variations, source behaviours estimated by CP decomposition are in concordance with the expected variations of fluorescence versus parameters x and y (figures 5 A & B, Mode 1 and 2 respectively). Moreover, the spectral identification of the 4 bioreporters (gfp, yfp, E2or et dsRx) is faithful with the recorded references (figure 5C). Peak positions and half-bandwidths of estimated spectra are accurate and spectral overlaps between fluorescent reporter proteins are not an issue. For example, in spite of their narrow emission peaks (18 nm), spectra of E2orange and dsRed express2 may still separate.



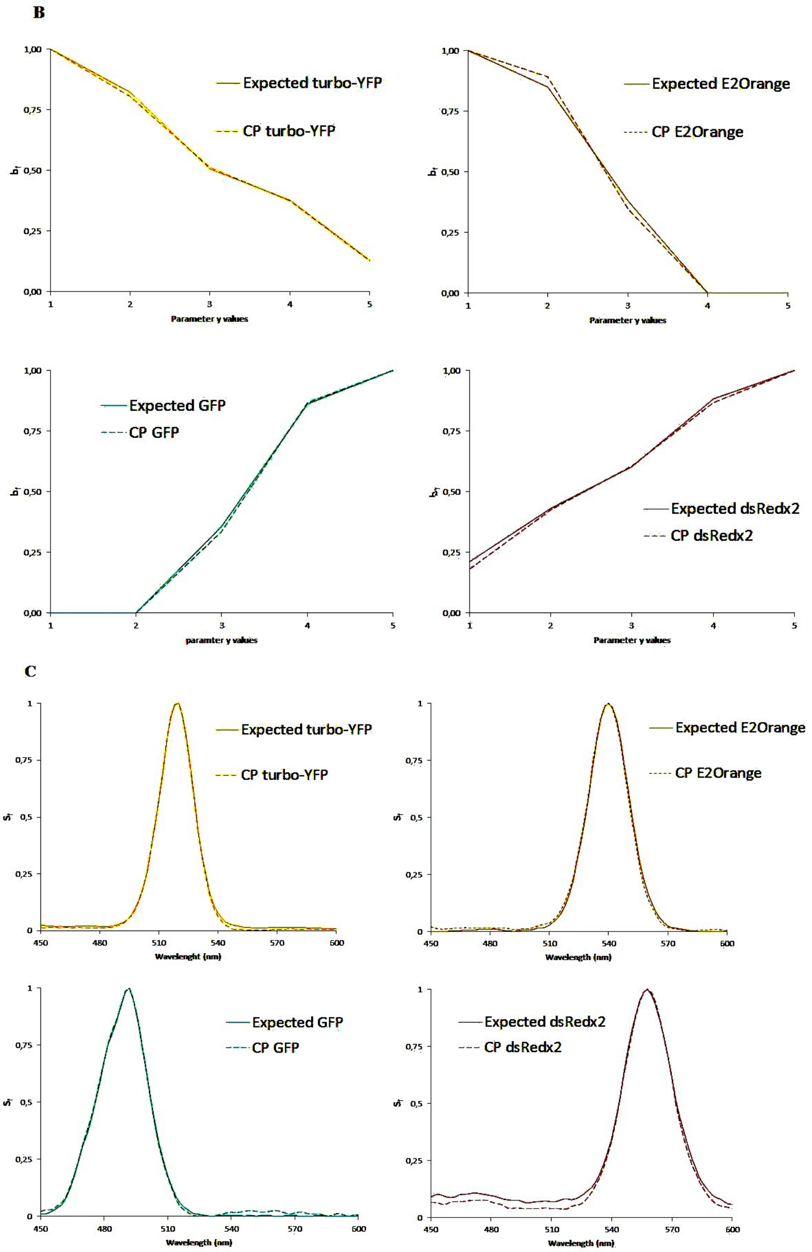


Figure 5: Concordance of CP decomposition results (dotted lines) with expected ones (solid lines) for pattern A. Mode 1 (parameter x) and Mode 2 (parameter y) are presented in box A and B. Box C showed the identified sources spectra (SFS).
]

Figure 6 shows the results of the CP decomposition with 2 sources corresponding to the experimental pattern B, designed to study the effect of co-linearity on one mode. The decomposition identified without any error the co-linearity on the first mode, but the estimated spectra do not perfectly match the references. Slight shoulders appear symmetrically on the right side of turbo-YFP spectra and on the left side of E2orange spectra. In details, the decomposition is incomplete and CP algorithm attributed randomly a part of the turbo-YFP signal to E2orange signal and vice versa. As a consequence, if two or more responses of genes are co-linear in one mode, their spectra will be supplemented by secondary bands corresponding to the collinear sources. Particularly in the case of full co-linearity, only one spectrum including an amalgam of the sources will be identified.

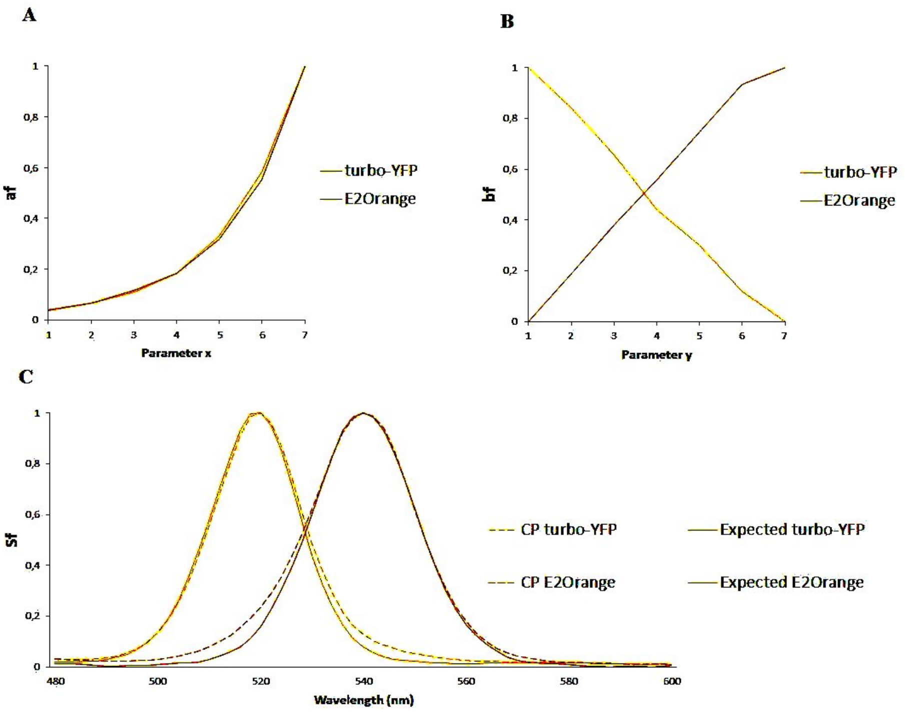


Figure 6: Concordance of CP decomposition for pattern B. Mode 1 (parameter x) and Mode 2 (parameter y) are presented in box A and B. Box C showed the identified sources spectra (SFS) in dotted lines and expected ones in solid lines

Figure 7 illustrates the results of the CP decomposition with 2 sources corresponding to the experimental pattern C, designed to study the effect of partial co-linearity on one mode with inter-modal co-linearity. In this case, the

CP decomposition identified correctly the two fluorescent sources and their respective behaviours vs. mode 1 and 2. The partial co-linearity is fixed by 4 points (out of 7).

Supplementary experiments demonstrated that (i) only one varying point can be necessary to properly separate two collinear sources and (ii) even if their maximum of emission are separated by 5 nm (data not shown).

As shown by the results, coupling synchronous fluorescence spectroscopy (SFS) and CP decomposition allows extending the potential use of fluorescent whole cell biosensors to the study of the relationship between genes in a systemic fashion. And if the maturation time of fluorescent proteins is taken into account, then a temporal dimension can be integrated in the experiments without disturbing the system integrity.

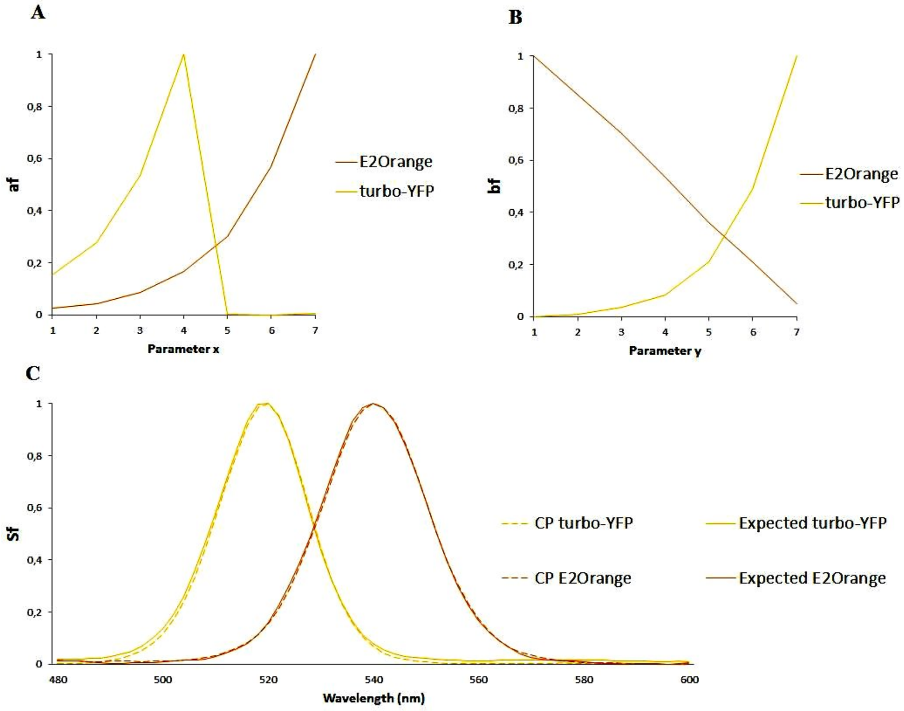


Figure 7: Concordance of CP decomposition for pattern C. Mode 1 (parameter x) and Mode 2 (parameter y) are presented in box A and B. Box C showed the identified sources spectra (SFS) in dotted lines and expected ones in solid lines.

3.3 Limits of the method: rank of decomposition

A well-known problem when using CP decomposition is the choice of the number of the components or sources with which the decomposition has to be done. In others words, how many relevant signal sources contribute to explain the variation of the data according to the different modes? The decomposition can be performed with infinity of sources but one cannot ensure that the supplementary sources fit either relevant behaviors or the noise.

However it is possible to estimate an optimal number of sources according to the energy component of each source. The energy component represents the level of contribution of each source in the explained variation. The explained variation indicates the fitting accuracy of the initial data tensor. For instance, an explained variance of 90% means that 90% of the variation in the data is described by the model, while the remaining 10% is related to the noise or data not included in the model. Thus we can soundly define that if the n$^{\text{th}}$ supplementary source is 2 log under the most explicative source, the optimal number of sources is (n-1). Figure 8 shows the evolution of the energy component versus rank of decomposition. The energy component stays almost stable for 4 sources with less than 1 log of difference between them. This means that 4 sources are needed to correctly fit the data. The 5$^{\text{th}}$ source is more than 2 log lower and thus not required to significantly improve the model.

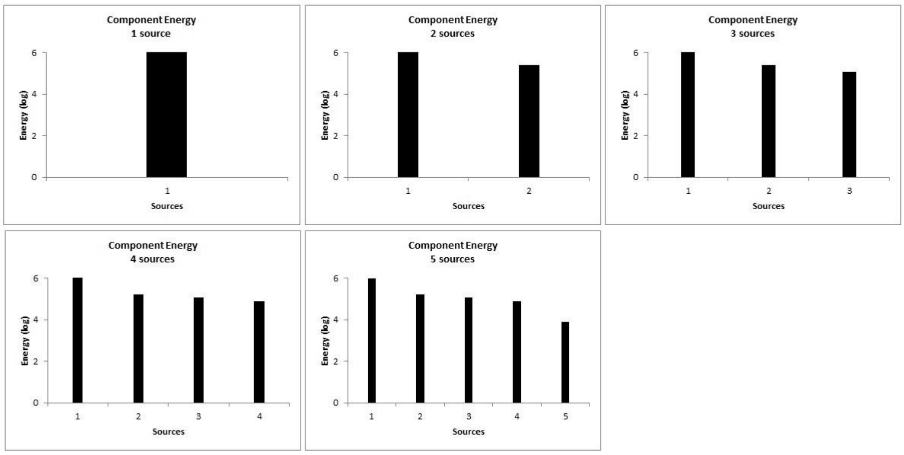


Figure 8: Component Energy for CP decomposition of dataset A achieved with 1 to 5 sources. Energy is expressed in logarithm.

4 Conclusion

We developed an original method at the interface of synthetic biology and systems biology in order to study topology of gene networks. From a natural set of genes representative of a cellular response, a linear synthetic circuit integrating the corresponding promoters coupled with reporter genes can be constructed and inserted into an organism of interest which can be used in multi-way experiments to capture genetic network toplogies. Spectral data acquisition realized with synchronous fluorescent spectroscopy helps to facilitate the production of diversified multiway dataset. CP-based signal separation methods solve simultaneously the overlapping issue linked to complex environments (autofluorescent medium, metabolites, etc.) and the use of neighbouring reporter signals. CP decomposition was chosen because of its signal separation potential but also because the models arising from it can be interpreted without ambiguity. The proposed methodology was used to study a three-way dataset that simulate different gene expression topologies. We demonstrated that the method could be used to study several overlapping signals and that the quality of the separation is linked to the interactions of these signals as a function of cross parameters (modes) from which the gene network dynamic can be described.

The proposed method cannot integrate as many genes as the classical transcriptomic approach but offer the opportunity to confirm the identified networks and monitor them temporally without perturbing the system. In an another work, we applied the method for the study of two antagonist genes involved in the iron homeostasis (bfrB and pvdA) present in *Pseudomonas aeruginosa* PAO1 strain (a model of soil bacterium) and described a sensibility threshold of this simple gene system to iron starvation which was not yet reported and thus confirmed the high potential of this methodology for the study of living systems.

Acknowledgements

This work has been supported by the French ANR program through grant ANR-09-BLAN-0336-04.

References

[1] Aderem, A. (2005). Systems biology: its practice and challenges. Cell, 121(4), 511-513.

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

[3] Daunert, S., Barrett, G., Feliciano, J. S., Shetty, R. S., Shrestha, S., & Smith-Spencer, W. (2000). Genetically engineered whole-cell sensing systems: coupling biological recognition with reporter genes. Chemical Reviews, 100(7), 2705-2738.

[4] Estienne, F., Matthijs, N., Massart, D. L., Ricoux, P., & Leibovici, D. (2001). Multi-way modelling of high-dimensionality electroencephalographic data.Chemometrics and Intelligent Laboratory Systems, 58(1), 59-72.

[5] Acar, E., Camtepe, S. A., Krishnamoorthy, M. S., & Yener, B. (2005). Modeling and multiway analysis of chatroom tensors. In Intelligence and Security Informatics (pp. 256-268). Springer Berlin Heidelberg.

[6] Yener, B., Acar, E., Aguis, P., Bennett, K., Vandenberg, S., & Plopper, G. (2008). Multiway modeling and analysis in stem cell systems biology. BMC Systems Biology, 2(1), 63.

[7] Bro, R. (1997). PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems, 38(2), 149-171.

[8] Miller, W. G., Leveau, J. H., & Lindow, S. E. (2000). Improved gfp and inaZ broad-host-range promoter-probe vectors. Molecular Plant-Microbe Interactions,13(11), 1243-1250.

[9] C. A. Andersson and R. Bro, ”The N-way toolbox for MATLAB,”Chemometrics and Intelligent Laboratory Systems, vol.52, pp. 1-4, 2000.

[10] Lloyd, J. B. F. (1971). Synchronized excitation of fluorescence emission spectra. Nature, 231(20), 64-65.

[11] Dinh, T. V. (1978). Multicomponent analysis by synchronous luminescence spectrometry. Analytical Chemistry, 50(3), 396-401.

[12] Patra, D., & Mishra, A. K. (2002). Recent developments in multicomponent synchronous fluorescence scan analysis. TrAC Trends in Analytical Chemistry,21(12), 787-798.

[13] Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an” explanatory” multimodal factor analysis.

[14] Kroonenberg, P. M. (1992). Three-mode component models: A review of the literature. Statistica, 4, 619.

[15] Harshman, R. A., & Lundy, M. E. (1994). PARAFAC: Parallel factor analysis.Computational Statistics & Data Analysis, 18(1), 39-72.

[16] Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: A literature survey. Knowledge and Data Engineering, IEEE Transactions on, 21(1), 6-20.

[17] Golub GH, Van Loan CF (1989). Matrix computations Baltimore: Johns HopkinsUniversity Press.

Characterization and modeling of an iron-sensitive system in Escherichia coli

Pierre Parutto[1*], Léni Le Goff[1*], Hippolyte Léger[1*], Guillaume Mercy[1], Emiel van der Kouwe[1], Audam Chhun[1], Baptiste Baudu[1], Gabriel Guillocheau[1], Florent Amiot[1], Louis Ujéda[1], William Rostain[1], Cyrille Pauthenier[1], Tristan Cerisy[1], Nicolas Pollet[1], Andrew Tolonen[1] and Alfonso Jaramillo[1]

[1]University of Evry, CNRS, Institute of Systems and Synthetic Biology, Évry, France

Abstract

Iron regulation is a fine tuned mechanism in living organisms. In bacteria, part of this mechanism involves the Ferric Uptake Regulator (Fur) protein possessing a dimeric form that in presence of iron ions acts as transcription factors. In this article, we present the *in vivo* and *in silico* characterizations of the *paceB* promoter, a Fur target promoter found in *Escherichia coli*. The emphasis of the work is put on building a simple model explaining the biological data obtained and analyzing this model.

1 Introduction

Iron homeostasis in the human body, is ensured through a finely tuned absorption mechanism. Most of the iron absorption happens in the small intestine and involves free iron ions. Once absorbed iron cannot be excreted from the body. Misregulation of the absorption mechanism leads to iron overabsorption related diseases such as hemochromatosis.

Within the framework of the iGEM 2013 competition, the Evry team proposed a new way to fight against these diseases by reducing the number of free iron in the intestine. To trap iron ions, we use siderophores, chelator molecules naturally produced by the bacteria *Escherichia coli* to capture iron in the environment in case of iron starvation. By inverting this mechanism one can obtain bacteria capable of producing more chelators as the concentration of iron in the environment increases.

To realize this goal, we designed a synthetic system composed of two plasmids: the sensor plasmid containing a *lacI-lva* gene, encoding the repressor of the lac promoter (*plac)*, under the control of the *paceB* promoter

[*] These authors contributed equally to the work

(a low-iron concentrations sensitive promoter) and the effector plasmid can contain any iron chelator gene under the control of a *plac* promoter. The *lacI/plac* communication system allows to invert the iron concentration signal to produce more effector molecules as the iron concentration increases.

To create such system, a first step is to test independently each parts: the *paceB* promoter, the *lacI/plac* system and the chelator production. The two-plasmids design makes the system modular, changing any of the two plasmid allows to create different systems and decouples the sensor from the effector system. Modeling is also facilitated as one can build a model of the whole system from models of the parts.

In silico analysis of these systems is necessary to find the good range of biological parameters to make *E. coli* bacteria produce the good quantity of chelator molecule for a given iron concentration. As creating too many chelator molecules would lead to underabsorption of iron by the body and too few chelator molecules would make the treatment inefficient.

This article focuses on the first part of the system: the iron-sensing. It presents its *in vivo* characterization and the mathematical models that have been derived from these data. Our purpose here is threefold: 1) present the characterization of the synthetic iron-sensing system in *E. coli*; 2) detail the construction of the models that have been derived for this system and analyze their output based on *in vivo* data; 3) propose a model for the response of the iron-sensitive *paceB* promoter.

2 In vivo iron-sensing system

Description Iron homeostasis is essential to the survival of bacteria [1] as iron is indispensable for the core mechanisms but in too high concentrations becomes toxic. For designing our device, we took advantage of the mechanisms naturally present in *E. coli* ensuring this homeostasis. The genes controlling the iron transport through the membrane in *E. coli* are found in an operon under the regulation of the Ferric Uptake Regulator (Fur) protein [2]. The Fur protein possesses two different binding sites [3]: one to bind to another Fur protein, creating Fur dimers and the other to bind to an iron ion. The Fur protein is thus found both in monomeric and dimeric forms in the cell [4]. The dimeric form in the presence of iron ions has the particularity of being able to specifically bind to a 19 base pairs DNA sequence motif called the Fur Binding Site (FBS) [4]. We searched the whole genome of *E. coli* for such FBS and kept only the sequences that were located in promoter regions. We cloned these sequences and we eventually selected the promoter of the *aceB* gene as the only successful clone.

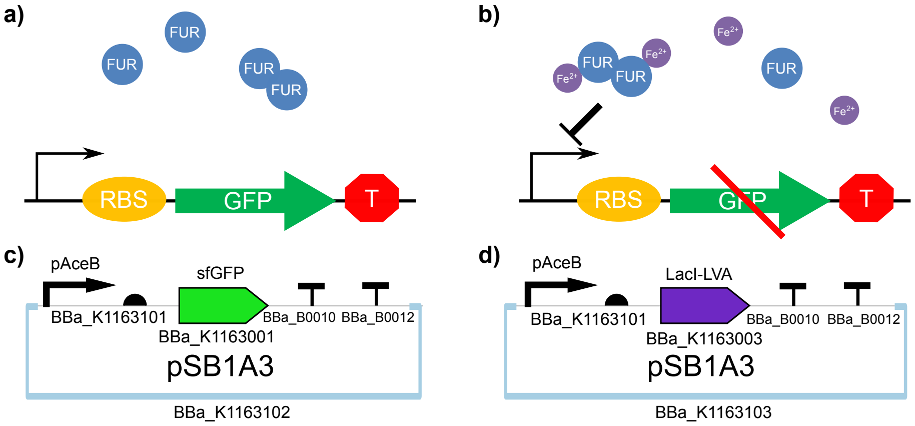


Figure 1: The *paceB* promoter under the control of the Fur protein in absence or presence of iron. **a)** At low iron concentrations, the Ferric Uptake Regulator protein can be found both in monomeric and dimeric forms, that cannot bind to the Fur Binding Site region of the *paceB* promoter allowing the transcription of the Green Fluorescent Protein gene (Gfp); **b)** At high iron concentrations, the Fur dimers can bind to two iron ions forming FeFur dimer complexes that can bind to the FBS region of the *paceB* promoter, preventing the transcription of the downstream *sfgfp* gene. The system is OFF. **c,d)** Map of the genetic constructs used.

The *paceB* promoter works in the following way: in its default state, the Fur dimer cannot bind to the FBS region of the DNA. However, when bound to ferrous (Fe^{2+}) or ferric (Fe^{3+}) ions, the FeFur dimer complexes are capable of binding to the FBS region of the *paceB* promoter preventing the RNA polymerase from transcribing the downstream gene. The behavior of the system is illustrated in Figure 1.

Genetic construction The *E. coli* Top10 strain (BBa_V1004) was chosen as the chassis organism. The different constructs were cloned into the pSB1A3 high copy number plasmid carrying an ampicillin resistance marker for selection. Figure 1b,c present the synthetic plasmids used for the experiments. These plasmids were submitted to the biobrick registry under the IDs: BBa_K1163102 and BBa_K1163103 respectively.

Experimental setup Bacteria containing the construct were grown overnight in $2mL$ of M9 minimal medium (10mL M9 salt (5x), 5μL $CaCl_2$ (1M), 100μL $MgSO_4$ (1M), 800μL glycerol (50%), 12.5μL NaOH(pH 7.4), 40mL H_2O) supplemented with iron (50μL of $FeSO_4$ (10nm) and 1mL of casamino acids (0.2%)) and ampicillin. After one night, the cultures were diluted 200 times in a fresh M9 medium with the same composition. These cultures were then incubated at $37°C$ for 8 hours and washed using iron-free M9 medium. The bacteria were transferred to a 96 well plate containing 4 different M9 media (24 wells each) with Fe^{2+} iron ion concentrations of 0.1μM, 1μM, 10μM and 100μM. For each well, the Optical Density at 600nm (OD600) and the GFP fluorescence at 530nm were recorded every 10 minutes for 820 minutes using a TECAN plate reader.

On the 96 well plate, 24 wells (2 rows) were used per medium in the following way: 4 wells (blank) were filed only with medium; 8 (control) wells were filed with *E. coli* Top10 bacteria possessing the construct of Figure 1d; the 12 remaining wells (data) were filed with *E. coli* Top10 bacteria possessing the construct of Figure 1c. For each set of wells, two different colonies were used. This amounts to 2 biological and 6 technical replicates per medium.

The raw values obtained from the data wells were corrected for the medium absorbance (background absorbance) using the blank wells and for the background fluorescence of *E. coli* using the control wells. These operations were conducted following the data analysis procedure presented by [5].

Biological characterization The results from the plate reader experiment are presented in Figure 2. The values correspond to the mean values over the 12 data wells for each medium. The values for the 100μM media were not exploited because the bacteria did not grow well in this condition.

Figure 2a presents the optical density at 600nm (OD600) curves obtained for the three media and corrected for background absorbance. The differences in growth between the different media is interpreted as being due to the quantity of iron molecules available for the cells growth. Figure 2b presents the fluorescence over OD600 corrected for background fluorescence and absorbance, corresponding to the fluorescence per cell for the three media. A correlation is observed between the cell fluorescence and the quantity of iron in the medium. This is the expected behavior of the system. Finally Figure 2c presents the growth rates mu for each medium, extracted from the data of figure 2a and used in the models.

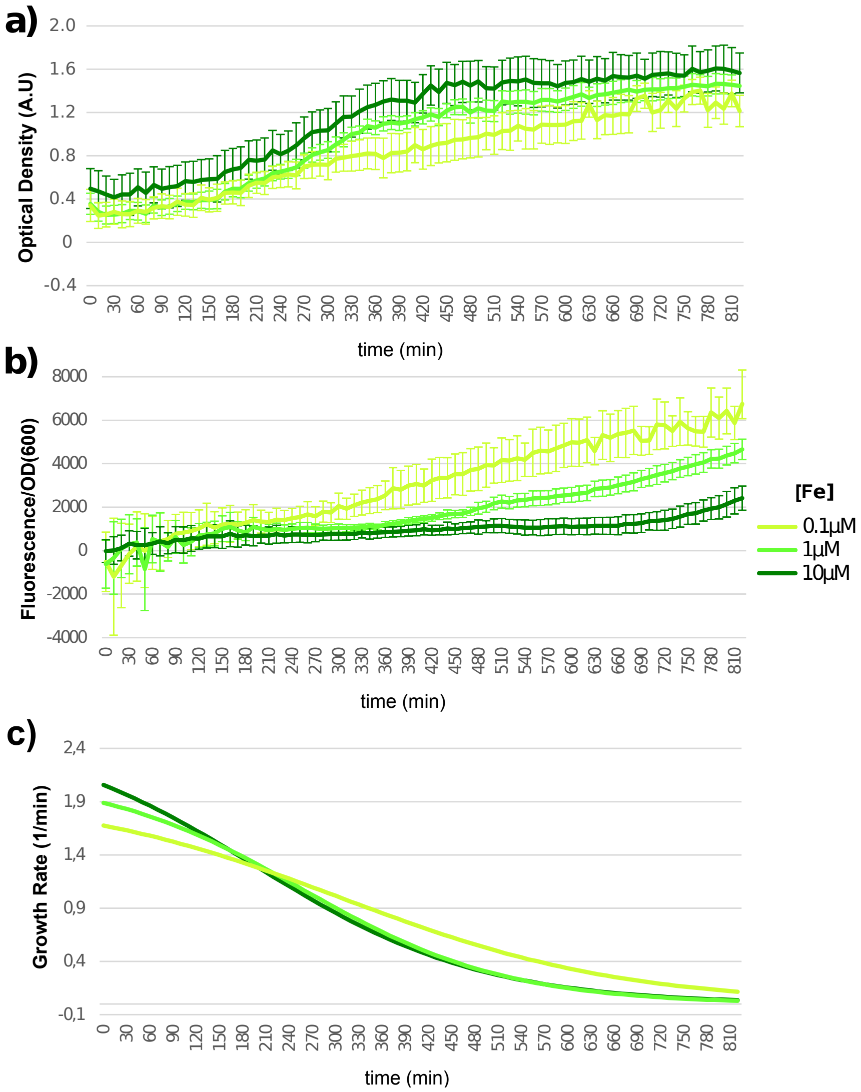


Figure 2: Characterization of the behavior of the *paceB* promoter. The values correspond to the average and error bars to the standard deviation of the values obtained on the 12 data wells. **a)** Mean Optical Density (OD600) corrected for background absorbance for the three different media. **b)** Mean fluorescence over OD600 corrected for background absorbance and fluorescence. **c)** Growth rate functions extracted from **a** and fitted by logistic regression.

3 Model construction

The goal of our mathematical models is to explain the biological data presented before in the simplest way possible. Only a qualitative model can be done with these data as they are expressed in terms of fluorescence but the models consider quantities of molecules. The mathematical framework used is the one of Ordinary Differential Equations where each process is represented as a set of reactions and modeled through mass action law.

Gfp Production The GFP production model represents the production of Gfp proteins that would be obtained by considering no regulation of the *sfgfp* gene. Its goal is to rule out the possibility that the differences in fluorescence observed in figure 2b be due to the differences of growth rates of the cells between the different medium shown in figure 2c.

$$\begin{cases} \frac{d}{dt}RNA(t) = & k_t \cdot N_p & - & (D_{RNA} + \mu(t)) \cdot RNA(t) \\ \frac{d}{dt}GFP_I(t) = & K_p \cdot RNA(t) & - & (D_{GFP} + \mu(t) + F) \cdot GFP_I(t) \\ \frac{d}{dt}GFP_A(t) = & F \cdot GFP_I(t) & - & (D_{GFP} + \mu(t)) \cdot GFP_A(t) \end{cases} \tag{1}$$

The system of equations derived for the GFP production model are presented in (1). In this model, the production of sfGfp mRNA molecules (RNA), was considered proportional to the number of synthetic plasmids (N_p) times the transcription rate (K_t) of one *sfgfp* gene. The quantity of inactive Gfp proteins (Gfp_I) produced by the translation of one sfGfp mRNA molecule by the ribosome depends on the quantity of sfGfp mRNA molecules times the translation rate (K_p) of one sfGfp mRNA molecule. Finally, with some time, the inactive sfGfp protein eventually folds into an active sfGfp protein (Gfp_A) its production is thus proportional to the quantity of inactive sfGfp proteins times the folding rate of one sfGfp protein (F).

For all the quantities, two degradation mechanisms were considered: the dilution due to cell division ($\mu(t)$) computed for each medium from the data (Figure 2c). And the natural degradation of the RNA molecules (D_{RNA}) and of the Gfp proteins (D_{Gfp}), the degradation rate was considered equal in both folded and unfolded states.

Iron sensing The iron-sensitive model adds to the production model the equations representing the behavior of the *paceB* promoter. In order to limit the complexity of the model two approximations were done: we considered a closed system for the total quantity of Fe and Fur dimers; The binding of the FeFur dimers on the DNA was approximated as a chemical reaction.

$$\begin{cases} \frac{d}{dt}p(t) = 2k_r^{Fur} \cdot C_F(t) - 2k_d^{Fur} \cdot p(t)^2 \cdot q(t) \\ \frac{d}{dt}q(t) = k_r^{Fur} \cdot C_F(t) - k_d^{Fur} \cdot p(t)^2 \cdot q(t) \\ \frac{d}{dt}C_F(t) = k_d^{Fur} \cdot p(t)^2 \cdot q(t) + k_r^{FBS} \cdot C_B(t) - (k_r^{Fur} + k_d^{FBS} \cdot (N_p - C_B(t))) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot C_F(t) \\ \frac{d}{dt}C_B(t) = (N_p - C_B(t)) \cdot k_d^{FBS} \cdot C_F(t) - k_r^{FBS} \cdot C_B(t) \\ \frac{d}{dt}RNA(t) = K_t \cdot (N_p - C_B(t)) - (D_{RNA} + \mu(t)) \cdot RNA(t) \\ \frac{d}{dt}GFP_I(t) = K_P \cdot RNA(t) - (D_{GFP} + \mu(t) + F) \cdot GFP_I(t) \\ \frac{d}{dt}GFP_A(t) = F \cdot GFP_I(t) - (D_{GFP} + \mu(t)) \cdot GFP_A(t) \end{cases} \tag{2}$$

The system of equations derived for the iron-sensitive model is presented in (2). The regulation behavior was modeled as follows: two iron ions (p) and one Fur dimer (q) can bind to from a free FeFur dimer complex (C_F) with an association rate of K_d^{Fur} and dissociate with a rate k_r^{Fur}. Free FeFur dimer complexes can then bind to the FBS region of the *paceB* promoter on the DNA of the synthetic plasmid creating a bound FeFur dimer complex (C_B). This binding was modeled as a chemical reaction with an association rate k_d^{FBS} and a dissociation rate k_r^{FBS}.

As each bound FeFur dimer complex occupies the FBS of one plasmid, preventing the transcription of the downstream genes, the number of synthetic *sfgfp* genes that can be transcribed is then $N_p - C_B$, the total number of synthetic plasmids minus the number of bound FeFur complexes. The last two equations for the Gfp production are identical to the ones presented in equation 1.

4 In silico iron-sensing system

The parameter values used in the models are presented in Table 1. These values are either directly extracted from the literature or computed from it. The column "Sim" corresponds to the values used in the following simulations. For both models, the growth rate parameter was computed from the measured optical densities following the method of [5].

Kinetic parameters The direct and reverse kinetic parameters (k_d^i, k_r^i), $i \in \{Fur, FBS\}$ were computed from the association constants (K_A^i) using the fact that:

Symbol	Unit	Range	Sim	Description	Ref.
N_p	$\emptyset$	$[15 - 300]$	232	Number of plasmids	[7]
k_t	min^{-1}	$[2.33 - 7.42]$	2.557	sfGfp DNA transcr. rate	[8]
K_P	min^{-1}	$[3 - 5.25]$	4.276	sfGfp mRNA transl. rate	[9]
F	min^{-1}	$[0.045 - 0.056]$	0.056	sfGfp folding rate	[10]
D_{mRNA}	min^{-1}	$[0.166 - 0.217]$	0.216	mRNA degrad. rate	[11]
D_{GFP}	min^{-1}	$[0.011 - 0.013]$	0.013	sfGfp degrad. rate	[5]
K_D^{FeFur}	μM	1.2 ± 0.06	1.2	Fe-Fur dissoc. cst	[12]
K_D^{FBS}	nM	20	20	FeFur-FBS dissoc. cst	[12]

Table 1: The parameters directly taken from the literature. The column Sim corresponds to the parameters used in the simulations. These values were obtained by running a minimization algorithm on the sum of square differences between the normalized data and model output.

$$K_A^i = \frac{k_d^i}{k_r^i} \Rightarrow k_d^i = K_A^i \cdot k_r^i$$

The K_A^i were computed from the literature following the formula $K_A^i = \frac{1}{K_D^i \cdot N_a \cdot v}$ where K_A^i is the association constant, N_a the Avogadro number and v the volume of an *E.coli* cell ($v = 1.2\mu$m^3 [6]).

Hence, only the k_r^i remain as a free parameters. In order to determine the influence of these parameters on the dynamic of the model, we conducted a specific sensitivity analysis for a range of possible k_r^i values and a logarithmic step of $10^{0.1} min^{-1}$. We analyzed two different outputs of the model for each k_r^i: the final quantity of active Gfp proteins (at t=820 min) and the mean quantity active Gfp proteins. The standard deviation was computed from all the different simulations and is presented in Table 2). The analysis was run for $Fe_0 = 10^{4.5}$ iron ions because this quantity leads to the more unstable case. The results can thus be more easily generalized. It revealed that the dynamics of the model was strictly not affected as long as $k_r^{Fur} > 10^{-4}$min^{-1} and $k_r^{FBS} > 10$min^{-1}. For the simulations, we have arbitrarily chosen values in this range: $k_r^{Fur} = 15$min^{-1} and $k_r^{FBS} = 19$min^{-1}.

Characterization To compare only the qualitative behaviors, both the biological and model data were normalized by dividing by the highest value reached on the three media. We show a comparison between the biological data and the production model (Figure 3a) and the iron-sensing model (Figure 3b). As can be seen, the production model alone does not explain at all the change

Modified parameter	Range	Unit	Output	Std. dev.
k_r^{Fur}	$[10^{-4} - 10^{4}]$	$\min^{-1}$	Final Gfp output	0.0031
k_r^{Fur}	$[10^{-4} - 10^{4}]$	$\min^{-1}$	Mean Gfp output	0.0098
k_r^{FBS}	$[10^{1} - 10^{4}]$	$\min^{-1}$	Final Gfp output	0.0029
k_r^{FBS}	$[10^{1} - 10^{4}]$	$\min^{-1}$	Mean Gfp output	0.0026

Table 2: Results of the sensitivity analysis on the kinetic parameters with $10^{4.5}$ iron ions

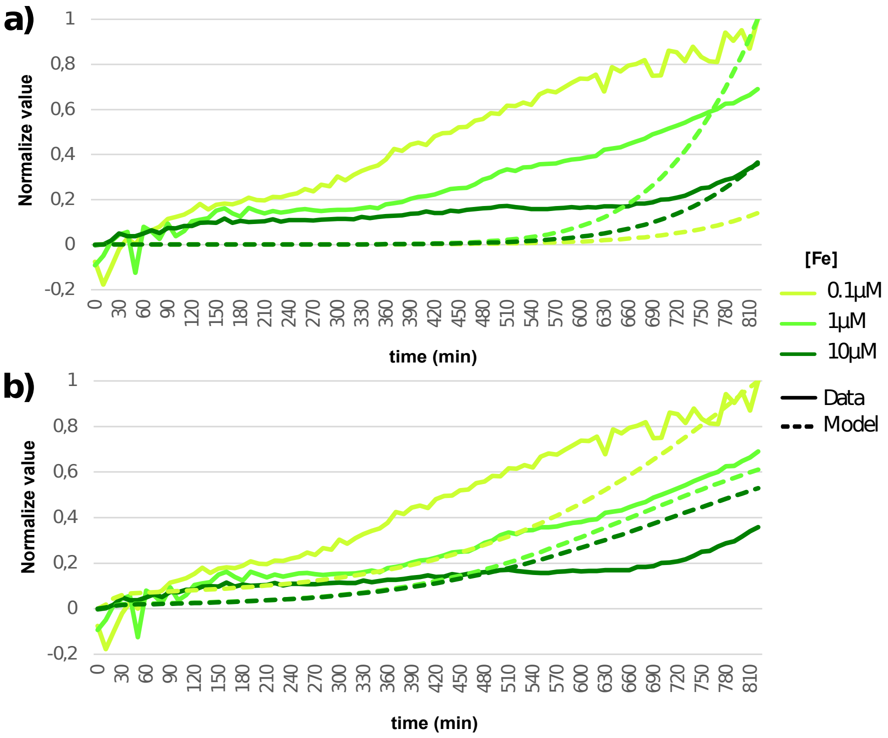


Figure 3: Comparison between the obtained biological data and the simulation output. The biological and model data are normalized by dividing by the highest value of each set of curves. a) Quantity of GFP_A proteins predicted by the GFP production model and b) Quantity of GFP_A molecules predicted by the iron-sensing model. Initial non-null parameters of the simulation: $Fe_0 = 10^3, 10^{3.5}, 10^4$ [13] respectively for the media $[Fe] = 0.1, 1, 10\mu$M, $Fur_0 = 1000$ [14]

in fluorescence and does not possess the same dynamics as the biological data. Even if not perfect, the iron sensing-model is capable of explaining the change in Gfp production due to the difference in iron concentration. It better simulates the biological data.

Sensitivity analysis Finally, we conducted a sensitivity analysis in order to know which parameters impact the most the behaviors of the models. We used the Elementary Effects method (EE) [15] which allows to estimate the contribution of each parameter and the extent of their coupling. The method is the following: construct a grid from the parameter space with p values per dimension uniformly separated by a distance Δ. Generate k trajectories on this grid starting each time at a random point and moving by modifying one parameter at a time until all parameters have been modified. For each point on the trajectory, the difference in the model output compared to the previous point is estimated. The method produces two measures: μ^* assesses the effect of the parameter on the model while σ assesses the coupling with the other parameters.

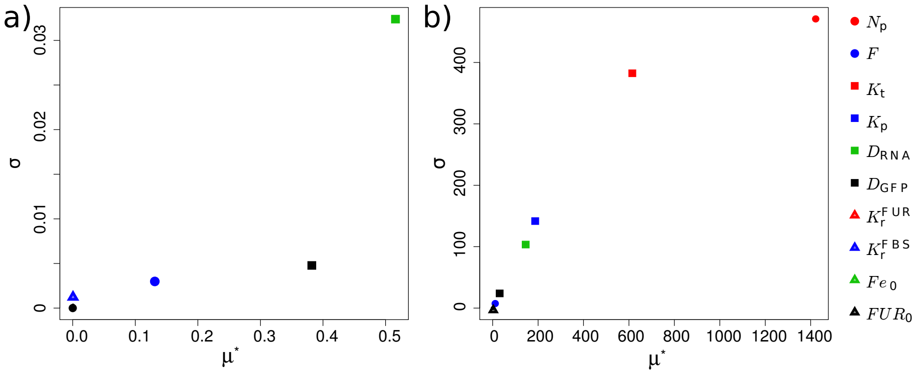


Figure 4: Sensitivity analysis plot, showing for each parameter its mean effect on the model μ^* against its variance σ. a) Analysis of the qualitative behavior of the model considering the area under the GFP_A curve; b) Analysis of the quantitative behavior by considering the final quantity of GFP_A produced. All the non-appearing symbols are superposed in (0,0). Elementary Effects parameters: $r = 75$, $p = 14$, $\Delta = 7$.

Two analyses were conducted on the iron-sensing model. In Figure 4a, the output of the model was considered to be the area under the activated Gfp proteins (GFP_A) curve. In Figure 4b, the output was set to be the final quantity (at t=820min) of activated Gfp proteins (GFP_A). The first output provides information about the effect of each parameter on the qualitative behavior of the model whereas the second approach gives an idea about the variations of the quantitative behavior. It can be seen that the qualitative behavior only depends on three parameters (K_p, F, D_{RNA}) that can be considered as limiting factors of the model. The quantitative behavior on the other hand depends on all the parameters with different degrees of coupling.

5 Discussion

We made the choice of preferring the simplicity of the equations and impose some more assumptions instead of increasing the complexity of the models mainly because of the few biological data at hand. This led to an oversimplification of some phenomenon, for example for determining the kinetic constants where a Fur protein is associated to a whole DNA molecule, resulting in a loss of accuracy. On the other hand, this level of complexity was enough to estimate the different levels of Gfp production at different concentrations of iron.

6 Conclusion

In this paper, we have shown the design and the characterization of the iron-sensitive promoter *paceB*. The biological experiments prove that the synthetic construction effectively works but shows some limitations because of the toxicity of iron in too high concentrations. Then we detailed the construction of a mathematical model to explain these biological data. This model is capable of reproducing the tendency of the biological system but is too simplistic to fully reproduce the behavior. Finally, the sensitivity analysis revealed that only three parameters impact the qualitative behavior of the model while all the parameters impact its quantitative behavior.

References

[1] Simon C Andrews, Andrea K Robinson, and Francisco Rodríguez-Quiñones. Bacterial iron homeostasis. *FEMS microbiology reviews*, 27(2-3):215–237, 2003.

[2] Anne Bagg and JB Neilands. Ferric uptake regulation protein acts as a repressor, employing iron (ii) as a cofactor to bind the operator of an iron transport operon in escherichia coli. *Biochemistry*, 26(17):5471–5477, 1987.

[3] Igor Stojiljkovic and Klaus Hantke. Functional domains of theescherichia coli ferric uptake regulator protein (fur). *Molecular and General Genetics MGG*, 247(2):199–205, 1995.

[4] Lucía Escolar, Jose Pérez-Martín, and Víctor de Lorenzo. Opening the iron box: transcriptional metalloregulation by the fur protein. *Journal of bacteriology*, 181(20):6223–6229, 1999.

[5] Hidde De Jong, Caroline Ranquet, Delphine Ropers, Corinne Pinel, and Johannes Geiselmann. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC systems biology*, 4(1):55, 2010.

[6] HE Kubitschek and JA Friske. Determination of bacterial cell volume with the coulter counter. *Journal of bacteriology*, 168(3):1466–1467, 1986.

[7] Reshma P Shetty, Drew Endy, and Thomas F Knight Jr. Engineering biobrick vectors from biobrick parts. *Journal of biological engineering*, 2(1):1–12, 2008.

[8] Ulla Vogel and Kaj Frank Jensen. The rna chain elongation rate in escherichia coli depends on the growth rate. *Journal of bacteriology*, 176(10):2807–2813, 1994.

[9] R Young and Hans Bremer. Polypeptide-chain-elongation rate in escherichia coli b/r as a function of growth rate. *Biochem. J*, 160:185–194, 1976.

[10] Adam C Fisher and Matthew P DeLisa. Laboratory evolution of fast-folding green fluorescent protein using secretory pathway quality control. *PLoS One*, 3(6):e2351, 2008.

[11] Jonathan A Bernstein, Pei-Hsun Lin, Stanley N Cohen, and Sue Lin-Chao. Global analysis of escherichia coli rna degradosome function using dna microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2758–2763, 2004.

[12] Stephen A Mills and Michael A Marletta. Metal binding characteristics and role of iron oxidation in the ferric uptake regulator from escherichia coli. *Biochemistry*, 44(41):13553–13559, 2005.

[13] Kay Keyer and James A Imlay. Superoxide accelerates dna damage by elevating free-iron levels. *Proceedings of the National Academy of Sciences*, 93(24):13635–13640, 1996.

[14] Ming Zheng, Bernard Doan, Thomas D Schneider, and Gisela Storz. Oxyr and soxrs regulation of fur. *Journal of bacteriology*, 181(15):4639–4643, 1999.

[15] Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

Designed modularity as an engineering principle in synthetic biology

Tina Lebar[1,2] and Roman Jerala[1,2*]

[1] Department of Biotechnology, National Institute of Chemistry, Ljubljana, Slovenia
[2] EN-FIST Centre of Excellence, Ljubljana, Slovenia

Abstract

Modularity is a widespread principle used in nature and even more prominently in engineering. Protein modules harvested from natural proteins are used in synthetic biology to construct new combinations with new functionalities. Our ability to design protein domain modules with defined properties opens exciting prospects for biological engineering. DNA binding domains can be designed to target almost any specific DNA sequence, which can be used for the design of designed orthogonal transcription factors. Designable DNA-binding TALE (Transcription activator-like effector) domains can be used to construct genetic logical NOR gates, which can be used, based in the same Boolean logic in electronics to build complex logic circuits (Gaber et al., NatChemBiol, 2014) in mammalian cells that could be used to engineer cellular response for new sensors, therapeutic or other applications. While the number of natural protein folds in the nature is limited, we can construct new polypeptide folds using modular building blocks. We have used coiled-coil modules to guide the self-assembly of the new topological protein folds. In this engineering approach the protein fold is defined by the sequential order of concatenated coiled-coil interacting pairs, which can self-assemble through interactions between segments into protein polyhedral cage. This principle was demonstrated on the construction of a nanoscale polypeptide tetrahedron, composed of a single polypeptide chain composed of 12 coiled-coil forming segments (Gradišar et al., NatChemBiol, 2013). This principle presents a new paradigm of structural scaffold formation, with potentials for many different applications.

1 Introduction

Biological systems are based on the self-assembly of natural polymers, such as DNA and proteins, and are one of the most complex forms of the "organized chaos". The sophisticated machinery of a cell is far from being understood completely, however, due to the spectacular progress of molecular biology in

[*] Corresponding author: roman.jerala@ki.si

the last few decades, we are now able to genetically alter living organisms, and endowing them with new characteristics and traits not found in nature. This knowledge has eventually lead to the genesis of a new discipline, synthetic biology, which uses the engineering principles for manipulation of biological systems [1]. While synthetic biology can be a powerful tool for studying biological systems and their molecular mechanisms, its main focus is creating new biological parts, devices and systems from well characterized elements.

Many natural proteins are intrinsically modular, composed of different protein domains, which facilitates wide combinatorial complexity. Ideal modular elements are characterized by their relative independence of the context. Modules can be rearranged for new purposes, while retaining similar basic properties as the original element in the initial context. Multimodular proteins, therefore, can be designed for new specific functions in the context with other modules, isolated from the endogenous cellular processes. A special subset of protein modules is particularly attractive for synthetic biology, the one where the structure of protein modules can be rationally designed to perform different functions, such as interactions with other specific proteins or DNA sequences. Wide basis set of designed modules therefore represents an extremely powerful toolbox for synthetic biology. Two types of designable modules have recently been discovered and elaborated, design of proteins that recognize any selected DNA sequence and coiled-coil interacting segments. These traits make modular proteins, such as zinc finger proteins (ZFPs), transcription activator-like effectors (TALEs) and coiled-coil proteins powerful tools for use in synthetic biology, with prospects in bionanotechnology and therapeutic, diagnostic, environmental and industrial applications.

2 Designable modular DNA-binding domains

Modular DNA-binding domains with a known DNA recognition code allow design and construction of proteins to bind virtually any nucleotide sequence [2,3]. Such designed proteins should display high DNA-binding affinities and low cross-reactivity due to the high specificity. They should be able to target DNA sequences that are sufficiently long to ascertain the absence of the target site in the host genome of high complexity and high orthogonality among several designed DNA domains used in synthetic devices. Recognition sequence of 18 nucleotides provides combinatorial complexity above 60 billion combinations and even taking into account the off target binding to sites having one to two mismatches they still provide large potential basis set.

Zinc finger proteins (ZFPs), which are the oldest well understood modular DNA binding domains represent a large repertoire of DNA-binding domains among the thousands of protein domains characterized up to now [4,5]. Each finger of the DNA-binding domain binds to the designated trinucleotide [6]. By combining several fingers, we can design and create proteins for specific targeting of DNA. However, individual finger domains and linkers between domains can affect the binding efficiency of the neighboring fingers, which demonstrated an important limitation to the designable modularity of ZFPs [7]. Transcription activator-like effectors (TALEs) are transcription factors, originating from plant pathogen bacteria. Their DNA-binding domain contains modular repeats of 34 aminoacids, where the 12th and 13th residue determine specific recognition of a single base pair [8,9]. Unlike ZFPs, these modules can be arranged into a single DNA-binding domain, designed for almost perfect reliability of binding to a specific DNA sequence. This designable modularity allows their combinations with other protein domains for a wide array of functions. For example designed DNA-binding domains can be fused to nucleases [10-13] with great potential for the genome engineering, while there have been reports of their fusion with biosynthetic enzymes [14], transcriptional repression and transcriptional activation domains [15-18] and many others.

2.1 Enhancing metabolic flow

Many organisms produce biologically active substances, useful e.g. for the pharmaceutical or food industry. In many cases, the productivity of the desired compound by combinations of natural biosynthetic enzymes needs to be improved. A branch of synthetic biology, known as metabolic engineering, strives to increase the production of these substances using genetically engineered organisms. A number of research groups have published reports on different strategies for the enhancement of the metabolic flow [19-21] and even construction of metabolic pathways, non-existent in nature [22].

The bottlenecks in biosynthetic processes often lie in the nonoptimal metabolic flow between sequential biosynthetic processes, including loss of intermediates to shunt reactions, toxicity of intermediates, imbalanced processivity of the biosynthetic enzymes etc. This can result in low production of the selected compound. Nature has, in some cases, found solutions for this limitation by forming enzyme complexes or by enclosing enzymes of the selected biosynthetic pathway in microcompartments, such as for example carboxysomes. This results in their close proximity, causing a faster diffusion of intermediates to the next enzyme. One of the most innovative synthetic biology-driven solutions to enhance and optimize metabolic flow is binding biosynthetic enzymes to a scaffold, which results in their fixed close proximity, similarly to natural enzyme complexes and microcompartments, yet adaptable to different protein combinations. Coupling of biosynthetic enzymes to protein [23,24] and RNA scaffolds [25] through protein dimerization domains and RNA-aptamer binding proteins, respectively, has been reported. Although these solutions produced encouraging results, one of the drawbacks of both approaches is the limited number of scaffold interaction domains (i.e. protein-protein interaction domains and RNA-aptamer binding proteins), while many biosynthetic pathways consist of a large number of enzymes.

An alternative approach to the same strategy, using DNA as a scaffold to bind modular DNA-binding domains fused to biosynthetic enzymes, was published by Conrado *et al*. [14] Over 1000 characterized ZFPs with defined target sequences are readily available [4,5], and several of those were selected based on the diversity of their binding sequences and previous reports on their binding affinity. Fusion of the selected ZFPs to biosynthetic enzymes in combination with the scaffold DNA that arranges the enzymes in a defined order based on the arrangement of the binding sites for the respective ZFPs. This approach provides the linear arrangement of biosynthetic enzymes taking into account the periodicity of the DNA helix. This approach was demonstrated for the production of resveratrol, mevalonate and 1,3-propanediol resulted in active chimeric enzymes, capable of producing selected compounds, while binding these chimeras to a DNA molecule with ZFP binding sites improved yields of biosynthesis up to 5-fold. Such an approach, based on modular designable elements such as ZFPs, can enable manipulation of biosynthetic pathways, consisting of large numbers of enzymes. In contrast to protein scaffolds, an additional advantage to this approach is spatial control of bound enzymes due to the highly predictable structure of a DNA molecule, while a DNA scaffold is also easy to design and produce in comparison to the highly complex RNA scaffolds.

2.2 Information processing through transcriptional regulation based on modular DNA binding domains

Cells are constantly exposed to different signals from the environment and from the internal cellular processes and adapt their state to these signals, based on different processing of those input signals. Synthetic biology aims to rewire cells for processing external and/or internal signals in new, complex and predictable ways. In this era of rapid (bio)technological advances, synthetic biologists created intricate designed systems, such as logic circuits [26-31], bistable switches [32,33], oscillators [34-36] and cellular counters [37]. The potentials for use of these genetic networks lie in applications, such as environmental or diagnostic sensing, and cell therapy, while the autonomous effective replication of cells provides the cost effective manufacturing of a cellular processor.

Artificial genetic circuits designed so far typically consist of multiple transcriptional regulators, wired into genetic networks to control cellular response. As such, they require introduction of several regulatory elements, which should ideally operate independent from the cell environment to ensure the reliable performance of the designed circuit. Therefore, mostly natural protein regulators have been used for the construction of synthetic biological systems [26-29,32-37]. Apart from their interaction with internal cell processes, the diverse biochemical properties of different proteins may require many adjustments and fine-tuning of the designed systems to ensure their unhindered and optimal behavior [38]. Synthetic transcription factors based on modular DNA-binding domains, can overcome these limitations due to their designability. Genetic fusion of the DNA-binding domain with transcriptional repression or activation domains result in creation of artificial transcription factors, able of highly specific and effective transcriptional regulation [15-18]. The ability to design these synthetic transcription factors to target a large sequential diversity of DNA operators supports construction of complex designed genetic circuits. In a recent publication by Gaber *et al.* [39] we employed repressors, based on transcriptional activator-like effectors (TALEs), to optimize construction of orthogonal NOR gates and implemented all 16 two-input logic functions from combinations of the same type of NOR gates in mammalian cells. In addition, we designed and implemented a genetic logic circuit, where one input is used to select between two logic functions for processing data within the same circuit. In addition to such static logic circuits, it would also be desirable to apply the designable, orthogonal building elements like TALEs as the building blocks for the construction of dynamic, time-dependent logic systems, such as bistable switches. This would provide a potent expansion of the biological engineering toolbox and promote construction of even more intricate genetic systems.

The approach for the construction of genetic circuits, described above, could also be applied for other modular DNA binding domains, such as ZFPs or the CRISPR/Cas system. The availability of a large number of designable orthogonal DNA binding domains supports construction of parallel orthogonal logic gates and switches. The introduction of designed, orthogonal logic operators can improve the predictability, reliability, modularity, and standardization of designed biological information processing systems [1]. Devices with designed logic functions in mammalian cells could be used to evaluate and respond to combinations of signals from the environment, to the physiological state of an organism, or for recognition of protein combinations, characteristic for cancer cells [41]. Cells, as self-replicating and cost-effective cellular factories, also open the prospect of massive information processing devices.

3 New protein folds from modular coiled-coil domains

DNA molecules have been used to rationally design bionanostructures of great complexity, adopting versatile two-and three-dimensional shapes, such as different polyhedra [42-44]. Proteins are the basic building blocks of all cells, and as they fold into the defined structures, they also represent the potent building blocks for artificial nanostructures, with an additional advantage of low-cost production in the recombinant form. However, the main disadvantage of using proteins for building nanostructures is the complexity of their structural information, which by far surpasses the complexity of any other biological molecule.

Creating new structural patterns of protein folding represents a big challenge, since protein structures are determined by a large number of weak, but cooperative long range interactions. De novo protein fold design has been, with few exceptions, successful only for folds, similar to those existing in nature [45-47]. A new synthetic biology approach to formation of new protein structures is based on the principle of modularity. Well characterized protein motifs, such as coiled-coils, can be designed to self-assemble into defined structures independently from other processes [48].

3.1 Self-assembly of new protein folds

The coiled-coil motifs are composed of two or more intertwined α-helices. The characteristic heptad repeats form two turns of a helix, spanning approximately 1nm. Most often coiled-coils form dimers in either parallel or antiparallel orientation. Combining two types of interacting coiled-coil segments can only result in assembly of one-dimensional fibrillar structures [49], while for the construction of two- and three-dimensional structures at least 3 different building elements are required.

In the publication by Gradišar *et al.* [50], the authors reasoned that it should be possible to self-assemble the designed three-dimensional objects from a single polypeptide chain, composed of coiled coil-forming segments (fig. 1). The key is sequential concatenation of the coiled-coil segments, separated by flexible linker peptides. The coiled coil segment pairs in a selected orientation with its complementary interacting segment within the same polypeptide chain thus driving the self-assembly. Each of those coiled coil-forming segments is in isolation unstructured and forms a coiled-coil helix only when it dimerizes with the corresponding complementary segment. This concept was demonstrated by the construction of a tetrahedron with edges of 5 nm, composed of 6 pairs of coiled-coil segments that self-assembles into the designed fold from the linear 12-segment polypeptide chain. Self-assembly of the tetrahedron was confirmed by TEM, AFM, DLS, CD and fluorescent protein reconstitution.

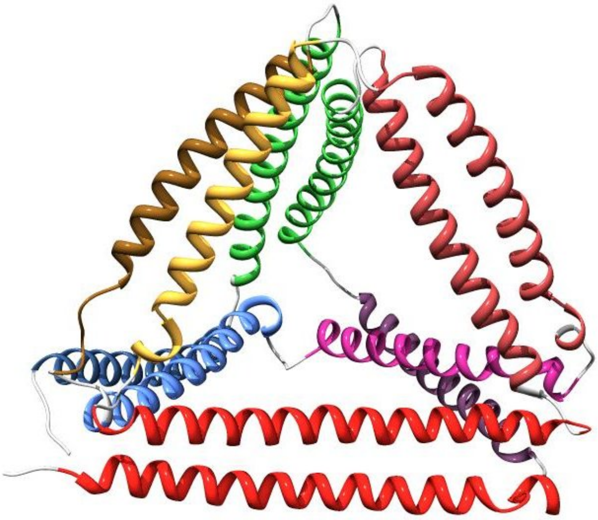


Figure 1: Design of the topological protein fold from designed coiled-coil modules. Tetrahedral nanostructure self-assembles from a single polypeptide chain composed of 12 segments that form 6 orthogonal coiled-coil dimers (Gradišar *et al.*, Nature Chemical Biology, 2013).

This strategy allows creation of new topological protein folds, which have not been observed in nature. The design, based on coiled-coil segments, can produce asymmetric structures, difficult to achieve by the assembly of protein domains or other nanoparticles. Due to their variable length and high length/width aspect ratio, coiled-coil dimers seem to be an almost ideal type of building element, allowing formation of large cavities within the polypeptide polyhedra, unlike interacting protein oligomerization domains, which occupy a major fraction of the volume. It is conceivable that such technological platform could be used for a range of applications, such as using designed cages for the drug delivery, designed vaccines or creation of artificial catalytic sites.

4 Conclusions

Modular proteins can be disassembled and redesigned for new specific functions, which makes them a particularly powerful tool for the use in synthetic biology. Two important areas of protein modularity are the designed DNA binding domains that underlay self-organization of protein domains along the DNA sequence and regulation of complex cellular devices and structural modularity based on orthogonal coiled-coil interactions underlying new type of topological protein folds.

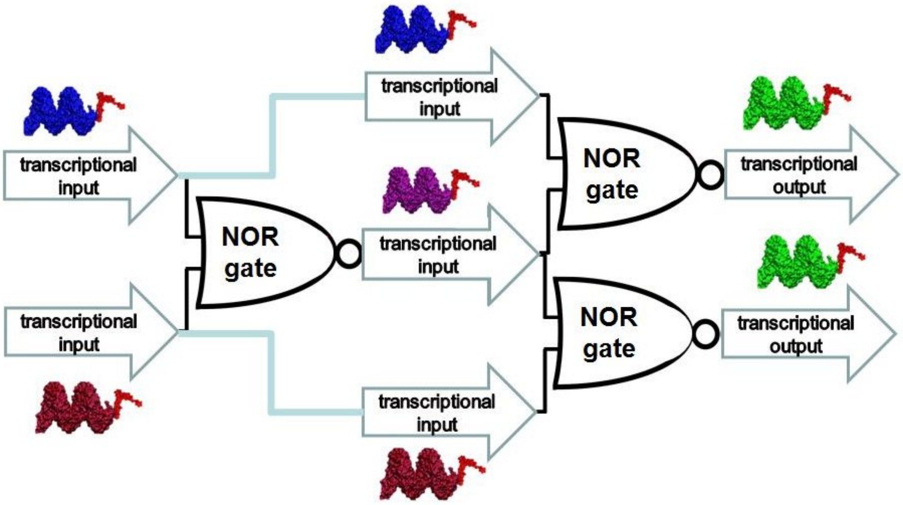


Figure 2: Schematic representation of the layred logic gate based on the designed modular TALE DNA binding domains. Orthogonal TALE-based repressors form NOR logica gates that can form the layered logic genetic circuit. (Gaber *et al*., Nature Chemical Biology, 2014).

Transcription activator-like effectors (TALEs) have been used to construct synthetic transcriptional regulators, implemented for construction of static logic gates and bistable switches in mammalian cells [39] (fig. 2). The prospects for use of these genetic circuits lie in applications such as sensing and medical therapy. Zinc finger protein domains, fused to biosynthetic enzymes, have been shown to improve biosynthetic yields through their binding to a DNA molecule [14]. This aspect is particularly important for the industrial production of biological substances, while such an approach could also be used for the construction of new metabolic pathways and biologically active compounds. Modularity is not only important for the construction of in vivo synthetic systems, but also for the design and construction of new, self-assembled protein structures, which was recently demonstrated on the self-assembly of a protein tetrahedron [50]. Designed, synthetic modular proteins will expand the existing toolbox of available elements and contribute to construction of complex synthetic biological systems and protein structures, composed of a large number of elements. In conclusion, protein modularity is of crucial importance for the future development of synthetic biology.

Acknowledgments

This work was supported by the funds from the Slovenian Research Agency and in part by the European structural funds to the Centre of excellence EN-FIST.

References

[1] Andrianantoandro, E., Basu, S., Karig, D. K. & Weiss, R. Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol* **2**, 2006 0028 (2006).

[2] Bergt, J. M. Toward rules relating zinc finger protein sequences and DNA. **89**, 7345-7349 (1992).

[3] Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* (80-.). **326**, 1509-1512 (2009).

[4] Fu, F. et al. Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res.* **37**, D279-83 (2009).

[5] Fu, F. & Voytas, D. F. Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C?H? zinc fingers and engineered zinc finger arrays. *Nucleic Acids Res.* **41**, D452-5 (2013).

[6] Pavletich, N. P. & Pabo, C. Zinc Structure of a Recognition?: Complex Zif268-DNA. **252**, 809-817 (2011).

[7] Ramirez, C. L. et al. Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat. Methods* **5**, 374-5 (2008).

[8] Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The Crystal Structure of TAL Effector PthXo1 Bound to Its DNA Target. (2012).

[9] Deng, D. et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720-3 (2012).

[10] Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1156-60 (1996).

[11] Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636-46 (2010).

[12] Miller, J. C. et al. A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143-8 (2011).

[13] Sander, J. D. et al. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nat. Biotechnol.* **29**, 697-698 (2011).

[14] Conrado, R. J. et al. DNA-guided assembly of biosynthetic pathways promotes improved catalytic efficiency. *Nucleic Acids Res.* **40**, 1879-89 (2012).

[15] Zhang, F. et al. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* **29**, 149-153 (2011).

[16] Cong, L., Zhou, R., Kuo, Y.-C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.* **3**, 968 (2012).

[17] Geissler, R. et al. Transcriptional activators of human genes with programmable DNA-specificity. *PLoS One* **6**, e19509 (2011).

[18] Garg, A., Lohmueller, J. J., Silver, P. A. & Armel, T. Z. Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.* **40**, 7584-7595 (2012).

[19] Anthony, J. R. et al. Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Escherichia coli for production of the anti-malarial drug precursor amorpha-4,11-diene. *Metab. Eng.* **11**, 13-9 (2009).

[20] Ro, D.-K. et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940-3 (2006).

[21] Steen, E. J. et al. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, 559-62 (2010).

[22] Tobias, A. V & Arnold, F. H. Biosynthesis of novel carotenoid families based on unnatural carbon backbones: a model for diversification of natural product pathways. *Biochim. Biophys. Acta* **1761**, 235-46 (2006).

[23] Dueber, J. E. et al. Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27**, 753-9 (2009).

[24] Agapakis, C. M. et al. Insulation of a synthetic hydrogen metabolism circuit in bacteria. *J. Biol. Eng.* **4**, 3 (2010).

[25] Delebecque, C. J., Lindner, A. B., Silver, P. a & Aldaye, F. a. Organization of intracellular reactions with rationally designed RNA assemblies. *Science* **333**, 470-4 (2011).

[26] Anderson, J. C., Voigt, C. a & Arkin, A. P. Environmental signal integration by a modular AND gate. *Mol. Syst. Biol.* **3**, 133 (2007).

[27] Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat Biotechnol* **31**, 448-452 (2013).

[28] Kramer, B. P., Fischer, C. & Fussenegger, M. BioLogic gates enable logical transcription control in mammalian cells. *Biotechnol. Bioeng.* **87**, 478-484 (2004).

[29] Auslander, S., Auslander, D., Muller, M., Wieland, M. & Fussenegger, M. Programmable single-cell mammalian biocomputers. *Nature* **487**, 123-127 (2012).

[30] Lohmueller, J. J., Armel, T. Z. & Silver, P. A. A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res.* **40**, 5180-5187 (2012).

[31] Lienert, F. et al. Two- and three-input TALE-based AND logic computation in embryonic stem cells. *Nucleic Acids Res.* **41**, 9967-75 (2013).

[32] Gardner, T. S. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli Supplementary Information.

[33] Kramer, B. P. et al. An engineered epigenetic transgene switch in mammalian cells. *Nat. Biotechnol.* **22**, 867-70 (2004).

[34] Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335-338 (2000).

[35] Tigges, M., Marquez-Lago, T. T., Stelling, J. & Fussenegger, M. A tunable synthetic mammalian oscillator. *Nature* **457**, 309-12 (2009).

[36] Tigges, M., Dénervaud, N., Greber, D., Stelling, J. & Fussenegger, M. A synthetic low-frequency mammalian oscillator. *Nucleic Acids Res.* **38**, 2702-11 (2010).

[37] Friedland, A. E. et al. Synthetic gene networks that count. *Science* **324**, 1199-202 (2009).

[38] Moon, T. S., Lou, C. B., Tamsir, A., Stanton, B. C. & Voigt, C. A. Genetic programs constructed from layered logic gates in single cells. *Nature* **491**, 249-253 (2012).

[39] Gaber, R. et al. Designable DNA-binding domains enable construction of logic circuits in mammalian cells. *Nat. Chem. Biol.* (2014). at `http://www.ncbi.nlm.nih.gov/pubmed/24413461`

[40] Cherry, J. L. & Adler, F. R. How to make a biological switch. *J. Theor. Biol.* **203**, 117-33 (2000).

[41] Nissim, L. & Bar-Ziv, R. H. A tunable dual-promoter integrator for targeting of cancer cells. *Mol Syst Biol* **6**, 444 (2010).

[42] Chen, J.H. & Seeman, N.C. Synthesis from DNA of a molecule with the connectivity of a cube. *Nature* **350**, 631-633 (1991)

[43] Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297-302 (2006).

[44] He, Y. et al. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* **452**, 198-201 (2008)..

[45] Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8 (2003).

[46] Regan, L. & DeGrado, W.F. Characterization of a helical protein designed from first principles. *Science* **241**, 976-978 (1988).

[47] Hecht, M.H., Richardson, J.S., Richardson, D.C. & Ogden, R.C. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science* **249**, 884-891 (1990).

[48] Woolfson, B. D. N. THE DESIGN OF COILED-COIL STRUCTURES AND ASSEMBLIES Abstract I . *Introduction to Protein Design.* **70**, 79-112 (2005).

[49] Fibers, R. et al. MagicWand?: A Single , Designed Peptide That Assembles to Stable , Ordered. (2008).

[50] Gradišar, H. et al. Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat. Chem. Biol.* **9**, 362-6 (2013).

De novo production of Turing activator generates polarity in pattern formation

Kartik Subramanian[1], Mark R. Paul[2] and John J. Tyson[3,4]

[1] Graduate Program in Genetics, Bioinformatics and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America,

[2] Department of Mechanical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America,

[3] Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America,

[4] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States of America

Abstract

Sixty years ago, Alan Turing showed that a system of reacting and diffusing chemicals could generate spontaneous, stable, stationary, spatial patterns of chemical concentrations. Since then "Turing patterns" have been used successfully to explain the development of patterns in both multicellular and unicellular organisms. In single-celled protists and bacteria, a common motif for spatial pattern formation is the autocatalytic production of filamentous protein polymers from globular protein subunits, which is an instance of the "activator-substrate depletion" (A-SD) mechanism of Turing patterning. A feature of A-SD models is that the peak of activator concentration tends to form at the center of the spatial domain rather than at the edges. Hence, a generic A-SD model cannot easily account for the generation of patterns with activator peaks at the poles of a cell. Here, we present a modified A-SD model that readily generates patterns with polar activator peaks. We explore the parameter combinations that determine whether the pattern has polar or central peaks. We propose that our mechanism may be used by living cells as a means to obtain polar localization of proteins.

1 Introduction

The impact of protein localization is evident at all levels of biological organization, from single cells to tissues, organs and whole organisms. Segregation of chromosomes during cell division, the positioning of flagella, and the specification of embryonic axes are some of the processes that depend on the timely placement of regulatory proteins at specific positions in the cell or organism. Advances in microscopy have dispelled the notion that protein localization is exclusive to the eukaryotic world. Bacteria such as *Escherichia coli*, *Bacillus subtilis* and *Caulobacter crescentus* all exhibit dynamic localization of specific proteins during the course of their cell division cycles . In addition to modern methods to visualize and identify these localized proteins, many experiments have sought the mechanisms behind dynamic protein localization. In many cases, a particular protein is found to localize by binding to a previously localized protein (an "upstream factor"), ultimately leading to the identification of "landmark" proteins that appear at specific locations in the cell in the absence of an upstream factor. How do landmark proteins spontaneously self-organize in the cell? Not surprisingly, a Turing-type mechanism has been proposed in many cases to explain autonomous pattern formation of landmark proteins .

The essence of Turing's idea was most clearly presented in a classical paper by Segel & Jackson , who showed that stable, stationary, spatial patterns can be generated in two-component reaction-diffusion systems under a few generic conditions: one component must be self-activating (autocatalytic) and slowly diffusing, and the other component—the fast-diffusing species—must inhibit the self-activation process in some way. In the Activator-Inhibitor Production (A-IP) mechanism, the activator produces an inhibitor that rapidly diffuses away from a zone of high activator concentration and restrains the self-activation process in surrounding regions. In the Activator-Substrate Depletion (A-SD) mechanism, the activator self-assembles from substrate molecules, which rapidly diffuse into the region of activator accumulation and thereby deplete the substrate from surrounding regions. In either case, the Turing instability generates patches of high activator concentration surrounded by regions of low activator concentration, either because inhibitor concentrations are high or substrate concentrations are low. A second activator patch can form only at some distance from the original patch, where inhibitor concentration is low enough or substrate concentration is high enough to allow for production of a new activator. This distance is the characteristic wavelength of the Turing pattern .

In this study we focus on pattern formation by A-SD models in growing domains of one spatial dimension (*i.e.*, a line segment, $0 \leq x \leq L(t)$, $\mathrm{d}L/\mathrm{d}t > 0$) to represent a growing, rod-shaped cell where x is the spatial coordinate and $L(t)$ is the length of the cell which varies with time t. A common feature of these patterns is that the first activator peaks forms at the center of the domain ($x = L/2$), and subsequent peaks bifurcate from previous peaks as the domain expands . Hence, the AS-D model fails to explain patterns that have activator peaks at the boundaries (the poles of the cell). To generate patterns with activator peaks at the poles, we add an additional reaction—*de novo* synthesis of activator—to the classical Gierer-Meinhardt reaction-diffusion equations (RDEs). Simulations of the modified RDEs exhibit patterns with polar maxima that are reminiscent of landmark protein localization in rod-shaped bacteria.

2 Methods

We model Turing patterns in one spatial dimension, x, by RDEs (Eq. 1-2) for activator concentration $a(x,t)$ and substrate concentration $s(x,t)$:

$$\frac{\partial a}{\partial t} = k_{\text{aut}} \cdot a^2 \cdot s - k_{\text{deg}} \cdot a + k_{\text{dnv}} \cdot s - k_{\text{as}} \cdot a + D_{\text{a}} \cdot \frac{\partial^2 a}{\partial x^2} \qquad (1)$$

$$\frac{\partial s}{\partial t} = k_{\text{syn}} - k_{\text{deg}} \cdot s - k_{\text{dnv}} \cdot s - k_{\text{aut}} \cdot a^2 \cdot s + k_{\text{as}} \cdot a + D_{\text{s}} \cdot \frac{\partial^2 s}{\partial x^2} \qquad (2)$$

The "classical Geirer-Meinhardt" equations lack the terms $k_{\text{dnv}} \cdot s$ and $k_{\text{as}} \cdot a$, which correspond to *de novo* synthesis of activator (polymer) from substrate (monomer) and to dissociation of polymer into monomers. We solve the RDEs (1-2) in one spatial dimension, $0 \le x \le L$, by the method of lines . That is, we discretize the spatial dimension into n compartments ($n = 100$) of length $h = L/n$, and approximate $\partial^2/\partial x^2$ by a central difference scheme (Eq. 3a). We implement no-flux boundary conditions as described in Eq. (3b,c). (In these equations, c_{i} represents the concentration of either activator or substrate in compartment *i.*)

$$\frac{\partial^2 c_i}{\partial x^2} = \frac{c_{i-1} - 2 \cdot c_i + c_{i+1}}{h^2} \qquad \textit{for } i = 2,\ldots, 99 \qquad (3a)$$

$$\frac{\partial^2 c_1}{\partial x^2} = \frac{c_2 - c_1}{h^2} \qquad (3b)$$

$$\frac{\partial^2 c_{100}}{\partial x^2} = \frac{c_{99} - c_{100}}{h^2} \qquad (3c)$$

In some cases we consider a growing spatial domain, $L = L(t)$, $\mathrm{d}L/\mathrm{d}t > 0$. In this case, we keep n fixed at 100 and let $h(t) = L(t)/n$ increase with growth

$$\frac{dh}{dt} = \mu \cdot h \tag{4}$$

The resulting 2*n*+1 ordinary differential equations are solved using MATLAB's ode15s solver. The parameter values used in our simulations are recorded in **Table 1**.

Table 1: Parameter set for the A-SD model*

$k_{\text{syn}} = 2\ \text{min}^{-1}$	$k_{\text{deg}} = 0.25\ \text{min}^{-1}$	$k_{\text{aut}} = 1.5\ \text{min}^{-1}$
$k_{\text{as}} = 1\ \text{min}^{-1}$	$D_{\text{s}} = 100\ \mu\text{m}^2.\ \text{min}^{-1}$	$D_{\text{a}} = 0.01\ \mu\text{m}^2.\ \text{min}^{-1}$
$\mu = 0.005\ \text{min}^{-1}$		

* The values of the *de novo* polymerization constant, k_{dnv} , are defined in the text and figure captions.

3 Results

3.1 De novo conversion of substrate to activator supports activator peaks at cell poles

A Turing-type RDE generates periodic activator peaks with a characteristic wavelength, λ_0 that depends on parameter values. If $L \approx \lambda_0/2$, the domain is large enough to accommodate a half-wave, with the activator maximum pinned at one end of the cell and the minimum at the other end. If $L(t)$ increases to ~ λ_0, the domain can accommodate either two activator half-peaks at the poles with a minimum in the center (polar peaks) or an activator peak at the center with minima at the poles (polar troughs). As $L(t)$ increases further, additional activator peaks are generated, but the patterns can still be classified as either "polar peaks" or "polar troughs". In A-IP models, either pattern is possible . In classical A-SD models, however, only polar trough patterns are observed.

In the classical A-SD model, Eq. (1-2) with $k_{\text{dnv}} = k_{\text{as}} = 0$, the only route to new activator production is the autocatalytic term, $k_{\text{aut}} \cdot a^2 \cdot s$. Hence, new activator tends to be produced where activator concentration is the highest, namely at the peak. However, the substrate necessary for activator production is more abundant at some distance from the peak. Hence, the activator peak tends to grow into the surrounding zone, where more substrate is available, and a depression tends to form in the center of the peak, where substrate concentration is lowest. When the activator peak reaches the center of the domain, substrate availability is equal on both sides of the peak, and the peak is anchored there. For this reason, classical A-SD models are unsuccessful in explaining patterns with polar peaks of landmark proteins.

To account for polar activator peaks, previous groups have modified the classical A-SD model in various ways. For the MinCD/DivIVA system in *B. subtilis*, Howard proposed that increased curvature of the cell's boundary at the poles might account for polar localization. To explain growth zone formation in fission yeast, Csikasz-Nagy *et al*. introduced a polarly localized nucleating factor that stimulates the autocatalytic production of activator from substrate at the poles. While these modifications are reasonable and yield appropriate results, they are case-specific. Examples of polar peaks in other biological systems would presumably need other case-specific explanations. We seek a more generic account of polar activator peaks in A-SD models.

Since the autocatalytic term biases A-SD models to produce central activator peaks, we conjectured that an alternative source of activator production might create a different pattern. Hence, we introduced the *de novo* activator production term, $k_{\mathrm{dnv}} \cdot s$, into Eq. (1-2). This term, which implies that new activator can be produced from substrate independent of pre-existing activator, is reasonable both chemically and biologically. The *de novo* term describes the spontaneous production of polymers (activator) from a handful of monomers (substrates), while the autocatalytic term describes the extension of pre-existing polymers by the addition of new monomeric subunits. Our model also contains a term, $k_{\mathrm{as}}{\cdot}a$ for the depolymerization of activator back to substrate monomers.

To test the impact of the *de novo* term, we simulated a cell of length $L \approx \lambda_0$, *i.e.*, long enough to accommodate either a central peak or central trough pattern. For the classical A-SD model, $k_{\mathrm{dnv}} = k_{\mathrm{as}} = 0$, we expected to find central peak patterns exclusively. By increasing the strength of *de novo* synthesis, we expected to find examples of spontaneous, stable polar peaks. We initialized the model with an excess of activator at one pole. When k_{dnv} is small, the activator peak moved towards the center (**Figure 1A and B**). When we initialized the model with excess activator at the center, the peak stayed in place (**Figure 1C**). Hence, for small values of k_{dnv}, our modified RDE behaves exactly as the classical A-SD model. For larger values of k_{dnv}, we saw that, for some initial conditions, the activator peak was no longer exclusively central. If the activator was initialized with a central bias, then the peak remained at the center (**Figure 2C**). However, for simulations that were initialized with an excess of activator at one pole, we observed activator half-peaks not only at the end with the initial bias but also at the other end (**Figure 2A and B**). While one of the half-peaks clearly arises from the initial bias, the other appears spontaneously, since the cell is long enough to hold two half-peaks at opposite poles.

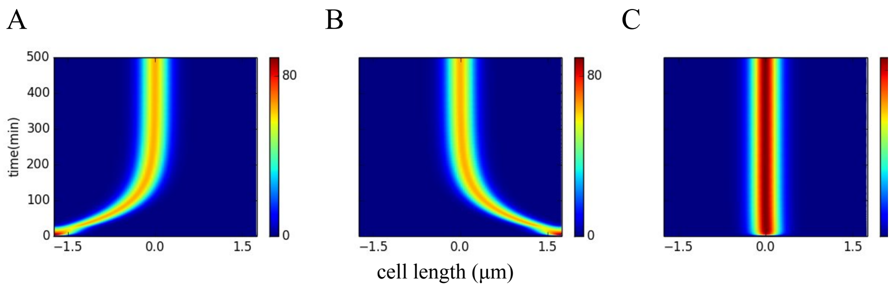


Figure 1: Space-time plot of activator concentration for varying initial conditions of the activator and k_{dnv} = 1 min^{-1}. Initial activator bias is provided by setting activator concentration as 1 dimensionless unit in one compartment and 0 dimensionless units in all other compartments. The color bars indicate the concentration of the activator. **(A)** Initial activator bias at the left pole *i.e.* $a(1,0)$ = 1 dimensionless unit. **(B)** Initial activator bias at the right pole *i.e.* $a(100,0)$ = 1 dimensionless unit. **(C)** Initial activator bias in the middle *i.e.* $a(51,0)$ = 1 dimensionless unit.

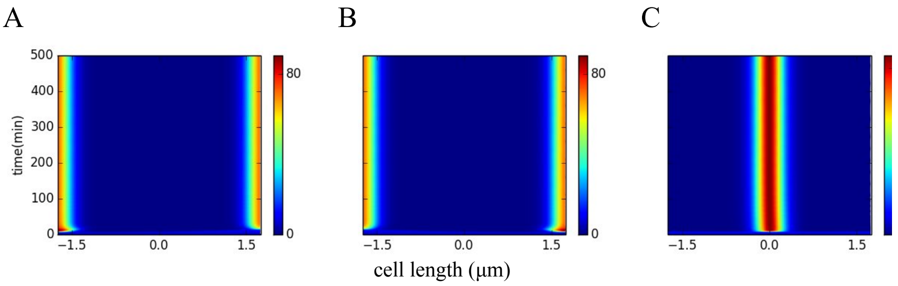


Figure 2: Space-time plot of activator concentration for varying initial conditions of the activator and k_{dnv} = 10 min^{-1}. Initial activator bias is provided by setting activator concentration as 1 dimensionless unit in one compartment and 0 dimensionless units in all other compartments. The color bars indicate the concentration of the activator. **(A)** Initial activator bias at the left pole *i.e.* $a(1,0)$ = 1 dimensionless unit **(B)** Initial activator bias at the right pole *i.e.* $a(100,0)$ = 1 dimensionless unit. **(C)** Initial activator bias in the middle *i.e.* $a(51,0)$ = 1 unit.

3.2 Central and polar peaks are alternative attractors

Irrespective of the value of k_{dnv}, a central activator peak was obtained if the initial bias was at the center, suggesting that the location of the activator peak is sensitive to initial conditions. To determine the effects of initial conditions on pattern formation, we simulated our model with random sets of initial conditions. In **Figure 3A**, we plot the final distribution of activator for 500 simulations, each initialized with different initial conditions selected randomly. For $k_{\text{dnv}} = 0$, only 2 of 500 runs showed polar activator peaks. In these two cases, we suspect, both ends of the cell were initialized with nearly equal activator concentrations. When k_{dnv} was increased from 0 to 1 or 10, a significantly larger number of runs (≈ 40%) showed polar peaks (**Figure 3B and C**). Further increase in the value of k_{dnv} showed moderate increase in polar peaks, indicating that central-peak patterns are dominant over polar-peak patterns. Importantly, both patterns are attractors, each with its own basin of attraction. The effect of increasing k_{dnv} is to increase the basin of attraction for the polar attractor, albeit only to a limited extent.

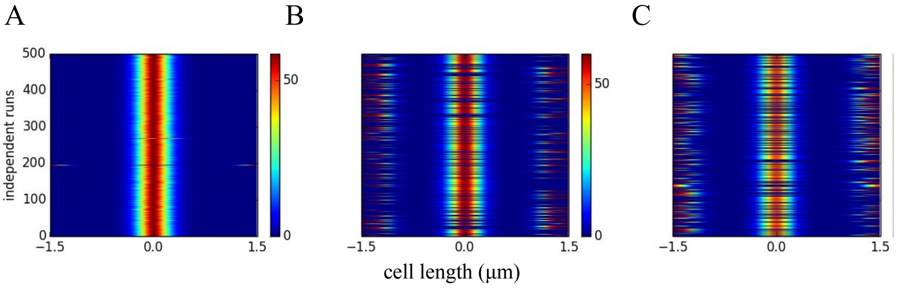


Figure 3: Final activator distribution (at t = 500 min) plotted for 500 independent runs, each with random initial conditions (at t = 0) of activator and substrate, ranging from 0 to 1 dimensionless units. The color bars indicate the concentration of the activator. **(A)** $k_{\text{dnv}} = 0$ min^{-1}. **(B)** $k_{\text{dnv}} = 1$ min^{-1}. **(C)** $k_{\text{dnv}} = 10$ min^{-1}.

3.3 Asymmetric pre-patterns lead to polar activator peaks above a critical value of k_{dnv}

As our simulations show, varying initial conditions give rise to different final patterns. How then can the model explain robust polar localization of landmark proteins? In most cases, cells inherit a pre-pattern or asymmetric polarity from the parent cell. We therefore investigated if polar peaks are robustly inherited from asymmetric pre-patterns, *i.e.* from an initial pattern with a peak at one pole and a trough at the other. For varying values of k_{dnv} , we simulated our model starting with an initial polar bias at one end of the cell. **Figure 4A** shows the final distribution of activator obtained for 1000 values of k_{dnv} ranging from 5.5 to 6.5. For each value of the parameter, the simulation was initialized with an initial activator peak at the right pole *i.e.* $a(100,0) = 1$ unit. Below a critical value of $k_{\mathrm{dnv},}$ the final distribution pattern shows a central peak, while above the critical value, two half-peaks are obtained at the poles. The results indicate that, given an asymmetric pre-pattern, polar activator peaks are robustly inherited above a critical value of $k_{\mathrm{dnv}} = 5.95$, On the other hand, if the pre-pattern is symmetric, then the corresponding final pattern is also symmetric, irrespective of the value of $k_{\mathrm{dnv}.}$ If the simulation is initialized with bias at both poles, then the corresponding final distribution pattern shows polar half-peaks for all values of k_{dnv} (**Figure 4B**). If the pre-pattern at $t = 0$ min has a central bias, then the corresponding final pattern also show central peaks, even above critical value of k_{dnv} (**Figure 4C**). Hence, above a critical rate of *de novo* activator synthesis, our model can account for the generation of polar peaks if the pre-pattern is asymmetric.

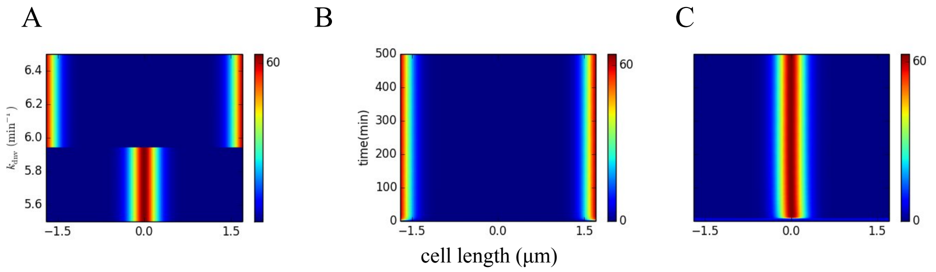


Figure 4: **(A)** Final activator distribution plotted for 1000 values of k_{dnv} between 5.5 and 6.5 min^{-1}. The initial pattern is asymmetric, with an activator peak at one end and trough at the other *i.e.* $a(100,0) = 1$ dimensionless unit. **(B)** Space-time plot of activator distribution for $k_{\mathrm{dnv}} = 5.5$ min^{-1} and for an initial bias at both poles *i.e.* $a(1,0) = a(1,0) = 1$ dimensionless unit. **(C)** Space-time plot of activator distribution for $k_{\mathrm{dnv}} = 6.5$ min^{-1} and the initial bias is in the middle of the cell *i.e.* $a(51,0) = 1$ dimensionless unit. The color bars indicate the concentration of the activator.

The condition for an asymmetric polar pre-pattern is easily realized in the relevant biological context. Imagine a cell with activator peaks at both poles. Upon cell division, each daughter cell inherits one of the polar peaks and a trough at the other end, which was the central trough in the mother cell. Newly born daughter cells with such a pre-pattern will give rise to a second activator peak at the opposite pole when they grow long enough to accommodate a second peak. The cycle of small daughter cells with a single polar peak growing into pre-divisional cells with two polar peaks will continue in each generation.

3.4 In growing cells, new activator peaks appear spontaneously

Our simulations up to this point have been for cells of fixed size where L is a constant. In reality though, cells grow and divide over time and therefore $L = L(t)$. The effect of cell growth on pattern formation in a classical A-SD model was previously studied by Crampin *et al.* . Additional activator peaks emerge when the size of the cell is a multiple of the characteristic wavelength of the Turing pattern. The transitions in the activator pattern have a characteristic feature: as cell size approaches a new multiple of λ_0, a pre-existent activator peak splits into additional activator peaks. Following this transition, the recently split activator peaks move away from each other as the cell continues to grow. This process of peak splitting occurs every time the cell becomes large enough to hold additional peaks, ultimately giving rise to a tree-like or branched activator pattern. For our set of equations, we found a similar branched pattern for low values of k_{dnv} (**Figure 5A**).

In sharp contrast, for larger values of the rate of *de novo* synthesis, we found that additional activator peaks appeared at locations where no peak existed previously (**Figure 5B**). Transitions in the number of activator peaks also occurred when cell size crossed a multiple of a critical length. Another significant deviation from the branched pattern is the number of peaks that appear at each transition. For the branched pattern, a cell of length $L \approx \lambda_0/2$ holds one half-peak at a pole. When length $L \approx \lambda_0$, a single peak is found at the center of the cell. Thereafter, the number of peaks doubles in subsequent transitions, giving rise to 2, 4, 8 and 16 peaks respectively (**Figure 5A**). For the case of *de novo* activator production in our model, the first transition, at $L \approx \lambda_0$, gives rise to two half-peaks at the poles. In the next transition, a third peak forms at the center. Subsequent transitions have 5, 9 and 17 peaks respectively. Hence, after the initial two transitions, the branched pattern shows an even number of peaks, while the *de novo* pattern shows an odd number of peaks. We propose that an odd number of activator peaks is a characteristic feature of an A-SD model with *de novo* synthesis of activator.

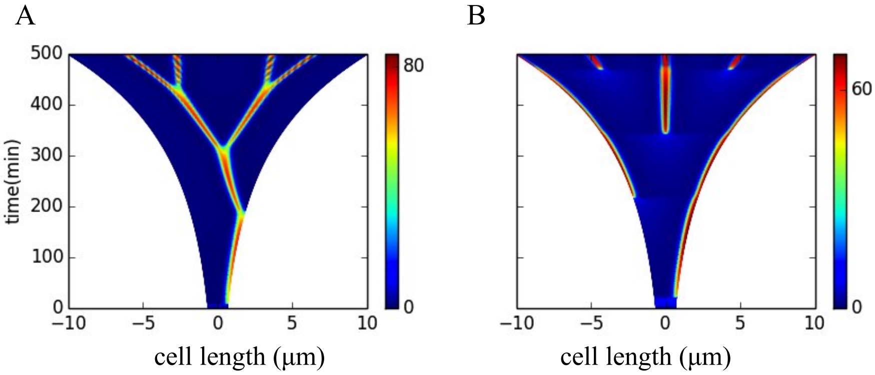


Figure 5: Space time plot of activator production in a growing cell, $L(t = 0) = 1.3$ μm initialized with a bias of activator at the right pole *i.e.* $a(100,0) = 1$ dimensionless unit. Activator concentration in other compartments is set to 0 dimensionless units. The color bars indicate the concentration of the activator. **(A)** $k_{\mathrm{dnv}} = 5$ min^{-1}. **(B)** $k_{\mathrm{dnv}} = 30$ min^{-1}.

At $t = 500$ minutes, the cell length $L = 100$ μm and compartment size $h = 0.1$ μm. A concern at large L is that the discretization may not be smooth enough. We therefore discretized the domain into 200 compartments and repeated the simulations. We found that the activator patterns obtained from the two discretization schemes were comparable to each other.

4 Conclusions

The classical A-SD model, proposed years ago by Gierer and Meinhardt, exhibits activator peaks that localize away from the poles of a one-dimensional field. We present a modified set of RDEs that includes additional terms for *de novo* production of activator and conversion of activator back to substrate molecules. We show that our modified RDEs can account for patterns with activator peaks at the poles.

How does the addition of the *de novo* synthesis term account for the phase shift in the activator pattern? When the cell is small, production of new activator peaks is inhibited due the rapid diffusion of substrate molecules into the activator peak, depleting the surrounding region of substrate. Only when the cell grows sufficiently long may a second activator peak be formed. In the classical A-SD model, activators are produced only by an autocatalytic route. Hence, new peaks are obtained by splitting a pre-existing peak, resulting in a branched pattern. In contrast, the *de novo* synthesis term in our model allows the production of activators at any site. Therefore, as the cell grows, new activator peaks may be produced in between pre-existing peaks, where substrate concentration is largest. The autocatalytic and *de novo* terms compete for activator production. At low k_{dnv} , the autocatalytic term dominates, and we observe movement of the first peak towards the center of the cell, followed by peak splitting in subsequent transitions. When k_{dnv} is large enough, *de novo* activator synthesis takes over, allowing the formation of new peaks between old ones. We propose that this mechanism, combining autocatalytic activator production with a *de novo* synthesis term, may be a means used by cells for polar localization of polymeric proteins.

Acknowledgements

This work was supported by a grant from the National Science Foundation, USA (DMS-1225160). The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Howard M (2004) A mechanism for polar protein localization in bacteria. J Mol Biol 335: 655–663.

2. Shapiro L, Losick R (1997) Protein localization and cell fate in bacteria. Science 276: 712–718. doi:10.1126/science.276.5313.712.

3. Johnson AS, van Horck S, Lewis PJ (2004) Dynamic localization of membrane proteins in Bacillus subtilis. Microbiology 150: 2815–2824. doi:10.1099/mic.0.27223-0.

4. Howard M, Kruse K (2005) Cellular organization by self-organization: mechanisms and models for Min protein dynamics. J Cell Biol 168: 533–536.

5. Hale CA, Meinhardt H, de Boer PA (2001) Dynamic localization cycle of the cell division regulator MinE in Escherichia coli. EMBO J 20: 1563–1572.

6. Kondo S, Miura T (2010) Reaction-diffusion model as a framework for understanding biological pattern formation. Science 329: 1616–1620. doi:10.1126/science.1179047.

7. Meinhardt H (1982) Models of biological pattern formation. Research Gate. pp. 1–10. doi:10.1016/S0070-2153(07)81001-5.

8. Meinhardt H (1982) Models of Biological Pattern Formation. Academic Press.

9. Meinhardt H, Gierer A (2000) Pattern formation by local self-activation and lateral inhibition. Bioessays 22: 753–760. doi:10.1002/1521-1878(200008)22:8<753::AID-BIES9>3.0.CO;2-Z.

10. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. Kybernetik 12: 30–39. doi:10.1007/BF00289234.

11. Turing AM (1952) The Chemical Basis of Morphogenesis. Philos Trans R Soc B Biol Sci 237: 37–72. doi:10.1098/rstb.1952.0012.

12. Segel LA, Jackson JL (1972) Dissipative structure: an explanation and an ecological example. J Theor Biol 37: 545–559. Available: http://www.ncbi.nlm.nih.gov/pubmed/4645361.

13. Edelstein-Keshet L (2005) Mathematical Models in Biology. SIAM.

14. Crampin EJ, Hackborn WW, Maini PK (2002) Pattern formation in reaction-diffusion models with nonuniform domain growth. Bull Math Biol 64: 747–769.

15. Schiesser WE (1991) The Numerical method of lines. Academic Press.

16. Csikász-Nagy A, Gyorffy B, Alt W, Tyson JJ, Novák B (2008) Spatial controls for growth zone formation during the fission yeast cell cycle. Yeast 25: 59–69. doi:10.1002/yea.1571.

Systematic study of a metabolic network

Bertrand Beauvoit[1], Sophie Colombié[1], Jean-Pierre Mazat[2], Christine Nazaret[3], Sabine Pérès[4] (Alphabetic order)

[1]INRA, UMR 1332 Biologie du Fruit et Pathologie, F33883 Villenave d'Ornon, France

Univ. Bordeaux 146 rue Léo-Saignat, F 33076 Bordeaux Cedex, France

[2]IBGC CNRS UMR 5095 & Université de Bordeaux, 1, rue Camille Saint-Saëns 33077 BORDEAUX-cedex, France e-mail.

[3] Institut de Mathématiques de Bordeaux, ENSTBB-Institut Polytechnique de Bordeaux .

[4]Laboratoire de Recherche en Informatique (LRI) Bât 650 Université Paris-Sud 11 91405 Orsay Cedex France

Abstract - Introduction

On an example, a simple version of Krebs cycle with transporters/exchangers between mitochondrion and cytosol, we will systematically apply, in a logical order, several theoretical approaches for metabolism. We will emphasize the advantages and the drawbacks of each approach and show how we can use all of them to gain a better understanding of the behavior of a metabolic network.

Metabolic Network. What is a metabolic network ?

Cells house a great number of chemical reactions, which split the nutriment we eat in smaller molecules and produce energy (ATP molecules for instance) from the oxygen we breathe in (catabolism).

From these molecules other reactions take place to synthesize the basic molecules of the cell, such as amino-acids, nucleotides and nucleosides, fatty acids, etc.(anabolism). With these elementary molecules, the cell can synthesize several macromolecules (protein, nucleic acid phospholipids etc.).

As a matter of fact, organisms are built almost entirely from water and about thirty small precursor molecules (amino acids, aromatic bases of nucleic acids, sugars, palmitate, glycerol and choline).

Metabolism is an open system. The entries are the nutriments we eat and the oxygen we breathe in; the output are the $CO_2$and H_2Owe breathe out and metabolites excreted from the cell.

An example of metabolic network: the simplified Krebs cycle with some transporters.

As an example, we will take a simplified version of the Krebs cycle represented in Fig. 1 [1]in which the three classical steps :
K5: AKG + NAD + Pi + ADP <=> Succinate + NADH + ATP . (α-keto-glutarate dehydrogenase).
K6: Succinate + FAD <=> fumarate + FADH2 .   (succinate dehydrogenase) .
K7: fumarate <=> malate. (fumarase).
have been lumped together in
K567: AKG + NAD + Pi + ADP<=> malate + NADH2 + ATP .

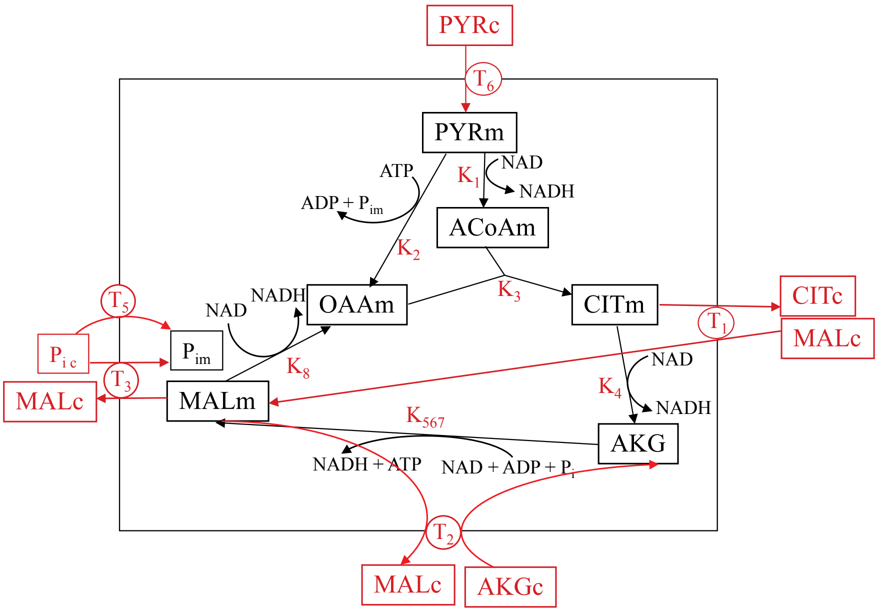


Figure 1. The metabolic network KS1

The reactions of this network can be written, with obvious abbreviations. “m” means mitochondrial and “c” cytosolic. CO_2 and H_2O have been omitted in the reactions.
K1 : PYRm + NADm + CoAm = ACoAm + NADH2m . (Pyruvate dehydrogenase with production of 1CO2.).
K2 : PYRm + ATPm = OAAm + Pim + ADPm . (Pyruvate carboxylase with consumption of 1CO2).
K3 : OAAm + ACoAm = CITm + CoAm . (Citrate synthase).
K4 : CITm + NADm = AKGm + NADH2m. (Aconitase + Isocitrate dehydrogenase).
K567 : AKGm + NADm + Pim + ADPm = MALm + NADH2m

+ ATPm (see above).
K8 : MALm + NADm = OAAm + NADH2m . (malate dehydrogenase).
T1 : CITm + MALc = CITc + MALm . (Malate/Citrate exchanger).
T2 : AKGc + MALm = AKGm + MALc . (Malate/ α-keto-glutarate exchanger).
T3 : MALm + Pic = MALc + Pim . (Malate/ Phosphate exchanger).
T5 : Pic = Pim . (Pi carrier)
T6 : PYRc = PYRm . (Pyruvate transporter).

In this open system, the internal metabolites have been framed in black. Their concentrations are the variables of the system. All is as if the concentrations of all other metabolites were constant. They are colored in red (external metabolites).

The stoichiometry matrix.

All metabolic network with *m* internal metabolites and *r* reactions can be represented by a stoichiometry matrix N of *m* rows and *r* columns. Its coefficients n_{ij} represents the number of molecules *i* consumed or produced by the reaction *j*. The coefficients are negative for the substrates and positive for the products. Note that the direction of the reaction is arbitrary. Only internal metabolites are counted.

	K1	K2	K3	K4	K567	K8	T1	T2	T3	T5	T6
PYRm	-1	-1	0	0	0	0	0	0	0	0	+1
OAAm	0	+1	-1	0	0	+1	0	0	0	0	0
Pim	0	0	0	0	0	0	0	0	+1	+1	0
CITm	0	0	+1	-1	0	0	-1	0	0	0	0
AKGm	0	0	0	+1	-1	0	0	+1	0	0	0
MALm	0	0	0	0	+1	-1	+1	-1	-1	0	0
ACoAm	1	0	-1	0	0	0	0	0	0	0	0

Table 1: Stoichiometry matrix of the network KS1

From the stoichiometry matrix it is possible to build the metabolic network. There is a unique stoichiometry matrix associated to a metabolic network and *vice versa*.
Exercise: draw the metabolic network associated to the following stoichiometry matrix:
[1 - 1]; [1 - 1 -1]; $\begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ (first you can write each reaction corresponding to each column then associate the reactions to constitute the network)

The Rank of the stoichiometry matrix.

It is an important concept from the mathematical point of view but also from the biochemical point of view, because it is the number of independent concentrations (variables). A variable is independent of another if it cannot be expressed as a function of the other.
Counter-example. NAD + $NADH_2$ = Nt (constant) so that NAD = Nt – $NADH_2$ and NAD and $NADH_2$ are not independent. If we add NAD and $NADH_2$ as variable, the stoichiometry matrix becomes:

	K1	**K2**	**K3**	**K4**	**K567**	**K8**	**T1**	**T2**	**T3**	**T5**	**T6**
PYRm	-1	-1	0	0	0	0	0	0	0	0	+1
OAAm	0	+1	-1	0	0	+1	0	0	0	0	0
Pim	0	0	0	0	0	0	0	0	+1	+1	0
CITm	0	0	+1	-1	0	0	-1	0	0	0	0
AKGm	0	0	0	+1	-1	0	0	+1	0	0	0
MALm	0	0	0	0	+1	-1	+1	-1	-1	0	0
ACoAm	1	0	-1	0	0	0	0	0	0	0	0
NAD	-1	0	0	-1	-1	-1	0	0	0	0	0
$NADH_2$	+1	0	0	+1	+1	+1	0	0	0	0	0

Table 2: Stoichiometry matrix comprising NAD and $NADH_2$ as new variable with NAD + $NADH_2$ = Nt

It is easy to see that (row NAD) + (row $NADH_2$) = 0 which corresponds to NAD + $NADH_2$ = Nt (in fact to d(NAD)/dt +d($NADH_2$)/dt = 0 (see below).
From the mathematical point of view, a column (resp. row) is independent of another column (resp. row) if it cannot be expressed from the other. For instance, K2 is independent of K1, because the stoichiometry factor of OAAm in K1 is zero which will never make +1 in K2(in other words, it is impossible to find a factor $\lambda \neq 0$ such that K2 = λ.K1). With the same reasoning one can see that K1 is independent of K3 and K4. In the same way it is easy to see that T5 is independent of all other vectors columns. On the contrary, because the column T2 = - column K567, T2 and K567 are not independent.
We will not go further in the calculation of the rank which is a complicated problem; let us give only some simple rules. The rank is ≤ the lower dimension of row and column. It means that in our example the rank ≤ 7, the number of rows. In our example it is easy to find 7 independent columns K1, K2, K3, K4, K567, K8 and T5, so that the rank is exactly 7 and all other

columns can be necessarily expressed from these seven. For instance, T3 = -K1 + K2 + K4 –K567 + T5.
Exercise: Express T1, T2 and T6 as the function of the previous seven independent columns vectors, K1, K2, etc..

Steady-state.

The time course of the metabolites concentrations can be described by differential equations.
The variations of the metabolites are:
d(PYRm)/dt = -VK1 –VK2 +VT6

……..
d(OAAm)/dt = VK2 – VK3 + VK8
or, in matrix notation : d(X)/dt = N.V, where

$$[X] = \begin{bmatrix} PYRm \\ OAAm \\ Pim \\ CITm \\ AKGm \\ MALm \\ ACoAm \end{bmatrix} \quad \text{and} \quad [d(X)/dt] = \begin{bmatrix} d(PYRm)/dt \\ d(OAAm)/dt \\ d(Pim)/dt \\ d(CITm)dt \\ d(AKGm)/dt \\ d(MALm)/dt \\ d(ACoAm)/dt \end{bmatrix} \text{and}$$

$$V = \begin{bmatrix} VK1 \\ VK2 \\ VK3 \\ VK4 \\ VK567 \\ VK8 \\ VT1 \\ VT2 \\ VT3 \\ VT5 \\ VT6 \end{bmatrix} \quad \text{N is the matrix of table 1.}$$

The steady- state is defined by d(X)/dt = 0, i.e. N.V = 0 which gives the relationships between the rates equation necessary to obtain a steady state. For instance the product of the first row N with V gives: -VK1 –VK2 +VT6 = 0 which insure the steady state of [PYRm] *etc.*
It is thus important to know all the vectors V such that N.V = 0. They form a vector space called **Null Space** of N or **Kernel** of N. We will denote it Ker(N). It is a vector space because if V’ and V” ∈Ker(N) (meaning that N.V’=0 and N.V” = 0) then V’+V”∈Ker(N), because N(V’+V”) = N.V’ + N.V” = 0. In the same way,λV’∈Ker(N). Its dimension is dim (V) – rank(N) = 4 in our case (11 – 7).

It is easy to find 4 independent vectors in Ker(N). (If you do not know the concept of "vector space" thinks, as an example in 2 dimensions geometry, of the vectors from the origin, see Fig. 2).

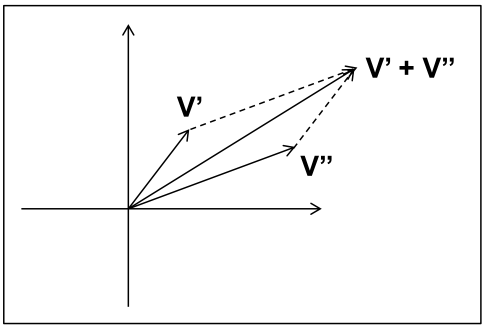


Figure 2: Vectorial space of vectors from the origin.

Exercise: Write the relationships between the rate at steady-state:
-VK1 –VK2 +VT6= 0
VK2-VK3 + VK8 = 0
……………………)
Use these relationships to derive 4 independent vectors of Ker(N).

Metatool (http://pinguin.biologie.uni-jena.de/bioinformatik/networks/) can be used to obtain a basis of the Kernel (see appendix 1 and 2). The entry file of Metatool (appendix 1) organizes the reactions in reversible and irreversible ones (ENZREV and ENZIRREV) and metabolites in internal and external metabolites (METINT and METEXT).(Anotice of Metatool is given in http://solea.quim.ucm.es/t4m/Manual_T4M.pdf).
The output file of Metatool (appendix 2) gives the stoichiometry matrix with first the reversible then irreversible reactions in the order of the input file, then a basis of the Kernel (entitled KERNEL), under the form of a matrix, the rows of which are the basis vectors. Because these vectors verify N.V = 0, they are possible pathways at steady-state. They are represented in Fig. 3. All the solutions at steady-state are linear combinations of these vectors : V = λ_1.Ker1+ λ_2.Ker2 +λ_3.Ker3 +λ_4.Ker4 .
However the vectors of the basis of the Kernel do not necessarily obey the irreversibility of some reactions. For example, Ker1 = (-2 -1 0 1 -2 -1 1 -1 0 0 0) means that vK1 = -1 (sixth value) while this reaction being irreversible its rate cannot be negative (indicating that it is taken in the opposite direction).

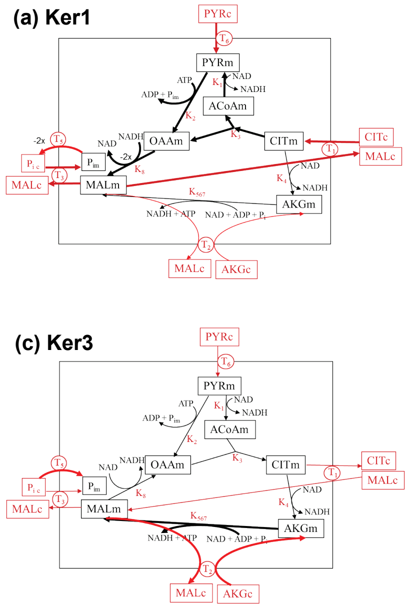


Figure 3: A basis of Ker(N)

Subsets of reactions (Enzyme subsets)

At steady state, an enzyme subset is a set of reactions which must always operate together with a fixed ratio in their rates. In our example, K1 K3 is a subset of reactions because when acetyl-CoA is formed by K1 its only use is through K3. The association of T3 with K2 is less obvious. It comes from the fact that the only output of the network are with MAL (T3) or through MAL exchange incorporating a molecule of MAL which has to go out through T3. This explains the association of T3 with K2 which is not a priori obvious (instead of the association of K1 with K3). It means that the flux through K2 will be always equal to the flux through T3. This constraint has to be kept in mind when examining the results of experiments. The subsets enable the reduction of the size of the system by lumping in one reaction all the reactions belonging to the same subset. It is exactly what we did at the beginning by lumping together the reactions K5, K6 and K7 in K567. The reactions which are not in a subset of several reactions remain isolated in the list of "subset".

Convex Basis

Within a system, several restrictions can appear. These may be due, for example, to irreversibility of the reactions. These restrictions limit the solution space, which becomes a polyhedral convex cone (Fig. 4 in the case of 3 reactions). In these conditions, any solution which is a vector of the convex cone can be expressed as a linear combination with non-null coefficients of the vectors representing the edges of the convex cone. Thus the basis of the convex cone are more interesting because they obey the restrictions imposed to the system which is not necessarily the case of the basis given above for Ker(N). The convex cone is included in Ker(N).

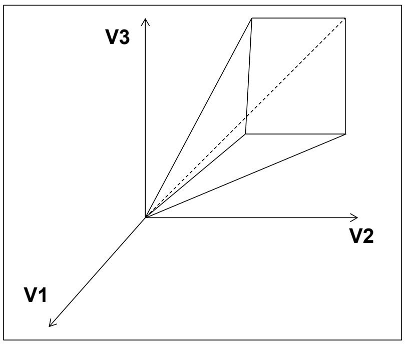


Figure 4: The convex cone of solutions.

In the case of KS1, Metatool gives the convex basis (Fig. 5):
C1: -T1 -T2 K4 irreversible = EFM1(see below. It is the only one of the Ci containing K4)
C2: T2 T5 K567 irreversible = EFM2 (see below. It is the only one of the Ci containing K567)
C3: K8 T1 K1 K3 T6 irreversible = EFM4 (see below. It is the only one of the Ci containing K1)
C4: -K8 T3 (-2 T5) K2 T6 irreversible = Ker4= EFM6 (see below. It is the only one of the Ci containing T3).

In this simple example the decomposition on the convex basis is easy if it is noticed that C1 is the only vector of the convex basis involving K4, C2 the only vector of the convex basis involving K567, C3 the only vector of the convex basis involving K1 and C4, the only vector of the convex basis involving T3. We have justto note the coefficient of these reactions to obtain the decomposition.

For instance the basis of the Kernel can be expressed on the convex basis:
Ker1 =-C3 + C4; Ker2 = - C1 ; Ker3 = C2 and Ker4 = C4 .

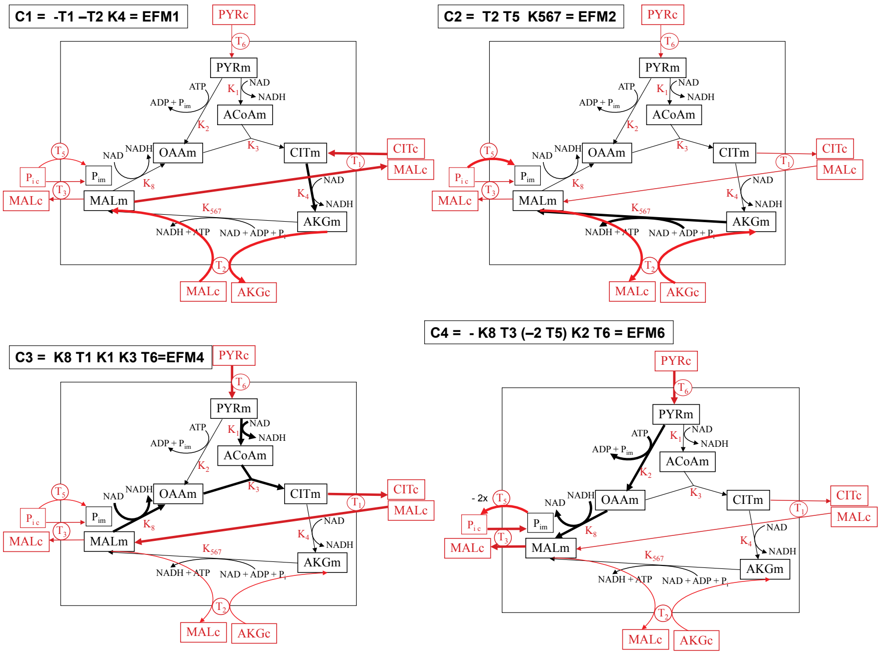


Figure 5: The basis of the convex cone

Elementary Modes

An **elementary flux mode (EFM)**[2-3]is a minimal set of enzymes that can operate at steady state with all irreversible reactions used in the appropriate direction. All flux distributions in the living cell are non-negative linear combinations of elementary modes. The decomposition is not necessarily unique.

A related concept was defined by the group of Palsson: **Extreme pathway** [4] in which every reversible internal reaction is split in two irreversible reactions (one is the forward reaction, the other is the reverse reaction).The number of EFMs is finite but can be great. Their comprehensive description gives all possibilities to browse the metabolic network. As we will see below, some of them are not trivial. In Metatool, the EFMs are given as rows of a matrix. Their description is given in table 3 and in the figures of Appendix 3. Among them we find the classical Krebs cycle (EFM 9). But they also evidence pathways which are less obvious (EFM 11 for instance)[5]. All the EFM can be decomposed on the convex basis.

Considering EFM is useful because they represent extreme situations of metabolic pathways: in EFM1 for instance there is only one metabolite entering, with AKG as the only output; in EFM2 there are two entering metabolites, Pi and AKG and only one going out, MAL. In the reality the situation is usually more complex with all metabolites going in or out, but with different extent. For each EFM, several indexes can be calculated such as the yield in carbon for ATP synthesis, the inputs and outputs etc. An actual set of fluxes at steady-state in a metabolic network can be interpreted as combination of some EFMs. It is particularly useful for understanding the passage between a flux pattern to another at steady-state [6].

EFM	*K8*	*T1*	*T2*	*T3*	*T5*	*K1*	*K2*	*K3*	*K4*	*K567*	*T6*
1	0	-1	-1	0	0	0	0	0	1	0	0
2	0	0	1	0	1	0	0	0	0	1	0
3	0	-1	0	0	1	0	0	0	1	1	0
4	1	1	0	0	0	1	0	1	0	0	1
5	1	0	-1	0	0	1	0	1	1	0	1
6	-1	0	0	1	-2	0	1	0	0	0	1
7	0	1	0	1	-2	1	1	1	0	0	2
8	0	0	-1	1	-2	1	1	1	1	0	2
9	1	0	0	0	1	1	0	1	1	1	1
10	0	0	0	1	-1	1	1	1	1	1	2
11	-1	0	2	1	0	0	1	0	0	2	1
12	-1	-2	0	1	0	0	1	0	2	2	1
13	1	0	0	1	0	2	1	2	2	2	3
14	0	1	2	1	0	1	1	1	0	2	2
15	0	-1	0	1	0	1	1	1	2	2	2
16	0	0	1	1	0	1	1	1	1	2	2

C1	0	-1	-1	0	0	0	0	0	1	0	0
C2	0	0	1	0	1	0	0	0	0	1	0
C3	1	1	0	0	0	1	0	1	0	0	1
C4	-1	0	0	1	-2	0	1	0	0	0	1

Table 3*:*Description of the EFM and of the convex basis.

Flux Balance Analysis

Flux Balance Analysis (FBA) is a method developed by the group of Palsson [7-8] aiming at optimizing the flux values in a metabolic network to fulfill a peculiar objective such as cell growth or ATP production *etc*... The objective is mathematized under the form of an "objective function" (a rate equation of ATP consumption in the case of optimizing ATP production). Known constraints on the fluxes can be added such as minimal and maximal values (otherwise maximizing a flux will lead to infinity).FBA can be applied in many other contexts to analyze the phenotypes and capabilities of organisms upon different environmental and genetic perturbations (KO genes for instance). The optimization is directed at the metabolic fluxes values, without any knowledge of the underlying rate functions so that it can be applied to big genome scale networks for which not all steps are known in detail.

We will give some applications of FBA to our metabolic network. We will suppose that all fluxes are between -1 and +1 for the reversible ones and between 0 and +1 for the irreversible ones. We will express the fluxes at steady state as a function of the convex basis. In this way we are sure that they satisfy N.F = 0.A vector F can be decomposed in F = $\lambda_1.C_1 + \lambda_2.C_2 + \lambda_3.C_3 + \lambda_4.C_4$. (with $0\leq \lambda_i$).

The components of the vector F on the convex basis are given in table 4:

	C1	C2	C3	C4	Fluxes
K8	**0**	**0**	**1**	**-1**	λ_3-λ_4
T1	**-1**	**0**	**1**	**0**	-λ_1+λ_3
T2	**-1**	**1**	**0**	**0**	-λ_1+λ_2
T3	**0**	**0**	**0**	**1**	λ_4
T5	**0**	**1**	**0**	**-2**	λ_2-2λ_4
K1	**0**	**0**	**1**	**0**	λ_3
K2	**0**	**0**	**0**	**1**	λ_4
K3	**0**	**0**	**1**	**0**	λ_3
K4	**1**	**0**	**0**	**0**	λ_1
K567	**0**	**1**	**0**	**0**	λ_2
T6	**0**	**0**	**1**	**1**	λ_3+λ_4

Table 4. Decomposition of a vector on the convex basis. F = $\lambda_1.C_1 + \lambda_2.C_2 + \lambda_3.C_3 + \lambda_4.C_4$. (with $0\leq \lambda_i$).

We can now look for some objective functions. Note that due to the irreversible flux K1, K2, K4 and K567, $0\leq \lambda_i \leq 1$ whatever i = 1, .., 4.

1 – Minimization of all fluxes (in absolute value) for an input by T_6 = 1(pyruvate entry). It means $\lambda_3+\lambda_4=1$. The minimization of all fluxes is somewhat ambiguous.We can take $\lambda_1= \lambda_2= 0$ (values of the fluxes K4 and K567). It means that T1 = λ_3. Taking into account $\lambda_3+\lambda_4=1$ and the search for minimal fluxes gives $\lambda_3 = \lambda_4 = 0.5$. It corresponds to EFM 7.However with this value T5= -1, which is certainly not the minimum of T5 in absolute value.

We can think of taking the minimum of the sum of the absolute value of all the fluxes : $S(\lambda_1,\lambda_2,\lambda_3,\lambda_4)= |\lambda_3-\lambda_4|+|-\lambda_1+\lambda_3|+|-\lambda_1+\lambda_2|+\lambda_4+|\lambda_2-2\lambda_4|+\lambda_3+\lambda_4+\lambda_3+\lambda_1+\lambda_2$with $0\leq \lambda_i\leq 1$ ($\lambda_3+\lambda_4=1$ all the time and does not enter in S)

vT6=1 means $\lambda_3+\lambda_4=1$ i.e. $\lambda_3=1-\lambda_4$, so that:

$S(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=3+|1-2\lambda_4|+|-\lambda_1+1-\lambda_4|+|-\lambda_1+\lambda_2|+|\lambda_2-2\lambda_4| + \lambda_1+\lambda_2$

The values of the $\lambda_i$that achieve this minimum are$\lambda_1 = \lambda_2$, = 0,349, λ_3= 0,504 and λ_4=0,496. (determined using a software for optimization of functions such as Excel solver). With these values T5 = -0,643. In this case the solution is unique.

2- Maximization of Citrate production, i.e. maximization of T1 orλ_3-λ_1 maximal, i.e. λ_1=0 and λ_3=1. The last flux T6 gives 0≤ $\lambda_3 + \lambda_4 \leq 1$ i.e. λ_4 =0. One checks that there is no constraint on λ_2 so that there are in this case an infinity of solutions : C3 + λ_2 .C2 with 0≤ λ_2≤1.

3 - Maximization of alpha-ketoglutarate production, i.e. maximization of – T2, i.e. λ_1-λ_2 maximal, i.e. λ_1=1 and λ_2=0 which corresponds to C1 + 0.C2. C3 and C4 which do not contain T2 can be taken at any value provided that the resulting fluxes are all in the interval [-1; +1] or [-2; +2] for T5. Flux K2 and K3 give respectively 0≤ λ_4≤1 and 0≤ λ_3≤1. T6 adds 0≤ λ_3+λ_4≤1 which replaces the two previous inequalities. One can check that all other fluxes are in their interval. Thus in this case there is infinity of solutions: C1 + λ_3.C3 + λ_4.C4 with 0≤ λ_3+λ_4≤1.

4 - Maximization of citrate and alpha-ketoglutarate production with equal flux, i.e. maximization of T1 and –T2 with T1 = -T2. The corresponding coefficients of these fluxes gives : λ_3-λ_1 = λ_1-λ_2. All these coefficients are between 0 and 1 so that we are in the situation represented below:

0$\lambda_2\lambda_1\lambda_3$ 1
___|_______|________|_______|___|

the solution to maximize λ_3-λ_1 = λ_1-λ_2 is clearlyλ_2 = 0, λ_1 = 0.5 and λ_3=1. Because the coefficient of T6, λ_3+λ_4 ≤1, one get λ_4 = 0.This can be shown mathematically:

Maximizing vT1and-vT2 is equivalent to maximize $S(\lambda_{1,}\lambda_{2,}\lambda_3)=\lambda_3-\lambda_1+\lambda_1-\lambda_2=\lambda_3$ –λ_2(assuming that both fluxes are positive). The maximum is thus obtained for λ_2=0 and λ_3 =1 and with the constraintλ_3+λ_2 = 2λ_1(vT1=-vT2), we have λ_1=0.5. (we verify that vT1 and –vT2 are both positive)

Here, because we impose in fact two conditions, the solution is unique.

The Dynamics of a metabolic network

(COPASI http://www.copasi.org;
Berkeley Madonna http://www.berkeleymadonna.com).

The dynamical system of our metabolic network is written below:

d/dt(OAAm) = vK2 + vK8 - vK3
d/dt (ACoAm) = vK1 -vK3
d/dt (CITm) = vK3 -vK4 - vT1
d/dt(AKGm) = vK4 - vK567 + vT2

d/dt(MALm) = vK567 - vK8 + vT1 - vT2 - vT3
d/dt(Pim) = vK2 -vK567 + vT3 + vT5
d/dt(PYRm) = vT6 - vK1 - vK2

We take mass action laws as rate functions, with the only dependence upon the internal metabolites (the external metabolites can be thought to be equal to one):
vK1 = kK1 *PYRm
vK2 = kK2 *PYRm
vK3 = kK3*ACoAm *OAAm
vK4 = kK4*CITm
vK567 = kK567*AKGm*Pim
vK8 = kK8*MALm - kK8M*OAAm
vT1 = kT1*CITm - kT1M*MALm
vT2 = kT2*MALm - kT2M*AKGm
vT3 = kT3*MALm - kT3M*Pim
vT5 = kT5 - kT5M*Pim
vT6 = kT6
As an example we put all rate constants, kK1, kK2, ..kT6 = 2. The initial concentrations of metabolites (t = 0) are set equal to zero.

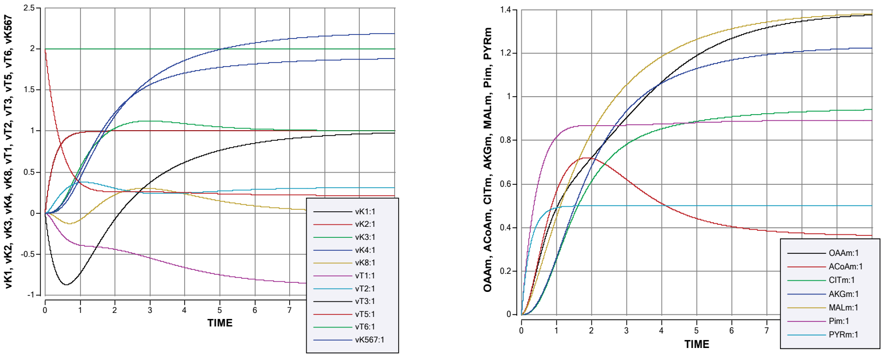


Figure 6: The time course of fluxes (left figure) and of metabolites concentrations (right figure) reaching a steady-state (calculations performed with Berkeley Madonna software).

Figure 6 gives the time course of the rates vK1, etc. and of the concentrations of metabolites. It appears clearly that we tend to a steady state for which the rates and the concentrations are constants. The values obtained for the rate are summarized in the table 5 below:

	K8	T1	T2	T3	T5	K1	K2	K3	K4	K567	T6
Steady-State	0	-0.894	0.317	1	0.211	1	1	1	1.894	2.211	2
1.894 x C1	0	-1.894	-1.894	0	0	0	0	0	1.894	0	0
2.211 x C2	0	0	2.211	0	2.211	0	0	0	0	2.211	0
C3	1	1	0	0	0	1	0	1	0	0	1
C4	-1	0	0	1	-2	0	1	0	0	0	1
Σ	0	-0.894	0.317	1	0.211	1	1	1	1.894	2.211	2

Table 5: Rates at steady-state and decomposition on the convex basis.

Thus the steady state is F = 1.894 C1 + 2.211 C2 + C3 + C4 which is close to EFM 15 = 2C1 +2C2 + C3 + C4.

What happens when we decrease the rate of pyruvate entry (T6)?

For T6 = 0.2, we obtain the flux at steady-state F = 0.76 C1 + 0.87 C2 + 0.1C3 + 0.1C4

And for T6 = 0.02, we obtain F = 0.65 C1 + 0.75 C2 + 0.01C3 + 0.01C4 which is close to 0.7 x (EFM 3) = 0.7 C1 + 0.7 C2. EFM 3 is characterized by entries of citrate and Pi and output of malate. Thus by decreasing the entry of pyruvate we pass from a steady-state close to EFM 15 to a new steady-state close to EFM 3.In the same way, one may expect that a decrease in the rate constant of K2 *versus* K1 will progressively shift the fluxes from EFM15, with equal fluxes in K2 and K1 towards EFM4, 5 or 9 (i.e. Krebs cycle with K8 ≠ 0 to regenerate OAA).Playing with the rate constants (and more generally with all the kinetic constants) is a convenient way to describe the possible reroutings of metabolism inside a fixed (from the point of view of structure) metabolic network.

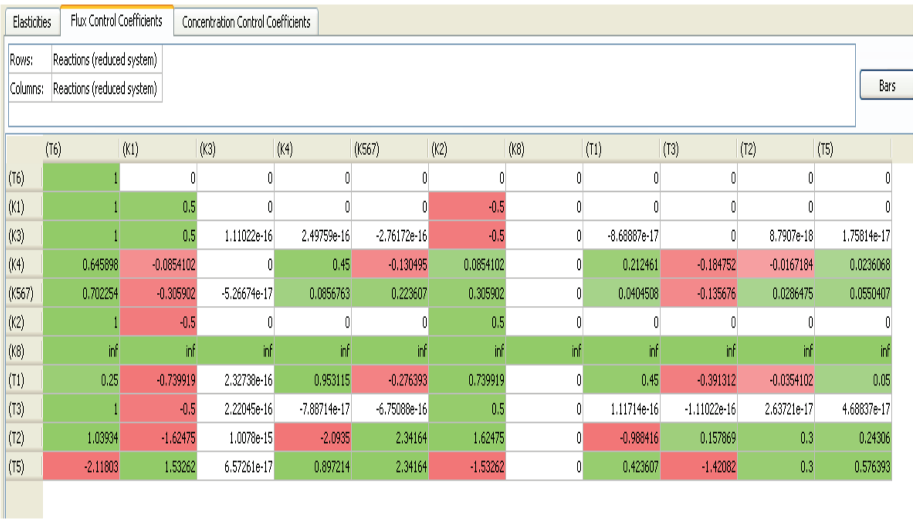
Control coefficients.

Copasi gives also the control coefficients on metabolite concentrations and on fluxes (which are the rates at steady state)[7-10)]. They are depicted in table 6:

The following points should be emphasized:

- The sum of all control coefficients on each row is equal to unity.
- The control coefficients of all steps except T6 on the flux through T6 are null, because T6 is irreversible.
- The control of T6 is high on all the steps.
- On the contrary the exchangers T2 and T5 have a low control.

These considerations may be used to select targets for the action of inhib (or activators), keeping in mind that when a step is inhibited (T6 inhib for instance, the flux and the control on the network will change as show table 7 below for T6 = 0.02. In this case it appears that the control of T K4, K567, T1, T2 and T5 is strongly decreased. If we



Rows: Reactions (reduced system)
Columns: Reactions (reduced system)

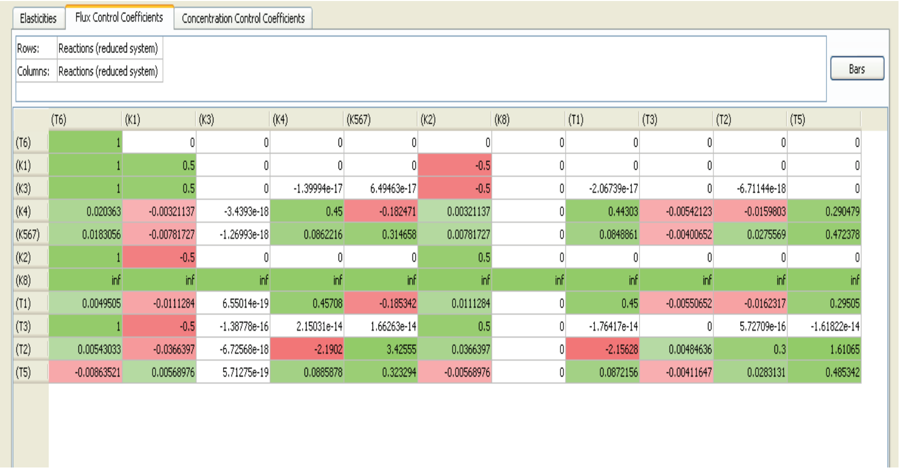
	(T6)	(K1)	(K3)	(K4)	(K567)	(K2)	(K8)	(T1)	(T3)	(T2)	(T5)
(T6)	1	0	0	0	0	0	0	0	0	0	0
(K1)	1	0.5	0	0	0	-0.5	0	0	0	0	0
(K3)	1	0.5	1.11022e-16	2.49759e-16	-2.76172e-16	-0.5	0	-8.68887e-17	0	8.7907e-18	1.75814e-17
(K4)	0.645898	-0.0854102	0	0.45	-0.130495	0.0854102	0	0.212461	-0.184752	-0.0167184	0.0236068
(K567)	0.702254	-0.305902	-5.26674e-17	0.0856763	0.223607	0.305902	0	0.0404508	-0.135676	0.0286475	0.0550407
(K2)	1	-0.5	0	0	0	0.5	0	0	0	0	0
(K8)	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf
(T1)	0.25	-0.739919	2.32738e-16	0.953115	-0.276393	0.739919	0	0.45	-0.391312	-0.0354102	0.05
(T3)	1	-0.5	2.22045e-16	-7.88714e-17	-6.75088e-16	0.5	0	1.11714e-16	-1.11022e-16	2.63721e-17	4.68837e-17
(T2)	1.03934	-1.62475	1.0078e-15	-2.0935	2.34164	1.62475	0	-0.988416	0.157869	0.3	0.24306
(T5)	-2.11803	1.53262	6.57261e-17	0.897214	2.34164	-1.53262	0	0.423607	-1.42082	0.3	0.576393

Table 6: Control coefficients of the steps (upper row) on the fluxes at steady-state (first column) for T6 = 2. (With Copasi)

want to further decrease the fluxes in the network we have to inhibit another controlling step such as K567 for instance. The high positive control coefficients are in green, the high negative control coefficients are in red.

Conclusion. We have detailed in this paper, several approaches on the same metabolic network representing the Krebs cycle with some exchangers between mitochondria and cytosol. It has to be emphasized that a large amount of information can be provided with the simple topological study of the metabolic network i.e. without considering the rate functions. This topological study should be a prerequisite to any more quantitative modeling of a metabolic network. The EFMs give a good description of all the possible pathways in the metabolic network. The only drawback of EFM analysis is their huge number in metabolic network with a number of stepsexceeding several hundreds. The flux balance analysis (FBA) approach can give quantitative values of fluxes satisfying a given objective without any knowledge of the rate functions. The problem is the choice of the objective function. If it could be simple for metabolic network of microorganisms (optimization of biomass, optimization of growth rate) defining an objective function in the case of eukaryotic organisms is more delicate, due to the interactions between cells and the possible changes of the objective function itself.

The introduction of the rate functions of all the steps lead to a set of differential equations describing the time course of the metabolites concentrations and of the fluxes. It permits to follow the dynamics of the system, to calculate the fluxes at steady state, to describe the possible reroutings of metabolic network. The determination of the control coefficient of fluxes on this quantitative description of the metabolic network may indicate targets for therapeutic drugs. The drawback of this approach is that very often not all parameters of the rate functions are known. In some cases they can be estimated by fitting the prediction of the model with some of the known fluxes (typically the input and output fluxes). A random sampling of the unknown parameters can also evidence coarse classes of similar pathways through the metabolic network.



Rows: Reactions (reduced system)
Columns: Reactions (reduced system)

	(T6)	(K1)	(K3)	(K4)	(K567)	(K2)	(K8)	(T1)	(T3)	(T2)	(T5)
(T6)	1	0	0	0	0	0	0	0	0	0	0
(K1)	1	0.5	0	0	0	-0.5	0	0	0	0	0
(K3)	1	0.5	0	-1.39994e-17	6.49463e-17	-0.5	0	-2.06739e-17	0	-6.71144e-18	0
(K4)	0.020363	-0.00321137	-3.4393e-18	0.45	-0.182471	0.00321137	0	0.44303	-0.00542123	-0.0159803	0.290479
(K567)	0.0183056	-0.00781727	-1.26993e-18	0.0862216	0.314658	0.00781727	0	0.0848861	-0.00400652	0.0275569	0.472378
(K2)	1	-0.5	0	0	0	0.5	0	0	0	0	0
(K8)	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf
(T1)	0.0049505	-0.0111284	6.55014e-19	0.45708	-0.185342	0.0111284	0	0.45	-0.00550652	-0.0162317	0.29505
(T3)	1	-0.5	-1.38778e-16	2.15031e-14	1.66263e-14	0.5	0	-1.76417e-14	0	5.72709e-16	-1.61822e-14
(T2)	0.00543033	-0.0366397	-6.72568e-18	-2.1902	3.42555	0.0366397	0	-2.15628	0.00484636	0.3	1.61065
(T5)	-0.00863521	0.00568976	5.71275e-19	0.0885878	0.323294	-0.00568976	0	0.0872156	-0.00411647	0.0283131	0.485342

Table 7 : Control coefficients of the steps (upper row) on the fluxes at steady-state (first column) for T6 = 0.02. (With Copasi)

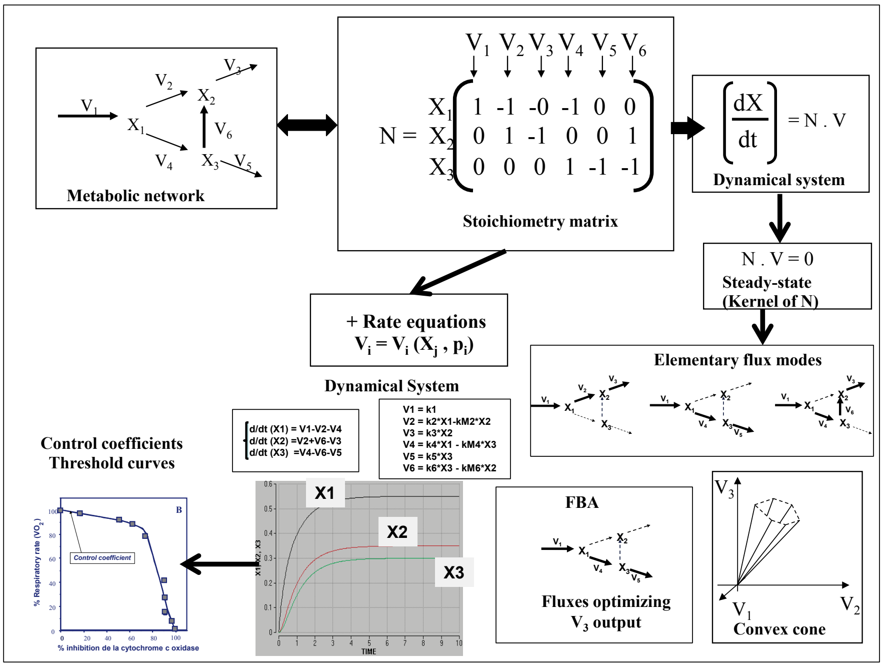


Figure 7: Plan study of metabolic networks. The first step is to write the stoichiometry matrix N from which a lot of knowledge can be already derived, including all the minimal pathways inside the metabolic network (EFM) and fluxes optimizing an objective function (FBA).
The addition of the rate functions gives a dynamical system describing the time course of fluxes and metabolites concentrations. The calculation of the control coefficients indicates the steps to which the network is the most sensitive.

Our suggestion is to approach all metabolic network in a systematic way (Fig. 7), beginning with the description of the stoichiometry matrix. The mathematical study of the stoichiometry matrix will give numerous important informations on the metabolic network such as the number of independent variables and the relationships between the dependent variables. The determination of the convex basis will be useful to describe any pathway in the metabolic network and the EFMs. If an objective is envisaged, a FBA approach could be useful to determine the fluxes inside the network and its robustness. For further quantitative study of the metabolic network it is necessary to define the rate functions of all steps. It permits the determination of the steady state, with the calculation of control coefficients but also to study dynamical behavior of the network which can be compared with experimental observations or with other theoretical predictions such as the ones of FBA.

Bibliography

[1] Christine Nazaret, Margit Heiske, Kevin Thurley, Jean-Pierre Mazat *: Mitochondrial energetic metabolism : A simplified model of TCA cycle with ATP production*. J Theor Biol. 2009;258(3):455-64.

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

[3] Schuster, S., Fell, D. A., & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology,* 18, 326-332.

[4]Schilling CH, Letscher, D and Palsson BO.*Theory for the systemic definition of metabolic pathway and their use in interpreting metabolic function from a pathway-oriented perspective* J. theor. Biol.203 (2000) 229-248.

[5]Lee J. Sweetlove, Katherine F.M. Beard, Adriano Nunes-Nesi, Alisdair R. Fernie and R. George Ratcliffe: *Not just a circle: flux modes in the plant TCA cycle*. Trends in Plant Science (2010) 15 : 390-406.

[6] F. Zamorano, A. Vande Wouwer, R.M. Jungers, G. Bastin : *Dynamic metabolic models of CHO cell cultures through minimal sets of elementary flux modes.* Journal of Biotechnology 164 (2013) 409– 422

[7] Jeffrey D Orth, Ines Thiele & Bernhard Ø Palsson *What is flux balance analysis?* Nature Biotechnology (2010) 28: 245-248.

[8]Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bölling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novère N, Malys N, Mazein A, Papin JA, Price ND, Selkov E Sr, Sigurdsson MI, Simeonidis E, Sonnenschein N, Smallbone K, Sorokin A, van Beek JH, Weichart D, Goryanin I, Nielsen

J, Westerhoff HV, Kell DB, Mendes P, Palsson BØ. *A community-driven global reconstruction of human metabolism.* Nat Biotechnol. 2013 May;31(5):419-25. doi: 10.1038/nbt.2488

[9] Kacser, H., Burns, J.A., *The control of flux*, Symp. Soc. Exp. Biol. 32 (1973) 65-104.

[10] Heinrich, R., Rapoport, T.A., *A linear steady-state treatment of enzymatic chains. General properties, control and effector strength,* Eur. J. Biochem. 42 (1974), 89-95.

[11] Reder, C., *Metabolic control theory: a structural approach*, J. Theor. Biol. 135 (1988) 175-201.

[12] D. Fell, *Understanding the control of metabolism* (Portland Press, Oxford, 1997).

APPENDIX 1 : METATOOL ENTRY FILE

```
-ENZREV
K8 T1 T2 T3 T5

-ENZIRREV
K1 K2 K3 K4 K567 T6

-METINT
OAAm ACoAm CITm AKGm MALm Pim PYRm

-METEXT
PYRc NADm NADH2m CoAm ADPm ATPm H2O CO2 MALc CITc
Pic AKGc


-CAT


K1 : PYRm + NADm + CoAm = ACoAm + NADH2m + CO2 .
K2 : PYRm + CO2 + ATPm = OAAm + Pim + ADPm .
K3 : OAAm + ACoAm + H2O = CITm + CoAm .
K4 : CITm + NADm = AKGm + NADH2m + CO2 .
K567 : AKGm + NADm + Pim + ADPm = MALm + NADH2m + CO2 +
ATPm .
K8 : MALm + NADm = OAAm + NADH2m .
T1 : CITm + MALc = CITc + MALm .
T2 : AKGc + MALm = AKGm + MALc .
T3 : MALm + Pic = MALc + Pim .
T5 : Pic = Pim .
T6 : PYRc = PYRm .
```

APPENDIX 2 : METATOOL OUTPUT FILE

```
METATOOL OUTPUT (int) Version 4.3 (25 October 2002) E:\ECOLE
THEMATIQUE EVRY 2014\TUTORIAL
METABOLISME\EFM\meta4.3_int.exe

INPUT FILE: KS1_in.dat

INTERNAL METABOLITES: 7
EXTERNAL METABOLITES: 12
```

```
REACTIONS: 11

  5 int      MALm
  4 external NADm
  4 external NADH2m
  4 external CO2
  4 int      Pim
  3 int      PYRm
  3 int      OAAm
  3 int      CITm
  3 int      AKGm
  3 external MALc
  2 external CoAm
  2 int      ACoAm
  2 external ATPm
  2 external ADPm
  2 external Pic
  1 external H2O
  1 external CITc
  1 external AKGc
  1 external PYRc
19 metabolites,  50 is the summarized frequency

edges    frequency of nodes
  1          4
  2          5
  3          5
  4          4
  5          1
freq_of_nodes = 5.7561 * edges^(-0.5765)
Linear correlation coefficient r = -0.539471
The dependency is not significant.

STOICHIOMETRIC MATRIX

 matrix dimension r7 x c11
 1  0  0  0  0  0  1 -1  0  0  0
0  0  0  0  0  1  0 -1  0  0  0
 0 -1  0  0  0  0  0  1 -1  0  0
 0  0  1  0  0  0  0  0  1 -1  0
-1  1 -1 -1  0  0  0  0  0  1  0
 0  0  0  1  1  0  1  0  0 -1  0
 0  0  0  0  0 -1 -1  0  0  0  1
The following line indicates reversible (0) and irreversible reactions (1)
 0  0  0  0  0  1  1  1  1  1  1
rows and columns are sorted as declared in the inputfile


KERNEL

 matrix dimension r4 x c11
-2 -1  0  1 -2 -1  1 -1  0  0  0
0  1  1  0  0  0  0  0 -1  0  0
 0  0  1  0  1  0  0  0  0  1  0
-1  0  0  1 -2  0  1  0  0  0  1
11 reactions (columns) are sorted in the same order as in the ENZREV
ENZIRREV section.

enzymes

 1:        (-2 K8) -T1 T3 (-2 T5) -K1 K2 -K3 irreversible
 2:        T1 T2 -K4 irreversible
 3:        T2 T5 K567 irreversible
 4:        -K8 T3 (-2 T5) K2 T6 irreversible

overall reaction

 1: 3 NADH2m + ATPm + 2 CO2 + CITc = 3 NADm + ADPm + H2O +
2 MALc + Pic
 2: NADH2m + CO2 + AKGc = NADm + CITc
 3: NADm + ADPm + Pic + AKGc = NADH2m + ATPm + CO2 + MALc
 4: PYRc + NADH2m + ATPm + CO2 = NADm + ADPm + MALc + Pic

BLOCK DIAGONALISATION
Reaction blocks were found from nullspace matrix (KERNEL).
1. block:
K8       T1       T2       T3       T5       K1       T6       K3       K4
         K567     K2

SUBSETS OF REACTIONS

 matrix dimension r9 x c11
 1  0  0  0  0  0  0  0  0  0  0
 0  1  0  0  0  0  0  0  0  0  0
 0  0  1  0  0  0  0  0  0  0  0
 0  0  0  1  0  0  1  0  0  0  0
 0  0  0  0  1  0  0  0  0  0  0
 0  0  0  0  0  1  0  1  0  0  0
 0  0  0  0  0  0  0  0  1  0  0
 0  0  0  0  0  0  0  0  0  1  0
 0  0  0  0  0  0  0  0  0  0  1
```

11 reactions (columns) are sorted in the same order as in the ENZREV ENZIRREV section.

enzymes

```
 1:      K8 reversible
 2:      T1 reversible
 3:      T2 reversible
 4:      T3 K2 irreversible
 5:      T5 reversible
 6:      K1 K3 irreversible
7:       K4 irreversible
 8:      K567 irreversible
 9:      T6 irreversible
```

overall reaction

1: MALm + NADm = OAAm + NADH2m
2: CITm + MALc = MALm + CITc
3: MALm + AKGc = AKGm + MALc
4: MALm + PYRm + ATPm + CO2 + Pic = OAAm + 2 Pim + ADPm + MALc
5: Pic = Pim
6: OAAm + PYRm + NADm + H2O = CITm + NADH2m + CO2
7: CITm + NADm = AKGm + NADH2m + CO2
8: AKGm + Pim + NADm + ADPm = MALm + NADH2m + ATPm + CO2
9: PYRc = PYRm

REDUCED SYSTEM with 6 branch point metabolites in 9 reactions (columns)

matrix dimension r6 x c9

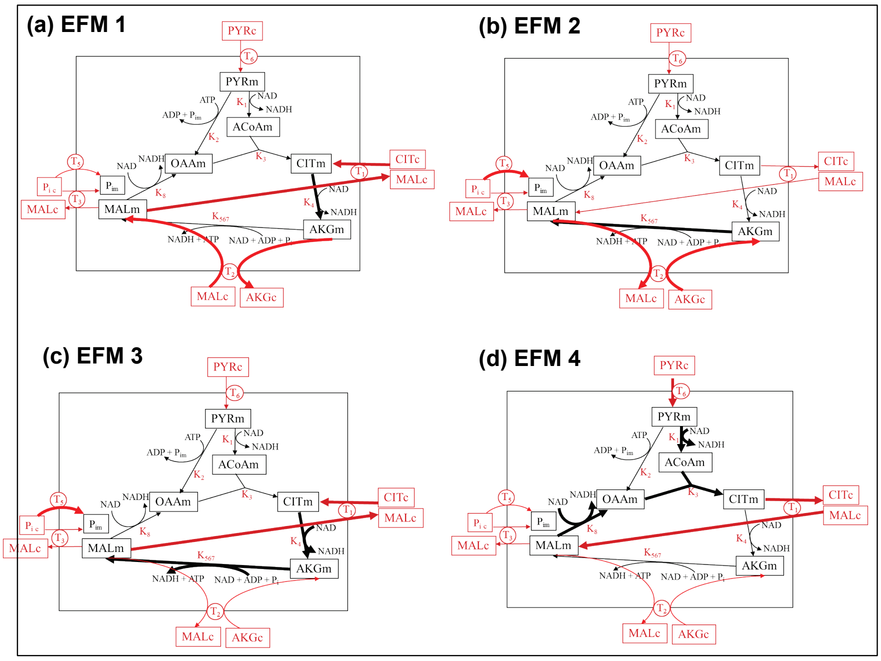
```
 1  0  0  1  0 -1  0  0  0
0 -1  0  0  0  1 -1  0  0
```

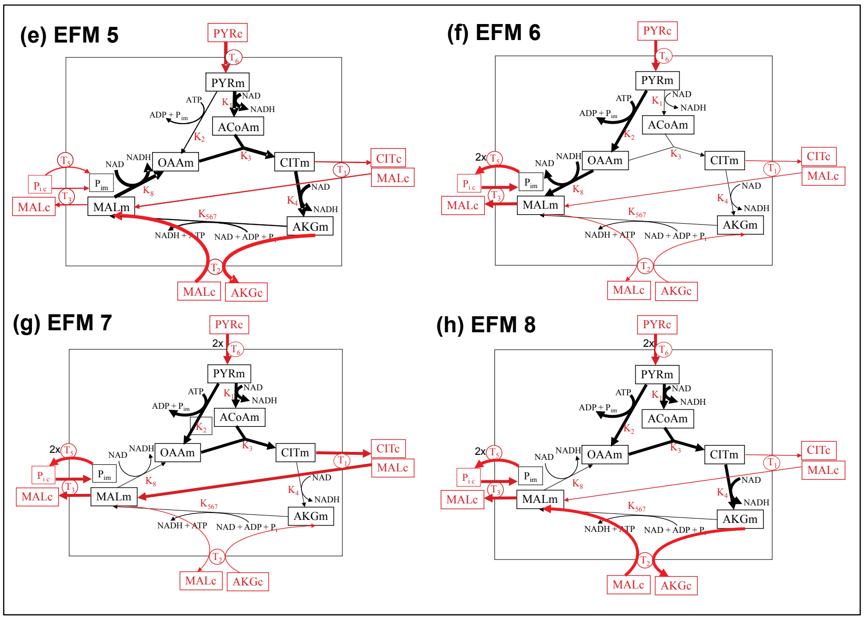
```
 0  0  1  0  0  0  1 -1  0
-1  1 -1 -1  0  0  0  1  0
 0  0  0  2  1  0  0 -1  0
 0  0  0 -1  0 -1  0  0  1
The following line indicates reversible (0) and irreversible reactions (1)
 0  0  0  1  0  1  1  1  1

-> Branch metabolites are :
met          cons    built    reactions
OAAm                 1        2        3        iir
CITm         2       1        3        iir
AKGm                 1        2        3        iir
MALm                 3        2        5        irrrr
Pim          1       3        4        iirr
PYRm                 2        1        3        iii

-> No branch metabolites are :
met          cons    built    reactions
ACoAm                1        1        2        ii

CONVEX BASIS

 matrix dimension r4 x c11
 0 -1 -1  0  0  0  0  0  1  0  0
 0  0  1  0  1  0  0  0  0  1  0
 1  1  0  0  0  1  0  1  0  0  1
-1  0  0  1 -2  0  1  0  0  0  1

 enzymes

 1:      -T1 -T2 K4 irreversible
 2:      T2 T5 K567 irreversible
 3:      K8 T1 K1 K3 T6 irreversible
 4:      -K8 T3 (-2 T5) K2 T6 irreversible

overall reaction

 1: NADm + CITc = NADH2m + CO2 + AKGc
 2: NADm + ADPm + Pic + AKGc = NADH2m + ATPm + CO2 + MALc
 3: PYRc + 2 NADm + H2O + MALc = 2 NADH2m + CO2 + CITc
 4: PYRc + NADH2m + ATPm + CO2 = NADm + ADPm + MALc + Pic
```

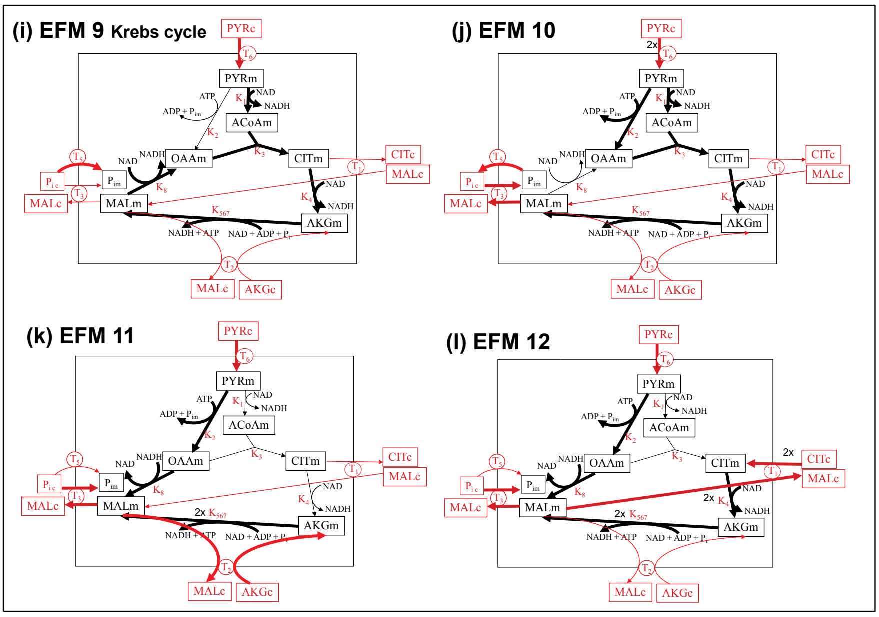
```
CONSERVATION RELATIONS
- not found -

ELEMENTARY MODES

matrix dimension r16 x c11
 0 -1 -1  0  0  0  0  0  1  0  0
0  0  1  0  1  0  0  0  0  1  0
 0 -1  0  0  1  0  0  0  1  1  0
 1  1  0  0  0  1  0  1  0  0  1
 1  0 -1  0  0  1  0  1  1  0  1
-1  0  0  1 -2  0  1  0  0  0  1
 0  1  0  1 -2  1  1  1  0  0  2
 0  0 -1  1 -2  1  1  1  1  0  2
 1  0  0  0  1  1  0  1  1  1  1
 0  0  0  1 -1  1  1  1  1  1  2
-1  0  2  1  0  0  1  0  0  2  1
-1 -2  0  1  0  0  1  0  2  2  1
 1  0  0  1  0  2  1  2  2  2  3
 0  1  2  1  0  1  1  1  0  2  2
 0 -1  0  1  0  1  1  1  2  2  2
 0  0  1  1  0  1  1  1  1  2  2
11 reactions (columns) are sorted in the same order as in the ENZREV
ENZIRREV section.
The following line indicates reversible (0) and irreversible reactions (1)
 0  0  0  0  0  1  1  1  1  1  1

 enzymes
# in () indicates # of enzymes used by the elementary mode
# in [] indicates the diagonal block of the kernel matrix to which the
elementary mode belongs
1: ( 3) [bl 1]  -T1 -T2 K4 irreversible
 2: ( 3) [bl 1]  T2 T5 K567 irreversible
 3: ( 4) [bl 1]  -T1 T5 K4 K567 irreversible
 4: ( 5) [bl 1]  K8 T1 K1 K3 T6 irreversible
 5: ( 6) [bl 1]  K8 -T2 K1 K3 K4 T6 irreversible
 6: ( 5) [bl 1]  -K8 T3 (-2 T5) K2 T6 irreversible
 7: ( 7) [bl 1]  T1 T3 (-2 T5) K1 K2 K3 (2 T6) irreversible
 8: ( 8) [bl 1]  -T2 T3 (-2 T5) K1 K2 K3 K4 (2 T6) irreversible
 9: ( 7) [bl 1]  K8 T5 K1 K3 K4 K567 T6 irreversible
 10: ( 8) [bl 1]  T3 -T5 K1 K2 K3 K4 K567 (2 T6) irreversible
 11: ( 6) [bl 1]  -K8 (2 T2) T3 K2 (2 K567) T6 irreversible
 12: ( 7) [bl 1]  -K8 (-2 T1) T3 K2 (2 K4) (2 K567) T6 irreversible
 13: ( 8) [bl 1]  K8 T3 (2 K1) K2 (2 K3) (2 K4) (2 K567) (3 T6)
irreversible
 14: ( 8) [bl 1]  T1 (2 T2) T3 K1 K2 K3 (2 K567) (2 T6) irreversible
 15: ( 8) [bl 1]  -T1 T3 K1 K2 K3 (2 K4) (2 K567) (2 T6) irreversible
 16: ( 8) [bl 1]  T2 T3 K1 K2 K3 K4 (2 K567) (2 T6) irreversible

overall reaction

 1: NADm + CITc = NADH2m + CO2 + AKGc
 2: NADm + ADPm + Pic + AKGc = NADH2m + ATPm + CO2 + MALc
 3: 2 NADm + ADPm + CITc + Pic = 2 NADH2m + ATPm + 2 CO2 +
MALc
 4: PYRc + 2 NADm + H2O + MALc = 2 NADH2m + CO2 + CITc
 5: PYRc + 3 NADm + H2O + MALc = 3 NADH2m + 2 CO2 + AKGc
 6: PYRc + NADH2m + ATPm + CO2 = NADm + ADPm + MALc + Pic
 7: 2 PYRc + NADm + ATPm + H2O = NADH2m + ADPm + CITc + Pic
 8: 2 PYRc + 2 NADm + ATPm + H2O = 2 NADH2m + ADPm + CO2 +
Pic + AKGc
 9: PYRc + 4 NADm + ADPm + H2O + Pic = 4 NADH2m + ATPm + 3
CO2
 10: 2 PYRc + 3 NADm + H2O = 3 NADH2m + 2 CO2 + MALc
 11: PYRc + NADm + ADPm + Pic + 2 AKGc = NADH2m + ATPm +
CO2 + 3 MALc
 12:      PYRc + 3 NADm + ADPm + 2 CITc + Pic = 3 NADH2m +
ATPm + 3 CO2 + 3 MALc
 13: 3 PYRc + 7 NADm + ADPm + 2 H2O + Pic = 7 NADH2m + ATPm
+ 5 CO2 + MALc
 14: 2 PYRc + 3 NADm + ADPm + H2O + Pic + 2 AKGc = 3 NADH2m
+ ATPm + 2 CO2 + 2 MALc + CITc
 15: 2 PYRc + 5 NADm + ADPm + H2O + CITc + Pic = 5 NADH2m +
ATPm + 4 CO2 + 2 MALc
 16: 2 PYRc + 4 NADm + ADPm + H2O + Pic + AKGc = 4 NADH2m +
ATPm + 3 CO2 + 2 MALc
```

The elementary modes (12) 3 5 7 8 9 10 11 12 13 14 15 16 are additional to the convex basis.

APPENDIX 3 : Description of all Elementary Flux Modes (EFM) of KS1





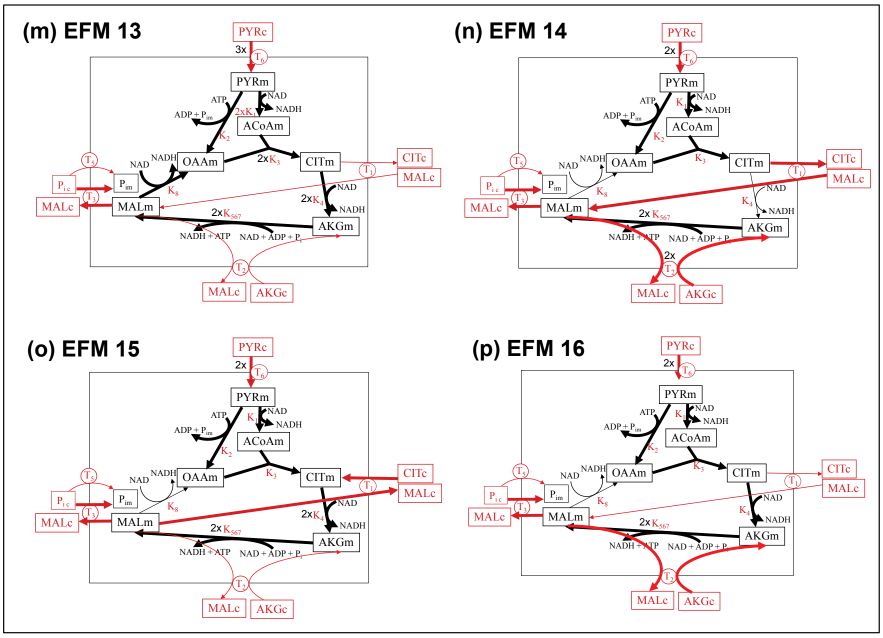


(i) EFM 9 Krebs cycle

(j) EFM 10

(k) EFM 11

(l) EFM 12



(m) EFM 13

(n) EFM 14

(o) EFM 15

(p) EFM 16

LIST OF ATTENDEES

(February 16th, 2014)

AMAR Patrick	(pa@lri.fr)
BASSO-BLANDIN Adrien	(abasso@ibisc.univ-evry.fr)
BERNOT Gilles	(bernot@unice.fr)
BERTAUX François	(francois.bertaux@inria.fr)
BEURTON-AIMAR Marie	(beurton@labri.fr)
BOCSKEI Zsolt	(zsolt.bocskei@sanofi.com)
BODINIER romain	(romainbodinier@gmail.com)
BONNARD Cécile	(cecile.bonnard@sobios.com)
BOUFFARD Marc	(marc.bouffard@lri.fr)
BOUYIOUKOS Costas	(costas.bouyioukos@issb.genopole.fr)
BRIL Antoine	(antoine.bril@fr.netgrs.com)
CARBONELL Pablo	(pablo.carbonell@issb.genopole.fr)
CASAGRANDA Stefano	(stefano.casagranda@inria.fr)
COLOMBIE sophie	(scolombi@bordeaux.inra.fr)
DELAPLACE Franck	(franck.delaplace@ibisc.fr)
DI GIUSTO Cinzia	(cinzia.digiusto@gmail.com)
DOULAZMI Mohamed	(mohamed.doulazmi@upmc.fr)
FIPPO FITIME Louis	(Louis.Fippo-Fitime@irccyn.ec-nantes.fr)
FROLOFF Nicolas	(Nicolas.FROLOFF@3ds.com)
GANGWE NANA Ghislain Yannick	(yghislain.gangwenana@yahoo.fr)
GENDRAULT Yves	(ygendrault@unistra.fr)
GILARD Françoise	(francoise.gilard@u-psud.fr)
HENRY Adrien	(adrien.ym.henry@gmail.com)
JANNIERE Laurent	(laurent.janniere@issb.genopole.fr)
JERALA Roman	(roman.jerala@KI.si)
JESTER Brian	(jester@issb.genopole.fr)

JUNIER Ivan	(i.junier@gmail.com)
KAUFMAN Marcelle	(mkaufman@ulb.ac.be)
KÉPÈS François	(francois.kepes@issb.genopole.fr)
LE GALL Pascale	(pascale.legall@issb.genopole.fr)
LEPAGE Thibaut	(lepage@issb.genopole.fr)
LLAMOSI Artemis	(artemis.llamosi@gmail.com)
MAZAT Jean-Pierre	(jean-pierre.mazat@phys-mito.u-bordeaux2.fr)
MORTEROL Martin	(morterol@lri.fr)
NORRIS Victor	(victor.norris@univ-rouen.fr)
PARRELLO Damien	(damien.parrello@univ-lorraine.fr)
PARUTTO Pierre	(pierreparutto@gmail.com)
PECCOUD Jean	(jpeccoud@vbi.vt.edu)
PERES Sabine	(sabine.peres@lri.fr)
POUPIN Nathalie	(nathalie.poupin@toulouse.inra.fr)
RIVOIRE Olivier	(olivier.rivoire@ujf-grenoble.fr)
ROBINSON Alan	(ajr@mrc-mbu.cam.ac.uk)
SIEBERT Heike	(hsiebert@zedat.fu-berlin.de)
SILVA Pedro Ângelo	(ppsilva@igc.gulbenkian.pt)
SOURNIA Pierre	(pierre.sournia@polytechnique.edu)
STAN Guy-Bart	(g.stan@imperial.ac.uk)
SUBRAMANIAN Kartik	(skartik@vt.edu)
THIELE Ines	(ines.thiele@uni.lu)
TROSSET Jean-Yves	(jytrosset@gmail.com)
VANOUSE Paul	(vanouse@buffalo.edu)
YOU Lingchong	(you@duke.edu)
ZELISZEWSKI Dominique	(dominique.zeliszewski@issb.genopole.fr)