Proceedings of The Nice Spring School
on

Avances in Systems and Synthetic Biology

March 25th - 29th, 2013

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.

Just such a multi-disciplinary group of scientists has been meeting regularly at Genopole, a leading centre for genomics in France. This, the *Epigenomics project*, is divided into five subgroups. The *GolgiTop* subgroup focuses on membrane deformations involved in the functionning of the Golgi. The *Hyperstructures* subgroup focuses on cell division, on the dynamics of the cytoskeleton, and on the dynamics of *hyperstructures* (which are extended multi-molecule assemblies that serve a particular function). The *Observability* subgroup addresses the question of which models are coherent and how can they best be tested by applying a formal system, originally used for testing computer programs, to an epigenetic model for mucus production by *Pseudomonas aeruginosa*, the bacterium involved in cystic fibrosis. The *Bioputing* group works on new approaches proposed to understand biological computing using computing machine made of biomolecules or bacterial colonies. The *SMABio* subgroup focuses on how multi-agents systems (MAS) can be used to model biological systems.

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, Reinhard Lipowsky, Jean-Pierre Mazat, Victor Norris, William Saurin, El Houssine Snoussi.

ACKNOWLEDGEMENTS

We would like to thank the conference participants, who have contributed in a way or another this book. It gathers overviews of the talks, discussions and roundtables, 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 *Club Belambra "La Bergerie"* 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 *Arc de Venet*, Jardin Albert 1^{ier} in the town center of Nice.

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/
- GDR CNRS 3003 Bioinformatique Moléculaire:
  http://www.gdr-bim.u-psud.fr

THE EDITORS

INVITED SPEAKERS

MARK BEDAU,	Reed College, Portland, OR, USA
HELEN BYRNE,	OCCAM, U. Oxford, UK
ANDREAS DRESS,	CAS-MPG, SIBS, Shanghai, CN
RACHEL GILES,	U. Medical center, Utrecht, NL
RICHARD KITNEY,	Imperial College London, UK
DOMINIQUE SCHNEIDER,	U. Joseph Fourier, Grenoble, FR
WALTER SCHUBERT,	CAS-MPG, Shanghai, CN & U. Magdeburg, DE
RICARD SOLÉ,	U. Pompeu Fabra, Barcelona, ES
ORKUN SOYER,	U. Exeter, UK
BIRGIT WILTSCHI,	Austrian Center of Industrial Biotechnology, Graz, AT

CONTENTS

To be announced

Richard Kitney[1]

[1] Imperial College London, UK

Abstract

Non-canonical amino acids as building blocks

Birgit Wiltschi[1]

[1] Junior Group Synthetic Biology, Austrian Centre of Industrial Biotechnology ACIB GmbH, Graz, Austria

Abstract

Non-canonical amino acids (ncAAs) can be used as building blocks for the biosynthesis of synthetic proteins. Though not encoded by the genetic code, these analogs of the canonical amino acids participate in ribosomal protein translation under tightly controlled conditions. Most of the ncAAs carry unusual side chains. Their translation into a target protein sequence can provoke structural, chemical, or functional modifications normally not found in nature. Thus, protein engineering with ncAAs offers an extension to classical genetic engineering approaches for protein modification. It is an emerging research area in the field of synthetic biology at the interface of biology and chemistry that bears unprecedented biotechnological potential.

The incorporation of ncAAs into proteins requires the reprogramming of protein biosynthesis. This can be achieved by skillful manipulation of the different components of the translational machinery. Aminoacyl-tRNA synthetases (AARSs) are crucial players in the genetic code interpretation and, therefore, represent a main target for engineering efforts. The manipulation of the catalytic activity of these enzymes provides the clue for the efficient incorporation of ncAAs into target proteins. Currently, two approaches for controlling amino acid selection and catalytic turnover are employed. The first exploits the natural substrate tolerance of the AARSs in the context of amino acid auxotrophies for global substitution of an amino acid by its non-canonical analog. Alternatively, site-specific introduction of an ncAA is achieved by in-frame stop codon suppression in combination with the mutation of the substrate specificity of an AARS. This approach requires the development of AARS/suppressor tRNA pairs that are orthogonal. The orthogonal pair must be specific for its cognate amino acid and must not exhibit cross-reactivity with AARS/tRNA pairs of the host.

In my presentation, I will focus on the basic requirements for the modification of proteins with ncAAs by the two complementary approaches. Using examples from the literature and from our own work, I will illustrate the potentials of protein engineering with ncAAs.

Beyond BioBricks: Using machine learning methods to discover and optimize complex systems in synthetic biology

Mark Bedau[1]

[1] Reed College, Portland OR, USA

Abstract

This talk has two main messages. The first is that emergence plays a central role in complex synthetic biology mechanisms. Emergence has a controversial history in both philosophy and science, but the controversy is now dissipating, in part because of growing awareness of a new conception of emergence (termed "weak" emergence) concerning global states produced by complex causal networks. Complex causal networks are characterized by high parallelism (many independent variables), high nonlinearity (of response of each variable), and high synergy (the response of a variable depends on the responses of other variables). One main way to understand and control the emergent properties produced by complex causal webs is through Edisonian trial and error, involving extensive empirical observation and experimentation. (Another is by means of computer simulations.) The mechanisms constructed in synthetic biology are typically very complex, and the resulting weak emergent properties explain why experimental troubleshooting dominates work in synthetic biology laboratories. Hence my first main message: synthetic biologists should embrace rather than ignore the emergent properties in the complex biochemical systems that they synthesize. My second main message is to demonstrate a new and powerful method to engineer systems to have desired emergent properties.

This method puts machine-learning algorithms in control of high-throughput experimental technology, in order to "program" an experimental system's desired emergent properties. The programming is indirect because it starts by specifying the desired goal state (the desired emergent properties) and then conducts a sequence of experiments that efficiently optimizes the desired properties. In addition to high experimental throughput, this method's keys to success are (i) a precisely defined library of potential experiments, (ii) an effective measurement of the degree to which a given experiment exhibits the desired emergent property, and (iii) an ability to create any experiment in the library on demand. Examples of successfully engineering the emergent properties of complex biochemical systems including programming synergistic drug combinations and formulations and programming experimental conditions for protein synthesis.

Novel applications in synthetic biology include programming genetic sequences, BioBrick designs, and minimal or refactored genomes.

References

[1] S. Rasmussen et al. 2008. A roadmap to protocells. In S. Rasmussen, et al., eds., *Protocells: bridging nonliving and living matter*, pp. 71-100. Cambridge: MIT Press.

[2] F. Caschera et al. 2011. Coping with complexity: machine learning optimization of cell-free protein synthesis. *Biotechnology and Bioengineering* 108 (9): 2218-2228.

To be announced

Ricard Solé[1]

[1] U. Pompeu Fabra, Barcelona, ES

Abstract

Evolutionary processes driving system properties and system properties channelling evolutionary processes

Orkun S. Soyer[1]

[1] University of Exeter, GB

Abstract

We can not claim a complete understanding of a biological system, predict its diversity among organisms, or hope to be able to reliable manipulate it (in the context of medicine and synthetic biology) without understanding its evolutionary history. In this talk, I will demonstrate mathematical and computational approaches towards deciphering the evolutionary processes that can lead to specific systems properties at the cellular level. In turn, these properties can channel further evolution creating an interesting interplay between systems level properties and evolutionary dynamics. Describing few recent projects in detail, I will highlight fluctuating environments and nonlinear dynamics as an example of this interplay and its effect of system robustness and evolvability. The talk will conclude with general remarks on the relevance of the emerging field of evolutionary systems biology in our quest to better understand (and manipulate) cellular systems.

Orkun S. Soyer is a senior lecturer of systems biology at the University of Exeter. He received a PhD from University of Michigan under the supervision of Richard Goldstein in 2004. His postdoctoral work focused on the evolution of signalling networks, and was done under the mentorship of Sebastian Bonhoeffer at the ETH, Zuerich. Orkun's lab is interested in deciphering the evolutionary and ecological principles that can explain the structure and dynamics of biological systems at molecular level. A connected aim is to use the resulting insights towards manipulating existing biological systems or designing novel ones.

Systems Biology during long-term evolution with *Escherichia coli*: dynamics of regulatory networks and mutation rates

Dominique Schneider[1]

[1] Lab. Adaptation et Pathogénie des Microorganismes, CNRS UMR5163, Université Joseph Fourier, Grenoble, France.

Abstract

Systems Biology, through its interdisciplinary nature, highlighted the multi-faceted complexity of living organisms by improving our understanding of the structure and function of genomes and cellular networks. Integrating evolutionary perspectives is fully complementary by providing a dynamic view of virtually all cellular processes. Experimental evolution is designed to reproduce evolution in controlled laboratory conditions and therefore provides such an evolutionary framework. During the longest running evolution experiment, twelve populations of *Escherichia coli* are independently propagated from a common ancestor in a defined environment for more than 55,000 generations. The full revivable fossil record of the entire evolution experiment is investigated for phenotypic, genomic and global expression changes. All populations achieved substantial fitness improvement during evolutionary time. Adaptive changes have been shown to be associated with complex effects on global gene expression, including widespread pleiotropy and epistasis, indicative of important changes in regulatory networks. In addition to their expression, bacterial genomes revealed an exceptional dynamics in their mutation rates, reflecting a tension between adaptation and genetic load that is strongly related to the fit of the bacterial populations to their environment.

The integration of Systems Biology into this evolutionary framework therefore addresses how evolvable are genomic features and the molecular bases of the evolvability of networks that has to be integrated in all the complex Synthetic Biology projects. These interdisciplinary approaches pave the way toward understanding precisely how the flow of information is transduced from the environment to bacterial cells, which in turn are able to cope with the various challenges imposed by the external conditions, including human behaviour and need.

The human toponome project: translating the spatial protein network code (toponome) into efficient therapies

Walter Schubert[1,2,3]

[1] Molecular pattern recognition research group, OvG-university Magdeburg, Germany
[2] International faculty, Max-Planck (CAS-MPG) partner institute for computational biology, Shanghai, China
[3] Human toponome project, TNL, Munich, Germany

Abstract

The recent development of parameter-unlimited super-resolution microscopy TIS^{TM} (Toponome Imaging System) provides direct access to protein neworks at nanometer 2D and 3D resolution in a single tissue section or inside cells. TIS^{TM} is a device that overcomes both the spectral and the resolving power of conventional light microscopy without having to change hardware. It is the first ready-to-use technology for dimension- and parameter-unlimited histological diagnostics and systematic decoding of the toponome at super-resolution (toponome: defined as the spatial protein network code in morphologically intact cells and tissues). TIS^{TM} is a highly flexible machine that can adapt to the needs of the researcher: a 4-in-one microscope including (1) routine transmitted light functions, (2) conventional fluorescence functionalities, (3) parameter-unlimited protein network visualization in real time, and (4) super-resolution of subcellular structures and protein clusters in tissue sections and in cultured cells (subnanometer to 40 nm resolution).

It is a novel platform providing the robustness needed for the human toponome project, combining industry partners and research institutions. The technology has shown to solve key problems in cell-, tissue-, and clinical toponomics by directly decoding cellular (disease) mechanisms *in situ/in vivo*, in particuar at the target sites of cancer in human tissue. Several next-generation toponome biomarkers and toponome drugs are on the way to clinic. The human toponome project has at its goal to unravel the complete toponome in all cell types and tissues in health and disease. The technology is scalable as large cooperative parallel screening devices extracting the most relevant disease targets from protein network hierarchies in situ: a novel efficient way to find selective drugs, by escaping the low content trap in current drug target and diagnostic marker discovery strategies, which, as yet, systematically disregard the spatial topology of the protein network code.

Using SOA and Cloud in a research project on CTCT (Cutaneous T-Cell Lymphoma)

Andreas DRESS[1]

[1] CAS-MPG, Shanghai Institutes for Biological Sciences (SIBS), China

Abstract

In this lecture, I will report on our vision (or dream) to develop:

- a CTCL model addressing the spatio-temporal dynamics of CTCL relevant cellular, protein, and miRNA networks within cancerous as well as non-cancerous tissue that integrates clinical, transcriptome, proteome, and toponome data,
- an expert platform for integrating such diverse data with clinical diagnosis and histopathology, providing —as a basis for expert discussion —the integrated data via the internet and, thus, allowing the project partners to continuously check and iteratively improve systems-biology models and their predictions in the light of new incoming data,
- and an IT platform supporting the use of our data and insights in *personalised* or —perhaps better —*precision medicine* for diagnosis as well as for devising and monitoring therapies in daily clinical routine.

A recent issue of Science celebrates 40 years of cancer research and its achievements over this time. While cancer research is certainly much older than that, the advent of molecular genetic technologies has revolutionised the field and ushered in a rapidly accelerating development of new technologies for elucidating the molecular biology of diseases.

Now, with increasing amount and refinement of *omics* data from genome, transcriptome, proteome, metabolome and other advanced and emerging high-throughput technologies applied to human diseases, we are ever more confronted with the task of processing large and continuously increasing *omics* data sets and projecting these data onto real-life patients, diseased organs, tissues, or cells. Of course, cancer – like any other complex disease or physiological state of life – is a complex biological phenomenon that cannot be described as an isolated event or reduced to a single factor. Hence, invoking systems biology and mathematical modelling for integrating the various data

from molecular and cell biology is indeed an inevitable requirement for success in current cancer research.

But the approach has to go beyond just these *omics* data sets and must incorporate morphological and histopathological features: Life is based on integrated systems of well-organised molecular and cellular networks, and diseases are deviations that can be described as alternative stable states thereof. In fact, clinical diagnostics for over 100 years has most successfully utilised morphological information from clinical descriptions and histopathology to define and classify diseases and to develop guidelines for therapy.

Such morphological approaches can now be augmented by new technologies based on vastly advanced (fluorescence) microscopy with resolutions from cell clusters down almost to single-molecule detection. In particular, TIS technology constitutes an outstanding example that offers the unprecedented and unique advantage of overcoming both the spectral limit and resolving power of standard fluorescence microscopy compared even to standard super-resolution imaging techniques such as STED, iPALM, and STORM. By providing fluorescence-microscopy images of multi-molecular distributions, it allows us to extract information on:

- the spatial organisation of whole protein networks,
- the hierarchical functionality of different proteins in these networks,
- and the control such networks exert on cells and cellular interaction,

thus establishing the basis for a whole new approach to cell biology, named toponomics.

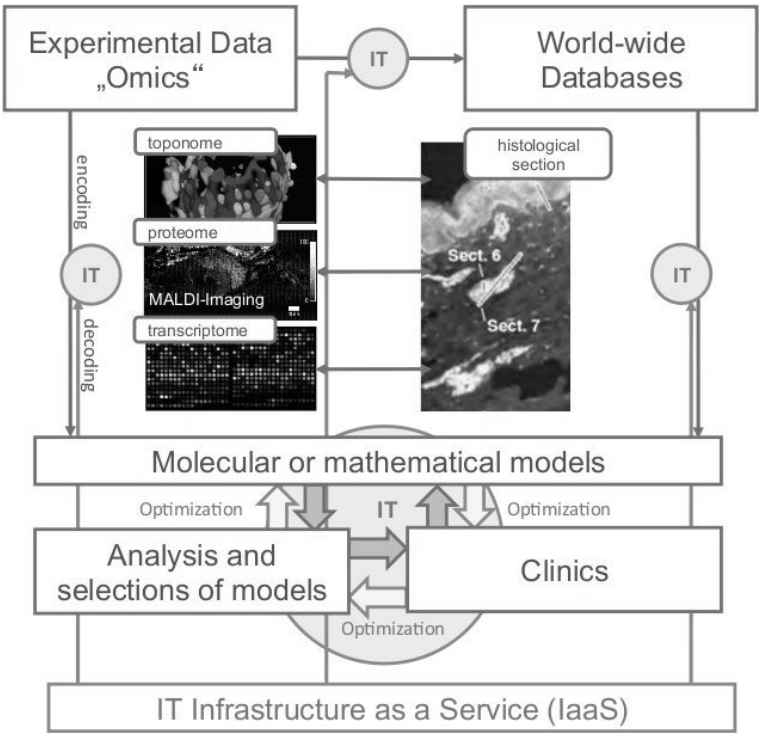
While toponome data can be aligned with other *omics* data, clinical knowledge and experience has been difficult to integrate with either of these. Problems are caused by the large variation of data formats and structures, levels of certainty and quantification, types of expertise required to comprehend and process the data, and means of communication.

This project addresses these problems in the particularly demanding context of applying systems biology to clinical research and development. We will:

- use known standard functional-genomics data that already exist,
- extend those data as required by the project's progress,
- complement and correlate them with toponome data and other morphological and histological information,
- while simultaneously involving clinicians right from the start

so that, in an iterative process, a validated systems-biology based model of CTCL will result.

The project will therefore not only produce new insights regarding the disease under investigation. It will also provide strategies and tools for clinically oriented systems biology in general. We regard our studies as an internationally unique approach to *clinical systems biology* or *systems medicine* that is likely to open up new avenues and provide new insights into individual tumour-specific mechanisms in cancer, and new ways to treat it.



Taking systems biology into the clinic: what we have learned from ciliopathies

Rachel Giles[1]

[1] Dept. Nephrology and Hypertension, University Medical Center Utrecht, Heidelberglaan 100, 3584CX Utrecht, The Netherlands

Abstract

In clinical medicine, chronic diseases are often oligogenic and complex to diagnose and treat. Transcriptomes, proteomes and metabolomes are currently being catalogued for such diseases yet the systems biology frameworks being generated by these datasets are primarily being used to study variation and function of the human genome and relating them to health and disease states. Recently enhanced efficiency of DNA sequencing allows powerful analytical computational and mathematical tools aimed at understanding functional and regulatory networks underlying the behavior of complex biological systems. Iterative systems approaches for specific human diseases such inherited renal cancer syndromes or renal ciliopathies have started making inroads to improving diagnostic and prognostic parameters. Because of their relatively contained yet oligogenic properties, in many ways the inherited renal cancer syndromes and ciliopathies offer an exemplary system to describe how systems approaches are transforming the way drugs are being developed based on complex interactions between distinct but overlapping pathways. Consequently, a perspective in which the interactions and dynamics are centrally integrated may steer medical intervention towards interrelationships of components. The optimal method for predictive, personalized, and preventive treatment of complex chronic diseases may therefore lie in systems medicine. We illustrate our arguments with a case report.

Multiscale Modelling: An Approach for Developing New Insight into Cancer?

Helen Byrne[1]

[1] OCCAM, Univ. Oxford, UK

Abstract

Until recently most experimental and theoretical studies of biological systems have focussed on a single scale. For example, at the tissue scale, one may measure how tumour size changes over time in response to different treatments. Alternatively, biomarkers for proliferation or apoptosis may provide information about whether a new drug reduces tumour mass by up-regulating apoptosis or down-regulating proliferation. Equally, proteomic and genomic data can be used to determine the subcellular signalling pathway by which a particular growth factor achieves its effect. In practice, these processes are inter-related and incredibly complex. For example, the rate at which a tumour grows depends on the rate of cell division; the rate of cell division depends on the rate at which cells progress through the cell cycle; the rate of progress through the cell cycle depends on the rate at which growth factors and vital nutrients are transported to the cells of interest, and this rate depends, in turn, on the size of the tumour!

In this talk, I will explain how multiscale models can be assembled and used to study and understand the complex interplay between processes that occur at different levels of spatial and temporal organisation. Attention will focus on two case studies: a multiscale model of vascular tumour growth and a multiscale model of the early, prevascular stages of colorectal cancer.

Modeling Gene Regulatory Networks: A Brief Overview of Current Approaches

Courtney Chancellor[1], Francisco Chinesta[2], Morgan Magnin[1], Olivier Roux[1]

[1] IRCCyN, Ecole Centrale de Nantes, France
[2] GeM, Ecole Centrale de Nantes, France

Abstract

In order to handle the increasing complexity of available data in the field of gene regulatory networks, computational tools have become invaluable. There exist many modeling frameworks to choose from, each bringing with it its own weaknesses and strengths, data requirements and fundamental assumptions. By casting a network in a particular model type, one inherently limits the resulting analysis to the bounds of that type. Here, we introduce three genres of frameworks, continuous, discrete and stochastic, and illustrate relative properties of each and apply it to a simple network of three genes.

1 Introduction

The mechanism by which a cell regulates genetic transcription and translation is often a complex network of dependencies broadly labeled as gene regulatory networks. As technology has advanced to aide the exploration and measurement of these previously unobservable systems, the quantity of available data has skyrocketed, drawing newfound connections between proteins and their encoding genes. As a result, the number of potentially interacting species in a regulatory network for any one protein may be too large to understand intuitively. In contrast, the kinetics of these reactions are often unknown and threaten to remain elusive since, at such a level, it becomes difficult to experimentally isolate individual reactions and quantify the contribution of a single component to the system as a whole. If, however, we are able to properly construct a framework with which to model the system, it may be possible to study the network on a deep, nonintuitive level and suggest experimental design for more practical in vitro or in vivo observation [4]. Computational biology and modeling tools have become invaluable to this end.

In this overview, we introduce three genres of modeling frameworks and illustrate relative properties of each. For the sake of continuity and to more easily compare the above model types, a single system is introduced in Section 1 and will be referenced throught the paper. While it represents no physical system, this simple toy network can demonstrate the weaknesses and potential

power of each framework type.The classical continuous models are in the form of differential equations in Section 2, specifically their peicewise linear simplifications and hybrid types. Logical models, also known as Generalized Logical Networks, as found in Section 3, consider discrete, qualitative change and, although they can be thought of as the most simplified model in this overview, have the potential to capture complex behaviors and come with their own hybrid expansions. Finally, stochastic models derived from the chemical master equation offer a contrasting paradigm to the previous two deterministic models and are discussed in Section 4. This paper does not represent an complete literature review of all modeling frameworks currently used [1] but, rather, gives a brief guide to three principle categories and the links existing between them. For example, not detailed here but certainly worth investigation are Petri nets and the recently developed Process Hitting [12], a kind of process algebra.

1.1 Feed Forward Circuit

A system whose product further propagates a particular signal, say, a protein which acts as its own activator, is known as a feed-forward circuit or, in bio-jargon, a feed forward loop. These structures are very common in the biological world and must be properly represented by any potential model. The considered network, however, is an example of an "incoherent" feed-forward circuit in that a single species, here called x_1, acts as both activator and, via a secondary species x_3, inhibitor of x_2 at given thresholds $\theta_{i,j}$. [9, 11]

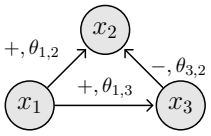


Figure 1: The toy network developed for this overview. In all examples we consider that $\theta_{1,2} > \theta_{1,3}$, that is, x_3 is transcribed before x_2. However, we wish to enforce behavior such that x_2 is present before the inhibiting effects of x_3 can take effect.

To represent the relationship described here, a weighted, directed graph proves to be an intuitive map of a gene regulatory network. Vertices signify the genes, proteins and cellular conditions which influence expression of a particular trait. The strength and type of interaction (inhibition and activation) are then denoted by weighted, directed edges between these variables. A visual representation of this graph can be found in Figure 1. On its own, the directed graph can give insight to the global behaviors and suggest possible connections within a regulatory network.

2 Continuous Models

There exists an extensive branch of modeling which considers a dynamic process in terms of a system of continuous, nonlinear differential equations [1]. This framework has the potential to model complex interactions between interacting species and gives rich, fully deterministic evolution through time. Here, species are represented by their concentrations, which are assumed to demonstrate continuous change as determined by the current state of the system described in a vecor, x. Thus, we may define the N individual species as $x_i(t), t \in \mathbb{R}_{\geq 0}$ which evolve as defined by the partial derivative $\frac{\partial x_i}{\partial t} = f_i(x)\ i = 1...N$, in which f_i may express complex interactions between species, the form of which must be known or derived from biological knowledge of the system. In the case that an analytical solution is possible, the system is quantitatively determined for all time and a wide range of tools are at ones disposal for analysis, including the determination of steady states, bifurcation analysis and the study of limit cycles. Unfortunately, the analytical solution is often impossible and, while numerical methods have been widely applied to nonlinear differential equations, these schemes do not always conserve realistic behaviors in gene regulatory networks. By making a few simplifying assumptions on the system, we hope to alleviate the worst of these problems.

2.1 Piecewise Linear Differential Equations

Piecewise Linear Differential Equations retain much of the same philosophy of their nonlinear counterparts in that the evolution of each species is governed by a partial derivative which outlines decay, growth, and interaction between species. However, in this simplified model we eliminate nonlinear terms by exploiting the characteristic form of regulation interactions.

The rate of change of a species is determined by the set of available resources in the current state, that is, the presence of its activators or absence of its inhibitors. In general, when one describes the interactions of biological regulators, one may say that the influence of a particular resource on its target is well represented by a sigmoidal function [1, 5]: factor x_j does not impact the target species x_i until reaching a given $\theta_{j,i}$, the threshold of influence of j on i, at which point it exhibits influence k_j^i. How sharp the threshold is can be captured with the choice of function. By using a step function such that change in influence is instantaneous and by assuming that the effects of resources, k_j^i, are additive, one eliminates nonlinearities from the differential system of equations and arrives at the PLDE network. Thus, we may rewrite the form of these equations [2] as

$$\frac{\partial x_i}{\partial t} = \left[k_0^i + \sum_{j \in \omega(i)} k_j^i \times \mathbb{F}_{x_j, \theta_{j,i}} \right] - \gamma x_i$$

in which we have explicitly defined k_0^i, a basal synthesis rate, and γ_i, a decomposition rate of x_i which is directly related to its concentration. $\mathbb{F}$ represents the sigmoid function of choice (though, in this paper, we will only utilize the simple step function), and $\omega(i)$, the resources of species i. The resulting differential system is segmented by the domains, but within each domain there always exists an analytical solution given an initial condition $x_i(0)$,

$$x_i(t) = \frac{k_0^i + \sum_{j\in\omega} k_j^i}{\gamma_i} - \left(\frac{k_0^i + \sum_{j\in\omega} k_j^i}{\gamma_i} - x_i(0) \right) e^{-\gamma_i t}$$

Although the state space of each variable is continuous, the nature of these interactions naturally divides this space into domains defined by the threshold values of resources. Within these segments, that is, when the available pool of resources do not pass any threshold values, all trajectories go to a focal state, $\frac{k_0^i+\sum_{j\in\omega} k_j^i}{\gamma_i}$, though not one necessarily contained within the current segment. Within the PLDE framework, one can observe many complex dynamical behaviors. Regular and singular steady states can be observed, with the assurance that all asymptotically stable steady states can be found [5]. Cyclical and aperiodic behaviors can also be demonstrated.

This complexity comes at the cost of computation. The problem of the PLDE becomes the search for the kinetic paramenters $\{k_j^i, \gamma_i, k_0^i\}$. Since these parameters are continuously valued, no automated, exhaustive search method is possible. Therefore, one must search for parameters which produce known behaviors with simulation techniques used to evaluate robustness, a fundamental characteristic of biological systems.To limit the possible space of parameters somewhat, constraints may be written under the guidance of biological knowledge and experimental data.

2.2 Hybrid Models

Because the PLDE framework contains an explicit definition of time, the incorporation of experimental temporal data is possible [2, 9]. It is often the case that how much time is required, say, between the instant an activator comes to its appropriate concentration and when its effects will be present in the system, is observable. In the case of our example model, let the threshold values be $\theta_{1,2} = 6$, $\theta_{1,3} = 5$ and $\theta_{3,2} = 7$. We wish to be sure that, although x_3 is transcribed first, in the time needed for it to reach its inhibiting concentration, x_2 will have already reached its active levels. We want to include a time delay such that this is always true. How much time is required to pass between each domain on an increasing order, $d_{x_i}^+$, (or decreasing order, $d_{x_i}^-$, respectively), can be analytically found by algebraic manipulation, and chains of constraints can be written to enforce a pathway in steps of two. So, in

our case, the desired pathway is $(2,0,0) \rightarrow (2,1,0)$, and we would intuitively write $d^+_{x_2}(2,0,0) < d^+_{x_3}(2,0,0)$. Since x_3 is activated slightly before x_2, we must incorporate the time it has already had into to transcribe into the delay constraint as $d^+_{x_2}(2,0,0) < d^+_{x_3}(2,0,0) - d^+_{x_1}(1,0,0)$. By incorporating time delays into the model, we introduce temporal domains based on the order of the system, effectively hybridizing it to a richer description at the cost of computation.

3 Discrete Models

First, recall the system as defined in the previous section, with each species conveyed by its current concentration at some time, $x_i(t)$, $i = 1 \cdots N$, with domains defined by thresholds $\theta_{i,j}$. In switching to the discrete paradigm, we must alter this definition in a fundamental way. Species are no longer continuous functions but, rather, are followed by qualitative– not quantitative– changes in state. Only when a variable crosses from one domain to another can be considered in describing the system. Although this seems to be a dramatic simplification, the Generalized Logical Network is capable of representing complex dynamics and comes with its own analytical tools [1].

3.1 Generalized Logical Network

In this framework, we concern ourselves only with qualtitative switches in a variable. We are given the current state of the system as described by what regions the factors relevant to a given variable are in- that is, its available resources- and want to determine, given those resources, to what region will the variable tend. This is called a logical image. The definition of resources remains the same from the previous section, but the variables are now defined by their domains, $x_i = 0, 1, \cdots m$. For every possible state, we must determine the image of each variable given the list of current resources, ω. The value given to each image is of the form $K_{i,\omega}$ and is equal to the numbered domain to which the variable tends to go. The search for these logical parameters defines the GLN completely, and, because its possible values are discrete and finite, an automated search is possible [2, 3]. The underlying graph usually supports a number of parameter sets exhibiting different kinds of behavior. As before, to limit the solution space, one may write contraints based on biological knowledge or the functionality of a positive or negative circuit [11].

The GLN as it applies to our example can be found in Figure 2. The choice of parameters can be influenced by our understanding of the system. Any interesting properties of the system depend on the strength of the negative influence of x_3 relative to the positive influence of x_1 on x_2. If the inhibition by x_3 far outweighs the activation by x_1, the parameters relevant to x_2 are ranked $K_{2,\{\}} \leq K_{2,\{1\}} \leq K_{2,\{3\}} \leq K_{2,\{1,3\}}$, the equality case representing

x_1	x_2	x_3	X_1	X_2	X_3
1	0	0	$K_1\{\}$	$K_2\{3\}$	$K_3\{1\}$
1	0	1	$K_1\{\}$	$K_2\{\}$	$K_3\{1\}$
1	1	1	$K_1\{\}$	$K_2\{\}$	$K_3\{1\}$
1	1	0	$K_1\{\}$	$K_2\{3\}$	$K_3\{1\}$
2	0	0	$K_1\{\}$	$K_2\{1,3\}$	$K_3\{1\}$
2	0	1	$K_1\{\}$	$K_2\{1\}$	$K_3\{1\}$
2	1	1	$K_1\{\}$	$K_2\{1\}$	$K_3\{1\}$
2	1	0	$K_1\{\}$	$K_2\{1,3\}$	$K_3\{1\}$

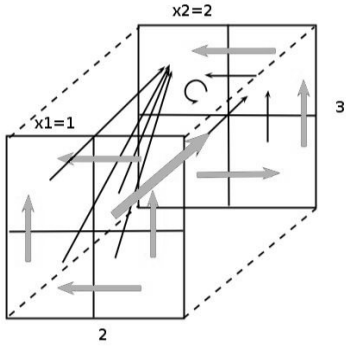


Figure 2: To solve a logical model, one must find the domain to which each species will tend given all possible combinations of resources (parameters $K_{i,\omega}$), also known as images. On the right can be found the table of images corresponding to the incoherent FFL with $\theta_{1,2} = 2, \theta_{1,3} = \theta_{3,2} = 1$. Naturally, the cases in which $x_1 = 0$ are of very little interest to us compared to those with active $x_1 = 1$ and 2, thus only these values are included in the table. Inhibition by x_3 is considered to be greater than activation by x_1, thus we arrive at solution $K_{1,\{\}} = 2$, $K_{2\{1,3\}} = K_{3\{1\}} = 1$, $K_{2,\{1\}} = K_{2,\{\}} = K_{2,\{3\}} = 0$ The values of the images correspond to the synchronous graph, demonstrated by the interior, black arrows. By allowing only one transition at a time, we come to the asynchronous graph, exterior blue arrows, in which any $x_1 = 1$ square may update to the $x_1 = 2$ plane, indicated by the large, blue arrow.

a weak influence not great enough to push the variable to its next domain. In contrast, the case of strong activation and weak inhibition is given by $K_{2,\{\}} \leq K_{2,\{3\}} \leq K_{2,\{1\}} \leq K_{2,\{1,3\}}$. Although this is not dramatically significant for this particular example, the writing of constraints is very important in the search for plausible parameters K. GLNs may also be subjected to their own sort of bifurcation analysis [10] in which parameters are modified one by one in accordance to some overarching property. This can be used to investigate deeper dynamic balances or isolate the most significant variables on the system.

A final modification must be made in the GLN framework: currently, the system evolves instantaneously from the current state to the image defined by the set of parameters $K_{i,\omega}$. In this representation, any number of variables are permitted to pass multiple thesholds in one time cycle. This is known as the synchronous state graph. While this may describe global behaviors, it does not fit our understanding of the physical system; There is little chance that any two reactive elements would pass thresholds simultaneously. To restore the plausibility of our model and the investigation of more local behaviors, we replace trajectories which pass through more than one transition state with single trajectories which modify only one element of the system at a time.

The new state graph is known as asynchronous, a comparison demonstrated by Figure 2, right hand side.

Certainly, the ability to use an automated search for the needed model parameters is a powerful feature of the logical framework. Although we have simplified our model and reduced how much we are able to know about local behavior, the GLN can represent rather complex dynamic behavior. The simplification may, in fact, be merited: too often a model overfits data and becomes a product of its framework rather than a representation of the actual system. There remain still, however, very strong ties between the PLDE and the GLN framework when one considers the division of the state space into domains, and one may move easily from the former to the later by defining parameters $K_{i,\omega}$ by their corresponding focal points $\frac{k_0^i+\sum_{j\in\omega} k_j^i}{\gamma_i}$. The relationship between PDE, PLDE and GLN frameworks is clearly visible.

3.2 Hybrid Models

The loss of an explicit definition of time, however, can be detrimental. In our example, it is impossible to enforce a pulse in x_2, that is, that the system always moves $(1,0,0) \to (2,0,0) \to (2,1,0) \to (2,1,1) \to (2,0,1)$. This is due to the structure of the asynchronous graph: the system will always have equal potential to increase x_3 as x_1 or x_2, thus is non-deterministic. Intuitively, we know that the effects of x_3 on x_2 will take time to sufficiently inhibit its expression. By incorporating time delays into the discrete model, we can effectively hybridize the system and enforce desirable physical behaviors in our model [2, 9].

This is a relatively straightforward process: each variable x_i is associated with a clock h_{x_i} and delays for passing the decreasing domain d^- and the increasing domain d^+, as was the case in the hybrid PLDE framework. When the state is eligible to make a transition between domains, the clock starts, keeping continuous time until reaching delay d^+ (or d^-), at which point the transition fires and the clock is reset. A clock is reset whenever its variable changes order since it is therefore subject to a different domain.

4 Stochastic Models

The previous models have made hidden, fundamental assumptions which may not reflect the biological reality of gene regulation. The first is that concentrations of species in the system vary continuously. While this may seem a straightforward assumption, most regulatory agents exist at very low concentrations such that a change in a very small number of individuals can have large effects on the system as a whole. Secondly, biological events are not truly deterministic: identical genotypes can lead to drastically different phenotypic expression, and genetic regulation has proven to be sensitive to noise. Two systems which have the same initial state may exhibit different global behaviors.

If we abandon these, most likely, flawed assumptions we are led to describe the system as stochastic in nature. Rather than tracking the concentration of species as it varies in time according to an underlying deterministic process, we think about the system in terms of the probability of existing at a certain state at any given time given some initial condition.

4.1 The Chemical Master Equation

Since we no longer concern ourselves with the concentrations of species in the stochastic framework, we change syntax for ease of differentiation. Thus, the system is described by the vector $z = \{z_1 \cdots z_N\}$ where $z_i \in \mathcal{N}$ is the count of individual molecules of species i. From a given state, reactions occur which move it to a new configuration according to a known stoichiometry. These reactions are subject to some propensity. For any forward reaction r_j which occurs with propensity a_j the system moves as demonstrated in Figure 3, where v_j is a vector containing the net changes in the state vector.

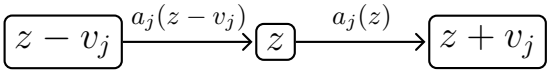


Figure 3: Visual representation of how the system evolves according to a single reaction, r_j with propensity function $a_j(z)$. When the reaction occurs, all species change according to the stoichiometry of the reaction, its net influence contained in v_j, to arrive at the new state.

The Chemical Master Equation describes the evolution of the probablity of the system existing at any given state z by considering the propensities of all reactions which leave z and those which enter z,

$$\frac{\partial P(z,t|z_0,t_0)}{\partial t} \equiv \sum_j [a_j(z - v_j)P(z - v_j, t|z_0, t_0) - a_j(z)P(z,t|z_0,t_0)]$$

To simplify this representation, we may aggregate these terms to express the CME in matrix form, $\frac{\partial P}{\partial t} = \mathcal{A}P$. Finding a solution to the CME has proven extremely difficult due to its high dimensionality: the number of degrees of freedom grows exponentially with the number of reacting species. Simulation methods are popular to this end, but do not necessarily resolve this issue. Monte Carlo methods demand large numbers of trials in order to approximate the posterior distribution, a barrier which can be prohibitive even in the light of approximation techniques such as time leaping or system partitioning. Here, we investigate Proper Generalized Decomposition, a method which approximates the probability of the system occupying a particular state as a finite sum of a product of separable functions, the form of which are a priori unknown [1, 6, 8].

4.2 Proper Generalized Decomposition

In PGD, the solution is assumed to be separable in terms of each species z_i and time. This is no small assumption but is, at worst, equivalent or superior to the dismissal of nonlinearities in the PLDE framework. The form of the solution is as follows:

$$P(z,t) = \sum_{j=1}^{n_F} \alpha^j F_1^j(z_1) \otimes F_2^j(z_2) \otimes \ldots \otimes F_N^j(z_N) \otimes F_t(t)$$

Given an initial guess for P, by plugging it back into the Chemical Master Equation one can find its residual, that is, how well or poorly it satisfies the equation. If it does not satisfy, one then searches for a new set of functions to be added to P which reduce the residual. In order to conserve properties of a probability distribution, these newfound functions are then projected to a normalized space by an appropriate α.
In general, this enrichment/projection algorithm is computationally inexpensive and flexible. Unknown parameters can easily be incorporated into PGD at the cost of an additional dimension, which, in the field of gene regulatory networks, is a considerable advantage. Note that the number of degrees of freedom in this new representation is of the order of $n_N \times N \times n_F$ rather than $(n_N)^N$. This is done by restricting the domain of the problem by assuming that the probability of occupying a state becomes negligible outside of some interval on the one dimentional grid, giving n_N degrees of freedom. We have thereby already substantially reduced the problem size. In addition, the presence of a function $F_t(t)$ allows us to solve for all time at each iteration.
Overcoming the computational barrier to a stochastic framework gives this modeling strategy a considerable advantage to its deterministic counterparts. However, even with the powerful PGD tool to guide its solution, the Chemical Master Equation depends on knowing the kinetics of a system and, therefore, may be confined to well-studied systems.

5 Conclusion

The question of gene regulatory networks remains open in the field of computational biology. Although there exists a variety of approaches, no one method fully satisfies in the depth of its description while remaining in the confines of available data and computational cost. The frameworks described here are representative of how each brings its own strengths and weaknesses, and even how some of these weaknesses might be overcome. New methods may be created by formally combining or linking existing ones, or possibly from the construction of novel paradigms. This search is of primary interest and will

be the topic of papers to come. Currently, connections between the Chemical Master Equation and a new formalism known as Process Hitting may be of interest [12].

References

[1] de Jong, H. (2002) Modeling and simulation of genetic regulatory networks: A literature review *Journal of Computational Biology* Vol 9,No1. pp 67-103.

[2] Comet, J.P. and Bernot, G. (2010) Introducing continuous time in discrete models of gene regulatory networks *Proc. of the Nice Spring school on Modelling and simulation of biological processes in the context of genomics. EDP Sciences, ISBN*.

[3] Bernot, G. and Comet, J.P. and Richard, A. and Guespin, J. & K. Skarstad, (2004) Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic *Journal of Theoretical Biology* Vol. 229, No 3:339-347.

[4] Filopon, D. and Mérieau, A. and Bernot, G. and Comet, J.P. and LeBerre, R. and Guery, B. and Polack, B. and Guespin-Michel, J. (2006) Epigenetic acquisition of inducibility of type III cytotoxicity in P. aeruginosa *BMC Bioinformatics*.

[5] El Houssine Snoussi. (1989) Qualitative dynamics of piecewise-linear differential equations: A discrete mapping approach. *Dynamics and Stability of Systems*, Vol 4, No 3.

[6] Ammar, A. and Cueto, E. and Chinesta, F. (2012) Reduction of the chemical master equation for gene regulatory networks using proper generalized decompositions *International Journal for Numerical Methods in Biomedical Engineering*.

[7] De Jong, H. and Gouzé, J.L. and Hernandez, C. and Page, M. and Sari, T. and Geiselmann, J. (2003) Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bulletin of Mathematical Biology*. Vol 66. No 2.

[8] Chinesta, F. and Ammar, A. and Leygue, A. and Keunings, R.(2011) An overview of the proper generalized decomposition with applications in computational rheology. *Journal of Non-Newtonian Fluid Mechanics*. Vol 166, No 11.

[9] Ahmad, J. and Bernot, G. and Comet, J.P. and Lime, D. and Roux, O.(2006) Hybrid modelling and dynamical analysis of gene regulatory networks with delays. *ComPlexUs*. Vol 3, No 4.

[10] Abou-Jaoude, Djomangan A; Ouattara, Marcelle Kaufman, (2009) Frequency tuning in the p53-Mdm2 network. I. Logical approach.*Journal of Theoretical Biology*. Vol 258.

[11] René Thomas, Denis Thieffry, and Marcelle Kaufman. (1995) Dynamical Behavior of Biological Regulatory Networks, I. Biological Role of Feedback Loops.*Bulletin of Mathematical Biology* Vol 57, No 2.

[12] Loic Paulevé, Morgan Magnin and Olivier Roux. (2011) Refining dynamics of gene regulatory networks in a stochastic π calculus. Transactions on computational systems biology. Vol 8

Introduction to the Process Hitting and inference of its underlying Biological Regulatory Network

Maxime Folschette[1]
Joint work: Loïc Paulevé, Katsumi Inoue, Morgan Magnin, Olivier Roux

[1] LUNAM Université, École Centrale de Nantes, IRCCyN UMR CNRS 6597
1 rue de la Noë – B.P. 92101 – 44321 Nantes Cedex 3, France.

Abstract

In this paper, the Process Hitting (PH), a recently introduced framework to model concurrent processes, is introduced. It is notably suitable to model Biological Regulatory Networks (BRNs) with partial knowledge of cooperations by defining the most permissive dynamics. On the other hand, the qualitative modeling of BRNs has been widely addressed using René Thomas' formalism, which is also depicted. A translation from PH to Thomas' representation of BRNs is finally presented. It relies on an analysis of all regulations to infer the Interaction Graph, then the possible parametrizations.

1 Introduction

As regulatory phenomena play a crucial role in biological systems, they need to be studied accurately. Biological Regulatory Networks (BRNs) consist in sets of either positive or negative mutual effects between the components. Besides continuous models of physicists, often designed through systems of ordinary differential equations, a discrete modeling approach was initiated by René Thomas in 1973 [16] allowing the representation of the different levels of a component, such as concentration or expression levels, as integer values. Nevertheless, these dynamics can be precisely established only with regard to some kind of "focal points", related to as Thomas' parameters, indicating the evolutionary tendency of each component. This modeling has motivated numerous works (see [12, 9, 15, 1]), and other approaches related to our work, which rely on temporal logic [7] and constraint programming [4, 5], aim at determining models consistent with partial data on the regulatory structure and dynamics. While the formal checking of dynamical properties is often limited to small networks because of the state graph explosion, the main drawback of this framework is the difficulty to specify Thomas' parameters, especially for large networks.

In order to address the formal checking of dynamical properties within very large BRNs, we recently introduced in [10] a new formalism, named the *"Process Hitting"* (PH), to model concurrent systems having components with a few qualitative levels. A PH describes, in an atomic manner, the possible

evolutions of a "process" (representing one component at one level) triggered by the hit of at most one other "process" in the system. This particular structure makes the formal analysis of BRNs with hundreds of components tractable [11]. PH is suitable, according to the precision of this information, to model BRNs with different levels of abstraction by capturing the most general dynamics.

In [6] we showed that starting from one PH model, it is possible to find the underlying interactions, then the underlying Thomas' parameters. This method relies on an exhaustive search of the interactions between components of the PH model, and an enumeration of the (possibly large) nesting set of valid parameters, so that the resulting dynamics are ensured to respect the PH dynamics, i.e. no spurious transitions are made possible. The first benefit of this approach is that it makes possible the construction refining of BRNs with a partial and progressively brought knowledge in PH, while being able to export such models in the Thomas' framework. Our second contribution is to enhance the knowledge of the formal links between both modelings. The method can be applied to large BRNs (up to 40 components).

2 Frameworks

2.1 The Process Hitting framework

A Process Hitting (PH) (Def. 1) gathers a finite number of concurrent *processes* grouped into a finite set of *sorts*. A sort stands for a component of the system while a process, which belongs to a unique sort, stands for one of its expression levels. A process is noted a_i where a is the sort and i is the process identifier within the sort a. At any time, exactly one process of each sort is present; a *state* of the PH corresponds to such a set of processes.

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 $a_i \rightarrow b_j \nearrow b_k$, 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* $a \neq b \Rightarrow L_a \cap L_b = \emptyset$*;*
- $\mathcal{H} = \{a_i \rightarrow b_j \nearrow b_k \in L_a \times L_b \times L_b \mid (a, b) \in \Sigma^2 \wedge b_j \neq b_k \wedge a = b \Rightarrow a_i = b_j\}$ *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 $h = a_i \to b_j \nearrow b_k \in \mathcal{H}$ 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 [10], 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 bc between sorts b and c, defined with 4 processes (one for each sub-state of the presence of processes b_1 and c_1). For the sake of clarity, processes of bc are indexed using the sub-state they represent. Hence, bc_{01} represents the sub-state $\langle b_0, c_1 \rangle$, and so on. Each process of sort b and c hit bc, which makes it bounce to the process reflecting the status of the sorts b and c (e.g., $b_1 \to bc_{00} \nearrow bc_{10}$ and $b_1 \to bc_{01} \nearrow bc_{11}$). Then, to represent the cooperation between processes b_1 and c_1, the process bc_{11} hits a_1 to make it bounce to a_2 instead of independent hits from b_1 and c_1. The same cooperative sort is used to make b_0 and c_0 cooperate to hit a_1 and make it bounce to a_0.

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

$$L_a = \{a_0, a_1, a_2\}, \qquad L_b = \{b_0, b_1\},$$
$$L_{bc} = \{bc_{00}, bc_{01}, bc_{10}, bc_{11}\}, \qquad L_c = \{c_0, c_1\}.$$

This example models a BRN where the component a has three qualitative levels, components b and c are Boolean and bc is a cooperative sort. In this BRN, a inhibits b at level 2 while b and c activate a with independent actions (e.g. $b_0 \to a_2 \nearrow a_1$) or through the cooperative sort bc (e.g. $bc_{11} \to a_1 \nearrow a_2$). Indeed, the reachability of a_2 and a_0 is conditioned by a cooperation of b and c, as explained above.

A Process Hitting model can be obtained from the literature or from a BRN as described in [10]. In both methods, the identification of interactions allows to define the set of actions leading to the desired dynamics, but an under- or over-approximation can also be built if the interactions are not precisely known (by adding or removing all actions allowing a given behavior). This can be used especially in cases where a cooperative sort cannot be built because of a lack of information.

2.2 Thomas' modeling

Thomas' formalism, here inspired by [13, 3], lies on two complementary descriptions of the system. First, the *Interaction Graph* (IG) models the structure

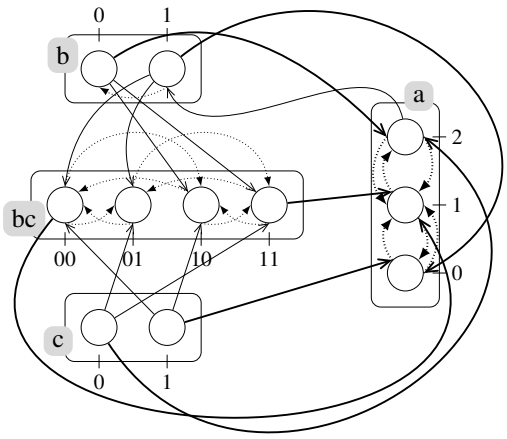


Figure 1: A PH example with four sorts: three components (a, b and c) and a cooperative sort (bc). Actions targeting processes of a are in thick lines.

of the system by defining the components' mutual influences. Its nodes represent components, while its edges labeled with a threshold stand for either positive or negative interactions (Def. 2); l_a denotes the maximum level of a component a.

Definition 2 (Interaction Graph) *An* Interaction Graph *(IG)* (Γ, E_+, E_-) *is a triple where:*

- Γ *is a finite number of* components*,*
- E_+ *(resp.* E_-*)* $\subset \{a \xrightarrow{t} b \mid a, b \in \Gamma \wedge t \in [1; l_a]\}$ *is the set of positive (resp. negative)* regulations *between two nodes, labeled with a* threshold*.*

A regulation from a *to* b *is unique, i.e. if* $a \xrightarrow{t} b \in E_+$ *(resp.* E_-*), then there is no regulation* $a \xrightarrow{t'} b$ *in* E_- *(resp.* E_+*), and no other regulation* $a \xrightarrow{t''} b$ *in* E_+ *(resp.* E_-*) with* $t'' \neq t$*.*

For an interaction of the IG to take place, the expression level of its head component has to be higher than its threshold; otherwise, the opposite influence is expressed. For any component $a \in \Gamma$, $\Gamma^{-1}(a)$ is the set of its regulators:

$$\Gamma^{-1}(a) = \{b \in \Gamma \mid \exists b \xrightarrow{t} a \in E_+ \cup E_-\} \ .$$

A *state* s of an IG (Γ, E_+, E_-) is an element in $\prod_{a\in\Gamma}[0; l_a]$ and $s[a]$ refers to the level of component a in s.

The specificity of Thomas' approach lies in the use of discrete *parameters* to represent focal level intervals (Def. 3). The use of intervals instead of single

values for parameters allows a wider range of expressiveness, by allowing behaviors impossible to define with single values.

Definition 3 (Discrete parameter $K_{x,A,B}$ and Parametrization K) *Let $x \in \Gamma$ be a given component and A (resp. B) $\subset \Gamma^{-1}(x)$ a set of its* activators *(resp.* inhibitors*), such that $A \cup B = \Gamma^{-1}(x)$ and $A \cap B = \emptyset$. The discrete* parameter *$K_{x,A,B} = [i; j]$ is a non-empty interval so that $0 \leq i \leq j \leq l_x$. With regard to the dynamics, x will tend towards $K_{x,A,B}$ in the states where its activators (resp. inhibitors) are the regulators in set A (resp. B), except in the case where $x \in K_{x,A,B}$ for which it does not evolve.*
The complete map $K = (K_{x,A,B})_{x,A,B}$ of discrete parameters for an IG is called a parametrization *of this IG.*

At last, dynamics are defined in BRN in a unitary and asynchronous way: from a given state s, a transition to another state s' is possible provided that only one component a will evolve of exactly one level towards $K_{a,A,B}$, where A (resp. B) is the set of activators (resp. inhibitors) of a in s, provided that $a \notin K_{a,A,B}$ in s.

Example 2 *Fig. 2(left) represents the Interaction Graph (Γ, E_+, E_-) with $\Gamma = \{a, b, c\}$, and:*

$$E_+ = \{b \xrightarrow{1} a, c \xrightarrow{1} a\} \qquad E_- = \{a \xrightarrow{2} b\} \ .$$

In particular, $\Gamma^{-1}(a) = \{b, c\}$. Fig. 2(right) gives a possible parametrization of this IG. In this BRN, the following transitions are possible:

$$\langle a_0, b_1, c_1\rangle \to \langle a_1, b_1, c_1\rangle \to \langle a_2, b_1, c_1\rangle \to \langle a_2, b_0, c_1\rangle \to \langle a_1, b_0, c_1\rangle,$$

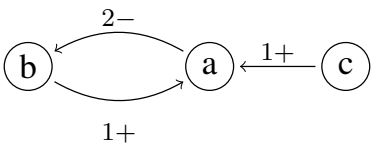
where a_i is the component a at level i.

3 BRN Inference

This section focuses on the inference of a complete BRN with Thomas' parameters from a given PH.

In order to infer a BRN, one has to find the Interaction Graph (IG) first, as some constraints on the parametrization rely on it. Inferring the IG is an abstraction step which consists, from atomistic actions of a PH, in determining the global influence of every component on each of its successors.

Then, given the IG inferred from a PH, one can find the discrete parameters that model the behavior of the studied PH. As some parameters may remain undetermined, another step allows to enumerate all parametrizations compatible with the inferred parameters.



$$K_{a,\{b,c\},\emptyset} = [2;2] \qquad K_{b,\{a\},\emptyset} = [0;1]$$
$$K_{a,\{b\},\{c\}} = [1;1] \qquad K_{b,\emptyset,\{a\}} = [0;0]$$
$$K_{a,\{c\},\{b\}} = [1;1]$$
$$K_{a,\emptyset,\{b,c\}} = [0;0] \qquad K_{c,\emptyset,\emptyset} = [0;1]$$

Figure 2: (left) IG example. Regulations are represented by the edges labeled with their sign and threshold. For instance, the edge from b to a is labeled "1+", which stands for: $b \xrightarrow{1} a \in E_+$. (right) One admissible parametrization of the left IG.

3.1 Interaction Graph inference

This step assumes that the studied PH defines two types of sorts: the sorts corresponding to BRN components, which will appear in the IG, and the cooperative sorts, as defined in Subsect. 2.1. The identification of these two sets of sorts relies on the observation of their possible behavior, which in both cases observe some rules, and can be automated. For instance, given the definition of cooperative sorts, if the actions hitting a sort lead to a unique fixed point for any configuration of its predecessors, then we can deduce that this sort is a cooperative sort. Conversely, because of the BRN dynamics explained in Subsect. 2.2, if all actions hitting a sort make its processes bounce at most one level away (e.g. if a_1 can bounce to a_0 or a_2 but not to a_3), then this sort is likely to correspond to a BRN component.

Inferring global influences of a predecessor b on a component a requires to find "local influences" from this predecessor first, by considering a given state of the PH and changing only the active process of b. The aim is to compare the set of processes towards which the component a will evolve, for each active process of b, leaving the active process of all the other sorts unchanged. Indeed, if after increasing the level of b, i.e. activating a higher process of b, we notice that a tends to reach a higher (resp. lower) level, we can then deduce that b activates (resp. inhibits) a in this selected state. Of course, only predecessors of a have to be considered.

This has to be observed on every possible state in order to infer a local influence. Indeed, if all local influences of b on a are the same (activations or inhibitions) we can deduce that the global influence of b on a is also the same, and the related threshold is the lowest level of b for which we can observe such an influence. An unsigned edge with no threshold is inferred if two different local influences are found, or in other particular cases (when a behavior cannot be represented as a BRN).

Example 3 *Consider, in the PH of Fig. 1, the sub-state $\sigma = \langle b_0, c_0, bc_{00}\rangle$ of predecessors of a. In this sub-state, a can be hit by the following actions:*

$$\mathcal{H}_a^\sigma = \{b_0 \to a_2 \Rsh a_1, c_0 \to a_2 \Rsh a_1, bc_{00} \to a_1 \Rsh a_0\} \ .$$

Thus, if a evolves, it will eventually reach process a_0. But if a higher process of b is activated, that is, b_1 instead of b_0, thus considering the sub-state $\sigma' = \langle b_1, c_0, bc_{10}\rangle$, then a can be hit by the two following actions:

$$\mathcal{H}_a^{\sigma'} = \{b_1 \to a_0 \Rsh a_1, c_0 \to a_2 \Rsh a_1\} \ ,$$

and will eventually reach process a_1.

Therefore, in this sub-state of predecessors of a, b locally activates a. Furthermore, if this analysis is carried for all possible sub-states of predecessors of a, only local activations are found, thus giving: $b \xrightarrow{1} a \in E_+$.

After applying this method to all pairs of influence, the IG given in Fig. 2 is inferred.

3.2 Parameters inference

This subsection presents some results related to the inference of independent discrete parameters from a given PH, equivalent to those presented in [10]. We suppose in the following that the considered PH is well-formed for parameters inference, i.e. its inferred IG does not contain any unsigned edge, and in each sort, all processes activating (resp. inhibiting) another component share the same behavior. Let $K_{a,A,B}$ be the parameter we want to infer for a given component $a \in \Gamma$, and $A \subset \Gamma^{-1}(a)$ (resp. $B \subset \Gamma^{-1}(a)$) a set of its activators (resp. inhibitors). This inference, as for the IG inference, relies on the search of processes of a towards which it will eventually evolve for the given configuration A, B of its regulators.

For each sort $b \in \Gamma^{-1}(a)$, we define a context that contains all processes of b activating (resp. inhibiting) a if $b \in A$ (resp. B). From all contexts of all predecessors of a, we create a global context $C_{A,B}$ that represents the configuration A, B (including the cooperative sorts involved). The parameter $K_{a,A,B}$ specifies towards which values a eventually evolves as long as the configuration A, B holds, which can now be computed by considering the dynamics of a in the global context $C_{A,B}$.

Example 4 *Consider the PH of Fig. 1, from which the IG of Fig. 2 is inferred. Inferring the parameter $K_{a,\{b,c\},\emptyset}$ requires to understand the behavior of a in the sub-state $\langle b_1, c_1, bc_{11}\rangle$. In this sub-state, a tends to eventually reach process a_2; thus, we can deduce the parameter: $K_{a,\{b,c\},\emptyset} = [2; 2]$. Inferring all parameters leads to the complete parametrization given in Fig. 2.*

3.3 Admissible parametrizations enumeration

The previous inference step may leave several parameters undetermined, due to missing cooperations or behaviors impossible to represent in a BRN. If it is not possible to change the PH model in order to remove these inconclusive cases, one can perform a last step to enumerate all valid values for each parameter that could not be inferred given the above results. We consider that a parameter is valid if any transition it involves in the resulting BRN is allowed by the studied PH by actions that represent this behavior. We also add some biological constraints on the whole parametrizations, given in [3]. These constraints lead to a family of admissible parametrizations which we can enumerate and are ensured to observe a coherent behavior that is included in the original PH.

Answer Set Programming (ASP) [2] turns out to be effective for the enumerative searches developed in this paper, as it efficiently tackles the inherent complexity of the models we use, thus allowing an efficient execution of the formal tools developed. Furthermore, ASP finds a particularly interesting application in the research of admissible parametrizations regarding the properties presented above, as this enumeration can be naturally formulated by using of aggregates and constraints.

3.4 Implementation

The inference method described in this paper has been implemented as a tool named `ph2thomas`, as part of PINT[1], a library gathering PH related tools. Our implementation mainly consists of ASP programs that are solved using Clingo[2].

In the previous sections, the methods and results are illustrated on a toy example considered as a very small network containing 3 components (a, b and c). But our approach can also successfully handle large PH models of BRNs found in the literature such as an ERBB receptor-regulated G1/S transition model from [14] which contains 20 components, and a T-cells receptor model from [8] which contains 40 components[3]. For each model, IG and parameters inferences are performed together in less than a second on a standard desktop computer.

4 Conclusion

This work establishes the abstraction relationship between PH, which is more abstract and allows incomplete knowledge on cooperations, and Thomas' approach for qualitative BRN modeling. This motivates the concretization of PH

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

[2] Available at `http://potassco.sourceforge.net`

[3] Both models are available as examples distributed with PINT.

models into a set of compatible Thomas' models in order to benefit from the complementary advantages of these two formal frameworks and extract some global information about the influences between components.

As an extension of the present work, we plan to explore new semantics of BRNs to be able to tackle influences currently represented by unsigned edges.

Acknowledgement

This work was partially supported by the Fondation Centrale Initiatives.

References

[1] Jamil Ahmad, Olivier Roux, Gilles Bernot, Jean-Paul Comet, and Adrien Richard. Analysing formal models of genetic regulatory networks with delays. *International Journal of Bioinformatics Research and Applications (IJBRA)*, 4(2), 2008.

[2] Chitta Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2003.

[3] Gilles Bernot, Franck Cassez, Jean-Paul Comet, Franck Delaplace, Céline Müller, and Olivier Roux. Semantics of biological regulatory networks. *Electronic Notes in Theoretical Computer Science*, 180(3):3 – 14, 2007.

[4] Fabien Corblin, Eric Fanchon, and Laurent Trilling. Applications of a formal approach to decipher discrete genetic networks. *BMC Bioinformatics*, 11(1):385, 2010.

[5] Fabien Corblin, Eric Fanchon, Laurent Trilling, Claudine Chaouiya, and Denis Thieffry. Automatic inference of regulatory and dynamical properties from incomplete gene interaction and expression data. In *IPCAT*, volume 7223 of *LNCS*, pages 25–30. Springer, 2012.

[6] Maxime Folschette, Loïc Paulevé, Katsumi Inoue, Morgan Magnin, and Olivier Roux. Concretizing the process hitting into biological regulatory networks. In David Gilbert and Monika Heiner, editors, *Computational Methods in Systems Biology*, Lecture Notes in Computer Science, pages 166–186. Springer Berlin Heidelberg, 2012.

[7] Z. Khalis, J.-P. Comet, A. Richard, and G. Bernot. The SMBioNet method for discovering models of gene regulatory networks. *Genes, Genomes and Genomics*, 3(special issue 1):15–22, 2009.

[8] Steffen Klamt, Julio Saez-Rodriguez, Jonathan Lindquist, Luca Simeoni, and Ernst Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(1):56, 2006.

[9] Aurélien Naldi, Elisabeth Remy, Denis Thieffry, and Claudine Chaouiya. A reduction of logical regulatory graphs preserving essential dynamical properties. In *Computational Methods in Systems Biology*, volume 5688 of *LNCS*, pages 266–280. Springer, 2009.

[10] 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.

[11] Loïc. Paulevé, Morgan Magnin, and Olivier Roux. Static analysis of biological regulatory networks dynamics using abstract interpretation. *Mathematical Structures in Computer Science*, in press, 2012. Preprint: `http://loicpauleve.name/mscs.pdf`.

[12] Adrien Richard and Jean-Paul Comet. Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, 155(18):2403 – 2413, 2007.

[13] Adrien Richard, Jean-Paul Comet, and Gilles Bernot. *Modern Formal Methods and App.*, chapter Formal Methods for Modeling Biological Regulatory Networks, pages 83–122. 2006.

[14] Ozgur Sahin, Holger Frohlich, Christian Lobke, Ulrike Korf, Sara Burmester, Meher Majety, Jens Mattern, Ingo Schupp, Claudine Chaouiya, Denis Thieffry, Annemarie Poustka, Stefan Wiemann, Tim Beissbarth, and Dorit Arlt. Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Systems Biology*, 3(1), 2009.

[15] Heike Siebert and Alexander Bockmayr. Incorporating time delays into the logical analysis of gene regulatory networks. In *Computational Methods in Systems Biology*, volume 4210 of *LNCS*, pages 169–183. Springer, 2006.

[16] René Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563 – 585, 1973.

Symbolic Modelling for Reverse Engineering of Thomas Networks Parameters

Emmanuelle Gallet[1], Matthieu Manceny[2], Pascale Le Gall[1], Paolo Ballarini[1]

[1]MAS, École Centrale Paris
[2]LISITE, ISEP, Paris

Abstract

In this paper we describe a computer-aided methodology for reverse engineering of parametric models of Genetic Regulatory Networks (GRN). In the discrete framework introduced by R.Thomas, when observations over time are translated into temporal logic properties, then classical model-checking techniques can state whether a model with instantiated parameters satisfies or not a given observation. In order to avoid the high level of combinatorial choices of parameters even for networks of small size, we customise model-checking techniques by manipulating classes of models by constraints over parameters: instances fulfilling the desired temporal property are defined by parameter values satisfying the resulting constraint. We provide insights concerning our prototype and we illustrate our approach with a 6-gene network governing the inducibility of cytotoxicity in *Pseudomonas Aeruginosa*.

1 Context

1.1 Thomas' discrete modelling of regulatory networks

Biological considerations. Gene expression is a biological process by means of which proteins are synthesised as the end result of a process that consists of two basic steps: *transcription* (through which an mRNA molecule is produced by a gene) and *translation* (through which the resulting protein is obtained from the transcribed mRNA). The dynamics of gene expression depends on the presence of other proteins, referred to as *transcription factors* (TF); if the concentration of a TF is sufficient, it may activate or inhibit the transcription of a gene into mRNA, hence it regulates the synthesis of the end proteins.

The expression of one gene may be regulated by several TFs, including the protein issued by the gene itself (*i.e.* self-regulation). The entire collection of regulatory inter-dependencies for a given set of genes is called a *Genetic Regulatory Network* (GRN).

Interaction graph. An interaction graph is a formal and abstract representation of these interactions, it is a directed graph whose vertices represent the

elements (*i.e.* the genes) of the GRN and whose arcs represent the *interactions* between them. Each edge $\alpha \to \beta$ is labelled with a *sign* ("+" or "−") to indicate if α can activate $(+)$ or inhibit $(-)$ the expression of β. In Thomas' models [11, 10], continuous concentrations are discretised into a finite set of so-called *levels of expression*. We denote by x_α the level of expression of the gene α. Thus, each edge $\alpha \to \beta$ also carries a positive integer, called *threshold*, which corresponds to the minimal level of expression α needs in order to influence the expression of β. If a value labels an outgoing edge from α, then all lower (non null) values must be used on edges outgoing from α.
An example of interaction graph is given in Figure 1, with two genes: α and β. α activates β with a threshold of 1 and itself with a threshold of 2, and β inhibits α with a threshold of 1.

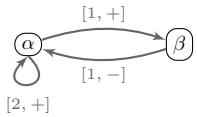


Figure 1: Example of interaction graph

Dynamics. We call dynamics (or models) of a GRN the evolution over time of the levels of expression of genes. For a gene α, the evolution of its level of expression (increase, decrease or stagnation) depends on the genes regulating it and is given by a set of *parameters* which indicates the value to which x_α tends. These parameters are of the form $K_\alpha(\omega)$ where ω denotes the subset of regulating genes β with a level of expression above the threshold of the corresponding action $\beta \to \alpha$. We consider an asynchronous model of dynamics, where the levels of expression change at most of one unit and one after one. From a state (*i.e.* a possibility of levels of expression of genes), there may be several possible next states; we do not know the speed of reactions (such as synthesis or binding of regulatory proteins), the model contains thus all the possibilities of evolution. Figure 2 gives an example of representation of one dynamic for the interaction graph in Figure 1 (with $K_\alpha(\{\}) = 2$, $K_\alpha(\{\alpha\}) = 2$, $K_\alpha(\{\beta\}) = 0$, $K_\alpha(\{\alpha, \beta\}) = 1$, $K_\beta(\{\}) = 0$ and $K_\beta(\{\alpha\}) = 1$), in the form of a *state transition graph*. Each vertex represents a state and the arcs represent the transitions between a state and its successors. For example, the successors of the state $(x_\alpha, x_\beta) = (1, 0)$ are the states $(2, 0)$ and $(1, 1)$ because x_α tends to $K_\alpha(\{\}) = 2$ (since neither x_α nor x_β are above their respective thresholds of regulation of α) and x_β tends to $K_\beta(\{\alpha\}) = 1$ (since x_α is above the threshold of $\alpha \to \beta$).

Exhaustive analysis of possible dynamics. Different values of the parameters lead to different dynamics. The actual values of such parameters

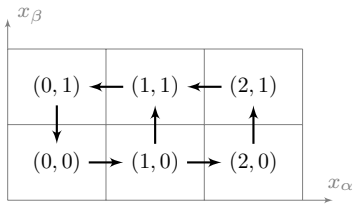


Figure 2: A state transition graph of the interaction graph in Figure 1

depend on a set of biological properties (for instance the affinity between proteins which can cause formation of protein complexes) and they are mostly unknown. Thus, in order to study the entire set of possible dynamics of a GRN, one has to consider each possible combination of values, *i.e.* more or less $\prod_{\alpha \in G}(|G^+(\alpha)| + 1)^{2^{|G^-(\alpha)|}}$, with $G^+(\alpha)$ the set of genes regulated by α and $G^-(\alpha)$ the set of genes regulating α. This explains the high level of combinatorial explosion of the problem of parameter identification of a GRN.

1.2 Computer-aided search of parameters

Biological knowledge and hypothesis. Let us stress that not all of the dynamics included within a GRN are consistent with observed (*in vivo* or *in vitro*) biological behaviours or with biological hypotheses. For instance, experimenters can detect homoeostasis and non-accessibility or ordered sequence of combinations of expression levels. This knowledge can be used directly to determine the value of some parameters or can be translated in the form of constraints that parameters should comply with.

Use of model-checking techniques. Computationally hard problems, such as that of exhaustively searching the parameter space of a GRN model, may be dealt thanks to computer science.

In Bernot *et al.*[1], a given model of the GRN is considered and model-checking techniques are applied to verify whether it fulfils some relevant biological properties. In order to exhaustively search the parameters' space, the model-checking procedure is re-iterated over all possible models defined by all possible parameters' values. As a consequence such a schema may be applied only to small GRNs with few unknown parameters. This approach is implemented in the SMBioNet tool [9], which supports the encoding of standard properties (such as functional cycles or parameter sets with shared transition graph) into formulae of *Computation Tree Logic* (CTL).

Klarner *et al.*[5] define an approach which considers different models of a GRN by using an encoding technique enabling the sharing of computations

between different models. Sets of models are encoded by a binary vector, one bit (or colour) per model, and (coloured) model-checking algorithms are extended with Boolean operations on vectors to identify models fitting time series measurements. These, in turns, are expressed through the temporal operator *Finally* ($\mathbf{F}$) of *Linear Temporal Logic* (LTL).

Similarly, our approach [8] considers several instances of a GRN at the same time. But unlike [5], instances are not manipulated using an explicit enumeration, but are implicitly referenced as the solutions of constraints defined on the parameters. For that, a *parametric* GRN modelling is considered (*i.e.* a modelling that represents a number of possible regulatory dependencies, hence a set of GRN models) alongside a *target behaviour* (a property that the model should meet). Such an approach may be referred to as *reverse engineering of a parametric GRN*. In Corblin *et al.* [3], the tool GNBox is based on constraint logic programming and target behaviours are expressed as some kinds of finite paths that models have to satisfy; a more recent tool of Corblin *et al.*, SysBiOX [2], uses also spatio-temporal data in order to make a selection within possible interaction graphs. Our approach also uses a parametric GRN modelling: biological parameters are processed as symbols within constraints and we use temporal formulae expressed in LTL over the set of expression levels of genes. These formulae are built from a set of atomic propositions using the usual logical operators in $\{\top, \bot, \neg, \wedge, \vee\}$ and the temporal operators $\mathbf{X}$ (for neXt time), $\mathbf{G}$ (Globally), $\mathbf{F}$, $\mathbf{U}$ (Until) and $\mathbf{R}$ (Release). For instance, the existence of a steady state (*i.e.* a state which is itself its own successor) in the dynamics of the interaction graph in Figure 1 in $(x_\alpha, x_\beta) = (2, 1)$ is encoded with the following formula: $\mathbf{G}\big((x_\alpha = 2 \wedge x_\beta = 1) \rightarrow \mathbf{X}(x_\alpha = 2 \wedge x_\beta = 1)\big)$.
Using parameterised formulae to denote a set of models allows us *a priori* to be less subject to the problem of size of GRNs.

2 Symbolic LTL model-checking

PTS models. The methodology for reverse engineering of Thomas' networks is based on a specific type of modelling formalism, namely that of *Parametric Transitions Systems* (PTSs) [7]. A PTS is essentially a *symbolic* form of state-transition graph: *i.e.* a transition system (TS), whose dynamics (*i.e.* transitions) are represented by means of some parameters.
Starting from the interaction graph of a GRN, a PTS is built so that it captures all the *static* and *dynamic constraints* of gene expression concisely. The *dynamics constraints* (encoded as logical formulae over biological parameters and expression levels labelling the transitions of a PTS) are the conditions representing the evolution of gene expression. The *static constraints*, on the other hand, are the conditions that are always fulfilled (encoded as logical formulae labelling the states of a PTS).

PTS instantiation through LTL model-checking. Given a PTS representing a certain (parametric) GRN modelling, the problem we face is to find out instances of the PTS (*i.e.* actual TSs obtained by assigning values to the GRN parameters) that fulfils a target LTL formula. This can be achieved through an adaptation of the LTL model checking problem to the PTS-based parameter search problem, thus resulting in the so-called symbolic LTL model-checking schema (Figure 3). Intuitively such a schema works as follows: given a PTS model ($\mathcal{M}$) and a target LTL formula (ϕ) , whose negation ($\neg\phi$) is encoded by a Büchi Automaton (BA, *i.e.* $\mathcal{A}_{\neg\phi}$), the cross-product model ($\mathcal{M} \times \mathcal{A}_{\neg\phi}$), named *Accepting PTS* (APTS), is built. The adaptation of LTL model checking to PTS-based parameter search, hence, boils down to find models that allow for paths in the APTS containing at least an accepting state occurring infinitely often in it, in accordance with the properties of the BA ($\mathcal{A}_{\neg\phi}$). Searching of paths (*i.e.* symbolic executions) that reach an accepting location of the APTS requires looking for solutions of the (symbolic) constraints labelling the edges of the PTS. To this aim, in every state of the ATPS (starting from the initial one), a constraint solving step is applied (specifically in our prototype tool we employ the CHOCO constraint solver [6]) to find out which transitions are enabled, and hence continuing in the construction of symbolic executions that lead to an accepting path. Each accepting path gives us a constraint on parameters (the set of formulae that has been verified to reach a node). With this constraints, we can determine the dynamical behaviours of the GRN compatible with the biological knowledge. Figure 3 summarises the different steps of such procedure.

Implementation. The PTS-based method has been implemented in a prototype software tool (in Java) called SPUTNIK[1]. In order to assess the SPUTNIK tool we considered a classical case-study of epigenetic switch in bacteriophage lambda (4 genes, 10 interactions) which has also been analysed in [5], through the coloured LTL model-checking approach. The number of solutions obtained with SPUTNIK (*i.e.* 480 solutions out of roughly 7 billion total possibilities) with a runtime of a few minutes, is comparable (same order of magnitude) to those illustrated in [5] (the difference in the number of solutions is probably due to the difference in the initial constraints).

3 Example

We develop an example taken from [4], concerning the inducibility of cytotoxicity in *Pseudomonas Aeruginosa*. The goal of this study is to test the possibility of an epigenetic switch of the *Type III Secretion System* (T3SS) of the bacterium, by computer modelling and experimentation (in [4] this has

[1] *Symbolic Parameters of Thomas' Networks Inference*

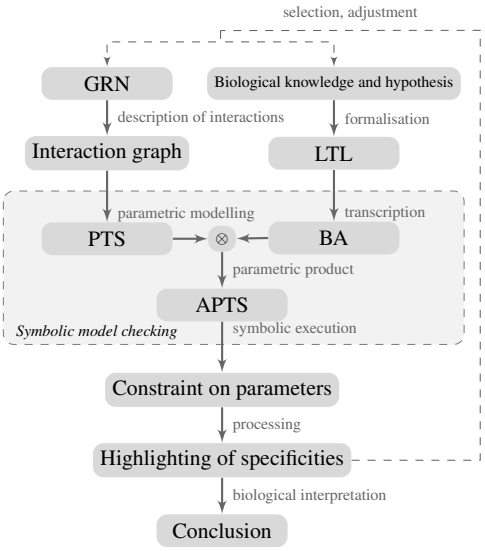


Figure 3: Resolution steps

been done with the SMBioNet [1] software tool). We begin by considering a minimal interaction graph of the system [4] consisting of only three elements and comparing the results computed with SPUTNIK with those obtained with SMBioNet [4]. Then we study a more comprehensive version of this GRN (from which the minimal graph is derived).

Such a GRN is involved in the secretion of toxins by the *Pseudomonas Aeruginosa* bacterium, via the T3SS. When the T3SS is activated, it causes the injection of toxic effectors directly in the target cell and thus the death or dysfunction of the cell. The role of T3SS is harmful to vulnerable people, especially those with cystic fibrosis. *In vivo*, this activation occurs in contact with the cell; *in vitro*, one of the factors might be the calcium depletion of the growth medium. In some cases, it seems that the T3SS is not activated despite the signals *in vivo* or *in vitro* and although no gene mutation of the system has been identified; this state is called *non inducible*. An epigenetic switch from *inducible* to *non inducible* is suspected to explain this phenomenon.

3.1 Minimal system

The minimal interaction graph (considered as being under calcium depletion) is presented in Figure 4. In fact we consider two variants of such minimal interaction graph: the one in Figure 4 and the other one obtained by inversion of the thresholds, with thresholds of 2 for $ExsA \rightarrow ExsD$ and of 1 for $ExsA \rightarrow ExsA$.

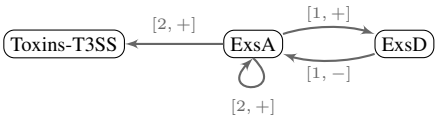


Figure 4: Interaction graph of minimal GRN

For such (minimal) interaction graph we obtain 324 different instantiations of parameters (648 in total, considering the two versions of interaction graph), *i.e.* 324 models complying with the dynamics of the GRN are possible (the complex *Toxins-T3SS* is not taken into account since it is not a regulator).

We have two biological properties about the system. In the case of an inducible strain, cytotoxicity of the bacterium (where expression level of *Toxins-T3SS* is maximum) is activated recurrently. In the case of a non inducible strain, expression level of *ExsA* is assumed too low to increase the expression level of *T3SS*: a non inducible strain remains non inducible. Such knowledge can be encoded in the following LTL formula:

$$\phi_1 : \mathbf{G}\big(x_{ExsA}{=}2 \rightarrow \mathbf{XF}(x_{ExsA}{=}2)\big) \wedge \mathbf{G}\big(x_{ExsA}{=}0 \rightarrow \neg\mathbf{F}(x_{ExsA}{=}2)\big)$$

To represent inducible state, we use a formula over *ExsA* ($x_{ExsA} = 2$) rather than over *Toxins-T3SS* ($x_{Toxins-T3SS} = 1$) but it is equivalent since we have necessarily $K_{Toxins-T3SS}(\{\}) = 0 \wedge K_{Toxins-T3SS}(\{ExsA\}) = 1$: thus $x_{ExsA} = 2$ causes activation of *T3SS* and secretion of toxins.

After computation, we obtain two models, one for each variant of the interaction graph in Figure 4, like in [4]. The state transition graphs obtained are represented in Figure 5.

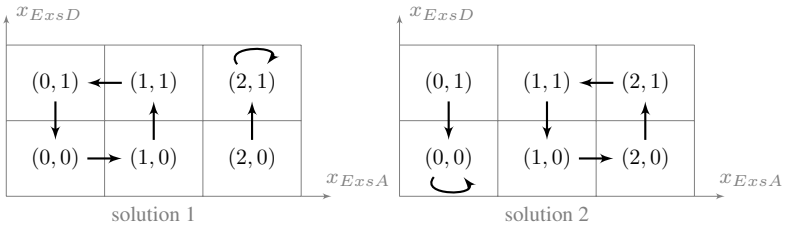


Figure 5: Models in the form of state transition graph

The two models are in accordance with the epigenetic hypothesis. Increase (resp. decrease) of *ExsA* modifies the phenotype of the bacterium, from a stable non-cytotoxic (resp. cytotoxic) state to a stable cytotoxic (resp. non-cytotoxic) state. The concentration in *ExsD* is not involved in this phenomenon.

Filopon *et al.* have proceeded to experimentations to test this behaviour *in vivo* and *in vitro* and confirmed the possibility of an epigenetic switch of the T3SS of *Pseudomonas Aeruginosa*.

3.2 Study of an extended GRN

The GRN presented above is a simplified model of a more complex network. From indications given in [4] and some discussions with one of its authors, we now study the interaction graph represented in Figure 6.

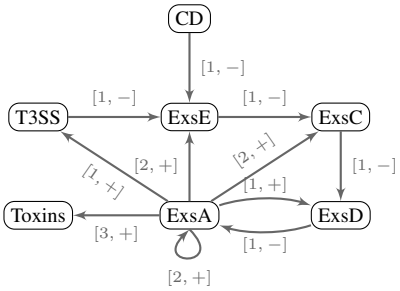


Figure 6: Interaction graph of extended GRN

In Figure 6, the genes constituting the T3SS are regulated by the ExsA TF. ExsA is also the regulator of three other genes (ExsC, ExsD and ExsE) and controls the genes coding the toxins. *CD* indicates if the growth medium is on *Calcium Depletion* or not. 4 million different models are possible.

To the biological knowledge that we have in the small GRN (see ϕ_1), we add two other informations to represent the role of calcium depletion in activation of the T3SS. We obtain the following formula:

$$\begin{aligned}\phi_2 \;:\; & \mathbf{G}\big(x_{ExsA}{\geq}2 \rightarrow \mathbf{XF}(x_{ExsA}{\geq}2)\big) \wedge \mathbf{G}\big(x_{ExsA}{=}0 \rightarrow \neg\mathbf{F}(x_{ExsA}{\geq}2)\big) \\ & \wedge \mathbf{G}\big((x_{ExsA}{\geq}2 \wedge CD{=}1) \rightarrow \mathbf{XF}(x_{ExsA}{=}3)\big) \\ & \wedge \mathbf{G}\big((x_{ExsA}{\geq}2 \wedge CD{=}0) \rightarrow \mathbf{XF}(x_{ExsA}{=}2)\big)\end{aligned}$$

We have also informations on parameters. The action of calcium and T3SS is joined, hence the following equalities: $K_{ExsE}(\{\}) = K_{ExsE}(\{CD\}) = K_{ExsE}(\{t3ss\})$ and $K_{ExsE}(\{ExsA\}) = K_{ExsE}(\{ExsA, t3ss\}) = K_{ExsE}(\{CD, ExsA\})$. If *ExsD* inhibits *ExsA*, then there is no toxins production, so its expression level is strictly below the production threshold of toxins, consequently we have: $K_{ExsA}(\{ExsA, ExsD\}) \leq 2$. Inhibition of *ExsC* on *ExsD* is stronger than activation of *ExsA*: *ExsC* makes a complex with *ExsD* as soon there are *ExsC*

and *ExsD*: thus $K_{ExsD}(\{ExsA, ExsC\}) = 0$. Similarly, inhibition of *ExsE* on *ExsC* is stronger than activation of *ExsA*: $K_{ExsC}(\{ExsA, ExsE\}) = 0$.

Taking into account all these biological hypotheses, we obtain two models with SPUTNIK. Thus, this extended GRN is consistent with results of [4]: both with the selected properties on the small GRN and with the experiments on epigenetic switch of T3SS.

The interaction graph represented in Figure 6 is not the only possibility to represent the GRN. We test also another interaction graph that combines *Toxins* and *T3SS* in one element *Toxins-T3SS* (in the same manner than the small example). Several thresholds are possible for the following interactions: $ExsA \to ExsE$, $ExsA \to ExsD$, $ExsA \to ExsA$ and $ExsA \to Toxins{-}T3SS$. By hypothesis, $ExsA \to ExsE$ and $ExsA \to ExsC$ have the same threshold and we assume that the threshold of $ExsA \to Toxins{-}T3SS$ is either the greatest level or the greatest level plus one (it changes the upper limit of $K_{ExsA}(ExsA, ExsD)$). With our hypotheses, it remains 26 different possibilities of interaction graphs. Figure 7 represents one of this possibilities (with respective thresholds: 2, 3, 1 and 3).

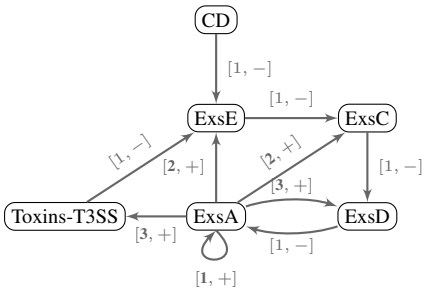


Figure 7: Another possibility of interaction graph of extended GRN

We tested all these possibilities, only 3 of them satisfy the expected properties. The combinations of values of unknown thresholds defining these 3 possibilities are given in Table 1.

Interactions	Threshold values		
$ExsA \to ExsE$	1	2	2
$ExsA \to ExsD$	1	3	3
$ExsA \to ExsA$	1	1	1
$ExsA \to Toxins{-}T3SS$	1	3	4

Table 1: Thresholds admitting solutions

Systems proposed in this section allow to obtain solutions corroborating the hypothesis of an epigenetic switch of T3SS. Properties highlighted in the minimal graph have been found in some larger interaction graphs. However, all 26 investigated configurations do not guarantee the existence of models satisfying the targeted properties: only some values of thresholds are possible. To identify the actual interaction graph, we need an additional feedback from biologists, either to confirm some choices of thresholds or to provide more informations on the expected behaviours of the GRN. However we have demonstrated that SPUTNIK provides flexibility to explore different possibilities.

4 Ongoing works

Beyond the first promising results, we want to look for optimisations to efficiently process larger GRNs. Some tracks are to investigate the use of other constraint solvers, of further simplifications of formulae during processing or of parallelisation mechanisms as Klarner *et al.* [5].

Acknowledgement

We thank Mrs. Janine Guespin for her help for the design of the extended model (Figure 6), especially for the discussions on possible threshold values.

References

[1] G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks: Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3):339-347, 2004.

[2] F. Corblin, E. Fanchon, L. Trilling, C. Chaouiya, and D. Thieffry. Automatic inference of regulatory and dynamical properties from incomplete gene interaction and expression data. In M.A. Lones et al., editor, *IPCAT*, volume 7223 of *LNCS*, pages 25-30. Springer Berlin Heidelberg, 2012.

[3] F. Corblin, S. Tripodi, E. Fanchon, D. Ropers, and L. Trilling. A declarative constraint-based method for analyzing discrete genetic regulatory networks. *Biosystems*, 98(2):91 - 104, 2009.

[4] D. Filopon, A. Merieau, G. Bernot, J.-P. Comet, R. Leberre, B. Guery, B. Polack, and J. Guespin-Michel. Epigenetic acquisition of inducibility of type III cytotoxicity in p. aeruginosa. *BMC bioinformatics*, 7:272, 2006.

[5] H. Klarner, A. Streck, D. Safranek, J. Kolcak, and H. Siebert. Parameter identification and model ranking of thomas networks. In David Gilbert and Monika Heiner, editors, *Computational Methods in Systems Biology*, LNCS, pages 207-226. Springer Berlin Heidelberg, 2012.

[6] F. Laburthe, and N. Jussien. *Choco solver Documentation*, 2012.

[7] M. Manceny, J.-P. Comet, J.-P. Gallois, and P. Le Gall. Model-checking for parametric transition systems based on symbolic execution - extended version. *Technical report of ECP*, 2008.

[8] D. Mateus, J.-P. Gallois, J.-P. Comet, and P. Le Gall. Symbolic modeling of genetic regulatory networks. *Journal of Bioinformatics and Computational Biology*, 5(2B):627-640, 2007.

[9] A. Richard. *SMBioNet User manual*, 2010.

[10] R. Thomas. Logical analysis of systems comprising feedback loops. *J.Theor. Biol.*, 73(4):631-56, 1978.

[11] R. Thomas and R. d'Ari. *Biological Feedback*. CRC Press, 1990.

Identification of psoriasis disease from transcriptomics data

Camelia Moldovan[1], George Viorel Popescu[2,3]

[1] Master ITEMS, University "Politehnica" Bucharest, Bucharest 060042,Romania
[2] National Institute for Laser, Plasma & Radiation Physics, Str. Atomistilor, Nr. 409, Magurele 077125, Bucharest, Romania
[3] University "Politehnica" Bucharest,313 Splaiul Independentei, Bucharest 060042, Romania

Abstract

Recent advances in genomics have led to new methods for identification of diseases from their "signature" at transcriptional level. The goal of this article is to design a computational method to classify clinical samples based on transcriptomics data analysis. The possible applications range from improved diagnosis and identification of disease-associated pathways, to prediction of response to new medicines [1] and prediction of development of a disease. The method designed was applied to identify normal vs psoriatic skin based only on the transcriptome of a skin biopsy from the GEO Dataset GSE13355 [2]. Using machine learning algorithms there were identified 14 probe sets out of 54,000 that might have an influence in psoriasis prediction.

Our results can be applied to improve diagnosis of psoriasis using gene marker analysis and to identify disease-associated pathways.

1 Introduction

Psoriasis is the most prevalent autoimmune disease in the U.S; according to current studies, as many as 7.5 million Americans —approximately 2.2 percent of the population —have psoriasis. This is a chronic inflammatory and hyperproliferative skin disease, which, in addition to cutaneous manifestation, is accompanied with inflammatory arthritis in up to 40% cases.

The economic burden of psoriasis is estimated to be approximately US$ 11.2 billion in the US, and CHF 314-458 million in Switzerland [3].

The disease is diagnosed following physical examination of the skin lesions. Microscopic analysis of psoriatic skin biopsy shows thick, red, flaky cells with no sign of inflammation ; blood tests can be used to differentiate psoriatic from rheumatoid arthritis.

Psoriasis is typically treated with topical application of both steroids and non-steroids and phototherapy: UVB and UVA with light-sensitizing medication. New drugs that target the autoimmune response and specific parts of the immune system (TNF, interleukin) are becoming available.

The data used in this study were obtained from [5] as part of Diagnostic Signature Challenge. The challenge was to verify that a robust diagnostic signature for Psoriasis can be extracted from gene expression data. Participants were asked to develop and then submit a classifier that could stratify skin samples into one of two phenotypic groups —Psoriasis or Control [5].

2 Materials and Methods

The sample dataset —GSE13355 —was obtained from "Gene expression data of skin from psoriatic patients and normal controls" [2].

Total RNA was extracted from punch biopsies taken from 58 psoriatic patients and 64 normal healthy controls. Punch biopsy is the primary technique to obtain diagnostic skin specimens. It is performed using a circular blade attached to a pencil-like handle rotated down through the epidermis and dermis into the subcutaneous fat [4].

Two biopsies were taken from each patient; one 6mm punch biopsy was obtained from lesional skin of each patient (involved sample) and the other from non-lesional skin (uninvolved sample), taken at least 10 cm away from any active plaque.

One biopsy was obtained from each healthy control. Totally 180 samples were run on Affymetrix HU133 Plus 2.0 microarrays containing >54,000 gene probes [2].
Affymetrix Human Genome U133A 2.0 Array annotation data (chip hgu133a2) is assembled using data from public repositories [15].

In the training data, control includes both healthy individuals and uninvolved skin of psoriasis patients. [2] All samples were acquired from living patients. Skin samples from control patients were mostly acquired from reduction surgeries due to obesity [5]. In total, 180 samples were run on Affymetrix HU133 Plus 2.0 microarrays (chip hgu133a2) [15] containing >54,000 gene probes [2].

The analysis reported here used the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array, which contains more than 54,000 probe sets [17]. A probe set is a collection of probes designed to interrogate a given sequence. Each gene transcript is represented on hgu133a2 by a probe set [18]. By design it is possible that multiple probe sets are mapping to the same gene [18]. Gene expression was measured in this study by extracting mRNA from the cells, obtaining the corresponding cDNA samples, and hybridising the cDNA to the probes on the microarray.

2.1 Statistical Analysis of microarray data

Both training and test data used in this article were obtained from [5] as part of Diagnostic Signature Challenge. Test dataset was licensed from GeneLogic (**http://www.genelogic.com**) [5].

The training dataset contains 180 samples —122 healthy skin samples, 58 lesional psoriatic skin. Each sample had an associated .cel file available for download from [2].
The test data consists of 62 samples that need to be classify as healthy/psoriatic. Additional information included age, gender, treatment, and comorbidities were available. The .CEL file is a binary file that store the results of the intensity calculations on the pixel values like intensity value, standard deviation of the intensity, the number of pixels used to calculate the intensity value [19]. The files were read using the Affymetrix [12] package of Bioconductor [11] and the result was saved as a 54675x181 matrix, the first column being the probeset id. A similar 54675x63 matrix was created for test data.

2.1.1 Data normalization

Data normalization was done in R [10] via Robust Multi-array Average (RMA) [6] on all training and test samples. The data processing steps applied were background correction, quantile normalization and summarization. The Affymetrix microarray data was first arranged in an expression matrix. Then the raw intensity values were background corrected, log2 transformed and quantile normalized [6].

2.1.2 Feature selection

Using the entire probe set for classification of samples could lead to incorrect predictions: the models are difficult to train and the estimated parameters will not return accurate results on test sets. One solution is to select a subset of features that are most "relevant" for gene expression analysis.

Here we selected a set of differentially expressed genes using linear regression analysis (LIMMA package [13]) to summarize the transcriptional profile of each sample and employed a machine learning algorithm for classification of samples. The first step in selecting the features necessary for sample classification was to fit a linear model (lmFit function in LIMMA) for each probe set. In its simplest form lmFit [13] uses as parameters an object containing the log-values of expression for a series of microarrays and the design matrix [20] of the microarray experiment. We created a design matrix in which each row corresponds to one of the 180 arrays (samples) in the experiment and each column corresponds to a class (psoriatic/control). Column values were defined as either one or zero, samples of healthy skin from psoriatic patients were considered healthy (0). The matrix defines a model of the relationship between the 180 samples as explanatory variables and the classes (psoriatic/control) as dependent variable. For example X(4, 2) =1 signifies that the 4th sample belongs to a healthy person.

Using regression analysis implemented in LIMMA package, we produced a table of the top ranking genes, sorted by default by their log-odds for dif-

ferential expression [13]. The top ranking probe sets were selected for further classification of the samples.

2.2 Sample Classification

We used as for sample classification Support Vector Machines [7], a supervised method introduced in the 1990s. SVMs are defined as hyperplanes that separate the training data by a maximal margin. The training instances closest to the hyperplane are called support vectors [7]. Compared to neural networks support vector machine are faster, can be used with a larger number of genes and are deterministic [8].

From the top ranking probeset list we found that 14 probe sets give the best classification. Tests were performed using up to 500 probe sets but the results were not satisfactory: most of the time all samples were classified either as psoriatic or healthy. Sample classification was performed using the Support Vector Machine(SVM) implementation in Matlab. We used svmclassify function to classifies each row of the data in the matrix corresponding to the test dataset, using the information in a support vector machine classifier structure SVMStruct, created with the svmtrain function on the training matrix.

3 Results and Discussion

The 14 differentially expressed probe sets retained for classification were converted to gene id using Bioconductor's annotation packages [15]. Some probesets from the short list of 14 were mapped to the same gene id. The most frequent genes in the list are: PI3 peptidase inhibitor 3, skin-derived, DEFB4A defensin, beta 4A and KYNU kynureninase. DEFB4A is expressed in the skin and respiratory tract and is induced by inflammation [21]. Since psoriasis is an autoimmune disease, it appears that DEFB4A selection for training set was appropriate.

The output of the svmtrain/svmclassify functions is an array of 62 elements, each element corresponding to a sample. There were 11 samples classified as psoriatic(value 1) the others were classified as healthy(value 0). According to the classification of test data, published later by the organizators, there were 35 control and 27 psoriasis samples [5]. Out of the 11 samples that we classified as psoriatic there were no false positives; however there was an important percentage of psoriatic samples left out (false negatives).

4 Conclusions

The performance of our psoriasis classification study illustrate the learning methods dilemma of generalization vs. specialization. Classifier performance depends on the optimal selection of features, on the size of the training set and on its capacity to capture discriminatory features of the input space. We

experimented here with a limited set of combinations of gene sets for sample classification. Further investigations of training and test datasets could explain the causes of misclassification. We intend to study optimal features selection for classification as well as other statistical learning methods to increase the robustness of classification for small training datasets.

Based on the gene list we have produced, we can perform a pathway and gene ontology analysis using information from bioinformatics databases. Clustering genes that share related pathways could help identify new functions for genes in the same cluster. Reverse engineering techniques (i.e bayesian networks) [16] can be used to infer gene regulatory networks based on gene expression levels measured by affymetrix microarrays. These gene regulatory networks could help identify key regulators for a list of genes.

References

[1] Suarez-Farinas et al.(2010), Personalized medicine in psoriasis: developing a genomic classifier to predict histological response to Alefacept. *BMC Dermatology* **10**:1

[2] Gudjonsson JE, Ding J, Nair R, Stuart P, Voorhees JJ, Elder JT, Abecasis G (2009) Gene expression data of skin from psoriatic patients and normal controls. *Gene Expression Omnibus (GEO)*, **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13355**

[3] National Psoriasis Foundation. *http://www.psoriasis.org*

[4] Zuber T.(2002 Mar) Punch Biopsy of the Skin Am Fam Physician. 15;65(6):1155-1158

[5] Improver Systems Biology Verification Challenge(2012). **http://www.sbvimprover.com/**

[6] Bolstad, B.M., Irizarry R. A., Astrand, M., Speed, T.P.(2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* **19(2)**:185-193

[7] Vapnik V,(1998), Statistical Learning Theory. *Wiley, New York, NY*

[8] S. N. Sivanandam, S. N Deepa(2006) Introduction to Neural Networks Using Matlab 6.0. *Tata McGraw-Hill Education*

[9] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar(2012), Foundations of Machine Learning. *The MIT Press, Cambridge, Massachusetts*, 63-75

[10] R Development Core TeamR: A language and environment for statistical computing. R Foundation for Statistical Computing, *http://www.R-project.org*

[11] Gentleman R., Carey V. Bates D.,(2004) Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* **5**:80

[12] Gautier L, Cope L. , Bolstad M, Irizarry R. (2004), affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. *Oxford University Press* **20**: 307–315.

[13] Smyth G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, *Springer, New York*, 39-420.

[14] Darius M Diuzda(2010) Data mining for genomics and proteomics, *Wiley, New York*,123

[15] Carlson M. hgu133a2.db: Affymetrix Human Genome U133A 2.0 Array annotation data (chip hgu133a2)

[16] Stumpf M, Balding D., Girolami M(2011). Handbook of statistical systems biology. Wiley, 39-66,153-154

[17] Affymetrix GeneChip Human Genome U133 Arrays, **http://www.affymetrix.com/support/technical/datasheets.affx**

[18] Affymetrix, Inc GeneChip Custom Express array design guide. Part No. 700506 Rev. 4. 2003. **http://www.affymetrix.com/support/technical/other/custom_design_manual.pdf**

[19] Affymetrix CEL data file format, **http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html**

[20] Smyth G. K., Ritchie M, Thorne N, Wettenhall J. (2012). Limma: linear models for microarray data user's guide. Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, 47-49.

[21] DFB4A_HUMAN, **http://www.uniprot.org/uniprot/O15263**

Minimal Cut Sets and Its Application to Study Metabolic Pathway Structures

Nguyen Vu Ngoc Tung[1,3], Beurton-Aimar Marie[1] and Colombié Sophie[2]

[1] Laboratoire Bordelais de Recherche en Informatique (LaBRI), Univ. Bordeaux, UMR 5800 UB1-IPB-UB2-CNRS-INRIA, 351 cours de la Libération. F-33405 Talence Cedex, France

[2] INRA Bordeaux Aquitaine, Univ. Bordeaux, UMR-A 1332 INRA-UB1-UB2-UB4-CNRS, Fruit Biology and Pathology BP 81, 71 Avenue Edouard Bourlaux. F-33140 Villenave d'Ornon, France

[3] Faculty of Science and Technology, Hoa Sen University, 93 Cao Thang St., Ward 3, Dist. 3, Ho Chi Minh City, Viet Nam

Abstract

From the last decade, researches on Systems Biology have been mainly interested in analysis of large networks. To analyze the function and dysfunction components of such systems by a simple visual inspection is infeasible due to their complex organization. As biological networks can be modeled as graph, algorithms coming from graph theory can be reused to write procedures to analyze them. In order to do it, we have concentrated on the analysis of metabolic networks and the way to discover structural properties of metabolic networks. In this context, Elementary Flux Modes (EFMs) analysis is one of the robust tools to give us a deeper insight into such metabolic networks. Using EFMs analysis, one can identify all the feasible routes through a given network. However, computing EFMs gives a large number of solutions of which analysis in detail is hard generally. Recently, a dual approach to EFMs has been developed with calculation of Minimal Cut Sets (MCSs), that is identification of the reaction sets which dismiss a given objective reaction. Giving a metabolic network, the results of computing MCSs are expected to be smaller than those of computing EFMs. In order to verify this hypothesis, we have tested this assumption with 4 networks each having different sizes and structures. We will see in the last section the hypothesis could be verified but only when EFMs computing provides a very huge result.

Introduction

Metabolic network structure is often described as one of the most complex cellular systems [37]. It represents biochemical reactions catalyzed by enzymes which connect one or several substrates to one or several products. In this context, needs of specific modeling and analysis methods can become a

bottleneck to work in System Biology. A lot of studies have concentrated on the characterization of metabolic networks by means of graph theory to gain an insight into the global network structure. In that way, finding topological properties can help to analyze biological structures [1, 5, 40]. Depending on the goals of the analysis, these works appear in different research themes: graph-theoretical analysis [17, 5, 16], flux balance analysis [19, 28], metabolic pathway analysis [32, 24, 26].

The purpose of our work is to study the way to compute MCSs and to evaluate the obtained results with the 4 concrete examples. First we introduce the concepts of cut set, definitions and notations. Then we present some classical algorithms for computing all cut sets. Next, we examined the specific algorithms to compute all the minimal cut sets in metabolic networks. Finally, the result of computing MCSs is compared to the one obtained with EFMs computation for the 4 different networks.

1 Graph cut sets

H. Whitney [38] is one of the precursors that used the concept of cut sets with planar graphs in the early 1930s. Then in 1956, Ford and Fulkerson [12] introduced the basic concepts of flow, cut, etc. and first stated the maximum-flow minimum-cut theorem. This theorem is considered as the fundamental contribution for developing theory of network flows. The main idea of MCSs theory can be resumed as follows: if there are several edges of a network to be failed with a certain probability and only these edges can disconnect the network, the set of the edges is called a *minimal cut set*.

Now we present some definitions and notations about *cut*, *cut set*, *s-t cut set*, and *minimal cut set* [15, 35]. Let $G = (V, E)$ be an undirected graph. Let $n = |V|, m = |E|$. For $S \subset V$, the set $\delta(S) = \{(u, v) \in E : u \in S, v \in \overline{S}\}$ is a *cut set* since their removal from G disconnects G into more than one subgraphs. The size of a cut is the number of the edges in $\delta(S)$. Formally, we define that a *cut* C of an undirected graph G is a partition of the vertices $V(G)$ into two separate non-empty subsets, that is, $C = \{S, \overline{S}\}$ where $S \cup \overline{S} = V(G)$ and $S \cap \overline{S} = \emptyset$. Consequently, the set of the edges (e.g., $\delta(S)$) that crossed between the two subsets is called a *cut set*. The number of the elements of $\delta(S)$ is the size of the cut set. A *minimum cut set* is a cut set of a certain minimum size. As an illustrative example, consider the undirected graph in Figure 1. In this example, $(\{1, 2, 4, 5\}, \{3, 6, 7\})$ is a minimum cut (the bold line and the minimum cut set corresponding with the minimum cut is $\{(2, 3); (5, 6)\}$ which the weight is 9.

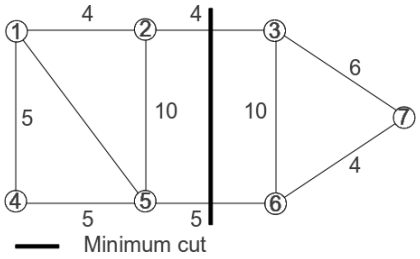
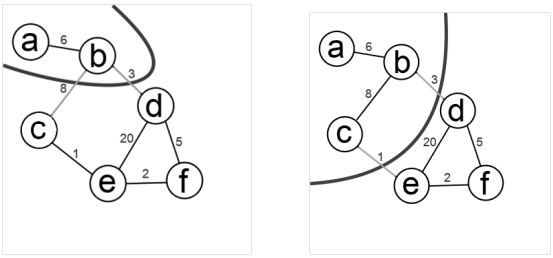


Figure 1: A minimum cut of an undirected graph G [39]

s-t cut set The *s-t cut* [4] is a *cut* with *s* and *t* in different partitions. In other words, a cut $s-t$ of an undirected graph G is simply a cut $C = \{S, \overline{S}\}$ with $s \in S$ and $t \in \overline{S}$. So, a cut set of the cut is the edge set which end points are in the separate subsets of the vertices. The removal (or "cut") of the edges disconnects the graph into two separate subgraphs. Let's consider the example



(a) Weight of this cut: 11 (b) Weight of this cut: 4

Figure 2: Examples of cuts of a graph

in Figure 2 depicts the *s-t cut* that one has the weight 11 and the other 4. Suppose that the *s-t* cut set of *s* = *a* and *t* = *d*. Then we can enumerate several *a-d* cut sets such as $\{bc, bd\}$ with the weight 11, $\{ce, bd\}$ with the weight 4, or $\{bd, de, ef\}$ with the weight 25. In this example, because of the less number of cut sets, the enumeration can be done manually. The minimum cut set is the one with the minimum weight (e.g, it is 4 as in Figure 2(b)).

Minimum cut set in directed graph Similarly, we define minimum cut set of a directed weighted graph. We denote $D = (N, A)$ a directed graph. Let $n = |N|, m = |A|$. A minimum cut set is a set of all the edges crossed through two subsets S and T. However, one should pay attention to how to compute minimum cut set value. Instead of summing all the weights of all the edges in the minimum cut set, the only crossed edges between the two subsets coming out S are taken into account.

Minimum cut sets have also been arisen in information retrieval [8], compilers for parallel languages [9], and in communication networks [20, 29] in routing of ATM networks. The theory of network reliability and survivability employed MCSs as the way to evaluate the stableness of complex systems can be found in [7, 30, 41]. In the same way, algorithms have been applied into computational biology and metabolic networks as well. The question to enumerate all MCSs shall be addressed in the next sections.

2 Finding minimal cut sets

The first approach for finding a MCS of a graph is based on the minimum $s-t$ cut problem as it has been mentioned previously. As given by the well known max flow [11, 12] theorem, a minimum $s-t$ cut can be found by computing the maximum flow between s and t. In 1961, Gomory and Hu (GH) [14] introduced a typical tree structure that can be able to find minimum $s-t$ cuts for all $\binom{2}{n}$ pairs of s and t in an undirected and weighted graph. They showed that the number of distinct cuts in the graph is at most $n-1$ (rather than the naïve $\binom{2}{n}$). Furthermore, there is an efficient tree structure that can be maintained to compute this set of distinct cuts using only $n-1$ maximum flow computations. A natural question to arise from GH algorithm is whether some of the information computed in one maximum flow computation can be reused in the next one or not. Hao and Orlin (HO) [15] answered this question in the affirmative. The key new idea is to use a push-relabel maximum flow algorithm to implement GH, and use the preflow and distance labeling from the last max-flow computation as a starting point for the current one. HO consider the problem of finding the minimum capacity cut in a directed network G with n nodes. One can use a maximum flow problem to find a minimum cut separating a designated source node s from a designated sink node t, and by varying the sink node one can find a minimum cut in G as a sequence of at most $2n-2$ maximum flow problems. They then showed how to reduce the running time of these $2n-2$ maximum flow algorithms to the running time for solving a single maximum flow problem. The resulting running time is $O(mnlog(n^2/m))$ for finding the minimum cut in either a directed or an undirected network. The implementation of HO algorithm can be found in LEMON library[1].

Nagamochi and Ibaraki [25] published the first deterministic minimum cut algorithm not based on the flow algorithm. This algorithm has the slightly better running time of $O(|V||E|+|V|^2log|V|)$, but it remains complicated to implement. Stoer and Wagner (SW) [35] gave a simplified version of the Nagamochi and Ibaraki algorithm with the same running time. This simplification

[1] LEMON library can be found at http://lemon.cs.elte.hu/pub/tutorial/index.html

was subsequently discovered independently by Frank [13]. SW proposed the following method for finding a minimum cut set of a graph G. The algorithm is based on the following statements: let s and t be two vertices of graph $G = (V, E)$. Let $G/s, t$ be the graph obtained by contracting s and t. Then, a minimum cut of G can be obtained by taking the smaller of minimum $s - t$ cut and minimum cut of $G/s, t$. The implementation of SW algorithm can be found in JGraphT library[2].

There exists a lot of algorithms to compute all minimal cut sets (MCSs) in graphs. In the 1970s, the theoretical algorithms [2, 3] had been proposed and proved formally. But the exponential growth of CPU time to compute MCSs makes difficult the implementation of these algorithms. Recently, several authors have proposed efficient algorithms to enumerate all MCSs of a graph. By using the dual maximum flow problem, Curet et al. [10] constructed a binary relation associated with an optimal maximum flow such that all minimum cost `s-t` are identified through the set of closures for this relation.

In summary, the algorithms to find all MCSs, which have just mentioned above, can be applied for systems modeled as classical graphs. But in metabolic networks, many biological processes have more than two participating elements and resulting graphs exhibit a high level of complexity [23]. Next section shows the applications of MCSs in the context of metabolic networks.

3 Minimal Cut Sets in context of Metabolic Networks

During the last decade, several methods and algorithms have been developed to analyze large metabolic networks. Among the available tools, we can mention the Extreme Pathways analysis defined by Schilling et al. [31] and the Elementary Flux Modes analysis defined by Schuster et al. [33]. Even they are not exactly similar, both use linear algebra algorithms to find the solutions. Unfortunately, most often when the metabolic network is a little bit large (more than 20 reactions), the number of Extreme Pathways or EFMs tend to be so high that their analysis is very hard to do by hand. So, the problem to study structure of big networks still remains. One alternative to study the structure of big networks could be the MCSs. Indeed, MCSs appears recently as an additional work to the EFMs giving a dual view of EFMs in the context of metabolic networks. Klamt et al. [22, 21] proposed an algorithm to compute all set of reactions which can disconnect the studied network in order to prevent a specific production of a metabolite. Basically, the concept is pretty similar than

[2] JGraphT library can be found at http://sourceforge.net/projects/jgrapht/

the one explains previously in graph theory but Klamt et al. [23] have added the same constraint than for EFMs computation, the steady state assumption. They explain that this new method can provide a number of solutions smaller than those of EFMs. In order to verify this hypothesis, we have computed EFMs and MCSs of 4 different networks: 3 of them modeling energetic metabolism of mitochondria into 3 tissues: muscle, liver, yeast [27], the last one modeling the central metabolism of heterotrophic plant cells (PC). The 3 mitochondrial metabolic networks exhibit small differences and their number of reactions, env 40[3], is not really high. The PC network is bigger, it includes several biological pathways (glycolysis, Pentose Phosphate pathway, Starch and Sucrose synthesis and degradation), and then it consists of 70 metabolites and 78 reactions[4]. As softwares to compute MCSs derived from EFMs computing softwares, the next paragraph presents them together.

CellNetAnalyzer CellNetAnalyzer (CNA)[5] comes from the previous software Metatool[6] written by the Jena Bioinformatics group. It is a package for MATLAB containing several modules to visualize networks and to analyze their structures. CNA enables users to compute both EFMs and MCSs. Thus we have used it to calculate EFMs and MCSs for 4 networks. That computation has often been time consuming, in some cases several hours or days are necessary. For example, to obtain MCSs of PC with CNA more than 10 days have been needed with a linux server. Second in the case of PC network memory requirements are larger than method can manage.

Efmtool and regEfmtool A couple of years ago, a new implementation of EFMs computation has been done with improvements of the original algorithm. This is efmtool[7] [36] implemented in the Java programming language. It supports multi-threading and seems to be robust to compute large networks. But even this software is freely available, with the source code, this program is not easy for use and lacks of a detailed documentation. Within recent years, a new software, named regEfmtool[8], written by C. Jungreuthmayer [18], provides a new way to use efmtool program. The package contains several scripts clearly documented. It also proposes to define some logical rules to compute EFMs containing or not some reactions and so reducing the size of the obtained solutions. The larger network that we have computed with this

[3] complete description files can be found at [27]
[4] described more detail in [6]
[5] http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html
[6] http://pinguin.biologie.uni-jena.de/bioinformatik/networks/
[7] http://www.csb.ethz.ch/tools/efmtool/
[8] http://www.biotec.boku.ac.at/regulatoryelementaryfluxmode.html

tool contains more than 80 reactions and we have obtained several millions of EFMs in only a couple of hours. Computing MCSs out of MATLAB and in C language will be available very soon from the same team. In preliminary tests, we have been able to obtain MCSs that have not ever been obtained before with MATLAB programs due to overload memory.

Results Hopefully, all those tools yield, of course, the same results. The table 1 shows the number of EFMs and MCSs obtained for the 4 networks. The column 1 gives the name of the networks and the column 2 the total number of reactions, the number of reversible reactions is given in the parentheses. The column 3 presents the number of the internal metabolites, we consider only internal ones because EFMs and MCSs are computed from the stoichiometric matrix (defined by the list of reactions and internal metabolites). The last two columns contain the final computing results.
Firstly, no obvious relationship can be observed between the number of reactions (or internal metabolites) and the number of EFMs. As it has been shown by Stelling et al. [34] more EFMs you have, more you can expect robust this network, because the size of the EFMs set is a measure of the network connectivity. Moreover, as we have previously mentioned, obtaining so huge results is not very useful for biologists. Secondly, the number of MCSs is unfortunately not at all lower than the number of EFMs, as it could expected for the 3 mitochondrial networks. When the result begins to be really huge as with the PC network, the number of MCSs begins to decrease. Finally, we can observe that the average length of EFMs increases with the number of reactions but the average length of MCSs remains stable. For example, for the muscle, the average length of EFMs and MCSs are 17.7 and 10.2 resp., comparing to the values obtained for the PC network, 37.7 and 11.1 resp. To conclude, a part of the goal has been reached with MCSs computation which provides a smaller set than EFMs only for large enough metabolic networks.

Tissues	Nb. React	Nb. Int. Meta	Nb. EFMs	Nb. MCSs
Muscle	37 (25)	31	3,253 (17.7)	42,534 (10.2)
Liver	44 (28)	36	2,307 (16.7)	47,203 (11.4)
Yeast	40 (29)	34	4,627 (15.3)	90,318 (11.6)
PC	78 (33)	55	114,614 (37.7)	93,009 (11.1)

Table 1: Global description of the 4 networks

Conclusion

The aim of this study was to investigate the computation of EFMs and MCSs on given metabolic networks. We have shown that the number of MCSs is higher than the number of EFMs on mitochondrial networks, when the network is not so big and the number of EFMs is not huge. But with bigger networks like the network of heterotrophic plant cells, the number of MCSs is lower than EFMs one and the length of MCSs does not increase with the number of reactions. So, it could be easier to analyze MCSs than EFMs, for instance, to classify them. A problem we are facing is the computation of MCSs algorithm, in time and memory size requirements. New research tracks are explored to introduce improvements coming from graph theory and parallelized techniques also.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, June 2001.

[2] H. Ariyoshi. Cut-set graph and systematic generation of separating sets. *IEEE Transactions on Circuit Theory*, 19(3):233 – 240, may 1972.

[3] S. Arunkumar and S.H. Lee. Enumeration of all minimal cut-sets for a node pair in a graph. *IEEE Transactions On Reliability*, R-28(1):51–55, April 1979.

[4] Ahmet Balcioglu. An algorithm for unumerating the near-minimum weigt s-t cuts of a graph. Master's thesis, NAVAL POSTGRADUATE SCHOOL, Monterey, California, December 2000.

[5] Albert-László Barabási and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, February 2004.

[6] Marie Beurton-Aimar, Bertrand Beauvoit, Antoine Monier, FranÃ§ois Vallée, Martine Dieuaide-Noubhani, and Sophie Colombié. Comparison between elementary flux modes analysis and 13c-metabolic fluxes measured in bacterial and plant cells. *BMC Systems Biology*, 5(95), 2011.

[7] Roy Billinton and Ronald Norman Allan. *Reliability Evaluation of Engineering Systems: Concepts and Techniques*, chapter 11, pages 347–350. Kluwer Academic/Plenum Publishers, New York, 2 edition, June 1992.

[8] R. A. Botafogo. Cluster analysis for hypertext systems. In *Proc. of the 16th Annual ACM SIGIR Conference of Res. and Dev. in Info. Retrieval*, pages 116–125, 1993.

[9] Siddhartha Chatterjee, John R. Gilbert, Robert Schreiber, and Thomas J. Shefl:ler. Array distribution in data-parallel programs. In *In Proceedings of the Seventh Workshop on Languages and Compilers for Parallel Computing*, pages 76–91. Springer-Verlag, 1994.

[10] Norman D. Curet, Jason DeVinney, and Matthew E. Gaston. An efficient network flow code for finding all minimum cost s-t cutsets. *Computers & Operations Research*, 29(3):205–219, 2002.

[11] P. Elias, A. Feinstein, and C. Shannon. A note on the maximum flow through a network. *IRE Transactions on Information Theory*, 2(4):117–119, December 1956.

[12] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[13] Andras Frank. On the edge-connectivity algorithm of nagamochi and ibaraki. Laboratoire Artemis, IMAG, Université J. Fourier, Grenoble, Switzerland, 1994.

[14] R. E. Gomory and T. C. Hu. Multi-terminal network flows. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):551–570, December 1961.

[15] Jianxiu Hao and James B. Orlin. A faster algorithm for finding the minimum cut in a graph. In *Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms*, SODA '92, pages 165–174, Philadelphia, PA, USA, 1992. Society for Industrial and Applied Mathematics.

[16] Ying-Jin Yuan Hong-Wu Ma, Xue-Ming Zhao and An-Ping Zeng. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, 20(12):1870–1876, March 2004.

[17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[18] C. Jungreuthmayer, D. E. Ruckerbauer, and J. Zanghellini. Utilizing gene regulatory information to speed up the calculation of elementary flux modes. *ArXiv e-prints*, August 2012.

[19] K.J. Kauffman, P. Prakash, and J.S. Edwards. Advances in flux balance analysis. *Curr Opin Biotechnol*, 14(5):491–496, 2003.

[20] W. H. Kim and R. T. Chien. *Topological Analysis and Synthesis of Communication Networks*. NewYork: Columbia University Press, 1962.

[21] Steffen Klamt. Generalized concept of minimal cut sets in biochemical networks. *Bio Systems*, 83(2-3):233–247, Apr 2005.

[22] Steffen Klamt and Ernst Dieter Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, 2004.

[23] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Comput Biol*, 5(5):e1000385+, May 2009.

[24] Steffen Klamt and Jörg Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2):64 – 69, 2003.

[25] Hiroshi Nagamochi and Toshihide Ibaraki. A linear-time algorithm for finding a sparse k-connected spanning subgraph of a k-connected graph. *Algorithmica*, 7:583–596, 1992. 10.1007/BF01758778.

[26] Jason A. Papin, Jöerg Stelling, Nathan D. Price, Steffen Klamt, Stefan Schuster, and Bernhard Ø. Palsson. Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8):400 – 405, 2004.

[27] S. Pérès, M. Beurton-Aimar, and J.P.M. Mazat. Analysis of large set of elementary modes: application to energetic mitochondrial metabolism. In *European Conference on Complex Systems*, 2005.

[28] Nathan D. Price, Jennifer L. Reed, and Bernhard Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11):886–897, November 2004.

[29] Suresh Rai. A cutset approach to reliability evaluation in communication networks. *IEEE Transactions on Reliability*, R-31(5):428–431, dec. 1982.

[30] V. Ramachandran, A.C. Raghuram, R.V. Krishnan, and S.K. Bhaumik. *Failure Analysis of Engineering Structures: Methodology and Case Histories*, chapter 5, pages 39–42. ASM International, 2005.

[31] C.H. Schilling, D. Letscher, and B.Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248, 2000.

[32] Stefan Schuster, David A. Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(3):326–332, 2000.

[33] Stefan Schuster and Claus Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(2):165–182, 1994.

[34] J. Stelling, U. Sauer, Z. Szallasi, F.J. Doyle 3rd, and J. Doyle. Robustness of cellular functions. *Cell*, 118(6):675–85, 2004.

[35] Mechthild Stoer and Frank Wagner. A simple min cut algorithm. *Journal of the ACM*, 44(4):585–591, July 1997.

[36] Marco Terzer and Jörg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235, 2008.

[37] Andreas Wagner and David A. Fell. The small world inside large metabolic networks. In *Proceedings of the Conference of The Royal Society in London B*, volume 268, pages 1803–1810, April 2001.

[38] Hassler Whitney. Planar graphs. *Fund. Math.*, 21:73–84, 1933.

[39] Li-Pu Yeh, Biing-Feng Wang, and Hsin-Hao Su. Efficient algorithms for the problems of enumerating cuts by non-decreasing weights. *Algorithmica*, 56(3):297–312, March 2010.

[40] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–24, 2007.

[41] Enrico Zio. *An introduction to the basics of reliability and risk analysis*, volume 13 of *Quality, Reliability and Engineering Statistics*, chapter 7, pages 128–132. World Scientific Publishing Co. Pte. Ltd., 2007.

Taking Systems Biology Into The Clinic: a Clinical Genetics Perspective

Priscilla H. Axwijk[1] and Rachel H. Giles[1]

[1] Dept. Nephrology and Hypertension, University Medical Center Utrecht, Heidelberglaan 100, 3584CX Utrecht, The Netherlands

Abstract

In clinical medicine, chronic diseases are often oligogenic and complex to diagnose and treat. Transcriptomes, proteomes and metabolomes are currently being catalogued for such diseases yet the systems biology frameworks being generated by these datasets are primarily being used to study variation and function of the human genome and relating them to health and disease states. Recently enhanced efficiency of DNA sequencing allows powerful analytical computational and mathematical tools aimed at understanding functional and regulatory networks underlying the behavior of complex biological systems. Iterative systems approaches for specific human diseases such inherited renal cancer syndromes or renal ciliopathies have started making inroads to improving diagnostic and prognostic parameters. Because of their relatively contained yet oligogenic properties, in many ways the inherited renal cancer syndromes and ciliopathies offer an exemplary system to describe how systems approaches are transforming the way drugs are being developed based on complex interactions between distinct but overlapping pathways. Consequently, a perspective in which the interactions and dynamics are centrally integrated may steer medical intervention towards interrelationships of components. The optimal method for predictive, personalized, and preventive treatment of complex chronic diseases may therefore lie in systems medicine. We illustrate our arguments with a case report.

1 Case Report

A family comprising a father, mother, two sons and one daughter visits the Department of Clinical Genetics at an Academic Hospital. The father, now age 65, was treated for renal cell carcinoma at the age of 45, and he has cysts in both kidneys. His eldest son, who is now 40 years of age, has also been treated for renal cell carcinoma at the age of 39. The daughter is 37 years and has been diagnosed with renal cysts. The youngest son, who is 35 years of age, wants to know if there is a hereditary cause of renal carcinoma and/or renal cysts in his family and if he is at risk of developing renal carcinoma and/or cysts. He also wants to know if he is eligible for screening and/or preventive

therapy. Pathology reports indicate two types of tumors: a papillary and a clear cell carcinoma in the father and a clear cell renal carcinoma in the eldest son.

2 Existing clinical guidelines

The Clinical Geneticist will currently consider two or more cancer cases in first-degree relatives seriously, especially if the type and subtype of the cancer fall within a known spectrum, or if at least one patient is under the age of 40. For the case of this family, the current decision tree would consider the young age of diagnosis of kidney cancer, the histological type of renal carcinoma and the co-existence of renal cysts and suggest single gene mutation analysis of the *VHL* and *FLCN* genes, which based on current knowledge would be the most likely candidates. No mutation was found in either gene. The decision tree at this point would recommend mutation analysis for the next most likely candidate genes associated with familial renal cell carcinoma: *MET*, *TSC1* and *TSC2*. A novel variant in *MET* was found in the father and the oldest son. Due to these findings the following questions immediately occur: does this *MET* mutation cause the renal carcinoma in the father and his eldest son? Do cysts occur in individuals with this *MET* mutation? Are the kidney cysts in the daughter a coincidence? Should the youngest son be tested for the *MET* mutation? Is he eligible for screening or is a one-time only examination sufficient?

This family is a means to illustrate the current approaches of genetic/hereditary diseases and also to underscore the value systems medicine will have in the relatively near future. A patient and/or his family are referred to a clinical geneticist because hereditary disease is suspected. Based on medical history, family history and sometimes pathology data, specific individual genes will be analyzed. When a disease-causing mutation is found, relatives are counselled optionally and individuals with the disease-causing mutation will undergo screening and/or preventive therapy. However, the clinic is reaching a point now where increasing knowledge and analytical laboratory technologies will certainly change the approach to most if not all genetic/hereditary diseases. Inexpensive and rapid molecular genetics coupled to linked data pertaining etiology and pathophysiology will provide new ways to approach or treat patients with genetic/hereditary diseases. Genetic diseases are usually complex and dynamic processes. Accordingly, a genetic disease should be analyzed as a complex network of components interacting with each other in time and physiology leading to development of (clinical symptoms of) the disease. Although these concepts are not revolutionary, the lack of tools and training available to Doctors or support staff has made it virtually impossible for clinical applications of systems medicine except in a few isolated events.

Treatment would ideally be accomplished through this approach by interfering in the network or interactions between components. Elucidating these networks or pathways can lead to more individualized treatment of the disease.

One common problem in the Genetics Clinic that will likely be aided by Systems Medicine is exemplified by our family. What is the exact consequence of the nucleotide variant in *MET*? Changes in nucleotides that do not result in a truncation of the transcript are often unclear as to whether they will contribute to disease or not. Assumptions are made based on evolutionary conservation, whether the variant has been previously associated with disease, and occasionally functional data is available. Modelling data based on protein structure are also often helpful in weighing the likelihood that a given variant is pathogenic. Because genetic testing is still usually analyzed one gene at a time, weighing variants is time consuming and can be even misleading, if the variant being weighed is only a small piece of a larger picture. However, we anticipate that many patients will have their genomes or exomes entirely sequenced in the near future and the full mutational load per individual will reflect one datapoint for that person which can be analyzed as a network or profile. With the advent of systems medicine, we anticipate that complex modelling will steer our understanding of individual mutational loads. Factors will be identified that have either a synergystic or cummulative effect on the development of a disease.

3 Transitioning to Systems Medicine

The clinical geneticist can fulfil a central role in the translation of systems biology to the clinic and can apply this approach for certain groups of patients. To return to our case study, the questions of the youngest son lead to important considerations and questions:

1. Do the patients with renal carcinoma and/or renal cysts have a germline mutation?
2. Are renal cysts part of the spectrum of this hereditary form of renal carcinoma?
3. Is there a correlation between features of renal cysts (number, size) and renal carcinoma?
4. Is there a correlation between histology and prognosis?
5. Which genes/pathways are involved in the development of a normal kidney cell into a renal cyst?
6. Which genes are involved in the development of a normal kidney cell into renal carcinoma?
7. Is there a common pathway in the transition of a normal kidney cell into a renal cyst and renal carcinoma?

8. Is there a genotype-phenotype correlation?
9. Do modifier genes or exogenous or endogenous factors influence symptoms, histology or prognosis?
10. Which genes are active during different stages of the disease (varying from normal to renal cyst and/or renal carcinoma: normal vs renal cysts vs renal carcinoma vs renal carcinoma and cysts?
11. Are there biomarkers available?
12. Can a specific screening program be advised based on histological data and biomarkers?
13. Can we use the individual's genetic passport to tailor therapy?

To answer these questions, the clinical geneticist would collect clinical data of both affected and unaffected relatives. Factors such as gender and known risk factors like smoking and blood pressure must be annotated for each individual and linked with the pedigree of the family. Retrospective and prospective links to histological data need to be available in the pedigree.

The Clinical Geneticist is trained to interpret mutational data, and evaluate the likelihood a given clinical symptom results from genetic variants. The father and the eldest son both had renal carcinoma (although different histological subtypes), and both also carry a *MET* mutation. Because the daughter without renal carcinoma but with renal cysts did not have the *MET* mutation, the *MET* mutation is more likely to be associated with renal carcinoma than with renal cysts. The location of this *MET* mutation is important for its significance: is this a highly conserved area in the genome, does this mutation lead to a change in mRNA (ie codon, splicing, structure)? Functional studies might be able to discriminate between proteins deriving from mRNA either with wild-type *MET* or with the *MET* mutation; however, the read-out (eg. tyrosine kinase activity or ligand binding) can be complex or uninformative. The Clinical Geneticist will have to tackle these same problems in the future as well, when the data available per patient will of an entirely different order of magnitude.

With new techniques like next generation sequencing, mutation analysis of a complete genome will almost certainly become commonplace. With this technique, known and unknown genes are analyzed in a short time and for a fraction of previous costs of mutation analysis. By analyzing the whole genome additional mutations will be identified. It is the responsibility of the Clinical Geneticist to analyze the role of these mutations on the development of the disease. Do (rare) mutations elsewhere in the genome (in other genes, introns and/or exons) influence the clinical constellation? Do they accelerate or delay the process of developing renal carcinoma or protect against another

disease like sickle cell anemia protects against malaria, conferring an evolutionary advantage? If these additional mutations lead to a difference in histology and/or prognosis and the presence of mutations in other or more genes are required for renal carcinoma, reconstitution of an abnormal protein or intervention in the abnormal protein interaction can delay of prevent the development of renal carcinoma.

The next step after viewing the genomes (DNA), transcriptomes (mRNA) and proteomes (protein) is to evaluate where gene products are expressed (localisomes) and in which metabolic network or enzyme it is involved (metabolomes). When common genes and pathways are assumed to play a role in the transition of a normal renal cell into renal cysts or renal carcinoma, evaluation of expression of genes in these different stages is necessary. Subsequently, an intervention of a specific network node or edge should lead to amelioration of one or more aspects of renal disease. In other words: when the genes and proteins are evaluated, expression of the proteins and their interactions must be evaluated in different stages of the disease, in this case renal carcinoma and renal cysts. By understanding these pathways, specific medications which modulate these specific processes at different stages can be developed so symptoms can be delayed or prevented.

It is clear that evaluation of these different systems comes with a mass of data. The challenge is to extract information that will optimally benefit patients. Essential for this transition is stringent data annotation: for example, a symptom, histological stage or mutation should be accurately and unambiguously labelled throughout the whole world. Semantic Web tools are currently being built to facilitate integration of relevant sources. Importantly these Semantic Web tools will in turn create ontologies which can be used to iterate relationships across individuals and query new hypotheses.

Beyond diagnostic uses, the personalized networks that will be generated for individuals will also generate predictive data that can be tested prospectively (Fig. 1). Furthermore, by combining clinical and histological data and data of molecular pathways and networks with functional annotation it is possible to develop useful biomarkers and treatment. For example, it would enormously benefit patients to have a renal specific antigen (RSA) like prostate specific antigen (PSA) to detect presymptomatic renal carcinoma. This biomarker could be used for (1) presymptomatic diagnosis, (2) stratification of disease, (3) assessment of the progression of the disease, (4) following patient's response to therapy and (5) identifying reoccurrences.

In the clinic of the future, let us speculate that whole genome screening demonstrates a subset of additional mutations in the father and the eldest son from our family. Given co-segregation of renal carcinoma and the *MET* mu-

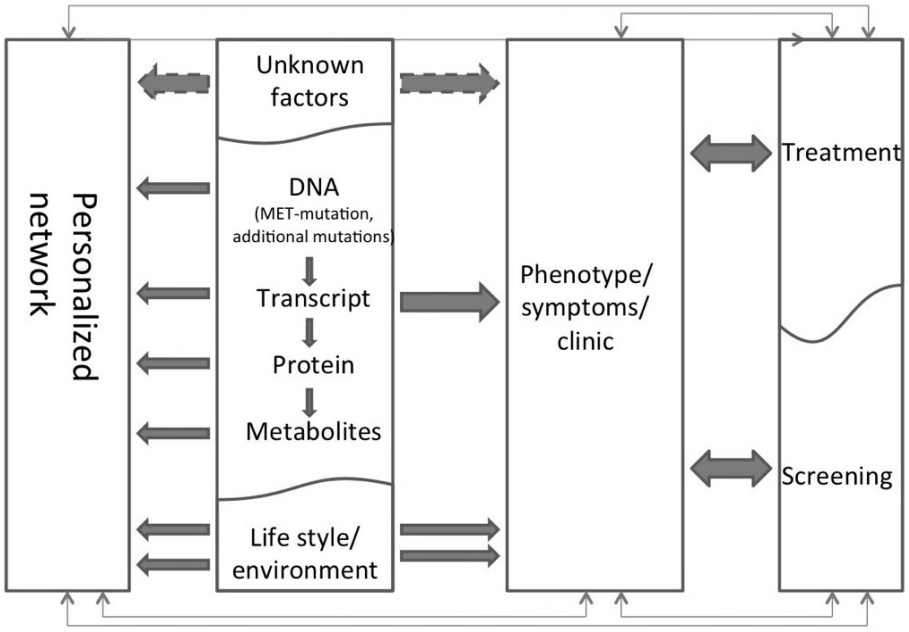


Figure 1: A theoretical view of a personalized network applied to a single patient. The arrows represent the relationship between the four vertical blocks (i.e. personalized network, causal factor, phenotype and treatment/screening). The dotted arrows indicate unknown (causal) factors. Causal factors lead to a specific phenotype. Symptoms drive treatment and/or screening prevention program, which in turn influence the phenotype. Collecting data on causal factors, phenotype and treatment/screening on different times of the period from normal condition to disease and vice versa gives a good picture of the interactions between the different parts of the network and can lead to a personalized profile for a patient, displayed by the arrows in the figure.

tation, the *MET* mutation is associated with an increased risk of renal carcinoma. So what do the additional mutations mean? Functional tests could help to discriminate the value of additional mutations in the development of renal carcinoma and/or renal cysts, but these are often laborious and outside the skill set of the Clinical Geneticist. If tests demonstrate that a specific subset of additional mutations either with of without the *MET* mutation give an increased risk of renal carcinoma, individuals without a *MET* mutation but with this subset of additional mutations could be considered eligible for preventative screening. Additional mutations linked to clinical outcome profile can be plotted. The type of screening can be subsequently adjusted to the profile. Individuals without the *MET* mutation, but with the additional mutations associated with an increased risk of renal cancer are eligible for screening, but perhaps less frequently if these additional mutations demonstrate a lower

risk of renal cancer compared with the risk of renal cancer in presence of the *MET* mutation. On the other hand, the daughter known with renal cysts has neither the subset of additional mutations, nor the *MET* mutation. Although the Clinical Geneticist will probably consider the daughter at be at low risk for kidney cancer he/she might want to compare the daughter's genomic profile to other patients with cystic kidney disease, in order to determine what risk that phenotype might carry.

Another expertise held by the Clinical Geneticist that will only become more necessary as Systems Medicine evolves, is the ability to interpret and communicate any incidental findings as a result of deep data acquisition. For example, in our family, the youngest son has decided to undergo whole genome sequencing to understand his risk of renal cell carcinoma. No mutation in either the *MET* gene or in the associated subset of genes known in his family was detected, and he is told by his Clinical Geneticist that he does not require further screening. However, imagine if a known deleterious mutation is identified in the son which indicates that he will develop Huntington's disease later in life. Clear guidelines for the handling of incidental findings have already been implemented in the Clinical Genetics Clinics and broad experience has been obtained. The data acquisition required to build any framework attempting Systems Medicine will need to rely on the expertise of several different specialists, with the Clinical Geneticist possibly taking a central or coordinating role.

Increasing knowledge and laboratory technologies already have and will still lead to immensely rich patient profiles, individualized for specific prognosis and treatment. The challenge is to extract the information that will optimally benefit patients. Annotation and ontology implementation will make data integration possible, from type of mutation in the genome to deep phenotyping, across hospitals and between groups of individuals. Approaching a genetic disease as a complex network of components and interactions between these components shall hopefully lead to better treatment of prevention of symptoms due to genetic/hereditary conditions.

Acknowledgements

The authors acknowledge funding from the EU FP7/2009 Systems Biology Grant 241955 "SYSCILIA".

Biological Complex Systems

Abdallah Zemirline[1] and Vic Norris[2]

[1] Lab-STICC, UMR 3192, Département Informatique, UFR Sciences et Techniques, Université
de Bretagne Occidentale, C.S. 93837 BP 809, F-29238 Brest Cedex 3, France.

[2] Theoretical Biology Unit, Department of Biology, Université de Rouen, 2829, F-76821 Mont Saint Aignan Cedex, France

Abstract

At all levels, biological complex systems have remarkable characteristics. In this paper, we describe some of these characteristics, such as the multi-level and nested structures of these complex systems, the multiple interactions between their constituents, their interaction functions which are often non-separable, their interaction networks which often are not graphs but hypergraphs, etc. We also emphasize the great difficulty —if not the impossibility —of defining a measure of complexity for these complex systems whose structure is itself dynamic.

1 Complex Systems

Complex systems [1, 23, 37, 39] are found in many business areas and are investigated in various scientific disciplines that include theoretical physics, systems theory, social sciences, mathematics, bioscience, and bioinformatics.

1.1 System

A finite non-empty set S is called a *dynamical system*, or simply a *system*, if it evolves in time, i.e., if its *behaviour* as well as its *properties* vary in time; where *behaviour* denotes all the actions by which the system can modify its relationship with its environment.

In general at each time step, a system has a *state* which is often characterized by the values of the parameters describing the system. The way that a system evolves from one state to another is usually

specified by a *transition function*. This function, or alternatively its generalization, sometimes allows the system behaviour to be understood. A system often has non-empty parts, called *entities*, which can interact with one another.

1.2 Observability

We say that an entity or a property is *observable* if the entity itself or its effects can be observed. The *observability* of a phenomenon does not necessarily imply its precise measurability; observability is simply the possibility that the phenomenon or its effects can be detected.

1.3 Essential entity

Let S be a system and T be an entity of S; T is said to be *essential* if its behaviour and its properties have a straightforwardly observable influence on the behaviour and properties respectively of S. An entity T of a system S is said to be *proper* if T is strictly contained in S. A proper essential entity of a complex system is sometimes also called a *sub-system* or a *component* of S.

1.4 Unpredictability

A behaviour is *unpredictable* if it is impossible to foresee it with certainty. Complex systems often behave in ways that are not predictable, and this unpredictability occurs even when the behaviours of the components of these systems are well known.

1.5 Interaction

Hereafter, an interaction may involve two or more proper essential entities or constituents of a system that are not necessarily located at the same level. An example of interaction between two entities not located at a same level is depicted in [24].

1.5.1 Binary interaction

A *simple interaction* or *binary interaction* denotes an interaction between two proper essential entities (constituents) of the system. An example of binary interaction is given by the Newton's law of universal

gravitation [36, 47]. This law states that two bodies exert on each other a force of attraction proportional to the product of their masses and inversely proportional to the square of the distance separating them.

1.5.2 Multiple interaction

A *multiple interaction* is, by definition, an interaction between at least three entities of a same system. For example, in a set of biochemical reactions, a reaction between one enzyme and one metabolite is a binary interaction whilst a reaction involving one enzyme and at least two metabolites is a multiple interaction. Another example of multiple interactions is the tide, which results from the joint attraction of the moon and the sun on the ocean [10, 42]. In this multiple sun-earth-moon interaction, the moon's influence is three times that of the sun because the earth is much closer to the moon than to the sun (even though the mass of the moon is much less than that of the sun). A biological example of multiple interactions is the interplay between around ten different proteins and other cellular constituents during cell division in *Escherichia coli* [21] and *Bacillus subtilis* [11]. In the latter, for example, the formation or stability of the complex between the FtsZ and DivC proteins is believed to require the interaction of this complex with the DivIB protein [11].

1.6 Interaction function

In exploring some complex system problems, we often need to consider an interaction function defined on the interaction set of the system in question. This function associates some value, not necessarily scalar, with each interaction. For example, an interaction function could be defined in terms of the *stoichiometric coefficients* of a biochemical reaction set describing a metabolism. Two other examples of interaction functions come from Newton's law of gravitation, and from Coulomb's law of electrical forces, which is used to calculate the magnitude of electrical force between two charged bodies [47]. For Newton's law, the interaction function is the set of all the functions of attraction between two bodies whilst, for Coulomb's law, the interaction function is the set of all the magnitudes of the electrical forces between two charged bodies.

An interaction function is fundamental to the construction of the full network describing the nature and effects of the interactions between the system's constituents. Obtaining an interaction function for biological processes, would allow better insights into the behaviours and properties of biological systems. For this reason, it is important to emphasize that an *interaction network* and its interaction function must be constructed with the appropriate accuracy.

1.7 Complex system

Definition 1.1 A system S is said to be *complex* if one of the following statements is true:

(i) S has at least one essential entity that has an unpredictable behaviour.

(ii) Among the essential entities constituting S at least one is not fully observable.

(iii) At least one effective interaction between essential entities of S is either not observable or not precisely known.

Note that in statement (i) the expression *essential entity* does not necessarily mean a *proper* essential entity. In characterizing complex systems, it is often said that: "The whole is more than the union of the parts". This will be considered in the next section devoted to the concept of *emergence*.

Remark 1.1 Assertions (ii) and (iii) deal with complex systems containing, respectively, an entity and an interaction either or both of which are non-observable. Such situations are perfectly plausible as can be seen in the two following examples:

a) By measuring the changes of the impedance during cell division [16], all the phases of mitosis were monitored in real time in a single cell. In particular, it was found that the use of electrodes of the same size as the cell increased the sensitivity of the measurements and allowed changes of properties to be observed that cannot be observed using optical methods.

b) Dark matter, which cannot be observed directly with telescopes, is hypothesized to account for a large part (more than 80 %) of the matter of the universe (which is a complex system of course) [46]. On July 2012, the ATLAS experiment announced that it had observed a new particle: a boson consistent with the Higgs boson [40]. This result is an advance in the understanding of the basic forces holding the universe together. In particular, this new boson provides support for the existence of the proposed Higgs field, which explains how some particles come to have mass and others don't. Indeed without mass, matter that is known until now could not exist. Does the discovery of the Higgs boson lighten the world of dark matter, as the physicist Sean Carroll thinks [41]?

Remark 1.2 What is not a complex system? Can, for example, a supersonic aircraft which consists of many interacting (mechanical, thermodynamic, electrical, automatic, electronic, software, etc) constituents be regarded as a complex system? According to Definition 1.1, it cannot. Indeed, none of the three statements of this definition hold: The behaviour of each component of the aircraft is readily predictable; each of its components is readily observable; and all the interactions between its components are well known and precisely measured. Furthermore, the behaviour of an aircraft cannot be the behaviour of a submarine nor of a sophisticated device other than an aircraft. When the behaviour of an aircraft is not what is expected this is due, except in very rare situations, to a technical failure or a pilot error.

1.8 Examples of complex systems

Examples of complex systems include single cells and multi-cellular organisms, the nervous system, the genome, an ant colony and an insect swarm, as well as human societies, social structures, market economies, ecosystems, and technology infrastructures like those in telecommunications, energy production and biotechnology. Hereafter, we describe a few examples of complex systems. The first example, taken from Physics, shows that complex systems are not necessarily complicated, whilst the last two, taken from Cell Biology, show that complex systems can be extremely complicated.

1) The *Heisenberg uncertainty principle* states that the more precisely the *position in time* of a particle (electron, neutron, proton) is

known the less certain is the calculation of its *momentum* (resp. *energy*), and conversely, cf. [18, 8]. This principle places a fundamental limit on the accuracy with which two physical properties, the *position* and *momentum* (or energy) of a particle, can be simultaneously known; i.e. the more precisely one property is measured, the less precisely the other can be determined. Thus, a particle cannot be precisely located; this is why it is not represented by a point but by a cloud of points. It follows that a particle has an unpredictable behaviour and is a complex system.

2) A prokaryotic or eukaryotic cell is obviously a complex system. All three statements of the above definition are true. Indeed, in Cell Biology, one of the main challenges is to understand how the functions of several molecular components combine to produce complex behaviours at the level of the cell [2]. In particular, the reason for certain patterns of organisation of entities is unclear despite the fact that this pattern is essential. For example, it remains unclear why mitochondria can sometime form dynamic networks [4]. The number of mitochondria in a cell depends on the organism and the tissue type; many cells have several thousand mitochondria, whereas others contain only a single mitochondrion. Mitochondria generate most of the cell's energy; they are also involved in other processes such as signalling, cell growth, apoptosis, and cellular differentiation. Thus, like the nucleus, a whole mitochondrion is an essential entity of a cell. The same is true for the Golgi apparatus.

3) Another example of biological complex system is that of *hyperstructures* [29] which constitute a level of organisation intermediate between macromolecules and cells. Many functions in both prokaryotes and eukaryotes are performed by large structures, alias hyperstructures, in which molecules, macromolecules and ions are physically associated. In the case of *E. coli*, *B. subtilis*, *Caulobacter crescentus* and other model bacteria, examples of such hyperstructures include:

- the array of chemotaxis-specific receptors (Tar, Tsr, Trg, Tap, and Aer);

- dynamic, coupled transcription-translation and transcription-translation-insertion (transertion) hyperstructures comprising active RNA polymerases and ribosomes along with the nascent mRNAs and nascent proteins and indeed the highly expressed genes themselves;

- the cytoskeletal filaments MreB, CreS and FtsZ;
- filaments of elongation factor EF-Tu;
- metabolons of glycolytic enzymes;
- foci of the enzymes E1 of the phosphoenolpyruvate:sugar phosphotransferase system;
- clusters of secretion enzymes such as SecA;
- nucleofilaments of recombination enzymes such as RecA;
- the cell division machinery (comprising a lipid domain and proteins such FtsZ, FtsA, FtsI, FtsK and AmiC);
- the DNA replication factory (comprising enzymes such as PolC, DnaB, DnaG and DnaE as well as enzymes responsible for the synthesis of precursors such as ribonucleoside diphosphate reductase);
- cellulosomes and polycellulosomes.

Ambiquitous enzymes can occupy two different positions in the cell and some hyperstructures depend on such enzymes and are *functioning-dependent structures* that assemble only when functioning and that disassemble when no longer functioning. Other hyperstructures are equilibrium or quasi-equilibrium structures that remain even in the absence of a flow of energy or nutrients.

Communication between hyperstructures would then take the form of changes in: DNA supercoiling, ion condensation on charged filaments, signalling molecules, water structures, and distribution of membrane domains. At this intermediate level of organization, hyperstructures would control the phenotype and, in particular, the bifurcations that occur, as during the cell cycle, so that events take place in the right place, at the right time and in the right order.

1.9 Nested complex systems

As the above definition suggests, complex systems may be nested, i.e. the components of a complex system may themselves be complex systems. For example, a colony of bacteria is a complex system made up

of bacteria, all of which are complex systems in their own right. In the same way, a eukaryotic cell comprises the nucleus, mitochondria, the Golgi apparatus, a set of chromosomes, a nucleolus, chromatin, etc., all of which are complex systems.

1.10 Hierarchical representation of a complex system

Hereafter, a directed graph G = (V, A) consists of a set V of nodes and a set A of arcs whose elements are ordered pairs of distinct nodes; i.e. every arc a of A represents a couple (u, v) of nodes and is directed from u to v; u is the *tail* of arc a = (u, v) and v is its *head*. The tail and head of any arc a will be respectively denoted $t(a)$ and $h(a)$. The *in-degree* of a node is the number of incoming arcs of that node, its *out-degree* is the number of its outgoing arcs, and its *degree* is the sum of its in-degree and out-degree. In a directed graph, a *chain* C is a sequence of arcs $\{a_1, a_2, \ldots, a_k\}$ such that $h(a_{i-1}) = t(a_i)$ or $h(a_{i-1}) = h(a_i)$ or $t(a_{i-1}) = t(a_i)$ or $t(a_{i-1}) = h(a_i)$ for $i = 2, \ldots, k$. Vertex a_1 is the initial end of C and a_k is its *terminal end*.

A graph G = (V, A) is said to be *connected* if for every pair of distinct nodes u and v there exists a chain which links u and v. A *path* P is a sequence of arcs $\{a_1, a_2, \ldots, a_k\}$ such that $h(a_{i-1}) = t(a_i)$ for $i = 2, \ldots, k$; a_1 and a_k are respectively the *initial* and *terminal* ends of P. A cycle is a chain such that its initial end and its terminal end are the same node. A *circuit* is a path $\{a_1, a_2, \ldots, a_k\}$ such that $h(a_k) = t(a_1)$. A graph is *acyclic* if it contains no cycle. A directed graph is said to be *acircuitic* if it contains no circuit, and a tree is a connected acyclic directed graph. Detailed fundamentals of graph theory can be found in [5].

The nested nature of complex systems allows them to be represented in a hierarchical form. Let S_1 and S_2 be two distinct subsystems of a system S such that $S_2 \subset S_1$, then S_2 is said to be a *son* of S_1 if there is no sub-system S3 distinct from S_1 and S_2 such that $S_2 \subset S_3 \subset S_1$. When a complex system S has more than one essential entity, it can be represented by a tree G, otherwise by a directed acircuitic graph G, in the following way:

- S is represented by a node s having a zero in-degree and with a *level* in G equal to zero;

- to each sub-system T of S is associated a node t; where its *level* in G is denoted by l;
- if T_1 and T_2 are two entities represented in G respectively by nodes t_1 and t_2 such that T_2 is the son of T_1 then (t_1, t_2) is an arc of G, directed from t_1 to t_2; moreover, if l_1 and l_2 denote respectively the levels of t_1 and t_2 then $l_2 > l_1$.

Let us observe that in modelling biological cells, there are several ways to take into account cell behaviour as diverse as shape change, motility, mitosis, apoptosis, etc. These include Virtual Cell [25], E-Cell [35], ChemCell [32] as well as the cell-centered approach of Merks and Glazier [27], and the cellular Potts model [17].

1.10.1 Measuring system complexity

Complex systems might be put into different classes of complexity if the criteria to use were clear. It seems evident that the complexity of biological organisations differs from that of inanimate systems such as the world climate, the underwater currents, and the complex movements of the vortices which animate the ocean surface [28]. But the exact nature of the difference is less evident [30]. How many sorts of complex systems exist? Are there levels of complexity? How are they related to one another? Unfortunately, as mentioned in [39] and [44] neither Shannon's entropy, nor Chaitin-Kolmogorov's algorithmic complexity, nor Bennett's logical depth, are adequate for measuring system complexity. Furthermore, approximate entropy (ApEn) is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data. ApEn was designed by S.M. Pincus [31] to work for small data samples, unlike accurate entropy calculation which requires vast amounts of data. But regrettably, ApEn is not really appropriate for measuring biological system complexity.

A different approach is to take into account the *nesting* of the system. Let us call the *nesting depth* of a system S the maximum number of levels in its graphical hierarchical representation G, and the *extreme subsystem* a subsystem represented in G by a vertex at the lowest level. That said, now the question is the following. Can we take the nesting depth of a complex system as a measure of its complexity? Again, the answer is unfortunately no. Indeed, how can we compare two complex

systems having the same nesting depth when we cannot make comparison between the complexities of their extreme subsystems? In general, two distinct complex systems do not necessarily have the same extreme subsystems since there are no building blocks which are complex systems, and from which we can build any complex system. As pointed out in [37], the quantum physical level is deemed to be irrelevant. A particle (electron, neutron, proton) is in itself a complex system, but effectively it is not an essential entity for any living microorganism (virus, bacterium, fungus); as a single neuron is not an essential entity for a brain. Therefore, the measure of complexity still remains an open question. Moreover, because of the absence of building blocks which are complex systems and from which we can build any complex system, a recursive definition of a complex system cannot exist.

2 Emergence

The concept of *emergence* dates back far into the past. As an indication, Claudius Galen (129-201 A.D.) has clearly distinguished between emergent properties and non-emergent properties of wholes [13]. Indeed, the idea that ”The whole is other than the sum of the parts”, better known as ”the whole is greater than the sum of its parts”, has been around for a long time, at least implicitly. In his book ”A system of logic” J. S. Mill wrote in 1840 that ”To whatever degree we might imagine our knowledge of the properties of the several ingredients of a living body to be extended and perfected, it is certain that no mere summing up of the separate actions of those elements will ever amount to the action of the living body itself”. The same idea is also present in Psychology where, in Gestalt theory [20], it is proposed that the human eye sees objects in their entirety before perceiving their individual parts, suggesting the whole is greater than the sum of its parts and allowing for the breakup of elements from the whole situation into what it really is. A variety of definitions, descriptions and examples of the emergence phenomenon have been given by biologists, computer scientists, logicians, mathematicians, philosophers, physicists, psychologists, etc. [7, 19] and [23]. In what follows, to make it simple, we do not distinguish between weak and strong emergence [9] and we confine ourselves to a succinct definition of the concept of emergence, where the term “characteristic” means a behaviour or else a property.

Definition 2.1 Emergence denotes the appearance of a characteristic of a complex system that it is not possible to predict from just a knowledge of the respective characteristics of the essential entities that constitute that system.

An unpredictable characteristic is said to be emergent. The concept of emergence is a straightforward consequence of the complex system definition given above; This is even what characterizes and distinguishes complex systems from other dynamic systems. Indeed, when a complex system S possesses an essential entity which is not observable or whose behaviour is unpredictable, or alternatively there is an interaction not precisely known, involving some essential entities of S, then any characteristic (behaviour or property) of S is emergent.

2.1 Examples

The following examples prove that all the life processes cannot be roughly reduced to only biochemical reactions and electrical signals occurring in the neurons and the cell body. In particular, consciousness, thought, qualia and other mental phenomena cannot be fully explained by contemporary physics and chemistry.

2.1.1 Living organisms

Living organisms are complex; from a fertilized ovum an embryo develops inside the uterus, and receives nutrition directly from the mother. Then this multiple interaction uterus-embryo-nutrients leads to the emergence and to the development of magnificent beings. Such a process is, for many of us, enchanting.

2.1.2 Giant Sequoia

The giant sequoia is an evergreen conifer, see Fig. 1. The biggest sequoia has a circumference near the ground of 31.1 m, and is 83.3 m high. The age of this, the biggest tree in the world, often called “the largest living thing on earth”, is estimated at 2000 to 2500 years. Its egg-shaped cones have roughly a size 4.5 cm x 3.5 cm and can contain some hundreds of light seeds of about 3.5 mm long. It is wonderful to think that a giant sequoia is the result of multiple interactions between a single seed the size of an oatmeal flake and natural substrates such as

humus, nutrients, cold water, fresh air etc. The sequoia is just one of many spectacular examples of emergence in the living world.

Figure 1: This picture which shows a giant sequoia comes from the following site: **http://upload.wikimedia.org/wikipedia/commons/e/e3/Sequoia_and_a_car.jpg**. It suggests to the observer that this wonder is one emergence in the living world, resulting from a multiple interaction between a very small seed and several natural substrates. A better picture of a giant sequoia may be found at the website hereafter: **http://www.monumentaltrees.com/fr/arbres/sequoiageant/californie/**

2.1.3 The generation of life from inanimate matter

The captivating riddle of the spontaneous increase in complexity from inanimate matter to the first forms of cellular life has been the focus of much thinking and research for decades, both philosophically and experimentally [26].

2.1.4 The emergence of consciousness

Here, the fundamental problem is to understand the relationship between the conscious mind and the physical body. This problem has also been the focus of much thinking and philosophical enquiry [14].

2.2 Reverse emergence

Can we consider as an emergence the flight behaviour of an aircraft, which can climb, cruise and descend when accelerating, maintaining a constant velocity and decelerating, respectively? This question seems legitimate because none of the aircraft's 'organs' like reactors, wings, rudder, landing gear, etc, is capable on its own of such flight behaviour. However, the answer is 'no' because each of the many organs of the aircraft was designed and built specifically for the purpose of making the aircraft capable of flight. The behaviour of an aircraft is not an emergence: it is what is called reverse emergence. Indeed, for a given dynamic system, reverse emergence refers to the problem of finding all the components, as well as their interactions and rules, required to generate the desired behaviour. Of course, reverse emergence is widely used in engineering [3]; It is also used in Business, Finance and Marketing. Insurance companies, banks and supermarkets make use of reverse emergence to encourage customers to adopt simple behaviours that maximize their profits (which is the desired outcome), frequently making use of Data Mining techniques to achieve this [6].

2.3 Non-Separability

In the following definition f denotes a mapping, defined on a n-dimensional normed space [33] like for example $\mathbb{R}^n$ the space of vectors whose n components are real numbers, which to every ordered n-tuple $(x_1, \ldots, x_n)$ associates the value $f(x_1, \ldots, x_n)$, not necessarily scalar. And the operator o is an internal composition law, like $+, *, /$ etc.

Definition 2.2

The function $f(x_1, \ldots, x_n)$ is said to be separable if there are n functions $f(x_1), \ldots, f(x_n)$ such that $f(x_1, \ldots, x_n) = f(x_1) o \ldots o f(x_n)$ for every $(x_1, \ldots, x_n)$; otherwise, $f(x_1, \ldots, x_n)$ is said to be non-separable [38].

Remark 2.1

1) A function $f(x_1, \ldots, x_n)$ is additively separable if
$f(x_1, \ldots, x_n) = f(x_1) + \ldots + f(x_n)$ for every $(x_1, \ldots, x_n)$.

An example of an additively non-separable function is given by what follows:

$f(x_1, x_2) = 3x_1^2 + 2x_1x_2 + 5x_2^2$.

2) The following assertion can be easily proven:
For $n \geq 2$, a function $f(x_1, \ldots, x_n)$ is separable if and only if there exists two functions $h(x_n)$ and $g(x_1, \ldots, x_{n-1})$ such that $g(x_1, \ldots, x_{n-1})$ is separable and $f(x_1, \ldots, x_n)$ = $g(x_1, \ldots, x_{n-1})$ *o* $h(x_n)$.

With this assertion, a *polynomial* [9] recursive algorithm is obtained for verifying whether a given function $f(x_1, \ldots, x_n)$ is separable. This algorithm, named $Separability(f, n)$, has two arguments which are respectively a function f and a parameter n representing the variable number of f, and returns *true* or *false* depending on whether the function f is separable or not.

$Separability(f, n)$
  if $n \leq 1$ then $true$
  else begin
    if $there\ exist\ two\ functions\ g(x_1, \ldots, x_{n-1})\ and\ h(x_n)$
      such that $f(x_1, \ldots, x_n)$ = $g(x_1, \ldots, x_{n-1})$ *o* $h(x_n)$
    then $Separability(g, n-1)$
    else $false$
  end.

3) The non-separability property may characterize some *undecidable* problems [12, 45], such as the following optimisation problem (P):

$$Maximize\ \{f(x_1, \ldots, x_n)\ subject\ to\ (x_1, \ldots, x_n)\ in\ \mathbb{R}^n\ and\ g_i(x_1, \ldots, x_n) \leq b_i\ i = 1, \ldots, m\}.$$

a) If the function f is non-separable and at least one of the two conditions (i) and (ii) is satisfied:

(i) f is neither *convex* nor *concave* [34];

(ii) the region $\{(x_1, ..., x_n)$ in $\mathbb{R}^n$ and $g_i(x_1, ..., x_n) \leq b_i\ i = 1, ..., m\}$ is non *convex*;

then (P) is *undecidable*, that is, there is no algorithm for solving (P).

Additional elements of optimization theory are found in [34].

b) If the functions f and g_i $i = 1, \ldots, m$ are linear and for $j = 1, \ldots, n$ x_j is constrained to be an integer, then (P) is *NP-complete* [15].

c) If the functions f and g_i $i = 1, \ldots, m$ are linear, then (P) is *polynomial* [22, 15].

Proposition 2.1
If an emergent characteristic of a complex system can be modelled by a function, this function is necessarily non-separable.

Indeed, let C be an emergent characteristic (behaviour or property) of a complex system S such that C is modelled by a function $f(x_1, \ldots, x_n)$, where for $j = 1, \ldots, n$, the variables x_j characterize respectively the subsystems S_j constituting S. Since C is emergent, $f(x_1, \ldots, x_n)$ cannot be obtained from n functions $f_1(x_1), \ldots, f_n(x_n)$, representing respectively n characteristics $C_1, \ldots, C_n$ of the subsystems $S_1, \ldots, S_n$, such that $f(x_1, \ldots, x_n)$ = $f_1(x_1)$ $o \ldots o$ $f_n(x_n)$. Thus, $f(x_1, \ldots, x_n)$ is non-separable.

This proof was given for showing that the emergence problem in complex systems cannot be solved by Von Neumann computing. Besides, in paragraph 1.6 of the previous section, it is mentioned that for better exploring some complex system problems, an interaction function is required. This function is defined on the set of all entities in interaction in the system in question. The following proposition deals with the non-separability of the interaction functions, where a *directed hypergraph* [5] is a generalization of a directed graph in which an arc can have several heads and/or several tails.

Definition 2.3
Let V a finite set. A *hypergraph* is a pair $H = (V, E)$ where E is the set of hyperedges $e_j, j = 1, \ldots, m$, such that e_j is a non-empty subset of V.
A *directed hypergraph* is a pair $H = (V, A)$ where A is the set of hyperarcs $a_j, j = 1, \ldots, m$ such that $a_j = (t_j, h_j)$, where t_j, h_j are two non-empty disjoint subsets of V, called respectively the tail and the head of a_j.
A hypergraph $H = (V, E)$ is said *complete* if its hyperedge set E is the power set $P(V)$ of the vertices; that is, $E = P(V) = \{U \subseteq V; U \neq \emptyset\}$.

A directed hypergraph $H = (V, A)$ is *complete* if its hyperarc set $A = \{(U, V \backslash U)$; for all $U \subseteq V$, $U \neq \emptyset$ and $V \backslash U \neq \emptyset\}$.

Definition 2.4
An interaction network of a complex system S is a triple $N = (V, I, f)$, where:

- (V, I) is a hypergraph or a graph, depending on whether S has a multiple interaction; (V, I) can be directed or not;
- f is an interaction function defined on I, which to each interaction i in I associates some value not necessarily scalar.

Proposition 2.2
The interaction network of a complex system having at least one multiple interaction is a hypergraph and its interaction function is generally non-separable.

Indeed, let S be a complex system having a multiple interaction called I, and let f be the interaction function of S. The interaction i involves at least three entities of S. In the interaction network N of S, then i is necessarily represented by a hyperedge or a hyperarc, and FN has a hypergraph structure. From this fact, it results that $f(i)$ has at least three arguments, each argument representing one entity among the entities involved by i. Now, suppose that $f(i)$ is separable, then every two entities a and b, among the entities involved by i, are interacting.

It follows that N is a hypergraph which has a very particular structure: every hyperedge or hyperarc of N is a *clique* [5]. Generally, the interaction network has not this special property, as it can be seen, for example, with the metabolic networks.

References

[1] C. Adami, *What is a Complex System*, BioEssays 24:1085-1094, 2002, Wiley Periodicals, Inc.
http://adamilab.mmg.msu.edu/wpcontent/uploads/Reprint/s/2002/Adami2002b.pdf

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, Garland Science, (2007).

[3] M. Batouche, S. Meshoul, A. Al Hussaini, *Image processing using quantum computing and reverse emergence*, Int. J. of Nano and Biomaterials, Vol.2, No.1/2/3/4/5, 136 - 142, 2009.

[4] G. Benard, N. Bellance, D. James, P. Parrone, H. Fernandez, T. Letellier and R. Rossignol, *Mitochondrial bioenergetics and structural network organization*, J. Cell Science 120, (2007), doi : 10.1242/jcs.03381

[5] C. Berge, *Graphs and Hypergraphs*, Elsevier Science Ltd, 1985, ISBN:0720404797.

[6] M. J. A. Berry, G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Wiley edit. 2004.

[7] H. Bersini, *"Qu'est-ce-que l'émergence ?"*, editions-ellipses, 2007.

[8] D. C. Cassidy, *Uncertainty - The life & science of Werner Heisenberg*, Freeman, (1992).

[9] D. J. Chalmers, *Strong and weak emergence*, In P. Davies & P. Clayton (eds.), The Re-Emergence of Emergence. Oxford University Press, 2006.

[10] R.H. Charlier, C.W. Finkl, *Ocean energy, Tide and tidal power*, Springer (2009).

[11] R. A. Daniel, M.-F. Noirot-Gros, P. Noirot, and J. Errington, *Multiple Interactions between the Transmembrane Division Proteins of Bacillus subtilis and the Role of FtsL Instability in Divisome Assembly*, Journal of Bacteriology, Vol. 188, No 21, (2006), p. 7396-7404, 0021-9193/06/$08.00+0 doi:10.1128/JB.01031-06

[12] M. Davis, *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions*, Dover edition, 2004.

[13] A. Debru, *Le corps respirant. La pensée physiologique chez Galien*, Brill, 1996.

[14] A. Freeman, *The Emergence of Consciousness (Journal of Consciousness Studies)* Imprint Academic, 2001.

[15] M. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., 1979.

[16] L. Ghenim, H. Kaji, Y. Hoshino, T. Ishibashi, V. Haguet, X. Gidrol, M. Nishizawa, *Monitoring impedance changes associated with motility and mitosis of a single cell*, Lab on a Chip - Miniaturisation for Chemistry and Biology, 10 (19): 2546-2550, 2010. doi: 10.1039/c004115g.

[17] F. Graner, J. A. Glazier, *Simulation of Biological Cell Sorting Using a Two-Dimensional Extended Potts Model*. Phys. Rev. Lett. 69 (13): 2013-2016. 1992. doi:10.1103/PhysRevLett.69.2013

[18] W. Heisenberg, *Les principes physiques de la théorie des quanta*, Gauthier-Villars, (1932).

[19] J. Holland, *"Emergence : From Chaos to Order"*, Perseus Books Group, 1999.

[20] G. Humphrey, *"Psychology of the Gestalt"*, Journal Educational Psychology, 15(7), 401-412, 1924, doi: 10.1037/h0070207

[21] G. Karimova, J. Pidoux and D. Ladant, *Interaction network among Escherichia coli membrane proteins involved in cell division as revealed by bactorial two-hybrid analysis*. Journal of Bacteriology, Vol. 187, No 7, (2005), p. 2233-2243

[22] L. G. Khachian, *A Polynomial Algorithm in Linear Programming*, Sov. Math. Doklady 20, 191-194, 1979.

[23] H. Kitano, *Foundations of Systems Biology*. MIT Press, (2001).

[24] A. Lesne, J.M. Victor. *Chromatin fiber functional organization: some plausible models* EPJE, vol 19, 279-290, Focus Point on "Structure and Dynamics of DNA".

[25] L.M. Loew and J.C. Schaff, *The virtual cell : a software environment for computational cell biology*, Trend in p. 401-406, 2001, doi : 10.1016/S0167-7799(01)01740-1

[26] P. L. Luisi, *The Emergence of Life : From Chemical Origins to Synthetic Biology*, Cambridge University Press; 2010.

[27] R. Merks and J.A. Glazier, *A cell-centered approach to developmental biology*, Physica A : Statistical Mechanics and its Applications, Vol. 352, n°1, pp 113-130, 2005. doi : 10.1016/j. physica.2004.12.028.

[28] H. Y. Nguyen, B. L. Hua, R. Schopp & X. Carton, *Slow quasigeostrophic unstable modes of a lens vortex in a continuously stratified flow*, Geophysical & Astrophysical Fluid Dynamics, Volume 106, Issue 3, p. 305-319, 2012. doi:10.1080/03091929.2011.620568

[29] P. Amar, P. Ballet, G. Barlovatz-Meimon, A. Benecke, G. Bernot, Y. Bouligand, P. Bourguine, F. Delaplace, J-M. Delosme, M. Demarty, I. Fishov, J. Fourmentin-Guilbert, J. Fralick, J-L. Giavitto, B. Gleyse, C. Godin, R. Incitti, F. Képès, C. Lange, L. Le Sceller, C. Loutellier, O. Michel, F. Molina, C. Monnier, R. Natowicz, V. Norris, N. Orange, H. Pollard, D. Raine, C. Ripoll, J. Rouviere-Yaniv, M. Saier, P. Soler, P. Tambourin, M. Thellier, P. Tracqui, D. Ussery, J-C. Vincent, J-P. Vannier, P. Wiggins, A. Zemirline. *Hyperstructures, genome analysis and I-cells*. Acta Biotheoretica. 50, (2002) 357-373.

[30] V. Norris, A. Cabin and A. Zemirline, *Hypercomplexity*. Acta Biotheoretica 53, (2005) 313-330.

[31] S. M. Pincus, *Approximate entropy as a measure of system complexity*, Proc. Nati. Acad. Sci. USA Vol. 88, pp. 2297-2301, March 1991 Mathematics.

[32] S. J. Plimpton and A. Slepoy, *Microbial cell modelling via reacting diffusive particles*, Journal of Physics, vol. 16, p. 305-309, 2005.

[33] B. P. Rynne and M. A. Youngson, *Linear Functional Analysis*, Springer-Verlag, 2008.

[34] W. Sun and Y. X. Yuan, *Optimization Theory and Methods*, Springer, 2006.

[35] M. Tomita et al, *E-Cell : software environment for whole-cell simulation*, Bioinformatics, vol. 15, n°1, p. 72-84, 1999.

[36] K. S. Udupa, *An Investigation of the Universal Gravitation Constant Based on the Celestial Clock*, Lavoisier (2004).

[37] M. H. V. Van Regenmortel, *Emergence in Biology*, In: P Amar, J-P Comet, F Kepes, V Norris (eds), Modelling and simulation of biological processes in the context of genomics. Genopole, 123-132, Evry 2004.

[38] A. Zemirline, V. Norris, *Non-separability, Multiple Interactions and Hypercomplexity*, In: P Amar, F Képès, V Norris, G. Bernot (eds) Modelling and simulation of biological processes in the context of genomics. Genopole, 99-102, Evry 2007.

[39] H. P. Zwirn, *Les Systèmes Complexes; Mathématiques et Biologie*, Odile Jacob sciences(2006).

[40] **http://atlas.ch/**

[41] **http://edition.cnn.com/2012/07/20/opinion/higgs-dark-matter-arroll/index.html**

[42] **http://www.encyclopedia.com/topic/Tides.aspx**

[43] **http://farside.ph.utexas.edu/teaching/em/lectures/node28.html**

[44] **http://www.maths.bristol.ac.uk/˜enxkw/Publications_files/what_is_complexity.pdf**

[45] **http://www-math.mit.edu/ poonen/papers/sampler.pdf**

[46] **http://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy/**

[47] **http://theory.uwinnipeg.ca/physics/circ/node7.html**

The human toponome project: translating the spatial protein network code (toponome) into efficient therapies

Walter Schubert[1,2,3]

[1] Molecular pattern recognition research group, OvG-university Magdeburg, Germany
[2] International faculty, Max-Planck (CAS-MPG) partner institute for computational biology, Shanghai, China
[3] Human toponome project (**www.huto.toposnomos.com**) TNL, Munich, Germany

Abstract

The recent development of parameter-unlimited functional super-resolution microscopy TIS^{TM} (Toponome Imaging System) provides direct access to protein networks at high 2D and 3D resolution in a single tissue section or inside cells. TIS^{TM} is a device that can overcome both the spectral and the resolving power of conventional light microscopy without having to change hardware. It is the first ready-to-use technology for dimension- and parameter-unlimited histological diagnostics and systematic decoding of the toponome at functional super-resolution (toponome: defined as the spatial protein network code in morphologically intact cells and tissues). TIS^{TM} is a highly flexible machine that can adapt to the needs of the researcher: a 4-in-one microscope including (1) routine transmitted light functions, (2) conventional epifluorescence functionalities, (3) paramater-unlimited protein network visualization in real time, and (4) functional super-resolution of subcellular structures and protein clusters in tissue sections and in cultured cells (approx. 40 nm resolution). It is a novel platform providing the robustness needed for the human toponome project, combining industry partners and research institutions. The technology has shown to solve key problems in cell-, tissue-, and clinical toponomics by directly decoding cellular (disease) mechanisms *in situ/in vivo*, in particular at the target sites of cancer in human tissue. Several next-generation toponome biomarkers and toponome drugs are on the way to clinic. The human toponome project has at its goal to unravel the complete toponome in all cell types and tissues in health and disease. The technology is scalable as large cooperative parallel screening devices extracting the most relevant disease targets from protein network hierarchies *in situ*: a novel efficient way to find selective drugs, by escaping the low content trap in current drug target and diagnostic marker discovery strategies, which, as yet, have disregarded the spatial topology of the protein network code.

1 Introduction

Cellular functionalities comprise at least four organizational levels: genome, transcriptome, proteome and toponome. The toponome is the spatial network code of proteins and other biomolecules (e.g. carbohydrates and nucleic acids) in morphologically intact cells and tissues [1, 13, 25]. The term toponome is derived from the ancient greek nouns "τόπος" (topos = place, position) and "νόμος" (nomos = law) [25]. It is substantiated by experimental insight from direct visualisation of toponome structures [1-9, 16, 17]. It indicates that a cell, which organizes this network of biomolecules, follows topological rules enabling coordinated interactions of its molecular components [3]: In a cell, or in the extracellular matrix, every single molecular component of such an interaction must be at the right time point at the right concentration and at the right subcellular location so that a specific molecular network can be formed. This interaction can take place either on the basis of strong or week physical associations of biomolecules, as well as indirectly, by means of diffusible molecules binding to other biomolecules, such as proteins, at distinct locations. Hence, biomolecular networks are characterized by a non-random spatial context of their molecular elements. Consequently, any biomolecular network exerting a concrete cellular functionality obeys rules of topologically defined assemblies of biomolecules —a spatial code (toponome) enabling a directed action, such as a specific information flow across a pathway. The toponome contains the code of conduct exerting these functionalities. Toponomics [25] is a discipline in systems biology, cell biology and histology, concerning the study of the toponome of organisms [14]. The human toponome project comprizes the complete decoding of the human toponome of 20,000 different proteins on a large scale of cell types, tissues and diseases [18] (**`www.huto.toposnoms.com`**).

2 TIS imaging can break the spectral limit and the resolving power of epifluorescence microscopy

Since the first description of the technology in 1990 [24] we look back to many technological and conceptual developments and biological insight, some of which are featured in Fig. 1. First, the only way to assess the combinatorial molecular structure of large molecular systems *in situ* (in a structurally intact cell or tissue) is to co-map dozens, hundreds or thousands of different proteins/biolomecules in one and the same morphologically intact fixed cell or tissue section.

Figure 1: Featured Toponome maps: **(a)** Cover image from [1]: over 7,000 Protein clusters in a single human liver cell; **(b)** Corresponding Research Highlight referring to [1] (text of this highlight is found in Abbot A. Nature, 443, 609, 2006); **(c)** Cover image from [2] showing a cell surface protein cluster network of a single human peripheral blood T lymphocyte.

The only basic principle that can be potentially used for this task, is fluorescence microscopy. However, to capture the high dimensional combinatorial molecular organization of the toponome in one and the same subcellular structure, many more proteins must be co-mapped simultaneously in a cell or a tissue than wavelengths are available within the spectrum of the visible light (between approx 300 and 700 nm). How to overcome this limit? The answer came from the toponome theory of dimension-unlimited molecular imaging [24], reviewed in [14]. This theory is based on large probe libraries (20; 50; 100; 1,000 a.s.o.), in which every probe, binding specifically to a given moiety or biomolecule *in situ* is conjugated to one and the same dye, e.g. fluoresceine-iso-thio-cyanate (FITC). More than one dye per cycle is also feasible. The principle then implies, that a caged robot runs automated and preprogrammed repetitive cycles of (i) probe-dye incubation on stage of an epifluorescence microscope; (ii) imaging the resulting signal; and (iii) inactivation of this signal by means of soft fluorescence bleaching (or inactivation using an appropriate agent) (patents reviewed in cite14). Fig. 2 illustrates this principle from cyclical data acquisition to 3D toponome mapping. This theory, based on the so called "Venice" and the "sunlight" hypotheses of dimension- and parameter-unlimited molecular imaging has been experimentally verified after having developed several generations of corresponding robots (multi-epitope-ligand-cartography, MELC, and toponome imaging systems, TIS) together referred to as imaging cyclers [1], reviewed in [12, 14].

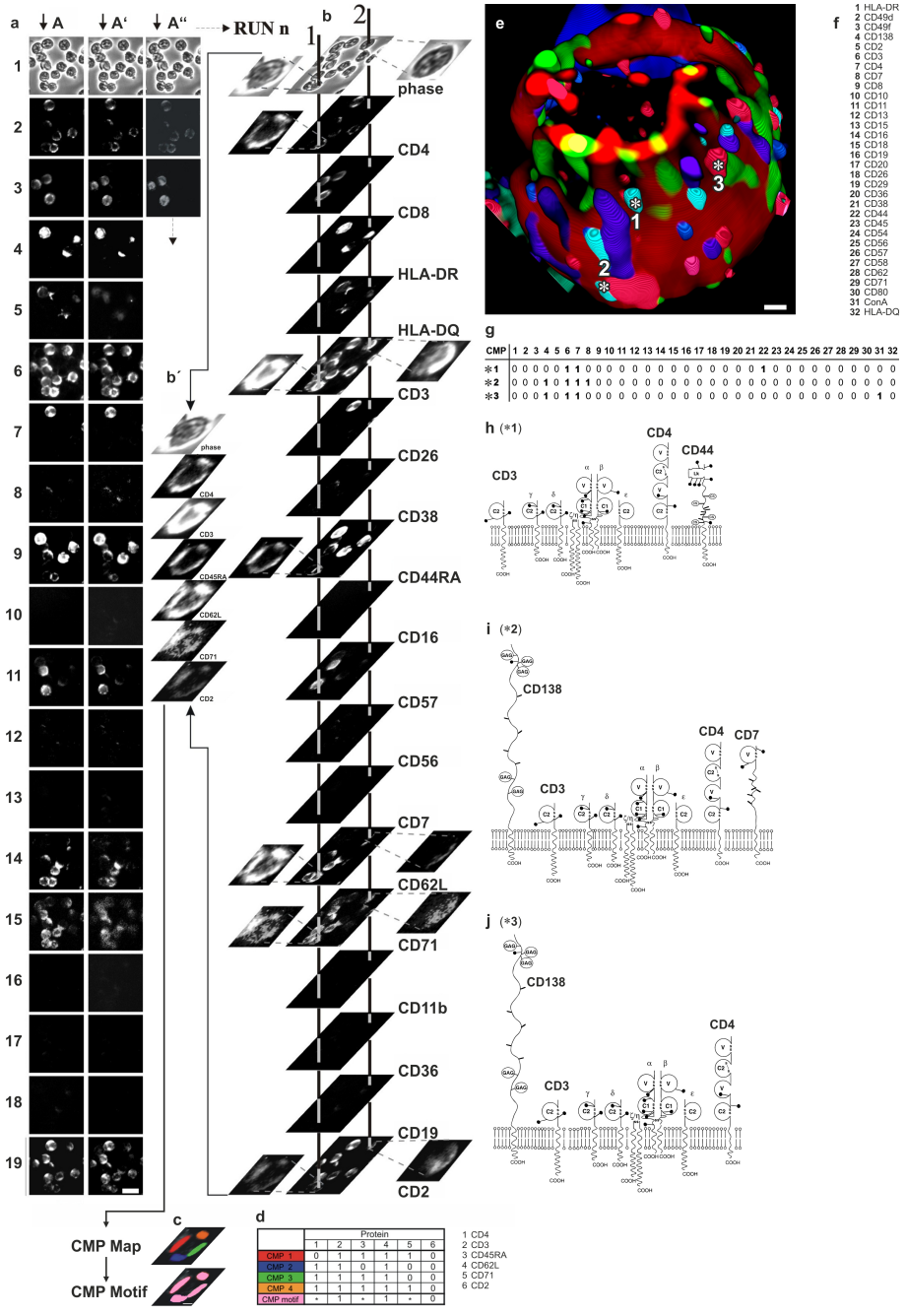


Figure 2: Cyclical assessment of many protein signals in one sample and spatial TIS digital imaging of multi-protein clusters on the cell surface. An example of

the cyclical TIS procedure on peripheral blood mononuclear lymphocytes (PBML) is shown to obtain a 2-dimensional (a - d) and a 3-dimensioal toponome map of cell surface multi-protein clusters (e) including an example of their functional annotation (g-j). **(a)** Altogether 9 TIS cycles, with 2 dye-conjugated probes/antibodies per cycle (FITC, PE as dyes), were run to label 18 different cell surface proteins in one fixed cell sample. The labelled proteins are specified on the right of (b, vertical list). (A, A´, A´´): The same probe library was used to run three repetitive TIS cycles on the identical cell sample (A through A´´), while the sequence of cycles remained unchanged (so called repetitive forward runs). Specificity of protein tagging at any location and lack of sterical hindrance of antibody binding during the TIS procedure is verified by (i) aligning each signal horizontally (A though A´´) and (ii) quantifying correlated signal intensities of the resulting pixel data set by means of mathematical methods: Note - by comparing A, A´, and A´´ (repetitive forward TIS runs) - that signal locations are identical from A through A´´ (horizontal panel of images), while the signal intensities decline due to progressive saturation of the corresponding antibody binding epitopes. Such sets of routine TIS validation procedures, involving repetitive forward, inverted and permuted TIS runs provide further evidence for quantitative precision of TIS. It is part of the so called logic high-end calibration procedure of TIS. **(b)** Illustration of the process of toponome mapping by depicting two cells (b, vertical lines 1 and 2; cells magnified and aligned in b´). Overlay and alignment is used to set thresholds for each fluorescence signal (d, expert based or automated) in order to identify regions of multi-protein clustering (CMPs) and the corresponding CMP-motif in **(c)**, color decoding list in **(d)**. **(e)** 3-D toponome map of a single CD4-PBML obtained by optical sectioning during TIS imaging: 32 TIS cycles were run at each out of 20 different optical planes across the cell (probe/antibody library in **(f)**. **(g)** CMP 1 - 3 are extracted from several thousand CMPs per this cell in total to illustrate which proteins (g, top line) are differentially associated as single protein clusters (CMPs) on the cell surface (e, asterisks 1 to 3 correspond to CMP 1 to 3 in g). **(h-j)** Illustration of the supramolecular assemblies (CMPs) revealed by TIS in (e, asterisks 1 - 3) by using a spatial model of the single co-mapped transmembrane proteins (h-j: corresponding to asterisks 1 to 3 in e, respectively). The corresponding cell surface structure in h to j of the single proteins was reconstructed after [7]. Note that non threshold-based TIS is illustrated in Fig. 4 (with kind permission from Springer Verlag, [14])

Second, the biological validation of this principle was given in many studies having shown that this technology can solve key problems in biology and medicine (reviewed in [9] and [14]).

For example it was shown in an early application [26] that endothelial cells invade the endomysial tube in humans during muscle regeneration, where they transdifferentiate to generate myogenic stem cells regenerating the ruptured muscle segment inside the endomysial tube formed by basal lamina structures.

This detection of transdifferentiation of adult endothelial cells to form muscle stem cells inside human tissue was confirmed by several experimental studies and has been further developed as novel cell therapy model to treat neuromuscular diseases. These latter and the consecutive achievements have been recently reviewed [14]; Similarly, it has discovered a new target protein in amyotrophic lateral sclerosis by hierarchical protein network analysis, a finding that has been confirmed by a mouse KO model. Moreover, it has uncovered a lead target protein in tumour cells that controls cell polarization/metastasis [1, 7], and (iv) it has found a new target protein that controls chronic neuropathic pain, a finding that has been confirmed by an independent KO mouse model [1]. Current research surprisingly showed that toponome fingerprinting of peripheral human blood lymphocytes can detect progressive neurological disease 5 years prior to its clinical onset (Schubert W, personal communication, to be published) and that a monogenetic disease can be successfully treated by conversion of the genotype specific toponome code into normal by using a small non toxic drug (M. Ruonala *et al.*, personal communication, to be published). Together these technological, biological and clinical validations show that MELC/TIS is a mature technology able to decode subcellular disease mechanisms, find novel drug targets and also novel efficient drugs, directly at the target sites of disease inside human tissue (biopsies and/or surgical material).

3 Basic structural and functional units of the toponome

As any system the toponome is composed of subunits (functional entities) on different scales - CMPs, CMP motifs and CMP superfamilies (Fig. 3) [3, 14]. Both the protein hierarchy and the lead proteins detectable within these structures are key to predictivity and therapeutic efficiency, as revealed by toponomic studies: Lead proteins, detected by protein co-mapping on the surface of tumour cells, can be first order key target molecules, because they control both the subcellular topology and the function of large cell surface protein networks, as revealed by the toponomics derived detection of the **A**minopeptidase **Po**larisation **Co**ntrol **N**etwork (APOCON): if the detected cell surface associated lead protein (CD13) is blocked or inhibited by a small molecule, the corresponding protein network, which is arranged along the cell surface membrane completely disassembles, leading to loss of function of the tumor cell to enter the explore state from the spherical state, and thus is inhibited to metastasize [1, 7]. This observation has indicated that the detection of lead proteins *in situ* (rather than by *ex vivo* large scale expression profiling) can be an essential first step in the drive to develop efficient therapies, by using the hypothesis free toponome decoding approach. Principally, toponomics has

the ability to detect relevant lead proteins on a large scale of human proteins in any type of tissue, by using a variety of threshold based and non-threshold based methods to capture the high dimensional combinatorial molecular code of protein systems, and in millions of subcellular data points simultaneously (reviewed in [14]). For example, it was possible to detect more than 2,000 different protein clusters in a single tissue section of prostate cancer [6] and more than 5,000 different protein clusters in colon cancer *in situ* [8]. This has shown that cancer mechanisms are often restricted to subcellular protein rearrangement rather than up or down regulation of their abundance [9, 13, 14]. This is not detectable by *ex vivo* expression profiling.

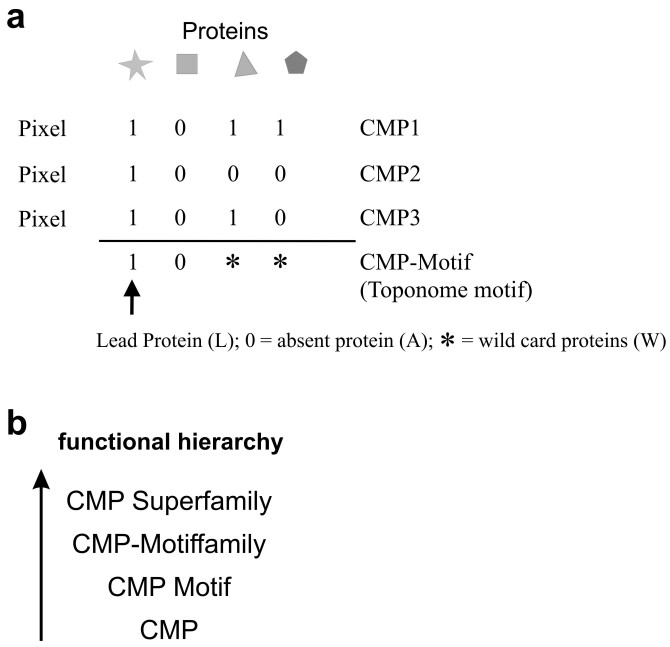


Figure 3: Schematic illustration of the topological and functional hierarchies of proteins within the toponome. **(a)** In a biological system, such as a cell or tissue, an arbitrary number of distinct proteins (symbols in the top line) can form different combinatorial molecular phenotypes (CMP 1, CMP 2,...) at one or several subcellular data points. They can have features in common, thereby forming a functional group termed CMP motif, with: L, lead protein(s) (common to all CMPs of a CMP motif); A, absent protein(s) (absent in all CMPs of a CMP motif); W, wild card proteins (proteins that are variably associated with the (L) and the (A) proteins of a motif [1, 5]). **(b)** Functional hierarchy of CMPs grouped into CMP motifs, CMP motif families, CMP superfamilies a.s.o., having at least one lead protein in common. (with kind permission from Springer Verlag, [14])

The fundamental aspect of the technology related to the decoding capability of protein networks and detection of their selectivity and specificity inside intact tissues is its power of combinatorial molecular discrimination (PCMD) per subcellular data point. For example, if a TIS measuring procedure comprizes the co-mapping of 100 distinct protein and carbohydrate moieties, the resulting PCMD is 2^{100}, under the prerequisite that every fluorescence signal of any given moiety out of these 100 distinct moieties is registered as being present or absent relative to a threshold (automatically determined, or determined by experts: 1 bit per protein [1]), reviewed in [14]. Given the signals are registered without any threshold by using an approach termed similarity mapping (SIM) [9], the resulting PCMD is 256^{100} per data point [9] (Fig. 4). This TIS-SIM approach is entirely performed in real time, allowing the histologist to use it like an electronic microscope in parallel with normal bright field microscopy in routine histological diagnostics, to detect pixels which express the identical protein profiles by being highlighted, while the manually controlled cursor is moving across the tissue [9]. Fig. 4 gives an example showing the high PCMD of 256^{100} per data point discriminating between the three layers of the basal lamina of the skin, having a diameter of altogether 120 nm, as known from transmission electron microscopy. This dimension is rather stable for all basal laminae across all human tissues, as well as for the ultrastructurally distinguishable three basal lamina (BL) layers: lamina lucida, lamina densa, and lamina fibroreticularis. As shown in Fig. 4, TIS-SIM can resolve all these three layers by co-mapping 100 distinct proteins at these sites. This indicates that TIS overcomes both the spectral limit and the resolving power of traditional fluorescence microscopy. In the present example, the resolving power of approx. 40nm is sufficient to discriminate the three layers of the BL.

Together, the information content of TIS data sets is orders of magnitudes greater than that of *ex vivo* large scale molecular profiling or low content few parameter fluorescence imaging [9, 14]. This is underscored also by many 3D toponome imaging data, for example by the toponome of peripheral blood lymphocytes (Fig. 2 e; Fig. 1,(c), cover story).

4 TIS detection of skin lamina densa as a giant supermolecule

TIS-SIM data show that the lamina densa is a giant supermolecule which is unique for this BL layer and inherently different from the suprabasal epidermis and the infrabasal dermis. This is indicated, because only the pixels which belong to the lamina densa structures are highlighted in the same color, displaying the identical protein profile alongside the whole lateral extension of this band like structure (Fig. 4, d, green profile). Hence, the pixel protein profiles highlighting the lamina densa is specific and selective for exactly this site.

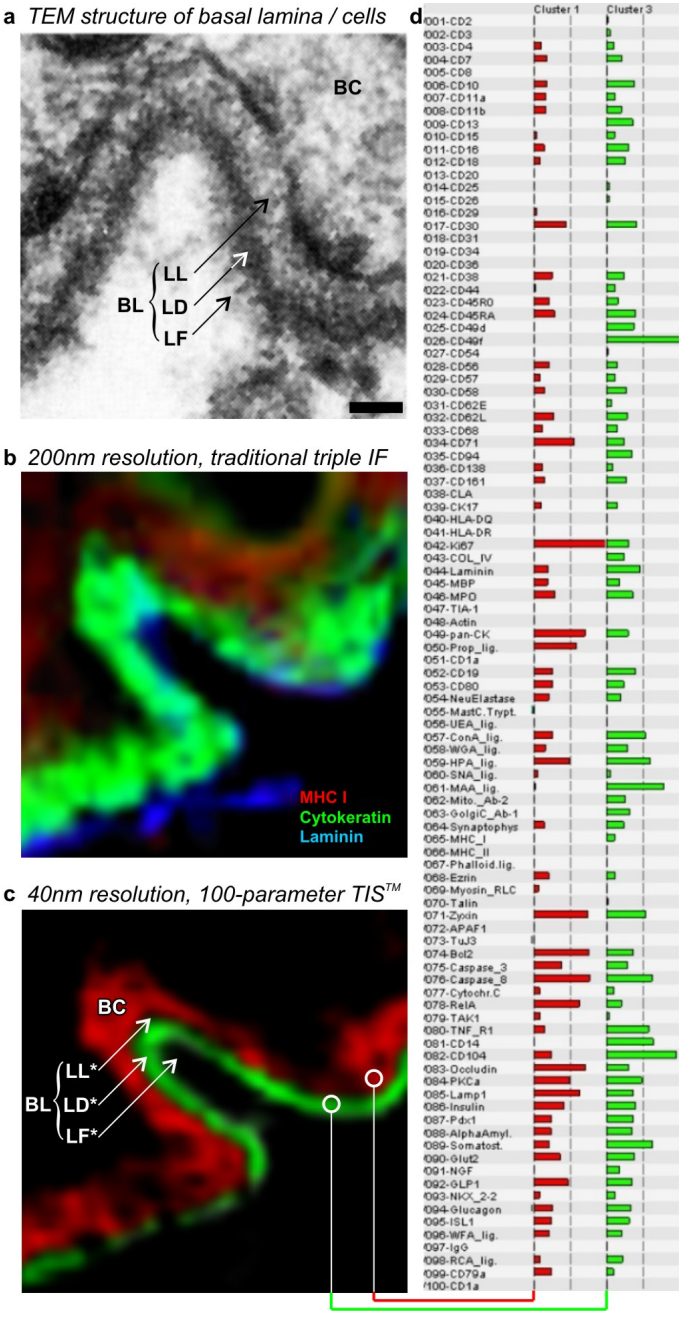


Figure 4: Dermoepithelial junction in human tissue: Visualization of the basal lamina densa as a giant stoichometrically controlled supermolecule. Compare low resolution traditional triple fluorescence microscopy **(b)** with 100-parameter TIS im-

age **(c)** and with corresponding transmission electron micrograph (TEM) (a) indicating that only the functional super-resolution of TIS microscopy (c) can capture the basal lamina densa (LD) as a giant supermolecule at a resolution close to that of TEM: TIS can distinguish the lamina densa (DD) from the lamina lucida (LL) as well as from the basal ceratinocyte (BC) layer, and the lamina fibroreticularis (LF). The protein profile of the corresponding 100 component supermolecule expressed in the LD is seen in (**d**, cluster 2, green), which is different from that expressed in the BC layer (d, cluster 1). Bar: 50 nm **(a)**. Note that images in (b, and c) are simultaneously taken from a 5 micron thick diagnostic frozen human skin tissue section (biopsy material).

Moreover, the highlighted pixels (Fig. 4, green profile) extend across many hundred microns, and the protein profile as well as the relative abundance of the co-mapped proteins is not altered across this site. This indicates that (i) there must be a highly controlled stoichiometry of the co-mapped proteins and of associated carbohydrate structures and (ii) the detected structure is a giant supermolecule expressed as a specific trait of the lamina densa. This has many important implications for the understanding of the system of biomolecules in human tissues *in vivo/in situ* in general: (i) The TIS technology is able to detect supermolecules inside cells and tissues at high functional and structural resolution; (ii) this technological advance enables researchers to uncover practically all existing modes and rules of the topological and functional organization of protein networks on a large scale of all human proteins across human tissues and diseases in sequential arrays of TIS hypercycles (to be published elsewhere); (iii) the TIS approach can be interlocked with genome sequencing. Hence, genome analysis can be combined with the analysis of the genome´s downstream "operational level", which is the toponome, while the toponome has its own spatial coding rules. This serves to understand the interrelationship between genome and toponome structures and functionalities in the same tissues directly.

5 International TIS clusters as virtual factories decoding the toponome

For this purpose TIS clusters will be established as virtual factories both inside any given facilities, but also as virtual factories, by linking several TIS robots located in distant sites (details to be published elsewhere). This will allow a rapid progress of decoding disease mechanisms and fundamental biological processes by combining specialized institutions worldwide. It follows the concept of both fostering and integrating different biological and other grown scientific cultures, such as mathematics, to rapidly gain biological insight. It is believed that the resulting power of functional insight and progressive

understanding will most likely lead to the discovery of the specific mechanisms driving chronic diseases inside the corresponding tissues.

6 Treatment of chronic diseases. Is there a logic of failure?

The steadily increasing inefficiency in treating chronic diseases —in spite of elegant and logic molecular biological studies and helpful applied mathematics —has recently evoked urgent warnings (World Alzheimers Report 2011, [20], commented in [19]). The overall question asked by seriously worried scientists and editors is "Where are we going wrong" [22], and similar concerns have been formulated by others [10, 11, 15], whilst the disillusioned pharmaceutical industry closes discovery facilities and dismisses thousands of employees. Surprisingly, corresponding early warnings in a report entitled "The fruits of genomics" (Lehman brothers and Mc Kinsey 2001) [21] were totally disavowed by the scientific community, whilst the very cause of the problem is obviously not yet known, and many concerns about the ongoing trend of high drug attrition rates in cancer [22] and failures in treating Alzheimers disease [23] remain without clear cut solutions. Obviously, the many elegant logic molecular cell biological and model studies with their convincing scientific rationales failed when the corresponding concepts were translated into Alzheimer´s therapies [23], and similar experiences exist in the field of cancer research [11, 22]. This indicates that the logic of current scientific practice, with well established cell biological and animal models, does not match the logic of chronic disease itself. Why?

A thought experiment (Gedankenexperiment): Given the giant supermolecule detected as specific and selective trait of the lamina densa in the skin (hence present in a morphologically intact tissue) (Fig. 4): If this shown set of 100 distinct biomolecules and the specific stoichiometry of these molecules inside this structure would not be known, but only *ex vivo* large scale expression profiles of the skin (from tissue homogenates) would be known, is it then possible to predict from this *ex vivo* profile the exact composition and stoichiometry of this lamina densa specific supermolecule, while predicting at the same time that other structures around (lamina fibroreticularis, lamina lucida, basal keratinocytes etc: Fig. 4) do not express it? The answer, of course, will most likely be negative. But let us further assume that this mentioned supermolecule would be a specific trait of a given cancer *in situ*, and further that this feature would be the very structure to be selectively targeted by a drug, it is clear that the corresponding selective drug will never be found by a rational approach on the basis of *ex vivo* profiling or *ex vivo* models. This is simply due to the fact, that the information content of the specific composition and topology of supermolecules in tissues is orders of magnitudes greater than

that of *ex vivo* expression data, and therefore cannot be predicted by *ex vivo* expression profiles, or current cell biological or any other models.

7 Toponomics: decoding and treating diseases within the context of histology

An interesting question: why did systems biology leave the tradition line, or, even never regarded the tradition line from the beginning: a line defined by the principle of histology founded by its pioneer Marcello Malpighi (1628 -1694), which has driven medicine into the successful anatomy based discipline? Why was histology as the very structural basis and historically successful field of diagnostics and research in human medicine left, or, even never regarded, in modern molecular systems biology as being important for understanding diseases? One of the obvious answers appears to be the widely accepted, but not uncontroverted view [10] that enormously large *ex vivo* data sets, such as genomics data or *ex vivo* expression profiling data sets, might bring about enough and even more functional insight than any structure bound information inside tissues. Not surprisingly, the resulting data sets, having no relationship to subcellular functional topologies, became so large that secondary technologies and algorithms had to be developed which "look on these data and explain them", and so forth: this system became self referential, establishing a new *ex vivo* data driven medical research, thereby accelerating the divergence between traditional microscopic anatomy and the new *ex vivo* disciplines. But: isn´t a cell or a tissue a large molecular pattern formation apparatus with clear cut topological rules of order to encode function? Don´t we need to understand these topology based modes and rules in all detail to find the Achilles´ heels of chronic diseases? And: can these fundamental principles be understood at all by reducing the cell to a large collection of biomolecules? Toponomics, having been developed independently from the routes of *ex vivo* large scale technologies, can contribute to revert this divergence, by reading out molecular systems as functional entities of morphologically intact tissue, thereby integrating the subcellular topology of protein network codes with large scale expression profiling. Toponomics technology was entirely histology based from the beginning of its invention [24] and, since then, looks back on a 22-years-old tradition of corresponding molecular histology. It builds on a strength that lies in the ability to detect disease specific supramolecular features and events without a priory mechanistic or molecular knowledge. It is largely hypothesis free. For this a high subcellular resolution generated by extreme dimension-unlimited PCMDs is mandatory (Fig. 4). The present status within the human toponome project allows researchers to systematically run large TIS clusters as worldwide virtual factories to decode the complete toponome of human proteins, and interlock these clusters with genome sequencing.

Acknowledgements

Supported by the Klaus Tschira Foundation (KTS), the Deutsche Forschungsgemeinschaft (DFG Schu627/10-1), the BMBF (grants CELLECT, NBL3, NGFN2, and NGFNplus), the DFG-Innovationskolleg (INK15) as well the EU project IMAGINT (Health-F5-2011-259881), as well as the human toponome project (**www.huto.toposnomos.com**). I thank ToposNomos Ltd. for providing access to the TIS^{TM} reference lab, and Andreas Krusche for help with formatting the manuscript.

References

[1] Schubert W, Bonnekoh B, Pommer AJ, Philipsen L, Boeckelmann R, Malykh Y, Gollnick H, Friedenberger M, Bode M., Dress AW. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature Biotechnology* 2006; 24(10): 1270 - 1278 (front cover story; Res highlight "Mapping togetherness" in *Nature* 443, 609, 2006).

[2] Friedenberger M, Bode M, Krusche A, Schubert W. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system (TIS): sample preparation and measuring procedures. *Nature Protocols* 2007;2(9):2285-94 (front cover story).

[3] Schubert W. A three symbol code for organized proteomes based on cyclical imaging of protein locations. *Cytometry A.* 2007 Jun;71(6):352-60.

[4] Schubert W, Friedenberger M, Bode M, Krusche A, Hillert R. Functional architecture of the cell nucleus: towards comprehensive toponome refrence meaps of apoptosis. *Biochim Biophys Acta.* 2008 Nov;1783(11):2080-8.

[5] Bode M, Irmler M, Friedenberger M, May C, Jung K, Stephan C, Meyer HE, Lach C, Hillert R, Krusche A, Beckers J, Marcus K, Schubert W. Interlocking transcriptomics, proteomics and toponomics technologies for brain tissue analysis in murine hippocamapus. *Proteomics.* 2008 Mar;8(6):1170-8.

[6] Schubert W, Gieseler A, Krusche A, Hillert R. Toponome mapping in prostate cancer: detection of 2000 protein clusters in a single tissue section and cell type specific annotation by using a three symbol code. *J Proteome Res.* 2009 Jun;8(6):2696-707.

[7] Schubert W. On the origin of cell function encoded in the toponome. *J Biotechnol.* 2010 Sep 15;149(4):252-9.

[8] Bhattacharya S, Mathew G, Ruban E, Epstein DB, Krusche A, Hillert R, Schubert W, Khan M. Toponome imaging system: *in situ* protein network mapping in normal and canerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their suncellular annotation by using a three symbol code. *J Proteome Res.* 2010 Dec 3;9(12):6112-25.

[9] Schubert W, Gieseler A, Krusche A, Serocka P, Hillert R. Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *Nat. Biotechnol.* 2012 Jun 15;29(5):599-610.

[10] Gatenby R. Finding cancer's first principles. Perspective. *Nature* 2012; Vol 491, No. 7425, S55.

[11] Editorial. Thoughts for a new year. *Nat Rev Cancer* 2013; 13, 1.

[12] Schubert W. Toponomanalyse. In: Lottspeich F., Engels J.W. (Eds.) *Bioanalytik* 3rd edn. Springer Spektrum, Berlin, Heidelberg: 1140-1151. 2012

[13] Schubert W, de Wit NCJ, Walden P. Systems Biology of Cancer. In: Pelengaris S., Khan M. (Eds.) *Molecular Biology of Cancer*. 2nd edn. John Wiley & Sons 552-582. 2013 (in press).

[14] Schubert W. Toponomics. In: Dubitzky W, Wolkenhauer O, Cho K, and Yokota H. (Eds.) *Encyclopedia of Systems Biology*. Springer Verlag, 2013 (in press)

[15] Michor F et al. What does physics have to do with cancer? *Nat Rev Cancer* 2011; 1, 657 - 670

[16] Murphy F. Putting proteins on the map. *Nat Biotechnol.* 2006 Oct;24(10):1223-4 (comment on ref 1, above)

[17] Sage, L The molecular face of prostate cancer. *J. Proteome Res.* 2009; 8, No 6, 2116 (editorial to ref 6, above)

[18] Cottingham, K. Human Toponome Project. *J. Proteome Res.* 2008; 7, No. 5, 1806. (editorial to invited lecture. HUPO world congr 2008, Seoul)

[19] Schubert W. Treatment of chronic diseases. Is there a logic of failure? *Dagstuhl Seminar "Structure Discovery in Biology: Motifs, Networks & Phylogenies"* Dagstuhl castle, Germany, 15-20 July 2012.

[20] *World Alzheimer Report* 2011,
http://www.alz.co.uk/research/world-report-2011

[21] Lehmann Brothers & Mc Kinsey. The fruits of genomics. 2011

[22] Hutchinson L, Kirck R. High drug attrition rates - where are we going wrong? *Nat Rev Clin Oncol* 2011; 8, 189 - 190

[23] Selkoe DJ. Resolving controversies on path to Alzheimer's therapeutics. *Nat Med.* 2011; 17(9): 1060-1065

[24] Schubert W. Multiple antigen-mapping microscopy of human tissue. *In Excerpta Medica* (Burger G, Oberholzer M, Voijs, GP., eds), pp 97- 98, Elsevier (1990)

[25] Schubert W. Topological proteomics, toponomics, MELK technology. *Adv.Bochem. Eng. Biotechnol.* 2003; 83, 198 - 209.

[26] Schubert W. Antigenic determinants of T lymphocyte receptor and other leukocyte surface proteins as differential markers of skeletal muscle regeneration: detection of spatially and timely restricted patterns by MAM microscopy. *Eur. J. Cell Biol.* 1992; 58, 395 - 410.

From automatic and expert annotation to the reconstruction of genome-scale metabolic networks and models

François Le Fèvre[1], Eugeni Belda[1], Damien Mornico[1], David Vallenet[1], Claudine Médigue[1]

[1] Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, UMR 8030 CEA/Genoscope - CNRS - Universite d'Évry, F-91057 Évry, France

1 Abstract

Metabolism is the set of enzyme-catalyzed biochemical reactions that occur in all cells of all living organisms. These reactions, often numbering in the order of thousands, operate together as a network to progressively transform chemicals from the surrounding environment into the energy and chemical species needed by the cells. While knowledge on metabolic reactions has been slowly accumulated for a few model organisms over the past century, the advent of genome sequencing and annotation now enables to transfer part of this knowledge to new organisms and quickly draft their own metabolic networks.

In this course, we will present current resources and tools for automatic and expert annotation of genomic object involved in metabolism and for reconstruction of the metabolic network of a newly sequenced organism, with a bias toward microbial species. The focus will be put on the comprehensiveness of the reconstruction, in the aim of identifying as exhaustively as possible the metabolic capabilities of the studied organism with a focus on resources and tools available in Microme. Microme, a 15-partner project funded by the Framework 7 programme of the European Commission, is a new bioinformatics infrastructure for the curation and integration of bacterial reactions and pathways, for genome annotation and for the reconstruction of metabolic networks.

2 Introduction, overall reconstruction process

Many applications motivate the reconstruction of global metabolic networks. From a descriptive point of view, they provide an overview of all metabolic reactions that can proceed in the cell. For instance, one can then study the structure of the metabolic network, locate each reaction within the whole network, or determine which enzymes are associated with a given metabolic pathway. Studying networks across several organisms introduce new evolutionary issues: how do metabolic pathways evolve? Various types of experimental data also benefit from being integrated on a reconstructed metabolic network [20]. Useful insights often only emerge from experimental data when they are interpreted in the context of the whole metabolism. At a larger scale,

studying the global metabolic network can help in relating metabolic functions found in the genome with observed cell phenotypes, such as the ability for the cell to grow on given chemical environments, or its tendency to excrete particular by-products. Comprehensiveness of the reconstruction is a crucial requirement in this type of application, as global phenotypes result from the operation of the whole set of reactions. The size and complexity of metabolic networks, however, often hinder direct deductions of their system-level properties, e.g. their dynamic behaviors or their global phenotypes. Such deductions are therefore mostly performed using mathematical or computational models of metabolism [9, 14]. An accurate knowledge of metabolic networks is here again key in building faithful models.

In this brief article, we will cover the most useful and fundamental parts of metabolic network reconstruction. Many more details can be found in other comprehensive reviews, which focus for instance on practical aspects [45], conversion to metabolic models [11], or reconstruction of biological networks of broader types [13].

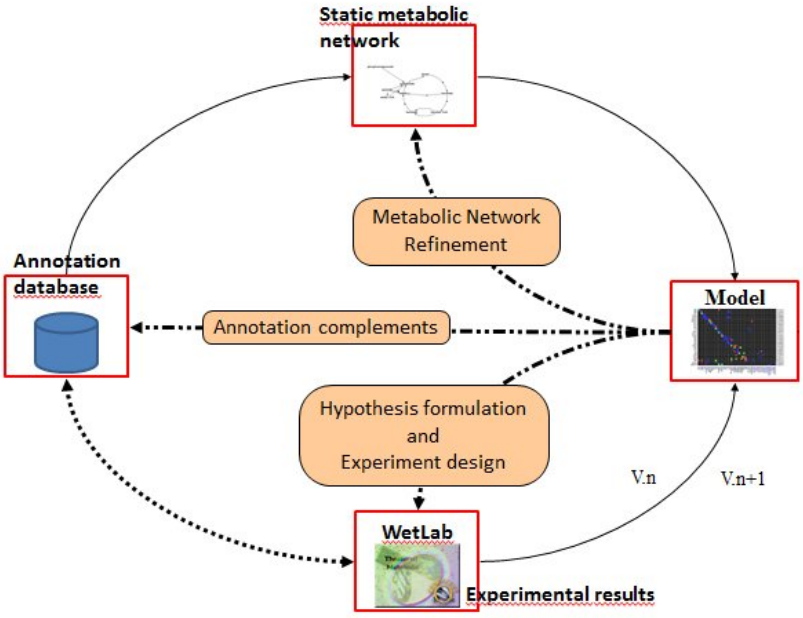


Figure 1: Iterative reconstruction of metabolic networks and models. Data extracted from the annotation database is used to build a static metabolic network, which is then converted to a metabolic model. Analysis of the model in regards of experimental results allows refining the metabolic network, completing genome annotation and formulating hypotheses that could be tested at the workbench.

Network reconstruction generally proceeds in three steps which gradually build and refine the metabolic network. First, information about metabolic reactions is extracted from the genome annotation. Depending on the amount of biochemical knowledge available for the organism, this step can account for a large fraction of the reconstructed network, especially for barely studied organisms. Then, the draft network is completed with organism-specific reactions and metabolic pathways which could not be extracted from the genome annotations. A significant amount of metabolic information is indeed present in the literature and needs to be incorporated in the network. Finally, the completeness of the network is evaluated and missing reactions inferred. These three steps are described in the following sections.

From this metabolic network, scientists can design a mathematical model that can be iteratively refined by comparing its predictions to experimental data, thereby increasing biological knowledge on the organism.

Several platforms allow confronting experimental results against in silico experiments. For instance, CycSim [25] supports the design of knockout experiments: simulation of growth phenotypes of single or multiple gene deletion mutants on specified media and comparison of these predictions with experimental phenotypes.

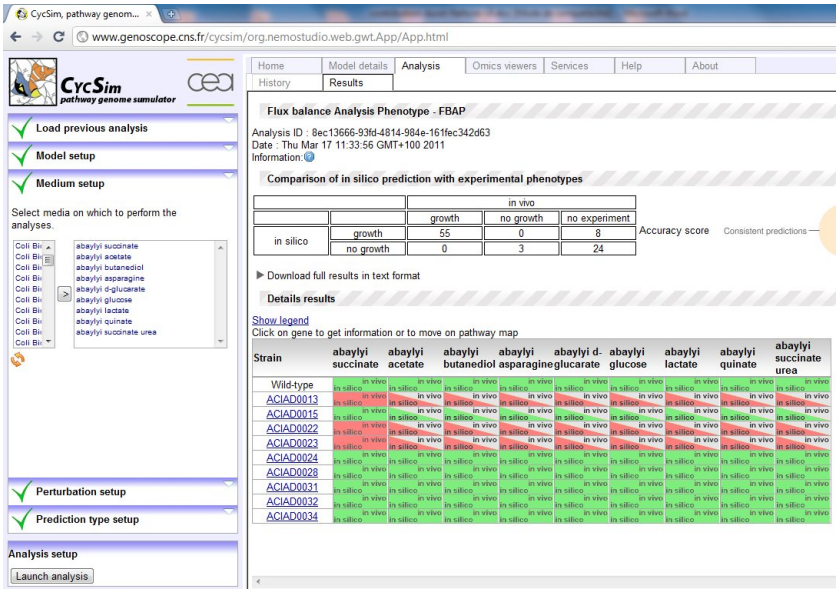


Figure 2: CycSim, a web application dedicated to in silico experiments with genome-scale metabolic models coupled to the exploration of knowledge from BioCyc and KEGG. **`http://www.genoscope.cns.fr/cycsim`**

3 From genome annotation to metabolic reactions

The first step of metabolic network reconstruction is mostly gene- centric. Leveraging on the ability to predict functions for a large fraction of genes, one can derive metabolic reactions from genes with enzymatic functions.

Thanks to dramatic improvements in genome sequencing technology, the number of available genome sequences is quickly increasing (see **http://genomeonline.org** for a review of past and ongoing sequencing projects). Since genome and protein databases GenBank, CMR, IMG, MicroScope, SEED, or UniProt for instance (see Table 1) collect genome information for a wide fraction of sequenced organisms, genome annotation is readily available for many organisms. If not, the raw genome sequence needs to be processed and annotated [29] using integrated annotation platforms such as Microscope [46], IMG [27] or RAST [3].

Focus on Microscope

The MicroScope platform, which provides support for the (re-)annotation of microbial genomes and their comparative analysis (**http://www.genoscope.cns.fr/agc/microscope**), currently hosts about 120 different projects covering roughly 1660 bacterial and archaeal genomes, 866 of which available in public databases.

Following a fully automated annotation and analysis process, data is made available through advanced graphical Web interfaces for manual refinement and for the exploration of comparative genomics analysis results. To these ends, various tools are made available to the platform user, such as dynamic procedures allowing one to analyse genomic islands, synteny conservation, "pan" and "core" genomes within microbial species, gene and metabolic phyloprofiles, etc [46]. The expert annotations gathered in MicroScope (an average of 50,000 a year, created by more than 1000 curators having a personal account on the platform) continue to improve the quality of bacterial genome annotations.

Following the main research activities at the Genoscope (i.e, in-depth reexamination of the metabolic functions of a bacterial organism and discovery of novel enzymatic activities), and in the context of an European collaborative project named Microme, recent evolutions of the MicroScope platform have focused on the prediction and curation of prokaryote metabolic pathways. New strategies for functional annotation of prokaryote genomes have been developed, especially a genomic and metabolic context-based inference method that can propose candidate genes for sequence-orphan enzymatic activities [40]. Results of these strategies are now made available through new MicroScope web interfaces. Procedures and graphical interfaces dedicated to the curation of enzymatic activities, of Gene- Protein-Reaction associations and of predicted metabolic pathways have also been developed. Expert annotations of

this metabolic data are directly imported into the Microme repository to form a core of high-confidence data used in the projection strategy (i.e, to predict new GPRs associations and metabolic pathways in microbial genomes).

These new tools and interfaces has been applied to the functional re-annotation of the genome of Bacillus subtilis 168, a model organism of the Firmicutes, with special focus on the curation of gene-reaction associations derived from experimental evidences and metabolic pathway projections. This work includes the curation of automatically predicted gene reaction associations by PathoLogic algorithm based on MetaCyc reaction and pathway repository, the creation of new compounds and reactions absents in MetaCyc repository in reference databases like RHEA and CHEBI and the validation of these new reactions associated to the corresponding CDSs, and the creation of new metabolic pathways and pathway variants absent in public MetaCyc repository into an internal pathway repository named MicroRefCyc that will be automatically projected into other complete genomes.

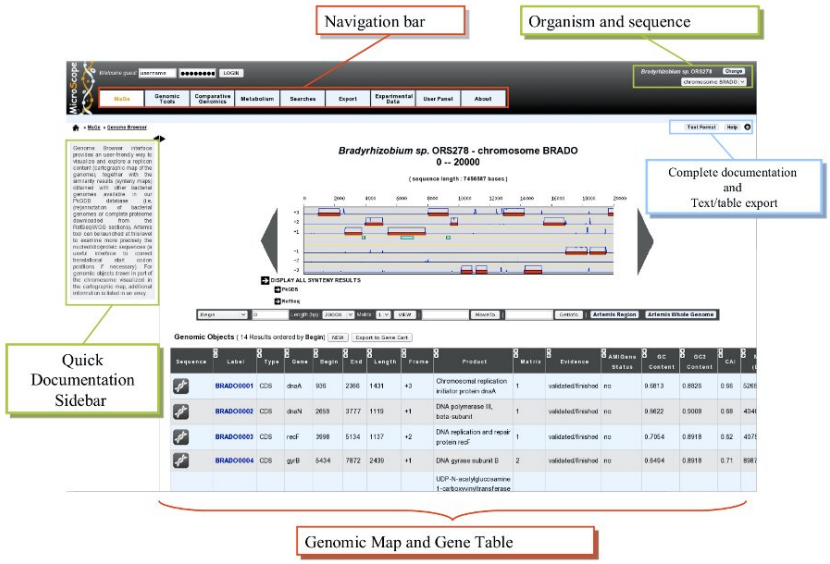


Figure 3: Overview of the Microscope platform: **`https://www.genoscope.cns.fr/agc/microscope`**.

The easiest method to translate gene functional annotations into metabolic reactions relies on Enzyme Commission (EC) numbers which classifies most enzyme activities [4]. Virtually all genome annotation databases associate EC numbers with gene functions and all metabolic databases relate existing EC numbers with the reactions they store (see Table 1). In this process, EC numbers act as direct links between genes and reactions.

Name	Link	Comment
MicroScope	**http://www.genoscope.cns.fr /agc/mage**	Microbial genome annotation database
IMG	**http://img.jgi.doe.gov**	Microbial genome annotation database
CMR	**http://cmr.jcvi.org**	Microbial genome annotation database
SEED	**http://seed-viewer.theseed.org**	Genome annotation database
UniProt	**http://www.uniprot.org**	Comprehensive database of curated and predicted protein functions
BRENDA	**http://www.brenda-enzymes.info**	Database of curated enzyme information
TransportDB	**http://www.membrane transport.org**	Database of predicted membrane transport proteins
Enzyme	**http://expazy.org/enzyme**	Database of enzymatic activities

Table 1: Resources and software tools for microbial metabolic network reconstruction: **Genome annotation/enzyme databases**

Yet, EC numbers present several drawbacks which require using further methods. First, a small set of recently identified enzymatic activities are lacking EC numbers, a consequence of the slow curation process of the EC classification. Second, full EC numbers are not always assigned to gene functions, bringing uncertainty about which activities are really predicted. And third, the specificity of enzyme activities is not explicitly defined for a few EC numbers. For these reasons, additional tools have been developed. They either directly translate textual annotations into metabolic reactions [21] or perform metabolic annotations de novo from genome sequence [43, 32, 2]. These latter tools do not benefit, however, from all improvements made to "classical" annotation platforms, a limitation which may restrict their efficiency in the future.

The BioCyc collection of metabolic databases rely on a such a software, Pathway Tools (see Table 2 and ref. [8, 21]), to build metabolic networks. Pathway Tools actually contains a module called PathoLogic which combines EC numbers with textual annotations and knowledge of metabolic pathways to infer metabolic reactions from genome annotation.

In order to solve those problems, reactions are now resolved not necessarily through their EC numbers but using of reactions database such as Rhea [1]. Rhea is a manually annotated database of chemical reactions; each reaction

Name	Link	Comment
ChEBI	**http://www.ebi.ac.uk /chebi**	Database of chemicals of biological interest
Rhea	**http://www.ebi.ac.uk /rhea**	Manually annotated database of chemical reactions
BioCyc	**http://www.biocyc.org**	Collection of organism-specific metabolic databases
MetaCyc	**http://www.metacyc.org**	Multiorganism, curated metabolic pathway database based on BioCyc
MicroCyc	**http://www.genoscope. cns.fr/agc/microcyc**	Collection of organism-specific metabolic databases based on BioCyc, for all organisms included in MicroScope. (more than 500 organisms available)
Reactome	**http://www.reactome.org**	Curated database of metabolic pathways.
Microme	**http://www.microme.eu**	A specific instance of Reactome dedicated to microbial world and a web portal to tools and resources of all Microme partners.
KEGG	**http://www.genome.jp /kegg**	Comprehensive database of compounds, reactions, metabolic pathways, and genes/proteins
UniPathway	**http://www.grenoble .prabi.fr/obiwarehouse /unipathway**	Database of curated metabolic pathways, linked to UniProt proteins
UM-BBD	**http://umbbd.msi .umn.edu**	Database of microbial biodegradation pathways

Table 2: Resources and software tools for microbial metabolic network reconstruction: **Metabolic databases**

is defined by its chemical equation: list of compounds and their associated stoichiometry (and eventually its compartment origin in/out for transport reactions) In Rhea, each compound is mapped to a compound of ChEBI [28], a database of chemicals of biological interest, ensuring that the reaction is balanced, unique and mapped to the real catalytic activity. Furthermore, each gene or protein functional annotation must be directly linked to their corresponding reaction(s). Several initiatives now focus on directly specifying metabolic reactions during the genome annotation step. The SEED system already associates genes to their metabolic roles [10], and UniProt is formaliz-

Name	Link	Comment
BIGG	**http://bigg.ucsd.edu**	Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions
The Model SEED	**http://www.theseed.org /models**	A resource for the generation, optimization, curation, and analysis of genome-scale metabolic models.
BioModels	**http://www.ebi.ac.uk /biomodels-main**	A repository of peer-reviewed, published, computational models
CycSim	**http://www.genoscope. cns.fr/cycsim**	An online tool for exploring and experimenting with genome-scale metabolic models.

Table 3: Resources and software tools for microbial metabolic network reconstruction: **Models databases**

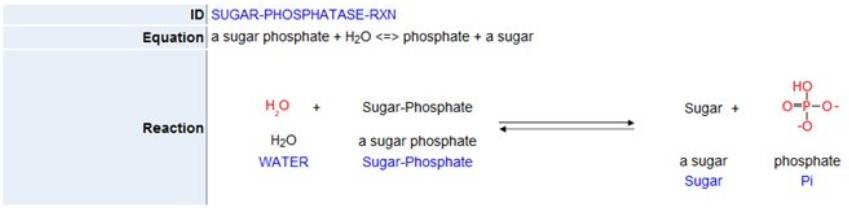


Figure 4: For instance, the reaction with EC#3.1.3.23 is not fully defined in MetaCyc: the main substract/product is labeled as Sugar-phosphate. Integrating this reaction needs to resolve which sugar-phosphate is the enzyme substrate.

ing the description of metabolic functions with the UniPathway resource [30] (see Table 2).

Thanks to such initiatives, the process of translating gene functions into metabolic reactions is likely to become much easier in the future. A new module in Microscope platform has been developed in order to directly link the MetaCyc reaction to gene without using the EC number as a key. This feature allows flagging the relationship between gene and reactions with the following evidence tags:

- "validated" : reaction has been manually linked to this gene by users,
- "annotated" : reaction has been linked to homologous gene and transferred here from a close genome,
- "predicted" : reaction has been linked to this gene by the pathway-tools algorithm.

Name	Link	Comment
Pathway Tools /Pathway Hole Filler	http://bioinformatics.ai.sri.com/ptools/	Automated reconstruction of metabolic pathways from genome annotation and MetaCyc pathway database. Pathway Hole Filler: identification of candidate genes for orphan metabolic activities
OptFlux	http://www.optflux.org	An open-source software platform for in silico metabolic engineering
YanaSquare	http://yana.bioapps.biozentrum.uni-wuerzburg.de	Software helping in metabolic model reconstruction, linked to KEGG data
GEMSiR	http://sb.nhri.org.tw/GEMSiR	A software platform for genome-scale Metabolic models Simulation, Reconstruction and Visualization.
Cobra tool box	http://opencobra.sourceforge.net	Quantitative prediction of cellular metabolism with constraint-based models
SurreyFBA	http://sysbio3.fhms.surrey.ac.uk/SurreyFBA.zip	A command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks

Table 4: Resources and software tools for microbial metabolic network reconstruction: **Software tools for metabolic network reconstruction**

A critical point in metabolic model reconstruction is the ability to distinguish isozymes from enzymatic complexes because the impact of a gene deletion will not have the same consequences. If the deleted gene belongs to a complex, the enzymatic activity will be lost, whereas in the case of isozymes, the enzymatic activity will remain present in the cell due to the existence of a "redundant" enzyme. Protein complexes could be built manually from literature or inferred by homology. One strategy is to use a pivot organism: for each protein complex experimentally demonstrated in the pivot organism, a similar complex is inferred in the studied organism if, for each subunit of the pivot complex, a

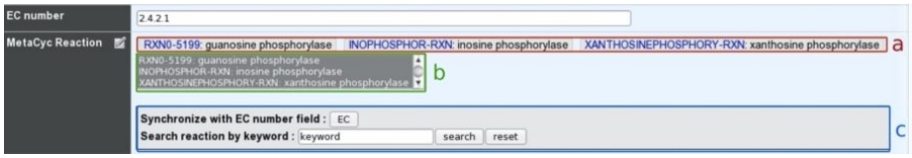


Figure 5: This field allows user to link one or more metabolic reactions from MetaCyc (BioCyc) to the current edited gene. A - Reactions presented at the top of the field have been manually curated by an annotator. B - A multiple selection list gives quick access to all predicted (unselected) or curated (selected) reactions linked to this gene. C - A search box allows one to quickly access MetaCyc reactions corresponding to either EC numbers from previous EC number field or a given keyword.

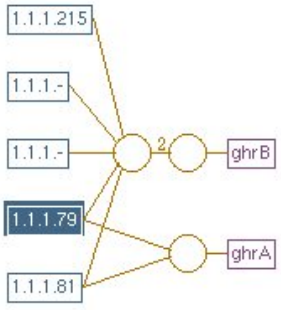


Figure 6: Protein complexes and isozymes representation in EcoCyc. At the right end, the genes (purple boxes) are translated into polypeptide (first orange circle). These polypeptides are able to make protein complexes that can aggregate to form the final catalyst.

candidate polypeptide can be detected by homology (e.g. Blast alignments with */bidirectional/* best hits) [47]. Finally, transport reactions are key reactions to be integrated to the metabolic network. They allow defining the interface between the cell and the environment. TransportDB provides tools to annotate a genome by finding the transporters.

4 Adding organism-specific metabolic pathways

Using an organism-specific metabolic pathways resource is a key point in delivering a well annotated network. Some databases enable to annotate the existence of each pathway: this is the first step to define an organism-specific metabolic pathways resource. Microscope platform provides the Pathway Curation tool which presents a list of predicted MicroCyc pathways in a given organism, coming from Pathway-tools software results, for which status can be curated by the annotator. Each pathway could have the following evidence status:

- predicted: Predicted by the BioCyc pathologic algorithm (default one),

- validated: Curated as a functional pathway (all the reactions of the pathway are supposed to exist in the organism),
- variant needed: The predicted pathway is not completely correct for the organism (i.e. some reactions may not be present in the organism but no better pathway definition exists in MetaCyc). Thus, a new pathway variant definition is needed,
- unknown: Not enough evidence to declare the pathway as functional (i.e. validated status),
- non-functionnal: The pathway has been lost in the organism and is no more functional (i.e. due to gene loss or pseudogenisation events),
- deleted: Curated as a false positive prediction.

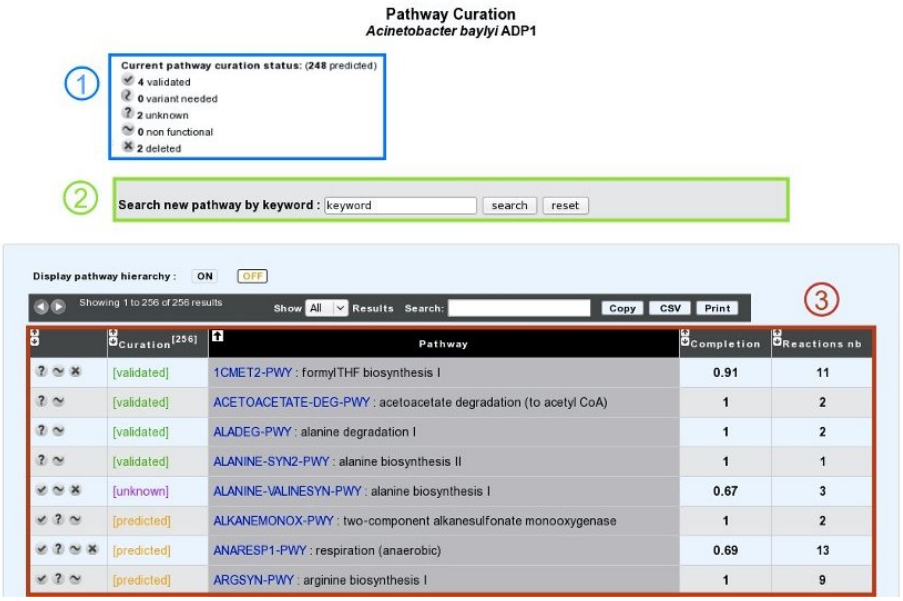


Figure 7: Main overview for Pathway Curation tool in Microscope.

Nevertheless, not all information on an organism's metabolic network can be found in its genome annotation. Previous experimental studies of its metabolism may well have identified metabolic pathways for which the corresponding genes remain unknown (e.g. an alanine biosynthesis pathway in *E. coli*, see **`http://biocyc.org/ECOLI/NEW-IMAGE?object=PWY0-41`**) or whose functions were not correctly propagated into their annotations. Also, not all metabolic pathways are represented in global metabolic databases, which are being built by reviewing the past literature, a still ongoing process.

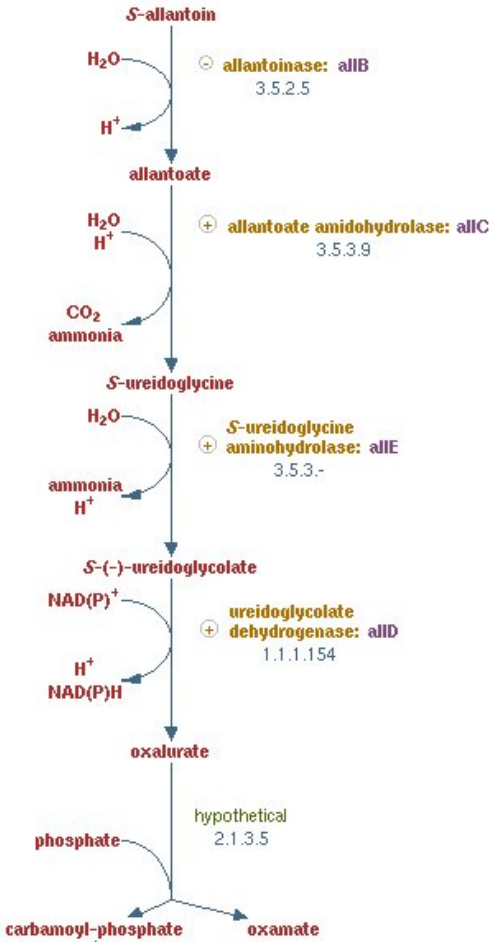


Figure 8: In EcoCyc, the allantoin degradation IV (anaerobic) presents a reaction (EC# 2.1.3.5) that do not have any associated enzyme protein.

For these reasons, the draft metabolic networks generated from genome annotation need to be curated and completed with the maximum amount of additional organism-specific biochemical information. Such information is mainly found in the literature, but also in some specialized biochemical databases. For instance, BRENDA extracts information on enzymes directly from the literature [23] and UM-BBD stores useful information on numerous microbial biocatalytic activities and biodegradation pathways [12] (see Table 2).

When including new reactions into the draft network, one must remain consistent with the reaction metabolite identifiers. Given the size of the metabolic networks, this technical issue is critical as designating a given metabolite

with distinct identifiers would prevent from finding connections between reactions that share the metabolite, and therefore would hinder the identification of pathways. Similarly, mathematical models of metabolism crucially depend on a consistent naming of metabolites. This issue is better solved by systematically referring to a unique database of metabolites, be it KEGG, MetaCyc, ChEBI or PubChem (see Table 2). These databases usually cross-reference their entries.

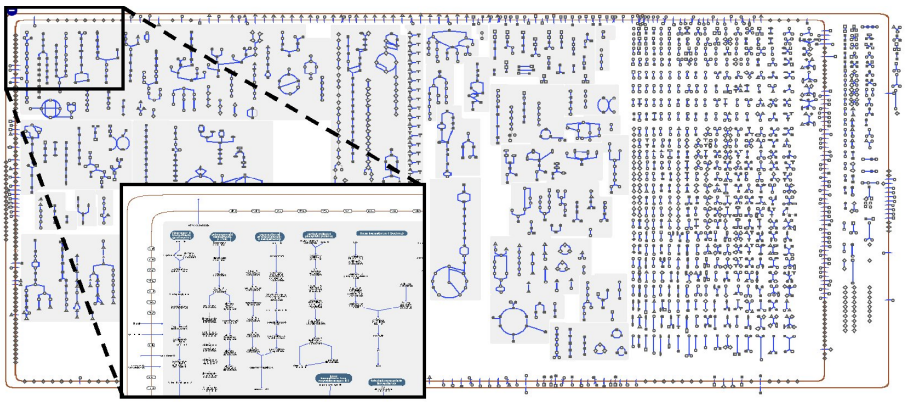


Figure 9: Interactive overview of Escherichia coli K12 metabolic network. Screenshots from EcoCyc [19] (**`http://ecocyc.org/overviewsWeb/celOv.shtml`**)

As regards software tools enabling curation of genome-scale metabolic networks, very few of them are currently available. Again, Pathway Tools provides an interface to curate, extend and visualize a locally reconstructed metabolic network [12] (see Figure 1). YanaSquare imports metabolic networks automatically generated in KEGG and allows curating it using a spread-sheet-like interface [38]. OptFlux [35] allows importing metabolic models in System Biology Markup Langage format [19]. General metabolic databases such as KEGG or SEED provide tools to visualize their reconstructed network, but do not allow curation.

Initiatives are being taken to develop further network curation tools (see for instance the EU-funded project Microme, **`http://www.microme.eu`**) and more convenient solutions will be hopefully available in a near future.

5 Identifying and filling gaps in metabolic networks

Once a global metabolic network is obtained, its completeness can be assessed by examining it globally or at the pathway level. Incomplete metabolic pathways, reactions disconnected from the rest of the network, or metabolites that

are only produced and never consumed (so-called dead-end metabolites) are actually as many clues that the reconstructed network include gaps that need to be filled for the network to represent a consistent functioning set of metabolic pathways.
A few methods have been designed to identify and fill such metabolic gaps. Pathway Tools makes use of predefined metabolic pathways: each pathway that is almost complete in the metabolic network is considered to be entirely present. The few missing reactions are therefore added to the network, thereby filling the gaps in the pathway [21]. Other more elaborate methods make use of graph- based or constraint-based mathematical representation of the network (see ref. [11] for a more thorough review). The GapFill method for instance attempts to fill metabolic gaps by adding minimal sets of reactions from a reference database that remove dead-end metabolites [36].

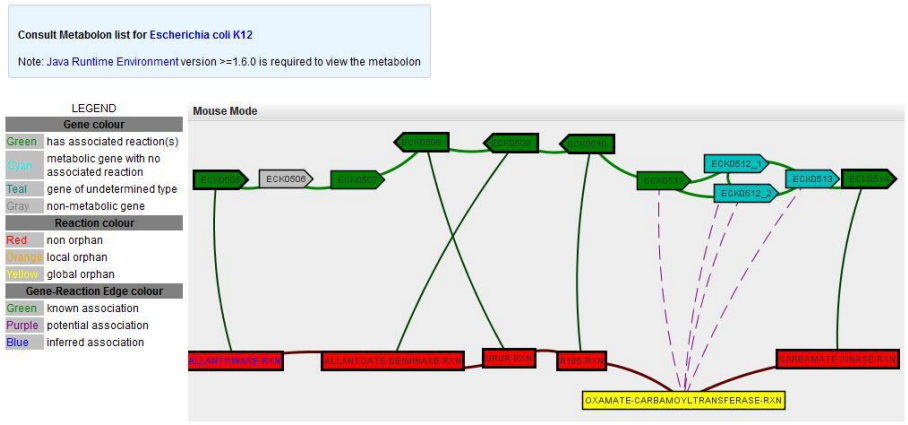


Figure 10: Screenshot of the CanOE Web interface. This metabolon is composed of 6 reactions (red and yellow) covering the complete pathway for the anaerobic degradation of allantoin, in which one reaction is orphans in *E. coli* (yellow): oxamate carbamoyltransferase (OXTCase). The missing activity in *E. coli* K-12 (the OXTCase) has yet to be associated to any genes in any organism and is thus a global orphan activity, despite its presence having been biochemically demonstrated in Streptococcus allantoicus, and even reported in *E. coli*. The CanOE metabolon bearing this reaction contained 5 gap genes (ECK0506-507 and ECK0511 to 0513 gray, blue and green) that could serve as candidate genes. The link between a reaction and a gene could be experimentally demonstrasted (green) or proposed by CanOE in purple.

Reactions that are inferred by pathway projection process constitute orphan activities, for which no related gene is known [26]. Additional methods have therefore been introduced to identify genes for orphan activities. While bioinformatic tools may help in identifying suitable candidate genes (e.g. Path-

way Hole Filler [17], implemented in Pathway Tools), promising experimental setups are being developed to systematically screen gene enzymatic functions [11, 7]. CanOE (Candidate genes for Orphan Enzymes) is a four-step bioinformatics strategy that proposes ranked candidate genes for sequence-orphan enzymatic activities (or orphan enzymes for short) [40]. The first step locates "genomic metabolons", i.e. groups of co-localized genes coding proteins catalyzing reactions linked by shared metabolites, in one genome at a time. These metabolons can be particularly helpful for aiding bioanalysts visualize relevant metabolic data. In the second step, they are used to generate candidate associations between un-annotated genes and gene-less reactions. The third step integrates these gene- reaction associations over several genomes using gene families, and summarizes the strength of family-reaction associations by several scores. In the final step, these scores are used to rank members of gene families which are proposed for metabolic reactions. These associations are of particular interest when the metabolic reaction is a sequence-orphan enzymatic activity. The CanOE strategy has been integrated in the MicroScope platform and is available at this URL:
http://www.genoscope.cns.fr/agc/microscope/metabolism/canoe.php.

6 Toward mathematic models of genome-scale metabolism

The end result of the network reconstruction process should be a comprehensive, consistent, and organized listing of metabolic reactions. It should provide an accurate overview of the organism's metabolism, especially when linked to convenient visualization software.

Reasoning on the metabolic network, however, requires going a step further as its size and complexity prevent from performing simple predictions from it. Several mathematical modeling frameworks have therefore been proposed to formalize metabolic networks, perform systems-level predictions and integrate various types of experimental data. Reviewing these frameworks would be out of the scope of this short article, and has already been extensively covered elsewhere [20, 41].

Deriving a metabolic model from a reconstructed network usually necessitates collecting additional data on reactions and metabolites and further checking their consistency. For instance, constraint-based models, which are widely used for genome-scale modeling, critically rely on stoichiometric coefficients and reaction reversibilities. Since subsequent model predictions directly depend on these data as well as on the correctness of the metabolic network, specific care should be taken when creating models from metabolic networks. Here again, the model reconstruction issue has been largely covered in a few reviews [11, 13].

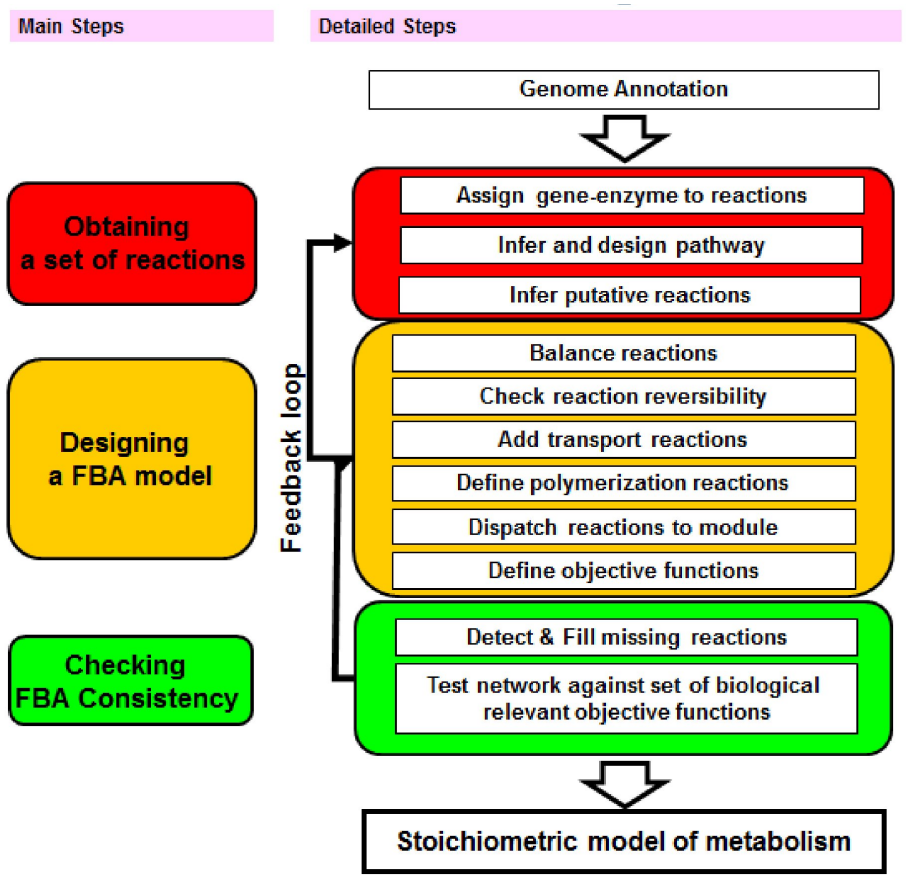


Figure 11: Standard protocol to design network and model from genome annotation.

7 Practical application: Functional re-annotation of Bacillus subtilis 168 genome

The complete strategy presented in the previous sections for the reconstruction of global metabolic networks from annotated genome sequences has been applied in the context of Microme project to the functional re-annotation of the genome of *Bacillus subtilis* 168 [6]. *B. subtilis* is considered a model organism for Gram-positive bacteria, forming part of the Firmicutes, a major phylum of the bacterial domain of life. Its genome was one of the first bacterial genomes that were completely sequenced in 1997 by an international consortium comprised by some 30 groups that sequenced and annotated different chromosomal segments covering the whole *B. subtilis* genome (4.2

megabases and 4100 protein coding genes in this original annotation) [22]. This initial annotation has been continuously updated in-depth, and in 2009, the genome was completely re-sequenced and re-annotated at the GenoScope, increasing the number of protein coding genes to 4244 CDSs, 48 percent of them with identified functions [5]. In parallel, at present 3 different genome-scale metabolic models of *B. subtilis* 168 has been published by different groups, based on the two most common pipelines for model reconstruction, the Biochemical Genetic and Genomic (BIGG) knowledbase [33] and the model SEED pipeline based on subsystem annotations stored in SEED database [18].

The different genome annotation projects does not considered explicitly gene annotation at the reaction level, meaning the specific associations of genes coding for enzymes to the biochemical reactions catalyzed by these enzymes, a basic step in the reconstruction of genome-scale metabolic networks and a starting point for the mathematical modeling of metabolic systems. This has been the primary goal of the last re-annotation of *B. subtilis* genome carried out in the context of the European FP7 project Microme (http://www.microme.eu/). Starting from the 2009 annotation, annotations from publications appeared after the 2009 sequence release as well as annotation of novel genomic objects like small untranslated RNAs, riboswitches and small CDSs has been integrated into MicroScope platform, putting specific emphasis on metabolic discoveries supported by experimental evidence. For this purpose, the metabolic network was generated by combining the automatic and manually curated annotations of *B. subtilis* stored in MicroScope by using Pathway Tools V16.0, the BioCyc pathway reconstruction software [24], and automatic pathway projections and gene-reaction associations were subsequently manually curated in the MicroScope environment by using the different tools and curation interfaces described in the previous sections. This curation step is essential in order to avoid over-predictions of the algorithm (multiple reactions associated to a particular CDSs) consequence of general or unclear annotations or unspecific EC number associations. In addition, automatic projections frequently render incomplete metabolic pathways with holes at different levels that can be consequence of false positive projections of the algorithm or can reflect missing enzymatic activities in current genome annotation (orphan enzymatic activities).

Manual curation at this level implies the search of candidate genes for these missing enzymatic activities together with the evaluation of the functionality of the corresponding pathways by using the pathway curation interface of MicroScope platform.

Finally, manual curation also allows the integration of novel metabolic knowledge in terms of enzymatic reactions and metabolic pathways that is currently absent in MetaCyc pathway and reaction repository. For this purpose,

new chemical compounds and reactions were created de-novo in the reference repositories CHEBI [28] and RHEA [1] respectively, these new reactions were validated associated to the corresponding CDSs by using the new Gene Editor interface of MicroScope, and new metabolic pathways were created with Pathway Tools and stored into an internal pathway database named MicroRefCyc. These new metabolic pathways will be subsequently projected not only on *B. subtilis* 168, but also over the whole set of genomes available in MicroScope together with metabolic pathways of the MetaCyc repository [8], allowing to extend the metabolic knowledge stored in the platform.

As result of this last-reannotation, the genome of *B. subtilis* contains 4458 genomic objects (4422 in the re-annotation of 2009), with 1083 CDSs that have been associated to 1097 chemical reactions representing a total number of 1751 gene-reaction associations.

These associations includes 45 chemical reactions that are absents in MetaCyc repository and that has been created de-novo in RHEA database, representing new metabolic knowledge product of the curation process.

In terms of metabolic pathways, 164 of the 283 MetaCyc pathways that were incomplete after initial projections were manually curated in order to fill pathway holes and define a functional status according to the pathway curation interface of MicroScope. As result of this curation process, 24 pathway holes were resolved (a candidate gene could be found by using different tools available at MicroScope platform), resulting in 22 initially incomplete pathways with validated status in MicroScope (functional pathways according with experimental evidence). However, in 14 of these validated pathways there is still remaining pathway holes for which no candidate gene can be found (12 pathway holes), representing orphan enzymatic activities that has been experimentally detected in *B. subtilis* but for which no candidate genes has been yet identified.

Finally, as result of manual curation process, 27 new metabolic pathways and pathway variants has been characterized and represented in the MicroRefCyc pathway repository. Among these new pathways, we can find a novel biotin biosynthesis pathway variant that allows to connect fatty acid metabolism with biotin biosynthesis through the oxidative cleavage of long-chain acyl-ACP and fatty acids to biotin precursors pimeloyl-ACP and pimelate respectively [16, 42], as well as other pathways related with polyketide biosynthesis [31], methionine salvage [39] or purine degradation [34]. These new metabolic pathways and pathway variants represent novel metabolic knowledge that can be propagated to other bacterial genomes through the MicroScope pipeline for automated metabolic network reconstruction.

This new metabolic network of *B. subtilis* 168 will be used as starting point for the reconstruction of an updated genome-scale stoichiometric model of *B. subtilis* metabolism that will allow understanding how these new pathways and reactions are integrated in the whole metabolic system of the cell. In this sense, a comparison of the gene content of the MicroScope metabolic network of *B. subtilis* 168 with the last published version of the metabolic model of *B. subtilis* 168 iBSU1103V2 [44] reveals that, although we are missing a substantial part of the model associated to transmembrane transport reactions, we are able to increase the coverage of the metabolic model by adding a significant number of CDSs associated to chemical reactions that are currently absent in iBSU1103V2 model (358 CDSs associated to chemical reactions in MicroScope, 100 of them linked to metabolic pathways).

References

[1] Alcántara, Rafael, Kristian B. Axelsen, Anne Morgat, Eugeni Belda, Elisabeth Coudert, Alan Bridge, Hong Cao, *et al.* 2012. "Rhea–a manually curated resource of biochemical reactions". *Nucleic Acids Research* 40 (Database issue): D754-760. doi:10.1093/nar/gkr1126.

[2] Arakawa, Kazuharu, Yohei Yamada, Kosaku Shinoda, Yoichi Nakayama, and Masaru Tomita. 2006. "GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes". *BMC Bioinformatics* 7: 168. doi:10.1186/1471-2105-7-168.

[3] Aziz, Ramy K, Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, *et al.* 2008. "The RAST Server: rapid annotations using subsystems technology". *BMC Genomics* 9: 75. doi:10.1186/1471-2164-9-75.

[4] Bairoch, Amos. 2000. "The ENZYME database in 2000". *Nucleic Acids Research* 28 (1): 304-305.

[5] Barbe, Valérie, Stéphane Cruveiller, Frank Kunst, Patricia Lenoble, Guillaume Meurice, Agnieszka Sekowska, David Vallenet, *et al.* 2009. "From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later". *Microbiology* 155 (Pt 6) (juin): 1758-1775. doi:10.1099/mic.0.027839-0.

[6] Belda, Eugeni, Agnieszka Sekowska, François Le Fèvre, Damien Mornico, Anne Morgat, Christos Ouzounis, David Vallenet, Claudine Médigue, Antoine Danchin. 2013. "An updated metabolic view of the Bacillus subtilis 168 genome". *Microbiology* (in-press)

[7] Beloqui, Ana, Mara-Eugenia Guazzaroni, Florencio Pazos, José M. Vieites, Marta Godoy, Olga V. Golyshina, Tatyana N. Chernikova, *et al.* 2009. "Reactome array: forging a link between metabolome and genome". *Science (New York, N.Y.)* 326 (5950): 252-257. doi:10.1126/science.1174094.

[8] Caspi, Ron, Tomer Altman, Kate Dreher, Carol A. Fulcher, Pallavi Subhraveti, Ingrid M. Keseler, Anamika Kothari, *et al.* 2012. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases". *Nucleic Acids Research* 40 (Database issue): D742-753. doi:10.1093/nar/gkr1014.

[9] Curran, Kathleen A, Nathan C. Crook, and Hal S. Alper. 2012. "Using flux balance analysis to guide microbial metabolic engineering". *Methods in Molecular Biology (Clifton, N.J.)* 834: 197-216. doi:10.1007/978-1-61779-483-4_13.

[10] Disz, Terry, Sajia Akhter, Daniel Cuevas, Robert Olson, Ross Overbeek, Veronika Vonstein, Rick Stevens, and Robert A. Edwards. 2010. "Accessing the SEED genome databases via Web services API: tools for programmers". *BMC Bioinformatics* 11: 319. doi:10.1186/1471-2105-11-319.

[11] Durot, Maxime, Pierre-Yves Bourguignon, and Vincent Schachter. 2009. "Genome-scale models of bacterial metabolism: reconstruction and applications". *FEMS Microbiology Reviews* 33 (1): 164-190. doi:10.1111/j.1574-6976.2008.00146.x.

[12] Ellis, Lynda B. M, Dave Roe, and Lawrence P. Wackett. 2006. "The University of Minnesota Biocatalysis/Biodegradation Database: the first decade". *Nucleic Acids Research* 34 (Database issue): D517-521. doi:10.1093/nar/gkj076.

[13] Feist, Adam M, Markus J. Herrgråd, Ines Thiele, Jennie L. Reed, and Bernhard Ø Palsson. 2009. "Reconstruction of biochemical networks in microorganisms". *Nature Reviews. Microbiology* 7 (2): 129-143. doi:10.1038/nrmicro1949.

[14] Feist, Adam M, and Bernhard Ø Palsson. 2008. "The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli". *Nature Biotechnology* 26 (6): 659-667. doi:10.1038/nbt1401.

[15] Gevorgyan, Albert, Michael E. Bushell, Claudio Avignone-Rossa, and Andrzej M. Kierzek. 2011. "SurreyFBA: A. Command Line Tool and

Graphics User Interface for Constraint-Based Modeling of Genome-Scale Metabolic Reaction Networks". *Bioinformatics* 27 (3): 433-434. doi:10.1093/bioinformatics/btq679.

[16] Green, A. J, S. L. Rivers, M. Cheeseman, G. A. Reid, L. G. Quaroni, I. D. Macdonald, S. K. Chapman, and A. W. Munro. 2001. "Expression, purification and characterization of cytochrome P450 Biol: a novel P450 involved in biotin synthesis in Bacillus subtilis". *Journal of biological inorganic chemistry* 6 (5-6) (juin): 523-533.

[17] Green, Michelle L, and Peter D. Karp. 2004. "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases". *BMC Bioinformatics* 5: 76. doi:10.1186/1471-2105-5-76.

[18] Henry, Christopher S, Jenifer F. Zinner, Matthew P. Cohoon, and Rick L. Stevens. 2009. "iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations". *Genome biology* 10 (6): R69. doi:10.1186/gb-2009-10-6-r69.

[19] Hucka, M, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, *et al.* 2003. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". *Bioinformatics (Oxford, England)* 19 (4): 524-531.

[20] Joyce, Andrew R, and Bernhard Ø Palsson. 2006a. "The model organism as a system: integrating "omics" data sets." *Nat Rev Mol Cell Biol* 7 (3): 198-210. doi:10.1038/nrm1857.

[21] Karp, Peter D, Suzanne M. Paley, Markus Krummenacker, Mario Latendresse, Joseph M. Dale, Thomas J. Lee, Pallavi Kaipa, *et al.* 2010. "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology". *Briefings in Bioinformatics* 11 (1): 40-79. doi:10.1093/bib/bbp043.

[22] Kunst, F, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, *et al.* 1997. "The complete genome sequence of the gram-positive bacterium Bacillus subtilis". *Nature* 390 (6657) (novembre 20): 249-256. doi:10.1038/36786.

[23] Lang, Maren, Michael Stelzer, and Dietmar Schomburg. 2011. "BKM-react, an integrated biochemical reaction database". *BMC Biochemistry* 12: 42. doi:10.1186/1471-2091-12-42.

[24] Latendresse, Mario, Suzanne Paley, and Peter D. Karp. 2012. "Browsing metabolic and regulatory networks with BioCyc". *Methods in molecular biology* 804: 197-216. doi:10.1007/978-1-61779-361-5_11.

[25] Le Fèvre, F, S. Smidtas, C. Combe, M. Durot, F. d'Alché-Buc, and V. Schachter. 2009. "CycSim–an online tool for exploring and experimenting with genome-scale metabolic models". *Bioinformatics (Oxford, England)* 25 (15): 1987-1988. doi:10.1093/bioinformatics/btp268.

[26] Lespinet, Olivier, and Bernard Labedan. 2005. "Orphan enzymes?" *Science (New York, N.Y.)* 307 (5706): 42. doi:10.1126/science.307.5706.42a.

[27] Markowitz, Victor M, Konstantinos Mavromatis, Natalia N. Ivanova, I-Min A. Chen, Ken Chu, and Nikos C. Kyrpides. 2009. "IMG ER: a system for microbial genome annotation expert review and curation". *Bioinformatics (Oxford, England)* 25 (17): 2271-2278. doi:10.1093/bioinformatics/btp393.

[28] de Matos, Paula, Nico Adams, Janna Hastings, Pablo Moreno, and Christoph Steinbeck. 2012. "A database for chemical proteomics: ChEBI". *Methods in Molecular Biology (Clifton, N.J.)* 803: 273-296. doi:10.1007/978-1-61779-364-6_19.

[29] Médigue, Claudine, and Ivan Moszer. 2007. "Annotation, comparison and databases for hundreds of bacterial genomes". *Research in Microbiology* 158 (10): 724-736. doi:10.1016/j.resmic.2007.09.009.

[30] Morgat, Anne, Eric Coissac, Elisabeth Coudert, Kristian B. Axelsen, Guillaume Keller, Amos Bairoch, Alan Bridge, Lydie Bougueleret, Ioannis Xenarios, and Alain Viari. 2012. "UniPathway: a resource for the exploration and annotation of metabolic pathways". *Nucleic Acids Research* 40 (D1): D761-D769. doi:10.1093/nar/gkr1023.

[31] Nakano, Chiaki, Hiroki Ozawa, Genki Akanuma, Nobutaka Funa, and Sueharu Horinouchi. 2009. "Biosynthesis of aliphatic polyketides by type III polyketide synthase and methyltransferase in Bacillus subtilis". *Journal of bacteriology* 191 (15) (août): 4916-4923. doi:10.1128/JB.00407-09.

[32] Notebaart, Richard A, Frank H. J. van Enckevort, Christof Francke, Roland J. Siezen, and Bas Teusink. 2006. "Accelerating the reconstruction of genome-scale metabolic networks". *BMC Bioinformatics* 7: 296. doi:10.1186/1471-2105-7-296.

[33] Oh, You-Kwan, Bernhard O. Palsson, Sung M. Park, Christophe H. Schilling, and Radhakrishnan Mahadevan. 2007. "Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data". *The Journal of biological chemistry* 282 (39) (septembre 28): 28791-28799. doi:10.1074/jbc.M703759200.

[34] Ramazzina, Ileana, Roberto Costa, Laura Cendron, Rodolfo Berni, Alessio Peracchi, Giuseppe Zanotti, and Riccardo Percudani. 2010. "An aminotransferase branch point connects purine catabolism to amino acid recycling". *Nature chemical biology* 6 (11) (novembre): 801-806. doi:10.1038/nchembio.445.

[35] Rocha, Isabel, Paulo Maia, Pedro Evangelista, Paulo Vilaa, Simo Soares, José P. Pinto, Jens Nielsen, Kiran R. Patil, Eugnio C. Ferreira, and Miguel Rocha. 2010. "OptFlux: an open-source software platform for in silico metabolic engineering". *BMC Systems Biology* 4: 45. doi:10.1186/1752-0509-4-45.

[36] Satish Kumar, Vinay, Madhukar S. Dasika, and Costas D. Maranas. 2007. "Optimization based automated curation of metabolic reconstructions". *BMC Bioinformatics* 8: 212. doi:10.1186/1471-2105-8-212.

[37] Schellenberger, Jan, Richard Que, Ronan M. T. Fleming, Ines Thiele, Jeffrey D. Orth, Adam M. Feist, Daniel C. Zielinski, *et al.* 2011. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0". *Nature Protocols* 6 (9):1290-1307. doi:10.1038/nprot.2011.308.

[38] Schwarz, Roland, Chunguang Liang, Christoph Kaleta, Mark Khnel, Eik Hoffmann, Sergei Kuznetsov, Michael Hecker, Gareth Griffiths, Stefan Schuster, and Thomas Dandekar. 2007. "Integrated network reconstruction, visualization and analysis using YANAsquare". *BMC Bioinformatics* 8: 313. doi:10.1186/1471-2105-8-313.

[39] Sekowska, Agnieszka, Valérie Dénervaud, Hiroki Ashida, Karine Michoud, Dieter Haas, Akiho Yokota, and Antoine Danchin. 2004. "Bacterial variations on the methionine salvage pathway". *BMC microbiology* 4 (mars 4): 9. doi:10.1186/1471-2180-4-9.

[40] Smith, Alexander, Claudine Médigue, Eugeni Belda, Alain Viari, and David Vallenet. 2012. "The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes". *PLoS Computational Biology*.

[41] Stelling, Jörg. 2004. "Mathematical models in microbial systems biology". *Current Opinion in Microbiology* 7 (5): 513-518. doi:10.1016/j.mib.2004.08.004.

[42] Stok, J. E, and J. De Voss. 2000. "Expression, purification, and characterization of BioI: a carbon-carbon bond cleaving cytochrome P450 involved

in biotin biosynthesis in Bacillus subtilis". *Archives of biochemistry and biophysics* 384 (2) (décembre 15): 351-360.

[43] Sun, Jibin, and An-Ping Zeng. 2004. "IdentiCS–identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence". *BMC Bioinformatics* 5: 112. doi:10.1186/1471-2105-5-112.

[44] Tanaka, Kosei, Christopher S. Henry, Jenifer F. Zinner, Edmond Jolivet, Matthew P. Cohoon, Fangfang Xia, Vladimir Bidnenko, S. Dusko Ehrlich, Rick L. Stevens, and Philippe Noirot. 2013. "Building the repertoire of dispensable chromosome regions in Bacillus subtilis entails major refinement of cognate large-scale metabolic model". *Nucleic acids research* 41 (1) (janvier 1): 687-699. doi:10.1093/nar/gks963.

[45] Thiele, Ines, and Bernhard Ø Palsson. 2010. "A protocol for generating a high-quality genome-scale metabolic reconstruction". *Nature Protocols* 5 (1): 93-121. doi:10.1038/nprot.2009.203.

[46] Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fevre F, Longin C, Mornico D, Roche D, Rouy Z, Salvignol G, Scarpelli C, Thil Smith AA, Weiman M, Médigue C., *et al.* 2012. "MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data.". *Database: Nucleic Acids Res* 2013: gks1194. doi:10.1093/nar/gks1194.

[47] Vieira, Gilles, Victor Sabarly, Pierre-Yves Bourguignon, Maxime Durot, Francois Le Fevre, Damien Mornico, David Vallenet, *et al.* 2011. "Core and panmetabolism in Escherichia coli". *Journal of Bacteriology* 193 (6): 1461-1472. doi:10.1128/JB.01192-10.

Metabolism Modelling

Jean-Pierre Mazat[1]

[1] IBGC CNRS UMR 5095 et Université Bordeaux 2

Glossary

- Metabolism: the metabolic network at work; often synonymous with metabolic network.
- Enzyme: protein with catalytic activity. Synthesized in the cell by other proteins and protein- RNA complexes.
- Enzymatic reaction: biochemical reaction catalysed by an enzyme.
- Metabolic network: The metabolic network or metabolism (which is more the metabolic network at work) is the set of all the reactions taking place in a cell.
- Metabolite: molecule synthesized, degraded or/and transformed in the cellular metabolism.

1 Introduction

1.1 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.

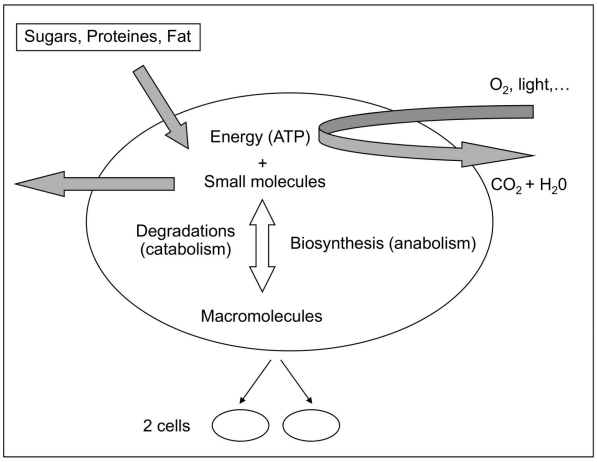


Figure 1: Cell metabolism, or the cell as an open system.

From these molecules other reactions take place to synthesize the basic molecules of the cell, such as amino-acids, nucleotides and nucleosides, lipids, etc. These molecules have a molecular weight (MW) between about twenty (18 for water) and about a thousand (NADH, lipids, etc.; see table 1). With these elementary molecules, the cell can synthesize several macromolecules (protein and nucleic acid biosynthesis for instance). 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).

All these synthesis and degradations constitute the *cell metabolism* or the *metabolic network* or the *metabolism*. In addition these reactions are regulated and the regulation network is presumably as or more complex than the cell metabolism itself.

In this chapter, for the sake of simplicity, we will limit ourselves to the part of metabolism which concerns the synthesis (anabolism) and the degradation (catabolism) of the small basic molecules of the cell (amino acids, aromatic bases of nucleic acids, sugars, palmitate, glycerol and choline etc.).
These molecules will be called *metabolites* in the following. We will not deal with the regulation of the metabolic network

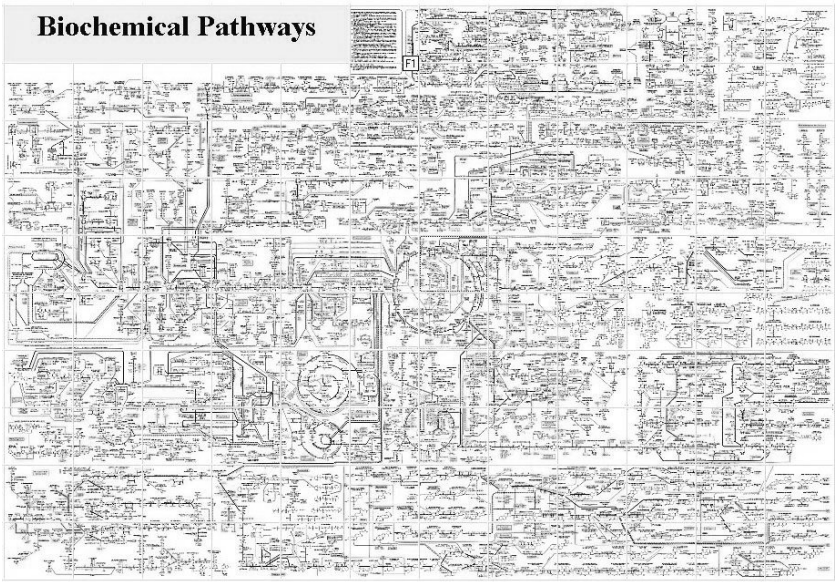


Figure 2: The metabolic network.

Even with all these restrictions, there remains a complex system (Fig. 2) with about 750 reaction and about 500 metabolites, slightly variable according to the cell type.

Computer scientists tend to derive a graph representation of the metabolic maps (Fig. 3). One has however to be aware that most of the metabolic reactions are bi-molecular, i.e. involve two substrates and two products, which complicate the graph representation because the two products of a reaction are usually not the two substrates of the following reaction.

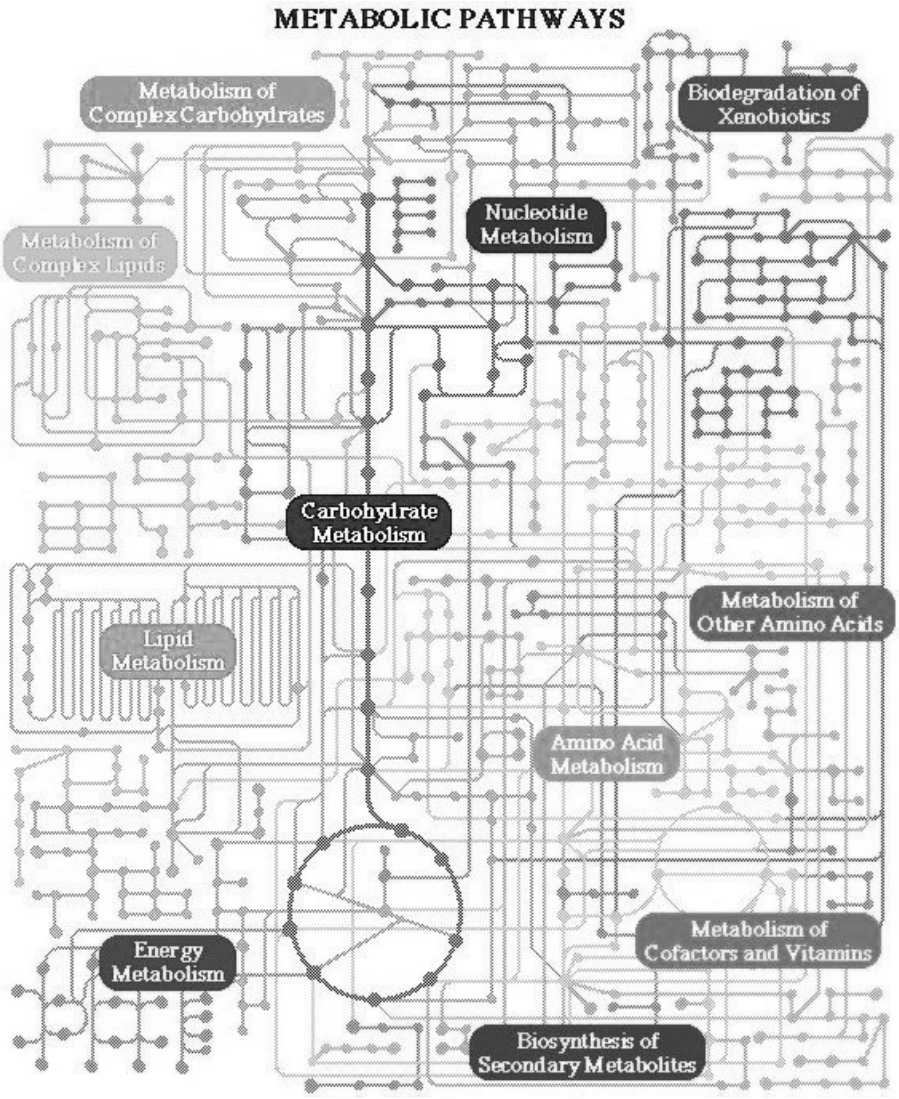


Figure 3: The structure of the metabolic network (from Kyoto Encyclopedia of Genes and Genomes). In this map, each dot represents an intermediate; each line represents an enzyme that acts on an intermediate

1.2 Why do cells use enzymes ?

Most of the cell reactions are either impossible in the living conditions or very slow. Thus they need to be catalyzed. The enzymes are these catalyzers. Enzymes are proteins, i.e. linear chain of amino acids which fold, adopting a special conformation able to favour a particular reaction.

The molecular weight of an enzyme is usually between 10000 and 1 million; it means the the MW of an enzyme is at least two order of magnitude higher than those of substrates and products (however, an enzyme can also act on other enzymes or more generally on macromolecules; we wil not deal with these cases here); the interactions of the substrates at the surface or in pockets of the enzyme put them in special conformations (analog to transition complex) and in this way favour the reaction.

It must be pointed out that the specificity of an enzyme is twofold. First there is a specificity of reaction : the enzyme E1 transforms the substrate S in P1 and another enzyme E2 transforms the substrate S in another product P2 etc. For instance pyruvate dehydrogenase (1.2.1.51) transforms pyruvate in acetyl-CoA, while lactate dehydrogenase (1.1.1.27) gives lactate from the same metabolite pyruvate. The second specificity is the one regarding substrates binding; for instance pyruvate dehydrogenase will bind only pyruvate; when metabolism needs to perform the same reaction on another substrate, α-ketoglutarate for instance, another enzyme has to be designed the α-ketoglutarate-dehydrogenase, specific of α-ketoglutarate (and not of pyruvate) and able to perform exactly the same chemical reaction.

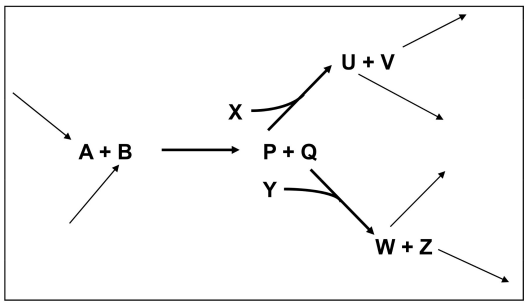


Figure 4: Entanglement of metabolic reactions

Let us nevertheless emphasise that specificities are not usually completely strict. Other metabolites, more or less chemically related to the "official" substrate are able to bind the same binding site on a given enzyme. This specificity notion (scale) is closely related to the problem of enzymes (genes) annotation : what is (are) the "official" substrate(s) of an enzyme? In other words which name will be given to a given enzyme.

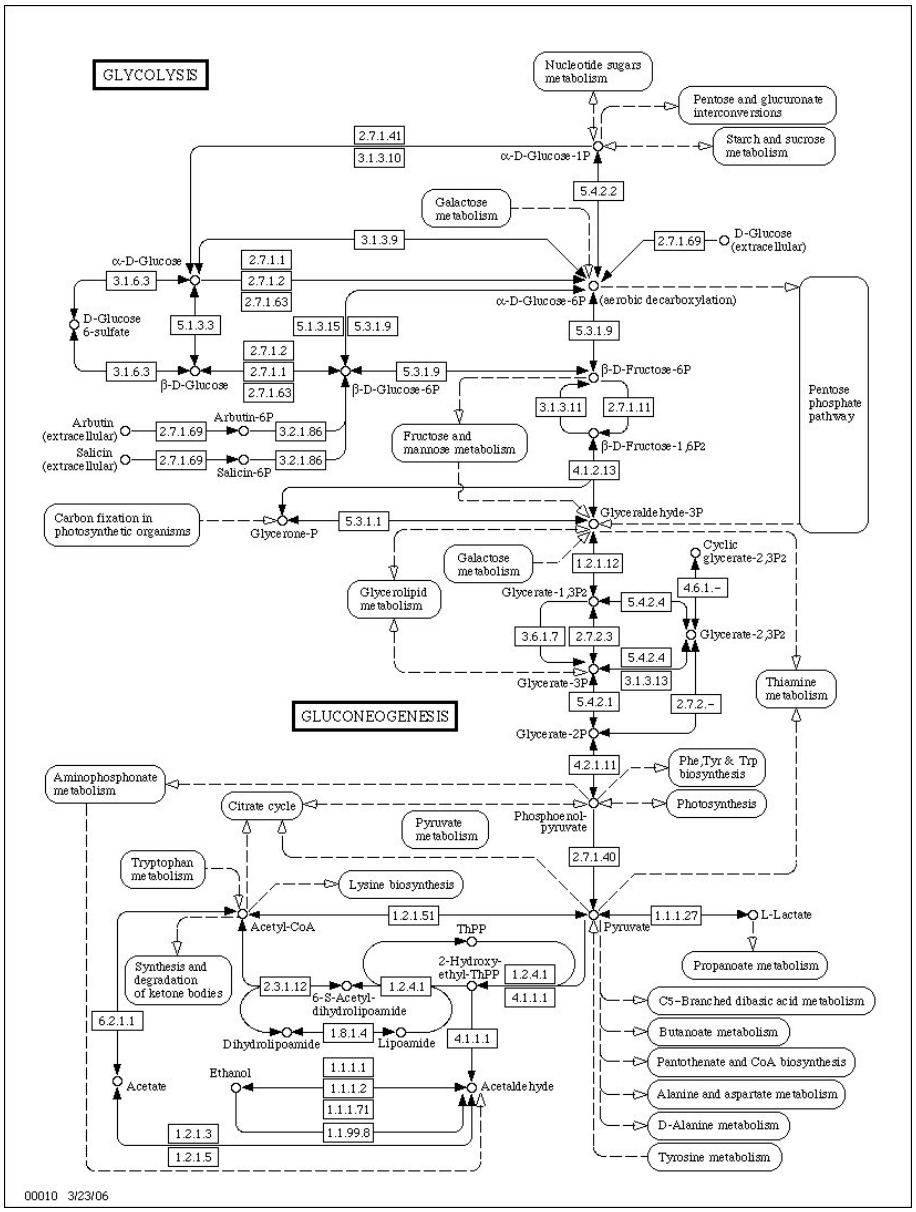


Figure 5: The pyruvate cross-roads (from KEGG)

Usually the substrate with the better affinity for an enzyme is the ”official” one, but it is not always the case due to historical reasons. Furthermore, some enzyme can have a wide specificity accommodating several related substrates (for instance hexokinase with hexoses, which are the family of sugars with six carbon atoms).

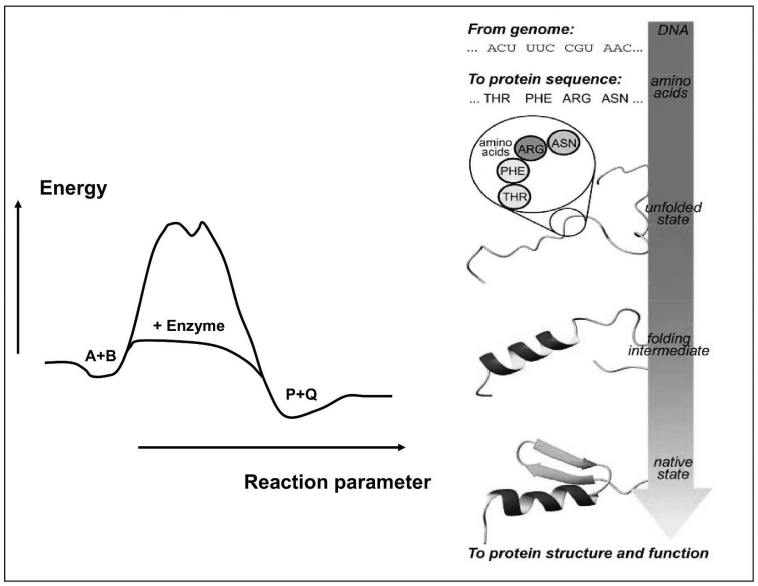


Figure 6: Enzyme folding

The last remark concerns the dimensions. On Fig. 7 are drawn, at the same scale, an average enzyme (diameter 5nm) and an average metabolite (0.5 nm). It is better understood how an enzyme can play its catalysing role in bending the substrate in an adequate, reacting, conformation. Table 1 gives also some dimension of biological objects.

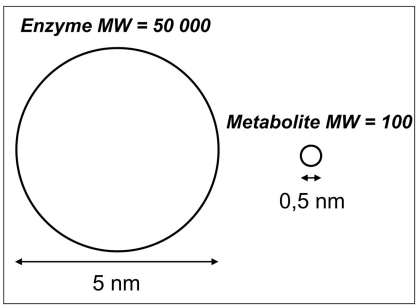


Figure 7: Average enzyme and average metabolite at the same scale.

Type	**Size** (nm)	**MW**	**Observations**
Water (H_2O)	0.26	18	
Alanine (Amino acid)	0.5	89	
Glucose	0.7	180	
ATH	1.6	507	MW of acid form
NADH		741	dipotassium salt
Phospholipid	3.5	750	
Myoglobin	3.6	16900	small protein
Hemoglobin	6.8	65000	medium protein
Cytochrom b		42592	part of complex of respiratory chain in mitochondria
Complex III	$7\times7\times11$	240000	
Ribosome (E. coli)	18	$2.8\ 10^6$	
Membrane	7-9		thickness
Lysosome	250-500		
Peroxysome	500		
Mitochodrion	1.5μ		
E. coli	2μ	2 pg	
Nucleus	4μ - 6μ		
Chloroplast	8μ	60 pg	spinach leaves
Hepatic cell	20μ	8μg	

Table 1: Sizes of molecules and cells.

2 Enzyme kinetics

2.1 Essential concepts

In metabolism modelling it is essential to know the *rates* with which enzymes catalyse the reactions. Biochemists often intermingle the reaction R_i catalyzed by the enzyme E_i with the rate V_i; it will be sometimes necessary in modelling to precise the variable actually under study. From analogy with the chemical catalysis, V. Henri (Henri, 1902 and 1903) proposed that the substrate bind firstly to the enzyme and then reacts to give the product; this scheme will be extended later by Briggs and Haldane in a fundamental paper of one and a half page long (Briggs and Haldane, 1925).

Modelling of this scheme in condition for which $P_0 = 0$ (the usual conditions of measurement which are different of the *in vivo* conditions) is written in the following dynamical system:

$$\begin{cases} \dfrac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] & (1) \\ v = \dfrac{d[P]}{dt} = k_2[ES] & (2) \end{cases}$$

with the conservation equations:

$$\begin{cases} [E] + [ES] = [E]_{total} & (3) \\ [S] + [ES] = [S]_{total} = [S]_0 & (4) \end{cases}$$

List of variables (4): [E], [ES], [S] and [P] with 2 conservation relationships $\rightarrow$ 2 independent variables [ES] and [P] for instance.
List of parameters (3): k_1, k_{-1} and k_2.
Initial conditions: $[E]_0 = [E]_{total}$, $[ES]_0 = 0$, $[S]_0 = [S]_{total}$, $[P] = 0$

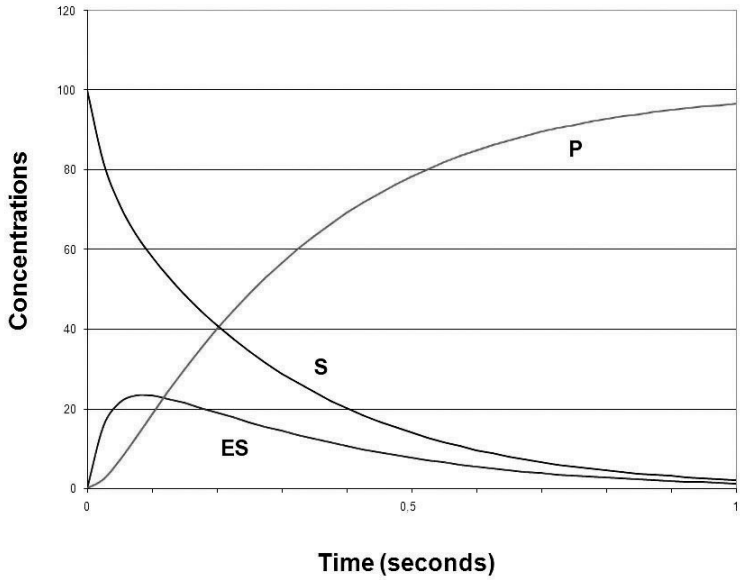


Figure 8: Time course of the different species of metabolites and enzymes in the simple enzymatic reaction: $E + S \leftrightarrow ES \rightarrow E + P$, where $E_{total} = 100, S_{total} = 100, k_1 = 1, k_{-1} = k_2 = 100$

At the beginning (Fig. 8), S binds E to form the complex [ES] which dissociates partially to form P. If the abscissa is time in seconds, one can see that the phenomenon is too fast to be easily observed. An experimental solution to this problem is to decrease the quantity of enzyme; because an enzyme is a

catalyst, its quantity can be in minute amount. One get the behaviour described in Fig. 9).

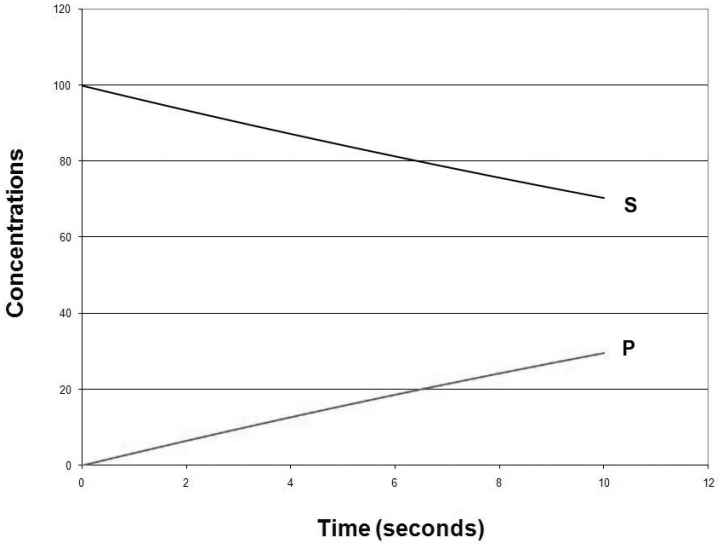


Figure 9: Time course of the different species of metabolites and enzymes in the simple enzymatic reaction: $E + S \leftrightarrow ES \rightarrow E + P$ with low enzyme concentration, where $E_{total} = 0.1, S_{total} = 100, k_1 = 1, k_{-1} = k_2 = 100$

The time course of P production is much longer and can be observed during several seconds. Furthermore the P production is quasi linear, i.e. is performed at a constant rate Fig. 9. This is due to the fact that the variation of [ES] is weak (between 0.033 and 0.026) and cannot be seen on the graph. In these conditions one can write $d[ES]/dt = 0$ which allows to derive the Henri-Michaelis-Menten equation proposed below in paragraph 1.

It should be pointed out also that in these conditions, the first reaction ($E + S \leftrightarrow ES$) is displaced from its thermodynamic equilibrium which gives $[ES]_{eq} = [E]_{eq} = 0.05$ ($K_{eq} = [E] \times [S]/[ES] = k_{-1}/k_1 = 100$). This is due to the high value of k_2 which is of the same order than k_{-1}.

In conclusion one can see that modelling with differential equations is easy. One can model metabolic networks with hundreds of kinetic equations and follow the time course of the metabolites concentrations towards a possible steady-state. However such a modelling assume that the cellular medium is homogeneous and that one can speak of concentrations, i.e. that the number of metabolites and enzyme molecules per volume unit is high enough. This is not always the case.

2.2 Henri-Michaelis equation

In the conditions of Fig. 9, i.e. $d[ES]/dt = 0$, one can solve the system of equations (1) to (4) shown in section 2.1, which becomes a system of algebraic equations to obtain:

$$V = \frac{d[P]}{dt} = \frac{V_M[S]_0}{K_m + [S]_0} \tag{5}$$

with:

$$K_m = \frac{k_{-1} + k_2}{k_1} \text{ and } V_M = k_2[E]_0$$

This rate equation due to V. Henri (Henri, 1902 and 1903) on the one hand and to Briggs and Haldane (Briggs and Haldane , 1925) on the other hand is usually known as the Michaelis-Menten equation.

The approximations which are underlying equation (5) are $[E]_0 \ll [S]_0$ and necessitate to be in a time range for which $d[ES]/dt$ is close to zero (quasi steady-state hypothesis). But Cornisch-Bowden & Hofmeyr (2005) properly show that this equation does not generally apply *in vivo*, because, in this case the product concentration is not zero (the product of a reaction is the substrate of the following reaction(s)). Furthermore, in general, a reaction deals with two substrates and two products:

$$A + B \leftrightarrow P + Q$$

One is thus conducted to write a differential system of the form:

$$\begin{cases} \dfrac{d[EA]}{dt} = k_A[E][A] - (k_{-A} + k_{AB})[ES] \\ \dfrac{d[EAB]}{dt} = k_B[EA][B] - (k_{-AB} + k_2)[EAB] \\ etc. \end{cases}$$

One can, according to the same sort of approximations as above, i.e. quasi steady-state of the enzymatic species E, EA, EB, EAB etc. , propose a rather general and suitable phenomenological equation (Chassagnole et al. 2001, Cornisch-Bowden & Hofmeyr, 2005):

$$v = \frac{V_{AB}\frac{[A][B]}{K_A K_B} - V_{BA}\frac{[P][Q]}{K_P K_q}}{\left[1 + \frac{[A]}{K_A} + \frac{[P]}{K_P}\right]\left[1 + \frac{[B]}{K_B} + \frac{[Q]}{K_Q}\right]} \tag{6}$$

or

$$v = \frac{\left[1 - \frac{[P][Q]}{K_{eq}[A][B]}\right]\left[\frac{V_{AB}}{K_A K_B}\right]}{\left[1 + \frac{[A]}{K_A} + \frac{[P]}{K_P}\right]\left[1 + \frac{[B]}{K_B} + \frac{[Q]}{K_Q}\right]} \tag{7}$$

K_A, K_B, K_P and K_Q, are the so called Michaelis constants for the metabolites A, B, P and Q, analog to the parameter K_m for the substrate S in equation (5). V_{AB} corresponds to the maximal rate of the reaction in the direction from A and B towards P and Q (P corresponds to A and Q to B in the reaction) and V_{BA} corresponds to the maximal rate in the reverse reaction. These parameters are linked through the Haldane relationship:

$$K_{eq} = \frac{[P]_{eq}[Q]_{eq}}{[A]_{eq}[B]_{eq}} = \frac{V_{AB}K_PK_Q}{V_{BA}K_AK_B}$$

which is introduced in eq (7).

In equation (7) the first term in brackets in the numerator expresses the distance of the concentrations [A], [B], [P] et [Q] to the equilibrium. The second term in brackets in the numerator expresses the efficacity of the enzyme due to its concentration and to the rate constants of products production (analog to the term $k_2[E]_0$ in the Michaelis-Menten equation).

3 Metabolic networks

3.1 Why is a metaoblic network more than the sum of its reactions?

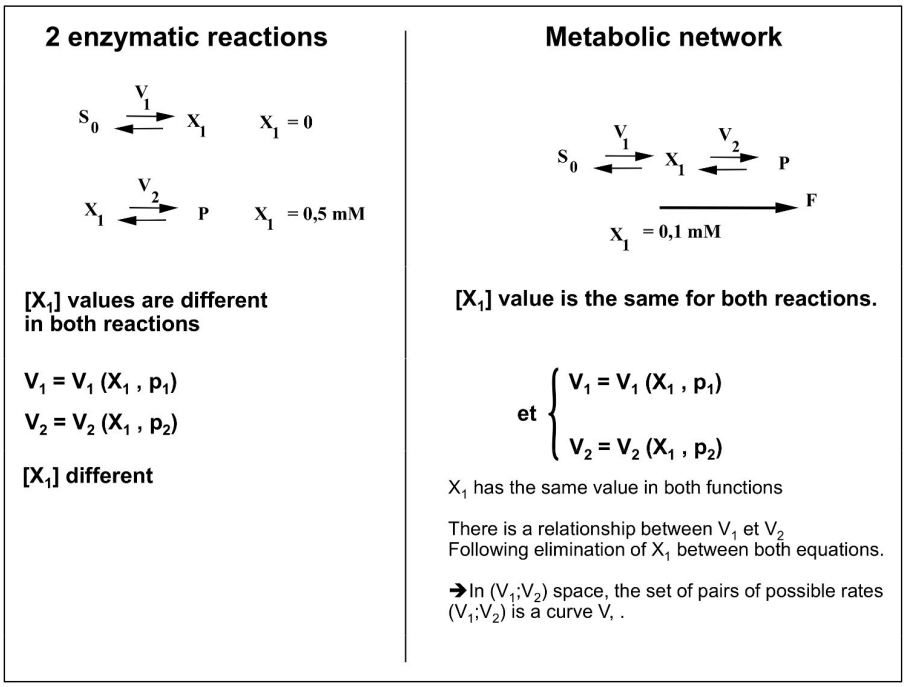


Figure 10: What is a metaoblic network?

The figure 10 gives the answer. It is also illustrated by the laboratory practice. When an enzyme is studied, one use substrate and product concentrations (product concentration equal to zero) which are not the one encountered *in vivo*. Typically in the case of Fig. 10, the first reaction will be studied at zero concentration of the product X_1, while the second reaction will be studied with variable non-zero concentrations of X_1. On the contrary in the cell, both reactions are linked by X_1 concentration, which, in the cell, is always the same (if the intracellular medium is supposed homogeneous; see the discussion later). In a metabolic network the metabolites concentrations are determined by the network structure and the rate equations. This point is shown in the right part of Fig. 10. In a metabolic network, the intermediate metabolites are the links between the reactions. From the mathematical point of view one can derive relationships between the rates by elimination of the intermediate metabolites concentrations, so that the rate of possible rates is a one-dimensional space in the case of example of Fig. 10 (see Fig. 11).

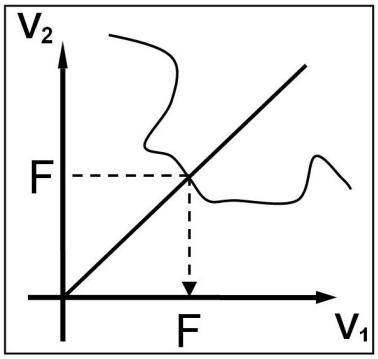


Figure 11: Phase space

3.2 Is cellular medium homogeneous?

The answer is most of the time probably not. Any view of a cell image obtained with an electronic or a photonic microscope indicates clearly an heterogeneous arrangement. Furthermore there are several examples of compartimentation and of channelling (Agius and Sherratt, 1997).

However, most of the modelling of metabolism assume (without saying it explicitly most of the time) that the cell medium is homogeneous. The reason is that the same study in an heterogeneous medium is much more difficult; a first study in an homogeneous medium can afford a first description of the main behaviour of the metabolic network. Differential equations are well adapted to concentrations supposed to be the same in all directions.

3.3 Cellular metabolisme as an open system

There are two ways to approach a metabolic network as shown in Fig. 12 below:

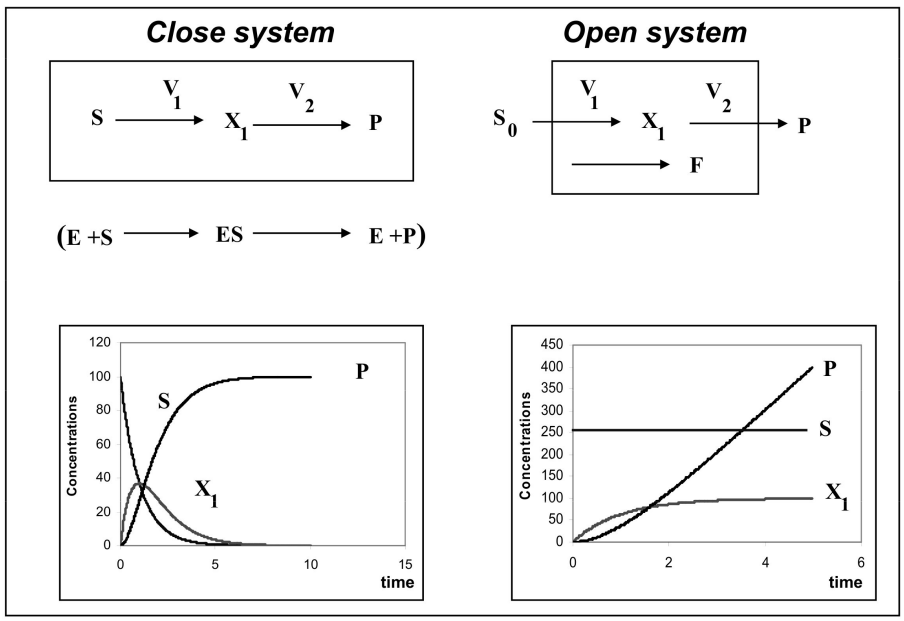


Figure 12: Two ways of considering a metabolic network

Either as a close system, which tends towards an equilibrium with usually several concentrations equal to zero due to (quasi) irreversible reactions. This situation does not correspond to the cellular situation for which the external metabolites are continuously imported and products exported. Even if one considere only part of metabolism one use to assume that the concentration of the metabolites outside the considered part of the network (S_0 and P in Fig. 12) are constants or produced at constant rate. In these conditions the metabolite concentrations inside the network reach a steady-state with non zero constant values.

3.4 The steady-state

For the calculation of the flux in a metabolic network, one use to consider that the intermediate metabolites have a constant value, i.e. that the production and the consumption of a given metabolite balances. This hypothesis is fairly well experimentally verified. On periods of time lasting several minutes and sometimes more, one can assume that most of the metabolite concentrations

are constant. One has however to be aware that this hypothesis does not hold when the physiological conditions are changed : rest to work transition for a muscle cell for instance. Some pathways are oscillatory : glycolosis, production of insulin by the β-cells in pancreas.

In the following we will accept the hypothesis of steady-state, knowing that it is not always satisfied.

Analysis of metabolic networks using elementary flux modes

Sabine Peres[1]

[1] LRI, Univervité Paris-Sud & UMR CNRS 8623, F-91405 Orsay, France

1 Formalisation of metabolic networks

The increasing quantity of data allows us to study of complex metabolic networks. The traditional modelling techniques of such metabolic networks are based on dynamical systems of ordinary differential equations allowing the analysis and the prediction of metabolic flux distributions under diverse physiological and genetic conditions. Dynamic mathematical modelling of large-scale networks meets difficulties because the necessary mechanistic details and kinetic parameters are not all available. However, independently of their dynamics, large metabolic networks exhibit a complex topological structure which can be studied in itself and produces already useful predictions. To understand the phenotypic capabilities of organisms, it is useful to characterize cellular metabolism through the analysis of pathway operations. Metabolic pathway analysis concentrates on the stoichiometric rather than kinetic properties of metabolic networks.

One important characteristics of a network is its boundaries. Related to this issue is the notion of internal and external species. Internal species are those which are explicitly considered in the network model and external species are thought to be sinks or sources which lie outside the system. All metabolic network with m internal metabolites and r reactions can be represented by a stoichiometric matrix N of m rows and r columns such that :

$$N_{ij} = \begin{cases} a & \text{number of molecules } i \text{ products by the reaction } j \\ -a & \text{number of molecules } i \text{ consumes by the reaction } j \\ 0 & \text{otherwise} \end{cases}$$

For example, let's consider a simple metabolic network (represented in Figure 1) and its stoichiometric matrix N which contains the following reactions :

$\mathrm{R}_1 : \mathrm{P}_1 \rightarrow \mathrm{X}$
$\mathrm{R}_2 : \mathrm{X} \leftrightarrow \mathrm{Y}$
$\mathrm{R}_3 : \mathrm{X} \rightarrow \mathrm{Z}$
$\mathrm{R}_4 : \mathrm{Y} + \mathrm{Z} \rightarrow \mathrm{P}_2 + \mathrm{P}_3$
$\mathrm{R}_5 : \mathrm{Y} \leftrightarrow \mathrm{P}_4$

$$N = \begin{array}{c} \begin{array}{ccccc} R1 & R2 & R3 & R4 & R5 \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix} \end{array} \begin{array}{c} X \\ Y \\ Z \end{array}$$

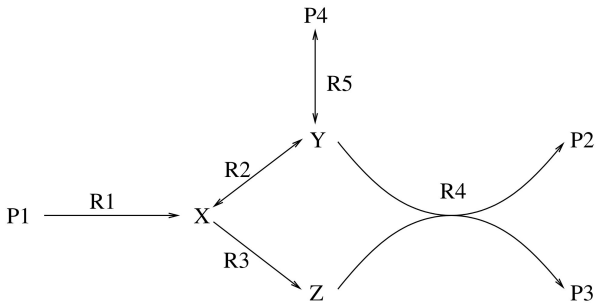


FIG. 1 – Metabolic network example. The metabolites X, Y and Z are the internal metabolites and P_1, P_2, P_3 and P_4 are external metabolites which are the sources or the sinks of the network. The metabolic network is represented by an hypergraph where the metabolites are represented by nodes and the reactions are represented by the edges. The reversible reactions are represented by a double arrow.

Such networks summarize the capabilities of a system and show how a set of source compounds can be converted into a set of target compounds. The main methods for identifying pathways in metabolic networks is the concept of elementary flux modes [16]. This constraint-based approach allows predicting fluxes [1] for various organisms. The main constraints that have been considered are the steady states (every metabolite that is produced has to be consumed) and thermodynamic constraints (irreversible reactions can only be taken in the appropriate direction).

2 Elementary flux modes

Elementary flux modes approach has been developed by Schuster *et al.* [14, 15], then by many others [9, 5]. Elementary flux modes (EM) can be defined as the smallest sub-network enabling the metabolic system to operate at steady state with all irreversible reactions proceeding in the appropriate direction [14, 15].

Let a metabolic network composed of r reactions e_i and m metabolites represented by a stoichiometric matrix N. A vector $e = (e_1, ..., e_r)^t$ is an EM if it fulfills the following conditions :

1. Steady state : $Ne = 0$.
2. Feasibility : For all index i of an irreversible reaction, $e_i \geq 0$.
3. Minimality : For all EM e' of N, $supp(e') \subseteq supp(e) \Rightarrow \exists \alpha \in \mathbb{R}$ such that $e' = \alpha e$ with $supp(v) = \{j \in \mathbb{N} : v_j \neq 0\}$

[1] A flux is defined as the production or consumption of mass per volume per unit time.

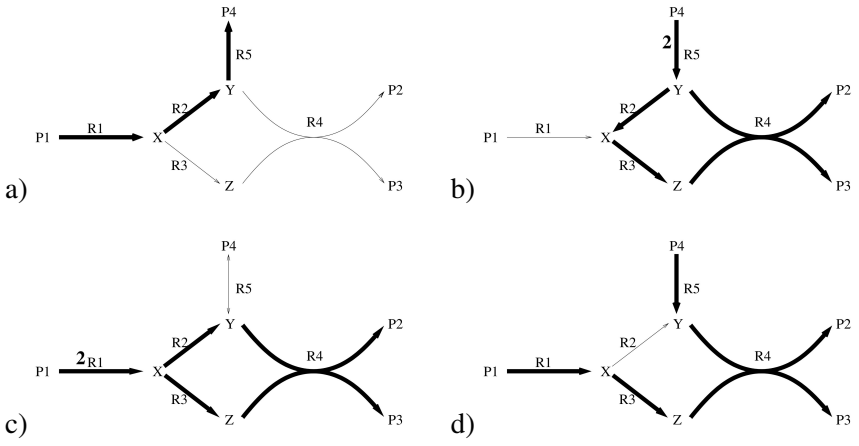


FIG. 2 – EMs of the metabolic network example

For example, the metabolic network represented in Figure 1 contains four EMs which are represented in Figure 2.

Such pathway definition provides a rigorous basis to systematically characterize cellular phenotypes, robustness and fragility that facilitate understanding of cell physiology. EM analysis is a useful metabolic pathway analysis tool to identify the structure of a metabolic network that links the cellular phenotype to the corresponding genotype. Applications of network-based pathway analyses have been presented for predicting functional properties of metabolic networks, measuring different aspects of robustness and flexibility, and even assessing gene regulatory features [18, 5]. In the context of cellular metabolism, robustness is defined as the ability of cells to achieve the optimal performance even under perturbations imposed by a gene knockout. The robustness of cellular metabolism is mainly due to the redundancy of pathway options that the wild type can choose to achieve similar performance. In our metabolic network example, there is no alternative pathway to produce P_4 if one enzyme of the EM 1 is inhibited. It is worth noting that there are three different ways to produce P_2 and P_3 and all the EMs contains the reaction R_3. So R_3 is an essential reaction for the production of P_2 and P_3. On the other hand, if the reaction R_2 is inhibited, it is still possible to find a pathway which P_2 and P_3 with the EM 4. In contrast to the structural robustness, the concept of minimal cut set has been introduced to determine the minimal set of reactions whose deletion completely blocks a target [8, 3]. In our metabolic network example, there are 5 minimal cut sets which prevent the productin of P_2 and P_3 : $\{R_3\}$, $\{R_4\}$, $\{R_1; R_5\}$, $\{R_1; R_2\}$ and $\{R_2; R_5\}$. With these concepts it is also possible to compare the pathways taken in different tissues in different

physiological conditions. It highlights the mutations that can be tolerated or not (essential genes). In biotechnologies it indicates pathways with high product yield. This approach has been applied to the mitochondrial energy metabolism in different tissues or organisms [12, 10], revealing non classical pathways allowing obviating some particular mutations [17].

3 Discussion

EMs can only be enumerated in small to medium-scale metabolic networks because the number of EMs increase exponentially with the network size [9]. The huge number of elementary flux modes associated with large biochemical networks prevents from drawing simple conclusions from their analysis. Estimation of the number of EMs has also been examined [9] in order to predict the complexity of the computational task to find all such metabolic pathways. Acuna *et al.* [1, 2] gave a systematic overview of the complexity of the optimisation problems related to EMs and showed that the counting of EMs is #P-complete. Studies have been carried out on the analysis of sub-networks. AcoM [10, 11] was a first attempt to deal with this problem. It is a biclustering method based on the Agglomeration of Common Motifs (ACoM). It was applied to the central carbon metabolism in *Bacillus subtilis* and to the yeast mitochondrial energy metabolism. It helped to give biological meaning to the different elementary flux modes and to the relatedness between reactions. Kaleta and co-workers [7] showed that the analysis of small sub-networks can be misleading. To overcome this problem, they introduced the concept of elementary flux patterns that takes into account the steady-state fluxes through a metabolic network at the genome scale when analyzing pathways in a sub-network. As an alternative, recently there have been attempts to find specific metabolic pathways from stoichiometric information by optimization modeling. They are able to cope with the scale and furthermore to identify pathways with specific requirements. Also, De Figueiredo *et al.* [4] proposed a procedure to determine the $K-$shortest EMs in large-scale metabolic networks by an optimization model that allows exploring a specific subset of EMs of interest. More recently, the $K-$shortest generating flux modes, a subset of EMs, have also been investigated by a similar model [13]. Being a computationally demanding task, several approaches to parallelize or distribute the computations of elementary modes have been proposed through parallelization techniques [6] or algorithmic reformulations [20, 19]. Nevertheless, although several improvements have been introduced for computing EMs in large networks, tools are still needed to allow their large-scale analysis and interpretation.

Références

[1] V. Acuna, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks : Complexity and algorithms. *BioSystems*, 95 :51–60, 2009.

[2] V. Acuna, A. Marchetti-Spaccamela, M-F. Sagot, and L. Stougie. A note on the complexity of finding and enumerating elementary modes. *BioSystems*, 99 :210–214, 2010.

[3] Kathrin Ballerstein, Axel von Kamp, Steffen Klamt, and Utz-Uwe Haus. Minimal cut sets in a metabolic network are elementary modes in a dual network. *Bioinformatics*, 28(3) :381–387, 2012.

[4] L.F. De Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J.E. Beasley, S. Schuster, and F.J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23) :3158–3165, 2009.

[5] J. Gagneur and S Klamt. Computation of elementary modes : a unifying framework and the new binary approach. *BMC Bioinformatics*, 5(175), 2004.

[6] Dimitrije Jevremovic, Cong T. Trinh, Friedrich Srienc, Carlos P. Sosa, and Daniel Boley. Parallelization of nullspace algorithm for the computation of metabolic pathways. *Parallel Computing*, 37(6-7) :261–278, 2011.

[7] C. Kaleta, L.F. De Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts ? pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Research*, 19 :1872–1883, 2009.

[8] S. Klamt and E.D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20 :226–234, 2004.

[9] S. Klamt and J. Stelling. Combinatorial complexity of pathway anaysis in metabolic networks. *Mol Bio Rep*, 29 :233–236, 2002.

[10] S. Peres, M. Beurton-Aimar, and J.P. Mazat. Classification des modes élémentaires : application au métabolisme énergétique mitochondrial. *Technique et Science Informatiques*, 26 :199–218, 2007.

[11] S. Peres, F. Vallée, M. Beurton-Aimar, and J.P. Mazat. Acom : a classification method for elementary flux modes based on motif finding. *Biosystems*, 103(3) :410–419, 2011.

[12] S. Pérès, M. Beurton-Aimar, and J-P. Mazat. Pathway classification of TCA cycle. *IEE Proc. Syst. Biol.*, 153(5) :369–371, 2006.

[13] A. Rezola, L. F. de Figueiredo, M. Brock, J. Pey, A. Podhorski, C. Wittmann, S. Schuster, A. Bockmayr, and F. J. Planes. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, 27(4) :534–540, 2011.

[14] S. Schuster, T. Dandekar, and D.A. Fell. Detection of elementary modes in biochemical networks : A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, 17 :53–60, 1999.

[15] S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, 18 :326–332, 2000.

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

[17] C. Schwimmer, L. Lefebvre-Legendre, M. Rak, A. Devin, P. Slonimski, J.P. di Rago, and M. Rigoulet. Increasing mitochondrial substrate-level phosphorylation can rescue respiratory growth of an atp synthase-deficient yeast. *J. Biol. Chem.*, 280(35) :30751–30759, 2005.

[18] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. Metabolic network structure determines key aspect of functionnality and regulation. *Nature*, 420 :190–193, 2002.

[19] Marco Terzer and Jörg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19) :2229–2235, 2008.

[20] A. Von Kamp and S. Schuster. Metatool 5.0 : Fast and flexible elementary modes analysis. *Bioinformatics*, 22 :1930–1931, 2006.

Introduction to Stochastic Simulation Algorithm

Fabien Campillo[1]

[1] Inria Méditerranée, projet MODEMIC Inria/INRA, INRA UMR MISTEA
F-34060 Montpellier, France

Abstract

Many discrete and random systems (like chemical dynamics, population dynamics or biochemical networks dynamics) are described in terms of Ordinary Differential Equations (ODE). The ODE formalism describes implicitly the deterministic limit of such systems: the final result is the average of all possible realizations.

In many theoretical and computational problems, we do not wish to ignore the intrinsic fluctuations of a particular solution and its discrete nature. Indeed, in many cases, the variability of one trajectory is needed for realism's sake. This question has lead to several algorithms that describe and compute *particular* and realistic trajectories.

The topic of this lecture will be focused on one of them: the Gillespie algorithm also called the Stochastic Simulation Algorithm (SSA). We will also present the tau-leaping technique and the diffusion approximation of such models.

We will present a simple version of the mathematical background behind the SSA and provide some indications about the convergence of such models toward stochastic differential equations and ordinary differential equation.

We will also introduce individual-based models that can be seen as extensions of the Gillespie algorithm.

LIST OF ATTENDEES

(February 15th, 2013)

ABDULLA Saman	(samankeley@yahoo.com)
ADAM Laura	(ladam@vbi.vt.edu)
AMAR Patrick	(pa@lri.fr)
ARANGANATHAN Naveen Kumar	(navyeinstein1991@outlook.com)
BEDAU Mark	(mark.bedau@mail.com)
BERNOT Gilles	(bernot@unice.fr)
BEURTON-AIMAR Marie	(beurton@labri.fr)
BOIS Frédéric	(fabienne.carette@ineris.fr)
BOUYIOUKOS Costas	(costas.bouyioukos@issb.genopole.fr)
BYRNE Helen	(helen.byrne@maths.ox.ac.uk)
CHANCELLOR Courtney	(Courtney.Chancellor@ec-nantes.fr)
COURBET Alexis	(alexis.courbet@sysdiag.cnrs.fr)
COURTADE-GAIANI Sophie	(sophie.courtade@fr.netgrs.com)
COX Robert	(sidney@dna.caltech.edu)
DA VEIGA MOREIRA Jorgelindo	(d_jorgelindo@yahoo.fr)
DOULAZMI Mohamed	(Mohamed.doulazmi@upmc.fr)
DRESS Andreas	(andreas.dress@mail.com)
EL SAYYED Hafez	(hafez.elsayyed@gmail.com)
FOLSCHETTE Maxime	(Maxime.Folschette@irccyn.ec-nantes.fr)
GALLET Emmanuelle	(emmanuelle.gallet@ecp.fr)
GILES Rachel	(rachel.giles@mail.com)
HEUX Stéphanie	(heux@insa-toulouse.fr)
IOVKOVA Alexandra	(alexandra.iovkova@tum.de)
JUNIER Ivan	(i.junier@gmail.com)
KÉPÈS François	(francois.kepes@issb.genopole.fr)

KITNEY Richard	(richard.kitney@mail.com)
KOUTROUMPAS Konstantinos	(konstantinos.koutroumpas@issb.genopole.fr)
LE GALL Pascale	(pascale.legall@issb.genopole.fr)
LEPAGE Thibaut	(thibaut.lepage@issb.genopole.fr)
MACOVEI Corina	(macovei_corina88@yahoo.com)
MAZAT Jean-Pierre	(jean-pierre.mazat@phys-mito.u-bordeaux2.fr)
MOLDOVAN Camelia	(camelia@moldovan.eu)
MOLINA Franck	(franck.molina@sysdiag.cnrs.fr)
NGUYEN Vu Ngoc Tung	(tnguyen@labri.fr)
NICOLAE Irina Emilia	(irinaemilia.nicolae@gmail.com)
NORRIS Victor	(victor.norris@univ-rouen.fr)
PECOU GAMBAUDO Elisabeth	(elisabeth.pecou@sobios.com)
PERES Sabine	(sabine.peres@lri.fr)
ROUX Jérémie	(jeremie_roux@hms.harvard.edu)
SAINZ DE MURIETA Iñaki	(inaki.sainzdemurieta@upm.es)
SCHAERLI Yolanda	(yolanda.schaerli@crg.eu)
SCHNEIDER Dominique	(dominique.schneider@mail.com)
SCHUBERT Walter	(walter.schubert@mail.com)
SOLÉ Ricard	(ricard.sole@mail.com)
SOYER Orkun	(orkun.soyer@mail.com)
SYLVAIN Benito	(sylvain.benito@sobios.com)
VIEIRA Gilles	(vieira@insa-toulouse.fr)
WILTSCHI Birgit	(birgit.wiltschi@acib.at)
ZELISZEWSKI Dominique	(dominique.zeliszewski@issb.genopole.fr)
ZEMIRLINE Abdallah	(zemirlin@univ-brest.fr)