Proceedings of
The Sophia-Antipolis Spring School
on

Modelling Complex Biological Systems in the Context of Genomics

May 23rd - 27th, 2011

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

What are the salient features of the new scientific context within which biological modelling and simulation will evolve from now on? The global project of high-throughput biology may be summarized as follows. After genome sequencing comes the annotation by 'classical' bioinformatics means. It then becomes important to interpret the annotations, to understand the interactions between biological functions, to predict the outcome of perturbations, while incorporating the results from post genomics studies (of course, sequencing and annotation do not stop when simulation comes into the picture). At that stage, a tight interplay between model, simulation and bench experimentation is crucial. Taking on this challenge therefore requires specialists from across the sciences to learn each other's language so as to collaborate effectively on defined projects.

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.

The works of subgroups underpinned the conferences organised in Autrans in 2002, in Dieppe in 2003, in Evry in 2004, in Montpelliers in 2005, in Bordeaux in 2006, back to Evry in 2007, in Lille in 2008, in Nice in 2009 and in Evry in 2010. The conferences in Sophia-Antipolis in 2011 which as reported here, brought together over a hundred participants, biologists, physical chemists, physicists, statisticians, mathematicians and computer scientists and gave leading specialists the opportunity to address an audience of doctoral and post-doctoral students as well as colleagues from other disciplines.

This book gathers overviews of the talks, original articles contributed by speakers and subgroups, 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, Eric Goles, Janine Guespin, Jürgen Jost, 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 Sophia-Antipolis *Mercure* 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 *Fort Carré* at Antibes.

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® Evry:
  http://www.genopole.fr
- GDRE CNRS 513 Biologie Systémique:
  http://www.mpi-magdeburg.mpg.de/CNRS_MPG
- Consortium BioIntelligence (OSEO)
- École doctorale STIC
- 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
- Fondation Scientifique Fourmentin-Guilbert:
  http://www.fourmentinguilbert.org

THE EDITORS

INVITED SPEAKERS

Anthony BLAU	Univ. of Washington, Seattle WA (USA)
Luca CARDELLI	Microsoft Research, Cambridge (UK)
Peter COOK	Oxford Univ., (UK)
Christoph CREMER	Kirchhoff It., Heidelberg (DE)
Attila CSIKASZ-NAGY	Univ. of Trento, (IT)
Rui DILÃO	It. Superior Tecnico, Lisboa (PT)
Geneviève DUPONT	ULB, Brussels, (BE)
Andrew GEWIRTZ	Emory Univ., Atlanta, GA (USA)
David HAREL	Weizmann It., Rehovot (IL)
Jeff HASTY	Univ. of San Diego, CA (USA)
Piet HERDEWIJN	iSSB & KU Leuven (BE)
Alfonso JARAMILLO	iSSB & Univ. d'Evry (F)
Ross KING	Aberystwyth Univ. (UK)
Klaus SCHERRER	Univ. Paris-Diderot (F)
Stefan SCHUSTER	Univ. of Jena (DE)

CONTENTS

PART I INVITED TALKS

3D Genome Organisation and Gene Expression: Unified Matrix Hypothesis and Genon Concept

Klaus Scherrer[1]

[1] Institut Jacques Monod, CNRS and University Paris Diderot, Paris, France.

Abstract

The Unified Matrix Hypothesis (UMH) proposed that genomes may be organised in space, and genomic domains be transcribed in specific sectors of the nucleus [1]. The UMH generalised for normal types of interphase cells, the pattern of *Ectopic Pairing* observed in drosophila salivary gland cells. The recent publication of the 3D structure of the yeast genome [2] is clear confirmation in lower eukaryotes of the UMH concept. Genomic domains of higher eukaryots are visible in polytene chromosomes of *sciaridae* and *drosophila* (C-value up to 10 000) as bands, representing units of transcription and meiotic recombination, which are held in 3D positions by ectopic cables linking distant interbands, within and in between chromosomes, as well as the nucleolus; in this type of cell, this pattern is genetically determined [1]. DNA in normal cells is flexible and able to link directly distant sites within and between chromosomes as obvious in yeast [2]. Of conceptual importance within the UMH concept is that DNA length *per se* represents genetic information, independent of sequence, as a basis of nuclear architecture and cellular morphogenesis.

Specific types of cells have changing patterns of hetero- and euchromatin; the phenomenon of “quantal mitosis” [1] shows that differentiation can be stopped by agents blocking chromatin remodelling. 3D DNA organisation determines *chromosome territories*, which are relatively stable including heterochromatin. Euchromatin, however, may participate in additional more flexible DNA interactions, which are conditional for individual gene expression, as shown in case of association of the distant TH2 cytokine and IFN-γ loci [3]; such “*3D gene regulation*” recalls the pattern observed in yeast [2].

Within the *Genon* concept of regulation [4, 5], 3D genome organisation may represent the highest-level program of gene expression encoded in the entire genome. Downstream, the expression of genomic domains and individual genes is implemented by transcription, differential splicing of pre-mRNA, mRNA transport as well as repression or activation, which are controlled *in cis* by the sequential expression of genetic programs termed *protogenon* for the domains, *pre-genon* for pre-mRNA, and genon for mRNA. This cis information

is controlled in trans by factors representing the corresponding *transgenons* in nucleus and cytoplasm. Cis and transgenon control, furthermore, mRNA sorting basic to protein biosynthesis in specific cell sectors.

The UMH proposed, furthermore, a logical link between cellular and supra-cellular morphogenesis. The latter is based on *programs of spindle orientation* defining the direction of subsequent cell divisions, prior to cell-cell interaction and selective apoptosis. Spindle re-orientation happens at critical steps of differentiation and morphogenesis and is, hence, based on the internal topological organisation of the mitotic cell.

Finally, 3D organisation of genome, transcripts and gene expression may explain most of the apparent excess of DNA observed in eukaryots and, in particular, the C-value paradox [1]. It may allow, furthermore, to link DNA polymorphism and supra-cellular morphogenesis in individuals as a paradigm of, e.g., genesis of facial patterns.

References

[1] Scherrer (1989) *Biosci Rep* **9**: 157-188; doi: 10.1007/BF01115994

[2] Duan et al. (2010) *Nature* **465**: 363-367; doi:10.1038/nature0897;

[3] Spilianakis et al. (2005) *Nature* **435**: 637-645; doi:10.1038/nature03574

[4] Scherrer and Jost (2007) *Molecular Systems Biology* **3; 87**
doi:10.1038/msb4100123

[5] Scherrer and Jost (2007) *Theory Biosci.* **126**: 65-113;
doi: 10.1007/s12064-007-0012-x
(see also discussion in doi: 10.1007/s12064-009-0027-1, etc.)

Active RNA polymerases are immobile molecular machines

Peter R. Cook[1]

[1] The Sir William Dunn School of Pathology, University of Oxford

Abstract

A parsimonious model for all genomes involving one major architectural motif will be presented: DNA/chromatin loops are tethered to transcription factories through active RNA polymerases and/or transcription factors. The polymerases are immobile and produce their transcripts by reeling in the DNA; this contrasts with the conventional view where polymerases track like locomotives down the template.

At least two theoretical mechanisms probably drive the required protein clustering and DNA looping - the dimerization of bound transcription factors and an (entropic) depletion attraction acting between engaged polymerizing complexes. We have also tested experimentally whether active polymerases are immobile using chromosome conformation capture and human genes switched on rapidly (i.e., within 10 min) and synchronously by tumor necrosis factor α. This potent cytokine signals through NFkB to stimulate and repress many genes. Two of the first to respond are *SAMD4A* (a 221-kbp gene that a polymerase takes $>$ 1h to transcribe), and *TNFAIP2* (a 10-kbp gene that is used as a reference and which is transcribed repeatedly). Ten minutes after stimulation, the reference gene develops new contacts with the *SAMD4A* promoter. Subsequently, these contacts are lost as new ones appear further downstream in *SAMD4A*; contacts are invariably between sequences being transcribed at that particular moment. Super-resolution microscopy confirms that nascent transcripts (detected by RNA fluorescence *in situ* hybridization) co-localize at relevant times. These results are consistent with active polymerases being immobilized. Moreover, many genes responding to TNFa often come together to be transcribed in specialized “NFkB” factories. In additional experiments, we have isolated complexes of $>$ 8 MDa that represent factory cores, and determined their proteomes by mass spectrometry.

Lightoptical Analysis of Nuclear Nanostructure Analysis and Modelling

Christoph Cremer[1,2,3]

[1] Applied Optics and Information Processing, University Heidelberg
[2] Interdisciplinary Center for Scientific Computing (IWR), University Heidelberg
Im Neuenheimer Feld, D-69120 Heidelberg, Germany
[3] Institute for Molecular Biophysics, The Jackson Laboratory,
Bar Harbor, ME 04609, USA

Abstract

The spatial organisation of the genome in the cell nucleus has emerged as a key element to understand gene expression. A wealth of molecular and microscopic information has been accumulated, resulting in a variety of - sometimes contradictory - models of nuclear architecture on the nanoscale. A major source of such unambiguities is due to the limits of conventional light microscopy (optical resolution about 200 nm laterally, 600 nm axially) which makes quantitative tests of model calculations on the nanoscale very difficult. To overcome this bottleneck, we have established a variety of superresolution microscopy ("nanoscopy") methods. Our present spectrum for nanoscopy of nuclear architecture comprises confocal laser scanning 4Pi-microscopy, Spatially Modulated Illumination (SMI), and Spectrally Assigned Localization Microscopy (SALM). Using a recently developed SALM technique, Spectral Precision Distance/Position Determination Microscopy (SPDM) with Physically Modifiable Fluorophores (SPDM$_{Phymod}$), nuclear nanostructures can now be studied on a large scale in 3D intact nuclei of mammalian cells down to a lateral optical resolution of individual molecules in the 20 nm range, using a variety of standard fluorescence proteins/fluorochromes. Examples are provided for the spatial distribution of individually resolved nuclear pore complex proteins, of histones, RNA Polymerase II and FISH labelled DNA sequences. First applications of such nanoscopy methods to modeling of nuclear nanostructure regard the use of statistical methods to infer structural features and density fluctuations an the nanoscale. The observed fluctuations were consistent with a recently proposed numerical chromatin model.

References

[1] M. Gunkel et al. (2009) Dual color localization microscopy of cellular nanostructures. *Biotechnology J.* **4**: 927 938.

[2] M. Bohn et al. (2010) Localization microscopy reveals expression dependent parameters of chromatin nanostructure. *Biophys. J.* **99**: 1358 1367.

[3] J. Rouquette et al. (2010) Functional nuclear architecture studied by microscopy. *International Review of Cell and Molecular Biolog* **282**: 1 90.

[4] C. Cremer et al. (2010) Far field fluorescence microscopy of cellular structures @ molecular resolution. *In: Nanoscopy and Multidimensional Optical Fluorescence Microscopy (A. Diaspro, Edit.)* pp. 3/1 3/35. Taylor & Francis.

[5] D. Hübschmann et al. Quantitative Approaches to Nuclear Architecture Analysis and Modelling. *in press*

[6] Y. Markaki et al. Chromatin Domains, Perichromatin Region and Interchromatin Compartment: A functional Marriage a Trois. *Cold Spring Harbor Symposia* **75**, in press

Unraveling Genome Architecture

Anthony BLAU[1]

[1] Department of Genome Sciences, University of Washington,
Seattle WA, USA

Abstract

The genome of a cell is organized non-randomly within the nucleus, with this three-dimensional organization serving to modulate various genomic functions, such as gene expression, DNA replication and maintenance. In mammalian cells, the spatial organization of genomes plays important roles in cellular and developmental events. Defects in genome architecture are linked to human diseases, including cancer. However, due to technological obstacles, little is known about how genomes are organized in vivo and what the principles are that guide chromatin folding and assembly.

Genome spatial organization was traditionally studied by microscopy-based DNA imaging technologies such as fluorescence in situ hybridization (FISH), which are limited in resolution and throughput. Chromosome conformation capture (3C) and its derivatives (4C, 5C, 6C, ChIA- PET, and e4C) have proved to be powerful molecular tools for characterizing locus-specific or protein complex-mediated structural properties of the genome. By combining 4C with next generation sequencing technology, we recently developed an ultra-high-throughput method that resulted in a high-resolution (kilobase) three-dimensional model of the haploid yeast genome. The map recapitulates known features of genome organization, thereby validating the method, and identifies new features. Extensive regional and higher order folding of individual chromosomes is observed. Chromosome XII exhibits a striking conformation that implicates the nucleolus as a formidable barrier to interaction between DNA sequences at either end. Inter- chromosomal contacts are anchored by centromeres and include interactions among transfer RNA genes, among origins of early DNA replication and among sites where chromosomal breakpoints occur. Our findings provide a glimpse of the interface between the form and function of a eukaryotic genome.

Modelling the spatio-temporal organization of intracellular Ca^{2+} signals : From mechanisms to physiology

Geneviève Dupont[1]

[1] Université Libre de Bruxelles. Unité de Chronobiologie Théorique.
Faculté des Sciences, CP231.

Abstract

Signal-induced Ca^{2+} oscillations have been observed in many cell types and play a primary role in cell physiology. They mediate vital physiological processes such as secretion, gene expression or fertilization. Specificity in the physiological responses is ensured by the high level of spatio-temporal organization of Ca^{2+} dynamics in the form of stochastic sub-cellular increases, regular oscillations and intra- or intercellular Ca^{2+} waves.

In the talk, I will first present the main features of the hierarchical organization of Ca^{2+} signalling and illustrate on some specific examples how the interplay between experiments and modelling allows for a detailed understanding of the regulatory feedbacks responsible for these phenomena. In the second part, mechanisms for the frequency encoding of Ca^{2+} oscillations will be discussed, with more emphasis on the process of glucagon secretion and on a Ca^{2+}-related pathology occurring at human fertilization.

From the glycolytic oscillations to the control of the cell cycle: a minimal biological oscillator

Rui Dilão[1]

[1] Nonlinear Dynamics Group, Instituto Superior Técnico, Lisboa, Portugal

Abstract

We introduce the basic modeling approach in order to describe chains of enzymatic reactions. We analyze the effects of activation feedback loops in these chains of reactions, and we derive the conditions for the existence of oscillations.

We show that enzymatic chain reactions with two sequential chains and one feedback activation loop describe the basic features of the cell cycle control in eukaryotes. This same enzymatic chain reaction also describes the glycolytic oscillations in yeast. From this modeling approach, it results that the S/G2 checkpoint of the cell cycle is under the control of the concentration of the Cdk protein Cdc25. The concentration of this protein tune several bifurcation parameters of the model equations and its variation can induce the crossing of a Hopf bifurcation, leading to stable oscillation in the concentrations of the Maturation Promoting Factor (MPF=cyclin B+Cdc2) and of its phosphorylated state. This model is consistent with the recent finding that the oscillation of a single Cdk module is sufficient to trigger the major cell cycle events (Coudreuse and Nurse, *Nature*, **468** (2010) 1074-1079).

Automating Biology using Robot Scientists

Ross D. King[1]

[1] Department of Computer Science, Aberystwyth University, UK

Abstract

A Robot Scientist is a physically implemented robotic system that applies techniques from artificial intelligence to execute cycles of automated scientific experimentation. A Robot Scientist can automatically execute cycles of: hypothesis formation, selection of efficient experiments to discriminate between hypotheses, execution of experiments using laboratory automation equipment, and analysis of results. We have developed the Robot Scientist “Adam” to investigate yeast (Saccharomyces cerevisiae) functional genomics. Adam has autonomously identified genes encoding locally “orphan” enzymes in yeast. This is the first time a machine has discovered novel scientific knowledge. To describe Adam’s research we have developed an ontology and logical language. Use of these produced a formal argument involving over 10,000 different research units that relates Adam’s 6.6 million biomass measurements to its conclusions. We are now developing the Robot Scientist “Eve” to automate drug screening and QSAR development.

Combining Metabolic Pathway Analysis with Evolutionary Game Theory

Stefan Schuster[1]

[1] Department of Bioinformatics, Friedrich Schiller University,
Ernst-Abbe-Platz 2, 07743 Jena, Germany

Abstract

Elementary modes in metabolic reaction networks are defined as miminal sets of enzymes that can operate at steady state with all irreversible reactions used in the correct direction. Elementary-modes analysis is a powerful method for detecting all potential pathways in a metabolic network and computing the associated molar yields; it has been applied successfully for a plethora of bacterial, fungal, plant and animal metabolic networks. Metabolic pathways (identified, for example, by elementary modes analysis) can be interpreted as different strategies of organisms. Thus, methods from evolutionary game theory can be employed. Pure and mixed evolutionarily stable strategies correspond to pure pathways and superimposed pathways (which are relevant for robustness), respectively. In Flux Balance Analysis, it is usually assumed that molar yields of relevant products (such as biomass or ATP) have been maximized during evolution. This has been questioned on game theoretical grounds. In particular, in situations that can be characterized as a Prisoner's Dilemma, maximization of flux is not in line with maximization of yield. Under other conditions (that is, for other parameter values of maximal velocities), a harmony game can result. Here, we analyse the optimal situations under varying conditions.

Intestinal-bacteria-immune cell interactions in health and disease

Andrew Gewirtz[1]

[1] Pathology & Laboratory Medicine, Emory U. School of Medicine,
Atlanta, GA, USA

Abstract

The intestinal mucosal immune system is charged with defending this key vast interface with the outside world from the enormous and diverse group of microbes that colonizes these surfaces. A key means by which the mucosal immune system protects the host from such diverse microbes is using germ-line-encoded molecules such as toll-like receptors (TLR) that target structurally conserved motifs that mediate important bacterial functions. The traditional view of TLR is that they are typically quiescent in the presence of commensal bacteria and are activated only upon detection of pathogens whereupon they initiate an inflammatory response that protects against the perturbing pathogen. This paradigm does indeed characterize the intestinal response to a number of acute pathogens but it is also now appreciated that the intestinal microbiota does not consist of mere pathogens or commensal bacteria but, rather is a continuum of microbes and that it is the job of the mucosal immune system to keep such microbes in check and maintain bacterial populations that benefit the host. TLRs play a key role in such policing of the gut microbiota. A particularly important TLR in defending the intestine is TLR5, which recognizes bacterial flagellin, the primary structural component of flagella, which afford bacteria the ability of directed locomotion. This presentation will discuss the roles of intestinal TLRs in host-bacterial interaction with a particular focus on the role, and mechanism, of TLR5 in host defense, chronic inflammatory disease, including inflammatory bowel disease and metabolic disorders. It will also discuss potential approaches to pharmacologically manipulate these pathways to benefit the host.

Speaking the Language of Molecules

Luca Cardelli[1]

[1] Microsoft Research, Cambridge, United Kingdom

Abstract

Computing has progressed through its history by relying on ever smaller programmable structures, inevitably leading us to devices assembled from individual molecules. Molecular systems, however, are not easily constructed, organized, or programmed, simply because they are the smallest possible. We shall look at some very effective 'natural languages' for molecular systems, found naturally in biochemistry, as well as artificial modeling languages used in systems biology. But none of those gives us the ability to flexibly execute molecular programs. Thanks to biotechnology, nucleic acids (DNA/RNA) are currently the only truly "user-programmable" entities at the molecular scale. They can be directed to assemble nano-scale structures, to produce physical forces, to act as sensors and actuators, and to do general computation in between. Eventually we will be able to interface them with biological machinery to detect and cure diseases at the cellular level under program control. Meanwhile, we need to engineer the molecular devices themselves and our ability to program them.

Can we Computerize an Elephant?

David Harel[1]

[1] Dept. Of Computer Science and Applied Mathematics,
Weizmann It., Rehovot, Israel

Abstract

This overview/concept/dream talk will discuss the idea of comprehensive and realistic modeling of biological systems, where we try to understand and analyze an entire system in detail, utilizing in the modeling effort all that is known about it. I will address the motivation for such modeling and the philosophy underlying the techniques for carrying it out, as well as the crucial question of when such models are to be deemed valid, or complete. The examples will be from among the biological modeling efforts our group has been involved in: T cell development, lymph node behavior, organogenesis of the pancreas, and fate determination in the *C. elegans* nematode. The ultimate grand challenge is to produce an interactive, dynamic, computerized model of an entire multicellular organism, such as *C. elegans*, which is complex, but well-defined in terms of anatomy and genetics.

Computational investigations of feedback and feed-forward controls of cell cycle transitions

Attila Csikasz-Nagy[1]

[1] Microsoft Research, Center for Computational Systems Biology,
University of Trento, Italy

Abstract

DNA replication, mitosis and mitotic exit are critical transitions of the cell cycle which should occur only once per cycle. The importance of various positive feedback and feed-forward loops in the irreversibility of these transitions has been investigated recently. By computational modeling we investigate how these loops ensure proper timing and order of cell cycle events. We will show the dynamical features of such regulatory loops and discuss their role in the robustness of the transitions. We will present how various modeling approaches (differential equations, Petri-nets, Model-checking) can highlight different features of the regulatory network.

References

[1] Novak, B., Tyson, J. J., Gyorffy, B., and Csikasz-Nagy, A. Irreversible cell-cycle transitions are due to systems-level feedback, *Nat Cell Biol* **9** (7), 724-8 (2007)

[2] Mura I, Csikasz-Nagy A. Stochastic Petri Net extension of a yeast cell cycle model. *J Theor Biol.* **254** (4), 850-60 (2008)

[3] Ballarini, P. et al. Studying Irreversible Transitions in a Model of Cell Cycle Regulation. *ENTCS* **232**, 39-53 (2009)

[4] Csikasz-Nagy A. Computational systems biology of the cell cycle. *Brief Bioinform* **10** (4), 424-34 (2009)

[5] Csikasz-Nagy, A. et al., Cell cycle regulation by feed-forward loops coupling transcription and phosphorylation, *Mol Syst Biol* **5**, 236 (2009)

[6] Romanel, A., Cardelli, L., Jensen, L. J., and Csikasz-Nagy, A. Universality of transcriptional and post-translational regulation of cell cycle transitions, *under review* (2011).

Genetic clocks from engineered oscillators

Jeff Hasty[1]

[1] Departments of Molecular Biology and Bioengineering BioCircuits Institute
University of California, San Diego, US

Abstract

One defining goal of synthetic biology is the development of engineering-based approaches that enable the construction of gene-regulatory networks according to design specs generated from computational modeling. This has resulted in the construction of several fundamental gene circuits, such as toggle switches and oscillators, which have been applied in novel contexts such as triggered biolm development and cellular population control. In this talk, I will first describe an engineered genetic oscillator in Escherichia coli that is fast, robust, and persistent, with tunable oscillatory periods as fast as 13 minutes. This oscillator was designed using a previously modeled network architecture comprising linked positive and negative feedback loops. Experiments show remarkable robustness and persistence of oscillations in the designed circuit; almost every cell exhibited large-amplitude fluorescence oscillations throughout observation runs. The period of oscillation can be tuned by altering inducer levels. Computational modeling reveals that the key design principle for constructing a robust oscillator is a small time delay in the negative feedback loop, which can mechanistically arise from the cascade of cellular processes involved in forming a functional transcription factor. I will then describe an engineered network with global intercellular coupling that is capable of generating synchronized oscillations in a growing population of cells. The network is based on the interaction of two quorum sensing genes; luxI, which produces an intercellular transcriptional activator (AHL, acylhomoserine lactone), and aiiA, which degrades AHL intracellularly. Microfluidic devices tailored for cellular populations at differing length scales are used to demonstrate collective synchronization properties along with spatiotemporal waves occurring on millimeter scales. The period of the bulk oscillations ranges from 55-90 minutes, depending on the effective degradation rate of the AHL coupling molecule. In large monolayer colonies of cells, the time scale for the diffusive coupling of AHL is characterized by wavefront velocities that range from 8-30 microns/min.

To Be Announced

Alfonso Jaramillo[1]

[1] Epigenomics Project, iSSB, Genopole®, F-91034 Evry, France

A chemistry-based strategy for the development of safe GMO's

Piet Herdewijn[1,2]

[1] Laboratory of Medicinal Chemistry, Rega Institute for Medical Research,
Katholieke Universiteit Leuven, B-3000 Leuven, Belgium
[2] iSBB, Université d'Evry-Val-D'Essonne, Evry, France

Abstract

The aim of the project is to propagate artificial nucleic acids in a microbial cell, for which an uptake system and a processing apparatus for exogenous activated precursors need to be developed. The key accomplishment would be the expression of an artificial aptazyme in the cell, catalyzing an essential metabolic reaction. Such reprogrammed microorganism could become a new instrument for avoiding genetic pollution when performing experiments in synthetic biology. The seminar will primarily deal with discussing the tools that need to be developed to reach this ambitious goal.

PART II ARTICLES

Architectural features of genome-scale metabolic networks and the role of both biochemical and functional constraints

Areejit Samal[1,2] and Olivier C. Martin[1,3]

[1] Laboratoire de Physique Théorique et Modèles Statistiques, CNRS and Univ. Paris-Sud, UMR 8626, F-91405 Orsay, France

[2] Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, D-04103 Leipzig, Germany

[3] INRA, UMR 0320/UMR 8120 Génétique Végétale, Univ. Paris-Sud, F-91190 Gif-sur-Yvette, France

Abstract

Numerous studies have revealed that biological networks exhibit salient structural features. These include fat tailed degree distributions and the small world property. We provide a closer look at these issues in the context of genome-scale metabolic networks, showing that the situation is more subtle than it seems from the most touted papers. Furthermore, recent work suggests that these architectural features may be by-products of biochemical and biological constraints which any live metabolism must comply with.

1 Introduction

Biological research in the last century became progressively dominated by a reductionist approach resulting in detailed understanding of molecular components. However, most system level properties of living systems arise as a result of complex interactions among their numerous constituents that are only beginning to be tackled at the intra-cellular level. The interactions among different cellular constituents lead to several kinds of molecular networks: transcriptional regulatory networks, metabolic networks, protein-protein interaction network, signalling networks, *etc.* All of these have cross interactions but nevertheless can be thought of as functional modules, associated with one or more functions carried out by the cell. Perturbations of these networks can have major consequences, in particular for the overall organism, leading to defense, disease, or death. An important goal for biology in this century is to understand the structure and dynamics of complex biological networks that contribute to the function and resilience of living cells and organisms [1, 2, 3, 4, 5, 6, 7].

Central to the core activities of the cell is its ability to run house-keeping tasks: it must replace molecular constituents upon natural degradation, or if it is to grow and divide, it must transform external "nutrients" into internal compounds that will then allow for cell division. Both tasks require biochemical transformations, and so in essence, a cell's metabolism is a central part of its life. In this article, we focus on metabolic networks; an organism's metabolic network is its set of biochemical reactions, available for converting nutrients into key molecular species required for the growth and maintenance of the cell. Advances in the development of high-throughput data collection techniques coupled with the systematic analysis of fully sequenced genomes have led to reconstruction of a number of organism-specific metabolic networks [8, 9, 10]. We will cover some research of the past decade that focuses on salient architectural features of metabolic networks. Also highlighted are more recent insights into the role of biochemical and functional constraints that may be essential driving forces shaping the structural characteristics of these networks.

2 Salient global structural properties of metabolic networks

2.1 Degree distribution

The degree k of a node in a graph is defined as the number of edges containing that node. The degree distribution, $P(k)$, gives the probability that a randomly selected node has exactly k edges – hereafter referred to as *links* – in the graph. Jeong *et al.* [11] and Wagner and Fell [12] have studied the degree distribution of metabolic networks using two different graphical representations. Jeong *et al.* [11] represented the metabolic network as a directed bipartite graph with two types of nodes: metabolites and reactions. In a directed bipartite graph representation of the metabolic network (cf. Fig.1(a)), each metabolite node can be associated with an in-degree and an out-degree. The in-degree (out-degree) of a metabolite node in such a bipartite graph is the number of reactions in the network that produce (consume) the metabolite. Jeong *et al.* found both the in-degree and out-degree distribution of metabolites to follow approximately a power law $P(k) \sim k^{-\gamma}$ for the metabolic networks of 43 organisms. Further, the degree exponent γ, for both the in-degree and out-degree distributions, was found to be universal and close to 2.2 for the 43 organisms. Independently, Wagner and Fell [12] studied the degree distribution of the *E. coli* metabolic network using the unipartite graph representation, and also found the metabolite connectivity distribution to follow a power law. Thus, both Jeong *et al.* and Wagner and Fell have shown that the degree distribution of metabolic networks follows a power law like many other real-world networks [1], a behavior that is very different from that arising in random graphs [13].

A power law degree distribution implies that although most metabolites participate in only a few reactions, a few instead participate in *many* reactions. The metabolites that have high degree and participate in a large number of reactions are called hubs of the network. Examples of hubs include ATP which pro-

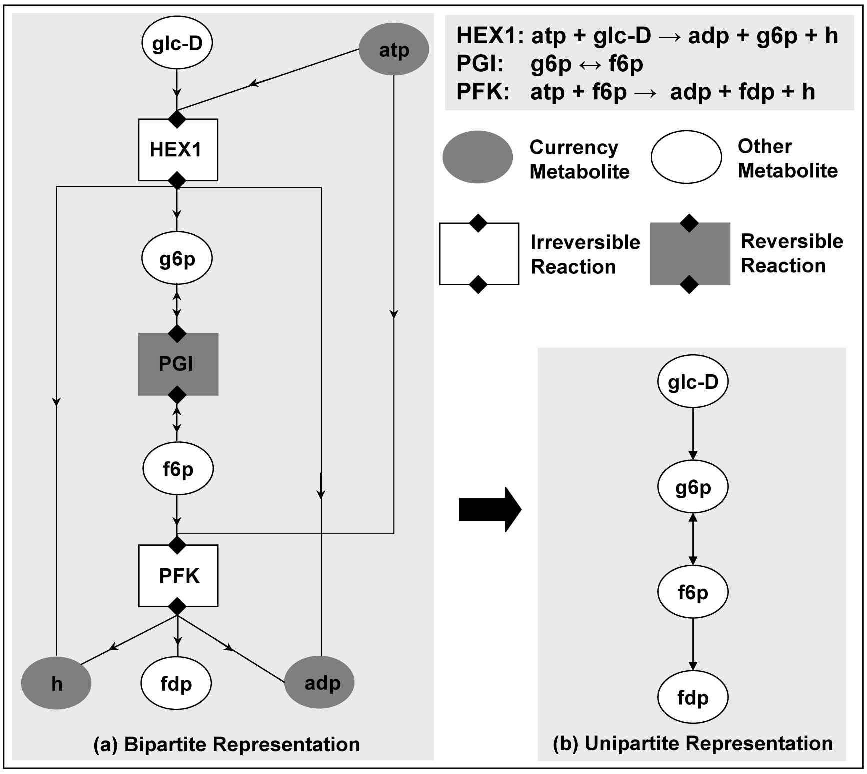

Figure 1: **Different graph-theoretic representations of a metabolic network.** (a) Bipartite graph representation for the three reactions, HEX1, PGI and PFK, in the glycolytic pathway. In the figure, reactions are depicted as rectangles and metabolites as ovals. Reversible reactions are shown in grey and irreversible reactions in white. The primary or other metabolites (white ovals) are distinguished from ubiquitous currency metabolites (grey ovals) in each reaction. If a reaction is reversible, then the links connecting the reaction to its reactant and product metabolites have arrows in both directions. (b) Unipartite metabolite graph representation for the three reactions in the glycolytic pathway obtained by omitting links associated with currency metabolites.

vides the transfer of a phosphate group, NADH which provides the transfer of electrons, etc. A closer look at high degree metabolites revealed that they were either carriers of small biochemical groups, or precursors linking catabolism to anabolism [14, 15]. Tanaka and Doyle have classified the metabolites in the metabolic networks of *H. pylori* and *E. coli* into three separate biochemical categories: carriers, precursors, and "others". They have analyzed the degree distribution of metabolites in the three categories separately, and found the distribution in each category to be close to exponential rather than power like [14, 15]. Recently, Samal and Martin [16] have studied the degree distribution of carriers, precursors and other metabolites separately for *all* reactions in the much larger KEGG database, and have also found the distributions to be different in the three categories.

The degree of a *reaction* in the bipartite metabolic graph is given by the number of metabolites that participate in it. Although the metabolite degree distribution follows a power law, the reaction degree distribution is found to be much different [14, 15, 16]. In contrast to metabolites, reactions do not have very high degree in the network, and most reactions involve exactly 4 metabolites. Of the 4 metabolites in a typical reaction, two metabolites belong to the category *other* (these have low degree) and the remaining two metabolites are carriers (with high degree) in most cases (cf. Fig.1(a)). Tanaka and Doyle have argued that this simple structure of a typical reaction involving 4 metabolites (two primary metabolites of type *other* along with two carrier metabolites transferring a small biochemical group) can explain the emergence of a broad degree distribution when considering all metabolites in the network [14, 15]. Samal and Martin [16] have shown that similar conclusions also hold when considering all reactions in the larger KEGG database.

2.2 Path length and clustering coefficient

The shortest path and thus distance between nodes i and j in a graph is defined via the minimum number of links that have to be traversed to reach node j from node i. The average path length of a graph is defined as the average of this length when considering all pairs of nodes in the graph. The diameter of a graph is defined as the supremum of the shortest paths between all pairs of nodes in the graph. Finally, the clustering coefficient of a node in a graph quantifies the extent to which its neighbours are connected to one another, and is given by the number of links between these nodes divided by the number of links that could possibly exist between them [17].

Both Jeong *et al* [11] and Wagner and Fell [12] found the average path length between metabolites in metabolic networks of different organisms to be between 3 and 4, which is close to value expected in a random graph with similar average connectivity. However, the average clustering coefficient of metabolites in metabolic networks of different organisms was found to be much higher than expected, *i.e.*, higher than would occur in a random graph with the same average connectivity [11, 12, 18]. These studies concluded that metabolic networks of organisms exhibited the small-world [17] property due to high clustering and small average path length between metabolites in the network.

Arita [19] constructed a modified directed metabolite graph from the *E. coli* metabolic network by connecting two metabolites if at least one carbon atom is transferred between them. This modification better accounts for the actual function of the biochemical reactions arising in the metabolic network. Using this more meaningful graph representation of metabolic network, Arita measured the average path length between metabolites to be 8.4. In a similar vein, Ma and Zeng [20] constructed a directed metabolite graph of the metabolic network by removing the connections through the ubiquitous currency metabolites and accounting for the preferred directionality of reactions in the network (cf. Fig.1(b)). Using this biochemically-motivated unipartite graph, Ma and Zeng also found the average path length between metabolites to be close to 8. These two studies show that when one accounts for the transfer of biochemical groups, directionality of reactions and activity of reactions in a more biochemically meaningful way, structural properties of the metabolic graph are changed. In particular, in this case the average path length is much larger than the one obtained by Jeong *et al* and Wagner and Fell. For the biochemically meaningful unipartite graph, Samal and Martin [16] also find the clustering coefficient between metabolites to be much smaller than that obtained by Wagner and Fell [12] and Ravasz *et al.* [18].

2.3 Topological versus functional robustness

Metabolic networks have been shown to follow a power law degree distribution, or at least to have heavy tailed distributions, much like what happens in many natural or man made networks. It has been suggested that one of the important consequences of power law degree distribution is the vulnerability of the network to selective attack on hubs while being robust to random deletion of nodes. Note that most nodes are of low degree and their deletion does not affect much the average path length between the remaining nodes in the network [21]. For example, in the case of the internet, the removal of high degree nodes corresponding to routers with many connections can turn out to be fatal for the communication system [21]. Similarly, for the *S. cerevisiae* protein-protein interaction network, the essentiality of a protein was also found to be correlated positively with the degree of the protein in the network [22]. Finally, Jeong *et al.* showed that the sequential removal of high degree metabolites from the metabolic network results in a sharp rise of the network diameter due to disintegration of the network into small isolated clusters while the removal of any randomly chosen set of metabolites from the network generally leaves the average path length between the remaining metabolites unaffected [11]. This observation led Jeong *et al.* to conclude that the hubs of the metabolic network are crucial for maintaining overall structure and function of the network. Although the role of high degree metabolites or hubs in maintaining the overall *topological* structure of the metabolic network was well emphasized in the study by Jeong *et al.*, the functionality of the different nodes in metabolic networks remained unexplored. Perhaps surprisingly, low degree metabolites can play an essential role in maintaining metabolic function [26]; we now explain this point.

In case of metabolic networks, metabolites participate in reactions where they are produced or consumed, and the reaction process can be controlled through the catalyzing enzyme which is a gene product. It is unclear how a biological process can lead to removal of metabolites from the network. A removal of high degree metabolite from the metabolic network would require the knockout of all genes whose products catalyze various reactions in which the metabolite participates. Instead, genetic mutations give rise to enzyme or reaction knockouts in the network. Thus, in case of metabolic networks, one is interested in determining the effect of removing a reaction rather than the effect of removing a metabolite. The computational technique of flux balance analysis (FBA) [23, 24] can be exploited to study fluxes through reactions and thus can be used to determine essential reactions for growth in metabolic networks. Using FBA, Mahadevan and Palsson [25] measured the lethality fraction for each metabolite in the metabolic network. The lethality fraction of a metabolite in the metabolic network is given by the fraction of the reactions in which the metabolite is involved as a substrate or a product that are essential for growth. The lethality fraction for different metabolites was shown to be uncorrelated with the degree of the metabolite, and low degree metabolites are just as likely to be critical to the overall network as the high degree metabolites [25]. However, Samal *et al.* [26] showed that almost all essential reactions are explained by their association with low degree metabolites; the essential reactions may involve other metabolites of higher degree, but their essentiality is due to their special production or consumption of an intermediate low degree metabolite that is needed for the eventual production of biomass. Thus, from a consideration of functional robustness or fragility of metabolic networks to naturally occurring perturbations, Samal *et al.* showed that it is the role of low degree metabolites that needs to be considered rather than high degree metabolites. This picture is opposite to what arises in protein interaction networks mentioned previously where high degree proteins are generally essential

[22]. These findings also suggest that the fundamental properties of a flow based network (such as metabolic network) can be significantly different from an “influence” based network (such as a protein-protein interaction network) [25, 26].

2.4 Bow-tie architecture

Given a directed graph, a *strongly connected component* is a maximal set of nodes such that for any pair of nodes i and j in that set there is a directed path from i to j and one from j to i [27]. In general, a directed graph may have one or many strong components. The strong components of a graph consist of disjoint sets of nodes. The strong component with the largest number of nodes is designated as the largest strong component or “giant component”. Given the giant component, one can define its “in-component” as the set of other nodes from which there exists some directed path to the giant component. Similarly one defines the “out-component” as the set of other nodes that can be reached from the giant component by following directed paths. The set of nodes that have no path to or from the nodes in the giant strong component forms the “isolated” subset. Broder *et al.* decomposed the nodes in the graph corresponding to the World Wide Web (WWW) into the four components: giant strong component, in-component, out-component and isolated subsets. They obtained a ‘bow-tie’ macroscopic structure for the WWW with the giant component accounting for more than 25% of the nodes in the graph [28].

Following the approach of Broder *et al.*, Ma and Zeng [20] have explored the global connectivity structure of the metabolic network by constructing a directed unipartite metabolite graph (cf. Fig.1(b)) and decomposing the metabolite nodes into the above mentioned four components. This study revealed a “bow-tie” macroscopic structure of the metabolic network with the giant component accounting for approximately 30% of the metabolites in the network [20], similar to that observed by Broder *et al.* for the World Wide Web [28]. Csete and Doyle have argued that such a bow-tie architecture of the metabolic network with a conserved core and plug-and-play modularity around the core can contribute toward robustness and evolvability of the system [29, 15].

3 In what sense are metabolic networks remarkable?

3.1 Meaningful randomization benchmarks

In the past decade, research on the large-scale structure of metabolic networks have revealed remarkable features such as power-law degree distribution, high clustering, small average path length and bow-tie architecture. To quantify the significance of such salient properties, it is appropriate to test the hypothesis that the observed value in the real network is not statistically different from that expected in a null model. The observed structural properties of metabolic networks clearly distinguish them from those arising in random graphs. However, random graphs are inappropriate null models to quantify the significance of observed properties in metabolic networks because they ignore all potential relevant underlying factors that constrain these networks.

The most commonly used null model to test the significance of observed properties in real biological networks uses randomization procedures based on edge exchange to generate randomized networks starting from the original network for comparison [30, 31]. This edge exchange randomization procedure preserves the degree of each node as given in the original network. But, in the case of metabolic networks, edge exchange randomization generates "random" fictitious reactions violating balance of mass, charge and atomic elements, and such reactions are biochemically meaningless. Hence, randomized metabolic networks generated by edge exchange randomization are inappropriate for comparison with real networks. To overcome this problem, Samal and Martin [16] have recently developed a new method to generate randomized ensembles of metabolic networks which properly takes into account biochemical and functional constraints arising in metabolic networks. These can then provide sensible benchmarks when asking which features of metabolic networks are "remarkable".

3.2 Role of biochemical and functional constraints in shaping metabolic network architecture

To test the significance of any property of real metabolic networks, Samal and Martin [16] have generated randomized ensembles of networks by successively imposing the following macroscopic constraints:

(a) The randomized networks contain only valid biochemical reactions which satisfy atom, mass and charge balance. This is achieved by restricting the set of allowed reactions to those in a validated database such as KEGG, *i.e.*, reactions that are known to occur in real organisms.

(b) The number of reactions and metabolites in each network in the randomized ensemble is fixed to that in the metabolic network of the reference organism to be benchmarked.

(c) Each network in the randomized ensemble satisfies the functional constraint of allowing growth under defined chemical environments. The computational technique of flux balance analysis (FBA) [23, 24] is used to determine the ability of each randomized metabolic network to produce all biomass components under each defined chemical environment. This constraint incorporates into the modeling and the benchmark ensemble the ability of living organisms to grow and reproduce.

Samal and Martin exploited a previously developed Markov Chain Monte Carlo (MCMC) based method [32] to sample spaces of metabolic networks. They found [16] that MCMC could be used to generate at will metabolic networks incorporating the above-mentioned biochemical and functional constraints. They then studied in this randomized ensemble global structural properties such as degree distribution, clustering coefficient, average path length and size of largest strong component. By comparing the structural properties of networks in the randomized ensemble with that of the *E. coli* metabolic network, they found the structural properties of the randomized ensemble to be close to that of the real organism. Thus, this study [16] indicates that the observed global structural properties of real metabolic networks are likely to be consequences of the simplest biochemical and functional constraints. Such a possibility was conjectured earlier in Refs. [33, 34] but direct evidence is now available, albeit from computations in an *in silico* framework.

Discussion and conclusions

On first glance, metabolic networks share common structural properties found in other "complex" networks. A striking such property is the fat tail in the distribution of degrees. Indeed, studies on different organisms have exhibited power law distributions for the metabolites in those organisms' metabolic networks [11, 12]. This feature can be traced to "currency" metabolites, thus called because reactions use them to transfer small groups; they reflect the way biochemistry functions. The fat tail in the metabolite degree distribution can then be considered to come from a "universal" use of these currency metabolites; nature seems to prefer to use these metabolites over and over again for transfers on different molecules rather than having different currency metabolites for different substrates. This justification is quite specific to metabolic networks and does not connect directly with a mechanism that could apply to general complex networks. Another feature that sets metabolic networks apart from other commonly studied networks is the dual nature of the network nodes: these correspond to either reactions or to metabolites, with very different characteristics. In particular, the degree distribution of the reaction nodes has *no* fat tail at all.

Further structural features like clustering or the small world property, as found in different natural and artificial networks, are also seen in metabolic networks [13, 1]. However the fact that metabolites fall into different biochemical categories (currency metabolites being one of these) means that many different treatments of the network are possible to reach a graph-based representation. Depending on the treatment, the conclusions for the structural properties can be different. In particular, in the simplest treatment which ignores metabolite categories [12], one has high clustering and the small world property, but both of these features go away with the more sophisticated treatment that handles currency metabolites separately [19, 20]. Thus again metabolic network structural properties are not those of the standard complex network picture, in spite of many early claims.

Clearly metabolism is complex, sufficiently to include numerous subtleties that set it apart from other network based systems. Nevertheless, one is left with the problem of explaining the salient features of these networks. At present, the power law distribution of the metabolite degrees remains unjustified, at least at a quantitative level. The situation seems better for the other structural properties: as argued by studying *in silico* genome-scale metabolic models [16], there is a good chance that the biochemical and functional constraints underlying cellular metabolism constrain the architecture of any metabolic network to have the characteristics found experimentally. As a cautionary note, this conclusion does not exclude the possibility that other forces such as robustness, evolutionary innovation *etc.* also shape to some extent structural properties of metabolic networks.

Acknowledgments

AS acknowledges support from CNRS GDRE513 and OCM from the Metacoli grant ANR-08-SYSC-011. The LPTMS is an Unité de Recherche de l'Université Paris-Sud associée au CNRS.

References

[1] Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 2004, **5**:101–113.

[2] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 1999, **402**:C47–C52.

[3] Alon U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall 2007.

[4] Wagner A. *Robustness and Evolvability in Living Systems*. Princeton University Press 2005.

[5] Bornholdt S, Schuster HG. *Handbook of Graphs and Networks: from the Genome to the Internet*. Wiley-VCH 2003.

[6] Képès F. *Biological Networks*. World Scientific 2007.

[7] Palsson BO. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press 2006.

[8] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, **28**:27–30.

[9] Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S. The Ecocyc Database. *Nucleic Acid Res* 2002, **30**:56–58.

[10] Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* 2008, **26**(6):659–667.

[11] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000, **407**:651–654.

[12] Wagner A, Fell DA. The small world inside large metabolic networks. *Proc R Soc Lond B* 2001, **268**:1803–1810.

[13] Albert R, Barabasi AL. Statistical mechanics of complex networks. *Rev Mod Phys* 2002, **74**:47–97.

[14] Tanaka R. Scale-rich metabolic networks. *Phys Rev Lett* 2005, **94**(16):168101.

[15] Tanaka R, Csete M, Doyle J. Highly optimised global organisation of metabolic networks. *Syst Biol* 2005, **152**(4):179–184.

[16] Samal A, Martin OC. Randomizing genome-scale metabolic networks. arXiv qbio:**1012.1473**.

[17] Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. *Nature* 1998, **393**:440–442.

[18] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002, **297**:1551–1555.

[19] Arita M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* 2004, **101**:1543–1547.

[20] Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 2003, **19**:1423–1430.

[21] Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000, **406**:378–382.

[22] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001, **411**:41–42.

[23] Varma A, Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology* 1994, **12**:994–998.

[24] Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004, **2**(11):886–897.

[25] Mahadevan R, Palsson BO. Properties of Metabolic Networks: Structure vs. Function. *Biophys J* 2005, **88**:L7–L9.

[26] Samal A, Singh S, Giri V, Krishna S, Raghuram N, Jain S. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* 2006, **7**:118.

[27] Harary F. *Graph Theory*. Addison-Wesley Publishing Company 1969.

[28] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the web. *Computer Networks* 2000, **33**:309–320.

[29] Csete M, Doyle J. Bow ties, metabolism and disease. *Trends Biotechnol* 2004, **22**:446–450.

[30] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science* 2002, **296**:910–913.

[31] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 2002, **298**:824–827.

[32] Samal A, Rodrigues JFM, Jost J, Martin OC, Wagner A. Genotype networks in metabolic reaction spaces. *BMC Syst Biol* 2010, **4**:30.

[33] Wagner A. *Evolutionary Genomics and Proteomics*, Sinauer Associates Inc., Sunderland, MA 2007 chap. Gene networks and natural selection.

[34] Papp B, Teusink B, Notebaart RA. A critical view of metabolic network adaptations. *HFSP J* 2009, **3**:24–35.

Synchronised ribosomes

Hongjun Gao[1], Patrick Amar[2] and Vic Norris[1]

[1] AMMIS Laboratory, EA 3829, Department of Biology, University of Rouen,
F-76821, Mont Saint Aignan, France

[2] L.R.I., University of Paris Sud Orsay & CNRS UMR 8623, 15 avenue
George Clémenceau, F-91405 Orsay Cedex, France

Abstract

The discoveries of periodic motions of yeast cells of around a kHz and of sonic communication between bacteria open up a new and exciting field. It is now apparent that prokaryotic, like eukaryotic, cells are highly structured and very crowded. In general, transcription and translation are coupled in bacteria and the density of ribosomes is high. When synthesizing proteins, these ribosomes go through the same limited cycle of movements. Here, we propose that these movements become synchronised to create periodic oscillations of the entire cell and speculate that this synchrony plays a role in determining the phenotype and in communication.

1 Introduction

Two pendulum clocks of identical frequency mounted on a common wall tend to synchronise such that they swing in opposite directions. Such coupled oscillations were first described by Huygens in 1665 and are still the subject of study [1, 2]. Cells are full of macromolecules going through limited cycles of conformational changes and these macromolecules are often either in contact with one another or very close in a crowded cytoplasm and membrane. The question therefore arises as to whether the periodic movements of these macromolecules become coupled. Evidence has been obtained for nanomechanical movements in eukaryotes [3] and for sonic communication in bacteria [4]. Do these phenomena result from coupled oscillations and, if so, of what?

The transcriptional and translational machinery and, in particular, ribosomes, make up the bulk of the mass of bacteria such as Escherichia coli during growth in rich media [5, 6]. Much of this machinery is organised into hyperstructures, extended macromolecular assemblies, that include hyperstructures in which transcription and translation are physically linked [7, 8]. We propose here that the coupled oscillations of ribosomes are important in the physiology of bacterial cells and we suggest how our proposal might be investigated.

2 The model

The movements of ribosomes become synchronised within the bacterial cell when ribosomes are translating the mRNAs within a transcription-translation hyperstructure. This occurs because these movements are coupled via the mRNA that connects them and via the changes in water structure that result from their movements. Consequently, the hyperstructure itself pulsates. Such hyperstructure pulsations themselves become synchronised and the result is a bacterium that oscillates. These oscillations may constitute the basis for physical communication between bacteria and between bacteria and eukaryotic cells.

3 The evidence

3.1 Ribosomes go through a limited cycle of movements

Each ribosome has three sites: the A site, the P site, and the E site. The A site is where the aminoacyl tRNA enters (except for the first aminoacyl tRNA, fMet-tRNAfMet, which enters at the P site). The P site is where the peptidyl tRNA is formed. The E site is where the now uncharged tRNA leaves after it has given its amino acid to the growing peptide chain. Elongation of the polypeptide chain starts when the fmet-tRNA enters the P site, causing a conformational change that opens the A site to allow the new aminoacyl-tRNA to bind. This binding is facilitated by elongation factor-Tu (EF-Tu), a small GTPase. At this stage, the P site contains the start of the peptide chain whilst the A site has the next amino acid to be added to this chain. Then the growing polypeptide connected to the tRNA in the P site is detached from the tRNA in the P site and a peptide bond is formed between the last amino acid of the peptide chain and the amino acid still attached to the tRNA in the A site. At this stage, the A site has the newly formed peptide, while the P site has an uncharged tRNA. In the final stage of elongation, translocation, the ribosome moves 3 nucleotides towards the 3' end of mRNA. Since tRNAs are linked to mRNA by codon-anticodon base-pairing, tRNAs move relative to the ribosome taking the nascent polypeptide from the A site to the P site and moving the uncharged tRNA to the E exit site.

3.2 Ribosomes are close to one another

The pioneering microscopy of Miller and coworkers showed that translation is completely coupled to transcription in *E. coli* [9, 10]. Calculations showed that when the *E. coli lac operon* is induced in an exponentially growing culture, where it is present at more than one copy, the numbers of transcripts of *lacZ* per cell are 32 full length, 32 decaying and 38 nascent [11]; *lacZ* is 3063 nucleotides long and under these conditions the RNA polymerases and ribosomes are 135 and 110 nucleotides apart respectively. This works out to around 300 ribosomes [12]. Ribosomes can be as little as 35 nucleotides apart along the mRNA but the question is, how close are they in 3-D? Recently, cryoelectron tomography and a template-matching approach have been used to localise ribosomes in vitrified bacterial translation extracts and in lysates of active *E. coli* spheroplasts [13]; neighbouring ribosomes in polysomes were densely packed and had a pseudo-helical organization along the mRNA. Nuclear magnetic resonance spectroscopy has shown that the NusG protein (which binds RNA polymerase) binds to a protein identical to the ribosomal protein S10 to link, it is proposed, transcription and translation in *E. coli* [14]. Also recently, fluorescence microscopy of *E. coli* and *Caulobacter crescentus* has shown that mRNA is colocalised with the gene that encodes it, consistent with the widespread existence of transcription-translation hyperstructures in which translating ribosomes are cheek-by-jowl [15].

3.3 Coupling between ribosomes

How might ribosomes be coupled so as to move in synchrony à la Huygens? Firstly, ribosomes are connected by a common mRNA. The average numbers of ribosomes within *E. coli* polysomes were estimated as 4, 8, and 11 depending on the length of the mRNA [13]. Naively, one would expect the displacement of the mRNA through one ribosome to influence the displacement of that same mRNA through the neighbouring ribosome. Secondly, there may be direct physical connections between ribosomes: "An additional density, presumably involving the L1 stalk region, appears to bridge the gap between two ribosomal neighbors in polysomes. This contact may induce a preferential orientation of polysomal neighbors" [13]. Thirdly, elongation factor EF-Tu, which is essential for protein synthesis, has long been suspected to be a bacterial actin [16]. Ground-breaking work has shown that EF-Tu forms cytoskeletal filaments with which the ribosomes are associated [17, 18], moreover, these filaments are dynamic and are associated with another cytoskeletal protein, the actin-like MreB [19]. Hence, an EF-Tu cytoskeleton (or, more exactly, an *enzoskeleton* [20]) could couple both the movements of actively translating ribosomes within the same hyperstructure and the movements of different hyperstructures. Fourthly, the insertion of nascent proteins into the cytoplasmic membrane during transertion increases its microviscosity [21]. The changes in the state of the phospholipids as proteins are inserted could couple the way they are inserted. Fifthly, there is water (see below).

3.4 Water structures change during translation

Water is believed by some specialists to form more than one structure within cells [22, 23, 24]. In the two-state model, water is considered as a temperature-dependent, fluctuating equilibrium between two types of local structures: low density water (LDW) and high density water (HDW); this equilibrium is driven by incompatible requirements for minimizing enthalpy via strong near-tetrahedral hydrogen-bonds to give LDW and for maximizing entropy via non-directional H-bonds to give HDW. LDW with its strong, straight H-bonds has a density of 0.91 g/ml whilst HDW with its bent, weak H-bonds has a density of 1.2 g/ml and, since their hydrogen bond strengths are different, microdomains of LDW and HDW differ in all physical and chemical properties [24]. It has been proposed that the folding and functioning of enzymes results from these properties and that catalysis by enzymes entails the redistribution of LDW and HDW [25]: "There is a crucial functional connection between the force that drives folding of an enzyme and reactions that it catalyses. When water can move to abolish osmotic pressure gradients created by selective uptake of solutes into HDW or LDW, it does so with some decrease in the partition coefficients of the reactants. When water is prevented from moving, partition coefficients are unchanged, increased or transiently inverted." Assuming that this holds for translating ribosomes, water structures should be moving as ATP and GTP are hydrolysed and as the peptide bond is made. Such movements in water structures might couple the movements of ribosomes to one another. Indeed, the synchronisation of ribosome movements and related changes in water structure might be expected to have co-evolved so as to minimise the energy needed for protein synthesis.

3.5 Eukaryotic cells exhibit coherent vibrations

In vivo, Saccharomyces cerevisiae (baker's yeast) has periodic motions in the range of 0.8 to 1.6 kHz with amplitudes of approximately 3 nm as measured using an atomic force microscope [3]. The magnitude of the forces observed (10 nN) led the authors to suggest that "concerted nanomechanical activity is operative in the cell" whilst the calculated activation energy of 58 kJ/mol was interpreted as implicating molecular motors such as kinesin, dynein, and myosin in the motions. Another, perhaps complementary, interpretation is that ribosomes were responsible.

3.6 Bacteria communicate sonically

In a bold and original series of experiments, it was found that a variety of bacteria can emit a physical signal that helps *Bacillus carbophilus* grow on agar containing erythromycin or streptomycin or high concentrations of salt [4, 26]. This signal was taken to be sonic since it could be transmitted through sealed Petri dishes and through an iron barrier [27] (although there are other possibilities [28]). Moreover, continuous single sine sound waves produced by a speaker at frequencies of 6-10, 18-22, and 28-38 kHz promoted colony formation by *B. carboniphilus* despite the stress of high KCl concentration and high temperature. Sound waves emitted from *Bacillus subtilis* at frequencies between 8 and 43 kHz with broad peaks at approximately 8.5, 19, 29, and 37 kHz could also be detected directly leading to the proposal that these sound waves function as a growth-regulatory signal between cells [29].

4 Predictions

1. Atomic Force Microscopy should confirm that individual bacteria also oscillate on the nm scale at sonic frequencies.
2. Sonic oscillations should be detectable in vitro during transcription and translation in extracts of bacteria [30].
3. The frequency of oscillations should be altered in vivo when drugs are added to inhibit translation, when ribosomes are altered by mutations to the translational machinery (rRNA, rproteins, tRNA, tRNA synthetases, EF-Tu etc.), and following a major change to the codon composition of the bacterium (for example, so as to slow translation). The T7 system might be used in the last case since this allows a gene (of chosen codon composition) to be expressed whilst the native genes are silenced [31].
4. Bacteria are in different states in different parts of a colony and these bacteria should have different periods of oscillation.
5. The generation of synchronised movements by ribosomes (and its possible consequences on cellular organisation) should be observed *in silico* in stochastic automata such as HSIM [32].

5 Discussion

The existence of periodic motions in the sonic range in bacteria raises interesting questions [28]. What generates such motions? Is it generated by contractile proteins such as myosin [3] (which has yet to be found in bacteria despite a search for it [33]), or by ribosomes (as proposed here), or by an ensemble of different enzymes in the cell (DNA gyrase, helicases, polymerases etc.)? What, if any, is the function of such motions? Is it in communication between cells? Does it allow communication between the same species of bacteria in a colony or even between different species (including between bacteria and yeast) in a mixed biofilm? Is it important in the determination of the phenotype and, in particular, the regulation of the cell cycle - for example, by contributing to a dialogue between hyperstructures [7]? If it has a function, how might periodic motions be altered so as to alter, for example, pathogenesis? Answering these questions will require bold, interdisciplinary collaborations.

References

[1] Strogatz, S.H. and I. Stewart, Coupled oscillators and biological synchronization. *Scientific American, 1993.* **269**:p. 102-109.

[2] Czolczynskia, K., P. Perlikowskia, A. Stefanskia, and T. Kapitaniak, Clustering and synchronization of n Huygens' clocks. *Physica A: Statistical Mechanics and its Applications, 2009.* **388**(24):p. 5013-5023.

[3] Pelling, A.E., S. Sehati, E.B. Gralla, J.S. Valentine, and J.K. Gimzewski, Local nanomechanical motion of the cell wall of Saccharomyces cerevisiae. *Science, 2004.* **305**:p. 1147-1150.

[4] Matsuhashi, M., A.N. Pankrushina, K. Endoh, H. Watanabe, Y. Mano, M. Hyodo, T. Fujita, K. Kunugita, T. Kaneko, and S. Otani, Studies on carbon material requirements for bacterial proliferation and spore germination under stress conditions: a new mechanism involving transmission of physical signals. *Journal of Bacteriology, 1995.* **177**:p. 688-693.

[5] Nomura, M., Regulation of ribosome biosynthesis in Escherichia coli and Saccharomyces cerevisiae: diversity and common principles. *J Bacteriol, 1999.* **181**(22):p. 6857-64.

[6] Bremer, H. and P. Dennis, Feedback control of ribosome function in Escherichia coli. *Biochimie, 2008.* **90**(3):p. 493-9.

[7] Norris, V., T.D. Blaauwen, R.H. Doi, R.M. Harshey, L. Janniere, A. Jimenez-Sanchez, D.J. Jin, P.A. Levin, E. Mileykovskaya, A. Minsky, G. Misevic, C. Ripoll, M. Saier Jnr., K. Skarstad, and M. Thellier, Toward a Hyperstructure Taxonomy. *Annu Rev Microbiol, 2007.* **61**:p. 309-329.

[8] Norris, V., Speculations on the initiation of chromosome replication in Escherichia coli: the dualism hypothesis. *Medical Hypotheses, 2011* in press.

[9] Miller, O.L., Jr., B.A. Hamkalo, and C.A. Thomas, Jr., Visualization of bacterial genes in action. *Science, 1970.* **169** (943):p. 392-5.

[10] French, S.L. and O.L. Miller, Jr., Transcription mapping of the Escherichia coli chromosome by electron microscopy. *J Bacteriol, 1989.* **171**(8):p. 4207-16.

[11] Kennell, D. and H. Riezman, Transcription and translation frequencies of the Escherichia coli lac operon. *Journal of Molecular Biology, 1977.* **114**:p. 1-21.

[12] Norris, V., T. Onoda, H. Pollaert, and G. Grehan, The mechanical origins of life. *BioSystems, 1999.* **49**:p. 71-78.

[13] Brandt, F., S.A. Etchells, J.O. Ortiz, A.H. Elcock, F.U. Hartl, and W. Baumeister, The native 3D organization of bacterial polysomes. *Cell, 2009.* **136**(2):p. 261-71.

[14] Burmann, B.M., K. Schweimer, X. Luo, M.C. Wahl, B.L. Stitt, M.E. Gottesman, and P. Rosch, A NusE:NusG complex links transcription and translation. *Science, 2010.* **328**(5977):p. 501-4.

[15] Llopis, P.M., A.F. Jackson, O. Sliusarenko, I. Surovtsev, J. Heinritz, T. Emonet, and C. Jacobs-Wagner, Spatial organization of the flow of genetic information in bacteria. *Nature, 2010.* **466**(7302):p. 77-81.

[16] Beck, B.D., Polymerization of the bacterial elongation factor for protein synthesis, EF-Tu. *European Journal of Biochemistry, 1979.* **97**:p. 495-502.

[17] Mayer, F., Cytoskeletons in prokaryotes. *Cell Biology International, 2003.* **27**:p. 429-438.

[18] Helms, M.K., G. Marriott, W.H. Sawyer, and D.M. Jameson, Dynamics and morphology of the in vitro polymeric form of elongation factor Tu from Escherichia coli. *Biochimica Biophysica Acta, 1996.* **1291**: p. 122-130.

[19] Defeu Soufo, H.J., C. Reimold, U. Linne, T. Knust, J. Gescher, and P.L. Graumann, Bacterial translation elongation factor EF-Tu interacts and colocalizes with actin-like MreB protein. *Proc Natl Acad Sci U S A, 2010.* **107**(7):p. 3163-8.

[20] Norris, V., G. Turnock, and D. Sigee, The Escherichia coli enzoskeleton. *Molecular Microbiology, 1996.* **19**:p. 197-204.

[21] Binenbaum, Z., A.H. Parola, A. Zaritsky, and I. Fishov, Transcription- and translation-dependent changes in membrane dynamics in bacteria: testing the transertion model for domain formation. *Molecular Microbiology, 1999.* **32**:p. 1173-1182.

[22] Robinson, G.W. and C.H. Cho, Role of hydration water in protein unfolding. *Biophysical Journal, 1999.* **77**:p. 3311-3318.

[23] Wichmann, C., P.T. Naumann, O. Spangenberg, M. Konrad, F. Mayer, and M. Hoppert, Liposomes for microcompartmentation of enzymes and their influence on catalytic activity. *Biochemical Biophysical Research Communications, 2003.* **310**:p. 1104-1110.

[24] Wiggins, P., Life depends upon two kinds of water. *PLoS One, 2008.* **3**(1):p. e1406.

[25] Wiggins, P., Enzymes and surface water. *Water, 2009.* **1**:p. 42-51.

[26] Matsuhashi, M., A. Shindo, H. Ohshima, M. Tobi, S. Endo, H. Watanabe, K. Endoh, and A.N. Pankrushina, Cellular signals regulating antibiotic sensitivities of bacteria. *Microbial Drug Resistance-Mechanisms Epidemiology and Disease, 1996.* **2**: p. 91-93.

[27] Matsuhashi, M., A.N. Pankrushina, K. Endoh, H. Watanabe, H. Ohshima, M. Tobi, S. Endo, Y. Mano, M. Hyodo, T. Kaneko, S. Otani, and S. Yoshimura, Bacillus carboniphilus cells respond to growth-promoting physical signals from cells of homologous and heterologous bacteria. *Journal of General and Applied Microbiology, 1996.* **42**:p. 315-323.

[28] Norris, V. and G.J. Hyland, Do bacteria "sing"? *Molecular Microbiology, 1997.* **24**:p. 879-880.

[29] Matsuhashi, M., A.N. Pankrushina, S. Takeuchi, H. Ohshima, H. Miyoi, K. Endoh, K. Murayama, H. Watanabe, S. Endo, M. Tobi, Y. Mano, M. Hyodo, T. Kobayashi, T. Kaneko, S. Otani, S. Yoshimura, A. Harata, and T. Sawada, Production of sound waves by bacterial cells and the response of bacterial cells to sound. *J Gen Appl Microbiol, 1998.* **44**(1):p. 49-55.

[30] Zubay, G., In vitro synthesis of protein in microbial systems. *Annual Review of Genetics, 1973.* **7**:p. 267-298.

[31] Studier, F.W., A.H. Rosenberg, J.J. Dunn, and J.W. Dubondorff, Use of T7 RNA polymerase to direct expression of cloned genes. *Methods in Enzymology, 1990.* **185**:p. 60-89.

[32] Amar, P., G. Legent, M. Thellier, C. Ripoll, G. Bernot, T. Nystrom, M.H. Saier, Jr., and V. Norris, A stochastic automaton shows how enzyme assemblies may contribute to metabolic efficiency. *BMC Syst Biol, 2008.* **2**:p. 27.

[33] Casaregola, S., V. Norris, M. Goldberg, and I.B. Holland, Identification of a 180 kDa protein from E. coli related to a yeast myosin heavy chain. *Molecular Microbiology, 1990.* **4**:p. 505-511.

A chemistry-based strategy for the development of safe GMO's

Piet Herdewijn[1,2] and Philippe Marlière[3]

[1] Laboratory of Medicinal Chemistry, Rega Institute for Medical Research,
Katholieke Universiteit Leuven, B-3000 Leuven, Belgium
[2] iSBB, Université d'Evry-Val-D'Essonne, Paris Sud, France
[3] Isthmus Sarl, F-75015 Paris, France

Abstract

The aim of the project is to propagate artificial nucleic acids in a microbial cell, for which an uptake system and a processing apparatus for exogenous activated precursors need to be developed. The key accomplishment would be the expression of an artificial aptazyme in the cell, catalyzing an essential metabolic reaction. Such reprogrammed microorganism could become a new instrument for avoiding genetic pollution when performing experiments in synthetic biology. The seminar will primarily deal with discussing the tools that need to be developed to reach this ambitious goal.

1 Introduction

The first gene was synthesized in 1971 by the group of H.G. Khorana [1]. In contrast to progress in gene-analysis, progress in the synthesis of genes and genomes has been very slow. 40 Years have elapsed between the synthesis of the first gene and the synthesis of the DNA of a Mycoplasma. The chemo-enzymatic method used for this synthesis is based on technologies that have been developed in the previous century. All technologies for genome synthesis are available, which means that more and more examples will show up of laboratories that will synthesize the DNA of always more complex organisms. In line with this, genetic reprogramming of organism and directing their evolution is also within reach of most microbiological laboratories because it is based on established technologies such as the use of synthetic oligonucleotides, directed mutagenesis and amplification techniques. The observation that the genome of microorganism can be synthesized and manipulated is considered by some organizations as a threat to the natural ecosystem and it asks for the development of radically new approaches to avoid genetic pollution when designing and using newly engineered microorganism. As the use of such microorganisms for the production of food, drugs, energy and chemicals has become inevitable, our choices are limited.

2 Discussion

Synthetic biology has been defined as a new science that is focused on the re- engineering of natural biology, based on circuit design, bioinformatics and systems biology. We propose that we could preserve the natural ecosystem in a better way than proposed before, by developing new biologicals, based on synthetic (bio)chemicals and evolutionary enzymes that need to be implemented in vivo, and that the resulting microorganism should be used as chassis to perform 'classical' synthetic biology experiments. Therefore, we would like to develop information systems that are synthetically and functionally isolated within a cell and that are not able to communicate its information with natural nucleic acids, which brings us to the principle of orthogonality (in this case, orthogonality is, in first instance, defined as lack of communication between information systems). Indeed, Nature works with only two types of nucleic acids (DNA and RNA) and only one type of building blocks (nucleoside triphosphates). A third type of nucleic acid (XNA) should be selected that could be synthesized within a cell starting from non-natural precursors. This XNA should be replicated and propagated in an autonomous way (without making use of cellular enzymes). It should form its own genetic enclave, not able to infiltrate the genome of the cell and vice versa. This orthogenetic system could form the foundation of a xenobiology.

To realize a xenobiology platform in its most rudimentary form, we have started four different scientific projects which need to be worked out in parallel and in a coherent way, as they need to be integrated. The first is the selection and in vitro replication of a third type of nucleic acids (XNA) with an alternative backbone motif. The second is the development of an uptake system for the precursors of XNA in reprogrammed host cells. The third project is to evolve a polymerase that is able to propagate XNA, but not DNA and RNA. The fourth initiative is to design an aptazyme which contains information for one selectable function which is indispensable for the survival of the host cell. The three cornerstones of a living system, information, metabolism and catalysis, need to be implemented in a new biological network.

For the selection of XNA, initially we prefer to develop sugar modified nucleic acids, as a reliable base pairing system for communication with natural DNA is important in the first stage of the project i.e. when there is still a need to use information from a natural cell to develop the necessary tools (polymerases). Chemically, the first XNA should resemble the natural polymers but differ structurally enough from DNA and RNA so that its function and biosynthesis can be uncoupled from the natural system. The first examples which will be studied are HNA and CeNA. These are chemically and enzymatically stable information systems, able to communicate its information with natural nucleic acids. dsHNA [2] and dsCeNA [3] form a helical structure which slightly deviate from the classical type A and type B. The presence of an additional carbon atom in the sugar ring could result in a steric advantage to evolve to an orthogonal system. The analysis of the helical parameters of dsHNA and dsCeNA, which forms the basis of this selection, is based on x-ray studies. An important difference between HNA and CeNA is that the latter information system is expected to be conformationally more flexible, would this be important to function in biology. This flexibility has been demonstrated as well by modeling experiments, NMR and x-ray analysis. Further selection criteria for an orthogonal information system are based on the helical parameters of its duplex structure. We have analyzed helical parameters such as slide, shift or twist in function of the helicalization process itself, which is a selection criterium for orthogonality [4]. One of the interesting new examples coming out of this study is xylo-DNA, which is as well a very flexible information system as structurally orthogonal to the natural nucleic acids [5].

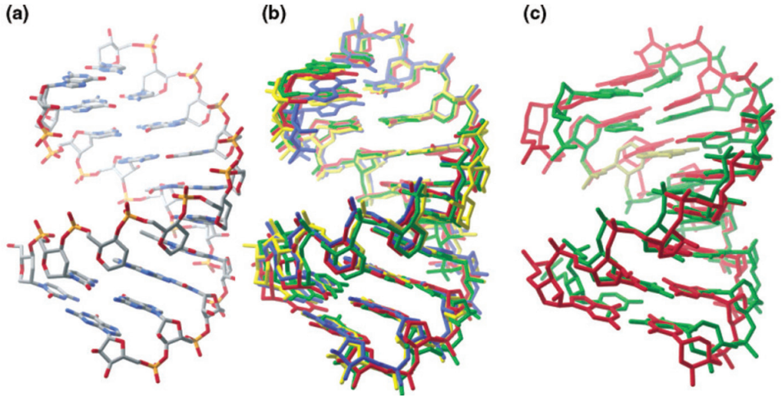

Figure 1: Structural overview and comparison to NMR solution structure of a RNA-HNA duplex [6] h(3'GCGATGCG5') r(5'CGCUACGC3')

The active site of a polymerase positions the triphosphate moiety of an incoming nucleotide in line with the attacking secondary hydroxyl group at the end of the growing nucleic acid chain. A wide variety of sugar modified nucleotides are accepted as substrate by several polymerases. Even pyranosyl-type nucleotides can be incorporated in DNA in an enzymatic way, although they are not prime candidates as orthogonal information system [8].

When using HNA as prototype, we have demonstrated that type B DNA polymerase and terminal transferase shows DNA dependent HNA polymerase activity and that DNA polymerase I (E.Coli) demonstrates as well HNA de-

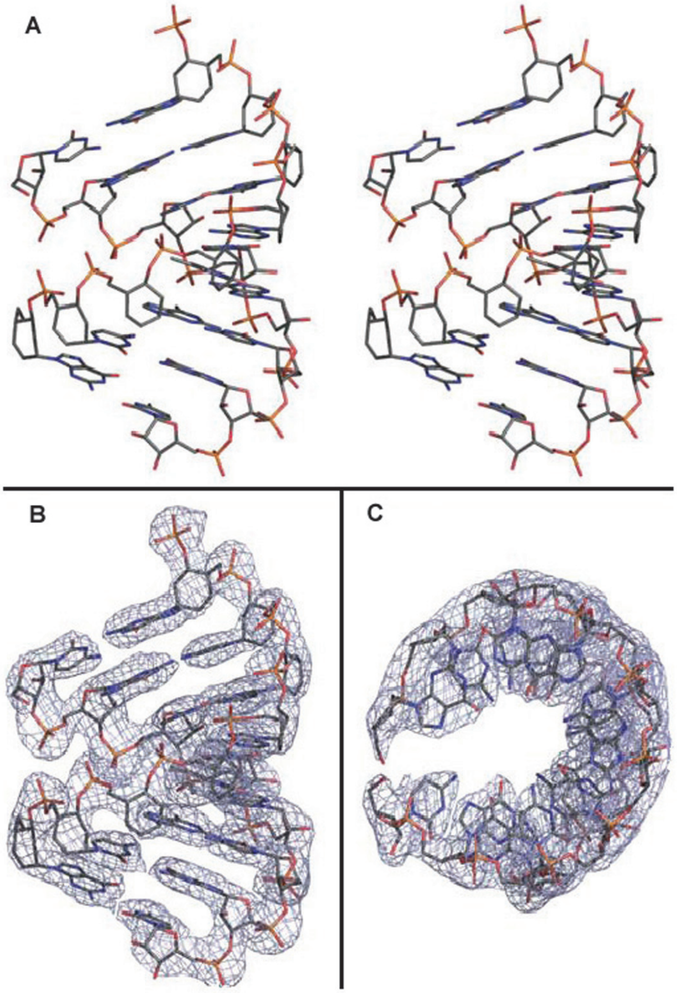

Figure 2: Side (top) and top (bottom) view of the molecular structure for the CeNA:RNA hybrid [7].

pendent DNA polymerase activity as initial HNA dependent HNA polymerase activity [9]. DNA polymerase B and HIV reverse transcriptase shows DNA dependent CeNA polymerase and CeNA dependent DNA polymerase activity. The kinetic parameters for incorporation of one or two modified nucleotides are generally very similar to the parameters found for natural nucleotides. However, polymerases generally hold after the incorporation of two to three modified nucleotides. Therefore, to facilitate development of XNA-based replicons and episomes, it will be necessary to develop mutant polymerases able to synthesize new polymers of gene length and to use them as template for the propagation of information. For that reason, a program for the directed evolution of polymerase for the replication of XNA has been established, based on the compartmentalized self-replication process. This process is based on a selection system of mutant libraries where a polymerase replicates its own encoding gene and the compartmentalization is responsible for linking phenotype to genotype. An in vivo evaluation system is based on the availability of prototrophic transformants of an E.coli strain lacking an active gene for thymidylate synthase [9]. The correct copying of a message encoded in XNA is mandatory for survival of the microorganism. Another very helpful enzyme for diversifying nucleic acids in vivo would be an XNA ligase i.e. a ligase that accepts XNA as substrate to produce long stretches of XNA for making whole genes and episomes.

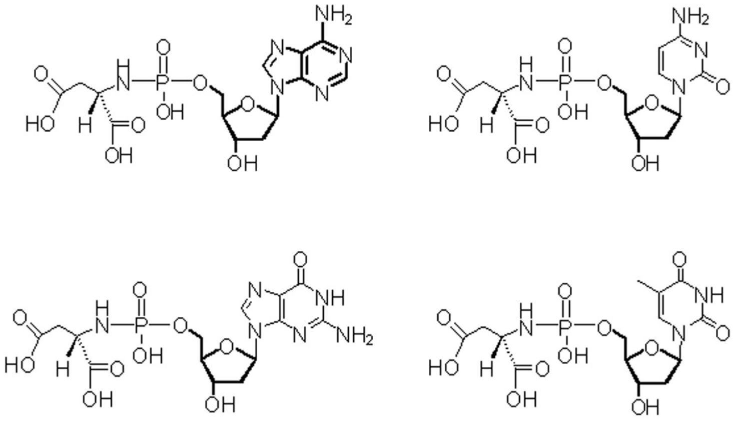

Figure 3: Structures of amino acid 2'-deoxynucleoside-5'-monophosphate derivatives used in the HIV RT incorporation assays.

Moreover, it will be difficult to install additional nucleoside triphosphates in a cell without interfering with DNA and RNA metabolism, cell energy supply via respiration or substrate-level phosphorylation. Indeed, energy storage and genetic functions in a cell, both rely on phosphoanhydride formation and pyrophosphate (phosphate) release. The use of alternative leaving groups could result in an additional level of synthetic and functional isolation, distinct from canonical nucleic acids without having to physically separate precursors from XNA from these of DNA and RNA. The ideal properties of a leaving group to function for the enzymatic synthesis of XNA are: soluble in water and chemically not too unstable, to be accommodated in the active site of polymerases and to react as substrate for the enzyme, to undergo productive elongation, the chemical choice of the leaving group should be mechanism-based, the leaving group should be actively degraded or recycled to common metabolites so as to render the polymerization irreversible. Based on the model structure of the accommodation of dATP in the active site of polymerases and on the knowledge of the mechanism of the polymerization process itself, we have predicted and evaluated new series of leaving groups for the enzymatic synthesis of DNA using reverse transcriptase as catalyst [10]. We observed that aspartate could function as alternative for pyrophosphate in the polymerization of DNA. Watson-Crick rules and Michaelis-Menten kinetics are respected and the process is stereospecific. Initial concerns about incorporation kinetics (the Vmax is similar to that of dATP, the Km is lower) and the observed stalling after incorporation of two to three nucleotides, could be overcome by selecting other chemical entities as leaving group such as phosphono-alanin and iminodiacetate. Molecular modeling demonstrates that the amino acids located in the active site of the polymerase and involved in the binding of these leaving groups are very conserved. An important issue is that those leaving groups are potential metabolically accessible. Further optimization of the kinetic properties and selectivity of the polymerases will be done by in vivo evolution.

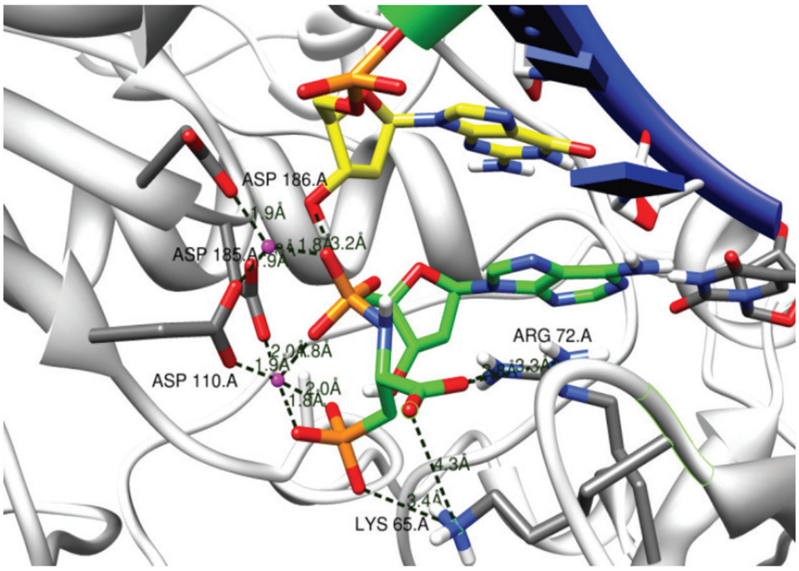

Figure 4: Model structures of 3-phosphono-L-Ala-dAMP in the RT dNTP pocket. The residues Asp 110, 185 and 186 anchor the 2 Mg2+ ions. Possible stabilization of the carboxyl function and the phosphonate function in the leaving group by Arg 72 and Lys 65 is indicated [11].

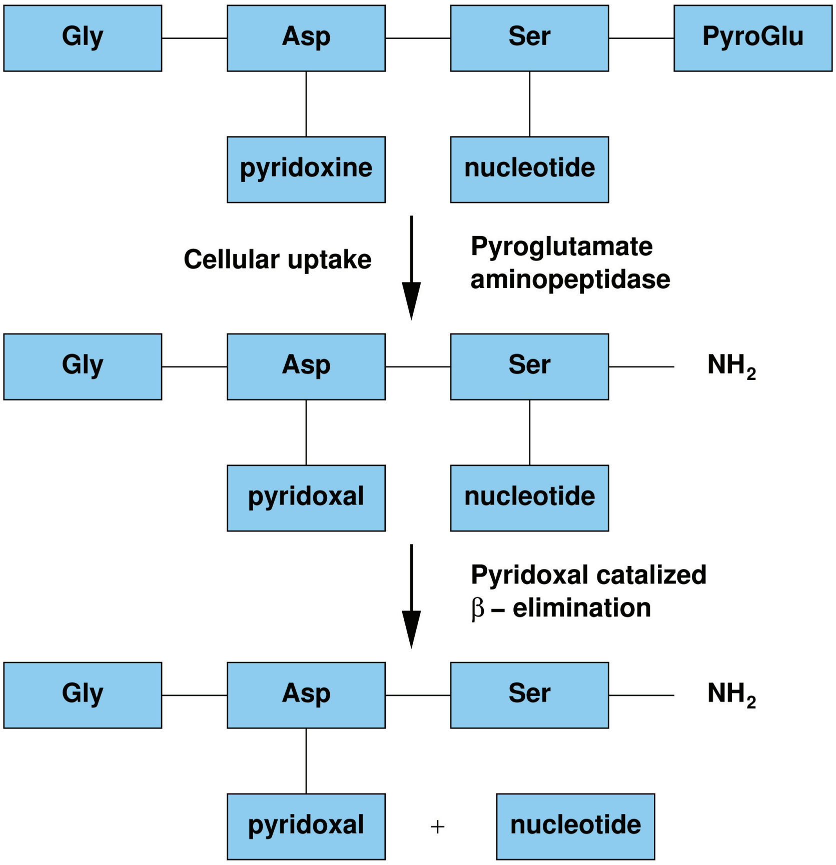

Figure 5: Example of a pyridine-peptide delivery system as potential substrate for oligopeptide permease.

It will be important that the nucleotide precursors can be taken up by the host cell using an active uptake process. A delivery system for nucleotides could make use of oligopeptide permeases, the ligand-binding site shows a broad substrate specificity and accepts chemical groups of large diversity (di- to pentapeptides). Those permeases are part of a larger group of transport systems i.e. ATP-binding cassette transporters. We have obtained a first prototype for delivery of nucleotides that could function as substrate for the transporter. This consists of a pyroglutamyl protected tripeptide, a lateral pyridoxal moiety and a nucleotide loaded on a serine residue. It is hypothesized that, following transport into the bacterial cytoplasm, the pyroglutamyl group could be deblocked by a specific aminopeptidase liberating a free amino group that could be involved in the catalytic process to deliver the laterally attached nucleotide. Intracellular delivery of the nucleotide could be accomplished by a pyridoxal-catalyzed elimination of the nucleotide, bound to the free amino group of the serine residue, via formation of a Schiff base [12].

Obtaining metabolic dependency via the transcription of DNA into a catalytic XNA (xenozyme), making use of the above described DNA dependent XNA polymerase could be realized in several ways. For example, this xenozyme could either catalyse an essential metabolic reaction, either catalyse the synthesis of an essential cofactor, or the catalytic activity of the xenozyme could become dependent on the availability of a synthetic cofactor. We have started the first efforts to generate a ribozyme that catalyzes an essential reaction for the synthesis of amino acids, making use of a synthetic cofactor. The reaction itself is an aldol condensation reaction , which is a carbon- carbon forming reaction, which creates beta-hydroxy-carbonyl compounds. In nature aldolases catalyses this reaction through an imine mechanism. The selected aldol reaction involves glycinate Shiff base formation with the help of a synthetic cofactor (salicylaldehyde), followed by reaction with an aldehyde in the presence of aluminium trichloride, giving rise to threonine- like compounds. For selecting the catalytic RNA, a highly diverse RNA library is used following iterative cycles of in vitro selection (Systematic Evolution of Ligands by Exponential enrichment).

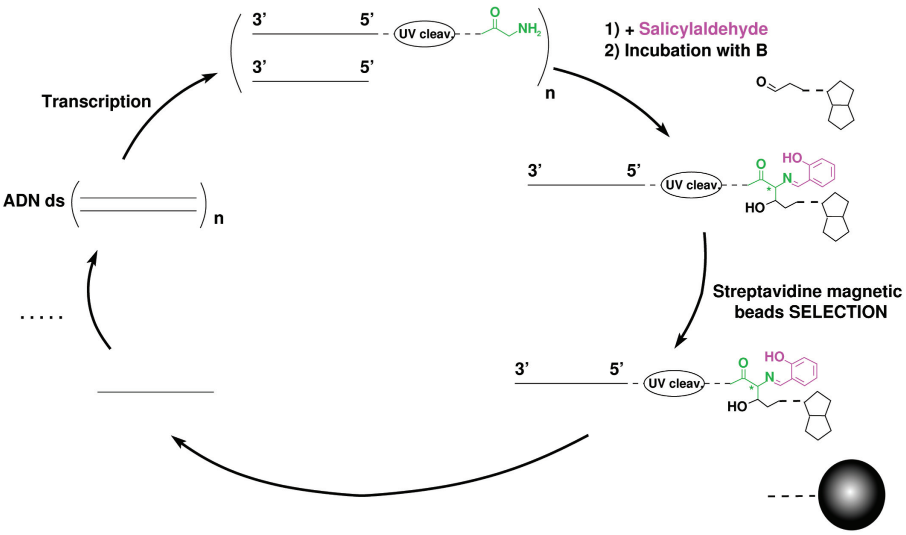

Figure 6: Systematic Evolution of Ligands by Exponential enrichment

3 Conclusion

Realization of a xenobiology in a microbial cell, for applications in energy, medicine, environment, food, requires a close collaboration between chemistry, biochemistry, biotechnology, genetica, microbiology. The final aim of this project is to have access to a safe level of informational transactions in engineered life forms.

References

[1] Khorana H.G. (1971) Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Pure Appl Chem.* **25**: 91 - 118.

[2] Hendrix C., I. Verheggen, H. Rosemeyer, F. Seela, A. Van Aerschot, and P. Herdewijn (1997) 1',5'-Anhydrohexitol oligonucleotides: synthesis, base pairing and recognition by regular oligodeoxyribonucleotides and oligoribonucleotides. *Chemistry Eur. J.* **3**: 110-120.

[3] Wang J., B. Verbeure, I. Luyten, E. Lescrinier, M. Froeyen, C. Hendrix, H. Rosemeyer, F. Seela, A. Van Aerschot and P. Herdewijn, (2000) Cyclohexene nucleic acids (CeNA): serum stable oligonucleotides that activate RNase H and increase duplex stability with complementary RNA. *J.Am.Chem.Soc.* **122**:8595-8602.

[4] Nauwelaerts K., E. Lescrinier and P. Herdewijn (2007) Structure of ?-homo DNA:RNA duplex and the function of twist and slide to catalogue nucleic acids duplexes. *Chem. Eur. J.* **13**: 90-98.

[5] Ramaswamy A., M. Froeyen, P. Herdewijn and A. Ceulemans (2010) The helical structure of xylose-DNA. *J. Am. Chem. Soc.* **132**:587-595.

[6] Maier T., I.Przylas, N. Strater, P. Herdewijn and W. Saenger (2005) Reinforced HNA backbone hydration in the x-ray crystal structure of a decameric HNA/RNA hybrid. *J. Am. Chem. Soc.* **127**: 2937-2943

[7] Ovaere M., P. Herdewijn and L. Van Meervelt 2011) The crystal structure of the CeNA:RNA hybrid ce(GCGYAGCG): r(CGCUACGC) *Chem. Eur. J.* (submitted)

[8] Renders M., M. Abramov, M. Froeyen and P. Herdewijn (2009) Polymerase-catalysed incorporation of glucose nucleotides into a DNA duplex. *Chem. Eur. J.* **15**: 5463-5470.

[9] Pochet S., P.A. Kaminski, A. Van Aerschot, P. Herdewijn and P. Marlière (2003) Replication of hexitol oligonucleteotides as a prelude to the propagation of a third type of nucleic acid *in vivo*. *Comptes Rendus Biologies* **326**: 1175-1184.

[10] Terrazas M., P. Marlière and P. Herdewijn (2008) Enzymatically catalyzed DNA synthesis using L-Asp-dGMP, L-Asp-dCMP and L-Asp-dTMP *Chem. Biodiv.* **5**: 31-39.

[11] Yang S., M. Froeyen, E. Lescrinier, P. Marlière and P. Herdewijn (2011) 3-Phosphono-L-alanine as Pyrophosphate Mimic for DNA Synthesis Using HIV-1 Reverse Transcriptase. *Org. Biomol. Chem.* **9**: 111-119.

[12] Marchand A., D. Marchand, R. Busson, P. Marlière, P. Herdewijn (2007) Synthesis of a Pyridoxine-Peptide Based Delivery System for Nucleotides. *Chem. Biodiv.* **4**: 1450-1465.

Evolution of spatially optimized gene networks

Thimo Rohlf[1,2,3], Ivan Junier[1,4] and François Képès[1]

[1] Epigenomics Project, iSSB, Genopole®, F-91034 Evry, France
[2] Max-Planck-Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany
[3] IZBI, University of Leipzig, D-04107 Leipzig, Germany
[4] ISC-PIF, F-75005 Paris, France

Abstract

Recent results from experiments, bioinformatics and theory indicate a strong impact of spatial genome organization on the machinery of gene regulation. One example is the periodic positioning of certain co-regulated genes on DNA, which is found both for prokaryotes and eukaryotes [12, 11]. We follow the hypothesis that distance minimization in 3D space under a solenoidal epi-organization of the chromosome leads to weight maximization of regulatory interactions. Using Boolean Threshold dynamics, we show that inhomogeneity in interactions (”weak” and ”strong” links) increases robustness of regulatory dynamics. Finally, we study the evolution of periodic organization under different mutation operators, and different types of selective constraints.

1 Introduction

The exploration of *epigenetic organization* in biological organisms is at the core of the post-genomic era. In itself, it is a typical Systems Biology endeavor, aiming to disentangle the complex interplay of different mechanisms at different spatial and temporal scales - beyond sequence information in DNA and Proteins - at work to establish information processing in living beings.

One particular example is the optimization of DNA transcription into RNA, and its regulation by transcription factors (TFs). The ensemble of TFs in organisms build up complex gene regulatory networks (GRN). Yet, many details of the genotype-phenotype map that arises from global GRN dynamics are unknown. To explore the possible state space of GRN from first principles, theoretical approaches based on ensembles of Random Boolean Networks (RBN) have been proposed 4 decades ago [10] and thoroughly explored by methods adopted from statistical mechanics and nonlinear dynamics [6, 2, 16, 21, 25]. Recently, Boolean network-based approaches have been successfully applied to reproduce state space and robustness of real GRN [15, 5] and, using evolutionary optimization techniques, predictions on the relationships between the robustness against different types of perturbations have been made [4]. Yet, these studies are limited by several shortcomings immanent to the

RBN abstraction. In particular, in these models constraints originating from sequence structure and spatial extension of DNA are not considered, though they play key roles in genome organization, adding an important layer of epigenetic information to GRN. For example, it has been shown that both in eukaryotes [8] and in bacteria [3] gene transcription occurs in discrete foci (”transcription factories”) where several RNA polymerases are co-localized. Hence we expect that, in order to optimize transcriptional control, genes should also tend to co-localize in space. This idea is indeed supported by genomic and transcriptomic analyses [11] that have shown that genes regulated by a given TF and the regulator gene coding for this TF tend to be periodically located along the DNA. This periodicity is consistent with a solenoidal epi-organization of the chromosome, which would dynamically gather the interacting partners into foci [12, 9] and thereby enhance the effect of TFs by induction of local concentration effects [19]. Recent studies support the idea that the genome as a whole dynamically self-organizes in space to optimize information processing [7].

The organization of this paper is twofold: In section 2, we first introduce the notion of hybrid GRN, modeled with discrete threshold networks that are characterized by two classes of interactions: weak (or ordinary) links, and strong (or privileged) links. The latter represent interactions (TFs and regulated genes) that profit from local concentration effects. Using random network ensembles, we investigate parameter ranges (in particular, the fraction of strong links) where robustness of network dynamics is optimized. In the second part of the paper, we extend an existing *artificial genome model* [22, 26] by including information about 3D-distances between genes with respect to a solenoidal genome organization. While different artificial genome models based on sequence-matching mechanisms have been studied by several authors [22, 1, 14], also addressing evolutionary optimization problems [13, 20, 26] constraints induced by the spatial genome organization are usually neglected. Using genetic algorithms, we investigate in section 3 the evolutionary optimization of GRN interaction weights with respect to a solenoidal epi-organization. We identify combinations of mutation operators particularly successful for evolving this epigenetic organization, and formulate constraints we have to impose on selection (fitness functions). In section 4, a discussion of our results and concluding remarks are provided.

2 Effect of inhomogeneous weights on robustness of regulatory dynamics

As a first step, we will develop a discrete dynamical network model that takes into account spatial effects on regulatory interactions, parametrized in the strength of interaction weights. Similar approaches have been investigated before, e.g., by assigning ”privileged interactions” to Boolean networks and studying the dynamical constraints that emerge from this extension of the Boolean network paradigm [17]. We will follow similar ideas, however, applying a more intuitive approach (rooted in statistical physics) for investigation of the resulting phase space of network dynamics.

In certain contexts, e.g. assuming bivalent binding of TFs to distal binding sites on DNA, DNA looping can induce strongly cooperative effects and lead to fold increases of up to 100 in the efficiency of transcription regulation [28], furthermore, models predict that at the same time fluctuations of transcription are drastically reduced. More generally, without assuming specific binding mechanisms, spatial proximity of reactive groups can induce *local concentration effects* [19], whereby molecules that are close to each other interact more efficiently. We now aim at integrating this type of effect into a discrete dynamical network model, using a binary state space. Evidently, one problem arises immediately: since concentrations of gene products are reduced to an ”on-off” description, concentration effects cannot be considered at the side of regulators. However, it is well possible to consider effects at the side of the targets, i.e. the regulated genes. More specifically, we assume that local concentration effects due to spatial organization of the genome strongly increase the *weights* of ”correctly positioned” regulatory interactions in gene regulatory networks. In our model, regulatory interactions are discrete with $w_{ij} = \pm 1$ for ”ordinary” links, while interactions exploiting positional effects take higher values $w_{ij} = \pm W$ (e.g. $W = 100$). Note that this distinction into two classes of interactions (”ordinary” and ”strong” links) - mainly to facilitate analytical treatment - represents a strong idealization of the situation in real genomes, yet, to some degree it is justified by the fold increase in local concentrations that can be induced by spatial effects, as explained above. The transcriptional state $\sigma_i \in \{-1, 1\}$ of a gene i at time t (-1 meaning ”untranscribed”, $+1$ ”transcribed”) depends on its regulatory inputs at time $t-1$ through the transfer function:

$$\sigma_i(t) = \text{sgn}\left(\sum_j w_{ij}\sigma_j(t) + h_i(t)\right). \tag{1}$$

The threshold $h_i(t)$ is typically fixed to a constant value (e.g. zero), however, we will later also consider fluctuations of h_i as a model for the impact of *extrinsic noise* on regulatory dynamics.

In the following, we consider randomly constructed GRN with N genes, a total of K_{tot} interactions between those genes and $K_w \leq K_{tot}$ strong interactions ($w_{ij} = \pm W$). We fix the *average connectivity* $\bar{K} = K/N$ and vary the fraction $\rho_w = K_w/K_{tot}$ of strong interactions.

2.1 Damage propagation and canalizing inputs in hybrid GRNs

Robustness of regulatory states, i.e. insensitivity against small, random perturbations of dynamics, is an important determinant of the persistence of phenotypes in biological organisms. Different types of perturbations have been studied in this context. In *damage propagation* studies, the binary state of a small subset of ”genes” (typically one gene) is inverted temporarily, and the divergence between perturbed and unperturbed state trajectories (damage) is measured. If the damage typically vanishes, the network is said to be in the ”ordered regime”, if damage typically increases and leads to a different dynamical attractor, the network is said to be ”chaotic”. Here, initial bit flips (state inversions) can be interpreted as transient gene knock-outs (e.g. by blocking their transcription for a limited time). A second type of perturbation arises from random fluctuations (noise) in state updates. For discrete networks, it is notoriously hard to give a biological interpretation of state noise due to the missing time scale separation between elementary updates and global network states. Hence, we shall apply a different concept and study *threshold noise*, i.e. fluctuations of h_i (cf. Eqn. 1), which have a well-defined interpretation in terms of extrinsic noise, i.e. noise in regulatory dynamics that comes from diverse environmental and intracellular influences [27].

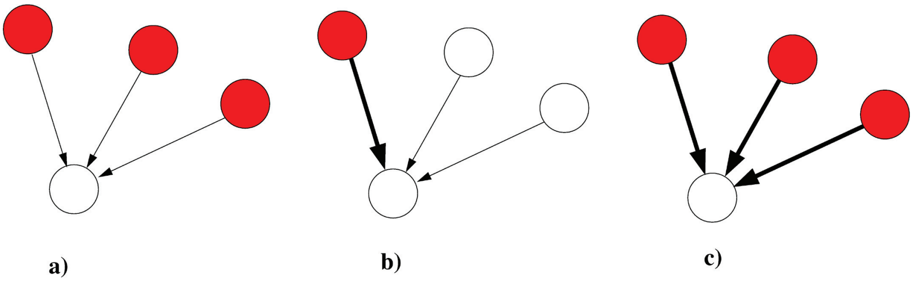

Figure 1: Effect of input perturbations in hybrid GRN; perturbations that can change the state of the regulated gene are shown in red. a) If only weak regulatory inputs are present (thin arrows), all inputs can induce state changes. b) For mixed inputs, only those with strong interaction weights (thick arrows) can induce state changes - they act as *canalizing inputs*. c) The case when only strong inputs are present, is equivalent to a).

Depending on their number of strong regulatory inputs, genes respond differently to damage (bit flips of inputs). If only inputs of one type (weak or strong) are present, any perturbation can change the state σ_i (cf. Fig. 1, case a) and c)). For mixed inputs (Fig. 1 case b)), in most cases only the strong interactions can lead to damage propagation. Hence, they act as *canalizing inputs* [18] that completely determine the state of the regulated gene. In the following, we will apply the so-called *annealed approximation* [6] to derive the average damage propagation behavior over the whole network as a function of ρ_w.

2.2 Damage calculation: annealed approximation

We assume that all links have equal probability ρ_w to have a strong weight. It follows that the density $\rho(k, k_w)$ of nodes that have k inputs, of which $k_w \leq k$ are strong ones, is given by

$$\rho(k, k_w) = \rho(k) \cdot \binom{k}{k_w} (1 - \rho_w)^{k-k_w} \cdot \rho_w^{k_w}, \tag{2}$$

where $\rho(k)$ is the in-degree distribution of the underlying network graph (i.e., in our case, a Poissonian with mean $\bar{K}$). Using an *annealed approximation* [6, 23], it can be shown that the average damage $\bar{d}$ following a one-bit perturbation (state flip) at time $t = 0$ is

$$\begin{aligned} \bar{d}(t+1) = \quad & \sum_{k=1}^{N} \left(k\,\rho(k,0)\,p_s(k) + \sum_{k_w=1}^{k} k_w\,\rho(k,k_w)\,p_s(k_w) \right) \\ + \quad & \sum_{k=3}^{N} \sum_{l=1}^{[k/2]} \frac{1}{2}(k-2l) \cdot \rho(k,2l)\,p_s(k-2l) \end{aligned} \tag{3}$$

Here, $p_s(k)$ is the damage propagation rate for nodes with in-degree k, which can be calculated analytically with combinatorial methods [23, 24]; it approximately decays $\sim 1/\sqrt{k}$. The second term on the right hand side of Eqn. 3 takes into account input configurations where strong weight inputs of opposite sign exactly cancel out, and hence perturbations of weak inputs can contribute to damage. These cases can be avoided by random assignment of strong weights from an interval $[W - \Delta W, W - \Delta W + 1, ..., W, ..., W + \Delta W - 1, W + \Delta W]$, where ΔW is an integer with $\Delta W \ll W$. Fig. 2 shows both cases (i.e. identical values $W = 100$ for all strong weights (1), and random assignment of strong weights from an interval $[80, 120]$ (2), as explained above), for random networks with average connectivity $\bar{K} = 2.1$. In both cases a transition from chaotic ($\bar{d} > 1$ to ordered ($\bar{d} < 1$) dynamics is found for intermediate values of ρ_w, however, for case (2) it is more pronounced than for case (1).

Finally, we investigate the effect of threshold fluctuations. The threshold $h_i(t)$ (see Eqn. 1), usually set to zero, now can take values -1 or $+1$ with probability $p_{tf}/2$, respectively, and 0 with probability $1 - p_{tf}$, where p_{tf} is chosen at the order of $1/N$. This type of fluctuation could be interpreted, e.g., as extrinsic noise [27]. Here, we find a picture that is slightly different

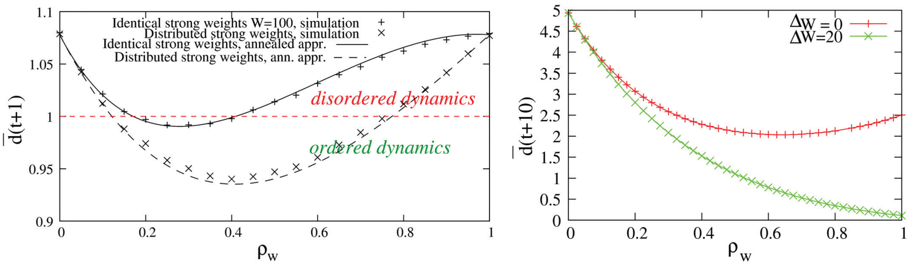

Figure 2: *Left panel:* Damage one time step after a one-bit state perturbation, as a function of ρ_w, for $\bar{K} = 2.1$ and $N = 128$. Data points for $W = 100$ (+) and $W \in [90, 110]$ (X) where averaged over 100000 network realizations each. Lined curves are the respective analytical predictions from Eqn. 3. *Right panel:* Damage due to threshold noise ($p_{tf} = 0.05$), 10 time steps after dynamics was started from identical initial states, for $W = 100$ (+) and $W \in [90, 110]$ (X). Details are explained in the text.

from the systems behavior for state flips: while for identical W a minimum of trajectory divergence (i.e. maximal robustness) is found at intermediate values, for case (2) it monotonously decreases and becomes maximal at $\rho_w = 1$. To summarize, our damage propagation studies on ensembles of random, "hybrid" GRN (i.e. networks with mixed weak and strong interactions) suggest that robustness is maximized either at intermediate fractions of strong interactions - when state flips (transient knock-outs) are considered - or for networks with a majority of strong links, when threshold fluctuations (extrinsic noise) dominate. Hence, the optimal density ρ_w depends on which type of fluctuations dominates the dynamics. From statistical network ensembles, however, we gain only limited insight into possible optima of epigenetic organization; in particular, it cannot be decided if those can be reached in an evolutionary process, or not. This will strongly depend e.g. on the *spatial constraints* in the genome, and on the types of mutations that can occur. These questions will be addressed in the following section.

3 An artificial genome model with solenoidal epi-organization

3.1 Development of the spatial genome model

From a pure network model only limited insight into the interplay between regulatory dynamics and spatial, epigenetic organization of the genome can be gained. Hence, we improve the model by including 1) a sequence-based artificial genome model that encodes both TF-DNA binding, and the positions of genes and non-coding regions on DNA and 2) an abstract representation of the 3D distances between elements of the genome, based on the assumption of a global genome organization according to a solenoidal structure.

Let us first define the underlying, basic artificial genome model (for details, cf. e.g. [22, 26]). Randomly string together S integers drawn uniformly between 0 and 3 (to provide correspondence to the ATGC alphabet of DNA). Next, define a base promoter sequence of length l_p to indicate the position of genes in the genome, e.g. '01010'. Wherever the promoter sequence occurs, the next l_g digits are specified as a "gene" (coding sequence). Transcription and translation into a protein sequence are abstracted into the transformation $s \mapsto \{(s+1) \mod 4\}$ for each digit of the coding sequence. Binding sites are determined by searching the genome for the protein sequence. If a match is found, then the protein is a transcription factor (TF) that binds to that site and that regulates the next downstream gene. In case there are multiple binding sites of this TF for this gene, only one of them is counted for network construction (the one which is closest to the gene coding for the TF, with respect to the 3-dimensional distance defined in the following).

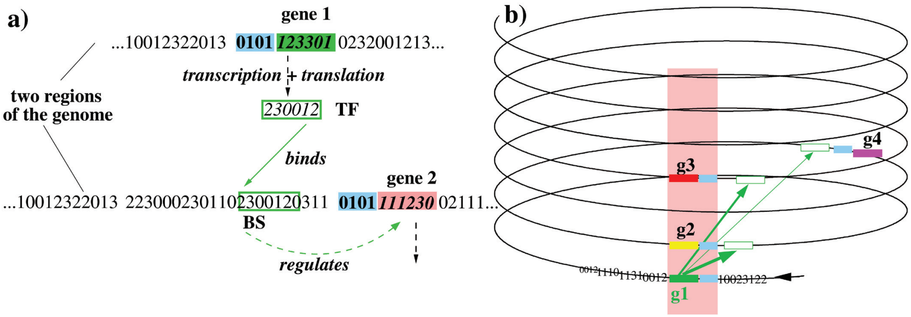

Figure 3: a) Schematic description of the artificial genome model (after [22, 26]). Base promotor sequences are marked in light blue; the next l_g digits define the gene (the coding sequence). Gene 1 produces a TF that binds to a matching binding site (BS) upstream of gene 2, and regulates transcription of gene 2. Iteration of this construction for all genes leads to an emergent, global GRN structure. b) On top of a), we impose a solenoidal epi-organization with period P. g1 regulates transcription of g2, g3 and g4, however only genes g2 and g3 are aligned at distances of P bases, and hence in phase with the solenoidal organization, while g4 is not. Therefore, interactions between g1, g2 and g3 are stronger (thick green arrows) than between g1 and g4 (thin green arrow).

Fig. 3 demonstrates the transformation of the 1D sequence of the artificial genome into a 3D solenoidal structure. Assuming DNA is folded according to a solenoid with periodicity P and height h per turn, the 3D-distance between to points i and j with 1D distance $d_{1D}(i,j)$ (counting the number of bases between i and j) on DNA is

$$d_{3D}(i,j) = \sqrt{\frac{P^2 - h^2}{2\pi^2}\left\{1 - \cos\left(\frac{2\pi d_{1D}(i,j)}{P}\right)\right\} + \left(\frac{hd_{1D}(i,j)}{P}\right)^2} \quad (4)$$

As indicated in Fig. 3 b), efficiency (strength) of interactions will decay fast with the 3D distance between the regulator (the gene coding for a TF) and its target (binding site). Assuming that TFs find their target in a diffusive random walk, we approximate the distance dependence of regulatory interaction weights as

$$|w_{ij}|(d_{3D}(i,j)) = (W-1)\exp\left[-\mu d_{3D}^2(i,j)\right] + 1, \quad (5)$$

where W is the maximum weight. Fig. 4 shows both the solenoidal distance $d_{3D}(i,j)$ (a) and the weight function $|w_{ij}|(d_{3D}(i,j))$ (b) for $h = 2$ and $P = 1024$. Evidently, for this choice of parameters, the weight function is sharply peaked at periodic intervals, such that only co-regulated genes (and their respective binding sites) aligned according to the scheme shown in Fig 3 (b) will contribute large values of $|w_{ij}| \approx W$.

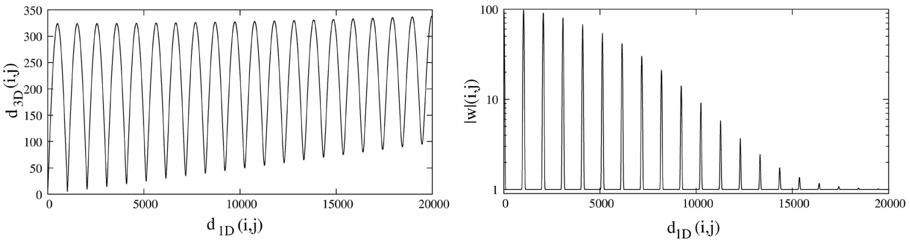

Figure 4: *Left panel:* Solenoidal 3D distance, as a function of the 1D distance (number of bases) between two points i and j on DNA, calculated according to Eqn 4 with $P = 1024$ and $h = 2$. *Right panel:* Interaction weights as a function of solenoidal 3D distance, calculated according to Eqn. 5.

3.2 Evolutionary optimization of the GRN with respect to solenoidal organization

We now ask the question how a GRN structure optimized with respect to a solenoidal organization may arise in an evolutionary process. Selective pressure will tend to increase interaction weights to optimize reliability and efficiency of gene regulation. Hence, we begin our evolutionary study with selection for increasing interaction weights. We create a mother genome, construct its GRN according to the procedure explained in the previous section, and determine $|w_{ij}|(d_{3D}(i,j))$ for all network connections according to Eqn 5, for fixed P and h. Next, we calculate the *average absolute interaction weight*

$$\langle |w_{ij}| \rangle = \frac{1}{K_{tot}} \sum_{i=1}^{N} \sum_{j=1}^{N} |w_{ij}|(d_{3D}(i,j)). \tag{6}$$

We apply one of the following combinations of mutation operators to the mother genome: 1) point mutations with probability p_{pm} per base, 2) transposition of one random subsequence of length $0 < l_{seq} < l_{max}$ and 3) a combination of both, i.e. with probability 1/2 either 1) or 2) is applied [1]. Now the GRN for the daughter genome is inferred from the mutated sequence and all interaction weights are calculated. The daughter genome replaces the mother genome if $\langle |w_{ij}| \rangle_{daughter} \geq \langle |w_{ij}| \rangle_{mother}$, otherwise it is discarded and the mother is kept. Figure 5 summarizes the results of evolutionary simulations obtained for this fitness function.

We observe that optimization is most efficient for combination 3) of mutation operators (point mutations and transpositions), while point mutations alone are least efficient; however, in all three cases optimization towards increasing average weights (Fig. 5, left upper panel) and ρ_w (Fig. 5, right upper panel) is achieved. Additionally, we find that the average network connectivity decreases during evolution, an effect which is most pronounced when only point mutations are at work (Fig. 5, right lower panel). Obviously, it is very hard to co-adapt the positions of regulators and their targets on DNA with point mutations alone, such that mainly semi-destructive mutations (i.e. mutations that delete weak interactions, while they keep strong ones) are exploited, as can be concluded from the strong decrease in average connectivity. While this observation is interesting in that it may provide a novel explanation for the relatively sparse connectivity of real GRN, in real systems regulatory demands will certainly limit destructive mutations. In our study, network disconnection is mainly a consequence of the rather weak selective pressure applied by using Eqn 6 to define the fitness function, which does not impose any constraints on network connectivity. Hence, we now refine the fitness function and demand that

$$f(\rho_w, \langle W \rangle, \bar{K}) = \rho_w \cdot \bar{K} \cdot \langle W \rangle \tag{7}$$

is optimized, where ρ_w is the density of strong interactions, $\bar{K}$ the average network connectivity and $\langle W \rangle$ the average weight of strong interactions. The

[1] Note that mutations are random, however they have to respect the constraint that the number of genes N remains constant, i.e. that no new base promotor sequences are created and no existing ones are deleted.

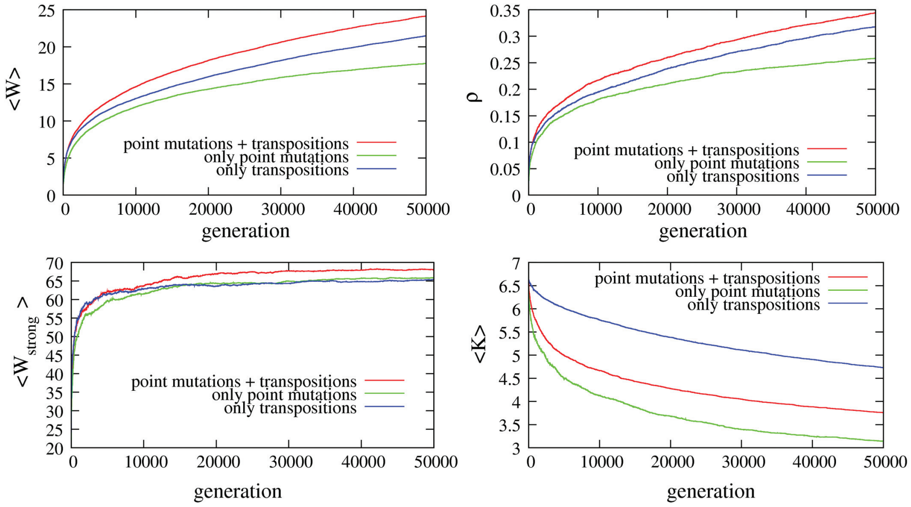

Figure 5: Results of evolutionary optimization for increased average interaction weights (as defined by Eqn 6), averaged over 50 different genetic algorithm runs. *Left upper panel:* Evolution of the average interaction weight (average over all network links). *Right upper panel:* Evolution of the density of strong interactions. *Left lower panel:* Evolution of the average weight of strong interactions. *Right lower panel:* Changes of the average connectivity during evolution.

product structure of f enforces optimization of both the fraction and the average strength of strong interactions, while at the same time, due to the dependence on $\bar{K}$, network disconnection is disfavored. We find that evolutionary optimization also works for this new fitness function, as can appreciated from the strong increase of ρ_w (Fig. 6, right panel) in evolutionary runs, while network disconnection is avoided (in fact, even a slight increase in connectivity is found, cf. Fig. 6, left panel).

Let us now have closer look on the evolved solenoidal organizations. We find that the probability distributions for the 3D distances between regulators (TF-coding genes) and their *target binding sites*, after 100000 generations of the genetic algorithm, indeed exhibit sharp peaks at intervals that are multiples of the imposed period P (Fig. 7, left panel), while for the distances between the co-regulated *genes* no periodic pattern is found (Fig. 7, right panel). How can we explain this seemingly counter-intuitive result? For typical parameter choices of the artificial genome model (in our study, alphabet size $\lambda = 4$, base promotor length $l_p = 5$ and gene length $l_g = 6$), genome structure is dominated by intergenic (non-coding) sequences, which make up almost 99% of genome content. This leaves ample space to optimize distances between

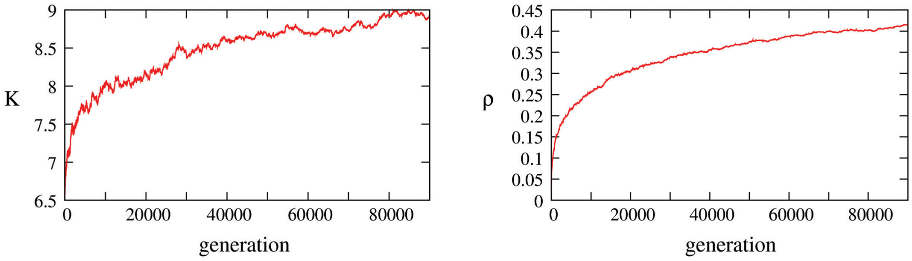

Figure 6: Genome evolution using a product fitness function (Eqn. 7). Changes in the average connectivity (left) and average density of strong links (right) in the course of GA generations are shown.

regulator genes and each of their target binding sites (BS), by just moving BS positions with respect to the regulator gene. This does not impose any direct constraint on the relative positions of genes that are regulated by this TF (in fact, they are often quite far away from the BS and not in phase with each other). While this situation is quite realistic for eukaryotes, where a similarly large fraction of non-coding DNA is found and TF binding sites can be at distance from target genes, the situation in bacteria is very different. Here, the distance between BS and target genes is typically short and quasi-constant. This dependence implies that indeed both types of distances - regulator gene to TF binding site, and relative distances between regulator gene and target gene - are minimized simultaneously. This joint optimization can lead to very focused transcription factories that facilitate, e.g., the coordinated binding of RNA polymerases and hence the synchronization of transcription of co-regulated genes, which will further enhance local concentration effects. To take into account this joint optimization problem, we now also impose a product structure on interaction weights, and define

$$|w_{ij}|(d_{3D}(i,j,k)) = (W-1)\exp\left[-\mu d_{3D}^{2}(i,j)\right]\exp\left[-\mu d_{3D}^{2}(i,k)\right]+1 \quad (8)$$

for two genes with positions i and l on the DNA, where i is the position of the regulator, l the position of the regulated gene, and j the position of the target binding site. For genome evolution, we again apply a genetic algorithm with the fitness function defined in Eqn. 7. Our results indicate that it is indeed possible to optimize both distances of co-regulated genes, and the respective target binding sites simultaneously with respect to the distance measure imposed by the solenoidal organization of the chromosome (Fig. 8). Interestingly, evolution of gene positions appears in turn to impose additional constraints on binding site evolution: we find that the corresponding distribution now decays much faster, i.e. favor small multiples of the solenoidal period, as compared to the relatively flat decay when fitness depends on the weight function not taking into account gene distances (compare Fig. 7, left panel to Fig. 8, left panel). This indicates that, under combined selective pressure, strong interactions in GRN are preferentially found in a very localized neighborhood of the regulating gene, while weaker interactions (that still make up a considerable part of the GRN) tend to connect to more distant genes. This suggests that the evolved GRN exhibit the small-world property [29], however, this needs further investigation. In principle, the predictions of our model can be addressed by means of whole genome data on the positions of TF bindings sites available today, and thereby provide hints on the (relative) selective pressures and mutational mechanisms at work that have shaped the overall organization of the (epi-) genome.

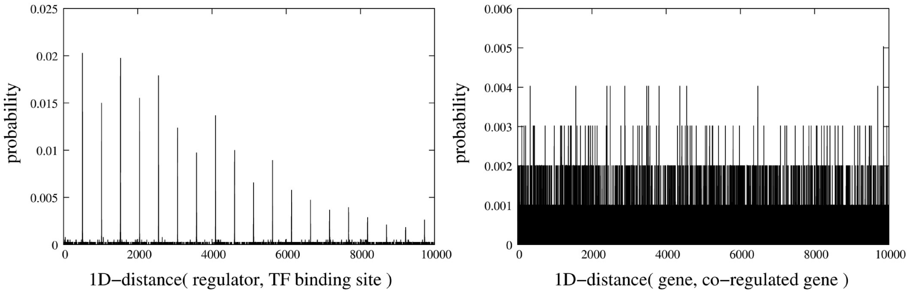

Figure 7: Probability distributions for 1D distances (leading to short 3D distances) between regulator genes and the respective TF binding sites (left) and between co-regulated genes (right) after 100000 generations, using the fitness function defined by Eqn. 7 and the weight function given by Eqn. 5.

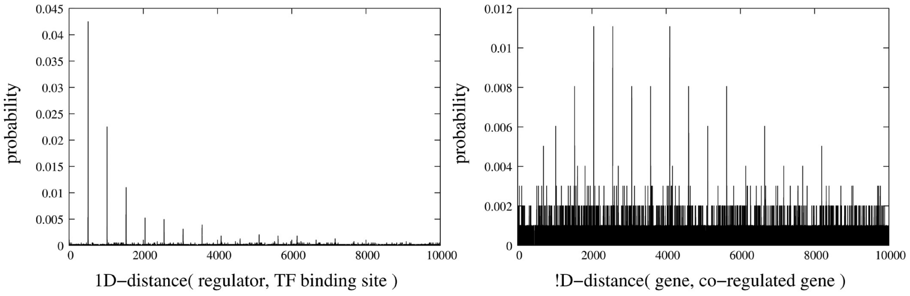

Figure 8: Probability distributions for 1D distances (leading to short 3D distances) between regulator genes and the respective TF binding sites (left) and between co-regulated genes (right) after 100000 generations, using the fitness function defined by Eqn. 7 and the product weight function given by Eqn. 8.

4 Discussion

We extended GRN models based on discrete dynamical networks, and artificial genome models based on a combinatorial description of TF-to-DNA binding by inclusion of an additional layer of epigenetic information: the spatial organization of chromosome structure. First, we studied ensembles of randomly generated GRN without an explicit representation of sequences and space, but with two classes of weights: weak (or ”normal”) interactions which are not optimized with respect to space, and strong (or privileged) interactions that are assumed to be optimized with respect to spatial vicinity of interaction partners and hence to exploit local concentration effects [19] efficiently. In particular, we investigated robustness of these ”hybrid GRN” with respect to two types of perturbations: random state flips of genes (damage), and threshold fluctuations that may be considered as a model of extrinsic transcriptional noise. Our results indicate that the optimal density ρ_w of strong interactions depends on the type of fluctuations considered: while damage propagation is typically minimized at intermediate values of ρ_w (where strong interactions act as canalizing inputs), maximal robustness with respect to threshold fluctuations is found at values of ρ_w close to 1. Next, we investigated evolutionary optimization of GRN under an explicit spatial representation, based on artificial genome model taking into account a solenoidal epi-organization of the chromosome. We find that optimization of GRN structure with respect to a periodic organization works best when both point mutations and transpositions are applied in the genetic algorithm (both types of mutations alone work, too, however, lead to slower convergence). The evolved network topology depends on the details of selective pressure formalized in the fitness function: selection for increased average interaction weights alone leads to strong disconnection of networks. A product fitness function - depending on both the absolute number and the average weight of strong interactions - avoids disconnection, and leads to even more pronounced periodic organizations. Furthermore, we showed that positions of co-regulated genes and TF binding sites can be optimized simultaneously, even when no particular constraints to the lengths of intergenic regions are applied. Together, these results suggest that it is possible to optimize global genome structure, including several layers of genetic and epigenetic information, in a gradual evolutionary process under multiple (sometimes even conflicting) constraints imposed by the different layers of organization (DNA sequence, GRN topology and dynamics, spatial development of the GRN). Future lines of research will lead to more elaborate multi-scale models of genetic information processing, taking into account more realistic constraints (e.g. selection for particular cellular phenotypes, or switching between different phenotypes) and additional elements of epigenetic organization, e.g. chromatin structure, and more detailed models of DNA looping.

Acknowledgments

T. R. gratefully acknowledges support from the CNRS-MPG program on Systems Biology (GDRE 513 SysBio), and the warm hospitality of the iSSB in Evry where most of this work was done.

References

[1] Wolfgang Banzhaf. On the dynamics of an artificial regulatory network. In W. Banzhaf, T. Christaller, P. Dittrich, J. Kim, and J. Ziegler, editors, *Advances in Artificial Life, Proceedings of the 7th European Conference (ECAL-2003), Dortmund, September 15-17, 2003*, Lecture Notes in Artificial Intelligence, LNAI 2801, pages 217–227. Springer, Berlin, 2003.

[2] Amsrtya Bhattacharjya and Shoudan Liang. Power-law distributions in some random boolean networks. *Physical Review Letters*, 77:1644–1646, 1996.

[3] J. E. Cabrera and D. J. Jin. The distribution of RNA polymerase in *Escherichia Coli* is dynamic and sensitive to environmental cues. *Mol. Microbiol.*, 50:1493–1505, 2003.

[4] Stefano Ciliberti, Oliver C. Martin, and Andreas Wagner. Innovation and robustness in complex regulatory networks. *Proc. Natl. Acad. Sci.*, 104:13591–13596, 2007.

[5] Maria I. Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS ONE*, 3:e1672, 2008.

[6] B. Derrida and Y. Pomeau. Random networks of automata: a simple annealed approximation. *Europhys. Lett.*, 1:45–49, 1986.

[7] Z. Duan, M. Andronescu, K. Schutz, S. Mcllwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Balu, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

[8] D. A. Jackson, A. B. Hassan, R. J. Errington, and P.R. Cook. Visualization of focal sites of transcription within human nuclei. *J. Cell. Biol.*, 164:515–524, 2004.

[9] Ivan Junier, Olivier Martin, and François Képès. Spatial and topological organization of dna chains induced by gene co-localization. *PLoS Comput Biol*, 6(2):e1000678, 02 2010.

[10] S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22:437–467, 1969.

[11] François Képès. Periodic transcriptional organization of the *E. Coli* genome. *J. Mol. Biol.*, 340:957–964, 2004.

[12] François Képès and C. Vaillant. Transcription-based solenoidal model of chromosomes. *Complexus*, 1:171–180, 2003.

[13] Paul Dwight Kuo, Andre Leier, and Wolfgang Banzhaf. Evolving dynamics in an artificial regulatory network model. In Yao X., Burke E., Lozano J.A., Smith J., Merelo-Guervos J.J., Bullinaria J.A., Rowe J., Tino P., Kaban A., and Schwefel H.-P., editors, *Proc. of the Parallel Problem Solving from Nature Conference (PPSN-04), Birmingham, UK, September 2004*, pages 571–580. Springer, LNCS 3242, Berlin, 2004.

[14] A. Leier, D.P. Kuo, and W. Banzhaf. Analysis of preferential network motif generation in an artificial regulatory network model created by duplication and divergence. *Advances in Complex Systems*, 10:155 – 172, 2007.

[15] F. Li, T. Long, Y. Lu, . Quyang, Q, and C. Tang. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. USA*, 101(14):4781–4786, 2004.

[16] B. Luque and R. V. Sole. Controlling chaos in random boolean networks. *Europhys. Lett.*, 37(9):597–602, MAR 20 1997.

[17] M. Manceny, M. Aiguierand P. Le Gall, I. Junier, J. Hérisson, and F. Képès. Spatial information and boolean genetic regulatory networks. *BICoB*, 5462:270–281, 2009.

[18] Andre Auto Moreira and Luis A. Nunes Amaral. Canalizing kauffman networks: Nonergodicity and its effect on their critical behavior. *Phys. Rev. Lett.*, 94:218702, 2005.

[19] B. Muller-Hill. The function of auxiliary operators. *Mol. Microbiol.*, 29:13–18, 1998.

[20] A. P. Quayle and S. Bullock. Modelling the evolution of genetic regulatory networks. *J. Theor. Biol.*, 238(4):737–753, FEB 21 2006.

[21] C. J. Olson Reichhardt and Kevin E. Bassler. Canalization and symmetry in boolean models for genetic regulatory networks. *J. Phys. A*, 40:4339, 2007.

[22] T. Reil. Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In *Proceedings of the 5th European Conference on Artificial Life*, pages 457–466. Springer, 1999.

[23] T. Rohlf and S. Bornholdt. Criticality in random threshold networks: Annealed approximation and beyond. *Physica A*, 310:245–259, 2002.

[24] Thimo Rohlf. Critical line in random threshold networks with inhomogeneous thresholds. *Phys. Rev. E*, 78:066118, 2008.

[25] Thimo Rohlf, Natali Gulbahce, and Christof Teuscher. Damage spreading and criticality in finite dynamical networks. *Phys. Rev. Lett.*, 99:248701, 2007.

[26] Thimo Rohlf and Christopher R. Winkler. Emergent network structure, evolvable robustness, and nonlinear effects of point mutations in an artificial genome model. *Adv. Comp. Sys.*, 12:293–310, 2009.

[27] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[28] Jose M. G. Vilar and Stanislas Leibler. DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.*, 331:981–989, 2003.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393:440–442, 1998.

Logic, Automation, and the Future of Biology

Ross D. King[1]

[1] Department of Computer Science Aberystwyth University, UK

Abstract

I present my vision of the future of laboratory biology based on using logic to represent biological knowledge and hypotheses, and advanced computers / robotics to automate the formation and testing of hypotheses. The advantages of using logic to represent scientific knowledge have long been understood. Despite this very little scientific knowledge has ever been represented using logic. This is now changing, and the application of the Semantic Web to science is developing a logic-based distributed infrastructure that is integrating large amounts of scientific knowledge. General purpose scientific reasoning tolls are also being developed to reason across the semantic web. These advances opens up the possibility of utilising the Semantic Web to provide a logical foundation for computational biology, and then using this foundation to develop novel tools and services. High-throughput laboratory automation is transforming biology and revealing vast amounts of new scientific knowledge. A natural extension of the trend is the concept of a Robot Scientist: this is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experimentation. If the trend to increased automation is continue laboratory automation hardware/software will have to overcome a number of existing limitations: flexibility, reliability, improved integration, etc. The greatest limitation of Robot Scientists is the lack of intelligence of the software. Improving this software is intimately linked to the goal of using logic to represent biological knowledge and hypotheses, and developing general purpose scientific reasoning tools.

1 State-of-the-Art

1.1 Logic and Biology

With a two and half thousand year tradition logic is the best understood way of representing scientific knowledge. Only logic provides the semantic clarity necessary to ensure the comprehensibility, reproducibility, and free exchange of knowledge [20]. Use of logic is also necessary to enable computers to play a full part in science: it removes the intractable difficulties with understanding natural language, and enables computational reasoning. Although the advantages of logic for science have long been understood [5], very little scientific knowledge has ever been represented using logic.

1.1.1 The Semantic Web

The Semantic Web was born out of a confluence of ideas from computer science, logic, and library science [2]. The best way to understand the Semantic Web, is not as the standard Web with an extra semantic layer, but rather as a world-wide knowledge base represented in logic. The Semantic Web is becoming a universal publishing platform for scientific knowledge [18]. The focus of Semantic Web development is now on the logical layer and developing applications.

1.1.2 Reasoning and the Semantic Web

Like the standard Web, the Semantic Web it can be used to search for information [2]. The advantage of the Semantic Web is that its information has clearer semantics, enabling information to be found easier. For example, if a human user or a computer are searching for information on ”RIF” (the rule interchange format), using the Semantic Web both should be able to easily avoid getting information on the Rif region in Morocco, the company RIF Worldwide, etc. For science the Semantic Web can also provide facilities such as integrating metadata, providing provenance information, integrating publications with original data and analysis methods, etc. Important as these advantages of the Semantic Web will be for science, the real benefits will be in enabling new inferences to be made from the knowledge available on the Semantic Web. *This is because it is these inferences that will enable new types of tools and services.*

There are three basic form of logical inference: deduction, abductions, and induction, and these along with probabilistic reasoning are the basis of all scientific inference. Deduction is the basis of traditional logic, mathematics, and computer science. It is a valid form of reasoning, so if a knowledge base is consistent then only new truths can be inferred. An example of a bioinformatic deduction is the following:

rule) if a cell grows it can synthesise tryptophan (P → Q);
fact) cell cannot synthesise tryptophan (¬Q);
then infer) cell cannot grow (P).

Research on deduction has until recently dominated research on inference for the Semantic Web (e.g. [11] is typical). There are now stable open source and commercial reasoning engines.

Deductive reasoning is insufficient for science as it cannot infer any knowledge that isn't already implicit in a knowledge base. This means that abductive and inductive inference are required to advance science. The easiest way to think about abduction is as deduction in reverse. An example of abduction is:

> rule) if a cell grows it can synthesise tryptophan ($P \rightarrow Q$);
> fact) cell cannot grow ($\neg P$);
> then infer) cell cannot synthesise tryptophan (Q).

Abductive reasoning is not valid, and therefore new empirical observations are required to ensure the truth of abductive inferences. Very little research has been done on developing abduction for the Semantic Web, but see e.g. [4].

More work has been done on developing induction for the Semantic Web (e.g. [12]), but it is still an under researched area. In relational learning (RL) there exists a technology which is "pre-adapted" for inductive reasoning over the Semantic Web [17]. The main technical challenge of adopting RL for the the Semantic Web are: the large amounts of data involved, engineering the inference methods to work over an open, and distributed environment of the Web, and the previous focus of RL on Datalog [21] rather than description logics [1]. Within machine learning RL's position is unusual. It is generally agreed to be theoretically important, yet its practical impact has been low. *The main reason for this is that very little data has been natively represented using logic, this is now changing with the Semantic Web, and RL is becoming a central technology.*

Logical inference and the Semantic Web fit well together. However, as James Clerk Maxwell pointed out "the true logic of this world is in the calculus of probabilities". By this he meant that all scientific knowledge is essentially probabilistic. The integration of relational learning with probability theory is one of the most exciting areas in machine learning [8, 7]. The main theoretical issue is that the traditional foundation of probability theory is propositional logic, while some variety of 1st-order predicate logic is required for RL and the Semantic Web.

1.1.3 Biology and the Semantic Web

Computers are essential to modern biology. Typical computational biological tasks are: genome annotation, analysing gene expression, protein structure prediction, phylogenetics, metabolomic analysis, systems modelling, etc. The state-of-the-art in computational biology is to use sophisticated scripting languages and Web services. This enables the zoo of existing bioinformatic programs to be integrated together, and enables some form of reproducibility.

Biological knowledge makes up a large percentage of the scientific Semantic Web, and many of the problems that makes general Semantic Web reasoning difficult don't apply to bioinformatics:

- A ground truth of scientific knowledge exists.
- A top level ontology have been agreed - the Basic Formal Ontology (BFO). This ensures that specific bioinformatic ontologies are logically compatible, and promotes cross-domain reasoning.
- The bioinformatic Semantic Web is large, but not as large as many other areas of the Semantic Web. It is therefore more computationally tractable.

These advantages have enabled work to proceed on describing biological knowledge using logic, and the European Bioinformatics Institute (EBI), and other large providers of bioinformatic data are now routinely publishing biological knowledge on the Semantic Web.

However, there is a mismatch between the growing use of the Semantic Web to represent biological knowledge, and the tools and scripts currently used for bioinformatic inference. Traditional biological software uses ad hoc inference, and the assumptions (logical and biological) they make are rarely explicit. This is unsatisfactory, as the hard-coding of scientific assumptions makes them obscure, difficult to understand, and difficult to change. It also precludes biologists checking these assumptions. From a formal point of view bioinformatic programs are invariably making logical inferences: deductions, abductions, inductions, with perhaps a probabilistic element. The form of these inferences need to be clarified if bioinformatics is ever to have a solid scientific foundation.

1.2 Automation and Biology

The use of computers to control the execution of experiments contributes to a vast expansion in the production of biological data [9]. This growth in data, in turn, requires the increased use of computers for analysis and modelling. High-throughput laboratory automation is transforming biology and revealing vast amounts of new scientific knowledge [10]. Nevertheless, existing high-throughput methods are currently inadequate for areas such as Systems Biology. This is because, even though very large numbers of experiments can be executed, each individual experiment cannot be designed to test a hypothesis about a model.

A natural extension of the trend to ever-greater computer involvement in the automation of experiments is the concept of a Robot Scientist [14, 15]. This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence [3, 16, 22] to execute cycles of scientific experimentation. A Robot Scientist automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using laboratory robotics, interprets the results, and then repeats the cycle. Robot Scientists have the potential to automate high-throughput hypothesis led experimentation.

1.2.1 Adam

The Robot Scientist Adam was designed to investigate the functional genomics of *S. cerevisiae*. Adam is the first machine demonstrated to have autonomously discovered novel scientific knowledge [15]. Adam is physically one of the most advanced laboratory automation systems in existence (Fig 1). (We are aware of larger and more expensive automated systems in a few academic labs, and in many companies, but we are unaware of any more flexible system). The advances that distinguish Adam from other complex laboratory systems such as high-throughput drug screening pipelines, and X-ray crystallography crystal screening systems, are its AI software, its many complex internal cycles, and is its ability in high-throughput to execute individually planned cycles of experiments.

Adam is designed to measure, in high-throughput, growth curves (phenotypes) of selected microbial strains (genotypes) in a defined media (environment). Adam is fully automated and there is no essential requirement for a technician except to periodically add laboratory consumables and remove waste. (However, the system is a prototype and it is advisable to have a technician nearby in case of minor problems.) Adam is able to run "lights out" for days at a time, and is capable of designing and initiating $> 1{,}000$ new strain / defined growth-medium experiments each day (from a selection of 1,000s of yeast strains), with each experiment lasting up to 4 days. The design enables optical density (OD) measurement for each experiment every 20 minutes, enabling accurate growth curves to be obtained ($>$10,000 growth measurements a day) - plus associated metadata.

Adam has autonomously generated functional genomics hypotheses about the yeast *S. cerevisiae*, and experimentally tested these hypotheses using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research we have developed an ontology and logical language. The resulting formalisation involves over 10,000 different research units in a nested tree-like structure, ten levels deep, that relates the 6.6

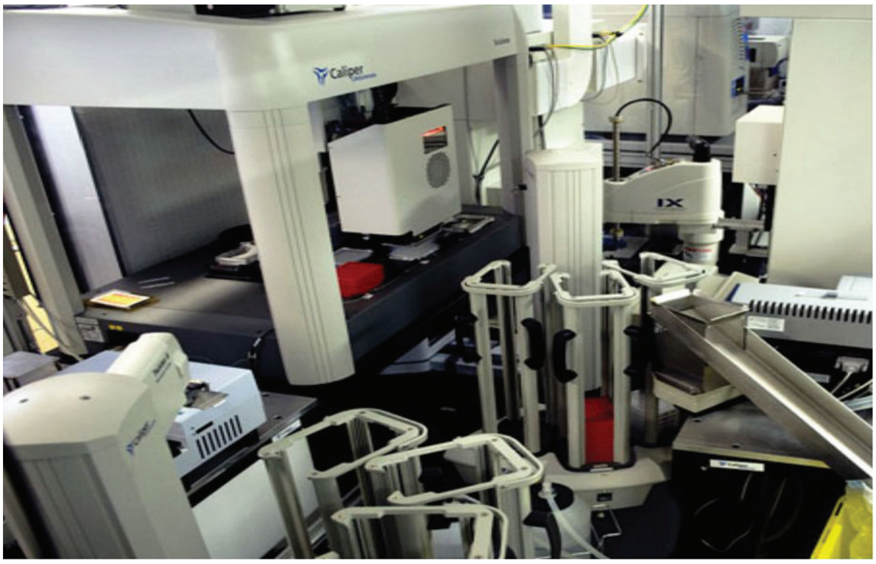

Figure 1: Part of Adam's integrated robotics and instrumentation

Figure 2: Part of Eve's integrated robotics and instrumentation

million biomass measurements to their logical description. This formalisation describes how a machine discovered new scientific knowledge.

1.2.2 Eve

Our second Robot Scientist, Eve, is a prototype system designed to demonstrate the automation of drug design and discovery [19]. Eve's robotic system is capable of moderately high-throughput compound screening (greater than 10,000 compounds per day depending on assay time) and is designed to be flexible enough such that it can be rapidly re-configured to run a number of different biological assays. Eve is designed to use chimeric yeast strains as the assay system. These strains are designed in collaboration with Steve Oliver's group in Cambridge. The main drug targets we are focussing on are enzymes from parasites such as *Plasmodium falciparum* and *Schistosoma mansoni*. Our assay approach is to create chimeric yeast strains that have yeast enzyme(s) removed and replaced by human EOR parasite ones.

A key objective Eve is to demonstrate the utility of integrating automated Quantitative Structure-activity relationship (QSAR) learning in the screening process. The idea is that once enough "hits" have been found (compounds found to be active through random screening of the compound library), then Eve will switch over to QSAR hypothesis formation and testing. The benefits of this are: lower attrition of the compound library, faster lead identification, lower costs, and better record taking.

2 The Future

2.1 Logic and Biology

2.1.1 A Logical Foundation for Computational Biology

The vision is to semantically integrate the existing computational biology service infrastructure with the growing amount of biological knowledge available on the Semantic Web. This will have two parts:

- Clarification of the semantics of existing computational biology software. The assumptions and inference mechanisms used by most existing computational biology software are not explicit. The aim is to make them explicit for the main classes of computational biology software.

- Formation of general purpose implementations of existing computational biology software. Given known assumptions and using general purpose Semantic Web inference tools implement standard computational biology tools.

To illustrate what I mean I will sketch what this would mean for two separate problem classes of computational biology software:

1. Predicting the structure of a protein domain based on sequence homology. This is typically the first step in a structural bioinformatics investigation. The computation is as follows: the distance between the target domain's sequence and all the domain sequences in the database of known structure is first calculated, then the target's structure is predicted to be the same as that of the closest sequence in the database. The biological rationale for this is based on the conservation of domain structure by evolution. Logical analysis reveals that many assumptions are made concerning the conservation of structure during evolution. It also reveals that the inference method is abductive. What is being abduced is the existence of a common ancestral domain shared by both the target domain and the domain with the closest sequence in the database, but by no other domains in the database

2. Predicting protein function from a micro-array profile. This is a common task in functional genomics. The goal is to predict the function of a gene by generalising patterns observed in transcriptomic experiments. The problem is technically interesting for machine learning as protein functions are organised in class hierarchy using gene ontology, and proteins may have more than one function. Logical analysis reveals that a number of implicit assumptions are made when applying machine learning to this problem. The most important of these is the closed-world assumption: if a protein is not known to have a specific function then it doesn't have that function. This assumption makes learning much more efficient as it generates large amounts of negative examples. However, it is in general a false assumption, as proteins may have functions which we do not yet know. This closed-world assumption clashes with the use of the semantic web language OWL. For the prediction task the inference mechanism is induction, as transcriptomic patterns associated with gene ontology classes are generalised.

2.1.2 Scientific Reasoning for the Semantic Web

There is a need to develop new inference mechanisms designed that takes full advantage of the logical infrastructure of the Semantic Web. These non-deductive reasoning methods will necessarily be based on Relational Learning [6] to be powerful enough to be able to reason using the logics used to represent scientific knowledge in the Semantic Web. The Relational Learning methods will include: Abductive Logic Programming, Relational Machine Learning, and Probabilistic methods.

2.1.3 Novel Computational Biology Tools

The key motivation for providing a logical foundation for computational biology and developing general purpose scientific inference mechanisms is not simply to improve our understanding of computational biology software, important as that is, but rather to use this understanding to develop new improved computational biology tools and services. Taking the same two examples from above, logical analysis will enable new variants to be envisaged, and these can be implemented using the general purpose scientific reasoning methods developed in the following ways:

1. Predicting protein domain structure on sequence homology. When it is realised that the basic logical inference involved is abduction of a common ancestral sequence, plus an assumption of conservation of function, it is possible to envisage variants of the basic bioinformatic method which are biologically more realistic, and which will result in more accurate predictions. For example it is clear that it should not be just a single ancestral sequence should be abduced, but rather a population of ancestors; and that the use of this population for prediction should be weighted by their evolutionary distance as estimated by the sequence metric. This produces a complex probabilistic relational graph similar to that generated in probabilistic relational learning [7]. Logical analysis of the problem can therefore be used to develop a representation then be solved using general purpose statistical relational learning methods.

2. The problem of predicting protein function from a micro-array profile is currently normally tackled using propositional learning methods, and these methods are generally limited to using only a limited set of attributes for prediction [6]. Logical analysis reveals that there is a large number of important sources of information that should be used in prediction: the gene ontology hierarchical class structure, the existence of multiple functions for the same protein (multiple-labels), that the micro-array data comes from multiple experiments often consisting of small time-series, the metabolic network that integrates the enzymes, the signalling networks that integrate the signalling pathways, the genome structure, etc. The bioinformatic semantic web will make collection and logical and biological integration of these sources simple to do. Then general purpose relational learning algorithms, plus the closed-world assumption, can be used to exploit all available sources of information for prediction - a basic law of reasoning is to use all available relevant information [13].

2.2 Automation and Biology

2.2.1 Hardware

Modern laboratory biology would be impossible without automation, for example high-throughput laboratory automation makes possible: the sequencing of DNA; the measurement of mRNA, proteins, of metabolites; identification of protein-protein and protein-DNA interactions; creation of deletant libraries; inhibition of gene expression, etc. [10]. Continuing advances in laboratory automation hardware mean that currently most biological manipulations can be done both faster and more accurately using automation than by hand. I expect this trend to continue, and for automation to increasingly dominate laboratory biology. If this to occur laboratory automation will have to overcome a number of existing limitations:

- Lack of flexibility. Almost all existing laboratory automation equipment is designed to do one or a few related tasks. This contrasts both with human scientists / technicians who when trained to use their hands to execute a vast array of laboratory operations; and to computers, which are capable of general purpose computing. Therefore, one the great challenges for laboratory automation is the design of equipment that can be reconfigured to execute a wide range of experimental tasks.

- Poor reliability. Almost all existing laboratory automation equipment is "brittle", that is if something goes wrong then the whole system ceases to function.

- Lack of standards. There are few agreed laboratory automation standards, and this greatly hinders the integration of different pieces of equipment. In addition even some the agreed standards are poorly designed for automation, e.g. the size / shape of 96-well microtire plates is not designed to be manipulated by equipment, unlike say wooden-pallets.

- High costs. Laboratory automation is very expensive relative to the sophistication of the equipment purchased. Our Robot Scientist Adam cost ~\$1,000,000 but its hardware is less sophisticated than that in a \$30,000 car. This is because of the small market for laboratory automation equipment - and possibly the ability of pharmaceutical companies to pass on their costs to consumers.

2.2.2 Software

Improved software is the key to the future of laboratory automation. Existing laboratory automation software is very limited, just like laboratory automation hardware, it lacks flexibility, is unreliable, there are few standards, and it is very expensive for what you get. To tackle these problems my colleagues (especially Dr. Amanda Clare) and myself have been trying to promote the open-source software for the control of laboratory automation. Recently Caliper Life Sciences kindly donated their software (formerly known as iLink or Clara) to Aberystwyth University in order that we can make it available to the open source community. This was the software that was used to control the Robot Scientist Adam here at Aberystwyth, and is in use in many other lab automation projects around the world:
see **`http://www.aber.ac.uk/en/cs/research/cb/projects/labux`**

The most exciting areas of software research for laboratory automation is the development of AI software. In my view the most fundamental limiting factor in developing Robot Scientists is the lack of intelligence of the software. The development of this software is very closely related to the development of the vision in section 3.1: a logical foundation for computational biology, scientific reasoning for the semantic web, and new computational biology tools. Robot Scientist software can be improved in the following ways:

- Improved background knowledge: This is currently represented as logic programs (1 st-order logic). This needs to be extended to include probabilistic knowledge, perhaps through the use of 1 st-order probabilistic logics (FOPLs). It will also be essential to augment the Robot Scientist's background knowledge with core knowledge about biology; currently Robot Scientists are idiot savants which have no real understanding what they are doing. This research is closely connection with developing a logical foundation for computational biology (2.1.1.).

- Improved methods of hypothesis formation: This is currently done using both pure abduction and bioinformatics, but the type of hypotheses that can be generated are limited. This research is closely connection with developing scientific reasoning for the semantic web (2.1.2.).

- Improved experiment formation: The current method is limited by assuming the execution of only one experiment at a time and does not properly take time into account. This research is closely connection with developing new computational biology tools (2.1.3.).

References

[1] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds. *The Description Logic Handbook*. Cambridge University Press (2003)

[2] Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web. *Sci. Am.* 284, 34-43.

[3] Buchanan, B.G., Sutherland, G.L. & Feigenbaum, E.A. Rediscovering some problems of artificial intelligence in the context of organic chemistry. In *Machine Intelligence* Vol. **5**(Eds. Meltzer, B. & Michie, D.) 253-280 (Edinburgh University Press, Edinburgh, 1969).

[4] Colucci, S., Di Noia, T., Di Sciascio, E., Donini, M.F., & Mongiello, M. (2005) Concept abduction and contraction for semantic-based discovery of matches and negotiation spaces in an e-marketplace. *Electronic Commerce Research and Applications* **4**, 345-361

[5] Davis, M. (2000) *The Universal Computer: The Road from Leibniz to Turing*. WW Norton.

[6] De Raedt, L. (2008) *Logical and Relational Learning*. Springer-Verlag

[7] De Raedt, L., Frasconi, P., Kersting, K., Stephen Muggleton, S. (2008) *Probabilistic Inductive Logic Programming*. Springer-Verlag.

[8] Getoor L. and Taskar B.(eds) *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[9] Hey, T. & Trefethen, A. The Data Deluge: An e-Science Perspective. In *Grid Computing - Making the Global Infrastructure a Reality*, **36** (Eds. Berman, F., Fox, G.C., & Hey, A.J.G.) 809-824 (John Wiley & Sons, New York, 2003).

[10] Hood, L., Heath, J.R., Phelps, M.E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science.* **306**, 640-643 (2004).

[11] Horrocks I., and Sattler U. A Tableau Decision Procedure for SHOIQ. *J. of Automated Reasoning*, 39(3):249-276, 2007.

[12] Iannone, L., Palmisano, I., Fanizzi, N. (2007) DL-FOIL Concept Learning in Description logics. *Applied Intelligence.* 26, 139-159.

[13] Jaynes, E.T. (2003) *Probability theory: The logic of science*. Cambridge

[14] King, R.D. *et al.* Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427, 247-252 (2004).

[15] King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E., Clare, C. (2009) The Automation of Science. *Science*. 324, 85-89.

[16] Langley, P., Simon, H.A., Bradshaw, G.L. & Zytkow, J.M. *Scientific Discovery: Computational Explorations of the Creative Process* (The MIT Press, Cambridge, Massachusetts, 1987).

[17] Lisi, F. A. and Esposito, F. 2008. Foundations of onto-relational learning. In Proc. of ILP'2008, F. Zelezny and N. Lavrac, Eds. Lecture Notes in Computer Science, vol. 5194. Springer, 158-175.

[18] Shadbolt, N., Hall, W., Berners-Lee, T., (2006). The semantic web revisited. *IEEE Intelligent Systems*.

[19] Sparkes, A., Aubrey, A., Byrne, E., Clare, A.,, Khan, K.N., Liakata, M., Markham, M., Rowland, J., Soldatova, L.N., Whelan, K., Young, M., & King, R.D. (2010) Towards Robot Scientists for Autonomous Scientific Discovery. *Automated Experimentation* **2**, 1.

[20] Toulmin, S. (2003) The Philosophy of Science. In *Encyclopaedia Britannica Deluxe Edition 2004 CD* (Encyclopaedia Britannica UK, London).

[21] J.D. Ullman. *Principles of Database and Knowledge-Base Systems*. Computer Science Press, 1989.

[22] Zytkow, J.M., Zhu, J. & Hussam, A. Automated discovery in a chemistry laboratory. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, 889-894 (AAAI Press, Menlo Park, CA, 1990).

The European Research Network in Systems Biology

Centre National de la Recherche Scientifique
CNRS
Coordinator
François KEPES (Genopole, CNRS, Evry)
Scientific Board
Vincent HAKIM (ENS, CNRS, Paris)
Vic J. NORRIS (Univ. Rouen)
Philippe TRACQUI (CNRS, Grenoble)

Max Planck Gesellschaft
MPG
Coordinator
Udo REICHL (MPI-Magdeburg)
Scientific Board
Jürgen JOST (MPI-MIS, Leipzig)
Frank JÜLICHER (MPI-PKS, Dresden)

Introduction

In 2006, the CNRS and the MPG began to think about the value of close collaborations in the field of ”Systems Biology”, which may be defined as the attempt to understand the behaviour of biological networks of interaction and, in particular, their spatio-temporal dynamics. This field typically requires cross-disciplinary import of concepts and crosstalk between benchwork, modelling and simulation. It turns out that research in Systems Biology is very vigorous and of a high standard on both sides of the Rhine and that there is an urgent need for collaboration.

After a small and successful first scientific meeting in Evry in February 2007, a bigger and more diverse meeting was organised in Berlin in September 2007 to broaden the appeal of MPG-CNRS cooperation in Systems Biology. Since then, general meetings have been organised each year, alternating between France and Germany. In January 2008, soon after the Berlin meeting, the European Research Network in Systems Biology (Groupement De Recherche Européen – CNRS GDRE 513) was created.

The main activities that were proposed by the scientific board in Berlin and validated by both CNRS and MPG are:

- Contribution to this yearly Thematic Research School on Systems Biology;
- Organisation of a general workshop held once a year alternatively in France and in Germany;
- Organisation of small focused workshops decided bottom-up to initiate or pursue specific CNRS- MPG collaborations.

This programme goes hand-in-hand with the CNRS-MPG post-doctoral programme, also established in 2008, which has appointed so far six high-level post-doctoral fellows subsidised by either the CNRS or the MPG. These fellows have either a major host in a CNRS laboratory and a minor host in a MPI, or vice-versa. The synergy between the two programmes became evident in 2009 when the fellows went beyond their own bi-institutional research projects to participate actively in fostering CNRS-MPG relations in systems biology. In particular, the fellows invested heavily in the organisation of some of the exceptionally successful, focused workshops.

Objectives

The GDRE exists to coordinate and improve CNRS-MPG collaboration in Systems Biology. Its specific objectives are to organise a yearly general conference as well as more focused workshops. The latter are proposed by scientists from both countries and allow concrete collaborations to be set up. The GDRE also contributes to a yearly Spring School on Systems Biology.

Report for 2008-2010

The yearly schools and conferences on Systems Biology, together with focused workshops (2 in 2009 and 3 in 2010) and exchange visits by senior scientists (1 in 2009 and 5 in 2010), have led to the creation of three ”small world networks” of strong collaboration:

1. Paris - Évry - Orsay / Leipzig - Halle
2. Montpellier - Bordeaux / Berlin
3. Lille / Saarbrücken - Dresden

The meetings included:

- 3 annual conferences in Grenoble, Leipzig et Paris which brought together a total of 180 participants.
- 3 thematic schools in Sophia and Evry which brought together a total of 260 participants.
- 5 focused workshops, proposed by the members of the network, namely, two small meetings in Evry and Berlin involving 8 groups and three bigger meetings in Lille, Leipzig and Paris bringing together 180 people.

It should be noted that the focused meeting in 2009 resulted in a publication that has been highly accessed and that several of these meetings have led to applications for international funding. The schools have each produced a book with an ISBN, totalling 500 pages. More details are given in the Annexes.

Perspectives

It is anticipated that the funds at the disposal of the GDRE will continue to be used to ”pump prime” new small, strong, long-term collaborative networks as well as to reinforce the collaborations already built. The agenda for 2011 includes:

- Organisation of a general Systems Biology conference in Dresden
- Partial support of a yearly thematic school
- Promotion of closer links between the CNRS and the MPG by providing information on the GDRE
- A call for focused workshops
- Encouraging further involvement of the GDRE postdocs in its activities
- Possibly setting up a visitors program for high level scientists

ANNEX 1

3^{rd} CNRS-MPG workshop on *Systems Biology*

ORGANIZATION CHAIR
Philippe Tracqui

PROGRAMME CHAIRS

Philippe Tracqui CNRS, Lab. TIMC- IMAG/DynaCell , Grenoble	Frank Jülicher Max Planck Institut für Physik komplexer Systeme, Dresden

More than fifty scientists from France and Germany gathered in Grenoble between 24 and 25 November 2008 for the 3rd edition of the CNRS-MPG workshop on Systems Biology.

It was the third time, after the launching meeting in Evry (February, 2007) and the following larger Berlin workshop (September, 2007), that scientists from CNRS and Max Plank Institutes have opportunities for discussions and exchanges on Systems Biologyresearch advances.

The “Centre de Congrès Europole” in Grenoble proved a fitting place for a workshop that brought together some of the leading scientists in France and Germany for intellectually stimulating debates and discussions on a wide range of timely topics, from the mechanical properties of the cell cytoskeleton and the response of mechanosensitive genes to the emergence of forms and functions in developing tissues and model organisms, from the integration and analysis of “omics” data to the development of models of gene and enzymes networks regulation.

Top ranking presentations of the keynote speakers, followed by questions drawn from a talented audience, helped to move the discussions with a brisk pace. Workshop schedule allowed very significant time for discussions, notably stimulated by the permanent exhibition of posters by junior scientists.

This 2008 workshop will certainly contribute to foster and facilitate interdisciplinary collaborations between both CNRS and MPG research organizations, already sustained by the perspective of the 2009 edition that will be hosted in Germany.

ANNEX 2

4th CNRS-MPG joint workshop on *Systems Biology*

November 23 24, 2009 Leipzig, Germany

ORGANIZATION CHAIR
Jürgen JOST

PROGRAMME CHAIRS

Jürgen JOST MPI MiS, Leipzig	Victor J. NORRIS University of Rouen

More than thirty scientists from France and Germany gathered in Leipzig on 23 and 24 November 2009 for the 4th edition of the CNRS-MPG workshop on *Systems Biology*.

It inaugurated a slight change in the mode of scientific interaction, as compared to previous general meetings. Indeed, it was a bit more focused, with the hope to increase the chances of finding common issues for collaborative work.

This workshop has in particular addressed the fundamental question in Systems Biology of how the interaction, regulation, and coordination of molecular processes leads to (a diversity of) coherent phenotypes at the cellular level. The workshop programme has been aptly distributed along three main lines of investigation, looking at control by structures, by molecules and by network properties.

The MPI MiS in Leipzig proved a fitting place for a workshop that brought together some of the leading scientists in France and Germany and abroad, for intellectually stimulating debates and discussions on timely topics in a cozy atmosphere. Top ranking presentations of the speakers, followed by questions from a talented audience, helped to move the discussions with a brisk pace. Workshop schedule allowed very significant time for discussions. The organizers would have hoped to see a higher number of posters by junior scientists, and stronger advertizing for this efficient means of direct interactions is envisioned for the next editions.

This 2009 workshop will certainly contribute to foster and facilitate interdisciplinary collaborations between both CNRS and MPG research organizations, already sustained by the perspective of the 2010 edition to be hosted in France.

ANNEX 3

Report of the meeting "Understanding robustness via dynamical transitions"

July 20-21, 2009, Berlin

Organizers:

Ovidiu Radulescu and Markus Ralser

Participants:

From CNRS : Ovidiu Radulescu (Rennes), Vincent Noel (Rennes), Jean-Pierre Mazat (Bordeaux), Christine Nazaret (Bordeaux)

From MPIMG Berlin: Markus Ralser, Christoph Wierling, Martin Kirch, Wasko Wruck, Marc Jung, Raed Abu Dawud, Anirban Banerjee, Hendrik Hache.

Summary:

The objective of the meeting was to organize a first contact between French modellers and German biologists and computer scientists interested in pioneering a new modeling approach to biological robustness. The focus of the discussions was the modelling of the transitions of the central carbohydrate metabolism as well as the general framework for studying robustness. Other topics have also been discussed such as: stemness, fate switching and robustness by differentiation, reverse engineering and parameter finding for pathway models, metabolomics technology.

Consequences of the meeting:

French and German participants to this meeting had not physically met before. Sociologically, new possibilities for interactions have been created. The Bordeaux group (JP Mazat and C Nazaret) interact with other persons in Berlin (Edda Klipp). The new contacts in MPIMG will strengthen and enrich the already existing collaboration.

O.Radulescu and C.Wierling will answer a general FP7 methodological call in systems biology.

We are seeking for common ressources (a joint post-doc) to continue the collaboration on modeling the glycolysis-PPP transition.

Remarks:

MPIMG Berlin was not aware of the MPG-CNRS agreement; however they financed the meeting at the same level as CNRS.

Both MPIMG and French parts would have been interested to support the post-doctoral programme of the CNRS-MPG systems biology consortium, by offering a collaborative environment and complementary infrastructures, but none of the institutions of the participants is authorized to do that.

ANNEX 4

Report of the meeting "Challenges in experimental data integration within genome-scale metabolic models"

October 10-11, 2009, Paris, Institut Henri Poincaré

Organizers:

Prof. J. Jost, MPI for Mathematics in the Sciences, Leipzig
Prof. O. Martin, LPTMS, UMR 8626 CNRS/Univ. Paris XI
Dr. P.-Y. Bourguignon, MPI MiS, Leipzig
Dr. A. Samal, MPI MiS, Leipzig.

Participants:

Country	Speakers	Participants
Germany	3+2	17
France	2	31
United States	2	4
Israel	2	5
United Kingdom	2	3
Switzerland	1	1
Hungary	1	2
Denmark	1	3
Spain	0	4

An awaited and timely event:

The main objective of the meeting was to bring together scientists working with constraints-based and kinetic models of metabolism. Aimed at bridging the gap between biochemistry and physiology using a combination of mathematics and computer science techniques, this subfield of systems biology is built upon a set of specific mathematical structures and computational methodologies. Researchers presenting their work in non-specialized conferences are frustrated by the need to introduce this background material at the expense of the originality of their contribution; because of this, they prefer by far more specialized audiences.

At a time when such modeling frameworks are moving towards ambitious endeavors such as integrating different omics data, we felt it was important to remedy the above problem and organize a specialized event. We thus seized the opportunity to invite world-class scientists who have recently contributed significant advances in the field to have interactive and in depth exchanges. We also wanted them to stimulate one another on promising alternative points

of view that are emerging (e.g. statistical vs. optimization-based predictions). Finally, it was an excellent means to have younger scientists (Ph.D.s and postdocs) to get in a mere 2 days a complete view of the field and its actors.
The conference was divided into 8 sessions, plus two panel discussions held at the end of each day. The latter benefited from a very active participation, allowing very diverse issues to be debated. This success was to be expected given the quality and vivacity of the participants, which also translated into good discussions after the talks. We feel that all this owed much to the friendly atmosphere that dominated the whole event. It should be stressed here that several talks presented yet unpublished works.

Some figures:

Following the aforementioned motivations, 14 invited speakers had been selected based on their commitment to experimental data integration using metabolic models. Given the limited number of time slots available, only two high-level contributed talks were accepted. The very high rate of positive answers to the organizers' invitations, as well as the multiple requests to organize such a conference again, testify to the usefulness of such an event.
The requests for participation was very high and so the registration had to be closed within a month of the announcement, when the maximum number of participants (70) was reached (limited by the conference hall capacity). The organizers thus had to reject a number of registration requests. The distribution of participants and speakers across countries is shown below. It is noteworthy that five participants were actually employed by companies (3 french biotechs - Genomining, Metabolic Explorer, Global Bioenergies- and Dupont).

Expected Impacts:

While Germany harbors the largest number of research groups working with metabolic models in Europe, this research topic has not yet attracted a comparable interest in France. However, the large number of attendees working in French research centers indicates that this situation is likely to change in the near future. This impression was confirmed by informal discussions with French participants, among which, interestingly enough, several consider themselves as potential end-users. It is thus foreseen that this workshop will have helped French researchers interested by these approaches identify potential collaborators in Germany.
From a more general perspective, many speakers (including the foremost actors of the field) said they had been delighted to have been part of the workshop and had found it very stimulating; in fact several requested that a follow-up workshop be organized. Having brought together people developing different approaches (small scale and detailed models vs. genome-scale coarser ones, optimization- based predictions vs. statistical frameworks) made the

connections between all these approaches more understandable to everyone. An impact of the workshop will be an increased level of interaction between the various types of models of metabolism in the near future, in particular in France and Germany. Finally, given the strong interest shown by participants that were relatively new to the field, it would be useful to invest in some training on the subject, for instance via a thematic school for Ph.Ds and postdocs.

ANNEX 5

5th MPG-CNRS joint workshop on *Systems Biology*

December 9-10, 2010, Paris, France

INTEGRATIVE NEUROSCIENCES

The meeting was held at Institut Henri Poincaré on December 9 and 10 and was dedicated to integrative neurosciences, an important domain of systems biology that had not been explored in previous meetings. Talks were delivered by fourteen well-known speakers, seven belonging to MPIs and seven from CNRS, also split about equally between theorists and experimentalists. The meeting was attended by more than seventy registered participants coming of course mainly from France and Germany, but also from other countries (mainly the UK). About a dozen presented posters on their work. It was also attended by about twenty unregistered participants. The unexpected bad weather conditions and allied flight and train cancellations at the meeting start, only resulted in one speaker cancellation, perhaps a sign of the speakers motivation to attend (one reported a 26 hour travel from Leipzig).

The topics ranged from basic issues of neural connectivity and anatomy as well as single neuron description to information processing and memory in different neural structures and animals (from invertebrates to mammals). The response to the meeting was very positive both from the audience and speakers. The talks were judged of high scientific quality by the participants and suscited a lot of interest as testified from the numerous questions and animated scientific discussions. The speakers also particularly liked the variety of topics and approaches that were described and discussed, and the meeting allowed extensive exchanges between theoreticians and experimentalists. The various scientific exchanges started at the meeting will undoubtly deepen existing collaborations and also promote new ones. The meeting success is expected to lead to strengthened links between CNRS and MPG scientists.

ORGANIZATION CHAIR: Vincent HAKIM
PROGRAMME CHAIRS: Vincent Hakim (CNRS, ENS Paris), **Jürgen Jost** (MPI für Mathematik in den Naturwissenschaften, Leipzig), **Fred Wolf** (MPI für Dynamik und Selbstorganisation, Göttingen).

Event URL:
`http://www.mis.mpg.de/calendar/conferences/2010/cnrs-mpg5.html`

ANNEX 6

Report of the meeting "Multi-scale dynamics and evolvability of biological networks"

October 4-6, 2010, Leipzig

Organizers:

Jürgen Jost, MPI-MIS, Leipzig
François Képès, ISSB, Evry
Olivier C. Martin, LPTMS, Orsay
Thimo Rohlf, ISSB, Evry
Areejit Samal, LPTMS, Orsay

Venue: Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
Event URL:
`http://www.mis.mpg.de/calendar/conferences/2010/musebio10.html`

A major challenge in systems biology is to understand the dynamics of biological networks at different scales of organization, and to integrate this knowledge into models, thereby exhibiting functional sub- networks embedded in larger dynamical systems. Multi-scale dynamics is at the heart of biological function: proteins and RNA molecules, for example, may be seen as elementary computational devices that capture various types of information from the cellular environment, providing the bottom layer of cellular dynamics from which emerge the functional networks of metabolism, signal transduction and gene regulation. Similarly, the genome not only codes for proteins, but it also determines the dynamical processing of this information in space and time via gene regulatory networks and in the epigenetic organization of the genome. The multi-scale architecture of biological networks has been shaped by evolution, and it clearly influences strongly the evolvability of organisms, i.e. their potential to adopt new functions or new phenotypes. Thus evolutionary frameworks are also necessary for us to reach a good understanding of the how and why of cellular dynamics.

The main objective of this interdisciplinary workshop was to bring together leading MPG and CNRS scientists in key fields for integrated modeling of the function and evolution of biological networks. While the main focus was on the theoretical (modelling) side, recent advances from experiments were also presented at the meeting.

Some figures

This two day conference was divided into 10 sessions, of which 8 were devoted to 18 invited talks, plus one poster session and one panel discussion session. The panel discussion benefitted from a very active participation, allowing very diverse issues to be debated in a fairly thorough manner. This success was in agreement with the quality and vivacity of the discussions occurring after most of the talks, and obviously owed much to the friendly atmosphere that dominated the whole event.

In order to initiate and foster collaboration between CNRS and MPG scientists, 18 leading scientists, mostly from Germany and France, were invited to present their work at the event. Of the 18 invited speakers, 6 were from CNRS/French institutions while 10 were from MPG/German institutions. The organizers enjoyed an overwhelming response from scientific community for participation at this projected medium sized event. The registration had to be closed before scheduled deadline, since the maximum number of participants (70) was reached before the closing date (the conference theatre had limited capacity of 70 people). Further, there were 25 poster presentations by various participants at the event which was way beyond what was initially planned. The organizers regret rejecting several registration and poster requests due to limited space at the conference venue. The distribution of participants' and speakers' host countries is shown below:

Country	Speakers	Participants
Germany	10	45
France	6	18
Rest of the World	2	7

Expected Impacts

It is foreseen that this workshop will lead to long-term collaborations between CNRS and MPG researchers pursuing modeling of biological networks. Furthermore, this event has communicated very efficiently the goals and scope of the CNRS-MPG programme on Systems Biology to the research community, certainly leading to an increased interest and future applications of postdoctoral scientists to the program. Given the overwhelming response to this year's event from scientists in France and Germany, we believe it would be very useful to organize a similar event in 2011 in France to further strengthen ties between CNRS and MPG researchers.

ANNEX 7

Report of the meeting “Chromatin Days: chromatin remodeling”

October 7-8, 2010, IRI, Lille

Organizers:

Christophe Lavelle, Ralf Blossey (IRI CNRS 3078).

Venue: Interdisciplinary Research Institute, Lille

Event URL:

http://www.iri.univ-lille1.fr/doc/chromatin_days_2010/

The meeting was held on October 7 & 8 at the Interdisciplinary Research Institute in Lille. The second edition of the IRI Chromatin Days, which was supported by the CNRS GDRE 513 “Biologie Systémique”, the IFR 147 of Lille 1 University, and the Fédération de Physique et Interfaces, University Lille 1, was dedicated to the topic of chromatin remodeling. Eight speakers with backgrounds in molecular and structural biology, bioimaging, single-molecule biophysics and theoretical physics addressed this topic, giving ample proof of the high activity the field currently witnesses. About 40 participants subscribed to the meeting. Although mostly a topic of interest to molecular biologists, we are particularly proud that a substantial number of bioinformaticians were attending the meeting.

The response to the meeting was very positive, both from the audience and the speakers. The latter particularly enjoyed that the focus of this small meeting allowed them to interact and exchange on their scientific interest in a very concentrated fashion. In order to foster this exchange further, the organizers have sollicited a minireview series in the FEBS Journal, to which seven of the eight speakers will work together to publish four minireviews focusing on the themes of the meeting : molecular and structural biology of chromatin remodelers, single-molecule methods, imaging and a theoretical article on the dynamics of nucleosome displacement.

ANNEX 8

Report of the meeting “Biological networks”

December 7, 2010, iSSB, Evry

Organizers:

Pierre-Yves Bourguignon (MPI-MIS, Leipzig), Thimo Rohlf (iSSB, Evry) and Areejit Samal (LPTMS, Orsay).

Venue: institute for Systems and Synthetic Biology, Evry, France.

Outcome

A novel aspect about this event was the involvement of junior researchers from iSSB (Evry), LPTMS (Orsay) and MPI (Leipzig). Through this gathering, the biologists at iSSB (Evry) became aware of the new mathematical approaches developed at the MPI (Leipzig). This interdisciplinary event witnessed in-depth and stimulating discussions between theorists and experimentalists at the meeting. Further, the speakers from Max Planck Institute realized many new applications for their mathematical methods.

Program

15:45 - 16:15 Informal introduction over coffee and tea

16:15 - 16:40 Frank Bauer, MPI-MIS, Leipzig
On the synchronization of coupled oscillators in directed and signed networks

In this talk I will consider synchronizability of coupled oscillators in directed networks whose links may carry weights of mixed signs. I will show how the normalized Laplace operator naturally arises in case the coupling function does not vanish at the origin. I will study network synchronizability as characterized by the smallest real part of the Laplacian eigenvalues, with respect to the presence of directed links and signed weights and characterize cases when directed links improve synchronizability in comparison to undirected links.

16:40 - 17:05 Davide Fichera, iSSB, Evry
Enumeration of pathways in metabolic networks

17:05 - 17:30 Nils Bertschinger, MPI-MIS, Leipzig
Statistical complexity, exponential random graphs and motif statistics

In the context of time series analysis, statistical complexity is a well developed method to quantify dependencies of probability distributions. Here, we investigate the related idea of exponential random graph ensembles as a framework for quantifying the structure of networks. For these graphs, the counts of subgraphs with at most k links are a sufficient statistics for graph ensembles (exponential families) of order k. This framework allows to systematically study relation between cluster coefficient and assortativity, which are commonly used to quantify structure in networks. Finally, we present a principled way to construct null models for motif analysis.

17:30 - 19:00 Discussions on possibilities for collaborations.

From the glycolytic oscillations to the control of the cell cycle: a minimal biological oscillator

Rui Dilão[1]

[1] Nonlinear Dynamics Group, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisbon, Portugal

Abstract

We introduce the basic modeling approach in order to describe chains of enzymatic reactions. We analyze the effects of activation feedback loops in these chains of reactions, and we derive the conditions for the existence of oscillations. We show that enzymatic chain reactions with two sequential chains and one feedback activation loop describe the basic features of the cell cycle control in eukaryotes. This same enzymatic chain reaction also describes the glycolytic oscillations in yeast. From this modeling approach, it results that the S/G2 checkpoint of the cell cycle is under the control of the concentration of the Cdk protein Cdc25. The concentration of this protein tune several bifurcation parameters of the model equations and its variation can induce the crossing of a Hopf bifurcation, leading to stable oscillation in the concentrations of the Maturation Promoting Factor (MPF=cyclin B+Cdc2) and of its phosphorylated state. This model is consistent with the recent finding that the oscillation of a single Cdk module is sufficient to trigger the major cell cycle events (Coudreuse and Nurse, Nature, 468 (2010) 1074-1079).

1 Introduction

Oscillatory behavior in biological systems and in aggregates of cells and tissues is observed as periodic variations over time of protein concentrations. Examples of biological systems with this time behavior are oscillations in the concentrations of cyclin proteins and of the Cdk enzymes controlling the cell cycle [15, 1, 7, 18, 19]; glycolytic oscillations in yeast anaerobic respiration [9, 10]; calcium oscillations controlling several cellular processes [8]; oscillation in the expression of proteins that trigger morphogenetic signals in mammals and responsible for the definition of the animal body plans [16]; oscillatory signals in the spatial aggregation patterns of the *Dictyostelium Discoideum* amoeba [12]; circadian rhythms in eukaryotes [17], bacteria [20] and plants [6].

There are several mathematical models aiming to describe oscillations in the concentration of specific enzymes and proteins in cells and tissues. All these mathematical models are based on the same biological assumptions and observations. For example, in order to describe glycolytic oscillations in yeast, Higgins [10], Selkov [21] and Goldbeter [7], among others, provided different models describing quantitatively the oscillatory behavior in time of the concentration of glycolysis in yeast cells. The differences between these models is due to different assumption on the molecular interaction mechanisms and processes. The interaction mechanisms involve the choice of different *ad-hoc* rate functions and threshold mechanisms. Other differences can result from different technical and simulation approaches as, for example, deterministic versus stochastic approach or delay versus non-delay differential equation approach.

One of the processes that is common to all the organisms and that has been conserved through evolution is the regulation and control of the cell cycle. One of the features of this process is its periodicity together with the existence of checkpoints in order to determine the order of transitions between the different phases of the cell. As this process is transversal to all life phenomena, it is admissible to assume that the basic cell cycle control mechanism has been conserved across evolution. For that, it must be minimal and the role of evolution has been to fine-tune the core system, [14, 3].

One of the first attempts to obtain a classification of the main type of cellular regulatory mechanisms was done in 1961 by Monod and Jacob, [14]. In this classification, they have proposed six different structural mechanisms present in bacteria that can exist and have been preserved by evolution in higher organisms. According to Monod and Jacob, one of these models produce cyclic phenomena. The key ingredients of this Monod and Jacob oscillatory mechanisms is the existence of one activating and one co-repressing loop in the production of two oscillating proteins. Their analysis is largely qualitative.

To explain glycolytic oscillations in yeast, Higgins, [10], proposed a model that produces oscillations through a feedback activation loop. Latter, Higgins, [11], classified the simplest sequences of enzymatic chain reactions with negative and positive feedback and feedforward loops.

One of our goals here is to arrive at a simple, minimal and realistic biochemical model leading to oscillatory time behavior of the concentrations of proteins or other substances as observed in cells, tissues and higher organisms. In our modeling context, simple and minimal means to have a minimum number of elementary chemical reactions, and realistic means to have only first and second order reactions.

Here, we show that the enzymatic chain reaction with two sequential chains and a feedback activation loop describes the basic features of the cell cycle control in eukaryotes, as well as the glycolytic oscillations in yeast. Moreover, we show that the S/G2 checkpoint of the cell cycle is under the control of the concentration of the Cdk protein Cdc25. The concentration of this protein tune several bifurcation parameters of the model equations and its variation can induce the crossing of a Hopf bifurcation, leading to stable oscillation in the concentrations of the Maturation Promoting Factor (MPF=cyclin B+Cdc2) and of its phosphorylated state.

In the next section, we briefly overview and derive the properties of the minimal model for a chain of two enzymatic reactions showing oscillatory behavior in time. This model is derived solely from the mass action law. The mass action law is the only theoretical mechanism that has a bottom-up justification based on molecular dynamics. In sections 3 and 4, we apply the basic features of the model to the particular cases of the control of the cell cycle and of glycolytic oscillations in yeast. In the case of the cell cycle, our model is not specific to any eukaryote organism (yeast, frog, humans). With this model, we pretend to describe a cytoplasmic oscillator, [15, p. 26], the core engine of the cell cycle. In the last section, we summarize the main conclusions of this paper.

2 Enzymatic chains of reactions with feedback and feedforward loops

We distinguish a kinetic reaction from an enzymatic reaction. Kinetic reactions are represented by collisional diagrams of the type,

$$\mathrm{A} + \mathrm{B} \xrightarrow{k_1} \mathrm{C} \tag{1}$$

where A, B and C represents atoms or molecules. The diagram (1) means that when the molecules A and B collide, they can bind and form a third molecule C. The rate constant k_1 measures the speed of the reaction in the media. If this reaction occurs in a well mixed media, then the evolution in time of the concentrations of the three chemicals can easily be calculated by the mass action law, [2, 4].

A Michaelis-Menten enzymatic chain reaction, is described by the sequence of kinetic reactions,

$$S + \mathrm{E} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{C} \xrightarrow{k_2} \mathrm{E} + P \tag{2}$$

where the enzyme E is a catalytic variable and C is a complex. In enzymatic chemistry, S is the substrate and P is the product of the reactions. Due to the lack of information about the role and rates of the intermediate steps in (2), this kinetic reaction is sometimes simplified and it is represent by the schematic diagram,

$$S \overset{E}{\dashrightarrow}_{MM} P \tag{3}$$

In order to determine the equations describing the variation in time of the concentrations of S and P, it is necessary to consider the concentrations of E and C, as well as the rates that are explicit in (2), but are implicit in (3). This information is hidden from the diagram (3) because most of the time it is unknown. However, for every chains of enzymatic reactions these parameters must be taken into account.

One of the basic simplifications that it is used in enzymatic chemistry is to consider the steady state approximation. It is generally assumed that the concentration of the complex C is constant in time and therefore the equations describing the kinetic mechanism (2) greatly simplify. For the simple 1-chain enzymatic reaction (2), the existence of an additional conservation law together with the steady state approximation derived from the mass action law implies that the enzyme concentration E is also constant over time. The validity of this approach when we compare the solutions of the full and of the approximate equations are justified by a theorem of Tihonov, [13].

The process of simplification just described will be done for every collision diagram involving different types arrows as the ones represented in (2) and (3). In chains of enzymatic and kinetic equations, the simplifications is important and must be carried out for the full set of chemical reactions. In general, in the description of biological mechanism there is a ambiguity on the meaning of interaction arrows. Here, we will use arrows with a precise meaning, and all the arrows used in diagrams will be associated with a specific kinetic mechanism.

The classification of the simplest linear chains of enzymatic reactions with feedback and feedforward activation and inhibition loops can be done by exhaustively analyzing all the possible interaction diagrams. The simplest cases, involving feedback and feedforward loops are the four enzymatic reactions represented in figure 1.

The activation and inhibition mechanism represented by the signed arrows in figure 1 are associated with the reversible kinetic mechanisms,

$$\begin{array}{ll} (+) & \mathrm{P} + \mathrm{E}^- \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{E} \\ (-) & \mathrm{P} + \mathrm{E} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{E}^- \end{array} \tag{4}$$

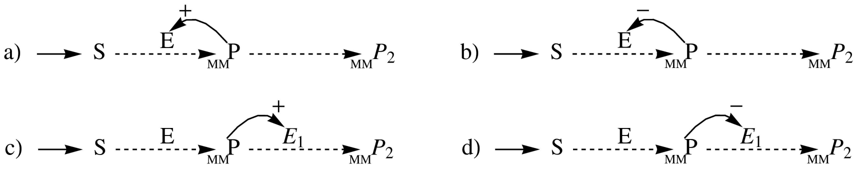

Figure 1: Linear chains of enzymatic reactions with feedback (a) and b)) and feedforward (c) and d)) activation (+) and inhibitory loops (−). Full line arrows without signals represent collisional kinetic mechanisms as in (1) and (2). Dashed line arrows represent enzymatic reactions as in (3). Signed arrows represent activation and inhibitory interactions and their meaning are explicited in (4).

where E^- is a non-active state of the enzyme E. If an enzyme is not affected by other mechanisms, it is assumed that it is in an active state.

It can be shown, that the enzymatic mechanisms b), c) and d) of figure1 cannot lead to oscillatory motion for all the values of the parameters[1], but the reaction a) has parameters values for which the dynamics oscillates in time. To be more specific, with the equivalences between (2) and (3) and the meaning of the signed arrows in (4), to the mechanisms in figure 1a), we associate the kinetic reactions,

$$\begin{aligned}
&\mathrm{G} \xrightarrow{k_1} \mathrm{G} + \mathrm{S} \\
&\mathrm{S} + \mathrm{E} \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} \mathrm{C} \xrightarrow{k_3} \mathrm{E} + \mathrm{P} \\
&\mathrm{P} + \mathrm{E}^- \underset{k_{-4}}{\overset{k_4}{\rightleftarrows}} \mathrm{E} \\
&\mathrm{P} + \mathrm{E}_2 \underset{k_{-5}}{\overset{k_5}{\rightleftarrows}} \mathrm{D} \xrightarrow{k_6} \mathrm{E}_2 + \mathrm{P}_2
\end{aligned} \tag{5}$$

The first reaction in (5) represents protein production from the gene G, [2]. The second reaction corresponds to the first Michaelis-Menten enzymatic reaction in figure1a), and the third reaction is the feedback activation loop. Note that, the forth reaction is also an enzymatic reaction, where we have explicitly introduced the additional enzyme E_2 and the new complex D. These substances are absent from the diagrams in figure1, however they must be taken into account. If we substitute this last reaction by a direct kinetic mechanism similar to the first reaction in (5), it can be exactly proved that the associated differential equations cannot have stable self-sustained oscillatory behavior (limit cycles).

[1] The proofs of these facts can be done as nonlinear dynamics exercises.

By the mass action law, [4, 5], the differential equations describing the time variation of the concentrations of the substances in the kinetic mechanisms (5) or in the diagram in figure 1a) are,

$$\left\{\begin{array}{lcl} \mathrm{C}'(t) & = & k_{-2}(-\mathrm{C}(t)) - k_3\mathrm{C}(t) + k_2\mathrm{E}(t)\mathrm{S}(t) \\ \mathrm{D}'(t) & = & k_{-5}(-\mathrm{D}(t)) - k_6\mathrm{D}(t) + k_5\mathrm{E}_2(t)\mathrm{P}(t) \\ \mathrm{E}^{-\prime}(t) & = & k_{-4}\mathrm{E}(t) - k_4\mathrm{E}^-(t)\mathrm{P}(t) \\ \mathrm{P}'(t) & = & k_3\mathrm{C}(t) + k_{-5}\mathrm{D}(t) + k_{-4}\mathrm{E}(t) - k_4\mathrm{E}^-(t)\mathrm{P}(t) - k_5\mathrm{E}_2(t)\mathrm{P}(t) \\ \mathrm{S}'(t) & = & k_{-2}\mathrm{C}(t) - k_2\mathrm{E}(t)\mathrm{S}(t) + k_1G(t) \end{array}\right. \tag{6}$$

with conservation laws,

$$\begin{array}{l} \mathrm{G}(t) = \mathrm{G}_0 \\ \mathrm{D}(t) + \mathrm{E}_2(t) = \mathrm{E}_2(0) \\ \mathrm{C}(t) + \mathrm{E}(t) + \mathrm{E}^-(t) = \mathrm{E}(0) \end{array} \tag{7}$$

where G_0 is the concentration of the gene G associated with the production of the protein S, and $\mathrm{E}(0)$ and $\mathrm{E}_2(0)$ are the concentrations of the enzymes E and E_2, respectively.

Solving the conservation equations (7) and the steady state conditions

$$\mathrm{C}'(t) = 0\,, \mathrm{D}'(t) = 0 \text{ and } \mathrm{E}^{-\prime}(t) = 0$$

in order to E, E_2, C, D and E^-, we obtain,

$$\left\{\begin{array}{l} \mathrm{E}_2(t) = \dfrac{\mathrm{E}_2(0)\,(k_{-5} + k_6)}{k_5P(t) + k_{-5} + k_6} \\ \mathrm{E}(t) = \dfrac{\mathrm{E}(0)\,(k_{-2} + k_3)\,k_4P(t)}{k_4P(t)\,(k_2S(t) + k_{-2} + k_3) + k_{-4}\,(k_{-2} + k_3)} \\ \mathrm{C}(t) = \dfrac{\mathrm{E}(0)k_2k_4P(t)S(t)}{k_4P(t)\,(k_2S(t) + k_{-2} + k_3) + k_{-4}\,(k_{-2} + k_3)} \\ \mathrm{D}(t) \to \dfrac{\mathrm{E}_2(0)k_5P(t)}{k_5P(t) + k_{-5} + k_6} \\ \mathrm{E}^-(t) = \dfrac{\mathrm{E}(0)k_{-4}\,(k_{-2} + k_3)}{k_4P(t)\,(k_2S(t) + k_{-2} + k_3) + k_{-4}\,(k_{-2} + k_3)}\,. \end{array}\right. \tag{8}$$

Introducing (8) into (6), the system of equations (6) simplify to,

$$\left\{\begin{array}{l} \mathrm{S}'(t) = \nu - f(\mathrm{S},\mathrm{P}) \\ \mathrm{P}'(t) = f(\mathrm{S},\mathrm{P}) - \dfrac{\beta_2\mathrm{P}(t)}{\mathrm{P}(t) + \alpha_3} \end{array}\right. \tag{9}$$

where,

$$f(\mathrm{S},\mathrm{P}) = \frac{\beta_1 \mathrm{P}(t)\mathrm{S}(t)}{\mathrm{P}(t)\left(\mathrm{S}(t)+\alpha_1\right)+\alpha_2}$$

and

$$\begin{gathered} \nu = \mathrm{G}_0 k_1,\ \beta_1 = \mathrm{E}(0)k_3,\ \beta_2 = \mathrm{E}_2(0)k_6 \\ \alpha_1 = \frac{k_{-2}+k_3}{k_2},\ \alpha_2 = \frac{k_{-4}}{k_4}\alpha_2,\ \alpha_3 = \frac{k_{-5}+k_6}{k_5} \end{gathered} \tag{10}$$

are positive parameters.

The differential equation (9) has a unique and positive fixed point with coordinates,

$$\begin{cases} \mathrm{S}^* = \dfrac{\alpha_1\alpha_3\nu + \alpha_2(\beta_2-\nu)}{\alpha_3(\beta_1-\nu)} \\ \mathrm{P}^* = \dfrac{\alpha_3\nu}{(\beta_2-\nu)} \end{cases} \tag{11}$$

provided $(\beta_1-\nu) > 0$ and $(\beta_2-\nu) > 0$. The differential equation (9) has a Hopf bifurcation in the vicinity of the fixed point (11) if the trace of the Jacobian matrix of the vector field defined by (9) and calculated at fixed point $(\mathrm{S}^*,\mathrm{P}^*)$ is zero and the corresponding determinant is positive. So, from this condition, it follows easily that equation (9) has a stable limit cycle in phase space if,

$$\alpha_2(\beta_1-\beta_2)(\beta_2-\nu)^2 > \alpha_3^2\beta_2(\beta_1-\nu)^2 + \alpha_1\alpha_3\beta_1(\beta_2-\nu)^2 \tag{12}$$

where necessarily, $\beta_1 > \beta_2$ and $\alpha_3 > 0$.

In figure 2, we show the region of the parameter space (β_1,β_2) for which equation (9) has a stable limit cycle in phase space and therefore its solutions show stable sustained oscillations. Numerical simulations have shown that the solutions differential equation (9) has stiffness behaviour for α_3 close to zero. In the case of existence of stable limit cycles this is an indication of existence of relaxation oscillations.

In figure 3, we show three limit cycle solutions of equation (9) and the corresponding dependence on time of S and P, for the points A, B and C indicated in the bifurcation diagram of figure 2.

From the condition (12), a necessary condition for existence of sustained oscillations is $\beta_1 > \beta_2$, which, by (10), implies that,

$$\mathrm{E}(0)k_3 > \mathrm{E}_2(0)k_6 \tag{13}$$

As we see from figure 2, with the parameters ν, α_1, α_2 and α_3 fixed, it is possible to control the existence of stable oscillations by changing β_1 or β_2.

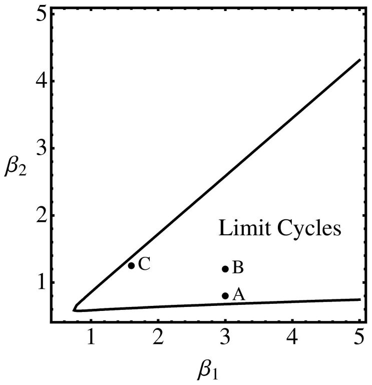

Figure 2: Region of the parameter space (β_1, β_2) for which equation (9) has a stable limit cycle in phase space. The region of existence of stable limit cycles has been calculated with condition (12). The other parameters are fixed and have the values: $\nu = 0.5$, $\alpha_1 = 1.0$, $\alpha_2 = 1.0$ and $\alpha_3 = 0.12$. Point A has coordinates, $\beta_1 = 3.0$ and $\beta_2 = 0.8$. Point B has coordinates, $\beta_1 = 3.0$ and $\beta_2 = 1.2$, and for point C, $\beta_1 = 1.6$ and $\beta_2 = 1.25$.

From the biological point of view, by (13), this change is equivalent to a change in the concentration of the enzymes of the mechanism of figure 1a). This will be discussed in more detail below.

3 A minimal model for the cell cycle

The cell cycle is divided into four phases: the phases G1, S, G2 and M. During the phase G1, the new cell absorb nutrients, synthetize mRNA and proteins and grows. During the phase S, the synthesis phase, chromosome replication occurs. The phase G2 is the preparation for mitosis and the cell continues to synthetize mRNAs and proteins. In the phase M, in most eukaryotes, the nuclear envelope breaks and begins a complicated process of separation of chromosomes preparing the cellular division (cytokinesis) at the end of phase M. After these processes, the cell enters again in the phase G1.

Human somatic cells complete a full cell cycle in 24h. Mitosis (M) takes 30 minutes, the phase G1 last 9h, the phase S last 10h and the phase G2 takes 4.5h to complete. In yeast cells, the full cell cycle takes 90 minutes.

The transitions between the different developmental stages of the cell are controlled by checkpoints. The checkpoint mechanisms determines if the cell remains in its actual state or if it makes the transition to the next cell stage.

All these processes are controlled by two families of proteins, the cyclins and the Cdk enzymes or cyclin dependent kinases.

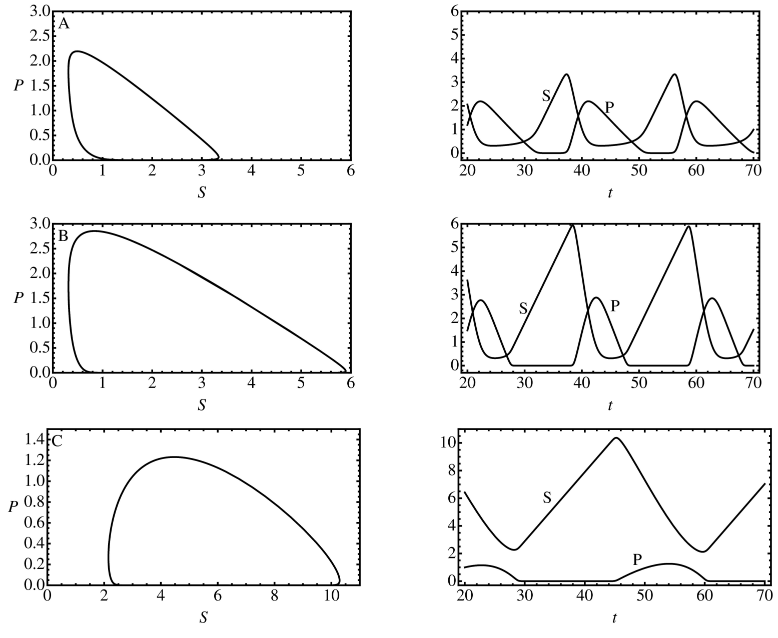

Figure 3: Limit cycle solutions of equation (9), for the points A, B and C shown in figure 2. On the right side we show the time evolution of P(t) and S(t). In A: $\beta_1 = 3.0$ and $\beta_2 = 0.8$. In B: $\beta_1 = 3.0$ and $\beta_2 = 1.2$. In C: $\beta_1 = 1.6$ and $\beta_2 = 1.25$. The other parameter values are $\nu = 0.5$, $\alpha_1 = 1.0$, $\alpha_2 = 1.0$ and $\alpha_3 = 0.12$.

The Cdk enzymes are kinase proteins (enzymes), also called phosphotransferases, whose function is to transfer phosphate groups in a process called phosphorylation. The cyclins control the progression of the cell cycle. Their effect is to activate the Cdk enzymes, forming an enzymatic complex. These complexes in its phosphorylated form trigger the different processes of the cell cycle, including protein synthesis, chromosome duplication, mitotic spindler formation, protease, etc.. The oscillations of different cyclins determine the cell fate and, for example, dictate if a cell stays longer in a specific phase or proceeds to the next stage. In particular, the mechanism of cell death is strongly dependent on the effectiveness of the cell cycle control. For biological details see [15, 1].

The starting points of the control of cell cycle is the the formation of the Maturation Promoting Factor (MPF). This complex is responsible for the signaling of mitosis initiation. As it is known from experiment from frog eggs, MPF can drive cells into mitosis without finishing the DNA replication. This suggest the existence of a cytoplasmic oscillator, [15, p. 26].

This MPF complex is a compound formed with one cyclin-B molecule and one Cdk molecule Cdc2 (MPF=cyclin-B + Cdc2). The beginning of mitosis is marked by the phosphorylation or activation of this complex. This phosphorylated or active form of MPF is denote by $(\mathrm{MPF})^+$. During the cell cycle, the concentration of cyclin-B increases during interphase but drops down at the exit of mitosis, after the initiation of anaphase.

During interphase, we can assume that the concentration of cyclin-B is under genetic control, and MPF is formed as cyclin-B is available. These processes can be described by the kinetic mechanisms,

$$\begin{array}{l} \mathrm{G} \xrightarrow{k_1} \mathrm{G} + \mathrm{cyclinB} \\ \mathrm{cyclinB} \xrightarrow{k_2} \\ \mathrm{Cdc2} + \mathrm{cyclinB} \xrightarrow{k_3} \mathrm{MPF} \end{array} \tag{14}$$

The activation of the MPF is a complex process involving the Cdc25 protein. It is known that the Cdc25 promotes the activation (phosphorylation) of MPF, and the removal of Cdc25 prevents the entry in mitosis of the cell. This process is enzymatic and we can assume that it follows a Michalis-Menten type kinetics, occurring at a faster time scale when compared with the cellular processes. Therefore, we can represent the MPF activation by the mechanism,

$$\mathrm{MPF} + \mathrm{Cdc25} \underset{k_{-4}}{\overset{k_4}{\rightleftarrows}} \mathrm{C} \xrightarrow{k_5} \mathrm{Cdc25} + (\mathrm{MPF})^+ \tag{15}$$

The activate or phosphorylated state of MPF — MPF^+ —- activates the activity of Cdc25. This feedback activation loop is represented by the mechanism, [15, p. 62],

$$(\mathrm{MPF})^+ + \mathrm{Cdc25}^- \underset{k_{-6}}{\overset{k_6}{\rightleftarrows}} \mathrm{Cdc25} \tag{16}$$

The MPF^+ complex then promotes the formation of the Anaphase Promoting Complex (APC) that triggers the beginning of anaphase. So, this subsequent enzymatic reaction can be represented by the mechanism,

$$\mathrm{MPF}^+ + \mathrm{E}_1 \underset{k_7}{\overset{k_{-7}}{\leftrightarrows}} \mathrm{D} \xrightarrow{k_8} \mathrm{E}_1 + \mathrm{APC} \tag{17}$$

where E_1 represent a non specific enzyme.

The mechanisms just described form the basic process that determines the entry of a cell into mitosis. In figure 4, we show the schematic diagram of the basic process just described, including the passage from the S/G2 checkpoint. In the diagram of the figure, arrows represent chemical transformations as in chemical kinetics. Dashed arrows represent irreversible Michaelis-Menten complex mechanisms.

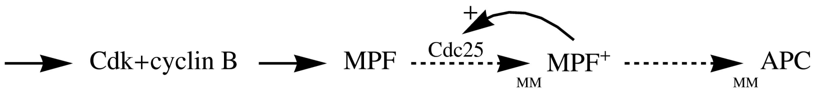

Figure 4: Minimal mechanism of control of the cell cycle. The passage from the checkpoint in the transition from phase S to phase G2 should emerge from the properties of this minimal model.

The reaction mechanisms in (14)-(17) are described by the set of independent differential equations,

$$\left\{ \begin{array}{lcl} \mathrm{C}'(t) & = & -k_{-4}\mathrm{C}(t) - k_5\mathrm{C}(t) + k_4\mathrm{Cdc25}(t)\mathrm{MPF}(t) \\ \mathrm{D}'(t) & = & -k_{-7}\mathrm{D}(t) - k_8\mathrm{D}(t) + k_7\mathrm{E}_1(t)\mathrm{MPF}^+(t) \\ \mathrm{Cdc25}'(t) & = & k_{-4}\mathrm{C}(t) + k_5\mathrm{C}(t) - k_4\mathrm{Cdc25}(t)\mathrm{MPF}(t) \\ & & -k_{-6}\mathrm{Cdc25}(t) + k_6\mathrm{Cdc25}^-(t)\mathrm{MPF}^+(t) \\ \mathrm{Cdc25}^{-\prime}(t) & = & k_{-6}\mathrm{Cdc25}(t) - k_6\mathrm{Cdc25}^-(t)\mathrm{MPF}^+(t) \\ \mathrm{cyclinB}'(t) & = & -k_3\mathrm{Cdc2}(0)\mathrm{cyclinB}(t) - k_2\mathrm{cyclinB}(t) + k_1\mathrm{G}(t) \\ \mathrm{E}_1'(t) & = & k_{-7}\mathrm{D}(t) + k_8\mathrm{D}(t) - k_7\mathrm{E}_1(t)\mathrm{MPF}^+(t) \\ \mathrm{MPF}'(t) & = & k_{-4}\mathrm{C}(t) + k_3\mathrm{Cdc2}(0)\mathrm{cyclinB}(t) - k_4\mathrm{Cdc25}(t)\mathrm{MPF}(t) \\ \mathrm{MPF}^{+\prime}(t) & = & k_5\mathrm{C}(t) + k_{-7}\mathrm{D}(t) + k_{-6}\mathrm{Cdc25}(t) \\ & & -k_6\mathrm{Cdc25}^-(t)\mathrm{MPF}^+(t) - k_7\mathrm{E1}(t)\mathrm{MPF}^+(t) \end{array} \right. \tag{18}$$

with conservation laws,

$$\begin{array}{l} \mathrm{G}(t) = \mathrm{G}_0 \\ \mathrm{D}(t) + \mathrm{E}_1(t) = \mathrm{E}_{10} \\ \mathrm{C}(t) + \mathrm{Cdc25}(t) + \mathrm{Cdc25}^-(t) = \mathrm{Cdc25}(0) \end{array} \tag{19}$$

where $\mathrm{Cdc25}(0)$ and E_{10} are the initial concentrations of the enzymes Cdc25 and E_1, respectively. The constant G_0 is the concentration of the gene that produces cyclin-B and the concentration of Cdc2 is considered constant along all the cell cycle, $\mathrm{Cdc2}(t) = \mathrm{Cdc2}(0)$.

In order to simplify the model equations (18) we use the conservation laws (19) together with the additional steady state assumptions: $\mathrm{C}'(t) = 0$, $\mathrm{D}'(t) =$

0 and $\mathrm{Cdc25}^{-\prime}(t) = 0$. Solving all together the steady state equations with the conservation laws (19) in order to E_1, C, D, Cdc25 and $\mathrm{Cdc25}^{-}$, we obtain,

$$\left\{\begin{array}{l} \mathrm{E}_1(t) = \dfrac{\mathrm{E}_{10}\,(k_{-7}+k_8)}{k_7\mathrm{MPF}^+(t)+k_{-7}+k_8} \\ \mathrm{Cdc25}(t) = \dfrac{\mathrm{Cdc25}(0)\,(k_{-4}+k_5)\,k_6\mathrm{MPF}^+(t)}{k_6\mathrm{MPF}^+(t)\,(k_4\mathrm{MPF}(t)+k_{-4}+k_5)+k_{-6}\,(k_{-4}+k_5)} \\ \mathrm{C}(t) = \dfrac{\mathrm{Cdc25}(0)k_4k_6\mathrm{MPF}(t)\mathrm{MPF}^+(t)}{k_6\mathrm{MPF}^+(t)\,(k_4\mathrm{MPF}(t)+k_{-4}+k_5)+k_{-6}\,(k_{-4}+k_5)} \\ \mathrm{D}(t) = \dfrac{\mathrm{E}_{10}k_7\mathrm{MPF}^+(t)}{k_7\mathrm{MPF}^+(t)+k_{-7}+k_8} \\ \mathrm{Cdc25}^-(t) = \dfrac{\mathrm{Cdc25}(0)k_{-6}\,(k_{-4}+k_5)}{k_6\mathrm{MPF}^+(t)\,(k_4\mathrm{MPF}(t)+k_{-4}+k_5)+k_{-6}\,(k_{-4}+k_5)}\,. \end{array}\right. \tag{20}$$

Introducing (20) into (18), the model equations simplify to,

$$\left\{\begin{array}{lcl} \mathrm{cyclinB}'(t) & = & = -k_3\mathrm{Cdc2}(0)\mathrm{cyclinB}(t) - k_2\mathrm{cyclinB}(t) + k_1\mathrm{G}_0 \\ \mathrm{MPF}'(t) & = & k_3\mathrm{Cdc2}(0)\mathrm{cyclinB}(t) - f(\mathrm{MPF},\mathrm{MPF}^+) \\ \mathrm{MPF}^{+\prime}(t) & = & f(\mathrm{MPF},\mathrm{MPF}^+) - \dfrac{\beta_2\mathrm{MPF}^+(t)}{\mathrm{MPF}^+(t)+\alpha_3} \end{array}\right. \tag{21}$$

where,

$$f(\mathrm{MPF},\mathrm{MPF}^+) = \frac{\beta_1\mathrm{MPF}(t)\mathrm{MPF}^+(t)}{\mathrm{MPF}^+(t)\,(\mathrm{MPF}(t)+\alpha_1)+\alpha_2}$$

and

$$\begin{array}{c} \beta_1 = \mathrm{Cdc25}(0)k_5,\ \beta_2 = \mathrm{E}_{10}k_8 \\ \alpha_1 = \dfrac{k_{-4}+k_5}{k_4},\ \alpha_2 = \dfrac{k_{-6}}{k_6}\alpha_2,\ \alpha_3 = \dfrac{k_{-7}+k_8}{k_7} \end{array} \tag{22}$$

are positive parameters. Model equations (21) describe the biological mechanism of figure 4, for the control of the cell cycle.

The system of equations (21) has a unique fixed point with coordinates,

$$\left\{\begin{array}{l} \mathrm{cyclinB}^* = \dfrac{k_1\mathrm{G}_0}{k_3\mathrm{Cdc2}(0)+k_2} \\ \mathrm{MPF}^* = \dfrac{\alpha_1\alpha_3\nu+\alpha_2(\beta_2-\nu)}{\alpha_3(\beta_1-\nu)} \\ \mathrm{MPF}^{+*} = \dfrac{\alpha_3\nu}{(\beta_2-\nu)} \end{array}\right. \tag{23}$$

where,

$$\nu = \frac{k_1 k_3 \mathrm{Cdc2}(0) \mathrm{G}_0}{k_3 \mathrm{Cdc2}(0) + k_2}$$

By the analysis we have performed in the previous section, necessary conditions for the existence of sustained oscillations in the solutions of equations (21) are,

$$\begin{array}{lcl} (\beta_1 > \nu) & \Longrightarrow & \mathrm{Cdc25}(0) k_5 > \dfrac{k_1 k_3 \mathrm{Cdc2}(0) \mathrm{G}_0}{k_3 \mathrm{Cdc2}(0) + k_2} \\ (\beta_2 > \nu) & \Longrightarrow & \mathrm{E}_{10} k_8 > \dfrac{k_1 k_3 \mathrm{Cdc2}(0) \mathrm{G}_0}{k_3 \mathrm{Cdc2}(0) + k_2} \\ (\beta_1 > \beta_2) & \Longrightarrow & \mathrm{Cdc25}(0) k_5 > \mathrm{E}_{10} k_8 \end{array} \tag{24}$$

In figure 5, we show the solutions of equations (21) for the parameters values of point A in figure 3, and the additional parameter values, $k_2 = 0.1$, $k_3 = 0.1$ and $\mathrm{Cdc2}(0) = 1.0$. After a transient time, the oscillations are established and the systems describe qualitatively the interphase and the mitose phases of the cell cycle. The mitose phase corresponds to the regions where MPF^+ is high. Concerning the concentration of cycline-B, in this minimal model, we have not considered other dynamical processes involved in the dynamics of cyclins. From observations, it is know that a cell once in mitosis, and after the entry in anaphase phase, APC has a proteolytic action on cyclin-B, lowering its concentration to very low values. In this model, the cyclin-B is considered to be permanently produced without other effects. This justifies the constant values shown in figure 5.

One of the important properties of this minimal model is modulation of the Hopf bifurcation by changing the concentration of $\mathrm{Cdc25}(0)$ as shown in the first and third conditions in (24). By changing the concentration of this Cdk protein, we can force, the crossing of the Hopf bifurcation boundary of figure 2, and therefore forcing cells to enter mitosis or to be arrested in some phase of the cell cycle. This effect explains cell fusion experiments where cells can enter mitosis without finishing phase S. On the other hand, this mechanism can explain the effect of the S/G2 checkpoint on the progression of the cell to mitosis.

For example, from (24), it follows that lowering the concentration of Cdc25 in cells and tissues, can inhibit the entry of a cell into mitosis, preventing, for example, the proliferation of damaged cells. A new class of drugs with these properties are being tested, [22].

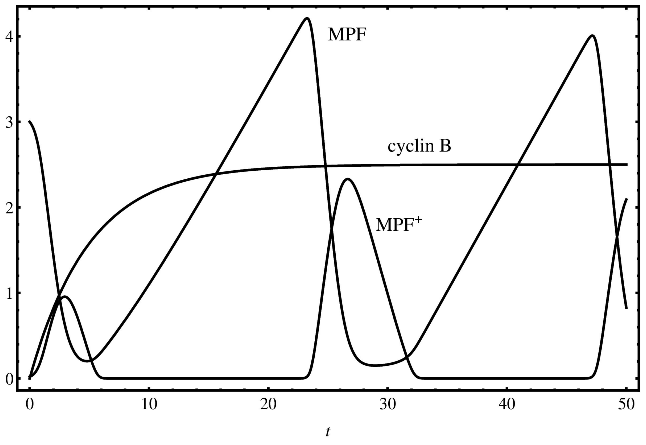

Figure 5: Limit cycle solutions of equation (21), with the parameters values: $\alpha_1 = 1.0$, $\alpha_2 = 1.0$ and $\alpha_3 = 0.12$, $\beta_1 = 3.0$, $\beta_2 = 0.8$, $k_2 = 0.1$, $k_3 = 0.1$, $k_1\text{G}_0 = 0.5$ and $\text{Cdc2}(0) = 1.0$. Fixing all the parameter values except β_2, the system of equations (21) has oscillatory solutions (stable limit cycle) for $\beta_2 \in [0.387, 2.588]$.

4 Glycolytic oscillations

In 1964, Higgins proposed that glycolytic oscillations in yeast cells could be understood as a sequential enzymatic mechanism, involving glucose (GLU), fructose-6-phosphate (F6P) and fructose diphosphate (FDP), [10]. The mechanism proposed by Higgins, can be summarize in the enzymatic chain reactions shown in figure 6. Comparing the diagrams in figure 6 and in figure 1a), we conclude that these mechanism are the same.

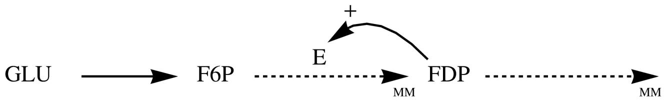

Figure 6: Higgins mechanism of glycolytic oscillations in yeast.

One of the conclusions we derive from this comparison is that this very simple biological mechanism is present in different biological system, showing that different biological systems can be described by the same biochemical mechanisms, even if their biological functions are different.

5 Conclusions

We have introduced a minimal model aiming to describe the cytoplasmic oscillator of the cell cycle. This cytoplasmic oscillator can drive cells into mitoses and insures that the transitions between cell phases are done in a specific order.

One of the control parameters of this model is the concentration of the Cdc25 protein. According to the analysis done here, the concentration of this protein simulates the effect of the S/G2 checkpoint of the cell cycle.

One of the consequences of this model is that lowering the concentration of Cdc25 in cells and tissues can inhibit the entry of a cell into mitosis, preventing, for example, the proliferation of damaged cells.

The cytoplasmic oscillator model derived here is structurally similar to a model describing glycolytic oscillations in yeast. This fact, together with the property that this model is associated with the simplest biochemical mechanism that we can imagine that explains sustained oscillations, imply that it is conceivable that the mechanism analyzed here is the biochemical back-bone for several biological systems exhibiting oscillations.

Acknowledgements

This work has been developed under the Fundação para a Ciência e Tecnologia (FCT) project PTDC/SAU-ENB/118044/2010.

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, (2008) Molecular Biology of the Cell, 5th Edition, *Taylor & Francis*, Abingdon.

[2] F. Alves and R. Dilão, (2005) A simple framework to describe the regulation of gene expression in prokaryotes, *Comptes Rendus - Biologies* **328**: 429-444.

[3] D. Coudreusse and P. Nurse, (2010) Driving the cell cycle with a minimal control network, *Nature* **468**: 1074-1078.

[4] R. Dilão and D. Muraro, (2010) A software tool to model genetic regulatory networks. Applications to the modeling of threshold phenomena and of spatial patterning in Drosophila, *PLoS ONE* **5** (5): e10743.

[5] http://sd.ist.utl.pt/NonLinear_Dynamics_Group/Software.html.

[6] A. N. Dodd, J. M. Gardner, C. T. Hotta, K. E. Hubbard, N. Dalchau, J. Love, J.-M. Assie, F. C. Robertson, M. K. Jakobsen, J. Gonçalves, D. Sanders, A. A. R. Webb, (2007) The Arabidopsis Circadian Clock Incorporates a cADPR-Based Feedback Loop, *Science* **318**: 1789-1792.

[7] A. Goldbeter, (1991) A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase, *Proc. Nat. Acad. Sci. USA* **88**: 107-111.

[8] A. Goldbeter, (1996) Biochemical Oscillations and Cellular Rhythms, *Cambridge University Press*, Cambridge.

[9] A. K. Ghosh and B. Chance, (1964) Oscillations of glycolytic intermediates in yeast cells. *Biochem. Biophys. Res. Commun.* **16**: 174-181.

[10] J. Higgins, (1964) A chemical mechanism for oscillations of glycolytic intermediates in yeast cells, *Proc. Nat. Acad. Sci. USA* **51**: 989-994.

[11] J. Higgins, (1967) The theory of oscillating reactions, *Ind. Eng. Chem.* **59**: 18-62.

[12] R. H. Kessin, (2001) Dictyostelium. Evolution, Cell Biology, and the Development of Multicellularity, *Cambridge University Press*, Cambridge.

[13] W. Klonowski, (1983) Simplifying principles for chemical and enzyme reaction kinetics, *Biophysical Chemistry* **18**: 73-87.

[14] J. Monod and F. Jacob, (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation, *Cold Spring Harbor symposia on quantitative biology* **26**: 389-401.

[15] A. Murray and T. Hunt, (1993) The Cell cycle, an introduction, *Oxford University Press*, Oxford.

[16] J. D. Murray, (1993) Mathematical Biology, *Springer*, Berlin.

[17] J. S. O'Neill and Akhilesh B. Reddy, (2011) Circadian clocks in human red blood cells, *Nature* **469**: 498-504.

[18] J. J. Tyson, (1991) Modeling the cell division cycle: cdc2 and cyclin interactions, *Proc. Nati. Acad. Sci. USA* **88**: 7328-7332.

[19] J. J. Tyson, (1996) Chemical kinetic theory: understanding the cell-cycle regulation, *Trends Biochem. Sci.* **21**: 89-96.

[20] F. Miyoshi, Y. Nakayama, K. Kaizu, H. Iwasaki, and M. Tomita, (2007) A mathematical model for the Kai-protein-based chemical oscillator and clock gene expression rhythms in cyanobacteria, *Journal of Biological Rhythms* **22**: 69-80.

[21] E. E. Selkov, (1968) Self-Oscillations in glycolysis, *European J. Biochem.* **4**: 79-86.

[22] A.M. Senderowicz, (1992) Flavopiridol: the first cyclin-dependent kinase inhibitor in human clinical trials, *Invest New Drugs* **17**(3): 313-320.

The penultimate goal of Synthetic Biology

Vic Norris[1,2]

[1] AMMIS Laboratory, EA 3829, Department of Biology, University of Rouen, F-76821 Mont Saint Aignan, France

[2] Epigenomics Project, Genopole Campus 1, Bât. Genavenir 6, 5 rue Henri Desbruères, F-91030 Évry Cedex, France

Abstract

There are several ways that our species might try to send a message to another species separated from us by space and/or time. Synthetic biology might be used to write "Kilroy was here" into the patterns of codons in the genome of a bacterium. I suggest here how this pattern might be used to create DNA movies. I also suggest that this may be a useful way to analyse DNA and I speculate unashamedly that a message from aliens may already exist in the genomes of cyanobacteria and other bacteria.

1 Introduction

The urge to leave traces of ourselves is revealed by the pictures in our museums, by the books in our libraries and by the tags on the walls in our cities. This urge led to the message written onto the gold-anodized aluminium plaques on the Pioneer 10 and 11 probes sent out by NASA. The same urge might be harnessed to send a message to future species on our own planet. The question would then arise as to how such a message might be written. Attempts to answer this question risk crossing the line that separates science from science fiction. Sometimes, however, breaching the divide between scientific speculation and science fiction can be desirable. Indeed, it has been encouraged by the physicist and science fiction writer, Gregory Benford, who proposed that there is "a link between the science I practise, and the fiction I deploy in order to think about the larger implications of my work, and of others'." [1]. In other words, allowing one's imagination to explore new possibilities in the writing of science fiction can be of value to real science. I use this to license the following series of speculations about intelligent life in the universe, its likely desire to communicate, and the use it may make of synthetic biology to write 'Kilroy was here' as an epitaph to *Homo sapiens* in the genome of bacteria.

2 The Problem

It is conceivable that "intelligent", dominant, life-forms like ours have arisen previously on Earth. It is even conceivable that they have arisen - and will continue to arise - many times. The problem for species such as *Homo sapiens* (or, as we might prefer to call it, *Homo systemicus*) is that the selection for tribalism, aggression, power-seeking and, above all, obedience to the group (i.e. uncritically adopting its beliefs and values), that leads to their dominance is also likely to lead to their destruction. It might be argued that no evident trace of such species has been found, as yet, in the fossil record. This might seem a powerful counter-argument given the effects of *Homo sapiens* on the ecosystem (e.g. via the relative abundance of pollens) or on fossilised artefacts (e.g. via our sophisticated tools). A possible explanation for this would be that such species have destroyed themselves so rapidly that they have left little trace behind. *Homo sapiens* may have lasted longer than most because its low intelligence relative to earlier species has retarded its development of weapons of mass destruction (e.g. of psychic, literally mind-blowing, weapons). Given awareness of its transience, an intelligent species (like many individuals) may want to want to leave a message for a future species, either just a "Kilroy was here" or some more interesting "message in a bottle". But how could they do it so as to ensure that it could be read after tens or hundreds of millions of years?

3 Possible approaches

One way would be to create artefacts on Earth along the lines of a modern equivalent of the pyramids. It is unlikely though that such artefacts could be constructed to last more than a few tens of thousands of years rather than a few hundreds of millions of years that be may needed for them to be interpreted [14]. The precursors of the pyramids, the mastabas, are already in a poor state despite the good conditions for preservation that have prevailed in Egypt. And little that we might construct is likely to survive a trip down a subduction zone. Another way would be to leave the message somewhere in the Solar System, perhaps to put it in a Lagrange Point, where one might hope it would stay for some time (I may be wrong here), or to send it off into the great black yonder as in the case of Pioneer 10 and 11. Yet another way would be to make use of biology.

4 The biological solution

Bacteria have the advantage of being able to maintain themselves unscathed over millions of year in different conditions in which, to take the extremes, they can either grow by faithful reproduction or survive by sporulation. How then might bacteria be used to send a message across the aeons or the light years? One possibility would be to write it in the DNA of an organism that was likely to be sequenced. This raises the question of how to encode the message. Suppose, for example, we were to take a circular bacterial chromosome like that of *Escherichia coli* and to use the sequences of the two replichores (i.e. the two *oriC-terC* sequences) as the axes of a 2-D matrix. (There are, of course, other possibilities such as taking the entire linear sequence from *oriC* back to *oriC* and then using that same sequence for both axes of the 2-D matrix). If one were to attribute a colour to each nucleotide base pair combination of x,y coordinates, this might be used to make a pretty pattern. It might be used to make a still prettier one if one were to use pairs of amino acids as combinations or simply pairs of similar/identical amino acids. It might be make a more interesting and accessible code if use were made of natural punctuation marks in the chromosome. Then the sequence it could be divided up so as to make a series of frames and a movie could be made out of it.

5 Implementation of DNA movies

“Punctuation marks” are the basis of the solenoid model for chromosome folding via the DNA-binding sites for sequence-specific transcriptional regulators which are located at regular distances that are multiples of 1/50th of the chromosome length [6]. Other periods such as the 33 kb in E. coli have been revealed by analysis of its ’core’ genes and may be based on requirements for translation and possibly transcription [8]. These are not the only results (see for example [3, 13]. Overall, periodicities may reflect two or more opposing constraints acting on the system. For example, there may be one constraint for unexpressed DNA and a different constraint for expressed DNA. The former constraint might correspond to maximising a reversible packing of DNA that would be obtained by the spontaneous adoption of a cholesteric structure as guided perhaps by sequences favouring high curvature [9] at the end of the loops or by other, as yet known, factors [4]. (An easily testable prediction is that this type of periodicity should be more evident in DNA that is largely untranscribed as in much of the DNA found in some dinoflagellates [7, 2].) The latter constraint might correspond to maximising the efficiency of translation by, for example, having all the codons for a particular amino acid translated near one another, which might be achieved by a particular set of 3-D distributions of codons within the cytoplasm and hence a particular set of 1-D distributions of these codons along the chromosome.

Irrespective of the exact nature and function of natural periodicities, the idea here is to exploit them as the frames in which a message is encoded. In one type of coding, this message would be in the position of pairs of amino acids in the two *ori-ter* arms or replichores. This would require extensive modification of existing coding sequences. It would require genes to be shifted from one location to another and for the sequences of individual genes to be extensively modified. But it is feasible. For example, the genes encoding nitrogen fixation in *Klebsiella* have been extensively modified with codons being exchanged and unwanted sites for regulation being removed [11].

6 The problems

One problem is that the synthetic messenger, which we might term *Escherichia nuntius* or *Nostoc nuntius* (depending on its origin), is likely to undergo so many recombinations, rearrangements and mutations that the message will be lost. A partial solution might be to use a slimmer version of E. coli from which elements that favour recombination have been removed [10]. In addition, key elements in the message might be carried by proteins that interact with several partners as in the case of ribosomal proteins since mutations in such proteins have an increased risk of disrupting an interaction important for survival.

Another problem is that *E. nuntius* is likely to be outcompeted by its natural competitors which have a billion years of selection on their side. This is not a problem if the conditions do not allow growth and only survival is important. (Note here that bacteria are reported to have been resuscitated after millions of years without growth [5].) Being outcompeted may not be a problem if *E. nuntius* can occupy fully the niche it is to grow in or if a new niche exists for which it can be designed to fill perfectly. Given the proportion of domestic animals compared to wild animals on Earth, *E. nuntius* could be added to feedstuffs so as to progressively replace the natural gut population.

7 Discussion

A first step in the construction of a synthetic messenger would be to make a matrix of the codons of a real bacterium and to experiment with the size and number of the frames to see what one gets. The joke here would be if this gave something non-random. The question would then be 'what does it mean?' In fact, this is not so silly. It is likely that bacteria have been selected so as to obtain optimal compromises between rates and fidelity of translation depending, for example, on growth conditions. One way to achieve this would be if ribosomes could tell the future - which, in principle, they could if a ribosome were to be informed of the codons that it would meet next by the preceding ribosomes (which have already met that codon). Suppose, for example, the tRNA used by ribosome$_t$ tells ribosome$_{t+1}$ which amino acid it will need next then, if recently used tRNAs were to increase the affinity of tRNA synthetases for one another, a functioning-dependent hyperstructure might form. Assembly of such a hyperstructure might profit greatly from a non-random distribution of codons in the group of genes that are expressed at any one time. The second joke would be to find that my proposal had already been acted on by an alien species and, funnier still, that the bacteria that made our world actually arrived via panspermia [12] - and contain a message. Reciprocally, absence of a message might be interpreted as indicating that there is no species out there - or back there - that wants to communicate with us.

Acknowledgements

For support, I thank the Epigenomics Project.

References

[1] Benford, G. (1995) Old legends. *In: New legends. G. Bear (ed). London: Legend Books (Random House UK)*, pp. 292-306.

[2] Chow, M. H., K. T. Yan, M. J. Bennett & J. T. Wong (2010) Liquid Crystalline Chromosomes: Birefringence and DNA Condensation. *Eukaryotic Cell* **9**: 1577-1587.

[3] Cook, P. R. (2002) Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.* **32**: 347-352.

[4] Danilowicz, C., C. H. Lee, K. Kim, K. Hatch, V. W. Coljee, N. Kleckner & M. Prentiss (2009) Single molecule detection of direct, homologous, DNA/DNA pairing. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 19824-19829.

[5] Fish, S. A., T. J. Shepherd, T. J. McGenity & W. D. Grant (2002) Recovery of 16S ribosomal RNA gene fragments from ancient halite. *Nature* **417**: 432-436.

[6] Kepes, F. (2004) Periodic transcriptional organization of the E.coli genome. *Journal of molecular biology* **340**: 957-964.

[7] Livolant, F. Y. & Y. Bouligand (1978) New observations on the twisted arrangement of dinoflagellate chromosomes. *Chromosoma* **68**: 21-44.

[8] Mathelier, A. & A. Carbone 2010) Chromosomal periodicity and positional networks of genes in *Escherichia coli*. *Molecular systems biology* **6**: 366.

[9] Pedersen, A. G., L. J. Jensen, S. Brunak, H.-H. Staerfeldt & D. W. Ussery (2000) A DNA structural atlas for *Escherichia coli*. *Journal of molecular biology* **299**: 907-930.

[10] Posfai, G., G. Plunkett III, T. Feher, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. de Arruda, V. Burland, S. W. Harcum & F. R. Blattner (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science (New York, N.Y)* **312**: 1044-1046.

[11] Temme, K., D. Zhao & C. A. Voigt (2010) Refactoring the nitrogen fixation gene cluster with synthetic biology tools. *In: 9th European Nitrogen Fixation Conference. Geneva*, pp. 29.

[12] Wickramasinghe, C. (2004) The universe: a cryogenic habitat for microbial life. *Cryobiology* **48**: 113-125.

[13] Wright, M. A., P. Kharchenko, G. M. Church & D. Segre (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 10559-10564.

[14] Zalasiewicz, J. (2009) The Earth after us. *Oxford University Press, Oxford* pp. 1-251.

PART III POSTERS

Development of computational methods for predictive simulation of TGF-β signaling pathway

Geoffroy Andrieux[1], Nolwenn Le Meur[1,2], Michel Le Borgne[2]
and Nathalie Theret[1]

[1] IRSET EA 4427 SeRAIC, IFR140, Univ. of Rennes1, F-35000, Rennes, France
[2] IRISA, Univ. of Rennes1, F-35000, Rennes, France

Abstract

Increasing evidence supports a role for the microenvironment as the major player in signaling pathways, however, lack of a dynamic integrated perspective constitute a strong impediment to the understanding of cell responses to microenvironment. Based on mathematical models and computational methods, systems biology have been recently developed to understand the interactions between components of a biological system and how these interactions give rise to the function and behavior of that system. Complex signaling by the transforming growth factor TGF-β is one of the most intriguing networks that governs complex multifunctional profiles and plays a pivotal role during chronic liver disease by activating the hepatic stellate cells and promoting tissue remodeling. TGF-β signals through a heteromeric complex of transmembrane serine/threonine kinases, the type I (TβRI) and type II (TβRII) receptors which transduces signals to downstream intracellular substrates, the Smad proteins. Alternatively, non-Smad pathways involved in TGF-β signaling include the Rho-like GTPase and PI3K/AKT pathways. Hence, combinations of SMAD and nonSMAD pathways might contribute to the diversity of cellular responses to TGF-β[4].

We develop differential model including the nuclear Transcriptional Intermediary Factor (Tif1γ) which was recently described as a new controversial regulator of Smad activity either through binding to Smad4 or Smad2-Smad3[3, 5, 6]. We took advantages of previous models of Smad shuttling[2] and receptors trafficking[7] to develop an integrated TGF-β signaling model which included Tifγ1. Dynamic simulation of the differential equation system demonstrated that controversial observations were compatible according to the Tif1γ/Smad4 ratio present in the cells. In addition we propose that Tif1γ, Smad4 and Smad2 might be transitory involved in a unique complex within the nucleus.

In a different way, we work on discrete model integrating both the Smad and non Smad dependant pathways to investigate the dual role of TGF-β on

cell cycle. Biological observations are extracted form literature and parsed in a database. To translate biological knowledge into a formal model built on the state concept, and to specify the evolution equation of each variable, we develop a new language based on State-chart formalism[1]. A computational environment for the design of statecharts graphs, software for the compilation of graphs in logic language and algorithms for simulation are under development.

References

[1] D. Harel, (1987) Statecharts : a visual formalism for complex systems. *Science of computer programming* **8**: 261-274.

[2] Schmierer, B., Tournier, A. L., Bates, P. A., and Hill, C. S., (2008) Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *PNAS* **105**: 6608-6613.

[3] He, W., Dom, D. C., Erdjument-Bromage, H., Tempst, P., Moore, M. A. S., and Massague, J., (2006) Hematopoiesis Controlled by Distinct TIF1γ and Smad4 Branches of the TGFb Pathway *Cell* **125**: 929-941.

[4] Massague, J., (2008) TGFβ in Cancer. *Cell* **134**: 215-230.

[5] Dupont, S., Zacchigna, L., Cordenonsi, M., Soligo, S., Adorno, M., Rugge, M., ans Piccolo, S., (2005) Germ-Layer Specification and Control of Cell Growth by Ectodermin, a Smad4 Ubiquitin Ligase *Cell* **121**: 87-99.

[6] Dupont, S., Mamidi, A., Cordenonsi, M.,Montagner, M., Zacchigna, L., Adorno, M., Martello, G., Stinchfield,M. J.,Soligo, S., Morsut, L., Inui, M., Moro, S., Modena, N., Argenton, F., Newfeld, S. J., and Piccolo, S., (2009) FAM/USP9x, a Deubiquitinating Enzyme Essential for TGFb Signaling, Controls Smad4 Monoubiquitination *Cell* **136**: 123-135.

[7] Vilar, J. M. G., Jansen, R., and Sander, C., (2006) Signal Processing in the TGF-β Superfamily Ligand-Receptor Network. *PLOS computational biology* **2**: 36-45

Randomizing genome-scale metabolic networks

Areejit Samal[1,2] and Olivier C. Martin[1,3]

[1] Laboratoire de Physique Théorique et Modèles Statistiques, CNRS UMR 8626,
Université Paris-Sud, F-91405 Orsay Cedex, France
[2] Max Planck Institute for Mathematics in the Sciences,
Inselstr. 22, 04103 Leipzig, Germany
[3] Laboratoire de Génétique Végétale du Moulon, UMR 0320/UMR 8120,
Université Paris-Sud, F-91190 Gif-sur-Yvette, France

Abstract

Networks coming from protein-protein interactions, transcriptional regulation, signaling, or metabolism may appear to have ”unusual” properties. To quantify this, it is appropriate to randomize the network and test the hypothesis that the network is not statistically different from expected in a motivated ensemble. However, when dealing with metabolic networks, the straightforward randomization of the network generates fictitious reactions that are biochemically meaningless. Here we provide several natural ensembles for randomizing such metabolic networks. A first constraint is to use valid biochemical reactions. Further constraints correspond to imposing appropriate functional constraints. We explain how to perform these randomizations and show how they allow one to approach the properties of biological metabolic networks. The implication of the present work is that the observed global structural properties of real metabolic networks are likely to be the consequence of simple biochemical and functional constraints.

References

A. Samal, O.C. Martin, arXiv:1012.1473

LIST OF ATTENDEES

(April 1st, 2011)

AMAR Patrick	(pa@lri.fr)
ANDRIEUX Geoffroy	(geoffroy.andrieux@irisa.fr)
BALLET Pascal	(pascal.ballet@univ-brest.fr)
BLAU Anthony	(tblau@u.washington.edu)
BASU ROY Sayantani	(sbasu@moulon.inra.fr)
BATMANOV Kirill	(Kirill.BATMANOV@lifl.fr)
BERNOT Gilles	(bernot@unice.fr)
BEURTON-AIMAR Marie	(beurton@labri.fr)
BONNET Muriel	(muriel.bonnet@clermont.inra.fr)
CAPPUCCIO Antonio	(Antonio.Cappuccio@curie.fr)
CARDELLI Luca	(luca@microsoft.com)
CARTA Alfonso	(alfonso.carta@inria.fr)
COOK Peter	(peter.cook@path.ox.ac.uk)
CREMER Christoph	(cremer@kip.uni-heidelberg.de)
CSIKASZ-NAGY Attila	(csikasz@cosbi.eu)
DILÃO Rui	(rui@sd.ist.utl.pt)
DOULAZMI Mohamed	(mohamed.doulazmi@upmc.fr)
DUPONT Geneviève	(gdupont@ulb.ac.be)
FOURMENTIN Éric	(fondation@fourmentinguilbert.org)
GEWIRTZ Andrew	(agewirt@emory.edu)
HAREL David	(dharel@weizmann.ac.il)
HASTY Jeff	(hasty@bioeng.ucsd.edu)
HERDEWYN Piet	(Piet.Herdewijn@rega.kuleuven.be)
JANNIÈRE Laurent	(janniere@issb.genopole.fr)
JARAMILLO Alfonso	(Alfonso.Jaramillo@issb.genopole.fr)

JUNIER Ivan	(i.junier@gmail.com)
KAUFMAN Marcelle	(mkaufman@ulb.ac.be)
KÉPÈS François	(francois.kepes@issb.genopole.fr)
KING Ross	(rdk@aber.ac.uk)
LE FÈVRE François	(flefevre@genoscope.cns.fr)
LE GALL Pascale	(pascale.legall@issb.genopole.fr)
MAZAT Jean-Pierre	(Jean-Pierre.Mazat@phys-mito.u-bordeaux2.fr)
MOLINA Franck	(franck.molina@sysdiag.cnrs.fr)
NAVAILLES Jean Paul	(navailles@msn.com)
NORRIS Victor	(victor.norris@univ-rouen.fr)
PERES Sabine	(sabine.peres@lri.fr)
RADULESCU Ovidiu	(ovidiu.radulescu@univ-montp2.fr)
RUSSO Christophe	(crusso@moulon.inra.fr)
SAMAL Areejit	(areejit.samal@gmail.com)
SCHERRER Klaus	(scherrer.klaus@ijm.univ-paris-diderot.fr)
SCHUSTER Stefan	(schuster@minet.uni-jena.de)
SEPULCHRE J.-Alexandre	(jacques-alexandre.sepulchre@inln.cnrs.fr)
TRUSSART Marie	(mtrussart@hotmail.com)
YOUSFI Haifa	(haifayousfi@yahoo.fr)
ZELISZEWSKI Dominique	(dominique.zeliszewski@issb.genopole.fr)