Proceedings of
The Evry Spring School
on

Modelling Complex Biological Systems in the Context of Genomics

April 30th - May 4th, 2007

Edited by

Patrick Amar, François Képès, Vic Norris, Gilles Bernot

"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 six 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 *Organisation* subgroup has adopted a systems biology approach with the application and development of new programming languages to describe biological systems which it has been applying to problems in the growth and differentiation of plants and in the structure and functioning of mitochondria. 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 and in Bordeaux in 2006. The conferences in Evry in 2007 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 attendees, 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, Marie Dutreix, Jean-Louis Giavitto, Christophe Godin, Janine Guespin, François Képès, Jean-Pierre Mazat, Franck Molina, Victor Norris, Vincent Schächter, Philippe Tracqui.

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 contributed by speakers, abstracts from attendees, posters and lectures proposed by the epigenesis group 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 *Mercure* Hotel at Evry 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.

We would like to thank Paul Hossenlopp (Responsable de la formation, CNRS) for his encouragements to this 6th edition of the spring school.

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:

- Genopole® Evry:
  http://www.genopole.fr
- Open Network of Centres of Excellence in Complex Systems (ONCE-CS):
  http://complexsystems.lri.fr/Portal/tiki-index.php
- Centre National de la Recherche Scientifique (CNRS):
  http://www.cnrs.fr
- General Integration of the Applications of Complexity in Science (GIACS):
  http://www.giacs.org
- GDR CNRS 3003 Bioinformatique Moléculaire:
  http://www.gdr-bim.u-psud.fr
- Institut National de Recherche en Informatique et en Automatique (INRIA):
  http://www.inria.fr
- Fondation Scientifique Fourmentin-Guilbert:
  http://www.fourmentinguilbert.org

THE EDITORS

INVITED SPEAKERS

ALESSANDRA CARBONE	Université Pierre et Marie Curie, INSERM U511 (F)
CHRISTOPHE CHIPOT	Institut nancéien de chimie moléculaire Université Henri Poincaré - Nancy I (F)
HIDDE DE JONG	The Helix Research Group, INRIA Rhône-Alpes (F)
JACQUES DEMONGEOT	Université Joseph Fourier Grenoble I - IMAG (F)
ALBERT GOLDBETER	Faculté des Sciences, Université Libre de Bruxelles, Brussels (B)
JAMES FOSTER	IBEST, University of Idaho (US)
MICHEL JACQUET	Institut de Génétique et Microbiologie, Université Paris Sud, Orsay (F)
KRISTIAN LINDGREN	Chalmers University of Technology, Göteborg (SE)
REINHARD LIPOWSKY	Max Planck Institute für Kolloid und Grenzflächenforschung, Potsdam, (D)
JOHN MCCASKILL	Biomolecular Information Processing (BioMIP), Ruhr-University Bochum (D)
PHILIPPE MARLIERE	Genoscope, Evry, (F)
ERIC SACKMANN	California Polytechnic State University College of Engineering, (US)
LAURENT TRILLING	Université Joseph Fourier Grenoble I, IMAG-LSR (F)
GÜNTER VON KIEDROWSKI	Lehrstuhl für Organische Chemie, Ruhr Univ. Bochum (D)
JEAN WEISSENBACH	Genoscope, Evry (F)
MICHAEL YUDKIN	Kellogg College, University of Oxford (UK)

CONTENTS

Qualitative Simulation of the Carbon Starvation Response in *Escherichia coli*

Delphine Ropers[1], Hidde de Jong[1], Johannes Geiselmann[2]

[1] INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, F-38334 St-Ismier cedex, France
[2] Laboratoire Adaptation et Pathogénie des Microorganismes CNRS UMR 5163, Université Joseph Fourier, Faculté de Médecine-Pharmacie Domaine de la Merci, F-38700 La Tronche, France

Abstract

The adaptation of living organisms to their environment is controlled at the molecular level by large and complex networks of genes, mRNAs, proteins, metabolites, and their mutual interactions. In order to understand the overall behavior of an organism, we must complement molecular biology with the dynamic analysis of cellular interaction networks, by constructing mathematical models derived from experimental data, and using simulation tools to predict the behavior of the system under a variety of conditions. Following this methodology, we have started the analysis of the network of global transcription regulators controlling the adaptation of the bacterium *Escherichia col*i to environmental stress conditions. Even though *E. coli* is one of the best studied organisms, it is currently little understood how a stress signal is sensed and propagated throughout the network of global regulators, so as to enable the cell to respond in an adequate way. Using a qualitative method that is able to overcome the current lack of quantitative data on kinetic parameters and molecular concentrations, we have modeled the carbon starvation response network and simulated the response of *E. coli* cells to carbon deprivation. This has allowed us to identify essential features of the transition between exponential and stationary phase and to make new predictions on the qualitative system behavior following a carbon upshift. The model predictions have been tested experimentally by means of gene reporter systems.

Logic and Constraint Programming for constructing and analysing Logical Models of Regulatory Networks

Laurent Trilling[1]

[1] Université Joseph Fourier (Grenoble I), IMAG-LSR, Boîte postale 72 F-38402, St. Martin d'Hères, France

Abstract

The main advantage of a logic programming approach is well-known: a lot of different functionalities become available from a unique logical specification. It seems very well suited to the multivalued logic models proposed by René Thomas. From a unique description of the logical equations, one can ideally perform simulation (all logical parameters are known) or inference of parameters (the model must be consistent with known behaviours) or possibly intermediate requests (the model and some behaviours are partially known). In the case where the givens are consistent, prediction of unknown behaviours and elaboration of new experiments, become possible, both being safe for all acceptable models. These advantages are emphasized by using constraints which can be solved by efficient algorithms. After a brief recall of CLP (Constraint Logic Programming) and of Thomas's models, we will present successively: the main lines of a precise specification of these models with CLP, some theoretical problems arising from this specification, the requirements for the termination of a request, the methodology for analysing a network given by its structure and behaviours, some biological models analysed and/or revisited by this method and finally computational performances of the approach.

Robustness in genetic regulation networks and micro-RNAs

Jacques Demongeot[1]

[1] TIMC-IMAG, UMR CNRS-UJF 5525

Abstract

The dynamical behaviour of the genetic regulatory networks is very sensitive to exogeneous influences, like apparition of chromosomic or genic abnormalities perturbing the genetic memory or the transcription/traduction machinery. In the modelling of genetic networks, we can take into account these external perturbations through classical tools like sensitivity indices to the boundary conditions or to the architecture parameters perturbations. The endogeneous role of the small RNAs like the micro-RNAs seems to be more difficult to study, but fluctuations in presence/absence or in concentrations of these effectors could play a role as important as those of the external factors. We propose a general frame for studying the sensitivity to both external and internal factors exerting an influence on the regulatory networks depending on their high or small robustness.

Analysis of wastewater processing using a metagenomic approach

Jean Weissenbach[1]

[1] Genoscope, Evry, France

Abstract

Wastewater is processed through a number of biochemical pathways by microorganisms which progressively decompose organic matter into mineral components. Mineralisation is driven to a certain extent through a succession of steps which take place in aerobic or anaerobic conditions. We have only a partial view of the biochemical aspects and we know little about the prokaryotic florae involved and their exact contribution to the overall process. We have therefore applied genomic and metagenomic approaches to get more knowledge about the microbial actors and the biochemistry at work in wastewater processing.

We have successively used 16 rDNA analysis and constructed metagenomic DNA libraries of large insert fragments extracted from different basins of a municipal wastewater treatment plant. In particular a fosmid library of more than one million clones of DNA extracted from an anaerobic mesophilic digester was screened by high density filter hybridization using various 16S rDNA probes. Sequences of 16S rDNAs identified in the fosmids were compared to those obtained from PCR products on the DNA extracted from the florae of the basins. New bacterial divisions that represent significant fractions ($> 10\%$) of the prokaryotic population of the digester were identified and are being further studied.

End sequences of the large insert clones have been obtained and are being analyzed using several approaches including assembly, definition of open reading frames, coding sequence alignments etc. Results of such analyses will be presented. The sequence dataset has also been used to identify missing genes in biochemical pathways as well as novel enzyme activities and fermentation pathways.

Adaptation studied with the self-consistent codon index: genomic spaces, metabolic network comparison, minimal gene sets and viral classification

Alessandra Carbone[1]

[1] Laboratory of Analytical Genomics, Université Pierre et Marie Curie, INSERM UMRS 511, 91, Bd de l'Hôpital, F-75013 Paris, France;
e-mail: Alessandra.Carbone@lip6.fr

Abstract

Facts and ideas presented in this short review concern some recent developments at the interface between microbial spaces, metabolic network comparison, minimal gene sets and viral classification. The guiding line to all results presented here is to derive biological information from genome sequences by means of a purely statistical analysis and an appropriate design of algorithms. The paper is an updated version of (Carbone 2005).

1 Some background and motivation

Proteins are formed out of 20 amino-acids which are coded in triplets of nucleotides, called codons. The four nucleotides (A, T, C, G) define 64 codons used in the cell. Codons are not uniformly employed in the cell, but at the contrary, certain codons are preferred and we speak about *codon bias*. There are several kinds of codon biases and some of them are linked to specific biological functions. Statistical analysis of DNA sequences and in particular of codon bias were performed from the moment that long chunks of DNA sequences were publicly available in the early eighties (Grantham et al. 1980; Wada et al. 1990), and the roots for these studies can be traced back to the sixties (Sueoka 1962; Zuckerkandl and Pauling 1965). However with the increasing number of bacterial genome sequences from a broad diversity of species, this field of research has been revivified in the last few years (Koonin and Galperin 1997; Lin and Gerstein 2000; Radomski and Slonimski 2001; Knight et al. 2001; Sicheritz-Pont´en and Andersson 2001; Daubin et al. 2002; Lin et al. 2002; Lobry and Chessel 2003; Sandberg et al. 2003; Jansen et al. 2003).

Biased codon usage may result from a diversity of factors: GC-content, preference for codons with G or C at the third nucleotide position (Lafay et al. 1999), a leading strand richer in $G + T$ than a lagging strand (Lafay et al. 1999), horizontal gene transfer which induces chromosome segments of unusual base composition (Moszer et al. 1999), and in particular, translational bias which has been frequently noticed in fast growing prokaryotes and eukaryotes (Sharp and Li 1987; Sharp et al. 1986; M´edigue et al. 1991; Shields and Sharp 1987; Sharp et al. 1988; Stenico et al. 1994). Three main facts support the idea of "translational impact": highly expressed genes tend to use only a limited number of codons and display a high codon bias (Grantham et al. 1980; Sharp and Li 1987), preferred codons and isoacceptor tRNA content exhibit a strong positive correlation (Ikemura 1985; Bennetzen and Hall 1982; Bulmer 1987; Gouy and Gautier 1982), and tRNA isoacceptor pools affect the rate of polypeptide chain elongation (Varenne et al. 1984; Buckingham and Grosjean 1986).

To study the effect of translational bias on gene expression, Sharp & Li (Sharp and Li 1987) proposed to associate to each gene of a given genome a numerical value, called *Codon Adaptation Index or* CAI for short, which expresses its synonymous codon bias (see appendix for the definition). The idea is to compute a weight (representing relative adaptiveness) for each codon from its frequency within a chosen small pool of highly expressed genes S, and combine these weights to define the $CAI(g)$ value of each gene g in the genome. For Sharp et al., the hypothesis driving the choice of S is that, for certain organisms, highly expressed genes in the cell have highest codon bias, and these genes, made out of frequent codons, are representative for the bias. Based on this rationale, one can select a pool of ribosomal proteins, elongation factors, proteins involved in glycolysis, possibly histone proteins (in eukaryotes) and outer membrane proteins (in prokaryotes) or other selections from known highly expressed genes, to form the representative set S. Then, CAI values are computed and are checked to be compatible with genes known to be highly or lowly expressed in the cell. If this is the case, then predictions are drawn with some confidence on expression levels for genes and open reading frames, even with no known homologues. Even if conceptually clear, this framework has been misused several times in the literature and incorrect biological consequences have been derived for gene expression levels of organisms which do not display a dominant translational bias, as discussed in (Grocock and Sharp 2002). This confusion motivated us to search for a methodology based on a precise mathematical formulation of the problem to detect the existence of translational bias.

But the main motivation for us came from the recognition that an increasing number of genome sequences will be available for organisms for which biological knowledge consists merely of a sketched morphological and ecological description. For these organisms, it might not be evident how to define the reference set S, nor how to identify a reliable testing set which can ensure that predictions meet a satisfiable confidence level. Still, one would like to detect if translational bias holds for these genomes and if so, to predict their gene expression levels. If not, one would like to know the origin of their dominating bias and use this information for genome comparison.

2 An automatic detection of codon bias

We proposed a simple algorithm to detect dominating synonymous codon usage bias in genomes (Carbone et al. 2003). The algorithm is based on a precise mathematical formulation of the problem that leads to use the Self-Consistent Codon Index (SCCI) (strongly correlated to the CAI measure in translationally biased organisms) as a *universal* measure of codon bias, that is a measure for biases of possibly different origins (and not only for translational bias, as CAI was originally introduced for). With the set of coding sequences as a sole source of biological information, the algorithm provides a reference set S of genes which is highly representative of the bias. This set is then used to compute the Codon Adaptation Index of genes of prokaryotic and eukaryotic organisms, including those whose functional annotation is *not* yet available. An important application concerns the detection of a reference set characterizing translational bias which is known to correlate to expression levels in many bacteria and small eukaryotes; it detects also leading-lagging strands bias, GC-content bias, GC3 bias, and horizontal gene transfer. In general, the algorithm becomes a key tool to predict gene expression levels and to compare species. The approach is validated on 96 slow-growing and fast-growing bacteria and archaeal genomes, *Saccharomyces cerevisiae, Plasmodium falciparum, Caenorhabditis elegans* and *Drosophila melanogaster*.

3 Genomic signatures and a space of genomes for genome comparison

Based on this analysis, we propose a novel formal framework to interpret genomic relationships derived from entire genome sequences rather than individual loci. This space allows to analyse sets of organisms related by a common *codon bias signature* (at times, more than one kind of bias influences the same genomic sequence and the ensemble of these overlapped biases defines what we call the *signature* of a genome) (Carbone et al. 2004). We give a number of numerical criteria to infer content bias, translational bias and strand bias for genome sequences. We show in a uniform framework that genomes of quite different phylogenetic relationship share similar codon bias; other genomes grouped together by various phylogenetic methods, appear to be subdivided in finer subgroups sharing different codon bias characteristics; Archaea and Eubacteria share the same codon preferences when $AT3$ or $GC3$ bias is their dominant bias; archaeal genomes satisfying translational bias use a sharply distinguished set of preferred codons than bacterial genomes. Our analysis, based on 96 eubacterial and archaeal genomes, opens the possibility that this space might reflect the geometry of a prokaryotic "physiology space". If this turns out to be the case, the combination of the upcoming sequencing of entire genomes and the detection of codon bias signatures will become a valuable tool to infer information on the physiology, ecology and possibly on the ecological conditions under which bacterial and archaeal organisms evolved. For many organisms, this information would be impossible to be detected otherwise.

4 Study of metabolic networks through sequence analysis and transcriptomic data

Genes with high codon bias describe in meaningful ways the biological characteristics of the organism and are representative of specific metabolic usage (Carbone and Madden 2005). In silico methods exploiting this basic principle are expected to become important in learning about the lifestyle of an organism and explain its evolution in the wild. We demonstrate that besides high expressivity during fast growth or glycolytic activities which have been very often reported, the necessity for survival under specific biological conditions has its traces in the genetic coding (Carbone and Madden 2005). This observation opens the possibility to predict rare but necessary metabolic activities from genome analysis.

High expression of certain classes of genes, like those constituting the translational machinery or those involved in glycolysis, are correlated particularly well in the case of fast growing organisms. By shifting the paradigm towards metabolic pathways, we notice that several energy metabolism pathways are correlated with high codon bias in organisms known to be driven by very different physiologies, which are not necessarily fast growing and whose genomes might be very homogeneous. More generally, we derive a classification of metabolic pathways induced by codon analysis, show that genetic coding for different organisms is tuned on specific pathways and that this is a universal fact. The codon composition of enzymes involved in glycolysis for instance, often required to be rapidly translated, is highly biased by dominant codon composition across species (this is indicated by the high CAI value of these enzymes). In fast growers, the numerical evidence is definitely far more striking than for other organisms (that is, the absolute difference between the CAI value of these enzymes and the average CAI value for genes in the genome is "large"), but even for *Helicobacter pylori*, a genome of rather homogeneous

codon composition, enzymes involved in glycolytic pathways happen to be biased above average. In the same manner, one detects the crucial role of photosynthetic pathways for *Synechocystis* or of methane metabolism for *Methanobacterium*.

mRNA transcriptional levels collected during the *Saccharomices cerevisiae* cell cycle under diauxic shift (deRisi et al. 1997) (here, glucose quantities decrease in the media during cell cycle and yeast goes from fermentation to aerobic respiration), have been used to analyze the yeast metabolic network in a similar spirit as done with codon analysis. A classification of metabolic pathways based on transcriptomic data has been proposed, and we show that the metabolic classification obtained through codon analysis essentially "coincides" with the one based on (a large and differentiated pool of) transcriptomic data. Such a result opens the way to explaining evolutionary pressure and natural selection for organisms grown in the wild, and hopefully, to explain metabolism for slow-growing bacteria, as well as to suggest best conditions of growth in the laboratory.

5 Genomic signatures and minimal gene sets

Computational and experimental attempts tried to characterize a universal core of genes representing the minimal set of functional needs for an organism. Based on the increasing number of available complete genomes, comparative genomics (Mushegian and Koonin 1996, Makarova et al. 2003, Nesbøet al. 2001, Harris et al. 2003, Brown et al. 2001, Koonin 2003, Charlebois and Doolittle 2004) has concluded that the universal core contains less than 50 genes. In contrast, experiments (Itaya 1995, Kobayashi et al. 2003, Hutchison et al. 1999, Glass et al. 2006, Akerley et al. 2002, Gerdes et al. 2003, Hashimoto et al. 2005, Salama et al. 2004, Ji et al. 2001, Forsyth et al. 2002, Thanassi et al. 2002, Winzeler et al. 1999, Giavier et al. 2002, Kamath et al. 2003) suggest a much large set of essential genes (certainly more than several hundreds, even under the most restrictive hypotheses) which is dependent on the biological complexity and the environmental specificity of the organism. Highly biased genes, which are generally also the most expressed in translationally biased organisms, tend to be over-represented in the class of genes deemed to be essential for any given bacterial species. This association is far from perfect, nevertheless it allows to propose a new computational method based on SCCI to detect to a certain extent ubiquitous genes, non-orthologous genes, environment specific genes, genes involved in stress response and genes with no identified function but highly likely to be essential for the cell. Most of these groups of genes cannot be identified with previously attempted computational and experimental approaches. The large spread of lifestyles and the unusually detectable functional signals characterizing translationally biased organisms suggest to use them as reference organisms to infer essentiality in other microbial species. In (Carbone 2006), we analyse in detail 27 organisms belonging to a large variety of phylogenetic taxa, γ and δ proteobacteria, firmicutes, actinobacteria, thermococcales and methanosarcinales; they do not display strong GC nor AT content and they are characterized by different optimal growth temperatures (Carbone et al. 2004). We also discuss the case of small parasitic genomes, and data issued by the analysis are compared to previous computational and experimental studies.

6 Viral adaptation to microbial hosts

The methodology presented in this paper and the notion of SCCI used to study bacterial species have been recently used to analyse viral genomes and adaptation to their host

(Carbone 2007, unpublished). We showed how viruses belonging to known phylogenetic organisations are localized in confined regions of codon space depending on their codon composition and demonstrate that codon bias is a highly refined measure that allows to reconstruct close relationships among viruses of the same species, being able to distinguish very clearly sequences of relatively small evolutionary distance.

Finding a convincing viral classification which is independent from morphology is becoming particularly important nowadays due to a large amount of metagenomic data already available and promised to be available in years to come. Numerical methods to approach these questions are sick.

Appendix: some comments on the mathematical methods

In this text, a coding sequence is represented by a 64-dimensional vector, whose entries correspond to the 64 relative codon frequencies in the sequence. Recall that the frequency of a codon i in a sequence g is the number of occurrences of i in g (where g is intended to be split in consecutive non-overlapping triplets corresponding to amino-acid decomposition), and that the *relative frequency* of i in g is the frequency of i in g divided by the number of codons in g. For each vector representing a coding sequence, the sum of its entries must equal 1. Hence, a coding sequence is a point in the 64-dimensional space $[0 \cdots 1]^{64}$, where no special assumption is made on the space nor on the coordinate system.

For each genome sequence G and some set of coding sequences S in G, *codon bias* is measured with respect to its synonymous codon usage. Given an amino-acid j, its synonymous codons might have different frequencies in S; if $x_{i,j}$ is the number of times that the codon i for the amino-acid j occurs in S, then one associates to i a *weight* $w_{i,j}$ relative to its sibling of maximal frequency y_i in $\mathcal{S}$

$$w_{i,j} = \frac{x_{i,j}}{y_j}.$$

A codon with maximal frequency in S is called preferred among its sibling codons. *Self-Consistent Codon Index* $(SCCI)$ associated to g in G, is a value in $[0, 1]$, defined as

$$SCCI(g) = (\Pi_{k=1}^{L} wk)1/L$$

where L is the number of codons in the gene, and w_k is the weight of the k-th codon gene sequence. Genes with $SCCI$ value close to 1 are made by highly frequent codons.

When the reference set S is predefined to be a set of highly expressed genes in the organism, then the index issued by the $SCCI$ formula corresponds to the known *Codon Adaptation Index* introduced by Sharp & Li (Sharp and Li 1987).

All results cited here are obtained using very simple mathematical and algorithmic notions which are fully described in (Carbone et al. 2003; Carbone et al. 2004; Carbone and Madden 2005). The statistical analysis and numerical thresholds we propose are realized in a 64-dimensional codon space. Multivariance statistical methods have been employed as visualisation tools, but none of the formal results nor the biological conclusions are inferred from the 3 dimensional projections. Both space of genes and space of organisms in 64 dimensions, and distances between organisms are defined as ℓ_1-distances.

References

[1] Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N., Mekalanos, J.J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences USA*, **99**, 966–971.

[2] Bennetzen, J.L., Hall, B.D. (1982) Codon selection in Yeast. *Journal of Biological Chemistry*, **257**, 3026-3031.

[3] Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J.(2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics*, **28**, 281–285.

[4] Buckingham, R.H., Grosjean, H. (1986) The accuracy of mRNA-tRNA recognition. In : *Accuracy in molecular processes: its control and relevance to living systems*, ed. T.B.L. Kirkwood, R. Rosenberger and D.J. Galas, Chapman & Hall Publishers, London, 83-126.

[5] Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728-730.

[6] Carbone, A., Zinovyev, A. and Képès, F. (2003) Codon Adaptation Index as a measure dominating codon bias. *Bioinformatics*, **19**, 2005-2015.

[7] Carbone, A., Képès, F. and Zinovyev, A. (2004) Microbial codon bias and the organisation microorganisms in codon space. *Molecular Biology and Evolution*, 22(3):547561.

[8] Carbone, A., Madden, D. (2005) Insights on the evolution of metabolic networks from data and sequence analysis. *Journal of Molecular Evolution*, 61:456469.

[9] Carbone,A. (2005) Revisiting the codon adaptation index from a whole-genome perspective: gene expression, codon bias, and metabolic networks in the context of genomes comparison", *Handelingen van de Koninklijke Vlaamse Academie van Belgie voor Wetenschappen en Kunsten*, Proceedings of the Belgian Royal Academy of Sciences and Arts, Study Day on Genomics, 18 October 2003.

[10] Carbone, A. (2006) Computational prediction of genomic functional cores specific to different microbes, *Journal of Molecular Evolution*, 63(6):733-746, 2006.

[11] Carbone, A. (2007) Phage signatures in codon space and viral lifestyle, in preparation.

[12] Charlebois, R.L., Doolittle, W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Research*, **14**, 2469–2477.

[13] Daubin, V., Gouy, M., Perrière, G. (2002) A phylogenetic approach to bacterial evidence of a core of genes sharing a comon history. *Genome Research*, **12**, 1080-

[14] DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997) Exploring the metabolic and genetic control expression on a genomic scale. *Science*, **278**, 680-686.

[15] Forsyth, R.A. *et al.* (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol Microbiol.* **43**, 1387–1400.

[16] Gerdes, S.Y., Scholle, M.D., Campbell, J.M., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., Bhattacharya, A. *et al.*(2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, **185**, 5673–5684.

[17] Giaever, G. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

[18] Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison III, C.A., Smith, H.O., Venter, J.C. (2006) Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences USA*, **103**, 425-430.

[19] Gouy, M. and Gautier, Ch. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10, 7055-7070.

[20] Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8, r49-r62.

[21] Grocock, R.J., Sharp, P.M. (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, **289**, 131-139.

[22] Harris, J.K., Kelley, S.T., Spiegelman, G.B., Pace, N.R. (2003) The genetic core of the universal ancestor. *Genome Research*, **13**, 407–412.

[23] Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., Katayama, T., Kato, J. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol Microbiol*, **55**:137-49.

[24] Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., Venter, J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169.

[25] Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13-34.

[26] Itaya, M. (1995) An estimation of the minimal genome size required for life. *FEBS Lett.*, **362**, 257–260.

[27] Jansen, R., Bussemaker, H.J., Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome prespective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Research*, **31**, 2242-2251.

[28] Ji, Y. *et al.* (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, **293**, 2266–2269.

[29] Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.

[30] Knight, R.D., Freeland, S.J., Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, **2**, at http://genomebiology.com/2001/2/4/research/0010.

[31] Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P. *et al.* (2003), Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A*, **100**, 4678-4683.

[32] Koonin, E.V. (2003) Comparative genomics, minimal gene sets and the last common ancestor. *Nature Reviews Microbiology*, **1**, 127–136.

[33] Koonin, E.V., Galperin, M.Y. (1997) Prokaryotic genomes: The emerging paradigm of genomebased microbiology. *Curr. Opin. Genet. Dev.*, **7**, 757763.

[34] Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research*, **27**, 1642-1649.

[35] Lin, J., Gerstein, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, **10**, 808-818.

[36] Lin, J., Qian, D.,Bertone, P., Das, R., Echols, N., Senes, A., Stenger, B., Gerstein, M. (2002)GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Research*, **30**, 4574-4582.

[37] Lobry, J.R., Chessel, D. (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J.Appl Genet*, **44**, 235-261.

[38] Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov,R.L., Wolf, Y.I. and Koonin, E.V. (2003) Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, **9**, 608–628 (2003).

[39] Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. *Journal of Molecular Biology*, **222**, 851-856.

[40] Moszer, I., Rocha, E.P.C., Danchin, A. (1999) Codon usage and lateral gene transfer in *Bacillus Subtilis. Current Opinion in Microbiology*, **2**, 524-528.

[41] Mushegian, A.R. and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Science USA*, *93*, 10268–10273.

[42] Nesbø, C.L., Boucher, Y., Doolittle W.F. (2001) Defining the core of non-transferable prokaryotic genes: the euryarchaeal core. *Journal of Molecular Evolution*, **53**, 340–350.

[43] Radomski, J.P., Slonimski, P.P. (2001) Genomic style of proteins: concepts, methods and analysis of ribosomal proteins from 16 microbial species. *FEMS Microbiology Reviews*, **25**, 425-435.

[44] Salama, N.R. *et al.* (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.*, **186**, 7926–7935.

[45] Sandberg, R., Bränden, C.I., Ernberg, I., Cöster, J. (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino-acids usage and G+C content. *Gene*, **311**, 35-42.

[46] Sharp, P.M. and Li, W-H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acid Research*, **15**, 1281-1295.

[47] Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiate highly and lowly expressed genes. *Nucleic Acids Research*, **14**, 8207-8211.

[48] Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both traditional selection and mutational biases. *Nucleic Acids Research*, **15**, 8023-8040.

[49] Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F. (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomices pombe, Drosophila melanogaster and Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research*, **16**, 8207-8211.

[50] Stenico, M., Loyd, A.T., Sharp, P.M. (1994) Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acid Research*, **22**, 2437-2446.

[51] Sicheritz-Pontén, T. and Andersson, Siv G.E. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, **29**, 545-552.

[52] Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad of Sci USA*, **48**, 582-592.

[53] Thanassi, J.A. *et al.* (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Res.*, **30**, 3152–3162.

[54] Varenne, S., Buc, J., Lloubès, R. and Lazdunski, C. (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology*, **180**, 549-576.

[55] Wada, K.S., Aota, R., Tsuchiya, F., Ishibashi, T., Gojobori, T. and Ikemura, T. (1990) Codon usage tabulated from GenBank genetic sequence data. *Nucleic Acids Research*, **18**(Suppl), 2367-2411.

[56] Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901-906.

[57] Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J Theor Biol*, **8**, 357-366.

An information dynamics approach for characterising pattern formation in spatiotemporal chemical systems:

applications to the Gray-Scott model and its extension to a flow reactor system

Kristian Lindgren[1]

[1] Chalmers University of Technology, Göteborg, Sweden

Abstract

Pattern formation in chemical systems is a dynamical process that has been extensively studied in the litterature since the original work by Turing. Given a certain set of reactions, the shape and spatio-temporal development of the patterns is determined by two different types of information flows. One flow is the information from the noise (fluctuations in concentrations) leading to symmetry breaking that determins the exact path in the dynamics. This flow is very small compared to the other information flow, associated with the driving force given by inflow free energy, which is the focus of this presentation.

The thermodynamic constraints on this type of self-organising system tells us that the system needs to be open (for inflow of free energy and removal of heat and waste products). We have developed an information-theoretic frame work that goes one step further in the analysis of physical constraints in chemical self-organisation. The formalism is based on a thermo dynamic information quantity (via statistical mech anics), and this makes it possible to connect an information-theoretic characterisation of a spatial pattern with the free energy driving the system. In this way, a consistent picture of the pattern formation process in terms of free energy being transformed into information in a spatial pattern and eventually destroyed by entropy production when reactions and diffusion processes tries to bring the system towards equilibrium.

In our analysis the information in the pattern is decomposed into contributions from both different positions and different length scales. The overall picture we get is an inflow of information at large length scales, due to the inflow of chemical free energy. Information then flows down in length scale (and also across space), where accumulation at certain positions is con nected with the pattern formation. Infor mation is lost from the system at the finest length scales. The whole process is summarised in a continuity equation for information.

In our current research we investigate the possibility to use this formalism to make predictions on how pattern formation may depend on the structure of the driving force, i.e., the inflow of free energy. Preliminary results indicate that the information flow is generally going in the direction described above – from larger to smaller length scales –which may be viewed as generalised “second law” of information destruction. If the characteristic length scale of the free energy inflow is reduced below the length scale of the patterns in the system, the flow will not be able to support the structures built up and neither will new structure emerge unless that happends on a smaller length scale. In the same way as ambient heat has too low energy quality to drive a physical process, a chemical free energy inflow of too low length scale characteristics may be insufficient to support pattern formation.

In the PACE project this analysis may be of importance for understanding limit ations on the formation of meso-scale structures under reaction-diffusion-convection dynamics when the reactor has a certain small length scale in its inflow, as is the case for the “fan” reactor in the Omega-machine. Recent work, involving an extension of the Gray-Scott model that allows for self-replicating pattern dynamics in a flow reactor is presented.

Membrane proteins under the computational microscope

Christophe Chipot[1]

[1] Institut nancéien de chimie moléculaire - Université Henri Poincaré - Nancy I, F-54506 Vandoeuvre-lès-Nancy, France

Abstract

Life sciences have greatly benefited from innovative techniques in structure determination, DNA sequencing at the level of entire genomes, and direct manipulation and observation of single molecules. Computational biophysics complements these efforts by means of cutting-edge molecular modeling in direct response to experimental advances. Its grand challenge is to attain the microscopic detail that cannot be easily accessed through conventional experimental techniques. Of topical interest are G protein-coupled receptors, which correspond to the third largest family of genes in the human genome, and, hence, represent privileged targets for rational drug design. When neither theory nor experiment alone can provide atomic-level three-dimensional structures of G protein-coupled receptors, their synergistic combination offers an interesting perspective to reach this goal. Such a self-consistent strategy has been applied successfully to elucidate the structure of the human receptor of cholecystokinin-1 in the presence of an agonist ligand.

We will show that the site-directed mutagenesis experiments designed to pinpoint key receptor-ligand interactions can be reproduced accurately employing the free energy perturbation methodology. We will further disclose how sufficiently long simulations can shed new light on the structural modifications undergone by the receptor upon transition from its activated state to its inactivated state. We will also demonstrate that novel computational approaches can be used fruitfully to investigate the reversible association of transmembrane helices, a key event in membrane protein folding. The paradigmatic example of glycophorin A is chosen to decipher the mechanisms of recognition and association reflected in the hypothesized two-stage model of membrane protein folding. The proposed free energy calculation illuminates the complementarity of a short- and a long-range regime in the formation of the native helix dimer, driven by forces of distinct nature. Last, the same methodology will be utilized to disentangle the intricate mechanism of glycerol conduction in the Escherichia coli facilitator GlpF.

Of particular interest, it will be demonstrated in the light of significantly long simulations, that transport across the aquaglyceroporin channel and both orientational and conformational relaxations are processes that span comparable time scales.

Oscillations in Saccharomyces cerevisiae

Michel Jacquet[1]

[1]Institut de Génétique et Microbiologie, Université Paris-Sud, Orsay, France

Abstract

The budding yeast Saccharomyces cerevisiae, which is one of the best-known cellular organisms in which practically all experimental approaches are feasible, has been chosen to study the functioning of many signaling systems. This unicellular organism, as most living organisms, is able respond to sudden changes in their environment by modifying its program of gene expression. Such responses involve several signaling devices. As an example, a given stress such as heat shock will trigger both a specific pathway leading to the activation of genes encoding chaperones and other protective components and in addition will also trigger a more general system which is activated under a wide variety of stresses as well as under metabolic transition: the so-called general stress system mediated by two transcription factors Msn2 and Msn4. The sensitivity to stress of Msn2 and Msn4 depends also from the metabolic state of the cell, which is monitored by the cyclic AMP system. We have been investigating this system as an interesting model of signaling system.

Genetic approaches, based on the fact that this system is required for normal growth, allowed us and others to identify all its components with their hierarchical relationships. Noteworthy, it was found that the production of cAMP is under a very strong feedback mechanism. cAMP positively controls the activity of the protein kinase A (PKA). Main targets for the growth control are Msn2 and Msn4, which are inactive and maintained in the cytoplasm when phosphorylated by this kinase. We took advantage of the development of rapid fluorescent video-microscopy to follow the kinetics of nucleocytoplasmic translocation of Msn2 and Msn4. We made an unexpected observation that these transcriptional activators shuttle periodically with an oscillatory behavior of the molecular population in and out the nucleus upon stress activation (1). Moreover, the illumination by the light of the microscope was sufficient to trigger the stress induced nuclear translocation. To explain this behavior we made several hypothesis based on potential systems able to generate oscillations.

A tempting model involving the modifications of Msn2, which occur during the formation of the transcriptional complex, was eliminated by experiments showing that the domains of Msn2 interacting in the complex were not required to produce the oscillatory behavior. Latter on, a small region of Msn2, controlled by PKA phosphorylation, was found to be necessary and sufficient to generate a periodic shuttling of the associated protein. This result let us with Albert Goldbeter, to investigate by means of a computational model the ability of the cAMP-PKA system to enter in an oscillatory regime. The strong feedback mechanism of the PKA on the accumulation of cAMP within the cell is able to generate sustained oscillation within a range of parameters compatible with the actual knowledge of the yeast components. The details of this model and its implications will be discussed during the presentation. To check this model, we have used mutants altered in the feedback mechanism and shown that the oscillatory behavior was lost while the stress-induced translocation was still occurring. This result confirms the need of an effective feedback mechanism of the cAMP-PKA system to produce oscillations and provide a mechanistic basis for this novel aspect of non-linear relationships in cellular signaling.

References

[1] Jacquet, M., Renault, G., Lallet, S., De Mey, J., and Goldbeter, A. (2003). Oscillatory nucleocytoplasmic shuttling of the general stress response transcriptional activators Msn2 and Msn4 in Saccharomyces cerevisiae. *The Journal of cell biology* **161, 497-505**.

Modeling Cellular Rhythms in Metabolic and Genetic Regulatory Networks

Albert Goldbeter[1]

[1] Unité de Chronobiologie théorique, Faculté des Sciences, Université libre de Bruxelles (ULB), B-1050 Brussels, Belgium

Abstract

Because of their close association with feedback processes in regulated genetic or metabolic networks, cellular rhythms represent a prototypic field of research for Systems Biology. After providing an overview of models for cellular rhythms, I will focus on two examples of rhythmic behavior associated, respectively, with enzymatic and genetic regulation. The first example pertains to metabolic oscillations in yeast glycolysis, which arise from the regulation of enzyme activity. The second pertains to circadian rhythms which originate from intertwined feedback processes in genetic regulatory networks. Computational models of increasing complexity have been proposed for the molecular mechanism of these rhythms, which occur spontaneously with a period of the order of 24 h. Models for circadian rhythms in Drosophila account for a variety of dynamical properties such as phase shifting or long-term suppression by light pulses, and entrainment by light-dark cycles. An extension of the model to the mammalian circadian clock allows us to address the dynamical bases of physiological disorders of the sleep-wake cycle in humans.

Initiation of differential gene expression in sporulating *Bacillus subtilis* - a mathematical mode

Michael D. Yudkin[1]

[1]Kellogg College, Oxford, OX2 6PM, UK.

Abstract

Early in sporulation *Bacillus subtilis* undergoes an asymmetric septation to give two compartments, a smaller prespore and a larger mother cell. Differential gene expression is established in these compartments as a result of the activity of compartment-specific transcription factors called sigma (σ) factors. The first of these is σ^F, which becomes active in the prespore (but not the mother cell) soon after asymmetric septation. The initiation of differential gene expression depends on interactions among σ^F and three specific regulatory proteins. In the past several years we and others have identified all these interactions, established the kinetic constants associated with them and measured the intracellular concentrations of the relevant proteins. We have now used these kinetic constants and concentrations to write a set of linked differential equations, which together constitute a mathematical model that successfully describes the regulation of σ^F.

1 The biological system

When *Bacillus subtilis*, a Gram-positive soil bacterium, is starved of certain nutrients, it enters a sequence of steps called sporulation, which lead to a heat-resistant spore. One of the earliest steps is the construction of a division septum in an unusual position. Whereas normal cell division in bacteria starts with the formation of a septum across the midline of the cell and yields daughter cells of equal sizes, the septation that is characteristic of sporulation is asymmetric: it gives rise to a small compartment named the prespore and a large compartment named the mother cell. The two compartments remain attached, and as sporulation proceeds the mother cell engulfs the prespore. The prespore eventually becomes the mature spore, and this is liberated from the mother cell when the latter lyses.

This series of morphological changes is dependent on the unfolding of a genetic programme which involves the expression, in a temporally ordered sequence, of sporulation-specific genes, some in the prespore and others in the mother cell. Such compartment-specific gene expression depends on the activity of certain transcription factors that are specific to sporulation. In bacteria in general, the specificity of the RNA polymerase is due to factors called sigma factors, which direct the enzyme to the promoters of particular genes. Different sigma factors can bind to the core RNA polymerase (depending on the cell's requirements at that moment), thus forming different promoter-specific RNA polymerase holoenzymes. In growing cells almost all transcription is due to holoenzyme containing the "house-keeping" sigma factor σ^A. The first sigma factor specific to sporulation is σ^F, but in growing cells this is in an inactive form. After asymmetric septation σ^F remains quiescent in the mother cell but is activated in the prespore. Active σ^F now starts to compete with σ^A, forming σ^F-containing holoenzyme, and the consequent transcription of σ^F-specific genes in the prespore sets in train all the subsequent events of sporulation. It is crucial to the success of sporulation that σ^F is regulated in the way that

I have just described — i.e. that before asymmetric septation it is kept inactive, and that after asymmetric septation it is activated in the prespore but not in the mother cell.

Since activation of the first sporulation-specific sigma factor in only one of the two compartments generated by asymmetric septation leads to differential gene expression, sporulation is widely regarded as a simple and tractable model of differentiation. The regulation of σ^F has been extensively studied in genetic, biochemical and biophysical experiments, the results of some of which will be summarised very briefly in what follows. A more extensive review, with references to the original literature, has been published by Yudkin and Clarkson (2005).

σ^F is regulated by three sporulation-specific proteins, SpoIIAB (here called AB), SpoIIAA (AA) and SpoIIE (IIE). AB is capable of three different interactions. In one interaction it binds (in the presence of ATP) to σ^F and prevents the latter from becoming attached to the core RNA polymerase. This binding of AB as an "anti-sigma factor" to σ^F prevents the transcription of σ^F -dependent genes. Alternatively, AB can interact with AA in either of two ways; the choice between these depends on the adenine nucleotide present. In the presence of ATP AB acts as a specific protein kinase, phosphorylating AA on one of its serine residues to yield AA-phosphate (AA-P). In the presence of ADP, by contrast, AB and AA interact to make a non-covalent ternary complex, AB.AA.ADP. The third of the proteins that regulate σ^F, IIE, is a specific phosphorprotein phosphatase, which hydrolyses AA-P to AA.

In growing cells σ^F, AB and AA are all present, although in low concentrations, but when sporulation begins their rate of synthesis increases dramatically. Before asymmmetric septation all the AA in the cell is phosphorylated. By contrast with AA, AA-P cannot make a non-covalent complex with AB, and the latter is free to bind to and thus inhibit σ^F. But soon after the beginning of sporulation IIE is synthesised, and as a result AA-P is hydrolysed to AA. This AA accumulates in the prespore, and it interacts with the σ^F.AB.ATP complex, liberating σ^F and initiating the genetic programme of sporulation. Thus the initiation of sporulation depends on AB's abandoning σ^F as a binding partner and binding instead to AA, and this change of partner depends in turn on the accumulation of AA (produced by hydrolysis of AA-P) in relatively high concentration in the prespore rather than the mother cell.

Results from my laboratory and elsewhere, particularly those gained by the use of fluorescence spectroscopy, have shown that the release of σ^F from its inactive complex with AB involves the following sequence of two reactions:

$$\text{AA} + \sigma^F.\text{AB.ATP} \rightleftharpoons \sigma^F + \text{AA.AB.ATP};$$
$$\text{AA.AB.ATP} \rightleftharpoons \text{AB.ADP} + \text{AA-P}.$$

Since the concentration of ATP in the cell is much higher than that of ADP, the AB.ADP thus formed will exchange its ADP for ATP, generating AB.ATP which would be expected to bind again to σ^F. From the relative rates of reaction and the intracellular concentrations of the proteins it can be shown that, as a result, σ^F will soon be inhibited once more — unless some means is found to maintain AA at a high concentration so that it is constantly available to disrupt the σ^F.AB.ATP complex. The only plausible means of maintaining the concentration of AA is via hydrolysis of the AA-P formed in the reactions shown above by IIE, a mechanism that implies that AA is continually being recycled

through its phosphorylated and non-phosphorylated forms. If this implication is correct, the cell must be engaging in a "futile cycle", which is potentially costly in ATP. How is the cost minimised?

Enzymological studies of the reaction in which the protein kinase AB catalyses the phosphorylation of AA have shown that it consists of two phases: first a moderately slow pre-steady state, and then a very slow steady state after two moles of AA have been phosphorylated per mole of AB. (AB is a homodimer). The slow step in the steady state has been identified as the loss of ADP from AB after each round of phosphorylation, and this finding has led to the discovery that the interaction of AA with AB induces a conformational change in the latter.

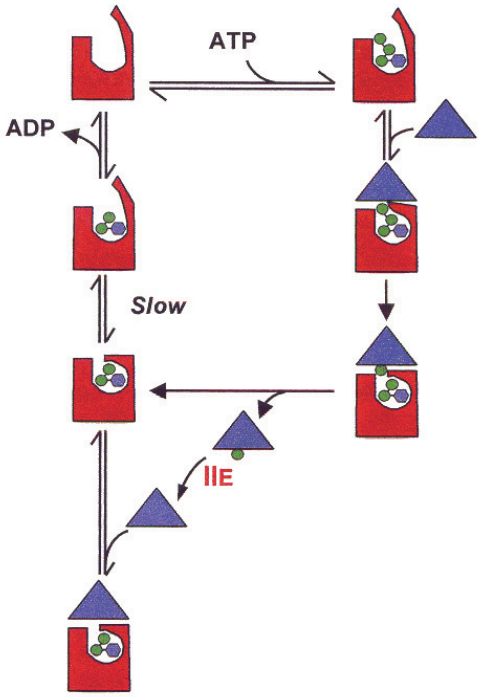

Figure 1: Cartoon diagram of the reaction scheme of the phosphorylation of AA by AB. AB is shown as a monomer with a flexible lid which is closed by the interaction with AA. Blue shape, AA; red shape, AB; purple hexagon, adenosine moiety of ADP and ATP; green circles, phosphate groups.

It is known from crystallographic studies that AB has a flexible fold (an "ATP-lid") covering a nucleotide-binding pocket; when this lid is "open" ATP or ADP can readily dissociate from (or bind to) the protein, but when the lid is covered by AA the nucleotide is trapped and its dissociation from the protein is very slow. Figure 1 shows, in cartoon form, the repeated phosphorylation of AA by AB, the regeneration of AA by hydrolysis, and the interaction of the resulting AA with AB.ADP to form the ternary complex AB.AA.ADP (which has been shown to accumulate during the reaction).

Although much information has been discovered about the regulation of σ^F, several questions remain. Two particular quantitative puzzles are these. First, how is the activation of σ^F confined to the prespore? It is known that IIE, whose activity is essential for the production of the AA that liberates σ^F from its complex with AB, is confined to

the asymmetric septum that separates the prespore from the mother cell. But the activity of IIE is displayed on both sides of the septum, with the result that AA is formed in the mother cell as well as in the prespore. A possible explanation of the fact that σ^F activity appears only in the prespore is the difference in volume of the two cells: since the volume of the prespore is at least four times less than that of the mother cell, the concentration of AA will be four times higher in the former than in the latter.

However, careful quantitative analysis would be necessary to show that this difference is sufficient to account for the fact that σ^F activity is essentially absent from the mother cell. Secondly, for RNA polymerase containing σ^F to become active in the prespore, the σ^F must compete with the sigma factor, σ^A, that is a component of RNA polymerase holoenzyme during the growth phase of the bacteria. However, the affinity of σ^A for the core RNA polymerase is 25-fold higher than that of σ^F, while the concentration of σ^F at the relevant time is only twice that of σ^A, so how can σ^F compete effectively with σ^A for binding to the core polymerase?

My colleagues and I have now constructed a mathematical model which solves these two puzzles and answers many other questions that have arisen during years of study of the regulation of σ^F.

2 Constructing and testing the mathematical model

In its original form the model consisted of a set of ordinary differential equations describing the interactions that had already been identified between the molecular species involved in regu-lating σ^F. To construct the model we used the known concentrations of the proteins and of ADP and ATP at the beginning of sporulation, the rates of synthesis of the proteins, and the rate constants for the molecular interactions. Most of these parameters are published; a few were determined in our lab specifically for the purpose of producing the model.

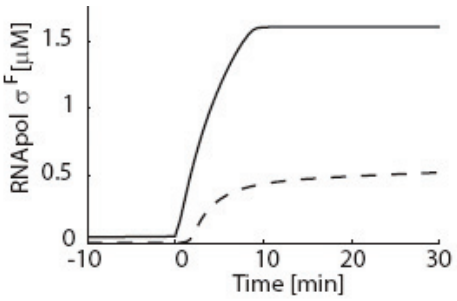

Figure 2: Predicted concentration of σ^F-RNA polymerase during sporulation if SpoAB is allosteric and SpoAA binds with positive cooperativity (continuous line), or if SpoAB is not allosteric (dashed line).

We then asked the model to predict the change in concentration of σ^F-holoenzyme around the time of asymmetric septation. But the prediction from this model (dashed line in Figure 2) shows that the holoenzyme never reaches a concentration sufficient to account for the transcription of σ^F-dependent genes in the prespore (at least 1 μmolar σ^F-holoenzyme). Clearly there is some feature that we have failed to take into account in constructing the model.

We wondered whether the missing factor was allostery in AB. Since AB is a dimer, it is possible that the binding of AA to AB is not linear with concentration but is subject to positive cooperativity. We therefore extended the model to include allosteric interactions in AB; the interactions in the revised model are shown in Figure 4.

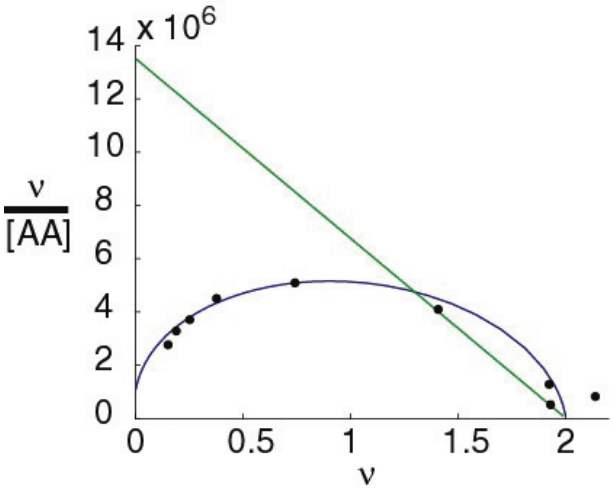

Figure 3: Scatchard plot of experimentally determined (filled symbols) or predicted (continuous lines) binding of SpoIIAA to SpoIIAB, if SpoIIAB is allosteric and SpoIIAA binds with positive cooperativity (blue line) or if SpoIIAB is non-allosteric (green line). ν is the fraction of SpoIIAB that has bound SpoIIAA.

But is AB in fact an allosteric protein, with the binding of AA to AB characterised by positive cooperativity? We studied the binding by suface plasmon resonance and generated a Scatchard plot, which showed that the answer to this question is Yes (Figure 3). It was thus reasonable to use the revised model to predict the change in concentration of σ^F-holoenzyme at the beginning of sporulation. The prediction was now in accordance with experimental findings (continuous line in Figure 2).

We next sought to see whether the revised model would simulate the results obtained from in vitro experiments with purified AA, AB, IIE and σ^F that had previously been published by us and others. We found that the model was successful in simulating: the binding of AA to AB.ADP; the binding of σ^F to AB.ATP; the disruption of σ^F.AB.ATP complexes by AA; the re-binding of σ^F to AB.ATP as the disrupting AA was inactivated by phosphorylation; and the response of this re-binding to the presence of IIE. In addition the model successfully simulated the distinctive biphasic time course of phosphorylation of AA by AB.

3 Application of the mathematical model to the sporulating cell

The success of these simulations of results obtained *in vitro* emboldened us to see whether the model could answer the two questions about the sporulating cell that we posed above. To study the first question — whether the difference in volume between the mother cell and the prespore accounts for the fact that the activation of σ^F is confined to the prespore — we modelled the system with the cellular concentrations of proteins at the time of asymmetric septation, and then imposed a fourfold increase in the concentration of

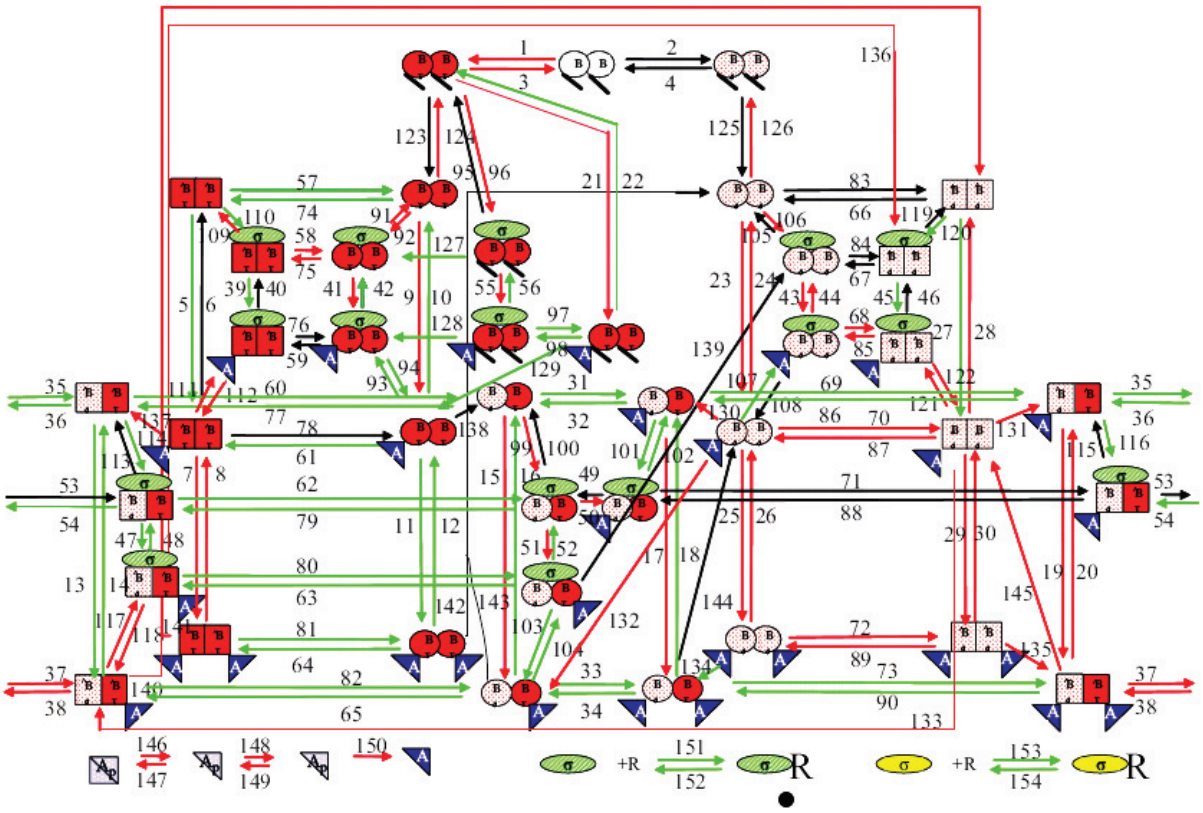

Figure 4: Reaction scheme for the regulation of σ^F release. B denotes SpoIIAB (red squares and circles), A SpoIIAA (blue triangles), and R the RNA polymerase core enzyme; σ^F is represented by a green oval, σ^A by a yellow oval. The SpoIIAB conformation that binds SpoIIAA with low affinity is depicted as circles and the high-affinity state as squares. ATP-bound SpoIIAB is depicted as a filled shape, ADP-bound SpoIIAB as a stippled shape, and nucleotide-free SpoIIAB as a white shape. Green arrows represent reactions to which the model is insensitive, red arrows those to which the model is highly sensitive, black arrows those that are considered neither sensitive nor insensitive (see Iber et al., 2006).

IIE. (This increase mimics the abrupt change in concentration of IIE brought about by concentrating a fixed number of molecules of IIE into a fourfold smaller volume).

The results (Figure 5) show that almost all of the σ^F is released — but only if the increase in IIE concentration is accompanied by an equimolar increase in the concentration of its substrate, AA-P. (An increase in the concentration of AA-P alone is not sufficient — results not shown). A calculation from the known rate of IIE activity, the known affinity between IIE and AA-P, and the known cellular concentrations of the two proteins, shows that IIE and AA-P will indeed accumulate together at the site of the asymmetric septum as the latter is being formed. We have thus solved one of the most important puzzles that lie at the heart of this simple differentiation system: it is now clear that compartment-specific activation of σ^F is dependent solely on the increase in concentration of IIE and AA-P generated in the prespore by the asymmetric placement of the division septum.

We next turned to the question of competition between σ^F and σ^A for binding to the core RNA polymerase. We included in our model the known concentrations of the two sigma factors and their affinities for the core RNA polymerase, and got the model to plot

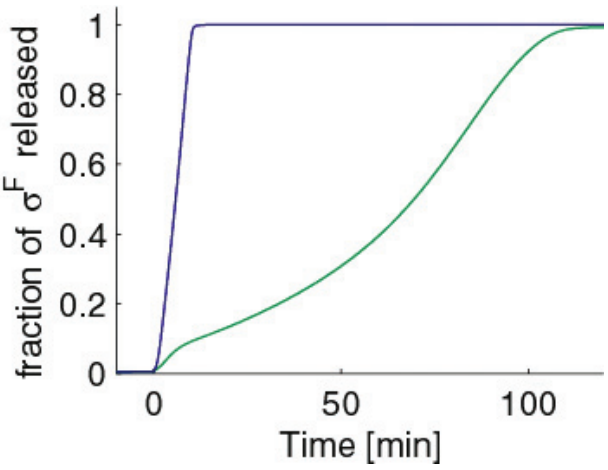

Figure 5: Predicted release of σ^F in response to a fourfold increase in the concentration of SpoIIE alone (green line) or of SpoIIE and SpoIIAA-P (blue line) imposed at zero time.

the predicted concentrations of the two holoenzymes against changes in the concentration of IIE. The prediction shows that, provided that the concentration of IIE increases by at least threefold over that found in the pre-septational cell, the concentration of σ^F-holoenzyme will reach a level sufficient to allow transcription of σ^F-specific genes (black curve in Figure 6). But even when σ^F-holoenzyme reaches its maximum concentration, σ^A-holoenzyme will still be active (red curve in Figure 6). This prediction exactly mimics the situation in the sporulating cell, where the two holoenzymes are active simultaneously in the prespore.

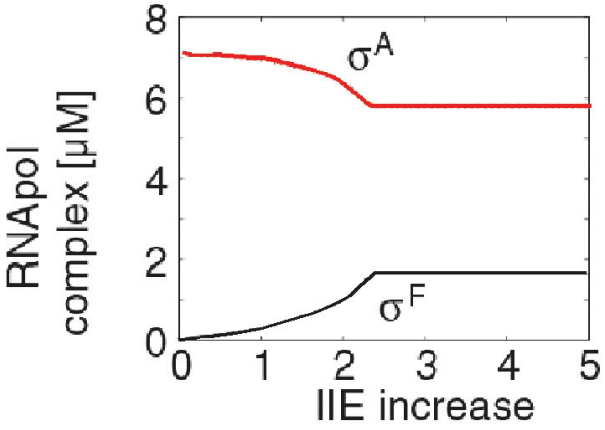

Figure 6: Predicted concentrations of holoenzymes containing σ^F or σ^A 90 minutes after asymmetric septation, as a function of increases in the concentration of SpoIIE including SpoIIE complexed with SpoIIAA-P.

4 Conclusion

The mathematical model summarised above is described in detail in a paper [1] by Iber et al. (2006). That paper includes many more results than I have had space to describe here; for example it accurately simulates the behaviour of a substantial number of sporulation mutants whose phenotype had not previously been readily explicable. The success of the

model in predicting all these phenomena, and in particular its success in accounting for the initiation of differential gene expression in terms of the volume difference between the two compartments of the sporulating cell, would not have been possible without the abundant information that has come both from studies of the interactions that regulate σ^F and from measurements of the intracellular concentrations of the relevant proteins (for a review of all this work see [2] Yudkin and Clarkson, 2005). It is remarkable that so fundamental a biological phenomenon as differential gene expression can be achieved with the help of only four proteins.

Acknowledgments

I am deeply grateful to my colleagues Iain Campbell, Joanna Clarkson and Dagmar Iber for their tireless work in devising and testing the model, as well as to the many students and postdoctoral research associates who have contributed to the study of this system in my laboratory over the years.

References

[1] Iber, D., Clarkson, J., Yudkin, M.D. and Campbell, I.D. (2006) The mechanism of cell differentiation in *Bacillus subtilis*. *Nature* **441**: 371-374.

[2] Yudkin, M.D. and Clarkson, J. (2005) Differential gene expression in genetically identical sister cells: the initiation of sporulation in *Bacillus subtilis*. *Mol. Microbiol.* **56**: 578-589.

Detection of emergent phenomena in multi-agent systems

Thomas Moncion[1], Guillaume Hutzler[1], Patrick Amar[2]

[1] Laboratoire IBISC FRE CNRS 2873, Université d'Évry Val d'Essonne, F-91000, France
[2] Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, F-91405, France

Abstract

Biological systems are composed of a large number of molecules. These systems are complex because numerous and diversified interactions exist between the molecules making their understanding difficult.

Due to these interactions, complex systems can show some phenomena considered as emergents because unpredictable from the simple description of molecules' behaviour. We particularly study self-organisation phenomena corresponding to a structural organisation without central control.

Many tools are available to model these biological complex systems. We use multi-agent systems which model each element of the system as an agent. We study their collective behaviour inside their environment. Multi-agent simulations exhibit phenomena which appear in the real systems they model. The characterisation of these phenomena is usually given by the users observing such systems, which introduces an important subjective bias.

The objective of our work is to automatically detect some emergent phenomena. We model multi-agent simulation as a graph where vertices represent agents and edges their interactions. This graph is modified at each time step and we study its properties.

Introduction

The high capabilities of computers enable the modelling of wide biological complex systems. In real systems or in the corresponding simulations, we detect some unpredictable phenomena called emergent phenomena, which are not easily explainable. Multi-agent systems, by their approach, are good tools for modelling complex systems, because they allow the simulation of such emergent phenomena.

Our objective is to detect in multi-agent simulations a kind of emergent phenomenon, namely self-organisation, which is described as the organisation of structures whithout central control. We have modelled agents and their interactions as a scalable graph to study some properties of the system.

1 Self-organisation in biological complex systems

1.1 Complex systems in biology

A complex system can be defined as a system composed of heterogeneous elements with multiple connections, which makes it difficult to apprehend by the mind.

Biological systems are complex systems composed of large number of molecules (DNA, RNA, proteins, ...). A metabolism is described by interactions between proteins whereas regulations processes use proteins/DNA interactions. Elements of these systems and their

interactions produce organised structures (organits, cells and organisms). These diverse structures have their particular functionality because of the diverse interaction networks in action.

To understand these biological systems, we have tried to isolate different elements in order to study them in depth. This is for example from this reductionist concept that molecular biology uprised, which have supplied a lot of answers to understand these systems. However, some functionalities of these systems are difficult to understand because they arise not only from the elements of the system themselves, but also from the interactions between them.

The complexity of these systems is due mostly on the one hand to the locality of the interactions between the elements of the system (the elements do not interact with all of the other elements) and on the other hand to the variability of the interactions along time (the duration of interactions is not the same for all the elements)

1.2 Self-organisation

The complexity of these systems can lead up to the formation of structural organisation whithout central control: we call this *self-organisation*. Most of the biological systems (ants, cells ...) exhibit this kind of phenomenon.

A set of mecanisms [8] enable the occurence of such phenomena:

- direct interactions based on information broadcast and localisation
- indirect interactions based on stigmergy (information sharing through the environment)
- reinforcement of agents behaviour based on a system of rewards and punishments
- cooperation behaviour of individual agents based on the composition and decomposition of agents

Many organisation scale levels can be found in these self-organisation processes. These differents scales give to the system a different behaviour and a super functionality any elements cannot bring taken separately.

2 Agent-based modelling for biological complex systems

Man has always tried to model his environment. These models have provided explanations about these systems but have also opened new directions of research, such as the validation of results in real systems, and have inspired new computing methods (for example, ant colony optimization in graphs [5]).

Many mathematical and computing tools can be used to model and to study complex systems, ranging from differential equations to cellular automata and multi-agent systems [4]. These tools can give different kind of answers to the same biological question, by means of different modellings.

Multi-agent systems are based on the modelling of each element of the system as an agent [3]. This modelling being also based on the interactions between agents, we can reproduce in the simulations some of the emergent phenomena observed in the real system, which is one of the main interest of this kind of approach.

3 Detection of self-organisation

3.1 Goal

In this context, our goal is to automatically detect self-organising phenomena that may arise in simulations , which is a preliminary before attempting to explain or predict them. The detection of these phenomena begins with the detection of the different levels of organisation which can be formed in these systems and also to study the functionality of the system as a whole. This structural organisation can be distinguished by differents points

- the organisation itself can remain stable for a more or less long period of time
- its functionality can be different from the one of the global system

We have designed methods allowing us to study global properties of a system and to detect differents clusters of agents which can be formed during simulations in order to study their evolution.

3.2 Multi-agent simulation as a graph

A multi-agent system can be seen as a graph where the vertices represent the agents and the edges their interactions. There are different kinds of interactions between two agents; We consider that two agents interact if they are close enough to each other and also if they share common attributes and/or functions. To implement these multiple kinds of interactions the edges of the graph are multi-labeled, each label representing an interaction criterion (see figure 1).

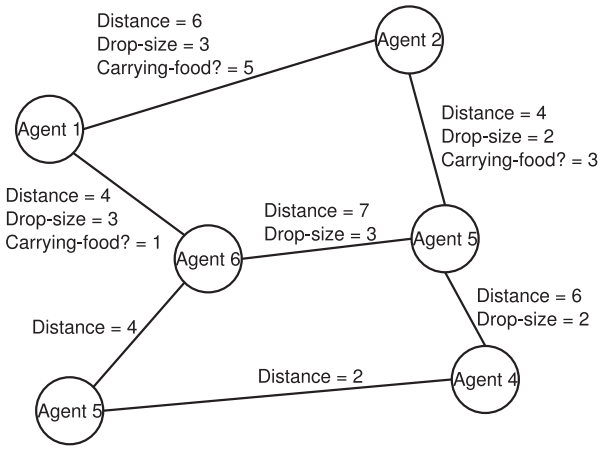

Figure 1: Model of an ant colony simulation. There are three interaction criteria: the distance between two ants, the quantity of pheromones dropped in the environment (*drop-size*) and the "return-to-nest" behaviour (*carrying-food?*). The label weights represent the intensity of each interactions

3.3 Construction and analysis of the graph

We consider multi-agent systems using a discrete time scale (constant time steps). The graph evolves at each time step, according to the current values of the attributes of the

agents. For each label, we use a weight representing the interaction intensity. This weight is increased or decreased according to the presence or the absence of a connection at the current time step between the two agents. A connection is set depending on which attribute is interesting; For example this can be based on:

- the comparaison of the value of the attribute with respect to the average value of the same attribute of all the agents
- the value of the attribute compared to a fixed interval of values
- the comparaison of the values of the attribute between the agents themselves.

The analysis of these systems is based on the graph properties. We study the global properties of the graph such as the research of topological particularities like those we can find in complex networks [6, 1]. We also study the global evolution of some indicators like the sum of the weights of the labels (see figure 2) or the degree of the edges. We also try to determine the various clusters that can appear according to each criterion using graph clustering techniques (see figure 3).

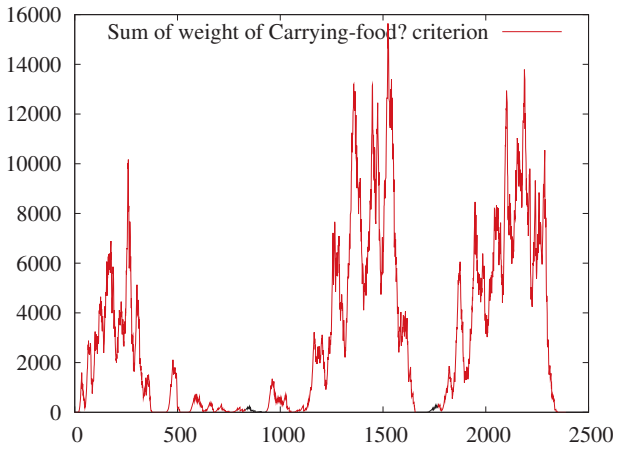

Figure 2: Sum of the weights of the *carrying-food?* criterion

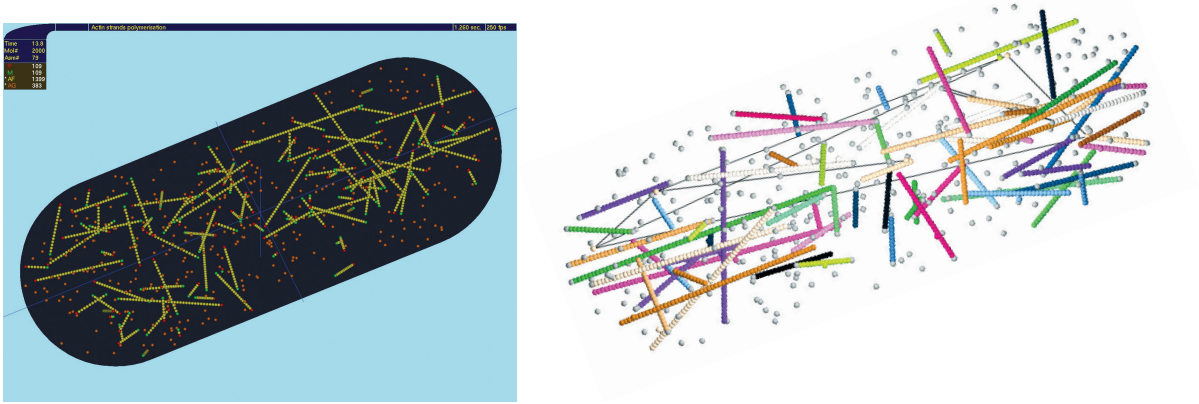

Figure 3: Detection of clusters in a simulation of actin filament polymerisation produced by Hsim[2].This detection is based on the distance criterion beetween agents.

Conclusion

The understanding of emergent phenomena and especially the processes leading to self-organisation, is important to apprehend complex systems. Self-organisation is also studied with the aim of being used in new computing applications based on these processes[7].

We have developed a tool that allows us to capture various interactions which can exist between the agents of a system. We think it is the key to the understanding of the processes leading to the creation of structures at different scales and to a superior functionality of the system.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews Of Modern Physics*, 74:47–97, 2002.

[2] Patrick Amar, Gilles Bernot, and Victor Norris. Hsim: a simulation programme to study large assemblies of proteins. *Journal of Biological Physics and Chemistry*, 4:79–84, 2004.

[3] Jochen Fromm. *The emergence of complexity*. 2004.

[4] Guillaume Hutzler. *Cellular Automata and Agent-Based approaches for the modelling and simulation of biological systems: application to the lambda phage*, pages 137–162. Platypus Press, 2006.

[5] Vittorio Maniezzo, Luca Maria Gambardella, and Fabio de Luigi. *Ant Colony Optimization in New Optimization Techniques in Engineering*, chapter 5, pages 101 – 117. Springer-Verlag Berlin and Heildelberg, 2004.

[6] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45:167 – 256, 2003.

[7] Manish Parashar and Salim Hariri. Autonomic computing: An overview. *Lecture notes in computer sciences*, pages 244 – 259, 2005.

[8] Giovanna Di Marzo Serugendo, Marie-Pierre Gleizes, and Anthony Karageorgos. Self-organisation and emergence in mas: An overview. *Informatica*, 30:45–54, 2006.

Inferring parameters of genetic regulatory networks with symbolic formal methods

Daniel Mateus[1], Jean-Paul Comet[2], Jean-Pierre Gallois[1] and Pascale Le Gall[2]

[1] CEA, LIST, Saclay, F-91191 Gif sur Yvette Cedex, France
[2] IBISC, FRE CNRS 2873, Université d'Évry Val d'Essonne & Epigenomics Project, Genopole®, 523 terrasses de l'Agora, 91000 Évry, France

Abstract

Understanding the functioning of genetic regulatory networks supposes a modeling of biological processes in order to simulate behaviors and to reason on the model. Unfortunately, the modeling task is confronted to incomplete knowledge about the system. To deal with this problem we propose a methodology that uses the qualitative approach developed by R. Thomas. A symbolic transition system can represent the set of all possible models in a concise and symbolic way. We introduce a new method based on model-checking techniques and symbolic execution to extract constraints on parameters leading to dynamics coherent with known behaviors. Our method allows us to efficiently respond to two kinds of questions: is there any model coherent with a certain hypothetic behavior? Are there behaviors common to all selected models? The first question is illustrated with the example of the mucus production in *Pseudomonas aeruginosa* while the second one is illustrated with the example of immunity control in bacteriophage lambda.

Keywords: Gene networks; qualitative dynamical models; symbolic execution; temporal properties; model-checking.

1 Introduction

Genetic regulatory networks are constituted of various interacting components, mainly genes and proteins, usually forming a complex network of interleaved feedback loops. As it is impossible to use intuitive reasoning to really understand these networks and predict their possible behaviors, modeling and simulation become necessary [1]. The lack of reliable quantitative data available about a given system is a typical difficulty of the modeling approach. To overpass this problem, qualitative models have been developed, whose goal consists in abstracting details of the system although preserving qualitative observations.

Boolean models of genetic regulatory networks [2] are one of such formalisms. In these models, the constituents of the network are represented by variables that can only take two values, 0 or 1, meaning that the associated component is absent or present (or that the associated gene is inactive or active). R. Thomas proposed an *asynchronous* boolean modeling [3]: his approach takes into account the fact that the delays of synthesis or degradation are different from one protein to another, whereas it is not the case in previous boolean models [4]. The relation between boolean models and piecewise-linear differential equations have been first discussed in [5]. R. Thomas' approach has been generalized to a multilevel discrete modeling [4, 6, 7]; in this generalized formalism the concentrations of the constituents of the network are represented by integer variables which can take a finite number of values. Such a discrete model can be seen as a precise

qualitative abstraction of a system of piecewise-linear differential equations, as demonstrated by E.H. Snoussi [8]. This formalism is described in Sec. 2, where the convenience of introducing more than two levels of expression for the variables is explained.

This generalized discrete approach has been used to model various gene networks (for example in [9, 10, 11, 12, 13]). H. de Jong et al. [14] have recently proposed a refinement of R. Thomas' discrete modeling that takes into account singular states (corresponding to frontiers between qualitative states).

Nevertheless, even in such a discrete and finite formalism there are usually more than one model compatible with the knowledge on the system. Knowledge generally consists, on the one hand, in inhibitions or activations between genes and other constituents of the network, and on the other hand, in behaviors, observed in experiments. Inhibitions or activations allow one to constrain the possible values of the parameters of the model, on which the evolution depends. It is more difficult to select the parameters corresponding to observed behaviors. For example, the properties relating homeostasis (stable cyclic behavior) or multi-stationarity to the steadiness of characteristic states of feedback circuits [15, 16] can be used to decrease the number of parameter values to be considered, as in the GINsim tool [17].

To go further, two main ideas have been proposed. The first one consists in using constraint logic programming, to manipulate partially known models [18]. As this approach does not allow one to describe all observed behaviors, the difficulty of selecting parameters according to observations remains. The other one consists in formally specifying temporal properties and in verifying if the constructed model satisfies the specification. For example Shaub et al. [19] proposed a method for determining all infinitely visited states for which the observed behaviors have to be verified. More generally the specification can be expressed in a formal temporal language (like computational tree logic – CTL) and verification of behavior specification is then studied for each possible complete model (*i.e.* where each parameter has a precise value) independently. Implementing this idea, the tool SMBioNet [20] selects the models with respect to a given specified behavior after having exhaustively generated all possible models. In the tool GNA [21], CTL is also used to specify behaviors but only one complete model can be simulated.

Description of the proposed method

In this chapter we propose a method combining the advantages of both approaches described above. The set of possible models can be represented by a unique formal model, a symbolic transition system (STS) [22]. Symbolic execution techniques allow the simulation of the STS, generating all possible behaviors. We specify behaviors using linear temporal logic (LTL) [23], and we select parameters with respect to LTL formulas by building constraints: parameters satisfying these constraints define the set of all models verifying the specified behavior.

Thus we propose a methodology to analyze partially known systems. On the one hand, an interaction graph of the system leads to a STS, representing the set of discrete models compatible with the interactions; on the other hand, the known behaviors of the system are translated into LTL formulas. Constraints associated with these formulas restrict the possible values of the parameters; then these constraints are added to the initial STS, which represents the set of discrete models with the specified behavior.

We will see in the sequel two different types of questions that can be asked after this construction:

- is there any model coherent with a certain hypothetic behavior? The hypothetic behavior is translated into LTL formulas, and the method finds the possible parameters coherent with this hypothesis or shows that this behavior is impossible over the set of selected models. This case is illustrated on the example of mucus production in *Pseudomonas aeruginosa*.
- Are there behaviors common to all selected models? We will see that the symbolic representation of possible parameters allows to exhibit common behaviors of the selected models, without having to enumerate the models. The example of immunity control in bacteriophage lambda illustrates this point.

After having described the R. Thomas' discrete modeling, we introduce, in Sec. 2, constraints deduced from gene interactions, and show their use in the system associated to mucus production in *Pseudomonas aeruginosa*. This system will be used as a running example to illustrate our method. Section 3 is divided in three parts. We firstly explain the translation of a set of models into a STS model. We secondly introduce symbolic execution techniques. We thirdly explain how behaviors can be specified with LTL formulas, and the way we extend usual model-checking techniques to characterize parameters coherent with the LTL formulas. Then we show how this framework can be fruitfully applied to discover the unknowns (parameters or behaviors) of the genetic regulatory network. Section 4 illustrates the whole methodology on the example of immunity control in bacteriophage lambda.

This chapter is a synthesis of recent works [24, 25, 26, 27] and is based on results presented in [28] that have been enriched and completed. From a practical point of view this proposed methodology has been implemented in the Agatha tool, which is also used for validation purposes of industrial specifications [29, 30].

2 Discrete modeling of genetic regulatory networks

In this section we first present the notion of discrete descriptions, also called complete or basic models in the sequel. They correspond to the generalized discrete models introduced by R. Thomas [4]. These models are based on the interaction graph of the system: interaction graphs are directed graphs whose nodes abstract genes and associated proteins (called variables in the sequel) and whose edges are labeled by signs and thresholds of interactions. The threshold of the interaction $a \xrightarrow{\theta,+} b$ (resp. $a \xrightarrow{\theta,-} b$) defines when the interaction takes place: variable a activates (resp. inhibits) variable b if its concentration level is above θ. The effect of a on b does not depend on the concentration of a as soon as the concentration of a is above θ.

This remark leads to the discretization of the concentration space of the system variables: if a has k outgoing edges labeled by different thresholds $\theta_1 < ... < \theta_k$, then the concentration space of a is discretized into $k + 1$ levels denoted by integers from 0 to k. Then the level i abstracts the concentrations which are above θ_i (if $i > 0$) and below θ_{i+1} (if $i < k$). Thus the real values of thresholds θ_i do not matter for the discrete dynamics. They are then modeled by integers which reflect their relative ranks.

This possibility of having different thresholds makes generalized discrete models more expressive than simpler boolean models: if a variable a has an effect on two other

variables b and c, the threshold of the two interactions are generally not equal; so the possible levels of a are 0, where no interaction is effective, 1 where only one interaction is effective, and 2 where the two interactions are effective. Boolean models can not distinguish different thresholds, as the level of a would be 0 (no effective interaction) or 1 (all interactions are effective).

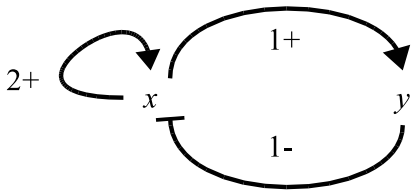

Figure 1: Example of interaction graph. Each arrow indicates an interaction from a regulator to a regulated variable; the sign indicates a positive or negative effect, and the integer is the rank of the threshold of the interaction. The blunt arrow indicates the negative interaction.

Example 1 *In Fig. 1, x and y represent two proteins produced by two genes. Variable x has two outgoing edges, with two different thresholds; the possible values for x are 0, 1 and 2; the threshold of the interaction on y is less than the threshold of the interaction on x itself, so the integer associated with the threshold of the interaction of x on y is 1, whereas the integer associated to the threshold of the interaction of x on itself is 2. The possible values for y are only 0 or 1. The genetic regulatory network corresponding to this interaction graph is described in Sec. 2.3.*

In Sec. 2.1 we present the possible discrete dynamics governed by parameters associated to discrete states. We then show how biological knowledge, in particular the gene interaction graph, can be used to construct a set of acceptable discrete descriptions.

2.1 Dynamics of a discrete description

In a discrete model, the genetic regulatory network is described by n variables, each representing the concentration of a constituent of the actual network. Each variable x_i can take an integer value between 0 and a maximum value max_i (this maximum value is deduced from the interaction graph of the system as explained above). A *state* $E = (E_1, \ldots, E_n)$ is a vector of values of the variables. With each state E, and each variable x_i, is associated a parameter $K(x_i, E)$, which has an integer value between 0 and max_i (the same maximum value than x_i). This parameter is the value toward which the associated variable tends in the associated state. It means that in the state E:

- If $K(x_i, E) > E_i$, then $(E_1, \ldots, E_i + 1, \ldots, E_n)$ is a successor of E;
- If $K(x_i, E) < E_i$, then $(E_1, \ldots, E_i - 1, \ldots, E_n)$ is a successor of E;
- If $K(x_i, E) = E_i$ for all i, then E is called a steady state, and has only itself as successor.

The associated *transition graph* is constituted of the states, and the transitions between each state and its successors. This complete model, for which each parameter has been instantiated, is called in the sequel a *discrete description*.

Let us remark that a successor of a state E differs from E in at most one coordinate: only one value from E_1 to E_n is modified (by adding or subtracting 1), if E is not a steady state. This property is called *asynchronous* updating of the variables. The reason is that when the concentration of two (or more) constituents of the network increase or decrease, there is no reason that these concentrations reach their threshold at the same time. So one of the concentration reach the threshold first; then the state of the system becomes different, with different interactions leading to different behaviors (*i.e.* the associated parameter can be different). Without knowledge on these delays, there can be more than one successor to a given state. See [4] for details about this point, and also the notion of desynchronization formally defined in [20].

Example 2 *We consider the system corresponding to Fig. 1. A state* $(E_1, E_2) \in \{0, 1, 2\} \times \{0, 1\}$ *is defined by the values of variables x and y. If* $K(x, (0, 0)) = 1$ *and* $K(y, (0, 0)) = 1$ *then the state* $(0, 0)$ *has two successors,* $(1, 0)$ *and* $(0, 1)$*. It means that if in the system the concentrations of the two proteins are at the lowest level, the concentrations increase to reach a state corresponding to* $(1, 0)$ *or* $(0, 1)$.

Until now the sign of the interactions between the constituents of the network is not taken into account. As shown in Sec. 2.2, equalities and inequalities between parameters can be deduced when positive or negative interactions between genes are known.

2.2 Constraints on parameters deduced from interactions

We have seen that each edge of the interaction graph is associated with a threshold. If a protein a activates a gene producing a protein b, the rate of synthesis of b is a sigmoid function of the concentration of a: it means that when the concentration of a is under a threshold θ, the rate of synthesis of b is not affected; but if the concentration of a is greater than the threshold θ, the rate of synthesis gets rapidly a maximal value. In piecewise-linear differential descriptions and associated qualitative models, these sigmoid functions are approximated by step functions [4, 5]. So, if in a discrete description a variable a has more than two discrete levels, and has an effect on b at the level 1, a has no effect on b when the level of a is 0, and a has the same effect on b when the level of a is 1, 2 or more.

More generally, the following equalities can be deduced from the interaction graph: we suppose that a variable x_i has one interaction on a variable x_j, and that the associated threshold has an integer level (or rank) t. Let $E = (E_1, \ldots, E_n)$ and $E' = (E'_1, \ldots, E'_n)$ be two states such that $E_i < t$, $E'_i < t$ and for every $k \neq i$, $E_k = E'_k$. E' differs from E at most in its i^{th} coordinate. Then $K(x_j, E) = K(x_j, E')$. Similarly, if $E_i \geq t$ and $E'_i \geq t$ then $K(x_j, E) = K(x_j, E')$.

These equalities allow the introduction of a new notation of the parameters: let Y be the subset of the variables $\{x_1, \ldots, x_n\}$ whose elements can have an action on x_j, and X a subset of Y; then if E is a state where the value of each variable in X is greater than or equal to the threshold of its interaction on x_j, and values of variables in $Y \backslash X$ are less than their thresholds, then the value of $K(x_j, E)$ is denoted by $K(x_j, X)$.

Example 3
In discrete descriptions associated to Fig. 1, $K(x, \emptyset) = K(x, (0, 0)) = K(x, (1, 0))$ *(the value of x, 0 or 1, is under the threshold of the interaction on itself, which is 2, and the value of y, 0, is under the threshold of the interaction on x, which is 1). Similarly* $K(x, \{y\}) = K(x, (0, 1)) = K(x, (1, 1))$ *(here the value of y, 1, is equal to the threshold).*

Moreover the sign of the interactions imply constraints on the parameters. We suppose again that a variable x_i has one interaction on a variable x_j, and X denotes a set of variables such that $x_i \notin X$. Then we have:

- $K(x_j, X) \leq K(x_j, X \cup \{x_i\})$ if x_i has a positive interaction on x_j;
- $K(x_j, X) \geq K(x_j, X \cup \{x_i\})$ if x_i has a negative interaction on x_j.

Let us point out that the inequalities are not strict: for example we can say that $K(x_j, X) \leq K(x_j, X \cup \{x_i\})$ rather than $K(x_j, X) < K(x_j, X \cup \{x_i\})$. The reason is that even if there is a positive or negative interaction, it is not sure that the interaction is sufficient to make the regulated variable reach a greater or lower threshold.

Example 4 *In discrete descriptions associated to Fig. 1,* $K(x, \{y\}) \leq K(x, \emptyset) \leq K(x, \{x\})$ *and* $K(x, \{y\}) \leq K(x, \{x, y\}) \leq K(x, \{x\})$ *(because* y *has a negative interaction on* x*, and* x *has a positive interaction on itself), and similarly* $K(y, \emptyset) \leq K(y, \{x\})$ *(*x *has a positive interaction on* y*).*

Sometimes more precise knowledge about the interactions is available. For example the presence of two different products x *and* y can be necessary to activate a gene z, or x can activate z but the simultaneous presence of x and y produces an inhibition. These two facts are respectively translated into constraints: $K(z, \{x\}) = K(z, \{y\}) = K(z, \emptyset)$ and $K(z, \{x, y\}) \geq K(z, \emptyset)$ in the first case, or $K(z, \{x, y\}) \leq K(z, \emptyset) \leq K(z, \{x\})$ in the second case.

2.3 *Mucus production in* Pseudomonas aeruginosa

Pseudomonas aeruginosa are bacteria that secrete mucus (alginate) in lungs affected by cystic fibrosis, but not in common environment. As this mucus increases respiratory deficiency, this phenomenon is a major cause of mortality. Details of the regulatory network associated with the mucus production are described by Govan and Deretic [31]. The simplified regulatory network, as proposed by Guespin and Kaufman [32], contains the protein AlgU (product of algU gene) and an inhibitor complex anti-sigma (product of muc genes). AlgU has a positive effect on anti-sigma and on itself, while anti-sigma has a negative effect on AlgU. A sufficient concentration of AlgU leads to the production of mucus (by activating different alg genes). If we consider that the threshold of the interaction of AlgU on anti-sigma is under the threshold of auto-activation of AlgU, then Fig. 1 is the interaction graph corresponding to the discrete descriptions where x and y represent respectively AlgU and anti-sigma. We consider that the production of mucus occurs precisely when the value of x is 2.

Constraints on parameters are described in the examples of Sec. 2.2. Moreover we assume that $K(x, \{y\}) = 0$ and $K(y, \emptyset) = 0$. This additional constraints mean that x tend toward its basal level (*i.e.* 0) without auto-activation and under inhibition of y and, similarly, that y tend toward its basal level when x does not activate it. The set of all these constraints will be denoted by $\mathcal{C}$ in the sequel.

It has been observed that mucoid *P. aeruginosa* can continue to produce mucus isolated from infected lungs. It is commonly thought that the mucoid state of *P. aeruginosa* is due to a mutation which cancels the inhibition of algU gene. An alternative hypothesis has been made: this mucoid state can occur in reason of an epigenetic modification, *i.e.* without mutation [32]. The models compatible with this hypothesis have been constructed in [33, 20]. We use the same example to explain our methodology in Sec. 3.

2.4 Manipulating sets of discrete descriptions

The only knowledge of the interaction graph is not sufficient to precisely determine which is the behavior of the biological system: numerous discrete descriptions can fit the constraints deduced from the interaction graph. In the example of Fig. 1, there are 6 states, so 6 parameters associated with x (with 3 possible values) and 6 with y (with 2 possible values). It results in $3^6 \times 2^6 = 46656$ different discrete descriptions.[1] With the equalities described in example 3, there remain $3^4 \times 2^2 = 324$ discrete descriptions, since parameters $K(x, \emptyset)$, $K(x, \{x\})$, $K(x, \{y\})$ and $K(x, \{x, y\})$ can take three different values (0, 1 or 2), and parameters $K(y, \emptyset)$ and $K(y, \{x\})$ can take two different values (0 or 1). The assumption that $K(x, \{y\}) = 0$ and $K(y, \emptyset) = 0$ reduce the set of possible discrete descriptions to $3^3 \times 2 = 54$ elements. Finally 28 of these discrete descriptions verify the inequalities deduced from the signs of interactions in example 4.

In order to precise the behavior of the biological system, complementary biological knowledge, different from previously used interaction graphs, have to be taken into consideration. To reduce the set of acceptable discrete descriptions we will express biological knowledge by temporal logic formulas involving equalities and inequalities on gene expression levels. Then model checking techniques combined with symbolic execution of the symbolic model denoting sets of acceptable discrete description will give will give us the set of acceptable parameters.

3 Symbolic formal methods

3.1 Symbolic transition systems

A symbolic transition system (STS) [22] is a transition system whose transitions are labeled by conditions on STS variables and assignments of STS variables. Each initialization of STS variables yields a basic model where each variable has a precise initial value, and all transitions are defined according to the STS transitions. Thus a STS is parameterized by an initialization function.

Let M be a STS, $V = \{v_1, v_2, \ldots, v_k\}$ the set of STS variables; then an initialization function of M is a map from V to the set of possible values of the variables. If σ is an initialization function, M_σ denotes a basic model whose first state is $(\sigma(v_1), \ldots, \sigma(v_k))$. So we can associate to M the set of all basic models obtained by applying an initialization function: $\{M_\sigma \mid \sigma \text{ initialization function}\}$ denotes this set.

A STS can represent a set of discrete descriptions associated to an interaction graph. In this case, STS variables are divided into two subsets:

- the set of variables $\{x_i \mid 0 \leq i \leq n\}$;
- the set of parameters $\{K(x_i, E) \mid 0 \leq i \leq n, E \in \{0, \ldots, max_1\} \times \cdots \times \{0, \ldots, max_n\}\}$ of the associated discrete descriptions (max_i is the maximal value of variable x_i).

The transitions are labeled according to the rules defined in Sec. 2.1. Nevertheless we need to take into account additional knowledge corresponding to constraints deduced

[1] Let us recall that a discrete description is completely defined by the values of parameters. However there are only $2^{10} \times 3^2 = 9216$ different dynamics, *i.e.* different transition graphs, for these discrete descriptions. Indeed, two different values of parameters can lead to the same dynamics because the parameters give only the directions of evolution.

from interactions. These constraints can naturally be expressed as first order formulas over the set of parameters. So we call symbolic model any couple $(M, \mathcal{C})$, where M is the STS with parameters $\{K(x_i, E)\}$ and variables $\{x_i\}$ as STS variables and $\mathcal{C}$ a set of constraints over parameters $\{K(x_i, E)\}$. It defines a set of basic models $\{M_\sigma \mid \sigma$ initialization function $\wedge \forall C \in \mathcal{C}, \sigma \models C\}$, where $\sigma \models C$ means that the parameters instantiated by σ satisfy the constraint C. Each basic model M_σ is then a discrete description associated to the values of parameters defined by σ (but with one distinguished initial state).

For the same instantiation of the parameters, every instantiation of the variables $\{x_i\}$ corresponds to the same discrete description; so a discrete description is completely defined by an initialization function σ' assigning a value only to parameters. Initialization of variables x_i allows one to specify initial states of the system if necessary.

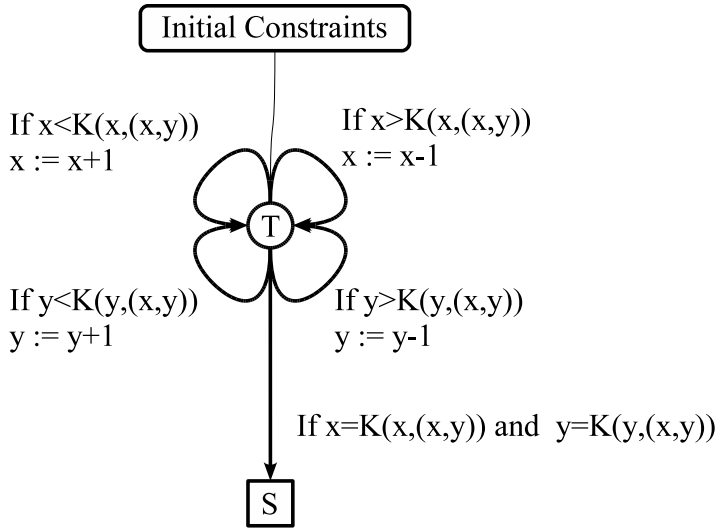

Figure 2: STS associated with Fig. 1. Arrows represent the transitions, labeled by a condition and an assignment.

Example 5 *Figure 2 represents the symbolic model associated with the interaction graph of Fig. 1, corresponding to the network of mucus production in* P. aeruginosa*. Initial constraints on parameters, denoted by $\mathcal{C}$, are specified in Secs. 2.2 and 2.3. The* control point *denoted by T in Fig. 2, indicates that the system is in a transient state (*i.e. *non-steady state), whereas the control point denoted by S indicates that the system has reached a steady state. We see that there are four different transitions from T to T: two of them correspond to a change of x and two of them correspond to a change of y. The transition from T to S occurs when all parameters of the current state are equal to the current values of the variables x and y.*

3.2 Symbolic execution

Symbolic execution has been introduced for analysis purposes of computer programs [34]. The method has been extended to STSs, and is used in the Agatha tool for behavioral analysis [35] and conformance testing [36]. As the known constraints and rules of evolution of a discrete description can easily been specified in a STS, we have adapted symbolic execution techniques to generate all behaviors compatible with the constraints on the parameters.

The method constructs a tree whose vertices are states labeled by constraints, with the following rules:

- The root of the tree is a state, associated with the initial constraints $\mathcal{C}$.
- Let us suppose that E is an already constructed state of the tree, labeled by the constraints $\mathcal{C}_E$, and that there is a STS transition from E to E' labeled by the condition D. The state E' provided with the constraint $\mathcal{C}_{E'} = \mathcal{C}_E \cup \{D\}$ is built if and only if the conjunction of the constraints of $\mathcal{C}_E \cup \{D\}$ is satisfiable. A new transition is built from $(E, \mathcal{C}_E)$ to $(E', \mathcal{C}_{E'})$.
- The process is repeated until the new state has already been encountered in the tree path from the root to the current state.

Let us point out that every state in the tree is associated with constraints whose conjunction is called *path condition*; this path condition is the condition on parameters under which the path exists.

Example 6 *Figure 3 shows the symbolic execution of the symbolic model associated with mucus production system in* P. aeruginosa*, with* $(x, y) = (0, 1)$ *as initial state, and* $\mathcal{C}$ *as initial constraints, as described in Sec. 2.3. The states in circles correspond to the control point T in the STS of Fig. 2, whereas states in squares correspond to the control point S,* i.e. *to steady states.*
Each state of the figure is associated with constraints; for example:

- $(0, 0)$ *is the only successor of* $(0, 1)$ *because initial constraints contain the equalities* $K(x, \{y\}) = 0$ *and* $K(y, \emptyset) = 0$*,* i.e. $K(x, (0, 1)) = 0$ *and* $K(y, (0, 1)) = 0$*. So the associated constraint associated with the state* $(0, 0)$ *in a circle is simply* $\mathcal{C}$*.*
- $(1, 0)$ *is a successor of* $(0, 0)$ *if* $K(x, \emptyset) > 0$*. So the set of constraints associated with* $(1, 0)$ *is* $\mathcal{C} \cup \{K(x, \emptyset) > 0\}$*.*
- $(0, 0)$ *is a steady state if* $(K(x, \emptyset) = 0 \wedge K(y, \emptyset) = 0)$*. So the set of constraints associated with the state* $(0, 0)$ *in a square is* $\mathcal{C} \cup \{(K(x, \emptyset) = 0 \wedge K(y, \emptyset) = 0)\}$ *which is equivalent to* $\mathcal{C} \cup \{(K(x, \emptyset) = 0)\}$ *as* $K(y, \emptyset) = 0$ *is contained in* $\mathcal{C}$*.*
- $(0, 1)$ *is not a successor of* $(0, 0)$ *because in this case* $K(y, \emptyset) > 0$*, which is not compatible with the initial constraint* $K(y, \emptyset) = 0$*.*

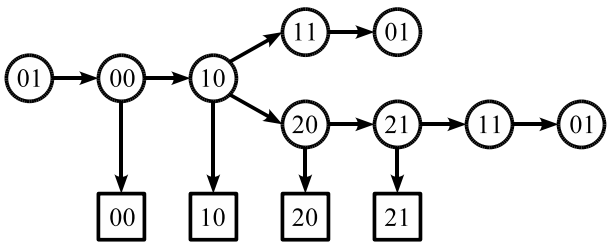

Figure 3: Symbolic execution of the STS of Fig. 2 from the initial state $(0, 1)$. Squares indicate steady states. For simplicity reason, the constraints labeling vertices are not represented in the figure.

Let us point out the reason why the construction of a path of the symbolic execution stops when the new state E has already been encountered in the tree path from the root. Actually, when this case occurs, the path condition of this new state is sufficient to lead to an infinite path repeating the states from E to E. For example in Fig. 3, under the constraints $\mathcal{C}_{01}$ associated to the last state of the path $01 \rightarrow 00 \rightarrow 10 \rightarrow 11 \rightarrow 01$, this path can be repeated infinitely because the constraints that are needed to make the path again are already contained in $\mathcal{C}_{01}$.

Very often, the construction of a path can be terminated before the occurrence of the previous condition (*i.e.* before than the new state has already been encountered in the tree path from the root). Actually, when the couple of the new state and its associated constraints have already been constructed in another path, we can be sure that the possible successors of this couple are precisely the same than the successors of the already constructed state. This case occurs when the same set of parameters leads to the same state by different pathways, which is usual in reason of the asynchronous updating of the variables. In this case the size of the symbolic execution tree can be reduced. The following example illustrates this point.

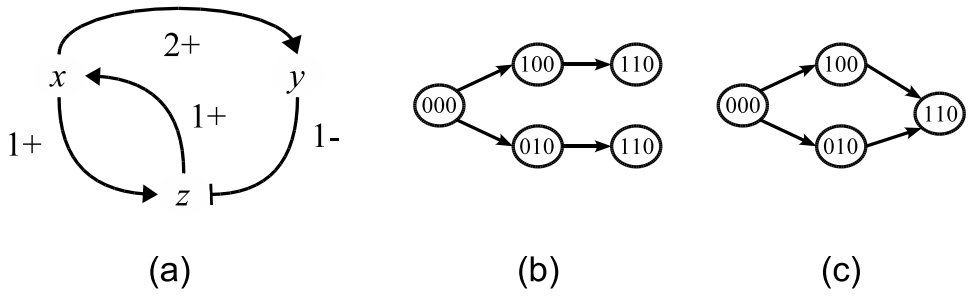

Figure 4: Illustration of the reduction of the symbolic execution. (a) is the interaction graph, (b) a part of the symbolic execution, (c) the same reduced symbolic execution.

Example 7 *We consider the system of three variables x, y, z associated with the interaction graph of Fig. 4(a). Part of the symbolic execution of the associated symbolic model from initial state $(x, y, z) = (0, 0, 0)$ is represented in the same figure (Fig. 4(b)). The condition associated to the path $000 \rightarrow 010 \rightarrow 110$ is $C = (K(y, 000) > 0 \wedge K(x, 010) > 0)$. The path condition of $000 \rightarrow 100 \rightarrow 110$ is $C' = (K(x, 000) > 0 \wedge K(y, 010) > 0)$. But from the interaction graph, we can deduce that $K(x, 000) = K(x, 010) = K(x, \emptyset)$ and that $K(y, 000) = K(y, 010) = K(y, \emptyset)$. Therefore, C and C' can be written $K(x, \emptyset) > 0 \wedge K(y, \emptyset) > 0$. Finally, as $(110, C) = (110, C')$, successors of one couple in the symbolic execution tree are exactly successors of the other; symbolic execution tree can be represented by Fig. 4(c).*

3.3 Specification of paths and synthesis of constraints on parameters

3.3.1 Linear temporal logic

To search a specific path in the symbolic execution tree we adapt model-checking techniques for linear temporal logic (LTL) [23]. Intuitively model-checking techniques consist in exploring all states of a basic model to state whether this model satisfies or not a given temporal logic formula [37].

A LTL formula expresses properties of a path. This logic adds to the classical operators of propositional logic[2] mainly two temporal operators, called Next (N), and Until (U). If f and g are formulas, Nf means that f is true in the following state of the path, and fUg means that f is true in each state of the path, until g becomes true (and g eventually happens). We can then define the operators Finally (F) and Globally (G); Ff means that f eventually happens (and can be written $\top Uf$); Gf means that f is always true (and can be written $\neg F(\neg f)$).

As a LTL formula expresses a property of a single path, there are two ways to use it to express a property of a discrete description. On the one hand we may want to express that *all* paths of the model have the specified behavior; we say that this property is universal. On the other hand, we may want to express that there exists at least a path in the model with the specified behavior; we say that this property is existential. The distinction is important because universal or existential properties can not be treated exactly by the same method (see Sec. 3.3.2).

Examples of temporal properties

Temporal properties of interest in a model include the existence of a path from a given set of states to another one. If for example there is a path from a state where a variable x is at its basal level 0 to a state where x is at its maximal value 2, it means that there is a path verifying $x = 0 \wedge F(x = 2)$, *i.e.* a path such that in its first state $x = 0$ and that eventually reaches a state where $x = 2$. Such properties can be known from experiments or can be hypotheses of interest. We will see in Sec. 4.2.1 examples of such properties.

The negation of the previous properties are also useful: they mean that a given set of states can not be reached from another one. This kind of property is used in Sec. 3.3.3.

Another current property can be the knowledge that a set of states is stable, *i.e.* that when the system is in these states, there is no path going out. This can include steady states, or stable cyclic behaviors. For example in a system of two variables (x, y), if S is the set of stable states, all paths must verify $(x, y) \in S \Rightarrow G((x, y) \in S)$. It means that there is no path verifying $(x, y) \in S \wedge F((x, y) \notin S)$. We will use in Sec. 4.2.1 examples of such properties.

More sophisticated properties can be expressed. For example, we can express that from a given set of states there exists a path such that this set will be infinitely revisited. Such paths verify the property $(x, y) \in S \wedge GF((x, y) \in S)$ (*i.e.* there is a path whose first state is in S, and from all states of the path, S will be reached in the future). This property can also hold for *all paths* beginning in S; then all paths verify $(x, y) \in S \Rightarrow GF((x, y) \in S)$. This is the type of property used in Sec. 3.3.4.

Let us suppose that the set A of states is an attractor of the system and S is its basin of attraction; then from every state in S, the set A will eventually be reached, and the system will then stay in this set A. It means that all paths verify $(x, y) \in S \Rightarrow FG((x, y) \in A)$ (*i.e.* paths beginning in S are such that after a certain time, all their states are in A; or Finally, all states are Globally in A).

[2] As $\neg$ (not), $\wedge$ (and), $\vee$ (or), $\Rightarrow$ (implies), $\top$ (true), $\bot$ (false).

3.3.2 Extended LTL model-checking

We extend classical LTL model-checking techniques designed for basic models to STSs. Just as classical LTL model-checking only considers pertinent paths according to the formula, our method also considers pertinent paths according to the formula, but in our case each state of a path is provided with constraints on parameters. The key point is that a path is eliminated as soon as the conjunction of constraints is no more satisfiable. This leads to a minimal tree construction and gives us the solutions in term of constraints: the disjunction of the path conditions associated to all remaining paths. The resulting constraint represents all parameter valuations compatible with the behavior specified by the formula. To summarize, given a symbolic model $(M, \mathcal{C})$, extended LTL model-checking allows us to compute all initialization functions (*i.e.* parameter valuations) leading to basic models satisfying a LTL formula. In other words, the extended LTL model-checking associates to any LTL formula a characteristic constraint defining the discrete descriptions satisfying it.

Let us remark that the developed technique constructs the disjunction of constraints on possible paths. Then satisfying a LTL formula for a model means that there exists at least a path satisfying the LTL formula. As said before, such a property is qualified as existential. On the contrary we may want to select models whose all paths satisfy the formula (universal property). In such a case the negation of the universal property is unsatisfiable. We have then to specify this impossible behavior as a LTL formula. It suffices to take the negation of the associated constraint to find all models compatible with the universal property. An example is given in next subsection (Sec. 3.3.3).

3.3.3 Adding knowledge to the symbolic model

When considering behaviors, expressed as LTL formulas, supposed to be known to occur in the actual system, we can add the corresponding characteristic constraints $\mathcal{D}$ to the symbolic model $(M, \mathcal{C})$. We get the symbolic model $(M, \mathcal{C} \cup \mathcal{D})$ restricting the set of discrete descriptions.

Example 8 *From a state where AlgU is at its basal level,* P. aeruginosa *will not produce mucus in a common environment, so there is no path from a state where $x = 0$ to a state where $x = 2$. That is clearly an universal property. In order to show that it is not possible to reach $x = 2$ from $x = 0$, we consider the formula $(x = 0) \wedge F(x = 2)$. The associated constraint, generated by our method, and added to initial constraints $\mathcal{C}$ is $K(x, \emptyset) > 1$. The negation is simply $K(x, \emptyset) \leq 1$. All discrete descriptions verifying $\mathcal{C}$ and the latter constraint satisfy the universal property. In the sequel we denote $\mathcal{C}' = \mathcal{C} \cup \{K(x, \emptyset) \leq 1\}$.*

3.3.4 Extracting knowledge from the symbolic model

Let us come back to the two central questions asked in the Introduction: is there any model coherent with a certain hypothetic behavior? Are there behaviors common to all possible models?

The first question consists in specifying the hypothesis with LTL formulas, and finding the associated constraints. When the constraints are not satisfiable, there is no model compatible with the LTL formulas. When they are satisfiable, the solutions of the constraints give all parameter valuations, each one corresponding to a discrete description satisfying the LTL formulas (see example 9).

The second question consists in finding properties common to all discrete descriptions associated to a symbolic model $(M, \mathcal{C})$. The set of constraints $\mathcal{C}$ precisely represents such common properties; then every behavior implied by these constraints is a common behavior to all selected discrete descriptions (see Sec. 4.2.3 for an illustration).

Example 9 *If the hypothesis of an epigenetic change in mucoid* P. aeruginosa *is verified, bacteria which produce mucus can continue to produce mucus in a common environment. A path beginning with $x = 2$ which revisits infinitely a state where $x = 2$ is described by the formula $(x = 2) \wedge GF(x = 2)$. The resulting constraint, added to $\mathcal{C}'$, is*

$$[K(x, \{x, y\}) = 2 \wedge K(y, \{x\}) = 1] \vee [K(x, \{x\}) = 2 \wedge K(y, \{x\}) = 0]$$

This constraint implies that the (mucoid) state $(2, 1)$ is a steady state, or that $(2, 0)$ is a steady state.
Let us point out that there is another path compatible with $\mathcal{C}$ and verifying $(x = 2) \wedge GF(x = 2)$ (given in Fig. 5). But in this path, $K(x, (1, 0)) > 1$, because there is a transition from the state $(1, 0)$ to the state $(2, 0)$; as $K(x, \emptyset) = K(x, (1, 0))$, it is not compatible with $K(x, \emptyset) \leq 1$, and therefore with $\mathcal{C}'$.

There are 8 discrete descriptions verifying the constraints; in these models the mucoid state can be related to an epigenetic modification. These constraints imply the existence of a stable mucoid state, but not that all paths from a mucoid state come back to a mucoid state. This more restrictive behavior, is achieved if $K(x, \{x, y\}) > 1$, i.e. *for 4 models from the 8.*

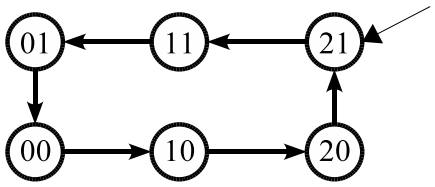

Figure 5: Example of a path of the STS of Fig. 2 verifying $(x = 2) \wedge GF(x = 2)$ (if $(2, 0)$ or $(2, 1)$ is the initial state of the path), compatible with $\mathcal{C}$ but not with $\mathcal{C}'$.

4 Application to immunity control in bacteriophage lambda

4.1 Immunity control in bacteriophage lambda

Bacteriophage lambda is a virus whose DNA can integrate into bacterial chromosome and be faithfully transmitted to the bacterial progeny. After infection, most of the bacteria display a lytic response and liberate new phages, but some display a lysogenic response, *i.e.* survive and carry lambda genome, becoming immune to infection. Figure 6 is the graph of interactions described by Thieffry and Thomas [9] which has also been studied in [38]. Four genes are involved, called cI, cro, cII and N. The states, represented by a vector (cI, cro, cII, N), are in $\{0, 1, 2\} \times \{0, 1, 2, 3\} \times \{0, 1\} \times \{0, 1\}$. Even with the constraints deduced following Sec. 2.2, the associated symbolic model represents 1 008 000 different discrete descriptions.

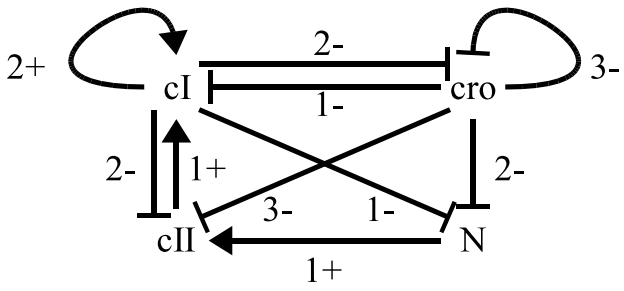

Figure 6: Graph of interactions associated with immunity control in bacteriophage lambda. Arrows are labeled by the threshold and sign of the corresponding interaction. For clarity blunt arrows indicate the negative interactions.

4.2 Lytic and lysogenic pathways of bacteriophage lambda

4.2.1 Specification of behaviors by LTL formulas

First we have to specify the set of states of interest. The lytic response leads to the states where cro is fully expressed, and other genes repressed. So $(0, 2, 0, 0)$ and $(0, 3, 0, 0)$ are called lytic states. To specify that the system is in one of these states, we use the following formula, called $lytic$:

$$lytic = (\text{cI} = 0 \wedge \text{cro} \geq 2 \wedge \text{cII} = 0 \wedge \text{N} = 0).$$

The lysogenic response leads to the state where cI is fully expressed, and the repressor produced by cI blocks the expression of the other viral genes, leading to immunity. So $(2, 0, 0, 0)$ is called lysogenic state. To specify that the system is in this state, we use the following formula, called $lysogenic$:

$$lysogenic = (\text{cI} = 2 \wedge \text{cro} = 0 \wedge \text{cII} = 0 \wedge \text{N} = 0).$$

The viral proteins are initially absent when the viral genome integrates a cell; so the initial state is $(0, 0, 0, 0)$. The system is in this initial state if it verifies the following $init$ formula:

$$init = (\text{cI} = 0 \wedge \text{cro} = 0 \wedge \text{cII} = 0 \wedge \text{N} = 0).$$

When the system reaches the set of lytic state it does not leave it; the stability of these states is an universal property. So we translate this property into the equivalent property $\mathcal{P}_1$:

- $\mathcal{P}_1$: there is no path verifying $lytic \wedge F(\neg lytic)$.

Similarly, the stability of the lysogenic state is an universal property, equivalent to the property $\mathcal{P}_2$:

- $\mathcal{P}_2$: there is no path verifying $lysogenic \wedge F(\neg lysogenic)$.

As lytic and lysogenic responses are possible from the initial state, it means that there exists at least a path from initial state to lytic states, and at least a path from initial state to lysogenic state. These properties are translated into $\mathcal{P}_3$ and $\mathcal{P}_4$:

- $\mathcal{P}_3$: there is a path verifying $init \wedge F(lytic)$;
- $\mathcal{P}_4$: there is a path verifying $init \wedge F(lysogenic)$.

4.2.2 Resulting constraints on parameters

In the sequel $\mathcal{C}_\lambda$ denotes the set of initial constraints associated with the interaction graph of Fig. 6 following the rules described in Sec. 2.2. We apply the extended model-checking method to the associated symbolic model, to find the constraints that have to be added to $\mathcal{C}_\lambda$.

To obtain the additional constraints associated with $\mathcal{P}_1$, we first generate the disjunction of the conditions leading to a path verifying $lytic \wedge F(\neg lytic)$. The negation of this disjunction is:

$$C_1 = [K(\text{cI}, \{\text{cro}\}) = 0 \wedge K(\text{cro}, \emptyset) > 1 \wedge K(\text{cII}, \emptyset) = 0 \wedge K(\text{N}, \{\text{cro}\}) = 0].$$

Similarly the negation of the constraints associated to $init \wedge F(lysogenic)$ is

$$C_2 = [K(\text{cI}, \{\text{cI}\}) = 2 \wedge K(\text{cro}, \{\text{cI}\}) = 0 \wedge K(\text{cII}, \{\text{cI}\}) = 0 \wedge K(\text{N}, \{\text{cI}\}) = 0].$$

These two constraints can be added to $\mathcal{C}_\lambda$ in the symbolic model. The discrete descriptions verifying these constraints verify $\mathcal{P}_1$ and $\mathcal{P}_2$.

By the same method applied on the symbolic model with the constraint $\mathcal{C}_\lambda \cup \{C_1, C_2\}$, we generate the additional constraint needed to verify $\mathcal{P}_3$. This constraint is $\top$ (the always true proposition): it means that all discrete descriptions whose parameters verify $\mathcal{C}_\lambda \cup \{C_1, C_2\}$ have a path verifying $init \wedge F(lytic)$.

Finally, the additional constraint associated with $\mathcal{P}_4$ and $init \wedge F(lysogenic)$ (obtained by disjunction of path conditions) is

$$C_4 = [K(\text{cI}, \emptyset) = 2] \vee [K(\text{cI}, \{\text{cII}\}) = 2 \wedge K(\text{cII}, \{\text{N}\}) = 1 \wedge K(\text{N}, \emptyset) = 1].$$

The discrete descriptions whose parameters verify $\mathcal{C}_\lambda \cup \{C_1, C_2, C_4\}$ are the discrete descriptions associated with immunity control that verify the properties $\mathcal{P}_1$ to $\mathcal{P}_4$.

4.2.3 Questioning the symbolic model

In this subsection we show that there are pathways to lysis or lysogeny common to all discrete descriptions whose parameters verify $\mathcal{C}_\lambda \cup \{C_1, C_2, C_4\}$. For simplicity, the states of values of $(\text{cI}, \text{cro}, \text{cII}, \text{N})$ are denoted by (0000), (0100), etc.

In all these discrete descriptions $K(\text{cro}, \emptyset) > 1$ (it is a consequence of C_1). But $K(\text{cro}, 0000) = K(\text{cro}, 0100)$ as these parameters are equal to $K(\text{cro}, \emptyset)$; so they are at least equal to 2. So it is clearly a sufficient condition to demonstrate that in all discrete descriptions there is the following path to lysis:

$$(0000) \rightarrow (0100) \rightarrow (0200) \tag{1}$$

The constraint C_4 is

$$[K(\text{cI}, \emptyset) = 2] \vee [K(\text{cI}, \{\text{cII}\}) = 2 \wedge K(\text{cII}, \{\text{N}\}) = 1 \wedge K(\text{N}, \emptyset) = 1].$$

So all discrete descriptions verify at least one of the properties C or C':

- $C = [K(\text{cI}, \emptyset) = 2]$;
- $C' = [K(\text{cI}, \{\text{cII}\}) = 2 \wedge K(\text{cII}, \{\text{N}\}) = 1 \wedge K(\text{N}, \emptyset) = 1]$.

First we look at the discrete descriptions verifying the first constraint C. As $K(\text{cI}, \emptyset) = K(\text{cro}, 0000) = K(\text{cro}, 1000)$, all discrete descriptions such that $K(\text{cI}, \emptyset) = 2$ have the following path to lysogeny:

$$(0000) \rightarrow (1000) \rightarrow (2000) \tag{2}$$

Now we consider the second constraint C'.

- As $K(\text{N}, \emptyset) = 1$, there is a transition $(0000) \rightarrow (0001)$.
- As $K(\text{cII}, \{\text{N}\}) = 1$, and $K(\text{cII}, \{\text{N}\}) = K(\text{cII}, 0001)$, there is a transition $(0001) \rightarrow (0011)$.
- As $K(\text{cI}, \{\text{cII}\}) = 2$, and $K(\text{cI}, \{\text{cII}\}) = K(\text{cI}, 0011) = K(\text{cI}, 1011)$, there is a path $(0011) \rightarrow (1011) \rightarrow (2011)$.
- The constraint C_2 implies that $K(\text{N}, \{\text{cI}\}) = 0$, then $K(\text{N}, 2011) = 0$. So there is a transition $(2011) \rightarrow (2010)$.
- The constraint C_2 implies that $K(\text{cII}, \{\text{cI}\}) = 0$, so $K(\text{cII}, 2010) = 0$. So there is a transition $(2010) \rightarrow (2000)$.

Therefore all discrete descriptions verifying C' have the following path to lysogeny:

$$(0000) \rightarrow (0001) \rightarrow (0011) \rightarrow (1011) \rightarrow (2011) \rightarrow (2010) \rightarrow (2000) \tag{3}$$

Interestingly, this last path is precisely the most likely pathway to lysogeny according to experimental knowledge, as described by Thieffry and Thomas [9].

A precise count of the number of discrete descriptions reveals that there are 2156 discrete descriptions verifying $\mathcal{C}_\lambda \cup \{C_1, C_2, C_4\}$. In all these discrete descriptions, there is a common pathway from initial state to lysis: pathway (1). There are 1176 of these discrete descriptions verifying C. They are discrete descriptions with a common pathway to lysogeny: pathway (2). Moreover there are 1470 discrete descriptions (out of the 2156) verifying C'; they have the common pathway (3) to lysogeny. 490 discrete descriptions verify C *and* C': they are the discrete descriptions with at least two different pathways to lysogeny, pathway (2) and pathway (3).

5 Conclusion

We have shown how a symbolic model representing a set of possible discrete descriptions of a genetic regulatory network permits one to deal with incomplete knowledge. Known interactions can be translated into constraints on the parameters, which can be specified in a symbolic transition system. This STS can be simulated with symbolic execution techniques. The known behaviors can be specified with LTL formulas, and then, model-checking techniques have been extended to select the constraints on parameters associated with these behaviors. Adding these contraints to the STS, a symbolic model representing all discrete descriptions coherent with the known behaviors is obtained.

Then we have explained how the symbolic model can be used to reveal new results: the possibility of hypothetic behaviors can be tested (as the epigenetic change in *P. aeruginosa*) or common behaviors between all selected descriptions can be found (as possible pathways to lysis or lysogeny in bacteriophage lambda).

By using SMBioNet to analyze the regulatory network of the cytotoxicity of *P. aeruginosa* [33], models coherent with the hypothesis of the existence of an epigenetic switch between non-inducible states and inducible ones have been constructed. The underlying interaction graph used was similar to the interaction graph associated with mucus production (in Fig. 1). This theoretical results have lead to new experimental results [39]. It is now interesting to take into account these new results into a more elaborated model, in particular by including other important proteins implicated in the network. The efficiency of the methods presented in this chapter should allow us to construct and analyze this more complex model. It is a work that we plan to do in the context of the observability working group of Epigenomics Project of Genopole®, Evry.

Acknowledgements

We gratefully acknowledge the members of the observability working group of Genopole® for stimulating interactions. In particular Gilles Bernot and Janine Guespin have encouraged us in the development of symbolic formal methods for systems biology. We thank also Epigenomics Project for constant support.

References

[1] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *J Comput Biol*, 9(1):67–103, 2002.

[2] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, 22(3):437–67, 1969.

[3] R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–585, 1973.

[4] R. Thomas and R. D'Ari. *Biological Feedback*. CRC Press, Boca Raton, Florida, 1990.

[5] L. Glass and S.A. Kauffman. The logical analysis of continuous non-linear biochemical control networks. *J Theor Biol*, 39:103–129, 1973.

[6] R. Thomas. Regulatory networks seen as asynchronous automata : a logical description. *J Theor Biol*, 153(1):1–23, 1991.

[7] R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. ii. logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, 11(1):180–95, 2001.

[8] E.H. Snoussi. Qualitative dynamics of piecewise-linear differential equations : A discrete mapping approach. *Dyn Stability Syst*, 4:189–207, 1989.

[9] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks –ii. immunity control in bacteriophage lambda. *Bull Math Biol*, 57(2):277–97, 1995.

[10] L. Mendoza, D. Thieffry, and E.R. Alvarez-Buylla. Genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15(7-8):593–606, 1999.

[11] L. Sanchez and D. Thieffry. Segmenting the fly embryo: a logical analysis of the pair-rule cross-regulatory module. *J Theor Biol*, 224(4):517–37, 2003.

[12] L. Mendoza. A network model for the control of the differentiation process in th cells. *Biosystems*, 84(2):104–14, 2006.

[13] C. Chaouiya, D. Thieffry, and L. Sanchez. From gradients to stripes: a logical analysis of drosophila segmentation genetic network. In N. Kolchanov, R. Hofestaedt, and L. Milanesi, editors, *Bioinformatics of Genome Regulation and Structure II*, pages 379–90. Springer, 2006.

[14] H. de Jong, J.-L. Gouzé, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol*, 66(2):301–340, 2004.

[15] E.H. Snoussi and R. Thomas. Logical identification of all steady states: the concept of feedback characteristic states. *Bull Math Biol*, 55:973–991, 1993.

[16] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks –i. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull Math Biol*, 57(2):247–76, 1995.

[17] A.G. Gonzalez, A. Naldi, L. Sanchez, D. Thieffry, and C. Chaouiya. Ginsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks. *Biosystems*, 84(2):91–100, 2006.

[18] E. Fanchon, F. Corblin, L. Trilling, B. Hermant, and D. Gulino. Modeling the molecular network controlling adhesion between human endothelial cells: Inference and simulation using constraint logic programming. In *Proceedings of Computational Methods in Systems Biology (CMSB 2004)*, volume 3082 of *Lecture Notes in Computer Science*, pages 104–118. Springer, 2005.

[19] M.A. Schaub, T.A. Henzinger, and J. Fisher. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology*, 1(4), 2007.

[20] G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks: extending thomas' asynchronous logical approach with temporal logic. *J Theor Biol*, 229(3):339–347, 2004.

[21] G. Batt, D. Ropers, H. de Jong, J. Geiselmann, R. Mateescu, M. Page, and D. Schneider. Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in escherichia coli. *Bioinformatics*, 21(suppl. 1):i19–i28, 2005.

[22] M. Aiguier, C. Gaston, P. Le Gall, D. Longuet, and A. Touil. A temporal logic for input output symbolic transition systems. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, pages 43–50. IEEE Computer Society Press, 2005.

[23] M.Y. Vardi. An automata-theoretic approach to linear temporal logic. In *Proceedings of the Banff Higher order workshop conference on Logics for concurrency : structure versus automata*, pages 238–266, Secaucus, NJ, USA, 1996. Springer-Verlag New York.

[24] D. Mateus, J.-P. Gallois, and P. Le Gall. Évaluation symbolique appliquée à l'étude de réseaux de régulation génétique. In *Journée thématique RIAMS, Réseaux d'interaction : analyse, modélisation et simulation*, pages 19–24. Rapport de Recherche LaMI 121, 2005.

[25] D. Mateus and J.-P. Gallois. Searching constraints in biological regulatory networks using symbolic analysis. In *Proceedings of the fifth international conference on Bioinformatics of Genome Regulation and Structure (BGRS'06)*, volume 3, pages 78–81, 2006.

[26] D. Mateus, J.-P. Comet, J.-P. Gallois, and P. Le Gall. Modeling genetic regulatory networks from specified behaviors. Oral presentation at Integrative Post-Genomics (IPG'06), Lyon, France, December 2006.

[27] D. Mateus, J.-P. Comet, J.-P. Gallois, and P. Le Gall. Modelling genetic regulatory networks from specified behaviours. Oral presentation at BioSysBio'07, Manchester, UK, January 2007.

[28] D. Mateus, J.-P. Comet, J.-P. Gallois, and P. Le Gall. Symbolic modeling of genetic regulatory networks. *J Bioinform Comput Biol*, (to appear), 2007.

[29] C. Bigot, A. Faivre, J.-P. Gallois, A. Lapitre, D. Lugato, J.-Y. Pierron, and N. Rapin. Automatic test generation with agatha. In *Proceedings of Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2003)*, volume 2619 of *Lecture Notes in Computer Science*, pages 591–596. Springer, 2003.

[30] D. Lugato, C. Bigot, Y. Valot, J.-P. Gallois, S. Gérard, and F. Terrier. Validation and automatic test generation on uml models: the agatha approach. *International Journal on Software Tools for Technology Transfer (STTT)*, 5(2-3):124–139, 2004.

[31] J.R. Govan and V. Deretic. Microbial pathogenesis in cystic fibrosis: mucoid pseudomonas aeruginosa and burkholderia cepacia. *Microbiol rev*, 60(3):539–74, 1996.

[32] J. Guespin-Michel and M. Kaufman. Positive feedback circuits and adaptive regulations in bacteria. *Acta Biotheor*, 49(4):207–218, 2001.

[33] J. Guespin-Michel, G. Bernot, J.-P. Comet, A. Mérieau, A. Richard, C. Hulen, and B. Polack. Epigenesis and dynamic similarity in two regulatory networks in pseudomonas aeruginosa. *Acta Biotheor*, 52(4):379–390, 2004.

[34] J.C. King. Symbolic execution and program testing. *Commun ACM*, 19(7):385–394, 1976.

[35] N. Rapin, C. Gaston, A. Lapitre, and J.-P. Gallois. Behavioral unfolding of formal specifications based on communicating extended automata. In *Proceedings of the first international workshop on Automated Technology for Verification and Analysis (ATVA'03)*, National Taipei University, Taiwan, 2003.

[36] C. Gaston, P. Le Gall, N. Rapin, and A. Touil. Symbolic execution techniques for test purpose definition. In *Proceedings of Testing of Communicating Systems (TestCom 2006)*, volume 3964 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2006.

[37] E.M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, Cambridge, Mass., 2000.

[38] A. Richard, J.-P. Comet, and G. Bernot. Formal methods for modeling biological regulatory networks. In A.H. Gabbar, editor, *Modern Formal Methods and Applications*, pages 83–122. Springer, 2006.

[39] D. Filopon, A. Mérieau, G. Bernot, J.-P. Comet, R. Leberre, B. Guery, B. Polack, and J. Guespin-Michel. Epigenetic acquisition of inducibility of type iii citoxicity in p. aeruginosa. *BMC Bioinformatics*, 7(272), 2006.

Hypothesis: Variations in the rate of DNA replication determine the phenotype of daughter cells

Vic Norris[1,4], Laurent Jannière[2] and Patrick Amar[3,4]

[1] UMR CNRS 6522, Department of Science, Univ. of Rouen, F-76821 Mont Saint Aignan, France
[2] Laboratoire de Génétique Microbienne, INRA, F-78352 Jouy en Josas, France
[3] LRI, Univervité Paris-Sud & UMR CNRS 8623, F-91405 Orsay, France
[4] Epigenomics Project, Genopole®, F-91000 Evry, France

Abstract

The existence of two identical chromosomes within the same cell in which genes and higher order structures compete for limited resources is a symmetry-breaking situation previously proposed to lead to differentiation. Recent experiments are consistent with an intimate relationship between metabolism and the rate of chromosome replication in bacteria. The process of chromosome replication progressively changes the copy number of genes and sites in a linear order. This raises the possibility that slowing or even pausing replication for different times at different sites in the chromosome might be combined with various mechanisms leading to local cooperation and global competition. If so, such replication-phenotype coupling could produce different patterns of gene expression. Indeed, replication-phenotype coupling may constitute a powerful and fundamental way of generating coherent phenotypes. As a prelude to testing this hypothesis, we discuss some of the parameters that will need to be explored by bench experimentation and computer simulation.

1 Introduction

One of the fundamental problems in biology, highlighted by Kauffman [15], is how cells integrate gene expression and environmental conditions to steer their phenotypes in a coherent, reproducible way through the vast space of possibilities apparently available to them. A possible solution is that the very existence of two chemically identical chromosomes in the same cytoplasm spontaneously leads to different patterns of gene expression and that this underpins differentiation [26]. This is based on the idea that if a gene attracts an RNA polymerase it has a greater chance of attracting a second one and hence, if two identical copies of a gene compete for a limited number of RNA polymerases, one copy is expressed and the other silent. Related ideas about the primordial role of the cell cycle in generating not just diversity but coherent diversity have also been developed [31, 25].

Such ideas need to be updated in the context in which gene expression takes place at the level of hyperstructures which are large, spatially extended assemblies of ions, molecules and macromolecules that are implicated in functions that range from DNA replication and cell division to chemotaxis and secretion [24]. These ideas also require updating due to the discovery that carbon metabolism in *Bacillus subtilis*, and almost certainly other bacteria, affects the enzymes responsible for the elongation step in chromosome replication [14]. In other words, metabolism appears to be exerting a direct control over the way the chromosome is replicated. This suggests to us a reciprocal relationship in which the way the chromosome is replicated determines the phenotype. Here we explore this idea.

2 Hypothesis

By slowing or accelerating the elongation step of DNA replication in different regions of the genome, a bacterium generates different patterns of copy numbers of different genetic elements and this results in different, coherent, phenotypes in daughter cells. This differentiation occurs via several mechanisms that are characterised by local synergistic relationships and global competitive ones.

3 Mechanisms

Sources of local positive feedback are based on cooperation and include:

1. the movement of genes during their transcription by RNA polymerase from the nucleoid where they are relatively inaccessible to RNA polymerase to a position on the periphery of the nucleoid where they are easily accessible.

2. the bringing together of different genes or sites on nucleic acids into a hyperstructure by factors such as protein binding that then increases the probability of these proteins binding (e.g. by raising their local concentration).

3. nucleation phenomena such that once a critical size has been reached subsequent assembly is faster (as observed in the polymerisation of eukaryotic actin *in vitro*).

Sources of global negative regulation are based on competition and include the limitations on the:

1. quantities of the transcriptional and translational machinery.

2. physical space within the cytoplasm and membrane.

Local positive feedback and global negative regulation can act via either activation or repression of gene expression (see Corollaries).

3.1 Activation mechanisms

Consider an activator of transcription that has two types of binding site, low affinity and high affinity (empty and filled circles, respectively) distributed as shown in Figure 1. This activator can form oligomers such that there is the possibility of *in cis* interactions between the activators binding to the four sites on the top daughter chromosome and, separately, between the activators binding to the four sites on the lower daughter chromosome. Replication must slow down or pause at one and only one of two rest-stops. The parameters here include the number of activators, the number and proximity of sites, the association constants between activator proteins and between activator proteins and their sites (low and high affinity), and diffusion coefficients, all of which contribute to the time taken for activators to interact to form an effective structure to activate transcription.

Other parameters include the length of time the replication fork remains at a rest-stop and the time between successive rest-stops (in this simple model, replication is effectively instantaneous between rest-stops and stationary at rest-stops; a more realistic model would have regions between rest-stops where replication is relatively fast and the rest-stops themselves would be other regions where replication is relatively slow) as well

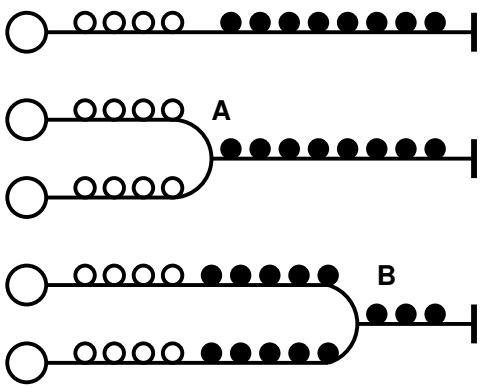

Figure 1: Transcriptional activation via binding sites. The large circle represents the origin of replication of the chromosome and the bar represents the terminus. The small empty circles are low affinity binding sites for transcriptional activators whilst the filled circles are high affinity binding sites. **A** and **B** represent two rest-stops for replication.

as the position of the gene encoding the limiting factor and whether its own expression is under control of this factor. Here we assume that the activator is being produced so as to yield a constant concentration (but see Corollaries). If replication pauses at rest-stop A, the two sets of four low affinity sites are in competition with the eight high affinity sites and, depending on the choice of parameter values, this can allow activation of transcription from *both* the top and bottom chromosomes. If pausing occurs instead at step B, the two sets of low affinity sites are in competition with thirteen high affinity sites and parameter values exist that allow activation of transcription from only one set of the four low affinity sites. The important prediction here is that a broad range of parameter values exists that results in expression from both daughter chromosomes if replication pauses at rest-stop A but only from one daughter chromosome if replication pauses instead at rest-stop B.

Figure 2: Transcriptional activation via genes. The small empty hexagons represent genes that can form part of the same hyperstructure whilst the filled empty hexagons represent other genes. Other symbols as in Fig. 1.

There are numerous variants on this theme. The activator need not be a specific protein but could be a species of phospholipid in a domain or localised structures dependent on divalent ions or polyamines or polyphosphates. Indeed, a more general activation mechanism based on a similar principle is when the activator that is limiting is RNA polymerase itself. Consider Figure 2 in which each of the small hexagons is a gene and in which the four empty hexagons on the top daughter chromosome can form part of a hyperstructure

going into one half of the cell (which will become a daughter cell) whilst the other four empty hexagons can form part of a similar hyperstructure going into the other half of the cell. Suppose that expression of a gene within a hyperstructure favours its chance of being expressed again and suppose that RNA polymerase is limiting. The competition for RNA polymerase at rest-stop A is between the two potential hyperstructures, each containing four genes, and the rest of the genome containing 24 genes whilst the competition at rest-stop B is between the two sets of four genes in these potential hyperstructures and 44 other genes. The prediction is then as above, namely, that parameter values can be found that allow expression of both hyperstructures if replication pauses at A and of only one hyperstructure if replication pauses at B.

3.2 Repression mechanisms

Now consider a repressor of transcription that has two types of binding site, high affinity and low affinity (filled and empty circles, respectively) distributed as shown in Figure 3. This repressor can form oligomers such that there is the possibility of *in cis* interactions between the repressors binding to the four high affinity sites on the top daughter chromosome and, separately, between the repressors binding to the four sites on the lower daughter chromosome.

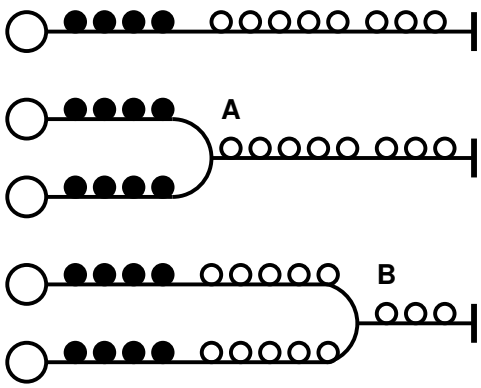

Figure 3: Transcriptional repression via binding sites. The small empty circles are low affinity binding sites whilst the filled circles are high affinity binding sites. Other symbols as in Fig. 1.

Again, replication must pause at one and only one of two rest-stops. If replication pauses at rest-stop A, the two sets of four high affinity sites are in competition for repressor with the eight low affinity sites and, depending on the choice of parameter values, this could allow two discrete repression hyperstructures to form (in which transcription would be repressed) containing the affected genes in both the top and bottom chromosomes. If pausing occurs instead at step B, the two sets of high affinity sites are in competition with thirteen low affinity sites and parameter values exist that allow only one repression hyperstructure to form and hence transcription to be repressed in only one of the future daughter cells. The important prediction here is that a broad range of parameter values exists that results in repression on both daughter chromosomes if replication pauses at rest-stop A but only on one daughter chromosome if replication pauses instead at rest-stop B.

As with the activator scenario, there are numerous variants on this theme. The repressor need not be a specific protein or RNA but could involve a preferential compaction or condensation of the regions containing the genes to be repressed into, for example, a

cholesteric phase. To continue in this vein, perhaps the most general repression mechanism would be when the repression is via denial to the space needed for transcription and translation to occur. Consider Figure 4 in which each of the small hexagons is a gene and in which the four genes (filled hexagons) on the top daughter chromosome can form part of a hyperstructure going into one half of the cell (which will become a daughter cell) whilst the other four genes (filled hexagons) can form part of a similar hyperstructure going into the other half of the cell.

Figure 4: Transcriptional repression via competition for space. The small filled hexagons represent genes that can form part of the same hyperstructure whilst the small empty hexagons represent other genes. Other symbols as in Fig. 1.

Suppose that expression of a gene within a hyperstructure favours its chance of being expressed again and suppose that RNA polymerase is limiting. The competition for RNA polymerase at rest-stop A is between the two potential hyperstructures, each containing four genes, and the rest of the genome containing 24 genes whilst the competition at rest-stop B is between the two sets of four genes in these potential hyperstructures and 44 other genes. The prediction is then as above, namely, that parameter values can be found that allow expression of both hyperstructures if replication pauses at A and of only one hyperstructure if replication pauses at B.

4 Evidence

A steadily accumulating body of evidence points to the universality of differentiation in the bacterial world [8, 4, 32]. Cell division gives a stalked and a swarmer cell in *Caulobacter crescentus*, a spore and a mother cell in *B. subtilis*, and a tetrad containing chromosomes in different states in *Deinococcus radiodurans*. Even populations of *Escherichia coli* reveal a heterogeneity that increases the probability that some cells will be ready to profit from new opportunities or survive new dangers [5]. We and others have argued that one of the primary functions of the cell cycle is to generate coherent diversity [26, 31, 25].

Studies of 'combed' chromosomes from a mutant of *E. coli* synchronised for replication reveal a heterogeneity in the pattern of replication, consistent with different rates of replication in different regions [9]. Sequences that slow or halt replication have been found in both *E. coli* and *B. subtilis*. In *E. coli*, a polar DNA replication barrier is formed when the DNA-binding protein Tus forms a complex with any of the four 23-base-pair terminator (ter) sites found in the terminus region of the chromosome (in addition to other systems [11]). In *B. subtilis*, a replication barrier exists near the origin of replication and

arrest is dependent upon the RelA protein, the action of which is correlated with high levels of the alarmone, ppGpp [3] but see [21]. Finally, there is evidence indicating a flow of information from metabolism to replication that could exploit phenotypic changes arising from changes in replication rate: firstly, in *E. coli*, the velocity of the replication fork may vary from about 1000 to 200 nt/s as a function of the energy contained in the nutrients [13, 19], secondly, in *B. subtilis*, the primase and the helicase (key enzymes in replication) appear to interact directly with metabolic enzymes such as pyruvate dehydrogenase [23, 24], an enzyme that might modulate the activity of the primase [34] and, thirdly, three *B. subtilis* enzymes known or proposed to act on the lagging-strand template in the replicating fork (the DNA polymerase DnaE, the helicase and the primase) are functionally connected to the five terminal reactions of glycolysis [14].

In the rest of this section, we give a few examples of the many candidates that could mediate replication-phenotype coupling. The oligomeric DnaA protein, which plays a key role in the initiation of replication [17], also acts as a transcription factor [18]. Simulations of DnaA activity in initiation have been made based on the distribution of low affinity binding sites in the origin region and higher affinity sites elsewhere [12]. However, the situation is much more complex than in our model as presented above. Not only are there are several classes of binding sites but also, depending on the position of its binding sites relative to promoters, DnaA can act as an activator, a repressor or a terminator of transcription. For example, it activates transcription from *nrd* (ribonucleoside diphosphate reductase), *glpD* (aerobic glycerol-3-phosphate dehydrogenase) and *fliC* (flagellin) whilst it represses transcription from *mioC* (biotin synthase), *rpoH* (heat shock sigma factor), *uvrB* (DNA repair), *proS* (prolyl-tRNA synthetase) and *dnaA* itself.

The heat-stable nucleoid-structuring protein (H-NS) is present in around 14000 copies in exponentially growing *E. coli*. In addition to its role in the compaction of the nucleoid [10], H-NS binds specifically to around 250 loci to cause transcriptional repression including that of its own gene; this repression involves an association with RNA polymerase [27]. Another regulatory protein, the leucine-responsive regulatory protein (Lrp), interacts with H-NS to form higher order, repressive nucleoprotein structures involved in the repression of rRNA transcription [29]. The formation of this regulatory structure appears to be directly affected by environmental changes.

Finally, the LacI repressor and its binding sites probably constitute a repression hyperstructure [24] that could in principle behave in accordance with our model. In the absence of an inducer such as lactose (or in the presence of the preferred sugar, glucose), the *lac* operon is not transcribed. This is because some of the ten copies per cell of the tetrameric LacI repressor bind with their dimers to the operator *O1* and to two auxiliary operators, *O2* and *O3*, nearby on the DNA; this on-off binding (which is an equilibrium process) increases the local concentration of LacI at these operators if they are close enough and brings them closer still to increase further the local concentration of LacI at *O1* [20]. LacI binding to the operators is in competition with that of RNA polymerase to the promoter (since these sites overlap) and, importantly for our hypothesis, "there is some finite level of affinity of the protein for the "correct" site and some lower (but nonzero) and progressively decreasing affinity for other sites with decreasing degrees of homology with the correct one. To the extent that the great preponderance of wrong sites can compete with the regulatory target for protein and thus reduce the free protein concentration, the effective affinity of protein for the correct sites will also be reduced" [38].

5 Tests

The lactose operon in *E. coli* is perhaps the best understood of all operons. Its expression can be manipulated and copies of the entire operon, as well as copies of just the operators, can be inserted into different parts of the chromosome. To monitor differential expression of the operon, use might be made of fusions between B-galactosidase and the Green Fluorescent protein. In addition, it may prove possible to use the *tus* system to induce fork arrest in specific regions [2, 36] as well as other systems [11] although this must be done prudently since it can prove lethal [7]. Note too that the lac system itself might be used as a way of slowing replication [28]. Such experimental approaches might be combined with a simulation approach. Multi-agent systems, cellular automata and stochastic automata are promising approaches to simulating the diffusion and interaction of the often large numbers of enzymes and metabolites present in biological cells [16, 6, 33]. A stochastic automaton, HSIM, has been developed and an early version used to model the assembly of cytoskeletal filaments in a virtual cell [1]. HSIM therefore seems well-suited to the exploration of the parameter space needed to see what is required for differentiation to occur.

6 Corollaries

1. Local positive feedback and global negative regulation can act not only via either activation or repression of gene expression but also via translation and degradation. In all cases, the end result is a hyperstructure that tends to maintain its own existence.

2. There is an epigenetic flavour to our hypothesis. If a bacterium replicates its chromosome, this is usually because it is growing. If it is growing, this is because it has the hyperstructures needed for growing. Hence, the mother cell already has one copy of a needed if the genes in this hyperstructure are all on one strand, one of the future daughters inherits an established hyperstructure. This creates a status quo situation that is likely to affect the chance of the other daughter generating a sister hyperstructure.

3. Factors that affect the synthesis or degradation of the activator or repressor are clearly important in replication-phenotype coupling and include the position of the gene encoding the activator or repressor (whether it is before rest-stop A, or between rest-stops A and B, or after rest-stop B) and whether this gene is itself regulated by its own product. Given that the ratio of RNA polymerase to genes is a key parameter, these factors also include the temporal pattern of synthesis of transcriptionally active RNA polymerase and the spatial distribution of this enzyme (note that RNA polymerase itself is subject to local concentration effects).

4. The hypothesis is relevant to eukaryotic cells too even those in which DNA replication and transcription can only occur at different times. This is because the role of transcription in the bacterial case may be replaced by modifications to the chromosomes and associated macromolecules.

To test the feasibility of this hypothesis, we propose to use a stochastic automaton, HSIM, that we have previously used to investigate the formation of cytoskeleton filaments and multi-enzyme assemblies within a cell [1]. In essence, the project will entail modifying

HSIM to represent linear genes that can bind RNA polymerases and that can be duplicated. Both genes and polymerases will be diffusible. The effects on gene expression will be investigating by varying the parameters of pausing during duplication, varying the quantity of RNA polymerase, positive feedback weightings for cooperative RNA polymerase binding, and chromosome condensation.

7 Discussion

How do cells manage to produce not only reproducible phenotypes (out of the hyperastronomical number apparently available to them [15]) but also coherent phenotypes? How do cells negotiate the cell cycle? Are these questions linked? We have suggested the phenotype is not decided at the level of individual macromolecules but at a higher level that of assemblies of molecules alias hyperstructures [24]. We and others have also suggested that a primary function of the cell cycle is to generated coherent diversity of phenotypes within a population of cells [26, 31, 25]. For example, in the strand-specific segregation model it is the association of each parental strand with a particular set of hyperstructures and the continued association once replication has occurred that ensures the separation of the daughter chromosomes [30] This is in line with the correlation between the location of genes on the chromosome and their position of genes along the long axis of the cell [35, 22, 37] and with evidence showing a highly asymmetric pattern of segregation of markers around the terminus [39]. It has also been proposed that the very fact of having two chemically identical chromosomes in the same cytoplasm creates a symmetry-breaking situation if the genes are in a global competition with one another for transcription by RNA polymerase and if local positive feedback circuits can operate such that a gene that is being transcribed has a greater chance of being transcribed again than one that is silent [26].

This ”differentiation for free” model can be revisited in the light of the recent discovery of a relationship between metabolism and the enzymes involved in the elongation step in chromosome replication [14]. There is an intriguing possibility that this relationship could affect phenotype by altering the rate of replication in different regions in the chromosome. In principle, a variety of physical mechanisms could combine to give the local positive cooperation and global negative competition needed for differentiation to be determined by replication-phenotype coupling. A judicious marriage between simulation using programs such as HSIM [1] and bench experimentation should identify the parameter values needed for such differentiation to occur.

References

[1] Amar, P., G. Bernot & V. Norris, (2004) HSIM: a simulation programme to study large assemblies of proteins. *Journal of Biological Physics and Chemistry* **4**: 79-84.

[2] Andersen, P. A., A. A. Griffiths, I. G. Duggin & R. G. Wake, (2000) Functional specificity of the replication fork-arrest complexes of *Bacillus subtilis* and Escherichia coli: significant specificity for Tus-Ter functioning in E. coli. *Molecular microbiology* **36**: 1327-1335.

[3] Autret, S., A. Levine, F. Vannier, Y. Fujita & S. J. Seror, (1999) The replication checkpoint control in *Bacillus subtilis*: identification of a novel RTP-binding

sequence essential for the replication fork arrest after induction of the stringent response. *Molecular microbiology* **31**: 1665-1679.

[4] Avery, S. V., (2006) Microbial cell individuality and the underlying sources of heterogeneity. *Nature reviews* **4**: 577-587.

[5] Balaban, N. Q., J. Merrin, R. Chait, L. Kowalik & S. Leibler, (2004) Bacterial persistence as a phenotypic switch. *Science* **305**: 1622-1625.

[6] Ballet, P., A. Zemirline & L. Marcé, (2004) The BioDyn language and simulator. *Application to an immune response and E. coli and phage interaction. Journal of Biological Physics and Chemistry* **4**: 93-101.

[7] Bidnenko, V., R. Lestini & B. Michel, (2006) The Escherichia coli UvrD helicase is essential for Tus removal during recombination-dependent replication restart from Ter sites. *Molecular microbiology* **62**: 382-396.

[8] Booth, I. R., (2002) Stress and the single cell: intrapopulation diversity is a mechanism to ensure survival upon exposure to stress. *International Journal of Food Microbiology* **78**: 19-30.

[9] Breier, A. M., H. U. Weier & N. R. Cozzarelli, (2005) Independence of replisomes in Escherichia coli chromosomal replication. *Proceedings of the National Academy of Science U.S.A.* **102**: 3942-3947.

[10] Dame, R. T., C. Wyman & N. Goosen, (2000) H-NS mediated compaction of DNA visualised by atomic force microscopy. *Nucleic acids research* **28**: 3504-3510.

[11] Guijo, M. I., J. Patte, M. del Mar Campos, J. M. Louarn & J. E. Rebollo, (2001) Localized remodeling of the Escherichia coli chromosome: the patchwork of segments refractory and tolerant to inversion near the replication terminus. *Genetics* **157**: 1413-1423.

[12] Hansen, F. G., B. B. Christensen & T. Atlung, (1991) The initiator titration model: computer simulation of chromosome and minichromosome control. *Research in Microbiology* **142**: 161-167.

[13] Helmstetter, C. E., (1991) Timing of synthetic activities in the cell cycle. In: *Escherichia coli and Salmonella*. F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter & H. E. Umbarger (eds). Washington D.C.: American Society for Microbiology, pp. 1594-1605.

[14] Jannière L., D. Canceill, C. Suski, S. Kanga, B. Dalmais, R. Lestini, A.-F. Monnier, J. Chapuis, A. Bolotin, M. Titok, E. Le Chatelier and S. D. Ehrlich, (2007) Genetic link between glycolysis and DNA synthesis. *PLoS ONE accepted upon minor modifications*.

[15] Kauffman, S., (1993) The origins of order. *Oxford University Press*, Oxford.

[16] Le Sceller, L., C. Ripoll, M. Demarty, A. Cabin-Flaman, T. Nyström, M. Saier Jnr. & V. Norris, (2000) Modelling bacterial hyperstructures with cellular automata. *Interjournal of Complex Systems* **Paper 366**: **`http://www.interjournal.org`**.

[17] Leonard, A. C. & J. E. Grimwade, (2005) Building a bacterial orisome: emergence of new regulatory features for replication origin unwinding. *Molecular microbiology* **55**: 978-985.

[18] Messer, W. & C. Weigel, (1997) DnaA initiator–also a transcription factor. *Molecular microbiology* **24**: 1-6.

[19] Michelsen, O., M. J. Teixeira de Mattos, P. R. Jensen & F. G. Hansen, (2003) Precise determinations of C and D periods by flow cytometry in Escherichia coli K-12 and B/r. *Microbiology* **149**: 1001-1010.

[20] Müller-Hill, B., (1998) The function of auxiliary operators. *Molecular microbiology* **29**: 13-18.

[21] Mulugu, S., A. Potnis, Shamsuzzaman, J. Taylor, K. Alexander & D. Bastia, (2001) Mechanism of termination of DNA replication of Escherichia coli involves helicase-contrahelicase interaction. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 9569-9574.

[22] Niki, H., Y. Yamaichi & S. Hiraga, (2000) Dynamic organization of chromosomal DNA in Escherichia coli. *Genes and Development* **14**: 212223.

[23] Noirot-Gros, M. F., E. Dervyn, L. J. Wu, P. Mervelet, J. Errington, S. D. Ehrlich & P. Noirot, (2002) An expanded view of bacterial DNA replication. *Proceedings of the National Academy of Science U.S.A.* **99**: 8342-8347.

[24] Norris, V., A. Cabin-Flaman, T. den Blaauwen, R. H. Doi, R. Harshey, L. Jannière, A. Jimenez- Sanchez, D. J. Jin, P. A. Levin, E. Mileykovskaya, A. Minsky, M. Saier Jnr. & K. Skarstad, (2007) A functional taxonomy of bacterial hyperstructures. *Microbiology and Molecular Biology* **Reviews**: in press.

[25] Norris, V., M. Demarty, D. Raine, A. Cabin-Flaman & L. Le Sceller, (2002) Hypothesis: hyperstructures regulate initiation in Escherichia coli and other bacteria. *Biochimie* **84**: 341-347.

[26] Norris, V. & M. S. Madsen, (1995) Autocatalytic gene expression occurs via transertion and membrane domain formation and underlies differentiation in bacteria: a model. *Journal of molecular biology* **253**: 739-748.

[27] Oshima, T., S. Ishikawa, K. Kurokawa, H. Aiba & N. Ogasawara, (2006) Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Res* **13**: 141-153.

[28] Payne, B. T., I. C. van Knippenberg, H. Bell, S. R. Filipe, D. J. Sherratt & P. McGlynn, (2006) Replication fork blockage by transcription factor-DNA complexes in Escherichia coli. *Nucleic acids research* **34**: 5194-5202.

[29] Pul, U., R. Wurm & R. Wagner, (2007) The Role of LRP and H-NS in Transcription Regulation: Involvement of Synergism, Allostery and Macromolecular Crowding. *Journal of molecular biology* **366**: 900-915.

[30] Rocha, E., J. Fralick, G. Vediyappan, A. Danchin & V. Norris, (2003) A strand-specific model for chromosome segregation in bacteria. *Molecular microbiology* **49**: 895-903.

[31] Segre, D., D. Ben-Eli & D. Lancet, (2000) Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proceedings of the National Academy of Science U.S.A.* **97**: 4112-4117.

[32] Smits, W. K., O. P. Kuipers & J. W. Veening, (2006) Phenotypic variation in bacteria: the role of feedback regulation. *Nature reviews* **4**: 259-271.

[33] Soula, H., C. Robardet, F. Perrin, S. Gripon, G. Beslon & O. Gandrillon, (2005) Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a Multi-Agent-based Software. *Biomed Central Bioinformatics* **6**: 228.

[34] Stein, A. & W. Firshein, (2000) Probable identification of a membrane-associated repressor of *Bacillus subtilis* DNA replication as the E2 subunit of the pyruvate dehydrogenase complex. *Journal of bacteriology* **182**: 2119-2124.

[35] Teleman, A. A., P. L. Graumann, D. C. Lin, A. D. Grossman & R. Losick, (1998) Chromosome arrangement within a bacterium. *Current Biology* **8**: 11021109.

[36] Valjavec-Gratian, M., T. A. Henderson & T. M. Hill, (2005) Tus-mediated arrest of DNA replication in Escherichia coli is modulated by DNA supercoiling. *Molecular microbiology* **58**: 758-773.

[37] Viollier, P. H., M. Thanbichler, P. T. McGrath, L. West, M. Meewan, H. H. McAdams & L. Shapiro, (2004) Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proceedings of the National Academy of Science U.S.A.* **101**: 9257-9262.

[38] von Hippel, P. H. & O. G. Berg, (1986) On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 1608-1612.

[39] Wang, X., C. Possoz & D. J. Sherratt, (2005) Dancing around the divisome: asymmetric chromosome segregation in Escherichia coli. *Genes and Development* **19**: 2367-2377.

Hybolites: novel therapeutical tools based on stochastic automata

Vic Norris[1,3], Patrick Amar[2,3], Michel Thellier[1,3], Guillaume Legent[1,3], Armelle Cabin[1,3], Jean-Christophe Plaquevent[4], Milton Saier jnr.[5], Gilles Bernot[3], Gradimir Misevic[1,3] and Camille Ripoll[1,3]

[1] Laboratoire d'Assemblages moléculaires: modélisation et imagerie SIMS, FRE CNRS 2829, Faculté des Sciences de l'Université de Rouen, 76821 Mont Saint Aignan Cedex, France
[2] Laboratoire de Recherche en Informatique, Université Paris Sud & UMR CNRS 8623, F-91405 Orsay cedex, France
[3] Epigenomics Project, Genopole®, F-91000 Evry, France
[4] Laboratoire de Synthèse et Physico-Chimie de Molécules d'Intérêt Biologique, UMR CNRS 5068, Université Paul Sabatier, F-31062, Toulouse cedex 9, France
[5] Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116 USA

Abstract

Bacterial disease is a major cause of suffering and the scarcity of new molecules that can act on bacteria is a major problem. New strategies for developing such molecules might be based on recent concepts in microbiology. Hyperstructures are large assemblies of molecules and macromolecules that perform functions such as DNA replication and chemotaxis and the interactions between hyperstructures have been proposed to constitute an intermediate level of organisation in cells. An entirely new therapy for bacterial diseases might therefore be devised based on the manipulation of hyperstructures. One way to do this would be to supply cells with hybrid metabolites or *hybolites* made by a pairwise combination of the thousands of small molecules involved in metabolism. Some of these hybolites would be substrates for two, very different, hyperstructures and might do much more than simply inhibit key enzymes and processes within the hyperstructures: they might provoke the assembly of two hyperstructures in the same space or lead to hyperstructures emitting misleading signals. It is conceivable that hybolites might even convert a pathogenic Mr Hyde into an inoffensive Dr Jekyll. The likely action of candidate hybolites on hyperstructure dynamics and hence phenotype might be explored cheaply in silico using stochastic automata such as HSIM.

1 Introduction

In a Howard Hughes lecture given in 1999, Don Ganem rebukes those eminent colleagues who, in the sixties, made optimistic predictions about how infectious disease was conquered and how funding should be redirected to chronic diseases: "Now we know, of course, that that notion (that infectious disease was no longer a problem) was foolish to begin with, that infectious disease, epidemic infection, is a part of the human condition. I'm going to show you that it's really a part of human evolution that we can never get away from infectious disease as a class. We can triumph over individual infectious diseases, but the concept that we're going to be free of infection as a species is a ridiculous one and one that nobody believes anymore". Infectious disease, usually bacterial, causes over 60% of total deaths in the developing world, infectious disease is the third leading cause of death in Europe (the elderly and weak are most vulnerable) and, despite existing

antibiotic therapies and vaccines, infectious disease remains the leading cause of mortality and morbidity worldwide (for references see [17]). In particular, the problem of bacterial resistance to antibiotics (often due to plasmids that can transfer resistance from one bacterium to another) is far from solved as illustrated by the recent deaths in Israel due to multi-drug resistant *Klebsiella* (in **news.yahoo.com/070308/43/6czlb.html**).

So what are we going to do about infectious disease? Vicente *et al.* [17] point out that "the antibiotics that were easy to discover have already been found, and it is likely that the search for new members of existing classes, and certainly for new classes of antibiotics, will involve a substantial amount of high-quality, expensive and laborious research". Indeed, the cost of bringing a new drug to market is estimated to be more than 800 million Euros whilst "a paradox of the effectiveness of antibiotics is their weak value as marketable goods: patients stop buying them once their health returns, after relatively short courses of treatment" [17]. Apparently, therapeutic success is not guaranteed even when pathogenic bacteria are sensitive to an antibiotic. One reason is persistence – the fact that a bacterial population is phenotypically heterogeneous such that some bacteria are not growing at the time of the antibiotic treatment and therefore survive [2]. Another reason is the antibiotic itself may lead to a biofilm forming in which the bacteria are more resistant to antibiotics [7].

Exploitation of concepts developed in the field of integrative biology is one approach to tackling bacterial and other diseases. Here, we consider a new therapy for bacterial diseases based on the manipulation of hyperstructures *alias* large assemblies of different molecules and macromolecules that, in our hypothesis, perform particular functions within cells [11]. We propose that hyperstructures may be manipulated by supplying bacteria with hybrid metabolites or *hybolites* and, to decide which hybolites to construct, we propose modelling their putative action on hyperstructure dynamics via the stochastic automaton HSIM [1].

2 Hyperstructures as targets

The concept of a hyperstructure is that of an extended assembly of molecules and macromolecules with a specific function within bacterial cells [11]. Certain hyperstructures are functioning-dependent structures that only form when the constituents of the structure function – and function *together* [16]. Examples of candidate hyperstructures include the prokaryotic equivalent of the eukaryotic nucleolus for making ribosomes, the giant factory for DNA replication, the array of chemotaxis receptors for interpreting gradients of attractants and repellents, transertion structures produced by the coupled transcription/translation/insertion of nascent proteins into membrane (which include nascent flagella and the expressed *lac* operon for metabolising lactose), a competence hyperstructure responsible for DNA uptake, a phosphoenolpyruvate:sugar phosphotransferase system (PTS) responsible for the sensing and uptake of a large number of extracellular sugars and for feeding their products, cytoplasmic sugar phosphates, directly to glycolytic enzymes that may even be part of the same hyperstructure, the divisome that executes cell division, and possibly structures involved in pili formation and virulence.

3 Hybolites

Hyperstructures and the interactions between them have been proposed to constitute an intermediate level of organisation in cells [11]. Importantly for therapy, this intermediate level, in our hypothesis, determines the phenotype. An entirely new therapy for bacterial diseases might therefore be devised based on the manipulation of hyperstructures. One way to do this would be to supply cells with hybrid metabolites or *hybolites* made by a systematic, high-throughput, pair-wise combination of the thousands of small molecules involved in metabolism or cell structure. Some of these hybolites would be substrates for two, very different, hyperstructures and might do much more than simply inhibit key enzymes and processes within the hyperstructures: they might induce the assembly of a functioning-dependent hyperstructure in conditions when normally no such structure should form, or they might provoke the assembly of two hyperstructures in the same space, or they might lead to hyperstructures emitting misleading signals resulting in the bacterium adopting (or trying to adopt) patterns of growth and rates of growth inappropriate for the environment. The hyperstructure hypothesis has corollaries in which hyperstructure dynamics regulate the cell cycle events of chromosome initiation [10], chromosome partitioning [15] and cell division [12]. Hybolites that affect these dynamics might well result in alterations to the rate of progress through the cell cycle or even to the order of cell cycle events with correspondingly serious consequences for the viability or virulence. Finally, it is conceivable that hybolites might convert a pathogenic Mr Hyde into an inoffensive Dr Jekyll [13] and hence avoid the problem of a killing that selects resistant mutants that survive or creates empty niches to be filled with other bacteria.

4 Feasability

There are several reports of hybrid molecules that might be used as hybolites. Different phospholipids, with saturated and unsaturated chains, have been linked to a cortisol derivative, novobiocin has been coupled through the 3' or 2" hydroxyl group and a linker to dioleoylphosphatidic acid, and functionalised lipids have been made with, as head groups, the DNA bases thymidine or adenosine [14, 8, 9]. In addition to a chemical synthesis of hybolites, it may prove possible to develop a biological system to produce them. For example, catalytic antibodies have been obtained for a wide variety of reactions including ester hydrolysis and transesterification, amide hydrolysis, glycosidic bond hydrolysis, and decarboxylation [18, 5].

5 Targets

In this section we give a few examples of the types of hybolite that might be constructed to manipulate particular hyperstructures.

1. *Sensing* Bacteria use quorum sensing molecules such as homoserine lactone to calculate the density of the population and to either continue growing or stop growing. Hybolites of such molecules linked to other molecules might result in bacteria either continuing to try to grow at high densities or stopping growing at low densities. Bacteria also sense attractants so as to swim up nutritional gradients or repellents so as to swim down gradients. This sensing involves a chemotaxis hyperstructure and hybolites to alter it might be constructed by fusing attractants and repellents.

2. *Transport and metabolism*. Sugars, bases, amino acids and many other small molecules can be imported by bacteria and used in metabolism. Fusing, for example, a sugar such as glucose to an amino acid, might interfere with either the membrane-bound enzyme $IIBC^{glc}$ or the cytoplasmic enzyme IIA^{glc} which are responsible for importing and phosphorylating glucose.

3. *DNA synthesis*. Ongoing replication requires feeding the hyperstructure with the four deoxyribonucleotides (dNTPs) at the rate of about 3000 nucleotides per second. Replication might then be readily perturbed by hybolites that involve fusions between nucleotides and other molecules. The involvement of the membrane in replication has a long history and it is tempting to speculate that the replication hyperstructure may also contain cardiolipin (for references see [4, 11]). Hence, it might be worth making hybolites from phospholipids such as cardiolipin combined with DNA precursors.

4. *Protein synthesis*. A ribosomal or 'nucleolar' hyperstructure forms in *Escherichia coli* at high growth rates when the synthesis of ribosomes consumes most of the bacterium's resources but is not apparent at lower growth rates [3]. Hybolites might modulate growth rate if they were to interfere with the assembly of this hyperstructure. Such hybolites might be made from the amino acids that constitute ribosomal proteins along with bases such as uracil that are form part of ribosomal RNA.

5. *Cell division*. The synthesis of phospholipids and of peptidoglycan leading to the invagination of the membrane and the cell wall must be coordinated between themselves and with ongoing chromosome replication and segregation. This coordination may be provided by a division hyperstructure [11]. Hence promising hybolites might be made from combinations of phospholipids and precursors of peptidoglycan and DNA

6 Discussion

Over the last decade, a wealth of experimental data on the existence of large intracellular structures or hyperstructures has led to a new but still speculative view of bacteria in which the dynamics of these hyperstructures determines the phenotype [11]. This view leads to the idea that manipulating hyperstructures should result in changes in phenotype and hence to the idea that molecules might be made to cause such changes. Indeed, such molecules might actually be made to prevent bacterial virulence. Here, we have explored briefly the notion of hybrid molecules or hybolites which comprise two molecules that participate in different hyperstructures.

A high throughput generation and testing of hybolites on bacterial pathogens might constitute an attractive strategy for the pharmaceutical industry. There are, however, tens of millions of potential hybolites that might be made and screened. Unfortunately, it is not clear which of them might be effective in an eventual therapy. A complementary rather than alternative strategy would be to develop *in silico* approaches based on stochastic automata such as HSIM [1]. HSIM is being used at present to model the dynamics of hyperstructures such as those involved in glucose transport and metabolism. Addition of virtual hybolites with different characteristics to HSIM might be used to help select those that are worth testing *in vivo*.

Hybolites have been proposed above as a possible panacea to bacterial diseases. Of course, it could be argued that there are eukaryotic equivalents to bacterial hyperstructures (see for example [6]). If so, and if hybolites really do alter bacterial hyperstructures to affect phenotypes, they may have similar actions on eukaryotic cells. In which case, hybolites might be of value in the treatment of certain chronic diseases.

Acknowledgements

For support, we thank the ACI *complexité du vivant* and the Epigenomics Project.

References

[1] Amar, P., G. Bernot & V. Norris, (2004) HSIM: a simulation programme to study large assemblies of proteins. *Journal of Biological Physics and Chemistry* **4**: 79-84.

[2] Balaban, N. Q., J. Merrin, R. Chait, L. Kowalik & S. Leibler, (2004) Bacterial persistence as a phenotypic switch. *Science* **305**: 1622-1625.

[3] Cabrera, J. E. & D. J. Jin, (2003) The distribution of RNA polymerase in *Escherichia col*i is dynamic and sensitive to environmental cues. *Molecular microbiology* **50**: 1493-1505.

[4] Firshein, W., (1989) Role of the DNA/membrane complex in prokaryotic DNA replication. *Annual Review of Microbiology* **43**: 89-120.

[5] Hanson, C. V., Y. Nishiyama & S. Paul, (2005) Catalytic antibodies and their applications. *Current opinion in biotechnology* **16**: 631-636.

[6] Hartwell, L. H., J. J. Hopfield, S. Leibler & A. W. Murray, (1999) From molecular to modular cell biology. *Nature* **402(6761 Suppl)**: C47-52.

[7] Hoffman, L. R., D. A. D'Argenio, M. J. MacCoss, Z. Zhang, R. A. Jones & S. I. Miller, (2005) Aminoglycoside antibiotics induce bacterial biofilm formation. *Nature* **436**: 1171-1175.

[8] Lebeau, L., C. Mioskowski & P. Oudet, (1988) Systematic study of phospholipids linked to a steroid derivative, spread into a monolayer at the air/water interface. *Biochimica et biophysica acta* **939**: 417-429.

[9] Lebeau, L., E. Regnier, P. Schultz, J. C. Wang, C. Mioskowski & P. Oudet, (1990) Two-dimensional crystallization of DNA gyrase B subunit on specifically designed lipid monolayers. *FEBS letters* **267**: 38-42.

[10] Norris, V., M. Demarty, D. Raine, A. Cabin-Flaman & L. Le Sceller, (2002) Hypothesis: hyperstructures regulate initiation in *Escherichia coli* and other bacteria. *Biochimie* **84**: 341-347.

[11] Norris, V., T. den Blaauwen, A. Cabin-Flaman, R. H. Doi, R. Harshey, L. Janniere, A. Jimenez-Sanchez, D. J. Jin, P. A. Levin, E. Mileykovskaya, A. Minsky, M. Saier, Jr. & K. Skarstad, (2007) Functional taxonomy of bacterial hyperstructures. *Microbiol Mol Biol Rev* **71**: 230-253.

[12] Norris, V. & I. Fishov, (2001) Hypothesis: membrane domains and hyperstructures control bacterial division. *Biochimie* **83**: 91-98.

[13] Norris, V., P. Freestone, S. Grant, I. Toth, M. Trinei, L. Le Sceller, T. Nyström, A. Oshima & T. Onoda, (1999) Tyrosine phosphorylation in *E. coli* has implications for the manipulation of intra-cellular signalling and structure in anti-bacterial therapies: the Jekyll and Hyde approach. *Emerging Therapeutic Targets* **3**: 89-92.

[14] Pincet, F., E. Perez, G. Bryant, L. Lebeau & C. Mioskowski, (1994) Long-range attraction between nucleosides with short-range specificity: Direct measurements. *Physical Review Letters* **73**: 2780-2783.

[15] Rocha, E., J. Fralick, G. Vediyappan, A. Danchin & V. Norris, (2003) A strand-specific model for chromosome segregation in bacteria. *Molecular microbiology* **49**: 895-903.

[16] Thellier, M., G. Legent, P. Amar, V. Norris & C. Ripoll, (2006) Steady-state kinetic behaviour of functioning-dependent structures. *The FEBS journal* **273**: 4287-4299.

[17] Vicente, M., J. Hodgson, O. Massidda, T. Tonjum, B. Henriques-Normark & E. Z. Ron, (2006) The fallacies of hope: will we discover new antibiotics to combat pathogenic bacteria in time? *FEMS microbiology reviews* **30**: 841-852.

[18] Xu, Y., N. Yamamoto & K. D. Janda, (2004) Catalytic antibodies: hapten design strategies and screening methods. *Bioorganic & medicinal chemistry* **12**: 5247-5268.

Could phase oscillations occur in bacteria?

Vic Norris[1,2]

[1] UMR CNRS 6522, Department of Science, Univ. of Rouen, F-76821 Mont Saint Aignan, France
[2] Epigenomics Project, Genopole®, F-91000 Evry, France

Abstract

Phase oscillations or phoscillations in one, two, three and more dimensions may exist in a wide variety of systems. In the case of biological systems such as bacteria, phoscillations might take the form of conformations travelling along DNA or protein filaments, density fluctuations travelling in the plane of the membrane, and changes in water structure travelling through the cytoplasm. Phoscillations in different dimensions might become coupled and thereby constitute a mechanism to generate an integrated and coherent phenotype.

1 Introduction

In June 2005 on a visit to the University of Brest funded by the Epigenomics Project, Pascal Ballet asked me which simple model system he might use to illustrate his 2-D, multi-agent simulator, BioDyn [1]. I had been working with physical chemists in Rouen on an *in vitro* Langmuir system to investigate how the formation of domains or patches of particular phospholipids within a planar membrane might instruct proteins when and where to act to perform division. I therefore suggested that he might simulate certain aspects of such systems and that, in a monolayer composed of a single species of phospholipid, he might simulate the formation of domains of ordered phospholipids in a surrounding sea of phospholipids in a disordered state. The idea was to represent a phospholipid with its fatty acid chains parallel to one another (i.e. ordered) as a blue circle occupying a unit area and a phospholipid with its chains splayed out (i.e. disordered) as a red circle occupying a larger area. The state of the phospholipid would depend on the frequency of collisions with other phospholipids. In Pascal's initial simulation, the phospholipids were relatively sparse, collisions were infrequent and the majority were red. A Langmuir-Blodgett trough has a bar to compress the phospholipid layer and at certain values of compression domains can form. We therefore decided to add a bar to the simulation and to compress the phospholipids. As the bar was moved into the phospholipids, a band of blue circles immediately formed and, at certain concentrations of phospholipids, this band could traverse the entire population. This was intriguing because the movement of the blue band corresponded to a phase change moving through the rather crowded population rather than to the diffusion of the individual phospholipids themselves. It seemed to me that if the bar itself were to oscillate a resonant frequency might be found such that waves of phospholipids in the ordered state might be generated. This raised the possibility that phase oscillations, or more exactly migrating bands of a different density, might occur in the cytoplasmic membrane of bacteria *in vivo*. It raised the further possibility that fluctuations in density might occur in any dimension in both animate and inanimate systems. Moreover, such multi-dimensional fluctuations might be coupled. Unfortunately, Pascal did not have time to work further on the simulation.

I therefore contacted Axel Hunding, who had modelled the oscillations of a division protein in bacteria and who had participated as a physical chemist in a think-tank on the origins of life organised by the Epigenomics Project. Not long after, Axel told me of a

model of domain formation in a membrane based on the diffusion of GTPases coupled with the attraction and adhesion among active GTPases [7]; the authors, who belonged to the same European network as Axel, argue that the structures generated by their model are not Turing structures [51]. It turns out that similar mechanisms were proposed a couple of years earlier to explain how waves of MinD-membrane density might be involved in cell division in *Escherichia coli* [24]. Another paper was published in 2005 in which domain formation in membrane was modelled based on lipids, GMC proteins, and attraction between acidic lipids and GMC proteins [21]; phase separation occurs as a result of free energy minimisation whilst driving the system away from equilibrium via enzyme activity results in waves of domains and, in this oscillatory system, an important role is given to reaction-diffusion as proposed by Turing and others [51, 15]. Finally, the concept of conformational spreading between proteins in one, two and three dimensional lattices has been proposed as a mechanism of general importance although the potential oscillatory aspect was not developed [6]. These interesting papers might be regarded as making a subsequent, related proposal pointless. Here, I argue that they should be regarded instead as the first (or at least the most recent [8]) in what I hope will be a long reflection on phase oscillations. A new field may be emerging. Its objects are what we may term *phoscillons* which are found in *phoscillatory* systems of any dimension. Exploration of this field, which may have an importance that is not limited to biology, will require collaborations across the disciplines. To illustrate this potential importance, I consider in what follows the possibility that bacteria are phoscillatory systems.

2 The hypothesis

Phase oscillation – phoscillatory – systems exist in biology and elsewhere that undergo phase oscillations in 1, 2, 3 and higher dimensions. In the case of cells, phoscillations in different dimensions may be coupled to provide signals that initiate cell cycle events and to generate coherent phenotypes.

3 The physical candidates

1D. Possible changes that could travel along DNA include solitons (a soliton is a self-reinforcing solitary wave) [44], B to A transitions (and other conformational changes) accompanied by ion condensation/decondensation [28] and changes in water structure [9]. (I leave it to the physical chemists to decide whether any of these '1D' changes can actually be classed as *phase* changes. I would argue though that they can be accompanied by a local change in density.) In bacteria, the energy needed to maintain phoscillations could come from the negative supercoiling created by DNA gyrase [17] or from the supercoiling created by RNA polymerases during transcription [27]. Phoscillations might also occur along the cytoskeletal filaments found in both prokaryotes and eukaryotes (for a review see [37]. These filaments correspond to the three cytoskeletal classes in eukaryotes, tubulin, actin and intermediate filaments as well as to other filaments such as those formed by EF-Tu. The FtsZ protein has a structural homology to tubulin [26] and, *in vitro*, forms a wide variety of polymeric structures depending on the presence and concentrations of lipids, divalent ions and GTP [16]. FtsZ is present in *E. coli* as helices (this is in addition to its presence in the 'ring' at the division site, see below) that have a dynamic activity on the scale of seconds along with slower oscillations of a minute or so (Thanedar & Margolin, 2004). In some bacteria, the Min system prevents aberrant divisions (Levin *et al.*, 1998, Norris *et al.*, 2004, Teather *et al.*, 1974) and in *E. coli* the MinD protein, which

binds cooperatively to membranes in the presence of ATP, forms a long, helical polymer that assembles from the pole and that is disassembled by MinE so generating oscillations (Shih *et al.*, 2003). Eukaryotic actin is proposed to undergo a conformational change following the binding of gelsolin and this may propagate throughout the entire actin filament (for references see [6]). Could this occur with bacterial actin equivalents such as MreB? It should be noted here that there is a long history of solitons and conformons in proteins (Ji, 1991, Sinkala, 2006). Flagella offer a fine example of the important of the propagation of conformations. Flagellin exists in two states and when the motor reverses some L-type flagellin subunits are converted into the R-type producing transient segments of a relatively tight right-handed helix that are proposed to propagate out toward the flagellar tip (for references see [6]).

2D. Conformational spreading between proteins in one, two and three dimensional lattices has been proposed as a mechanism to explain a multitude of phenomena including chemotaxis, which involves a 2D array of chemoreceptors in the pole of *E. coli* [6]. Phospholipids are distributed heterogeneously in the cytoplasmic membrane of *E. coli* and other bacteria (for reviews see (Mileykovskaya & Dowhan, 2005, Matsumoto *et al.*, 2006)). At least two types of lipid domains of different order and polarity are present in the *E. coli* membrane as revealed by studies using laurdan and 1,3-diphenyl-1,3,5-hexatriene (Vanounou *et al.*, 2002); moreover, pyrene-labeled analogues of phosphatidylethanolamine and phosphatidylglycerol are sequestered into separate pre-existing domains in the membrane (Vanounou *et al.*, 2003) but see (Nishibori *et al.*, 2005). The fluorescence of the hydrophobic dye FM 4-64 differs at the future septum from elsewhere at a very early stage in chromosome segregation (Fishov & Woldringh, 1999). The principal lipid constituents of these domains (which are unlikely to be pure (Galli Marxer *et al.*, 2005)) have been identified. Domains enriched in cardiolipin have been revealed at the centre and the poles of *E. coli* using 10-N-nonyl acridine orange (Mileykovskaya & Dowhan, 2000), findings that are reinforced by an elevated cardiolipin level in minicells which reflects that of isolated poles (Koppelman *et al.*, 2001). The cytoplasmic membrane of *Bacillus subtilis* also contains cardiolipin and phosphatidylethanolamine domains (Nishibori et al., 2005, Matsumoto *et al.*, 2006).

What creates these lipid domains? There are probably many factors including those responsible for lipid-lipid and lipid-protein interactions (Matsumoto et al., 2006). It should be noted that proteins are also distributed heterogeneously in the cytoplasmic membrane and the lipid domains are probably proteo-lipid domains (for references see (Norris, 1992, Norris et al., 2007)). The process of transertion (*alias* the coupled transcription, translation and insertion of proteins into and through membranes) has been proposed to create membrane domains (Norris & Madsen, 1995, Woldringh, 2002); the importance of this role for transertion is supported by experimental evidence that shows an increase in the fluidity of the membrane when transcription and translation are abolished (Binenbaum *et al.*, 1999). This means that the energy-consuming process of transertion is structuring the cytoplasmic membrane and the system is away from equilibrium. In addition, lipids are being inserted into the membrane, probably at the sites in the division septum where many of the phospholipid synthases are located (Matsumoto et al., 2006). Hence during some or all its cell cycle, the conditions may be met for the membrane of bacteria to be swept by waves of a particular physical state as perhaps foreshadowed in a seminal papers [8].

3D. The eukaryotic cytosol undergoes repeated gel/sol transitions as actin polymerises and depolymerises. Waves of NADP(H) traverse neutrophils [41] whilst waves of cal-

cium have been observed in many cell types. The cytoskeleton has been considered a tensegrity structure that may itself vibrate [19] and oscillations of the cell wall have been observed in *Saccharomyces cerevisiae* that have an frequency of around a kHz and an amplitude of 3 nm and that depend on ongoing metabolism and probably the coupled mechanical activity of many proteins [40]. What then of the 'simple' bacterial cytoplasm? Bacterial cells contain many large, spatially extended, assemblies of ions, molecules and macromolecules, which we have termed 'hyperstructures' [37], which are implicated in functions that range from DNA replication and cell division to chemotaxis and secretion. These hyperstructures include cytoskeletal filaments with significant similarities to microtubules, microfilaments and intermediate filaments as well as other types of filaments (see above). Could hyperstructures generate oscillations in density that would be propagated through the cytoplasm - and what might be their physical nature? Water is a seductive candidate. Some, if not all, of the water within the cell is structured by the molecules and macromolecules that constitute the cell (for references see [9]). One of the hypotheses in this controversial area is that of two-state water [10, 32, 33] which is proposed to consist of coexisting microdomains of water molecules of different densities and hydrogen bond strengths. Low density water has a density of 0.91 g/ml; whilst high density water has a density of 1.18 g/ml. The two types of microdomains, with different hydrogen-bonded structures would also differ in all their physical and chemical properties: melting points, boiling points, and solvent properties. These microdomains are in a rapidly-exchanging equilibrium and water at surfaces can be enriched in either type. It is also reported that water structures can be surprisingly stable [45] and the implications for biology are considerable [54]. The question we return to below is whether phoscillations depend on water structures.

4D. Phoscillations might occur, in principle, in more than three dimensions (plus time). This may only be relevant to subjects such as astrophysics and I mention it here for completeness and fun. How would a 4D oscillation appear to a human observer of a bacterium? The limiting case of a phoscillon in a system of 1, 2 and 3 dimensions is a point, a line, and a surface, respectively, so in 4 dimensions it should be a volume. The observer should therefore see a whole subvolume of the bacterium undergoing a simultaneous phase change and then another subvolume undergoing a phase change with no necessarily evident pattern.

4 Coupling between the dimensions

The coupling of oscillations can lead to their synchronisation as observed by Huygens in 1657 [48]. Bray and Duke argue for such coupling in the case of conformational spreading through protein lattices and use the movements of myosin molecules in a muscle sarcomere as an example [6]. Here, I propose that phoscillations become coupled in *all* three dimensions and that these involve *all* cellular constituents. Changes in phase resulting in changes in local density are generally likely to involve water and water itself – and water structures along with ions that can mediate them –are perhaps the best candidates for coupling phoscillations (see above). In 1D, fluctuations of ion condensation/decondensation from one side to the other of DNA [28] and consequent changes in water structure could accompany conformational changes. Much bacterial DNA is in a liquid crystalline state with the fibres lying roughly parallel to one another so phoscillations in neighbouring fibres might become synchronised [43, 4]. In 2D, water close to the membrane is generally believed to be structured – even if this may not extend far [2]. Phoscillations that occur in

the membrane must affect water in the adjacent cytoplasm. Finally, in 3D, transitions between gel and sol states entail polymer-water interactions giving way to polymer-polymer interactions and vice versa as the result, it has been proposed of divalent cation binding to anionic sites on polymers to bring them together and so exclude water [42]. Such ideas have been taken further following evidence that a surprisingly large number of proteins spend some of their time in the cytoplasm in a flexible and unstructured state - indeed, it has been speculated that such unfolding might lead to local regions becoming gel-like and even propagating regions of disorder [5]. As a specific example of coupling via water, consider the following speculation. Enzymes can undergo specific conformational changes on the order of a millisecond that are important in catalysis [11, 18] and the regular conformational changes of identical, neighbouring, enzymes as they perform their function might be coupled by movements of water structure between the enzymes so that these enzymes move together in a synchronous, low energy state. It may be worth considering how, at other frequencies, coupling via water dynamics might lead to a synchronization of ribosomes, which undergo major conformational changes as they translate mRNA and which can be packed very closely.

5 Discussion

Turing showed that chemical reactions between two or more species of which at least one diffuses are sufficient to form domains given that the kinetic terms and diffusion constants obey well defined relations [51] and Gierer and Meinhardt developed a closely related theory focusing on local activation and lateral inhibition [15]. The phoscillatory structures discussed here are, it seems to me, not Turing structures. It is a phase or state that is propagated rather than a molecular species and there is no chemical reaction.

What problems, if any, might phoscillation solve? There is a hunger for integrative principles to help explain how cells negotiate the apparent enormity of phenotype space confronting them [22]. Such principles might be found in conformational spreading along protein lattices [6] or travelling waves in membranes [21] (for examples of other candidates see [37]). Exploration of a role for phoscillation as an integrative principle might prove particularly rewarding in the context of the hyperstructures that we propose constitute an intermediate level of organisation in cells [37]. Despite decades of intense investigation, the nature of the signals that regulate the bacterial cell cycle is unknown. Oscillations have been implicated in this regulation [12], as have hyperstructures [36]. It is tempting to speculate that different hyperstructures with different compositions might act as independent phoscillating systems that could sometimes become coupled to constitute cell cycle signals.

Acknowledgements

For discussions, I thank Pascal Ballet, Patrick Amar, Axel Hunding, Camille Ripoll, Michel Thellier, Eugenia Mileykovskaya, Jean-Claude Vincent, Jean-Marc Valleton and Stephane Alexandre and for funding I thank the Epigenomics Project.

References

[1] Ballet, P., A. Zemirline & L. Marcé, (2004) The BioDyn language and simulator. Application to an immune response and *E. coli* and phage interaction. *Journal of*

Biological Physics and Chemistry **4**: 93-101.

[2] Berkowitz, M. L., D. L. Bostick & S. Pandit, (2006) Aqueous solutions next to phospholipid membrane surfaces: insights from simulations. *Chemical Reviews*: DOI: 10.1021/cr0403638.

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

[4] Bouligand, Y. & V. Norris, (2001) Chromosome separation and segregation in dinoflagellates and bacteria may depend on liquid crystalline states. *Biochimie* **83**: 187-192.

[5] Bray, D., (2005) Flexible peptides and cytoplasmic gels. *Genome biology* **6**: 106.

[6] Bray, D. & T. Duke, (2004) Conformational spread: the propagation of allosteric states in large multiprotein complexes. *Annual Review of Biophysical and Biomolecular Structure* **33**: 53-73.

[7] Brusch, L., P. Del Conte-Zerial, Y. Kalaidzidis, J. Rink, B. Habermann, M. Zerial & A. Deutsch, (2006) Protein Domains of GTPases on Membranes: Do They Rely on Turings Mechanism? In: Mathematical Methods and Models in Bioscience. G. de Vries, L. Brusch, A. Deutsch, H. Byrne & H. Herzel (eds). Boston: Birkhäuser, pp. 35-48.

[8] Changeux, J. P., J. Thiery, Y. Tung & C. Kittel, (1967) On the Cooperativity of Biological Membranes. *Proceedings of the National Academy of Sciences of the United States of America* **57**: 335-341.

[9] Chaplin, M., (2006) Do we underestimate the importance of water in cell biology? *Nat Rev Mol Cell Biol* **7**: 861-866.

[10] Cho, H. C., S. Singh & G. W. Robinson, (1997) Understanding all of water's anomalies with a non-local potential. *Journal of Chemical Physics* **107**: 7979-7988.

[11] Eisenmesser, E. Z., O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay & D. Kern, (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**: 117–121.

[12] Fishov, I., (1994) Do oscillations control the bacterial cell cycle? In: What is controlling Life? E. Gnaiger, F. N. Gellerich & M. Wyss (eds). Innsbruck, Austria: Innsbruck University Press, pp. 221-225.

[13] Fishov, I. & C. Woldringh, (1999) Visualization of membrane domains in *Escherichia coli*. *Molecular microbiology* **32**: 1166-1172.

[14] Galli Marxer, C., M. L. Kraft, P. K. Weber, I. D. Hutcheon & S. G. Boxer, (2005) Supported membrane composition analysis by secondary ion mass spectrometry with high lateral resolution. *Biophysical journal* **88**: 2965-2975.

[15] Gierer, A. & H. Meinhardt, (1972) A theory of biological pattern formation. *Kybernetik* **12**: 30-39.

[16] Gonzalez, J. M., M. Velez, M. Jimenez, C. Alfonso, P. Schuck, J. Mingorance, M. Vicente, A. P. Minton & G. Rivas, (2005) Cooperative behavior of *Escherichia coli* cell-division protein FtsZ assembly involves the preferential cyclization of long single-stranded fibrils. *Proceedings of the National Academy of Science U.S.A.* **102**: 1895-1900.

[17] Gore, J., Z. Bryant, M. D. Stone, M. Nollmann, N. R. Cozzarelli & C. Bustamante, (2006) Mechanochemical analysis of DNA gyrase using rotor bead tracking. *Nature* **439**: 100-104.

[18] Huang, Y. J. & G. T. Montelione, (2005) Proteins flex to function. *Nature* **438**: 36-37.

[19] Ingber, D. E., (1998) The architecture of life. *Scientific American* **278 January**: 30-39.

[20] Ji, S. I., (1991) "Biocybernetics": a machine theory of biology. In: Molecular theories of cell life and death. S. I. Ji (ed). New Brunswick: Rutgers University Press, pp. 2-237.

[21] John, K. & M. Bar, (2005) Travelling lipid domains in a dynamic model for protein-induced pattern formation in biomembranes. *Physical biology* **2**: 123-132.

[22] Kauffman, S., (1993) *The origins of order.* Oxford University Press, Oxford.

[23] Koppelman, C.-M., T. Den Blaauwen, M. C. Duursma, R. M. A. Heeren & N. Nanninga, (2001) *Escherichia coli* minicell membranes are enriched in cardiolipin. *Journal of bacteriology* **183**: 6144-6147.

[24] Kruse, K., (2002) A dynamic model for determining the middle of Escherichia coli. *Biophysical journal* **82**: 618-627.

[25] Levin, P. A., J. J. Shim & A. D. Grossman, (1998) Effect of minCD on FtsZ ring position and polar septation in *Bacillus subtilis*. *Journal of bacteriology* **180**: 6048-6051.

[26] Lowe, J. & L. A. Amos, (1998) Crystal structure of the bacterial cell-division protein FtsZ. *Nature* **391**: 203-206.

[27] Lynch, A. S. & J. C. Wang, (1993) Anchoring of DNA to the bacterial cytoplasmic membrane through co-transcriptional synthesis of polypeptides encoding membrane proteins or proteins for export : a mechanism of plasmid hypernegative supercoiling in mutants deficient in DNA topoisomerase I. *Journal of bacteriology* **175**: 1645-1655.

[28] Manning, G. S., (2006) The contribution of transient counterion imbalances to DNA bending fluctuations. *Biophysical journal* **90**: 3208-3215.

[29] Matsumoto, K., J. Kusaka, A. Nishibori & H. Hara, (2006) Lipid domains in bacterial membranes. *Molecular microbiology* **61**: 1110-1117.

[30] Mileykovskaya, E. & W. Dowhan, (2000) Visualization of phospholipid domains in *Escherichia coli* by using the cardiolipin-specific fluorescent dye 10-N-nonyl acridine orange. *Journal of bacteriology* **182**: 1172-1175.

[31] Mileykovskaya, E. & W. Dowhan, (2005) Role of membrane lipids in bacterial division-site selection. *Current Opinion in Microbiology* **8**: 135-142.

[32] Mishima, O. & H. E. Stanley, (1998a) Decompression-induced melting of ice IV and the liquid-liquid transition in water. *Nature* **392**: 164-168.

[33] Mishima, O. & H. E. Stanley, (1998b) The relationship between liquid, supercooled and glassy water. *Nature* **396**: 329-335.

[34] Nishibori, A., J. Kusaka, H. Hara, M. Umeda & K. Matsumoto, (2005) Phosphatidylethanolamine domains and localization of phospholipid synthases in *Bacillus subtilis* membranes. *Journal of bacteriology* **187**: 2163–2174.

[35] Norris, V., (1992) Phospholipid domains determine the spatial organization of the *Escherichia coli* cell cycle : the membrane tectonics model. *Journal of theoretical biology* **154**: 91-107.

[36] Norris, V., M. Demarty, D. Raine, A. Cabin-Flaman & L. Le Sceller, (2002) Hypothesis: hyperstructures regulate initiation in *Escherichia coli* and other bacteria. *Biochimie* **84**: 341-347.

[37] Norris, V., T. den Blaauwen, A. Cabin-Flaman, R. H. Doi, R. Harshey, L. Janniere, A. Jimenez-Sanchez, D. J. Jin, P. A. Levin, E. Mileykovskaya, A. Minsky, M. Saier, Jr. & K. Skarstad, (2007) Functional taxonomy of bacterial hyperstructures. *Microbiol Mol Biol Rev* **71**: 230-253.

[38] Norris, V. & M. S. Madsen, (1995) Autocatalytic gene expression occurs *via* transertion and membrane domain formation and underlies differentiation in bacteria: a model. *Journal of molecular biology* **253**: 739-748.

[39] Norris, V., C. Woldringh & E. Mileykovskaya, (2004) A hypothesis to explain division site selection in *Escherichia coli* by combining nucleoid occlusion and Min. *FEBS letters* **561**: 3-10.

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

[41] Petty, H. R., R. G. Worth & A. L. Kindzelskii, (2000) Imaging Sustained Dissipative Patterns in the Metabolism of Individual Living Cells. *Physical Review Letters* **84**: 2754-2757.

[42] Pollack, G. H., (2001) *Cells, gels and the engines of life: a new, unifying approach to cell function.* Ebner and Sons, Seattle, Washington.

[43] Reich, Z., E. J. Wachtel & A. Minsky, (1994) Liquid-crystalline mesophases of plasmid DNA in bacteria. *Science* **264**: 1460-1463.

[44] Scott, A. C., (1985) Soliton oscillations in DNA. *Physical Review. A* **31**: 3518-3519.

[45] Sedlak, M., (2006) Large-scale supramolecular structure in solutions of low molar mass compounds and mixtures of liquids: II. Kinetics of the formation and long-time stability. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys* **110**: 4339-4345.

[46] Shih, Y.-L., T. Le & L. Rothfield, (2003) Division site selection in *Escherichia coli* involves dynamic redistribution of Min proteins within coiled structures that extend between the two cell poles. *Proceedings of the National Academy of Science U.S.A.* **100**: 7865-7870.

[47] Sinkala, Z., (2006) Soliton/exciton transport in proteins. *Journal of theoretical biology* **241**: 919-927.

[48] Strogatz, S. H. & I. Stewart, (1993) Coupled oscillators and biological synchronization. *Scientific American* **269**: 102-109.

[49] Teather, R. M., J. F. Collins & W. D. Donachie, (1974) Quantal behavior of a diffusible factor which initiates septum formation at potential division sites in *E. coli*. *Journal of bacteriology* **118**: 407-413.

[50] Thanedar, S. & W. Margolin, (2004) FtsZ exhibits rapid movement and oscillation waves in helix-like patterns in *Escherichia coli*. *Current Biology* **14**: 1167-1173.

[51] Turing, A. M., (1952) The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* **237**: 37-72.

[52] Vanounou, S., A. H. Parola & I. Fishov, (2003) Phosphatidylethanolamine and phosphatidylglycerol are segregated into different domains in bacterial membrane. A study with pyrene-labelled phospholipids. *Molecular microbiology* **49**: 1067-1079.

[53] Vanounou, S., D. Pines, E. Pines, A. H. Parola & I. Fishov, (2002) Coexistence of domains with distinct order and polarity in fluid bacterial membranes. *Photochemistry and Photobiology* **76**: 1-11.

[54] Wiggins, P. M., (2002) Enzyme reactions and two-state water. *Journal of Biological Physics and Chemistry* **2**: 25-37.

[55] Woldringh, C. L., (2002) The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Molecular microbiology* **45**: 17-29.

Nonseparability, Multiple Interactions and Hypercomplexity

Abdallah Zemirline[1], Vic Norris[2]

[1] LISyC, EA3883, Département Informatique, UFR Sciences et Techniques, Université de Bretagne Occidentale, C.S. 93837 BP 809, F-29238 Brest Cedex 3, France.
[2] Laboratoire d'Assemblages moléculaires: modélisation et imagerie SIMS, FRE CNRS 2829, Faculté des Sciences de l'Université de Rouen, F-76821 Mont Saint Aignan Cedex, France

Abstract

In this paper, we describe two keys which characterize *hypercomplexity* and *emergence*. The first key is the *nonseparability* property, and the second is the process of *multiple interaction*.

1 Complex systems, hypercomplexity

Here, a *model* is a representation of some phenomenon by mathematical objects such as data sets, tables, matrices, relations, rules, equations, trees, graphs, hypergraphs, networks, functions, linear and nonlinear functional operators, etc. Typically a model will refer in general only to some aspects of the phenomenon in question. Having built a model for some perceived aspect of reality, this *modelling* can serve for some *computation* or as the basis for *simulation*, an efficient way to complement real-world experiments for non-invasive examination of physical reality.

In the following, *complex systems* are dynamic systems in which the components interact simultaneously in ways that are not precisely known. For example, these systems are a bacterium, cells in a body, animals in natural ecosystems, human agents in an economy or in a stock-market, etc. To gain perspicacity in these systems requires the construction of suitable models which in general will be *continuous* or *discrete* or even *hybrid*, often *multi-scale*, and which will be also *stochastic*, since a *deterministic* model is only one nice particular instance of stochastic models.

However, the complexity of biological systems differs from that of inanimate systems. First as mentioned in [1], the distribution of their components is very inhomogeneous. Furthermore, their structures evolve dynamically and their components are in permanent interactions, often at different levels, unlike inanimate systems. Another characteristic of these dynamic systems is the phenomenon of emergence [2]. In a biological system emerge properties and behaviours that cannot be readily predicted from a knowledge of properties and behaviours of the system's constituents. This very unpredictability makes emergence hard to model and to work with, by using concepts and considerations only appropriate to inanimate systems or only arising from mechanics, electronics or quantum physics [3]. To understand living systems and the mechanism of emergence, a new information processing theory would be required, with its own concepts, models and algorithms, and with efficient tools for computation; such a theory would be the basis for a brand-new and original technology. This theory may result from current attempts to harness the computational power of molecules [4], of bacteria [5], or of a larger field which could be termed "new computational biology" or "new science of information processing". Until such time, modelling and simulation will remain essential complements to wet experiments for understanding the complex systems behaviours.

Examples of just how complex a biological process can be and of why a new term, *hypercomplexity*, was used in [6], can be seen in [5], [6], [7], [8] and [9].

2 Nonseparability, Multiple Interactions

Although a few other terms such as non-linearity and strange attractors are employed in the field of complex studies, the vocabulary available to describe complex systems and hypercomplexity is still rather limited and, to begin a taxonomy of complex systems, it would be useful to have an idea of the parameters needed to capture the essence of complex biological organisations.

The model bacterium, *Escherichia coli*, is one of the best-understood of all organisms. It might therefore be expected that the process whereby this bacterium divides into two bacteria would be thoroughly understood. This is not the case, probably because bacterial division is dependent on the interplay between many different factors. The case can also be made that *E. coli* contains a level of organisation intermediate between macromolecules, such as genes and proteins, and the cell itself; this is the level of *hyperstructures* [7]. Other biological organisations also have intermediate levels and, to take account of hierarchical complexity, the vocabulary of complexity should include *levels* as a parameter. It could be argued that virtually all biological organisations, including social ones, have to undergo the vicissitudes of a fickle environment. Hence, additional parameters to characterise, and even quantify, hypercomplexity, might be derived and based on the essence of organisations subject to selection for growth in good conditions and survival in bad ones. This essence includes the existence of *quasi-equilibrium* and *non-equilibrium structures*.

Other parameters can be based on the process of *competitive coherence* [8] which underlies the operation of many biological organisations and which can be used to describe the way that a key subset of constituents are chosen to determine the behaviour of an organisation at a particular level. This choice results from a competition between the need of the organisation to behave in 1) a consistent way over time so as to maintain historical continuity via the status quo and 2) a coherent way at a particular time that makes sense in terms of both internal and environmental conditions and that is highly adaptive. This brought us, in [6], to a vision of biological organisations orbiting around two pairs of attractors. The first pair is the quasi-equilibrium versus non-equilibrium pair or, for example, spore versus growing cell. The second pair is the continuity versus coherence pair or the history versus the present.

In order to characterize *hypercomplexity* [6] and *emergence* [2], we are going to examine two keys which are behind these pairs of attractors. The first key is the *nonseparability* property, and the second is the process of *multiple interaction*.

2.1 The nonseparability property

In mathematics, a function $f(x_1, \ldots, x_n)$ is said *separable* if there exist n functions $f_1(x_1), \ldots, f_n(x_n)$ such that $f(x_1, \ldots, x_n) = f_1(x_1) + \ldots + f_n(x_n)$; otherwise it is *nonseparable*. This definition remains valid if the sum operator "+" is replaced by any other operator like product "x", division "/", etc.

The nonseparability property can characterize some undecidable problems, as the following optimization problem: Consider the optimization problem (*P*):

maximize $f_1(x_1), \ldots, f_n(x_n)$
under the constraints: $(x_1, \ldots, x_n) \in R^n; g_i(x_1, \ldots, x_n) \leq b_i, i = 1, \ldots, m$

- If the functions f and $g_i, i = 1, \ldots, m$ are linear, problem (*P*) is polynomial;
- if $x_i \in N$ and f and g_i are linear, $i = 1, \ldots, m$, then (*P*) is NP-complete [];
- but, if the function f is nonseparable, there is no algorithm which solves (*P*), that is if f is nonseparable, (*P*) is undecidable.

2.2 Multiple interaction

In the following, an interaction may involve two or more system's constituents not necessarily located at the same level. An example of interaction between two entities not located at a same level is given in [9]. A *simple interaction* denotes an interaction between two system's constituents. A *multiple interaction* is, by definition, an interaction between at least three system's constituents.

It is often defined, for the interaction set of a given biological system, an *interaction function* which associates some value to each interaction. Example: Let *S* be a biochemical reaction set; a simple interaction is a reaction between one enzyme and one metabolite, and a multiple interaction is a reaction involving one enzyme and at least two metabolites; besides, the stoechiometric coefficients of *S* define an interaction function.

The construction of the full *network* of interactions between the system's constituents is crucial for a better understanding of a biological process. This provides new insights into the structures and properties of biological systems, and allows useful computations. Several instances of networks of biomolecular interactions are depicted in [10].

Now, let us observe that, in a biological system, the presence of multiple interactions implies that the associated interaction function is nonseparable, and the network of interactions has not a *graph* structure but a *hypergraph* structure.

On the other hand, as mentioned in [11], biological function captures molecular architectures, and molecules need precise architectures and positions in space to function in an orderly fashion with specificity. This observation emphasizes the accuracy importance in the construction of a network of interactions and its interaction function.

In [12], an analysis of the protein interaction network among E. coli cell division proteins is described to indicate that the Fts proteins are connected to one another through multiple interactions.

2.3 Chaos, Emergence

In [13] it is noticed that in classical mechanics, a system always has a separable Hamiltonian H such that H is a sum of Hamiltonians H_i, one element in the sum for each subsystem; and as a consequence, this system is nonchaotic. Thus, the key to chaos in classical systems is better characterized as the nonseparability of the Hamiltonian rather than the nonlinearity of the equations of motion. We can make an analogous conclusion and say that, in biological processes, the emergence is characterized as the product of the nonseparability of the interaction function and of the multiple interactions.

References

[1] Y. Louzoun, S. Solomon, H. Atlan, and I. R. Cohen, *Modeling complexity in biology,* Physica A 297 (2001) 242-252

[2] J. Deguet, Y. Demazeau, and L. Magnin, *Elements about the Emergence Issue: A Survey of Emergence Definitions,* ComplexUs 2006, 3: 24-31, Complex Systems Part 2.

[3] A. K. H. Bengtsson A.K.H., *Quantum Computation: A Computer Science Perspective*, 2005, **http://arxiv.org/abs/quant-ph/0511274**

[4] A. Adamatzky, N. Rambidi, M. Conrad, T. Sienko, *Molecular Computing* (2003), MIT Press.

[5] E. Ben Jacob, Y. Aharonov and Y. Shapira, *Bacteria Harnessing Complexity*, Biofilms (2004) 1, 239-263.

[6] V. Norris, A. Cabin and A. Zemirline, *Hypercomplexity*. Acta Biotheoretica, V 54, N 4, 313-330, December 2005.

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

[8] V. Norris, *Modelling Escherichia coli: The concept of competitive coherence*, Comptes Rendus de l'Académie des Sciences Série III – Sciences de la Vie – Life Sciences 321 (9) : 777-787, (1998).

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

[10] F. Képès, *Complex Networks of Biomolecular interactions*, Advanced Introduction to Genomic Biology, (October 2006).

[11] E. Westhof, *Structural Biology Nucleic Acids*, Advanced Introduction to Genomic, (October 2006).

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

[13] F.M. Kronz, *Nonseparability and Quantum Chaos*, Philosophy of Science 65, 50-75, (1998)

PART III POSTERS

LifeExplorer: a visualization tool for molecular and synthetic biology

Éric Fourmentin[1], Damien Larivière[1]

[1] Fondation Scientifique Fourmentin-Guilbert, F-93160 Noisy-Le-Grand, France

Abstract

Biologists deal with large amounts of data on cellular components and processes. Therefore, repositories, models and simulation tools are necessary to provide a synthetic view of the inherent complexity of living organisms. As a step in this direction, the Foundation is developing a **3D visualization tool** enabling to map cell components and to navigate within the cell at various scales.

Audience and benefits

LifeExplorer, as an interactive visualization environment, is a tool for biologists to better understand how biological processes take place in the cell space. In particular it should help to analyse how a cell is organised in terms of compartments and internal crowding. The LifeExplorer tool will also support biologists and engineers to design virtual processes and cells.

First step and challenges

A LifeExplorer prototype is being developed by the Foundation and industrial/academic partners. It consists of the integration into a visual representation of all the components involved in the transcription process of the lactose operon in *Escherichia coli*.

The LifeExplorer prototype takes into consideration all actors of transcription, the resulting crowding and spacing between objects. Macromolecule localization and orientation are currently empirical, in agreement with actual knowledge deduced from molecular biology or in vitro single-molecule experiments.

A better understanding of the highly integrated sub-cellular organisation by a visual representation tool is achievable by the collaboration of IT experts and biologists. In conjunction with the development of the tool itself, a large part of its value will come from accurate and validated data by the biologists community. We welcome all biologists to participate into this effort.

Design of genetic networks with targeted behavior

Guillermo Rodrigo[1], Javier Carrera[1], Alfonso Jaramillo[2]

[1] Dep. Matematica Aplicada, Universidad Politecnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

[2] Laboratoire Biochimie, CNRS-UMR 7654, Ecole Polytechnique, Route de Saclay, F-91128 Palaiseau, France

Abstract

We describe the use of an automated optimization method to aid in the rational design of new synthetic genetic circuits. Biological circuits are harder to design than their electrical counterparts, as genetic circuits sharing the same topology could behave very differently depending on the kinetic parameters, furthermore the addition of new interactions could change the circuit dynamics in unexpected ways. Our automated procedure can design genetic circuits, composed of predefined genetic parts, having a desired time-response and a degree of robustness under stochastic perturbation of the parameters. This will allow the design of new genetic devices with desired transfer functions and robustness. Our procedure could be viewed as a genome evolution where a given genome would acquire mutations at the promoter, ribosome binding site or coding regions. Then one would select the fittest organism producing a desired output from a given set of inputs. This amounts to explore the space of all possible transcriptional regulation networks, where at each step we would add/subtract new interactions or modify kinetic parameters, to find the optimal circuit with specified system behavior. We apply our methodology to the design of specific genetic devices having a desired switching or oscillatory behavior. Our computational methodology will provide very valuable information for understanding natural circuits and for designing new synthetic ones.

Interdisciplinary approaches to life sciences master program: Combining Theory and Experiment

David Bikard[1], Timothée Flutre[1]

[1] Master M2 Recherche "AIV", ENS - Paris VII.
Paris Interdisciplinary College, Faculty of Medicine - Paris V, F-75014 Paris.

Abstract

The current challenges in biology involve ever more disciplinaries in order to tackle the complexity of living systems from different views, methods and tools. Thus it becomes necessary to learn how to collaborate efficiently with researchers from several scientific fields. In this way, the master program "Interdisciplinary approaches to life sciences" gives an innovative opportunity to students from numerous backgrounds (biology, medicine, physics, mathematics, computer science, chemistry) to develop skills in scientific curiosity and intellectual mobility. We begin with a one-week seminar during which students write a research project within interdisciplinary groups. Then along the year, we animate once a week a meeting with senior investigators to intensively analyse recent publications and synthesis. Moreover each student performs three-month internships in a laboratory, at least one theoretical and another experimental (one rotation can be replaced by specialized courses from a different master program).

We, David and Timothée, are currently following this master program as third and last year of our engineering school, the Institut National Agronomique Paris-Grignon (AgroParisTech). Following are detailed our projects realized during the master.

David made his first rotation in the laboratory *Variation and Abiotic Stress Tolerance* of O. Loudet at the INRA, working on an epistatic interaction within recombinant inbred lines in Arabidopsis thaliana. He is now working in the laboratory "Bacterial Genome Plasticity" of D. Mazel at the Pasteur Institute, on the use of the combinatorial capacities of integrons in synthetic biology: Integrons are a remarkable recombination platform involved, in particular, in antibiotic multi-resistance. We propose to use integrons as a combinatorial platform for the study of metabolic pathways. A synthetic integron will be build and integrated into E. Coli chromosome. Expression of the integrase and genes cassettes will be controlled through inducible promoters. Under the integrase expression, gene cassettes bored by a *library plasmid* will be randomly integrated into the integron platform. As a proof of principle, we want to show that such a synthetic integron is able to reconstruct E. Coli tryptophan operon out of a library of numerous genes cassettes.

Timothée made his first rotation in the laboratory *Bioinformatics and Genomics* of H. Quesneville at the Institut Jacques Monod, working on the de novo detection of transposable elements (TEs) in Drosophilidae genomes. Such methods are required to have detailed, quantitative inferences about the contribution of TEs to genome sequences. One possibility is to perform a self-alignement of the genome, cluster the high scoring pairs and derive consensus. We then propose to annotate consensus on the basis of biological features (long terminal repeats, terminal inverted repeats, ORF coding for a transposase...). Timothée is now working in the laboratory *Modeling in Integrative Biology* of K. Pakdaman at the same institute, working on the impact of transposable elements on the

robustness of gene networks. His model aims at describing the impact of TEs dynamics on host genome, particularly the restructuration of the genomic landscape by transposition and duplication.

Connection between the cell cycle and the circadian clock in mammalian cells

Judit Zámborszky[1], Chris I. Hong[2], Attila Csikász-Nagy[1]

[1] Materials Structure and Modeling Research Group of the Hungarian Academy of Sciences and Department of Applied Biotechnology and Food Science, Budapest University of Technology and Economics, Gellért tér 4, H-1521 Budapest, Hungary

[2] Department of Genetics, Dartmouth Medical School, Hanover, NH 03755, USA

Abstract

Mammalian cells contain two endogenous oscillatory systems: the circadian pacemaker and the control of cell division. Recent findings tell us that these systems are coupled. We investigate the effect of this coupling on timing of cell cycle processes with the tools of mathematical modeling.

1 Introduction

Cell cycle and circadian rhythms are conserved from cyanobacteria to humans with robust cyclic features. Recently, the molecular link between these two cyclic processes has been discovered: the circadian clock directly regulate Wee1 kinase that inhibits entry into mitosis. (*Matsuo et al., 2003, Science 302:255-259*)

1.1 Building the model

We built a model dealing with the molecular interaction of the circadian and cell cycles of mammalian cells. We started from the comprehensive Novak & Tyson model (*2004, J. Theor. Biol. 230:563-79*), extended it with the regulation of the G2/M transition, and connected it with a simplified circadian rhythm model. The interactions of the wiring diagram were turned into a set of differential equations by the rules of reaction kinetics and later we added a noise term to the equations. We fitted the new parameters of the nonlinear differential equations and analyzed the system by computational simulations.

1.2 Mode-locking

Variation in the coupling strength between these two systems shows different results. Based on our computational analysis, we report quantized cell cycles when *wee1* transcription is strongly influenced by the circadian clock. This occurs from "mode-locking" phenomena that create various periodic repetitions of cell division cycles with different cell growth rates.

1.3 Size control

It is questionable that mammalian cells use the same size control regulation as yeasts do. Our model suggests that critical cell mass control regulation in mammalian cells might depend on the circadian clock: our results show that there is no critical mass control if this coupling is weak, or the cells grow with a mass doubling time close to 24 hours, but strong coupling that induces quantized cell cycles can introduce strong size control into the system.

Structural and Evolutionary Study of Biological Network by Graph Laplacian Spectrum

A. Banerjee[1], J. Jost[1]

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

Abstract

We have developed a tool, spectra of normalized graph Laplacian that helps to understand the network structure with deep perception so that we could recognize the source of the network. We have explored the information about different topological properties of a graph carried by complete spectra of normalized graph Laplacian. And have investigated how and why structural properties are reflected by the spectra and how spectra change according to different networks from different sources. So we have inferred that spectral distribution is an excellent diagnostic to categorize the different networks from different sources. Different graph operation related to evolution of a network produce specific eigenvalue. Construction with those operation describe certain processes of graph formation that leave characteristic traces in the spectrum. So useful plausible hypothesis about evolutionary process could be made and it would be easy to take decision about the evolutionary assumption that is more relevance for the evolution of that system by investigating the spectra of a graph constructed from actual data.

A taxonomy of inter-level relationships in Systems Biology for multi-scale agent-based models

Chih-Chun Chen[1]

[1] Department of Computer Science, University College London, London, WC1E 6BT, UK

Abstract

One of the most challenging aspects of modelling dynamic biological systems is the fact that our understanding of these systems is often made up of explanations that cross several different spatio-temporal scales and levels of abstraction [2], [3]. Furthermore, many explanations are given in terms of emergence, where behaviour observed at a certain level (e.g. macroscopic) is the output of interactions between entities at a lower level (e.g. mesoscopic). Agent-based modelling (ABM) is a promising computational paradigm for capturing these inter-level relationships [4], [1]. In ABM, biological entities are represented by software agents. Agents each have a set of behavioural rules, which determine how their states change in response to their current state and the state of their local environment. The local environment can include other agents or environmental variables that are spatially (e.g. all agents within a certain area) and/or logically related (other agents in the same metabolic network). Because hierarchical relationships can be easily defined in an agent framework by agent nesting, ABM lends itself naturally to multi-level modelling.

However, for models to truly reflect biological understanding it is important that relationships between levels are defined precisely. For example, it is important to distinguish between an equivalence relationship (e.g. A at level x = (B, C, D) at level y) and a causal relationship (e.g. (B, C D) at level $y \rightarrow A$ at level x). It is also important to distinguish between entity/agent, state and event/behaviour mappings between different levels. Here, a practical taxonomy is suggested to help clarify these distinctions for the purpose of modelling biological systems with ABM.

References

[1] Khan, S., Makkena, R., McGeary, G., Decker, K., Gillis, W. and Schmidt, C. A Multi-Agent System for the Quantitative Simulation of Biological Networks. AAMAS '03. 2003.

[2] Kitano, H. Computational Systems Biology. Nature 420:6912, 206–210. 2002.

[3] Pierce, S. M., Skalak, T. C. and Papin, J. A. Multiscale biosystems integration: Coupling intracellular network analysis with tissue-patterning simulations. IBM J. Res. & Dev. 50:6, 601–615. 2006.

[4] Walker, D. C., Southgate, J., Hill, G., Halcombe, M., Hose, D. R., Wood, S. M., Mac Neil and Smallwood, R. H. The Epitheliome: Agent-Based Modelling of the Social Behaviour of Cells. BioSystems 76, 89–100. 2004.

Topological Approach of the Golgi Apparatus: Towards a Discriminating Modelling?

M. Poudret[1,2], J.-P. Comet[1,2], P. Le Gall[1,2], A. Arnould[3], P. Meseure[3]

[1] IBISC FRE 2873 CNRS, Université d'Évry
[2] Epigenomics Project, Genopole®, F-91000 Evry, France
[3] SIC, Université de Poitiers, 86962 Futuroscope, France

Abstract

The biological cell is a strongly compartmented environment. Though many biological mechanisms are the fact of molecules exchanges between membrane-bounded compartments, many others are based on the dynamics of the neighbouring relations within the compartmentation. For instance, in the well-known exocytosis, compartments divide, influencing the proteins concentrations. In order to take into account this compartmentation in a new model, we base our work on the topology-based geometric modelling (or topological modelling) and choose a sub-class of graphs, the n-dimensional generalised maps model [1] (n-G-maps for short) as a topological model. Here, the objects are decomposed into basic units (volumes, faces, edges, vertices, *etc.*) and both the geometry of those units and the dynamics of neighbouring relations are precisely described. Abstracting the biological compartments as volumes, we use the topological models to follow-up the neighbouring relations of the compartments. As an n-G-map is a graph, we translate the n-G-maps operations into graph transformation rules [2]. Thus, this formalism fits to the rule-based languages, which are well-adapted to model biochemical reactions. Moreover, for such a model, formal methods like model-checking can be applied in order to verify that the model satisfies a known property of the system.

This new cellular process modelling has led us to propose a mean to discriminate models that involved a strong geometrical and topological structure. In the secretory pathway of the living cell, where the excretion of proteins takes place, the Golgi apparatus is a complex organelle where the dynamics of the compartmentation is essential but not precisely described. The microscopy techniques do not allow the biologist to capture the entire structure of the apparatus.

The observations lead to at least two different models. On one hand, the apparatus consists in a pile of disconnected saccules and proteins move from one to another through a sequence of pack and unpack operations within transport vesicules. On the other hand, a tubular structure [3] that connects the saccules allows the proteins to diffuse into the apparatus. This state of the art knowledge allows us to define a topogical model for these hypotheses and parameterize them according to given biological input data. Then, we can simulate the models in order to estimate biological quantities (proteins production rates, energy consumption, *etc.*) through output values of the simulations. Results should differentiate these models and highlight the model which better captures the activities of Golgi apparatus.

Acknowledgements

The authors would like to thank F. Képès, J.-M. Verbavatz, J. Guespin and A. Rambourg for the numerous discussions about recent hypotheses on Golgi Apparatus.

References

[1] P. Lienhardt. Subdivision of n-dimensional spaces and n-dimensional generalized maps. In *SCG'89*, pages 228–236, Saarbrücken, Germany, 1989. ACM Press.

[2] M. Poudret, J.-P. Comet, P. Le Gall, A. Arnould and P. Meseure. Topology-based geometric modelling for biological cellular processes. In *LATA'07*, to appear in LNCS, April 2007.

[3] F. Képès, A. Rambourg, and B. Satiat-Jeunemaître. Morphodynamics of the secretory pathway. *International review of cytology*, 242:55–120, 2004.

3D simulation of biological membrane: mitochondrial application

Charles Lales[1], Nicolas Parisey[2], Jean-Pierre Mazat[2], Marie Beurton-Aimar[1,2]

[1] LaBRI CNRS UMR 58000, Université Bordeaux 1, 350 avenue de la libération, F-33405 Bordeaux, France

[2] Laboratoire de physiologie mitochondriale INSERM U688, Université Bordeaux 2, 146 rue Léo Saignat, F-33076 Bordeaux, France

Abstract

"In silico" simulations of biological processes must take into account several kinds of complex molecular behaviors. Simulations of membranes, due to interactions between phospholipid bilayers and enzymes embedded in them, are typically an example of such complexity.

In this context we are studying the specific case of the respiratory chain, a pool of five enzymatic complexes embedded into the inner mitochondrial membrane. We would be able to understand the emergence of inner membrane complex macro-structures and their impacts on the enzymatic chain reactions, especially the link between the cristae of the inner membrane and the raft of the enzymes of the respiratory chain.

The model implemented consists of several types of molecule (phospholipids, enzymes,...). The 3D space is continuous but a discretized grids optimize the neighbors research. A molecular abstraction by a single point is not capable to handle a dynamical 3D structure and its spatial orientation. On the other hand molecular dynamics at the atom level is not suitable for the expected time scale and length scale due to the huge number of molecular interactions to compute. Thus, we have chosed to reduce the granularity at the atom set level or grain level (coarse graining). Interactions between molecules are reduced to a set of forces (attractive, repulsive) due to Lennard Jones potential. Molecules are indeed subject to two distinct forces in the limit of large distance and short distance: an attractive force at long ranges (van der Waals force, or dispersion force) and a repulsive force at short ranges (the result of overlapping electron orbitals, referred to as Pauli repulsion from Pauli exclusion principle). The Lennard-Jones potential (also referred to as the L-J potential, 6-12 potential or, less commonly, 12-6 potential) is a simple mathematical model that represents this behavior. We would like to compare different models of the intramolecular interactions between the grains: from a rigid model with fixed grains around the gravitiy center of the molecule to a full linear spring-like model providing molecular flexibilty. We would also like to test different L-J potentials being able to give identical qualitative and quantitative global results and to estimate the impact of the granularity (the numbre of grain per molecule) on the emergent molecular organisation.

This kind of modelisation appears to be a great compromise between the "unreachable" complexity of molecular dynamics and the restrictive abstraction of molecule by points (or by spheres).

Automatic Calibration of Agent-Based Models

Benoît Calvez[1] and Guillaume Hutzler[1]

[1] IBISC Université d'Évry Val d'Essonne & CNRS FRE 2873, Tour Évry 2, F-91000, Évry

Abstract

When developping multi-agent systems (MAS) or models in the context of agent-based simulation (ABS), the tuning of the model constitutes a crucial step of the design process. Indeed, agent-based models are generally characterized by lots of parameters, which together determine the global dynamics of the system. Moreover, small changes made to a single parameter sometimes lead to a radical modification of the dynamics of the whole system. The development and the parameter setting of an agent-based model can thus become long and tedious if we have no accurate, automatic and systematic strategy to explore this parameter space. There are several different methods to explore the parameter space [1] [2].

That's the development of such a strategy that we are currently working on. We first suggested the use of genetic algorithms, with the aim of capturing the quality of the model in the fitness function (typically the consistency with experimental data or with the observed dynamics in the real system). This approach proved efficient [3] but it also raises specific difficulties. Furthermore, il only provides a single parameter set as a solution and doesn't give any insights about the parameter space as a whole.

To correct this drawback of the GA approach, we are now investigating an alternative strategy, in which the whole parameter space is explored in a parallel way. The basic idea is to take advantage of the fact that agent-based simulations rely on multiple agents. We propose to enable the parameterization of the different agents with different settings. In this method, each parameter is divided into intervals and the parameter space will be explored differentially, depending on the potential interest of the different regions of the space. Taking inspiration from dichotomic search and from octrees, we consider that a parameter space of dimension n (n independent parameters) is initially divided into hypercubes of dimension n. Then, for a each individual parameter, depending on rewards received by the different intervals, we may differentially merge or divide the intervals.

More precisely, when an agent is instanciated, the value of each parameter is chosen randomly among the intervals that divide the definition set of the parameter. After a model has been evaluated (by running a simulation and computing the fitness of the model), the intervals in which the parameters of the agents have been chosen are rewarded. For each parameter, and for each interval, the reward is proportional to the global fitness of the model and to the number of agents that have taken their parameter's value in the interval.

The method has been succefully applied to several test models and we are now in the process of applying it to more complex biological examples such as the simulation of the glycolysis and the phosphotranferase systems in Escherichia coli, or the migration of cancer cells.

References

[1] Sven A. Brueckner and H. Van Dyke Parunak. Resource-aware exploration of the emergent dynamics of simulated systems. In *AAMAS '03*, pages 781–788, 2003.

[2] Manuel Fehler, Franziska Klügl, and Frank Puppe. Techniques for analysis and calibration of multi-agent simulations. In *ESAW*, volume 3451 of LNCS, pages 305–321. Springer, 2004.

[3] Benoît Calvez and Guillaume Hutzler. Automatic tuning of agent-based models using genetic algorithms. In *Proceedings of the 6th International Workshop on Multi-Agent Based Simulation (MABS'05)*, volume 3891 fo LNAI, 2005.

Hybrid modelling of biological regulatory network with delays

Jamil Ahmad[1], Gilles Bernot[3], Jean-Paul Comet[2,3],
Olivier Roux [1] and Adrien Richard[2]

[1] IRCCyN UMR CNRS 6597, BP 92101, 1 rue de la Noë, F-44321 Nantes Cedex 3, France
[2] IBISC, Université d'Evry, Boulevard François Mitterrand, F-91025 Evry cedex, France
[3] Epigenomics Project, Genopole®, F-91000 Evry, France

Abstract

Biological regulatory networks (BRN) represent interactions between biological entities. For example, genetic regulatory networks are graphs where vertices represent genes or regulatory products e.g. RNA, proteins and edges represent interactions between them. These interactions are further directed (regulators are distinct from targets) and signed ($+$ for activation, $-$ for inhibition). Biologists often need to use the previously described regulatory graphs as a basis for generating dynamical models using either continuous representation or discrete ones [2]. The first approaches for studying the dynamics of BRN are based on the system of differential equations in which each equation describes the evolution of the concentration of a type of macro-molecules. In 1970s, René Thomas introduced a boolean approach that qualitatively captures the dynamics and where each entity is either on or off. This approach was then generalised to multiple values. It was then shown that the presence of a feedback circuit in BRN has a major role in the behavior of a system. For example to have multistationarity, the system must have a positive circuit (each entity of the circuit has a positive influence on its own evolution) and to have homeostasy (the equilibrium state at which the system converges or oscillates), there must be a negative circuit [3].

We use the approach of hybrid automata for the modeling of BRN. The formalism of René Thomas is very suitable for the discrete modeling of BRN [2]. We further improve this modeling approach by introducing time delays. We present two ways of using delays for analysing a dynamical model. We are interested to find the paths and their temporal regions and the infinite cycles in the model of BRN. We use HyTech [1], a verification tool for linear hybrid systems, to obtain the equivalent temporal regions of paths and the initial conditions, from where the system can enter into infinite cycles or invariance kernel.

Résumé

Les réseaux de régulation biologiques (RRB) représentent les interactions entre les objets biologiques. Par exemple, les réseaux de régulation génétiques sont des graphes où les sommets représentent les gènes ou les produits (ARN ou protéines) et les arcs représentent leurs interactions. Les interactions sont signées (+ pour activation, - pour inhibition) et orientées. Les biologistes établissent des modèles moléculaires représentant l'ensemble des relations connues pour un réseaux donné, qui correspondent aux gènes et aux protéines régulatrices qui, en se liant avec ces gènes favorisent (effet positif) ou empêchent (effet négatif) leur expression. Les biologistes utilisent les graphes d'interactions pour générer le modèle dynamique en utilisant les approches continues ou discrètes [2]. Les premières approches pour étudier les dynamiques de RRB sont basées sur les systèmes

d'équations différentielles dans lesquels chaque équation décrit l'évolution de la concentration d'un type de macro-molécules. Dans les années 70, René Thomas a introduit une approche booléenne dans laquelle chaque entité est allumée ou éteinte et qui capture qualitativement la dynamique. Puis cette approche a été généralisée aux valeurs multiples. Il a ensuite montré que la présence de circuits de rétroaction dans ces réseaux avait une importance capitale pour le comportement du système. Par exemple pour avoir une multistationnarité, le système doit présenter un circuit positif (chaque entité du circuit a une influence positive sur sa propre évolution) et pour avoir une homéostasie. – état d'équilibre vers lequel le système converge où autour duquel il oscille. – il doit présenter un circuit négatif [3].

Nous utilisons l'approche de modélisation par automate hybride pour la modélisation de RRB. Le formalisme de René Thomas est plus adéquat pour la modélisation discrète de RRB [2]. Nous améliorons ce formalisme en introduisant les temps de délais. Nous présentons deux façons d'utiliser les délais pour l'analyse du modèle dynamique. Nous sommes intéressés pour : trouver les chemins et leurs régions temporelles dans le modèle de RRB et aussi les propriétés caractéristiques des cycles infinis. Nous utilisons HyTech [1], un outil pour la vérification des systèmes hybrides linaires, qui permet d'obtenir les régions temporelles des chemins et les conditions initiales, à partir de ces conditions initiales, il existe des cycle infinis qui constituent le noyau d'invariance.

References

[1] T.-A. Henzinger, P.-H. Ho, and H. Wong-Toi, "HYTECH: A model checker for hybrid systems," *International Journal on Software Tools for Technology Transfer*, vol. 1, no. 1–2, pp. 110–122, 1997. [Online]. Available: citeseer.ist.psu.edu/henzinger97hytech.html

[2] R. Thomas and R. D'Ari, "Biological Feedback," *CRC Press, Boca Raton, FL*, 1990.

[3] R. Thomas, "Logical analysis of systems comprising feedback loops." *J. Theor. Biol.*, vol. 73, no. 4, pp. 631–56, 1978.

Introduction of a new type of nodes representing the biological processes into the genetic regulation graph

Zohra Khalis[1]

[1] Programme d'épigénomique Génopole et université d'Évry, F-91025 Évry cedex, France.

Abstract

The knowledge of a genetic regulation graph isn't sufficient to know the complete behavior of the system, because its dynamics is governed by a lot of parameters [2]. So, it's necessary to seek among the set of the possible parameter settings those which are coherent with the existing data and/or the assumptions formulated on the system. This approach is made tough because of the low number of really exploitable data, because of the presence of noise and of the complex nature of the studied biological processes. Current modelings of genetic regulation networks have a multitude of often unknown parameters, which result with an exponential number of models to study, and consequently limits the developed exhausive approach in [1].

Taking into consideration of underlying biochemical processes of interactions of the graph reduces number of possible models: for example the complex-forming can be represented by a "process", a transition taking both components of the complex at entry and being able to generate it. The regulation network is thus represented by a graph having 2 types of nodes: biological entities (genes or proteins) and "processes" whose actions are described by logical formula. Once these informations coded in the formulas, a process is active if its formula is evaluated with true in the current environment. When a process p_1 is included in the formula of another process p_2, the process p_1 is then replaced by its formula in the process p_2.

Taking into account this co-operation type information reduces the number of parameters. However, the presence of cycles can involve an indetermination for the evaluation of a formula: the formula of a process p_1 can contain a process p_2, which contains the process p_1. So for the representation of a genetic regulation network, only cycles containing at least a biological entity are authorized. From the graph, it is then possible to build a system of transitions reflecting system dynamics.

Finally, for each possible parameter setting, the coherence of dynamics with biological knowledge and/or assumptions formulated is checked. For that, the knowledge and/or the assumptions are translated into CTL temporal logical formulas. If the system of transitions does not satisfy CTL formulas (Model checking), the model is not considered any more for the sequel of the study.

Résumé

La connaissance d'un graphe de régulation génétique n'est pas suffisante pour connaître le comportement complet du système, car la dynamique du système est régie par tout un ensemble de paramètres [2]. Il est donc nécessaire de rechercher parmi l'ensemble des paramètrages possibles ceux qui sont cohèrents avec les données existantes et/ou avec les hypothèses émises sur le système. Cette démarche est rendue difficile à cause du faible nombre de données réellement exploitables, à la présence de bruit ainsi qu'à la nature

complexe des processus biologiques étudiés. Les modélisations actuelles des réseaux de régulation génétique possèdent une multitude de
paramètres souvent inconnus, ce qui conduit à un nombre exponentiel de modèles à considérer, et dès lors limite l'approche exhausive développée dans [1].

La réduction du nombre de modèles s'effectue grâce à la prise en compte des processus biochimiques sous-jacents aux interactions du graphe: par exemple la formation de complexe peut être représentée par un "processus", transition prenant en entrée les deux composants du complexe et pouvant générer celui-ci. Le réseau de régulation est donc représenté par un graphe ayant 2 types de nœuds : les entités biologiques (gènes ou protéines) et les "processus" dont l'action est décrite par une formule logique. Une fois ces informations codées dans les formules, un processus est actif si sa formule est évaluée à vraie dans l'environnement courant. Lorsqu'un processus p_1 intervient dans la formule d'un autre processus p_2, le processus p_1 est alors remplacé par sa formule dans le processus p_2.

La prise en compte de ces informations de type coopération, réduit le nombre de paramètres. Cependant, la présence de cycles peut entraîner une indétermination de lors de lévaluation d'une formule : la formule d'un processus p_1 peut contenir un processus p_2 qui lui même contient le processus p_1. De ce fait, lors de la représentation d'un réseau de régulation génétique, on ne s'autorisera que des cycles contenant au moins une entité biologique. À partir de ce graphe, il est alors possible de construire un système de transitions reflètant la dynamique du système.

Enfin, pour chacun des paramètrages possibles, on vérifie si la dynamique est cohérente avec les connaissances biologiques et/ou hypothèses émises. Pour cela, les connaissances et les hypothèses sont exprimées sous forme de formules temporelles CTL. Si le système de transitions ne satisfait pas les formules CTL (Model checking), le modèle n'est plus considéré pour la suite de l'étude.

References

[1] G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks: Extending thomas' asynchronous logical approach with temporal logic. *J.T.B.*, 229(3):339–347, 2004.

[2] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks. *Bull. Math. Biol.*, 57(2):247–276, 1995.

Analysing Gene Regulatory Networks by both constraint programming and model-checking

Jonathan Fromentin[1], Olivier Roux[1], Pascale Le Gall[2] and Jean-Paul Comet[2]

[1] IRCCyN UMR 6597, CNRS & École centrale de Nantes
[2] IBISC CNRS FRE 2873 & Université d'Évry

Abstract

The Gene Regulatory Networks (GRN) and their formal models [1] were conceived to describe the interactions between genes inside the cell. They allow foreseeing the complex evolutions of relative concentrations of several interacting genes. The work in [2], shows an example of GRN which are modeled by using the discrete modeling approach of R. Thomas. It has been shown that the computer tools can be used to analyse their structure and functional properties [3]. This work shows the model-checking method to explore a model that satisfies temporal properties. The model-checking was applied to the analysis of the GRN after the enumerations of the model, such that for each enumeration of the model, the properties estimated by biologists has been verified. The number of possible enumerations for the same model of the GRN is exponential.

The use of constraint programming [4] allows to bypass this problem by generating the enumerations of model directly validating the specified properties. Therefore, constraint programming appears to be a promising approach for the analysis of these dynamics. The goal of this work is to conceive and implement a combined approach (temporal logical formulas and constraint programming) to select enumerations of a model for GRN. In our context, the temporal properties are expressed in the form of constraints on our model that will represent a set of solutions. The transformation of a temporal logic formula to constraints on model is done by a mixed approach of model-checking and constraint programming.

Résumé

Les Réseaux de Régulation Biologique (RRB) et les modèles qui les représentent formellement [1] ont été conçus pour décrire les interactions entre gènes à l'intérieur de la cellule. Ils permettent de prévoir les évolutions complexes de concentrations relatives de plusieurs gènes interagissant. Des exemples [2] ont été mis à jour par des biologistes, et des travaux – utilisant la modélisation discrète de R. Thomas – sont menés pour traiter informatiquement ces réseaux afin de détecter des propriétés structurelles ou fonctionnelles [3]. Ces travaux utilisent la méthode de model-checking qui consiste à explorer un modèle fini par confrontation d'une propriété temporelle. Le model-checking a été appliqué à l'analyse des RRB par énumération des modèles, chacun donnant lieu à la vérification d'une propriété jugée pertinente par les biologistes. Mais les RRB étant des systèmes complexes, le nombre d'énumérations possibles pour une même modélisation est exponentiel.

L'utilisation de la programmation par contraintes [4] permet de pallier ce problème en générant directement les modèles validant les propriétés spécifiées. La programmation par contraintes s'avère donc être une approche prometteuse pour l'analyse de ces dynamiques. Le but de ce travail est de concevoir et mettre en œuvre une approche combinée (formules logiques temporelles et programmation par contraintes) permettant de

sélectionner des modèles pour des RRB. Dans notre contexte, les propriétés temporelles sont exprimées sous forme de contraintes sur notre modélisation et définissent un ensemble de solutions. La transformation d'une formule en logique temporelle à celle de contraintes sur la modélisation se fait par une méthode mixte de model-checking et de programmation par contraintes.

References

[1] R. Thomas, D. Thieffry and M. Kaufman. Dynamical behaviour of biological regulatory networks - I. Biological role of feedback loops an practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, **57(2)**:247–276, 1995.

[2] J. Guespin-Michel and M. Kaufman. Positive Feedback Circuits and Adaptive Regulations in Bacteria. *Acta Biotheoretica*, **49(4)**:207-218, 2001.

[3] A. Richard, J-P Comet and G. Bernot. *Modern Formal Methods and Applications*, chapter Formal methods for modelling biological regulatory networks. Springer, 2006.

[4] F. Fages. *Programmation Logique Par Contraintes*. Ellipses, 1996.

Autonomous interactions for *in virtuo* experience of multi-model and multi-scale complex biological systems

G. Desmeulles[1], P. Redou[1], JF. Abgrall[2], V. Rodin[3] and J Tisseau[1]

[1] Centre Européen de Réalité Virtuelle, Brest, France.
[2] Université de Bretagne Occidentale, Brest, France.
[3] Laboratoire d'hématologie, CHU de Brest, France.

Abstract

Nowadays, the Virtual Reality makes feasible the simulation and the (*in virtuo*) experience of complex phenomena, to complete the *in vivo* or the *in vitro* investigations. We apply this alternative method to different fields of biology : cutaneous neurobiology, vascular physiology, hematology, immunology and oncology. Here, we present the definition of a generic modelling framework and its implementation for the study of physiological systems.

In the first place, the generic model is based on the reification of interactions into autonous active objects. Thereby, the biological models can be organized in a layout of autonomous systems. Therefore, the generic model infers two conceptions of autonomy: the first one is used to design virtual reality systems and the second one is oriented towards biological modelling.

The generic model is specialized into several modelling tools for biology. Thereafter, the library composed by the generic models and the tools allows the building of applications.

Résumé

L'usage de la réalité virtuelle (RV) permet, aujourd'hui la simulation et l'expérimence (*in virtuo*) de phénomènes complexes, pour compléter la recherche *in vivo* ou *in vitro*. Nous appliquons cette méthode alternative à différents champs de la biologie : neurobiologie cutanée, physiologie vasculaire, hematologie, immunologie et oncologie. Nous présentons ici, un cadre générique de modélisation et d'implémentation adapté à l'étude des systèmes physiologiques.

En premier lieu, le modèle générique proposé s'appuie sur le principe de la réification des interactions en objets actifs autonomes. Ensuite, il permet l'organisation des modèles biologiques en un agencement de systèmes autonomes. Il rassemble alors deux conceptions de l'autonomie : l'une est destinée à concevoir les systèmes de réalité virtuelle et l'autre a pour objet la modélisation en biologie.

Le modèle générique est dérivé en un certain nombre d'outils de modélisation pour la biologique. La bibliothèque composée du modèle générique et des outils de modélisation permet alors de réaliser différentes applications.

Confrontation between models and real data for the DNA molecules space structure by the means of increased reality

M. Essabbah[1], R. Gherbi[1,2], M. Mallem[1]

[1] IBISC FRE 2873 du CNRS, Université d'Evry Val d'Essonne, F91000 Evry, France
[2] LIMSI-CNRS, Université Paris-Sud, F-91405 Orsay, France

Abstract

The genomic sequences are initially known under their linear form. However, it have also a three- dimensional structure which can be useful for genomes analysis. This 3D representation of the structure brings a new point of view for the sequences analysis. There are few studies about DNA geometrical properties, especially for large DNA segments. It is precisely on the latter point that synergy between data processing specialists and biologists can be extremely beneficial, exploiting multidisciplinarity. ADN-Viewer[1] is a complete software developed to explore DNA 3D modeling. It is based on a local conformation model. Our objective is to confront the results predicted by this software with real data (images) resulting from biological experiments. Thus, the deal is to try refining the 3D model by pairing real and predicted images. The molecules analysis under Atomic Force Microscopy (AFM) will enable us to make a structural analysis of various interesting genome areas. These images will be compared, paired and readjusted with the predicted images by using analysis and image processing techniques. It would be relevant to quantify the prediction error and to correct the three-dimensional conformation rules by returning to the model.

Keywords

DNA space structure, Microscopic images, Increased reality, Trajectory extraction, Pattern matching, Shape description.

[1]R. Gherbi and J. Hérisson, "3d Modeling tools for spatial-based in silico analysis of DNA". In International electronic journal on Computer Graphics and Geometry. Volume 3, N° 1, 2001

LIST OF ATTENDEES

(April 1th, 2007)

ADAMALA	Katarzyna	katarzyna@ewolucja.org
AHMAD	Jamil	jamil.ahmad@irccyn.ec-nantes.fr
AIGUIER	Marc	aiguier@lami.univ-evry.fr
AMAR	Patrick	pa@lri.fr
ARAUJO	Natalia	naraujo@vjf.cnrs.fr
BANERJEE	Anirban	banerjee@mis.mpg.de
BERNOT	Gilles	bernot@epigenomique.genopole.fr
BIKARD	David	dbikard@pasteur.fr
BOHLIN	Jon	jon.bohlin@veths.no
BOLOTIN-FUKUHARA	Monique	bolotin@igmors.u-psud.fr
BORKOWSKI	Wojciech	borkowsk@samba.iss.uw.edu.pl
BOTTANI	Samuel	bottani@paris7.jussieu.fr
BOURGEOIS	Ronan	ronan.bourgeois@crg.es
BRONNER	Gisele	gisele.bronner@univ-bpclermont.fr
CALVEZ	Benoît	benoit.calvez@ibisc.univ-evry.fr
CARRERA	Javier	jacarmon@etsii.upv.es
CARRERA	Javier	jphelp83@hotmail.com
CHEN	Chih-Chun	chihchun_chen@yahoo.co.uk
CHETTAOUI	Chafika	chafika.chettaoui@ibisc.fr
COMET	Jean-Paul	jp.comet@ibisc.univ-evry.fr
COREL	Eduardo	ecorel@genopole.cnrs.fr
COULON	Antoine	antoine.coulon@liris.cnrs.fr
CROS	Marie-Josée	cros@toulouse.inra.fr
D'ALCHÉ-BUC	Florence	florence.dalche@ibisc.fr
DELAPLACE	Franck	Franck.Delaplace@ibisc.fr
DESMEULLES	Gireg	desmeulles@enib.fr
ESSABBAH	Mouna	mouna.essabbah@ibisc.fr
FLUTRE	Timothée	timflutre@gmail.com
FOSTER	james	foster@uidaho.edu
FROMENTIN	Jonathan	jonathan.fromentin@irccyn.ec-nantes.fr
GRONDIN	Yohann	jg69@le.ac.uk
HAEGEMAN	Bart	bart.haegeman@gmail.com
HAREL-BELLAN	Annick	ahbellan@vjf.cnrs.fr
HEAMS	Thomas	thomas.heams@agroparistech.fr
HENEGAR	Corneliu	corneliu@henegar.info
JARAMILLO	Alfonso	Alfonso.Jaramillo@polytechnique.fr
JORG	Thomas	thomas.jorg@gmail.com
JUNIER	Ivan	i.junier@gmail.com

KÉPÈS	François	kepes@epigenomique.genopole.fr
KHALIS	zohra	zkhalis@epigenomique.genopole.fr
KOSTRZYÑSKA-MIERZWA	Lidia	rwmierzwa@poczta.onet.pl
LARIVIÈRE	Damien	damien.lariviere@fourmentinguilbert.org
LEE	Jong	jongsang@phya.snu.ac.kr
LEGALL	Pascale	legall@lami.univ-evry.fr
MABROUKI	Mbarka	mabrouki@ibisc.fr
MANCENY	Matthieu	matthieu.manceny@ibisc.univ-evry.fr
MARIE	Beurton-Aimar	aimar@labri.fr
MARTIN	Christine	martinc@limsi.fr
MATEUS	Daniel	daniel.mateus@cea.fr
MAZAT	Jean-Pierre	jpm@u-bordeaux2.fr
MAZURIE	Aurelien	amazurie@pasteur.fr
MESTIVIER	Denis	mestivier@ijm.jussieu.fr
MOLINA	Franck	fmolina@cpbs.univ-montp1.fr
MONCION	Thomas	tmoncion@ibisc.fr
MORIARTY	John	jm20@proba.ucc.ie
NONNE	Nora	nonne@vjf.cnrs.fr
NORRIS	Vic	victor.norris@wanadoo.fr
OZ	Aia	aia.oz@weizmann.ac.il
PERES	Sabine	sabine.peres@u-bordeaux2.fr
PINNA	Guillaume	gpinna@vjf.cnrs.fr
POLLET	Nicolas	pollet@epigenomique.genopole.fr
POUDRET	Mathieu	mathieu.poudret@ibisc.univ-evry.fr
RIALLE	Stéphanie	stephanie.rialle@cpbs.univ-montp1.fr
RODRIGO	Guillermo	guirodta@hotmail.com
ROLLAND	Thomas	thom_rolland@yahoo.fr
SEDWARDS	Sean	sedwards@cosbi.eu
SPICHER	Antoine	antoine.spicher@ibisc.univ-evry.fr
TAHI	Fariza	fariza.tahi@ibisc.univ-evry.fr
THYBERT	David	david.thybert@univ-rennes1.fr
TIKHONOVA	Irina	itikhonova@ucdavis.edu
TRONCALE	Sylvie	sylvie.troncale@ibisc.fr
VANDENBUNDER	Bernard	Bernard.Vandenbunder@ibl.fr
WANG	Shaoxiao	wang@moulon.inra.fr
ZÁMBORSZKY	Judit	z.judit@gmail.com
ZEHRAOUI	Farida	zehraoui@ibisc.univ-evry.fr
ZEMIRLINE	Abdallah	zemirlin@univ-brest.fr