# SYSTEMS BIOLOGICAL APPROACHES FOR UNDERSTANDING SPORULATION MECHANISMS OF *BACILLUS SUBTILIS*

A DISSERTATION

SUBMITTED TO THE SCHOOL OF FUNDAMENTAL SCIENCE AND

TECHNOLOGY OF KEIO UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

MINEO MOROHASHI

2007

# ACKNOWLEDGMENTS

First and foremost, I would like to thank my parents, to whom I dedicate this thesis, for bringing me up the way they did, and for having faith in the choices I have made. Without their support, I could not carry out this thesis project. I also really thank my brother Tomo, my sister Mari, and their family for their continuous support and help.

I am fortunate to have had chance to carry out my project under the supervision of Prof. Kotaro Oka, and as a principal advisor of this thesis. His optimistic and encourage motivated me whenever I was struggling to proceed. Prof. Hiroshi Yanagawa, Prof. Yasubumi Sakakibara, Dr. Rintaro Saito also help me a lot with fruitful discussions and helpful suggestions, and critical reading on my thesis.

I am most grateful to Prof. Yuichiro Anzai's generosity, by which I could start working on systems biology field, while other lab members were working on robotics. Dr. Hiroaki Kitano, who invited me to the world of systems biology, is for sure one of fantastic scientists I have ever met. He generously offered me great working environment when I was working at ERATO Kitano Symbiotic Systems Project. He provided me invaluable support at every step along the way. In addition, his sense of design made me sensible to those areas as well. I thank Ms. Yukiko Matsuoka, Ms. Chie Ushiwata, Ms. Mine Shioiri, who have provided me relaxed time while working.

I am greatly indebted to Dr. Yoshiaki Ohashi, a colleague in Human Metabolome Technologies. His visionary mind and attractive characters inspired me in many ways. His hard commit to work, yet attractive, made my paper accomplished in great manner.

I would also like to thank co-authors of my paper; Dr. Hamid Bolouri, Prof. John Doyle, Dr. Mark Borisuk, Dr. Amanda Winn, Ms. Kaori Shimizu, Dr. Junji Abe, Prof. Hirotada Mori, Ms. Saeka Tani, Mr. Kotaro Ishii, Prof. Mitsuhiro Itaya, Dr. Hideaki Nanamiya, and Prof. Fujio Kawamura.

Prof. Tomoyoshi Soga and Prof. Masaru Tomita provided me wonderful chance to work in a state-of-art field – metabolomics. While I work in Human Metabolome Technologies, their perspective and support helped me to go forward my project.

Along this long, long, long thesis project (almost seven years), I have been supported by many friends and colleagues; Ms. Nanae Mimura, Drs. Akira Funahashi, Noriko Hiroi, Tomomi Kimura, Theo Sabisch, Mike Hucka, Koji Kyoda, Shugo Hamahashi, Hiroki Ueda, Yasushi Hiraoka, Ayumu Yamamoto, Ding Da-Qiao, Martin Robert, Richard Baran, Masahiro Sugimoto, and Prof. Masatoshi Hagiwara. Their intelligence and kind support broadened my outlook to continue my thesis project.

My colleagues in Anzai Lab are also special to me; Drs. Sotaro Shimada, Nobuyuki Matsushita, Naohiko Kohtake, and Mr. Mitsuhiko Ohta. Sotaro was the only person who stayed in graduate school to get Ph.D., while all other members have left to get job. After few years, other two members have obtained their Ph.D. by chance – I am the next one following them.

Many thanks are also due to people by whom I have been supported at Human Metabolome Technologies. Mr. Takamasa Ishikawa, Mr. Hitoshi Sagawa, Mr. Seira Nakamura, Ms. Gin Maeta, Mr. Kosaku Shinoda, Mr. Atsushi Nagashima, Mr. Hajime Sato, Ms. Yuki Ueno, Ms. Mutsuko Sato, Ms. Miho Ikeda, Mr. Yuji Sakakibara, Mr. Masatomo Hirabayashi, Ms. Sumiko Kumaki, Ms. Aya Shinoda, Mr. Akiyoshi Hirayama, Mr. Kazunori Sasaki, Ms. Jun Imoto, Mr. Hideaki Murakami, Drs. Yoshihiro Ohtaki, Haruyuki Ohkishi, and Shizuo Ao.

PUBLICATIONS LIST

- **Morohashi, M., Ohashi, Y., Tani, S., Ishii, K., Itaya, M., Nanamiya, H., Kawamura, F., Tomita, M., and Soga, T.**

  Model based definition of population heterogeneity and its effects on metabolism in sporulating *Bacillus subtilis*.

  *J. Biochem*. 2007. (In press)

- **Morohashi, M., Shimizu, K., Ohashi, Y., Abe, J., Mori, H., Tomita, M., and Soga, T.**

  P-BOSS: A new filtering method for treasure hunting in metabolomics.

  *J. Chromatography A*. 2007. (In press)

- **Funahashi, A., Tanimura, N., Morohashi, M., and Kitano, H.**

  CellDesigner: a process diagram editor for gene-regulatory and biochemical networks.

  *BioSilico*, 1:159-162, 2003.

- **Morohashi, M., Winn, A. E., Borisuk, M. T., Doyle, J., Bolouri, H., and Kitano, H.**

  Robustness as a measure of plausibility in models of biochemical networks.

  *J. Theor. Biol.* 216:19-30, 2002.

# DEFINITIONS

| | |
|---|---|
| AIC | Akaike's Information Criterion |
| ANOVA | Analysis of variance |
| ATP | Adenosine 5'triphosphate |
| AUTO | A software tool for bifurcation analysis |
| CE | Capillary Electrophoresis |
| CellDesigner | A modeling tool for gene-regulatory and biochemical networks |
| IE | Intermediate enzyme |
| Java | An object oriented programming language |
| JWS | Java Web Start |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| MATLAB | A software tool for numerical analysis |
| MPF | Maturation promoting factor |
| MS | Mass spectrometry |
| ODE | Ordinary differential equation |
| PCA | Principal component analysis |
| PCR | polymerase chain reaction |
| P-BOSS | Peak filter based on orphan survival strategy |
| SBML | Systems Biology Markup Language |
| SBGN | Systems Biology Graphical Notation |
| SBW | Systems Biology Workbench |
| TOFMS | Time-of-flight mass spectrometry |
| UI | User interface |
| XML | Extensible Markup Language |

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# INTRODUCTION

*Science is organized knowledge. Wisdom is organized life.*

— Immanuel Kant

Recent biology is filled with complexity and flood of data. Since the discovery of molecular structure of DNA by Watson and Crick (Watson and Crick 1953), molecular biology has emerged as a methodology to understand biological systems from molecular viewpoint. Those approaches have enabled us to manipulate molecules in a way we would like to retrieve information out of them. Such approaches aimed primarily to know functions of each component (e.g., genes or proteins), and thus could have broaden our outlook in each. Their 'reductionist' approach is significant in listing all the parts of cells with detail function.

With the appearance of powerful computer processors and extensive data describing the mechanistic details of biological systems, there has been a shift toward 'integrated' approach – the focus is on understanding structure and dynamics (Kitano 2002). Besides, the advent of data-processing enabled high throughput data analyses, which resulted in completion of human genome sequence in 2001 (Venter et al. 2001). While those accomplishments are just a beginning toward system-level understanding of life, they are definitely significant milestones as a first step from systems biology perspectives.

What is next then?

Now that genomes over 500 species have already sequenced (e.g., human, mouse, *Drosophila*, *E. coli*, and *B. subtilis*), other -ome technologies have emerged. The primary fields of them are transcriptomics, proteomics, and metabolomics. Transcriptome is complement of mRNAs transcribed from genome, and transcriptomics refers to the study of the transcriptome using technologies of large-scale generation of mRNA expression profiles (Velculescu et al. 1997). Likewise, proteomics refers to the study of proteome (collection of proteins in the cells), and metabolomics to the study of metabolome (collection of metabolites in the cells (Soga et al. 2003; Morohashi et al. 2007)).

On one hand, systems biology is to infer knowledge from those various types of omics technologies, as mentioned above, which is literally 'integrated' approach – here we refer as "bottom up" approach. On the other hand, there is an utterly different approach, which we call as "top down" approach. The problems in biology are exacerbated by an increase in information complexity – no longer can systems be represented as isolated linear or hierarchical structures, instead we find complex interrelationships. Computer simulations can be used to study such systems, with the result that proposed models and hypothesis can be either validated or rejected. These methods can also complement experimental investigation, by testing experimentally measured data and highlighting future strategies of research. Although they complement each other, the top-down approach tends to focus on specific phenomena to understand mechanisms behind them. From the perspective, omics data is not necessary, yet only fraction of them is sufficient.

As mentioned above (see detail in next chapter as well), systems biology is diverse discipline, and one can take thousands of methodologies depending on what she/he would like to look into. As a common and significant fact, any approaches need to comprehensively utilize cutting edge measurement technologies and software infrastructure. Those technologies should be appropriately developed and well established, and also should be well linked toward efficient analyses thereafter. Such

efforts have been already underway in the world. One of them is Alliance for Cellular Signaling (AfCS, http://afcs.org), which is aiming at making large-scale measurements with the ultimate goal of creating an in-depth simulation model of cells.

Although we could now obtain large-scale and wide spectrum of data, we are still missing huge amount of components in analysis platform. We thus started to ask ourselves following three questions:

1. Can we perform more efficient analyses than before?

    - In order to facilitate systems biology research, various techniques must be employed, thus involving large amount of individual processing. We may need to convert data each time we proceed to next analyses manually. Such obstacles annoy ones to proceed in fast and cost effective manner, and also causing to speed down of research itself. We must keep in mind that any development should contribute to efficiency in research.

2. Can we obtain in-depth understanding of biological systems by employing both top-down and bottom-up approach?

    - As mentioned above, both approaches should be well linked to investigate biological systems. Those approaches will be seamlessly combined in future along systems biology research cycle (see next chapter), but we would like to know first that what is the outcome by employing both approach.

3. Can we apply our methods/tools to real cases?

    - Development of various tools/methods will speed up and facilitate our research, but at the same time we need to take care of its wide applicability to real cases. One of our aims is to provide the outcome to real cases as a “useful” one.

Those questions are simple, yet important starting point for examining systems biology research. To answer the questions, we undergo two steps of research:

1. To develop analysis platform

2. To utilize the platform upon test cases

Step 1 could enable us to evaluate question 1, whereas step 2 to evaluate questions 2 and 3. By taking on developing part of systems biology cycle, we believe that we could contribute to further analyses on systems biology field. Ultimately, using sporulation in *B. subtilis* as a case, our aim is to understand the basis for the bistable mechanisms utilizing above methods. Figure 1 illustrates the structure of this thesis.

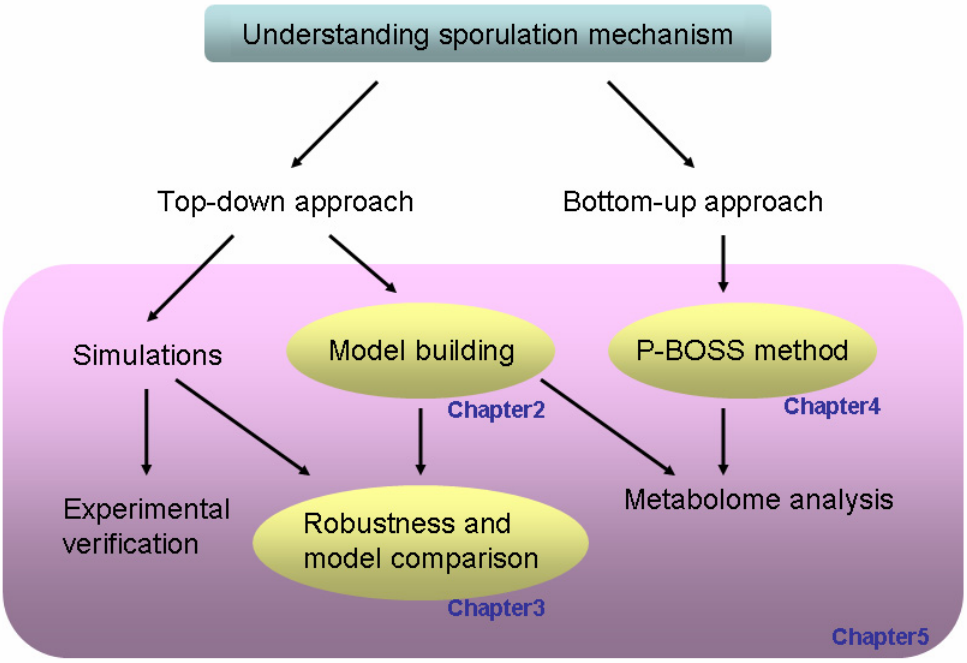


Figure 1: Structure of this thesis.

## STRUCTURE

This thesis consists of 6 chapters detailing my work. It begins with an introductory chapter that describes the motivation of research together with background information on systems biology and, in particular, simulation and metabolomics approach. Chapter 2 focuses to the development of modeling platform, which we call "CellDesigner." Chapter 3 attempts to examine simulation analysis by comparing two models of *Xenopus* using robustness as its plausibility measure. Chapter 4 shifts our focus to bottom-up approach, and describes how metabolome data processing method is developed for CE-MS based data. Chapter 5 applies above methods to examine mechanisms of sporulation in *B. subtilis*, and combines omics and model driven approach together. Chapter 6 summarizes the results of the work in previous chapters, and presents a vision for future research in systems biology field.

# CHAPTER 1: SYSTEMS BIOLOGY AND COMPUTATIONAL APPROACH

*The most incomprehensible thing about the world is that it is at all comprehensible.*

— Albert Einstein

Systems biology is defined as an approach to elucidate biological systems, such as cells, for "system-level" understanding (Kitano 2002, 2002; Hood et al. 2004). Progress in molecular and cell biology has led to the identification of complex biochemical networks involved in the normal functioning of cells, tissues and organs and even defects associated with many diseases. While those provide a complete list of factors, a building block, and relationship among each other, it is not enough to understand the system. Building them all together may lead to unexpected phenomena, because of its system characterstics – this cannot be identified by only knowing function of each factor. For instance, system may possibly cause to catastrophic status upon prescription of drug, which is because complete dynamics/kinetics of system is not understood. Those side effects are critical particularly in medical/pharmaceutical field, and thus we must examine how the individual components dynamically interact, and predict their outcome. Here comes the systems biology approach.

Figure 2 illustrates the basic research cycle of systems biology, proposed by Kitano (Kitano 2002). Although the cycle resembles that of other science field, even of biology, it is different in a way that comprises both “dry (computational)” and “wet (experimental)” experiments. It is apparent that wide spectrum of technologies is necessary to efficiently conduct the research cycle. We believe following three technologies are inevitable to go for the work:

- Experimental technologies
- Analysis technologies
- Computer technologies

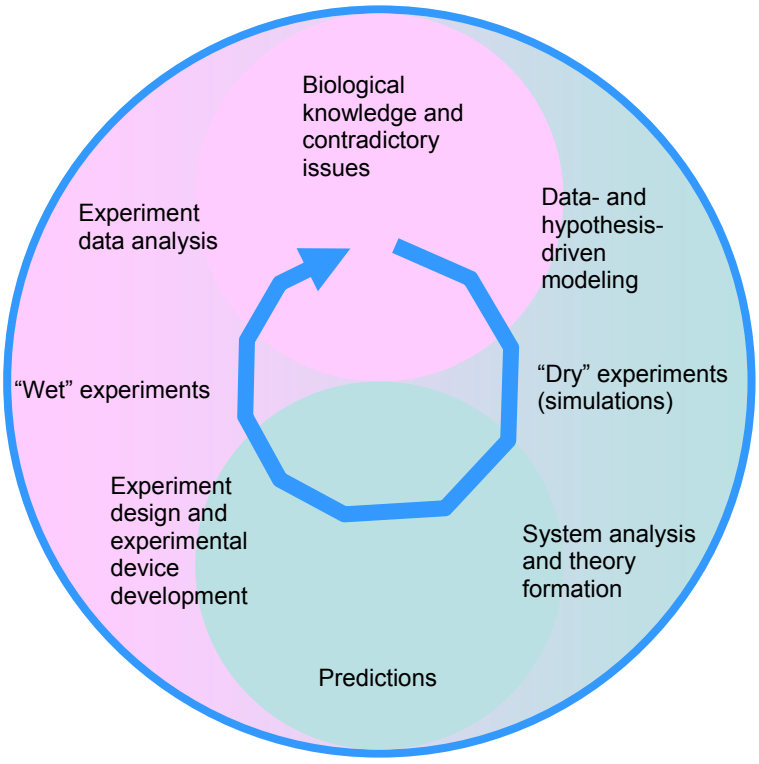


Figure 2: Hypothesis driven research in systems biology.

The image is altered from (Kitano 2002).

Here we will review each technology and discuss what is needed for further research.

# EXPERIMENTAL TECHNOLOGIES

Since the discovery of DNA structure by Watson and Crick (Watson and Crick 1953), following decades have been evolution of molecular biology field. Based on reductionism (original idea has been proposed by Dekart, a philosopher), cells were investigated by decomposing into fundamental components, particularly genes. One of traditional methods is to delete each component, and see phenotype of the mutants, comparing with wild type. Baba and colleagues have constructed whole knockout mutants of *E. coli* (Baba 2006), which could allow us to investigate the functions of components, and mechanisms of intra-cellular networks in detail. This method is plausible to construct, yet phenotype must be distinctly different from that of wild type. Silent mutations, DNA mutations that do not result in a change to the amino acid sequence of a protein, are representative phenomena.

After a half century of evolutionary progress in molecular biology, counterpart approach has appeared – holism. While approach of reductionism is aimed to investigate functions of parts of a system (a cell, in this context), holism is aimed to perform comprehensive analysis of intracellular components, or mathematical analysis to overview the mechanisms as a whole. One of the landmark projects is human genome project (Venter et al. 2001). This project has completed sequencing of human genome in few years, revealing 22,000 genes, which opened a gate to perform "omics" analysis in biological and medical fields. Currently genomes have been sequenced in more than 100 organisms. In addition to the genome, other -ome technologies have also emerged since then, e.g., transcriptome (mRNAs), proteome (proteins), and metabolome (metabolites).

One of the big differences between approach of reductionism and holism (omics analysis) is that former approach is hypothesis driven, while the latter approach is data driven. Similar to the cycle shown in Figure 2, analysis starts from biological knowledge, or observation of phenomena. Hypotheses are proposed based on the facts, and experiments are designed to verify the hypotheses. Analyses are performed, from

which hypotheses are either accepted/rejected. In case they are rejected, the data are accumulated as a feedback to design next experiments. New hypotheses are then proposed again, and research cycles are repeated until hypotheses are accepted. On the other hand, omics approach starts from measurement of data. Since omics data are quite large scale, ranging in order of thousands to ten thousands, data analysis is essential part in the analysis. Detail analysis will give an idea of hypothesis, from which additional experiments are designed. The rest of the cycles will be similar to those of former one. The key idea of omics approach is to overview the data from macro viewpoint. Without the data, appropriate hypotheses cannot be proposed, from micro viewpoint, unless many facts are accumulated upon certain targets. Recently, Ishii and colleagues have carried out variable omics analysis, which, in turn, were then combined, revealing robustness of *E. coli* in broad sense (Ishii et al. 2007). This kind of omics technologies must be taken carefully, because it contains massive data, and thus could easily lead to misunderstanding of the results (as an example, see Chapter 5).

## ANALYSIS TECHNOLOGIES

Systems biology is tightly coupled with mathematical analyses. In order to elucidate the complex mechanisms of cells, various computational and mathematical analyses are indispensable. In particular, control theory is expected to boost revealing fundamental mechanism of intracellular dynamics (Kitano 2004). There are lots of feedback loops exist in cellular networks, which seem to control stability of a biological system – in other words, robustness or homeostasis. Employing idea of control theory from engineering field, a number of applications have been investigated (Barkai and Leibler 1997; Becskei and Serrano 2000; Yi et al. 2000; Csete and Doyle 2002; El-Samad et al. 2005).

Bifurcation analysis is another tool to investigate the dynamics of a system in detail. Because of an intertwined network, complicated dynamics could emerge in most cases. Bifurcation analysis enables us to unravel the complication by decomposing huge parameter space to small spaces. Borisuk and colleagues have performed a detail investigation on *Xenopus* cell cycle model (Borisuk 1997; Borisuk and Tyson 1998), from which we extended the analysis to the comparison of two cell cycle models (this thesis). It only allows to perturb limited number of parameters at the same time (generally two at most), yet still useful to figure out the system dynamics. Sensitivity analysis might be another tool to be used for similar purpose (for example see (Ma and Iglesias 2002)).

Other than applying control theory to biological systems, which is focusing on dynamics aspect of the systems, topological analysis of networks have been well investigated. A scale-free network is a representative term describing tendency of topology in the Web, which was initially proposed by Barabasi (Barabasi and Albert 1999). In their study, some network nodes had many more connections than the average – Barabasi and colleagues called such highly connected nodes "hubs." In physics, such right-skewed or heavy-tailed distributions often have the form of a power law, i.e., the probability $P(k)$ that a node in the network connects with $k$ other nodes was roughly proportional to $k^{-\gamma}$, and this function gave a roughly good fit to their observed data. The idea has then been applied to intracellular networks, such as metabolic pathways (Jeong et al. 2000; Barabasi and Oltvai 2004). The works have been extensively investigated, some of which can be found in (Tanaka et al. 2005; Tanaka et al. 2005). The application to metabolic pathways, however, should be treated carefully, because definition of network connectivity could easily alter the results and explanation (Arita 2004, 2005).

## COMPUTATIONAL TECHNOLOGIES

Computer science plays a fundamental role in various aspects of systems biology research. It has wide spectrum of application area, from modeling to simulations, reverse engineering, visualization, parameter optimization, and database development.

Processing of computational/mathematical analyses needs extensive computing power. Development of high performance computing is thus necessary to proceed the systems biology research. Parallel computing such as "grid computing" is a solution to provide large-scale number of PCs to exhibit extraordinary performance.

- Folding@home (http://folding.stanford.edu)
- SETI@home (http://setiathome.berkeley.edu)

Above two are the examples, in which more than 1 million PCs join the project in former one. Other than that, ordinary super computers (e.g., Blue Gene in IBM) may play a significant role in extensive data processing.

Modeling and simulations are one of the hot topics in systems biology research. While those activities have been up for more than a decade, recent advances in development of software technologies and platform allow us to work in the field more extensively. An example is development of model exchange format, as represented by Systems Biology Markup Language (SBML, http://sbml.org) (Hucka et al. 2003), or BioPAX (http://biopax.org). Those formats have been designed to exchange computer models among various type of software tools, including simulators, databases. There are two possible approaches to develop software tools; one for integrating all functions and capabilities to handle by itself, and the other for communicating among various tools. The formats are for the latter purpose, and the approach seems to work well so far, being supported by over 100 tools. A reason why the former approach is not common (or even popular) is that there are still yet to overcome various issues, such as establishment of modeling theory for intracellular networks. While many attempts

have accomplished great results, which have also been verified experimentally, it can be applied to small range of areas, and still difficult to establish general theories (such as establishing theory of dynamics of gene expression). Without establishing those theories, various tactics or methodology must be tried – it would then be feasible to have a huge software platform which has all possible functions to handle data, and to perform simulations and various analyses.

Note that although simulations have been employed in many biological studies recently, they do not hold the all answers. Often, when a complex system is simulated, the results are equally difficult to interpret, depending on what question we are trying to answer. We may be able to demonstrate that a given model reproduces the experimentally observed behavior, but we may not understand why – in other word, what features of the model are responsible for the behavior of the system. For this reason, conventional methods of mathematical analysis may, at times, be more appropriate.

The term “reverse engineering” is the process of discovering the technological principles of a device or object or system through analysis of its structure, function and operation. It often involves taking something (e.g., a mechanical device, an electronic component, a software program) apart and analyzing its workings in detail, usually to try to make a new device or program that does the same thing without copying anything from the original. Employing the idea, biological systems are needed to be reverse engineered so that each unit (or module in other words) is to be investigated. At least unless relationship of wiring information being obtained, no further analysis can be carried out –input/output information only can tell nothing, but a just black box of the system. This approach can be readily applied to omics-based data, because omics data exhibit one aspect of a system in comprehensive manner. Since its progress in DNA chip, or microarray technologies (known as transcriptomics), there have been vast demand on reverse engineering. While there are primarily two type of data, time-series data, or steady-state data, some approaches can be found in following paper (Liang et al. 1998; Morohashi and Kitano 1999; Ideker et al. 2000; Kyoda et al. 2000; Kyoda et al. 2004).

It is indispensable to have sophisticated database for searching information on specific species (e.g., genes, proteins, and metabolites). Depending upon type of species and organisms, there are large numbers of database publicly available, some of which are as follows:

- The GDB Human Genome Database (http://www.gdb.org)
- *Saccharomyces* Genome Database (http://www.yeastgenome.org)
- Human Protein Reference Database (http://www.hprd.org)
- Reactome (http://www.reactome.org, reaction database)
- Brenda (http://www.brenda.uni-koeln.de, enzyme database)
- PubChem (http://pubchem.ncbi.nlm.nih.gov, small molecules database)
- KEGG (http://www.genome.jp/kegg)
- BioCyc (http://www.biocyc.org)

Yet, those databases are not fully curated, because of lack of data, or lack of coverage. KEGG, one of most comprehensive database in the world, has advantages in covering various type of information, from genes to metabolites to proteins, but still lacking in data – only half of metabolites have been assigned for *E. coli* (Ohashi, personal communication).

Bio-IT companies are interested in providing more sophisticated and more curated database. To cite some of them,

- MetaCyc (GeneGo, http://www.genego.com)
- Ingenuity Pathway Analysis (Ingenuity Systems, http://www.ingenuity.com)
- PathArt (Jubilant Biosys, http://www.jubilantbiosys.com)
- PathwayStudio (Ariadne Genomics, http://www.ariadnegenomics.com)

The former 3 products are based on manually curated database, while the latter one employs machine learning based text mining approach to gather publication information. They have advantages in terms of providing valuable information, and way to extract information out of database (for example, combining pathway information or experimental data into database, so that more broad view of intracellular mechanisms can be obtained).

## APPLICATIONS OF SYSTEMS BIOLOGY

What would be the applications in systems biology? A big impact would be to contribute in medical and pharmaceutical field (Kitano 2007). While genome-based drug discovery has been paid huge attention as a next generation in pharmaceutical field, no other approaches seem to have been employed successfully so far. This could be because systems biology approach is still a new approach, and takes time to be validated in the field (one pipeline takes over ten years in average). The other reason might be that the field is still too immature to be applied to those fields, although some omics approach have for sure been applied already. Some of industrial activities are introduced in final chapter.

More feasible applications are for basic and fundamental research field. As Kitano proposed in (Kitano 2002), there are diverse fields to aggregate, most of which are still in mature. This is more like interdisciplinary field, and needs wide variety of knowledge and technologies to put in, not only biology (such as molecular biology, genetics, cell biology), but also computer science, mathematics, physics, chemistry, and engineering. To be able to advance systems biology research, each field must be well established to successfully apply to systems biology field. Although this may take enormously long time to establish, it should allow us to investigate in much more fast and accurate manner, leading to principle of biological systems – ultimately to control them.

# CONCLUSION

We have introduced various technologies and methodologies, as a part of systems biology research. While there are huge efforts being performed, we are still underway to fully utilize or establish methodology of systems biology. As each individual relevant technology advances, we believe to be able to perform comprehensive analysis and application toward various fields, such as medical and pharmaceutical fields. Our work should be a big step for the systems biology approach from both computational and analytical viewpoint.

# CHAPTER 2: CELLDESIGNER: DEVELOPMENT OF GENETIC/BIOCHEMICAL NETWORK EDITOR

*If you want to understand life,*

*don't think about vibrant, throbbing gels and oozes,*

*think about information technology.*

— Richard Dawkins

Understanding of logic and dynamics of gene-regulatory and biochemical networks is a major challenge of systems biology. To facilitate this research topic, we developed CellDesigner, a modeling tool of gene-regulatory and biochemical networks. CellDesigner supports users to easily create such networks using solidly defined and comprehensive graphical representation (SBGN: Systems Biology Graphical Notation). CellDesigner is SBML compliant, and SBW-enabled software so that it could import/export SBML described documents, and could integrate with other SBW-enabled simulation/analysis software packages. CellDesigner is implemented in Java, thus it runs on various platforms such as Windows, Linux, and MacOS X.

## INTRODUCTION

While software infrastructure is one of the most crucial components of systems biology research, there has been no common infrastructure or standard to enable integration of computational resources. To solve this problem, the Systems Biology Markup Language (SBML, http://sbml.org) (Hucka et al. 2003) and the Systems Biology Workbench (SBW, http://sbw.kgi.edu) have been developed (Sauro et al. 2003). SBML is an open, XML-based format for representing biochemical reaction networks, and SBW is a modular, broker-based, message-passing framework for simplified intercommunication between applications. More than 110 (as of Jan 2007) simulation and analysis software packages already support SBML, or are in the process to support them.

Identification of logic and dynamics of gene-regulatory and biochemical networks is a major challenge of systems biology. We believe that the standardized technologies, such as SBML, SBW and SBGN, play an important role in the software platform of systems biology. As one such approach, we have developed CellDesigner (Funahashi et al. 2003), a process diagram editor for gene-regulatory and biochemical networks.

## DESIGN PRINCIPLES

Broadly classified, CellDesigner was designed according to following requirements:

- Representation of biochemical semantics
- Detailed description of state transition of proteins

- SBML compliant (SBML Level-1 and Level-2)
- Integration with SBW-enabled simulation/analysis modules
- Extreme portability as a Java application

Our aim in developing CellDesigner is to supply a process diagram editor with the standardized technology (SBML in this case) for every computing platform, so that it could confer benefits as many users as possible. By using the standardized technology, the model could be easily used with other applications, thereby reducing the cost to create a specific model from scratch. The main standardized features that CellDesigner supports could be summarized as "graphical notation", "model description", and "application integration environment." The standard for graphical notation plays an important role for efficient and accurate dissemination of knowledge (Kitano et al. 2005), and the standard for model description will enhance the portability of models between software tools. Similarly, the standard for application integration environment will help software developers to provide the ability for their applications to communicate with other tools.

### SYMBOLS AND EXPRESSIONS

CellDesigner supports graphical notation and listing of symbols based on a proposal by Kitano and colleagues (Kitano et al. 2005). The definition of graphical notation has now been developed as international community based activities called ‘Systems Biology Graphical Notation (SBGN, http://sbgn.org). Although several graphical notation systems have been already proposed (Pirson et al. 2000; Cook et al. 2001; Kohn 2001; Maimon and Browing 2001; Kohn et al. 2006), each has obstacles to become a standard. SBGN is proposed for biological networks designed to express

sufficient information in clearly visible and unambiguous way (Kitano et al. 2005). We expect that these features will become part of the standardized technology for systems biology. The key components of SBGN, which we propose, are as follows:

1. To allow representation of diverse biological objects and interactions
2. To be semantically and visually unambiguous
3. To be able to incorporate notations
4. To allow software tools to convert a graphically represented model into mathematical formulas for analysis and simulation
5. To have software support to draw diagrams
6. The notation scheme to be freely available

To accomplish above requirements for the notation, Kitano (Kitano et al. 2005) firstly decided to define a notation by using process diagram, which graphically represents state transitions of the molecules involved. In the process diagram representation, each node represents state of molecule and complex, and each arrow represents state transition among states of a molecule. In the conventional entity-relationship diagrams, arrow generally means “activation” of the molecule. However, it confuses semantic of the diagram as well as limiting possible molecular processes that could be represented. Process diagram is more intuitively understandable definition than the entity-relationship diagram – one of the reasons is that the process diagram could be explicitly represented as a temporal sequence of events which entity-relationship cannot. For example, a process of MPF factor activation in cell cycle, kinase such as Wee1 phosphorylates residues of Cdc2 that is one of the components of MPF (Figure 3: A process diagram representation of MPF cycle.). However, MPF is not yet activated by this phosphorylation. If we use an arrow for activation, we cannot properly represent the case. In the process diagram, on the other hand, whether a

molecule is “active” or not is represented as a state of the node, instead of “arrow” symbol for activation. Promoting and inhibition of catalysis are represented as a modifier of state transition using a circle-headed line and a bar-headed line, respectively.

While process diagram is suitable for representing temporal sequence, either process diagram or entity-relationship approach could be used, depending upon the purpose of the diagram. Both notations could actually maintain compatible information internally, but differ in visualization (Kitano et al. 2005). We propose, as a basis of SBGN, a set of notation that enhances the formality and richness of the information represented. The symbols used to represent molecules and interactions are shown in Figure 4.

The goal of SBGN is to define a comprehensive system of notation for visually describing biological networks and processes, thereby contributing to the eventual formation of a standard notation. For such a graphical notation to be practical and to be accepted by the community, it is essential that software tools and data resources to be made available. Even if the proposed notation system satisfies the requirements of biologists, lack of software support will drastically decrease its advantages. CellDesigner currently supports most of the process diagram notation proposed, and will fully implement the notation in the near future.

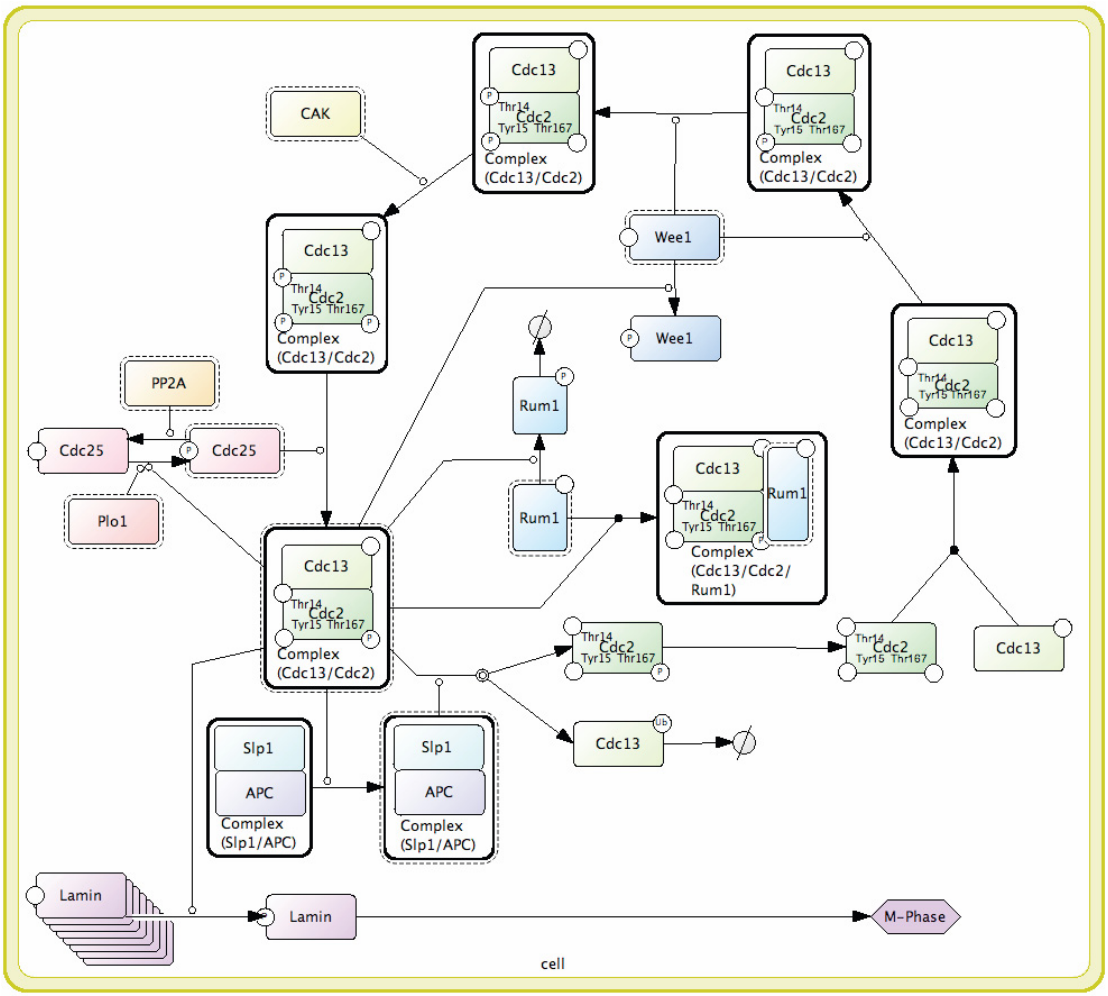


Figure 3: A process diagram representation of MPF cycle.

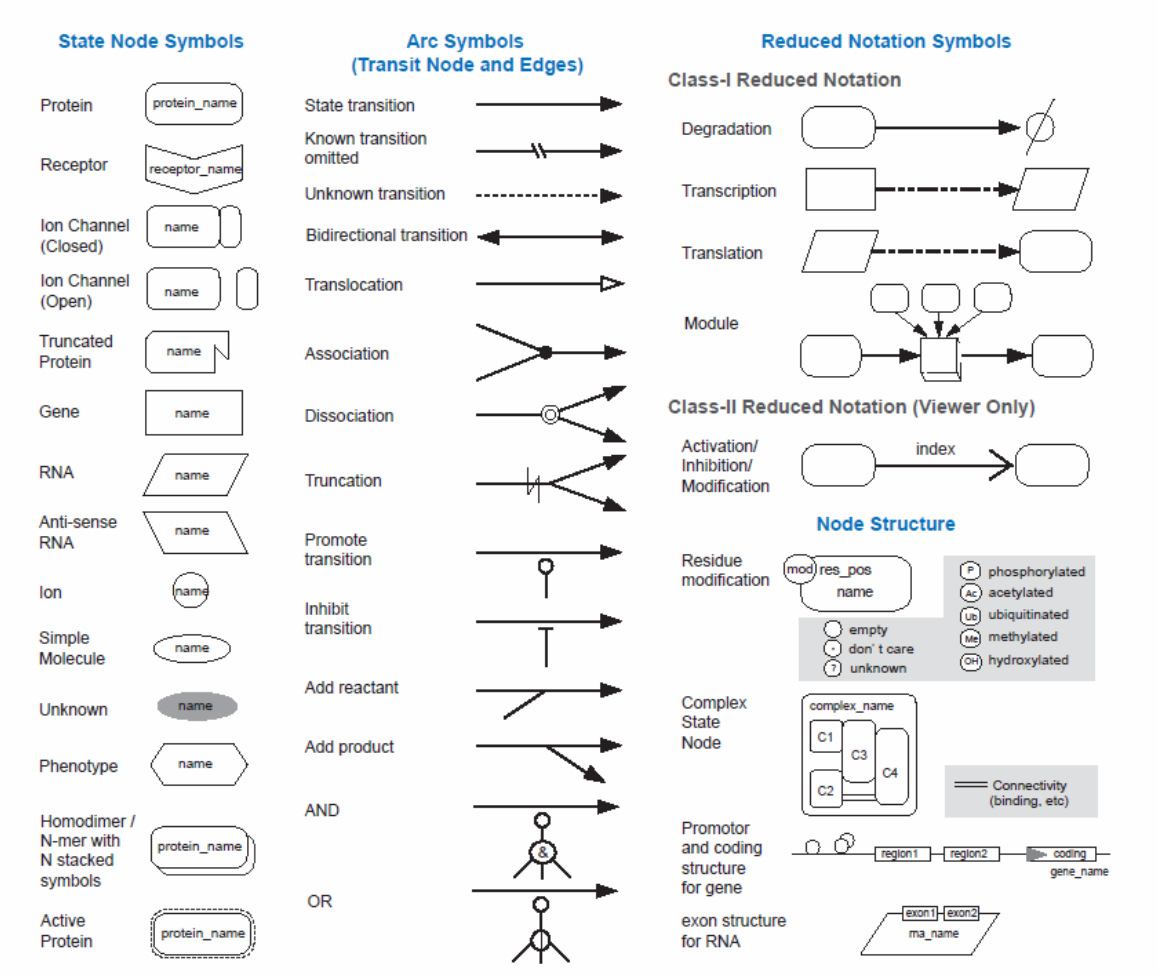


Figure 4: Proposed set of symbols for representing biological networks.

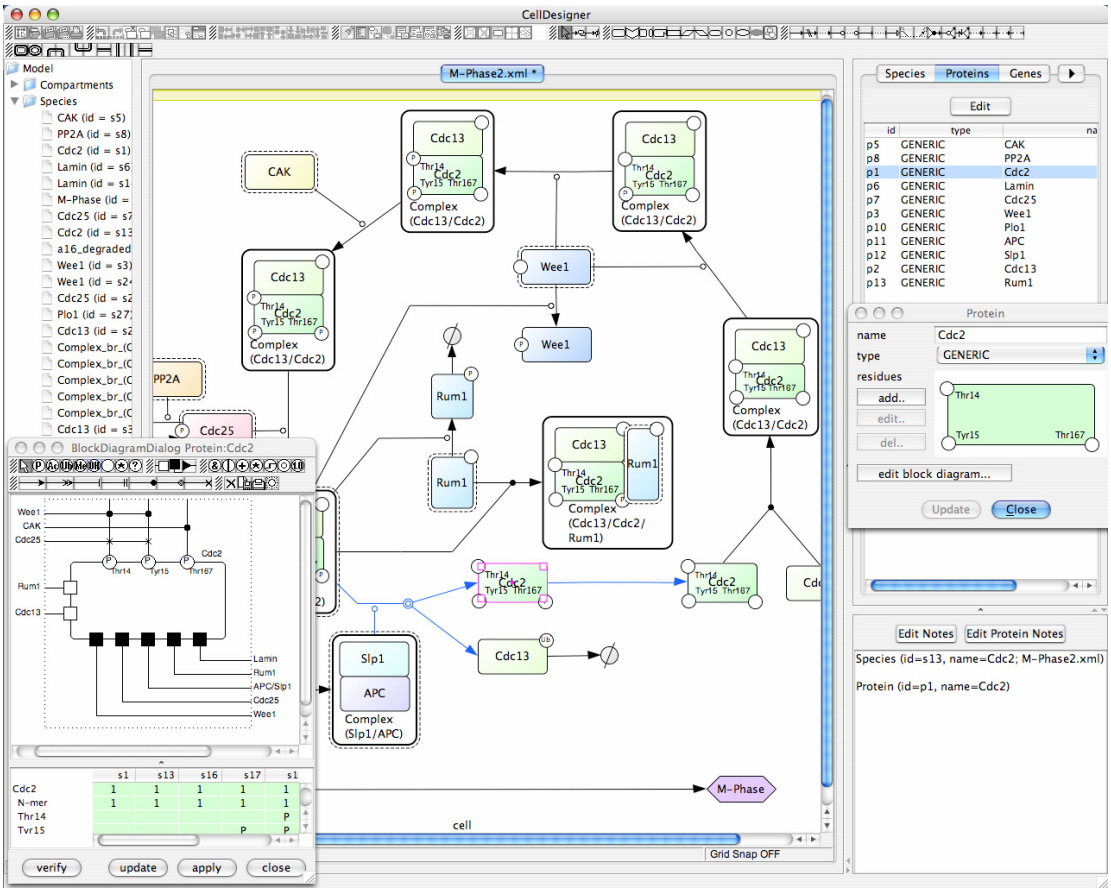


Figure 5: Screenshot of CellDesigner.

Main panel includes various panes, such as species list (left-sided), pathway (center), block diagram (Cdc2 in this case), and notes (right-sided).

## SBML COMPLIANT

CellDesigner is an SBML-compliant application – it supports SBML reading and writing capabilities. SBML is a tool-neutral, computer-readable format for representing models of biochemical reaction networks, applicable to metabolic networks, cell-signaling pathways, gene regulatory networks, and other modeling problems in systems biology. SBML is based on XML (eXtensible Markup

Language), a simple, flexible text format for exchanging a wide variety of data. The initial version of the specification was released on March 2001 as SBML Level-1. The most recent released version of SBML is Level-2 Version 2. Currently, SBML is supported by over 110 software systems and widely used. CellDesigner uses SBML as its native model description language, and thus once a user create a model by CellDesigner, all information inside the model will be stored in SBML and the model could be used by other software systems without any conversion of the model. As mentioned, CellDesigner draws a pathway with its specialized graphical notation. Since such layout information has not been supported by SBML, CellDesigner stores its layout information under "annotation" tag, which does not conflict with current SBML specification. There is a working group of layout extension for SBML, and will be incorporated to SBML Level-3. We are currently underway to implement a conversion module to export SBML layout extension from CellDesigner. CellDesigner has an auto layout function so that it could read all SBML Level-1 and Level-2 documents whether the model contains layout information or not. By using this function, users could use existing SBML models such as KEGG, BioModels database, and so forth. We have converted more than 12,000 metabolic pathways of KEGG to SBML (the pathways are available from http://systems-biology.org/). Other SBML models are available from the BioModels Database (http://www.ebi.ac.uk/biomodels/). We could also use our own SBML models created by CellDesigner on other SBML compliant applications (http://systems-biology.org/001/).

## SUPPORTED ENVIRONMENT

CellDesigner is implemented in Java, and could run on many platforms that support JRE (Java Runtime Environment). Currently CellDesigner runs on the following platforms:

- Windows (98SE or later)
- MacOS X (10.3 or later)
- Linux (Fedora Core 4 or later)

The current version of CellDesigner requires JRE1.4.2 or higher, and X Window System for UNIX platforms.

## EXPORTING CAPABILITY

Since CellDesigner is supposed to be a “design tool” for representing gene regulatory and biochemical networks, the pathways described by CellDesigner should be easily used in various situations. CellDesigner could thus export the pathways in various formats – currently in JPEG, PNG and SVG format.

HOW DOES IT WORK?

Building models with CellDesigner is quite straightforward. To create a model, the user selects "New" from "File" menu, inputs the name of an SBML document – a new canvas will then appear. The user could then place a species, such as a protein, gene, RNA, ion, simple molecule and so forth. A new window will appear asking the name of the species. The size of each species could be changed by clicking and dragging the corner of species. The user could also define the default size of each species from "Show Palette option" from the “Window” menu. Species could be moved by dragging and dropping

To draw reactions, a type of reaction should first be selected from the UI buttons, and a reactant species then clicked, followed by a product species. To add more reactants, the user could select "Add reactant" button, and then choose species and reaction.

As mentioned above briefly, the modeling process with CellDesigner is straightforward steps, which should not cause users any confusion.

CellDesigner could also represent common types of reactions, such as catalysis, inhibition activation and so forth. The procedure for representing such reactions is just as same as adding reactants or products to an existing reaction; that is, to select a species (modifier), followed by a reaction. The user could also easily edit the symbols for proteins with modification residues, and hence, could describe detailed state transitions between species of an identical protein by adding different modifications.

The models are stored in an SBML document, which contains all the necessary information referring to species, reactions, modifiers, layout information (geometry), state transitions of proteins, modification residues and so forth. These SBML models could be used on other SBML-compliant applications.

If users want to run simulation based on the SBML model, select Simulation menu, which, in turn, calls SBML ODE Solver directly. The Control Panel appears, enabling users to specify the details of parameters, to change amount of specific species, to conduct parameter search, and to run simulation interactively. To conduct time evolving simulation, users may need to know basics of the SBML specification (See http://sbml.org for detail).

If users select SBW menu, on the other hand, CellDesigner passes the SBML data to the SBML compliant tools via SBW, while you need to set up SBW before you invoke SBW connection.

## WHAT DISTINGUISHES CELLDESIGNER'S TECHNOLOGY FROM OTHERS CURRENTLY AVAILABLE?

Currently, many other applications exist that include pathway design features. The advantages of CellDesigner over other pathway design tools could be summarized as follows:

- Based on standard technology (i.e., SBML compliant and SBW enabled),
- Supports clearly expressive and unambiguous graphical notation systems (SBGN), which is aimed at contributing to eventual standard formation

- Runs on many platforms (e.g., Windows, MacOS X, Linux)

As described above, the aim of the development of CellDesigner is to supply a process diagram editor with standardized technology for every computing platform, so that it will benefit as many biological researchers as possible. For instance, tools such as E-Cell (Tomita et al. 1999) is SBML-compliant, and tools such as Cytoscape (Shannon et al. 2003) runs on multiple platforms

These tools are powerful in some aspects and they are not intended to support the features as CellDesigner. Some of them have the facility to create pathways, and some also include a simulation engine or database integration module. CellDesigner does include a simulation engine provided by SBML ODE Solver development team, and also it could connect to other SBW-enabled applications so that user could switch the simulation engine on the fly. Furthermore, we have been converting existing databases to SBML (e.g., KEGG), and all SBML-compliant applications could easily be browsed, edit the models, and even simulate via CellDesigner.

The overriding advantage of CellDesigner is that it uses open and standard technologies. The models created by CellDesigner could be used on many other (over 110) SBML compliant applications and its graphical notation system will make the representation of models in more efficient and accurate manner.

## FUTURE WORK

In future release of CellDesigner, we plan to implement further capabilities. Improvement of auto layout function is a big issue – the bigger (e.g. > few hundreds of nodes) the network diagram becomes, the slower the performance of CellDesigner

becomes, which causes our current version not to align each nodes and edges quite well. Integration with other modules is also underway, such as other simulation, analysis and database modules. Current version of CellDesigner has been implemented as a Java application, while we are developing a JWS (Java Web Start) version of CellDesigner so that it could be used as a web-based application as well.

To be widely used from biologists to theorists, we believe that it is essential to meet the standard. We are thus actively working as SBML and SBGN working group members, which aims to establish de facto standards in systems biology field – former one seems to have already become de facto as model description language. SBML Level-3 (next version) will include layout extension, and we will incorporate the functions in our new release of CellDesigner. BioPAX (http://ww.biopax.org) is another big activity, which tries to connect widely distributed data resources seamlessly. We also plan to connect CellDesigner with BioPAX data format so that users could use CellDesigner from BioPAX platform and vice versa.

From software development perspectives, providing API, plug-in interface or open source strategy might be a solution to speed up the development, and enable users to customize the software depending on users needs. While we have been providing binary program of CellDesigner so far, we are now working to extend our development scheme in such manner.

We wish CellDesigner to be used by anyone who is working on biology-related field. As described throughout this manuscript, CellDesigner is designed to be user-friendly as much as possible, thus allowing users to draw pathway diagrams quite easily as drawing with other drawing tools, such as Microsoft Visio, or Adobe Illustrator. Since our proposed notation itself is rigidly defined, the diagrams could be used for

presentation or even for knowledge base – the diagrams could be used as figures in manuscript, or pathway representation of databases. Since definition of the pathway diagram notation is now getting much attention, which has now resulted to form an SBGN working group (http://sbgn.org), we hope the notation will be much refined as a de facto standard representation, which will be reflected in the representation manner of CellDesigner as well.

Our concept for developing CellDesigner is "easy to create a model, to run simulation and to use analysis tools." This will be achieved by extending the development of corresponding native libraries or SBW-enabled modules. Improvement of the graphical-user interface is also required, including the mathematical equation editor, so that the user could easily write equations by selecting and dragging a species.

## CONCLUSION

We introduced CellDesigner, a process diagram editor for gene-regulatory and biochemical networks based on standardized technologies and with wide transportability to other SBML-compliant applications and SBW-enabled modules. Since first release of CellDesigner, 12,000 downloads has been already accomplished. CellDesigner also aims to support standard graphical notation. Since the standardization process is still underway, our technologies are still changing and evolving. As we are in partnership with SBML, SBW, and SBGN working groups, we will go through with these standardization projects and hence improve the quality of CellDesigner.

# CHAPTER 3: SIMULATION ANALYSIS OF CELL CYCLE MODEL OF *XENOPUS*

*Only those who dare to fail greatly can ever achieve greatly.*

— John F. Kennedy

Theory, experiment, and observation suggest that biochemical networks which are conserved across species are robust to variations in concentrations and kinetic parameters. Here, we exploit this expectation to propose an approach to model building and selection. We represent a model as a mapping from parameter space to behavior space, and utilize bifurcation analysis to study the robustness of each region of steady-state behavior to parameter variations. The hypothesis that potential errors in models will result in parameter sensitivities is tested by analysis of two models of the biochemical oscillator underlying the *Xenopus* cell cycle. Our analysis successfully identifies known weaknesses in the older model and suggests areas for further investigation in the more recent, more plausible model. It also correctly highlights why the more recent model is more plausible.

# INTRODUCTION

In recent years, a series of landmark papers have reported the existence of robust behaviors in a variety of biochemical networks (Alon et al. 1999; von Dassow et al. 2000; Yi et al. 2000). Indeed, robustness in metabolism (Fell 1997), the cell cycle (Borisuk and Tyson 1998), and inter-cellular signaling (Freeman 2000) is now widely accepted. Of course, nothing can be robust to absolutely all variations. Some variations may not matter in terms of the functionality of the system in question. For example, the process that specifies the geometric relationship between hair follicles on human heads need not be very exact or robust. Nor is there any guarantee that all biological systems are necessarily optimally organized. A well-known example of this is the apparently inverted layered structure of the human retina. In this paper, we are interested in robustness to variations in kinetic parameters. That biochemical networks will exhibit robustness to variations in their kinetic parameters was theoretically predicted long ago (Savageau et al. 1972; Kacser and Burns 1973). However, these issues have recently received more widespread attention (Hartwell et al. 1999; Dearden and Akam 2000) due to the growing need to understand the large volumes of data produced by the emerging biotechnologies.

While we tend to think primarily of functionally distinct cellular processes such as metabolism, or the cell cycle, the reality is that all cellular processes are highly interrelated and involve not only biochemical interactions, but also mechanical, electrophysiological, and other interdependencies across multiple time and space scales. Nonetheless, “if we are to comprehend [molecular biology], we must hope that it can be dissected into a series of modules or networks which can be studied in relative isolation” (Dearden and Akam 2000).

Recent discoveries of modular interspecies conserved networks suggest that such hope may not be in vain. The fact that such networks perform homologous functions

with similar but differing proteins (hence different reaction rates) and in different cellular contexts (hence different total concentrations of chemical species) suggests functional robustness to such variations.

The chemical oscillator underlying the control of cleavage-stage cell divisions in *Xenopus* embryos is a well-known example of a robust biochemical module: its component proteins can be replaced by proteins from other species (e.g. human) without affecting its function, and its oscillatory behavior can be reproduced in vitro (Murray and Hunt 1993). In this paper, we compare two models of the *Xenopus* cell cycle oscillator to evaluate the feasibility of using robustness as a means of identifying potential weaknesses in models. Our approach extends the use of bifurcation analysis for model evaluation by Ringland (Ringland 1991) and Clarke (Clarke 1980, 1994) to include observations about the shape, smoothness, and other features of behavior regions in parameter space. The results suggest that the approach can help with iterative development of increasingly detailed models of cellular processes, and selection between alternative explanations (models) of experimentally observed phenomena.

## MATERIALS AND METHODS

The analytical solution of the parameter space for the two-equation version of the 1991 model was derived using Maple (Maplesoft, Ontario, Canada). All other numerical characterizations of the parameter spaces of the two models were performed using the AUTO bifurcation analysis package (Ermentrout 2002). The frequency contour plots were generated as co-dimension two bifurcation plots on which the frequency of oscillation was superimposed post hoc. Oscillation frequencies were calculated by sampling the oscillatory region of each plot in a 100x 100 or a 50x50 grid, grouping the results into bins, and then using the AUTO to trace the loci of each frequency bin. Numerical parameter optimizations were carried out interactively using Berkeley Madonna (http://www.berkeleymadonna.com/).

# RESULTS

## WHAT SHOULD BIOCHEMICAL NETWORKS BE ROBUST TO?

It would be impractical and undesirable for systems to be equally robust to everything. For example, a system should be sensitive to particular types of variation in its inputs, otherwise, it would not respond to anything. On the other hand, there is also no reason to believe that all cellular processes will be optimally robust to everything. In this section, we delineate where one can expect robustness or sensitivity and discuss the implications.

To begin, we define a biochemical model as a mapping from parameter space to behavior space. The structure of a network is given by the set of all non-zero elements in its stoichiometry matrix (i.e. the set of interactions in the network). The parameters define reaction kinetics and total (initial) concentrations of the chemical species constituting the modeled network. Two types of parameters may be noted:

(A) Parameters whose values vary during the lifetime of an individual (e.g. temperature, regulated gene activity level, or amount of a protein in a particular state).

(B) Parameters that are constant for individuals, but variable across individuals/species (e.g. reaction rate constants ($k_{cat}$; $k_m$), initial/total concentrations).

Any "parameter" that does not vary across individuals or across species is considered a constant here. Inputs are parameters that control the system state trajectory. The inputs to a network can be type A or type B parameters. Sensitivity to type A inputs is useful for behavioral adaptation, while sensitivity to type B inputs can generate diversity in populations without loss of function.

Carlson & Doyle (Carlson and Doyle 2000) have proposed that robustness to common variations is achieved at the cost of added system complexity. The additional complexity will generally incur some new sensitivity. Optimally robust systems are those that achieve a useful balance between robustness to frequent variations and the concomitant sensitivity to some rare events. A corollary of this view is that natural systems tend to be highly robust to frequently occurring variations and, in counterbalance, fall catastrophically when some rare variations occur. We exploit this observation to say that if a model of a robust system (e.g. a conserved biochemical network) exhibits sensitivity to a parameter $p$; one of the following must hold (see also (Alves and Savageau 2000)):

(1) $p$ is a control input; in that case the model should be sensitive to $p$: The type of sensitivity will depend on the functionality of the modeled network. Systems that switch between a finite numbers of states tend to be sensitive to the level of inputs, but not the exact value of any input. On the other hand, systems with continuous outputs (e.g. an amplifier) tend to be sensitive to the exact value of the input(s).

(2) $p$ is regulated (held constant) elsewhere in the system. A familiar example from engineering is power supply provision in electronic circuits: sub-circuits depend critically on receiving a supply voltage held constant by dedicated circuitry. An analogous biochemical example may be the provision of metabolic "services" in cells.

(3) $p$ is not regulated, but the system as a whole is insensitive to $p$ (e.g. soot buildup in a heater will tend to affect heater performance, but not room temperature). In that case, the modeled network is actually a part of a larger system and should be studied in this larger context.

(4) We have misunderstood the function of the network. For example, suppose a system is designed to provide pressure and temperature compensation signals to other systems on an aircraft. We might model the network as only a pressure compensator, and then discover that it is also sensitive to temperature. In such a case, it is not that our model of pressure compensation is wrong, but rather that we have misunderstood the full function of the system.

(5) The model structure is incorrect (e.g. there may be missing components, or incorrect interactions between existing components).

It is often possible to guess whether a model parameter may be a control input from the nature of the processes it controls. For example, the rate of transcription of a gene, the rate of synthesis of a protein, and the initial concentration of a maternally inherited factor are all parameters which are often controlled by upstream biochemical processes and which can usefully control processes such as developmental cell fate specification.

On the other hand, enzyme-mediated reaction rates vary widely among individuals and species (Eanes 1999), so any biochemical network whose function is conserved across individuals and species may be expected to be highly robust to variations in reaction rates. Similarly, variations in total concentrations of locally synthesized chemical species should not affect the behavior of a structurally correct model dramatically.

When a biochemical model exhibits sensitivity to some of its parameters, one of conditions (1)-(5) above must hold. One may then investigate each possibility in turn. However, sensitivity and robustness are not “all or none”, binary characteristics. Below, we define quantitative measures that allow more exact characterization of the type and extent of sensitivity/robustness exhibited. This greater resolution in turn provides greater insight into the potential cause of the observed sensitivity, as illustrated by our example analysis of models of the *Xenopus* cell cycle.

## MEASURING ROBUSTNESS AND SENSITIVITY

Consider an example system with only two parameters $P_1$ and $P_2$. Suppose the system has a steady state which can be characterized by a single variable, say a concentration level, or an oscillation frequency. Two-parameter bifurcation plots delineate the range of $P_1$ and $P_2$ for which the system exhibits the measurable behavior. Figure 6 shows two example behavior loci for such a system. The figure is drawn such that the colored regions in (A) and (B) are roughly equal in area. The crosses represent example operating points, that is, the mapping from the particular values of $P_1$ and $P_2$ to a particular value for the measurable system characteristics. The arrows show the effect of example variations (noise) in $P_1$ on the location of the operating point. The model in (A) has two important features:

(1) Define the minimum distance between an operating point and the boundary of the behavior locus as the stability margin (SM) of the operating point. The optimum stability margin (OSM) of the model is then defined as the maximum stability margin achievable by judicious placement of the operating point. The OSM is greater for the convex locus in (A) than for the concave locus in (B). Moreover, the sum of all stability margins is greater for (A) than for (B). Therefore, the model exhibiting characteristic (A) has greater overall stability than the model exhibiting characteristic (B).

(2) For the particular drawings in this example, we note that the rate of change of the measured characteristic with changes in $P_2$ is lower in (A) than in parts of (B). Which of the two models is more plausible depends on the extent of behavioral variability observed experimentally.

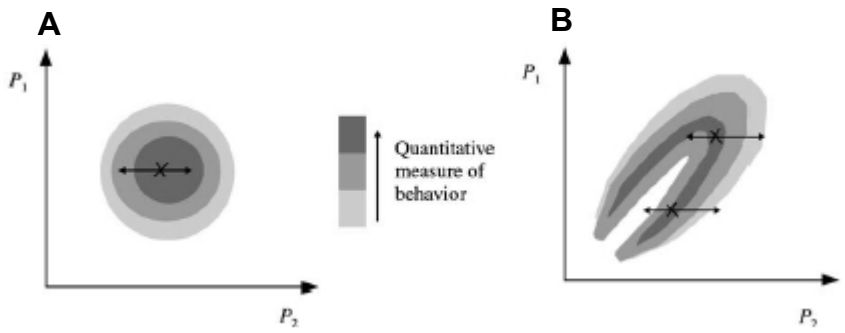


Figure 6: Schematic representation of two example behavior loci.

The symbol 'x' represents example operating points, mapping from the particular values of $P_1$ and $P_2$ to a particular value for the measurable characteristics. As the operating points shift from the points, quantitative measure of the behavior changes – sensitivity to the variation of parameters greatly affect behavior depending on the shape of loci.

Where a modeled system exhibits multiple steady state behaviors, there will be one or more loci for each behavior in parameter space and it would be necessary to consider issues such as (1) and (2) (above) for each locus. Often, the multiple behaviors exhibited by a model border each other. Clearly, in such cases convexity of one region would imply concavity in the neighboring region(s). In such cases (as for example in the cell cycle models below), optimum robustness for all model behaviors requires that the boundaries between behavioral regions in parameter space be flat (i.e. neither concave nor convex). The boundaries between neighboring behavior regions are parameter bifurcation loci and can be computed and plotted in two dimensional slices for visual assessment. For examples, see our cell cycle oscillator analysis below.

For parameters acting as state switch (control) inputs, once a system has switched states, it should be robust to small variations ("noise") in the input signals, i.e., we require large stability margins for each switched state. Finally, we use Ockham's Razor to distinguish between any two models which may match experimental

observations equally well: the model with the greatest parameter robustness – as defined by the above considerations – is the more plausible.

In the remainder of this chapter, we explore the above ideas by applying them to two well-known models. Where Ringland (Ringland 1991) and Clarke (Clarke 1980, 1994) used bifurcation analysis to obtain models capable of exhibiting experimentally observed steady-state behaviors, we start with models that meet steady-state experimental observations in some qualitative manner (in the examples below, both models produce two cell cycle arrest states and an oscillatory state whose frequency is close to observations). We analyze and compare models on the basis of the size, shape and degree of variability within each steady-state behavior region.

## CASE STUDY: THE *XENOPUS* CELL CYCLE OSCILLATOR

To illustrate and demonstrate the above concepts, we use two models of the cell cycle oscillator that regulates cleavage in early *Xenopus* embryos (Tyson 1991; Marlovits et al. 1998). Both models were developed by Tyson and colleagues, and replicate the wild-type *in vivo* and *in vitro* oscillatory behavior and arrest states well. Indeed, at this superficial level they are not distinguishable. The earlier model was essentially theoretical (Tyson 1991). Its structure is abstract and some interactions within it do not correspond to specific chemical reactions. It was written before experimental data on the structure and kinetics of the system were available. The later model has experimentally validated structure; most of its kinetic parameters have experimentally measured values, and correctly predict the phenotypes of a large range of experimental interventions (Marlovits et al. 1998). With the benefit of hindsight, the limitations of the older model are known. We compare the dynamics of the two models to demonstrate the manner in which robustness analysis can highlight important systematic differences between structurally correct and incorrect models.

## WHY USE THE CELL CYCLE AS OUR EXAMPLE CASE STUDY?

The cell cycle oscillator is highly conserved in all eukaryotes (Murray and Hunt 1993), so here is good reason to believe it is robust to small mutations. There are several additional reasons for our choice.

(1) The basic dynamics observed *in vivo* in *Xenopus* embryos can also be reproduced in vitro using cytoplasmic extracts. There is also no growth during cleavage stages, so growth directed control of the cell cycle, or other unknown cellular processes are not necessary to explain the fundamental features of the *Xenopus* cell cycle oscillator.

(2) *Xenopus* eggs are large and the embryos lend themselves well to experimental analysis. There is therefore a wealth of experimental evidence used by Tyson and colleagues to ensure the plausibility of the more recent structurally detailed model.

(3) Known defects in the earlier model have been experimentally pinpointed.

(4) Analytic solutions of the parameter space are obtainable for the simpler earlier model.

(5) In an extensive study, Borisuk (Borisuk 1997) and Borisuk & Tyson (Borisuk and Tyson 1998) fully characterized the multidimensional parameter space of the later, more complex model, thus providing unique insights into its behavior as a mapping from parameter space.

## OVERVIEW OF THE TWO MODELS

Figure 7 presents an overview of the behavior of cell cycle determinants in *Xenopus* eggs and embryos. The concentration of an active form of a cyclin – CDC2 dimer –

known as the maturation promoting factor (MPF) – controls cell division activity. The regulation of active MPF concentration is the subject of the two models studied here. Prior to fertilization, active MPF levels are arrested at low concentration in immature eggs and at high concentration in mature eggs. At fertilization, after an initial delay, a series of 12 equal-period, synchronous cell divisions ensue. Thus the system has three steady state behaviors: low MPF arrest, high MPF arrest, and oscillations in MPF concentration.

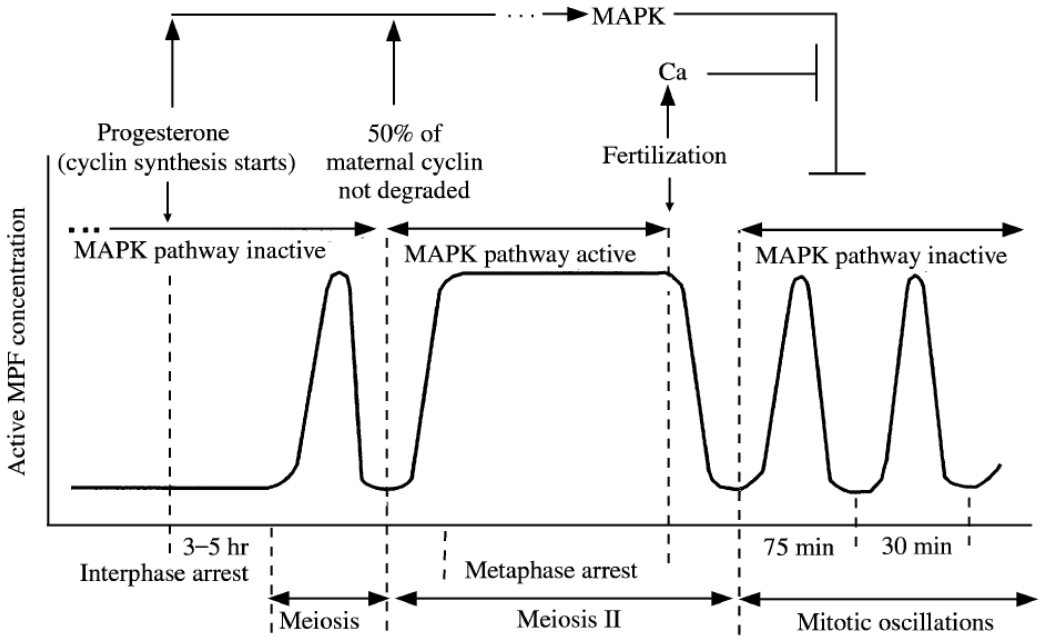


Figure 7: Schematic representation of major events in *Xenopus* eggs and embryos. Note the role of the MAP-kinase-mediated pathway that blocks active MPF degradation (and hence oscillations) until after fertilization (see text for further description).

Figure 8 (A) and (B) are schematic representations of the two models. Both models are based on a cyclic set of reactions involving cyclin-CDC2 dimerization, followed by phosphorylation/dephosphorylatlon and a positive feedback loop which creates hysteretic dynamics. However, the models are otherwise different. In particular, the positive feedback on active MPF is modeled phenomenologically in the 1991 model.

In the 1998 model, on the other hand, the positive feedback loop is defined in terms of a set of specific molecular interactions discussed later and shown in Figure 15. In addition, the 1998 model includes another feedback loop through which active MPF promotes its own degradation. Both of these added structures turn out to have a significant impact on the robustness of the network behavior as discussed below.

## CHARACTERSTICS OF THE 1991 MODEL

The full 1991 model requires six equations and 10 kinetic parameters. But as Tyson showed in 1991, to a good approximation, the model can be reduced to two equations and four kinetic parameters. As illustrated in Figure 9, the system has three operating regimes corresponding to cell cycle arrests in immature and mature eggs (low and high MPF levels, respectively), and an oscillatory regime corresponding to the cleavage cycles in early embryos. The bifurcation loci between the three behavioral regions can be characterized analytically. Figure 10 (A) and (B) show the variations in the shape and size of these three operating regions as a function of the values of the four kinetic parameters of the system. The surfaces at the boundaries between these regions represent bifurcation loci in parameter space.

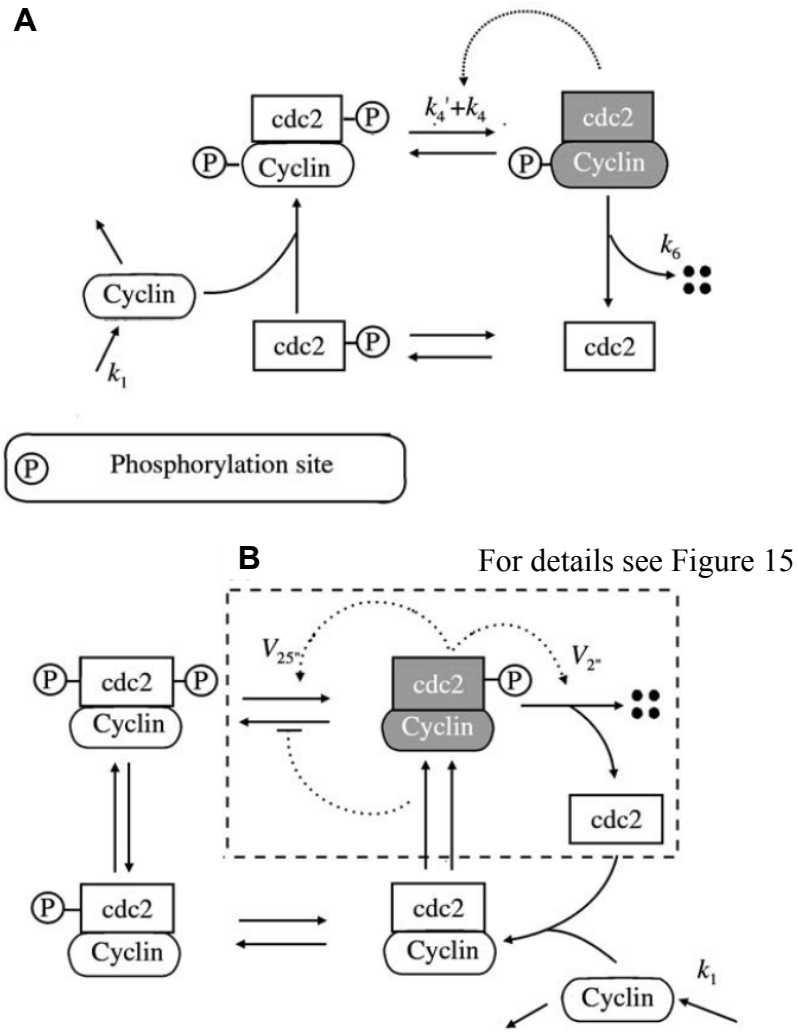


Figure 8: Schematic representations of two models of the *Xenopus* cell cycle.

Both models share a basic reaction loop in which cyclin dimerization with CDC2 is followed by a series of phosphorylation/dephosphorylation events. (A) The 1991 model: at that time, details of the (de)phosphorylation events were not known and were hypothesized. Moreover, the mechanism underlying the positive feedback of active MPF (gray-filled dimer) on its own production was not known and was only modeled phenomenologically. $k_1$ is the rate of cyclin synthesis. The rate of active MPF formation is modeled as the sum of two components: $k_4$ is the high rate of active MPF formation proportional to active MPF concentration. $k_4'$ is the low rate of active MPF production proportional to inactive MPF concentration. $k_6$ is the rate of dimer breakdown. (B) The 1998 model: the dimerization and (de)phosphorylation sequence of events have been corrected and the single positive feedback effect of MPF on itself has been replaced by three feedback mechanisms (dotted arrows) each of which is modeled as a set of detailed molecular interactions (see Figure 15 for details), $k_1$; $V_{25}''$; and $V_2''$ correspond to $k_1$; $k_4$ and $k_6$; respectively, in the 1991 model.

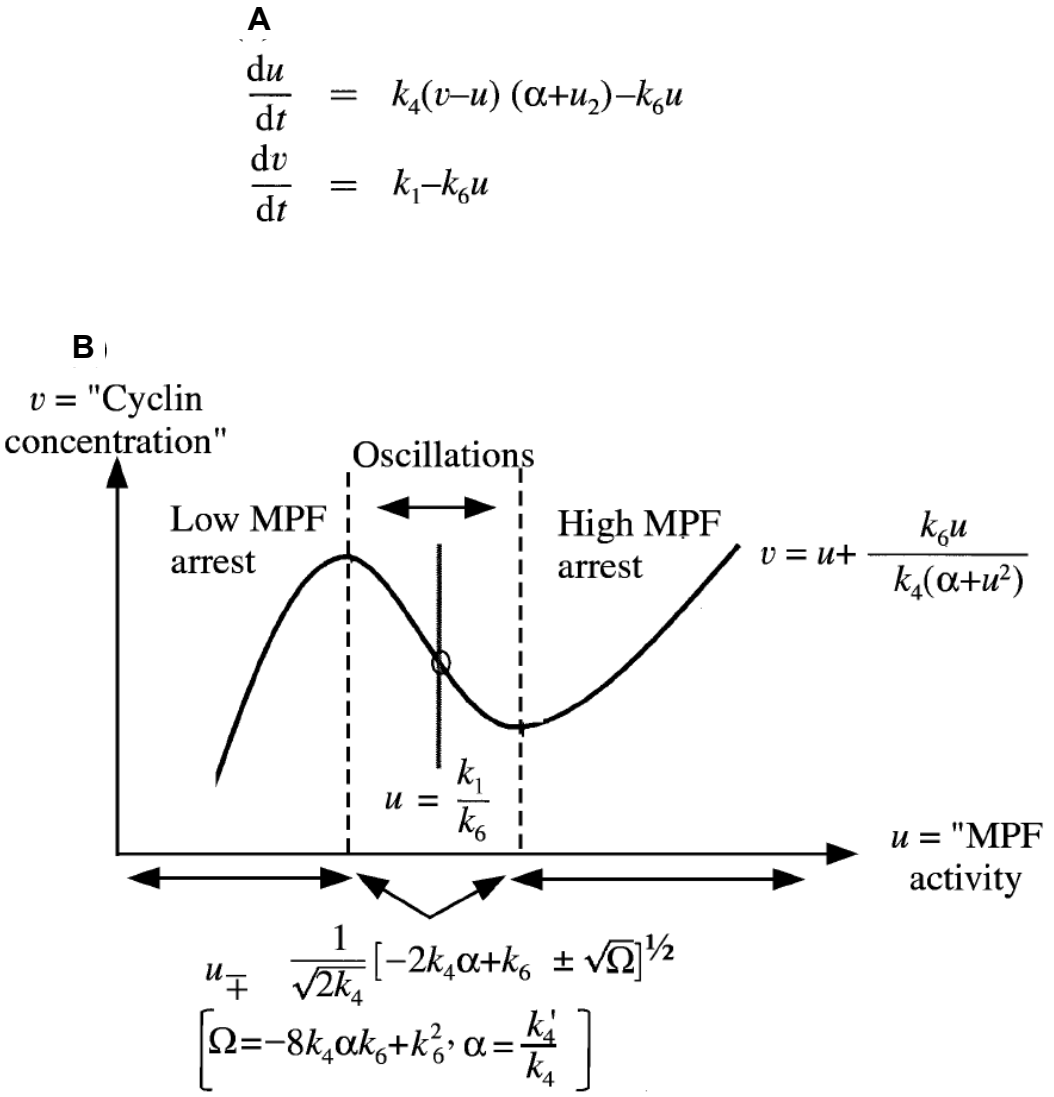


Figure 9: Overview of the reduced, two-equation version of the 1991 model.
(A) The two-equation, four-parameter model. (B) Phase portrait of the two-equation model. The v nullcline is a vertical line whose location is given by parameters $k_1$ and $k_6$: The u and v nullclines cross only once, giving a single steady state that is either stable (cell cycle arrest states to the left and right of the maximum and minimum of the u nullcline), or unstable (oscillations corresponding to repeated embryonic cell divisions, region between the arrest regions). The loci of the boundaries between these three behavioral regions can be derived analytically from the nullcline equations and are shown below. This allows exhaustive characterization of the model behavior as a function of its four kinetic parameters (see Figure 10, Figure 11 and text).

Because the reduced 1991 model has only four kinetic parameters and is amenable to analytic exploration, we were able to exhaustively plot its behavior in parameter space. As the example in Figure 10 (A) illustrates, the model's three regions of steady-state

behavior in any two-parameter plot are broad regions with approximately flat boundaries indicating robustness to parameter variations. This is not true for plots involving the rate of cyclin synthesis ($k_1$). For example, the $k_4$–$k_1$ plot in Figure 10 (B) shows that the system behavior depends critically on the value of $k_1$. Note how changing the value of $k_4$ affects the choice of $k_1$ for which the system is in any one particular steady state (seen most readily in the sharp curvature of the boundaries of the oscillating region).

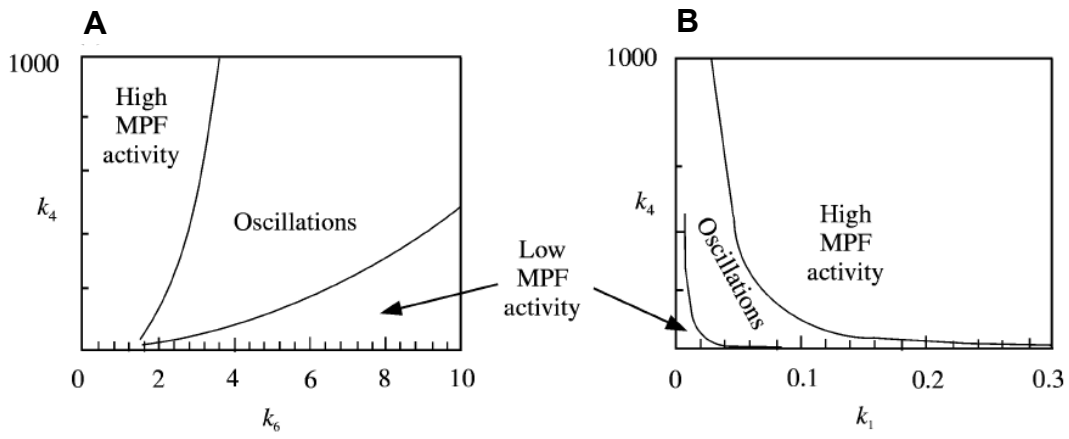


Figure 10: Two-parameter plots showing the regions in parameter space.

(A) The region between the two curves corresponds to repeated cell divisions in embryos, the area left of this corresponds to the high active-MPF arrest state of mature eggs, and the right hand region to the low active-MPF arrest state of immature eggs. (B) The orientation is reversed. Except for $k_4$–$k_1$ plots, as in (B), the characteristics in (A) are typical of all other plots: three approximately equal regions separated by roughly flat boundaries, as would be expected for optimal robustness to parameter variation, (B) demonstrates the nonlinear dependence of system behavior on $k_1$:

The observation that the system is sensitive to $k_1$ is not surprising: the oscillatory behavior of the system can be shown to depend on the steady-state concentration of cyclin, which in turn depends on $k_1$ and $k_6$: In vivo, control of cyclin concentration is achieved through a dual control mechanism consisting of (a) the regulation of cyclin synthesis and (b) the activity of a MAP-kinase-mediated pathway which acts as a binary switch, blocking active-MPF (and hence cyclin) degradation until after fertilization (see (Ferrell and Machleder 1998)). Ferrell Jr et al. (Ferrell et al. 1991)

and Groisman et al. (Groisman et al. 2000) discuss experimental evidence of the role of cyclin synthesis in the control of the cell cycle. Therefore, we focus on sensitivity to $k_1$ rather than $k_6$:

The rate of cyclin synthesis also exerts a strong control on the size of the three regions. With high values of $k_1$ – Figure 11(D) – the arrest state for mature eggs dominates. So long as $k_1$ is high, the system is highly robust to variations in the values of the other three parameters. At the opposite extreme, when $k_1$ is small – Figure 11(A) – the size of the regime corresponding to cell cycle arrest in immature eggs is by far the biggest. So with $k_1$ very small, the immature egg cell cycle arrest state is very robust to variations in the other three kinetic parameters. As the value of $k_1$ is varied from very low to very high, we see that the size of the middle region (cleavage oscillations) first grows – Figure 11(B) – and then shrinks again – Figure 11(C). Figure 11(B) shows an example value for $k_1$ that results in a very wide oscillatory region occupying most of the parameter space. So with this value, the cell undergoes cleavage oscillations in a manner highly robust to variations in the other three parameters. It is now known that the *Xenopus* egg inherits large amounts of maternal cyclin that enables the two meiotic divisions of the egg prior to fertilization. Mitotic oscillations prior to fertilization are prevented by a MAP-kinase-mediated biochemical switch (see the cartoon illustration in Figure 7 and (Ferrell and Machleder 1998)). The sensitivity of the 1991 model's behavior to $k_1$ reveals the role of $k_1$ as a control input for the mitotic oscillator, acting to generate the capacity for oscillations which are later triggered by fertilization (the biological case for control of the embryonic cell cycle by cyclin synthesis was first put forward by Murray & Kirschner (1989) and Murray et al. (Murray and Kirschner 1989). The nonlinear ($k_4$-related) dependence of the system behavior on $k_1$ reveals a weakness in the model: the state of the system cannot be predicted from the value of the control input ($k_1$) alone.

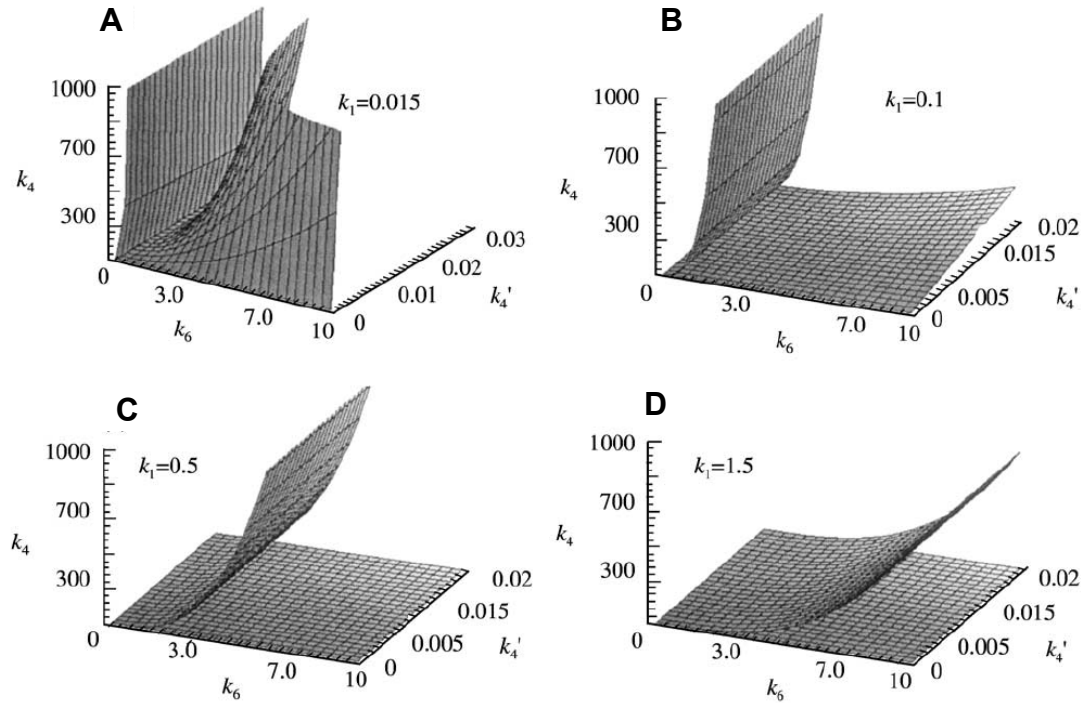


Figure 11: The effect of $k_1$ on the shape of the model behavior in parameter space. This suggests that the rate of cyclin synthesis may be a state control input for the cell cycle oscillator. (A) For low values of $k_1$; the region to the right of both planes (low active-MPF immature-egg arrest) occupies most of the volume of the parameter space. So when $k_1$ (rate of cyclin synthesis) is low, immature egg cell cycle arrest is highly robust to variations (noise) in the values of the other system parameters ($k_4$; $k_4'$; $k_6$). (B) For intermediate values of $k_1$ (here 0.1), the oscillatory region dominates the parameter space and oscillatory behavior is highly robust to changes in the other system parameters. (C) As $k_1$ is increased further, the size of the region corresponding to high active-MPF mature-egg cell cycle arrest grows. (D) For high values of $k_1$; the region corresponding to high active-MPF mature-egg cell cycle arrest dominates the parameter space. A cell in this state would be highly robust to variations in the other system parameters.

Note that in Figure 11, $k_4$ ranges from 0 to 1000. To be comparable to the experimentally measured values of the corresponding parameters used in the 1998 model, $k_4$ should be limited to <10. However, if we limit the value of $k_4$ to this smaller range, the robust model behavior observed in Figure 11 can only be replicated if $k_1$ is increased to values beyond its plausible range (here taken as nominal ± one order of magnitude). Thus, with the benefit of hindsight, we note that the structural weakness of a single (phenomenological) feedback loop in the 1991 model results in a need for unfeasibly large parameter ranges in the model.

Figure 12 is a plot of the oscillation frequency of the model as a function of parameters $k_1$ and $k_4$: As expected, the oscillation frequency is zero in the dark-blue regions corresponding to the two cell cycle arrest states discussed above (low- MPF immature-egg arrest to the left, high-MPF mature-egg arrest to the right). In the region in between these, the value of $k_1$ determines the cleavage oscillation frequency. The values of the other parameters are set to those recommended in Tyson (Tyson 1991). The period of the resulting oscillations ranges from 10 to 50 min: The in vivo period for *Xenopus* cleavage cycles is 30 min (Masui and Wang 1998). The in vitro period is around 60 minutes (Murray and Hunt 1993). So the model includes the observed in vivo and in vitro behaviors, but its exact oscillation period varies with changes in $k_1$: Since $k_1$ — the protein synthesis rate — cannot be controlled very tightly in vivo, this sensitivity suggests that the model has structural deficiencies.

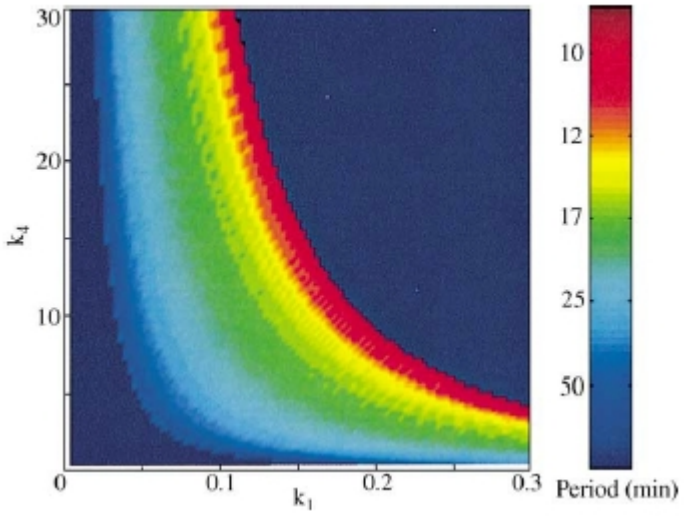


Figure 12: Contour plot of the frequency of oscillations in the 1991 model.
According to this model, the cleavage period would vary between 10 and 50 min from individual to individual, even when all "environmental conditions" are held constant. This contradicts experimental observations of a stereotypic cell division period in *Xenopus* embryos.

It is possible to optimize the model parameters to constrain the frequency range of the oscillatory region. However, in this case the oscillatory region becomes very narrow

and the sensitivity of the model to $k_1$ variations is even more pronounced. As we show below, the 1998 model does not suffer from this problem.

## CHARACTERISTICS OF THE 1998 MODEL

The full 1998 model is represented by nine differential equations and 26 kinetic parameters. It is clearly far too complicated to study analytically. We used the numerical bifurcation analysis tool AUTO (Doedel 1981) to characterize this model in the same manner as the 1991 model. Based on the earlier results of Borisuk & Tyson (Borisuk and Tyson 1998), we knew that the model has robustness characteristics similar to the 1991 model, and that $k_1$ continues to control system state. Figure 13(A) and (B) illustrate this point. In Figure 13(A), $k_1$ is set to a high value (corresponding to large amounts of maternal cyclin being present in the egg in which the degradation of cyclin is blocked by the MAP kinase pathway) and we see that virtually all of the plausible parameter space (the volume to the left of the plotted surface) is taken up by the region corresponding to high-MPF cell cycle arrest in mature eggs. In Figure 13(B), $k_1$ is reduced to 0.01 (corresponding to fertilized eggs, where the MAP kinase pathway is disabled and maternal cyclin has been degraded), and we see that now the same volume in parameter space represents oscillatory behavior (the volume between the two surfaces). Note that the above state control characteristic of $k_1$ is highly robust to variations in other kinetic parameters: changes in $V_{25}''$; $V_{25}'$; and $V_2''$ (corresponding to $k_4$; $k_4'$ and $k_6$ in the 1991 model, respectively, see Figure 8) have virtually no effect on this behavior.

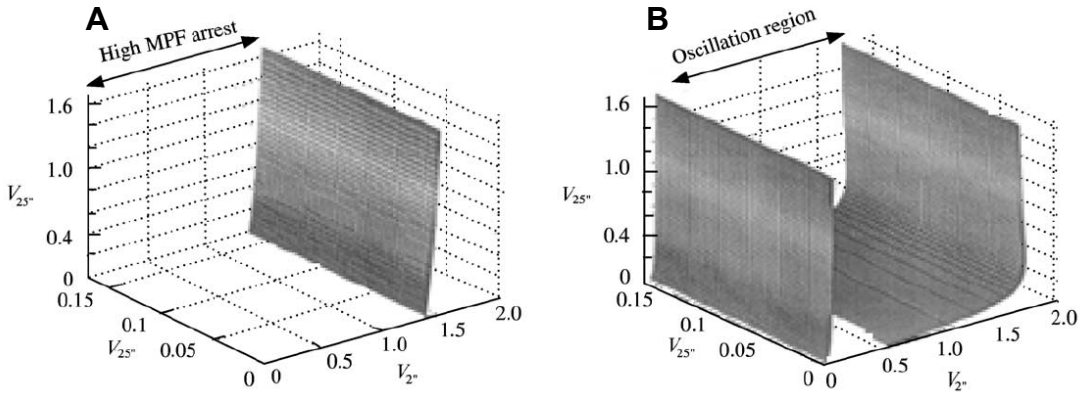


Figure 13: The effect of $k_1$ on the size/shape of the regions in the 1998 model.
The flat, nearly vertical surfaces separating the regions are close to the ideal for optimum robustness to variations in the other system parameters. Compared to the corresponding characteristics in the 1991 model (Figure 11), the 1998 regions are also much larger, thus offering greater robustness. The values of $k_1$ are (A) 1.0, and (B) 0.01, respectively.

A critical difference between the 1991 and 1998 models is that the period of oscillations is much less variable in the 1998 model. In the latter, the value of $k_1$ determines whether the system oscillates. However, the period of oscillations is fixed by the combination of the values of the other parameters in the system. Since these parameters would be expected to be constant in any individual, the cleavage period would be fixed and not vary with small fluctuations in the regulated value of $k_1$: In a sense, $k_1$ control behaves like a multi-level switch. Its value is interpreted in three discrete levels: low, medium and high. These in turn determine the mode of operation of the cell cycle engine. In comparison with the 1991 model, the 1998 model is not only less sensitive to parameters other than $k_1$; but also operates with greater stability margins on $k_1$:

As shown in Figure 14, the parameter values of the 1998 model – which are based on in vitro experimental measurements – result in 45-50 min period oscillations similar to in vitro preparations. Note, however, the existence of a triangular region of frequency instability where $k_1$ is small. Moreover, the range of $k_1$ values for which the system oscillates seemed surprisingly small to us. On further investigation, we found that the positive feedback loop through which MPF facilitates its own degradation

(shown in Figure 15(B)) is not experimentally specified. Moreover, Tyson and colleagues did not optimize the parameters of these reactions for any particular behavior, but rather used nominal values. As shown in Figure 16, optimizing these unknown parameters for oscillation periods in the in vitro range dramatically improves the robustness characteristics of the 1998 model. The oscillatory region is now much wider than that of the 1991 model. The oscillation period is remarkably constant, and $k_1$ control of cell state no longer depends on co-variation with other rates ($V_{25}''$ in the 1998 model corresponds to $k_4$ in the 1991 model). Thus, robustness analysis of the 1998 model not only highlighted a potential weakness in the model, but also pinpointed where the problem may lie and allowed us to remedy it.

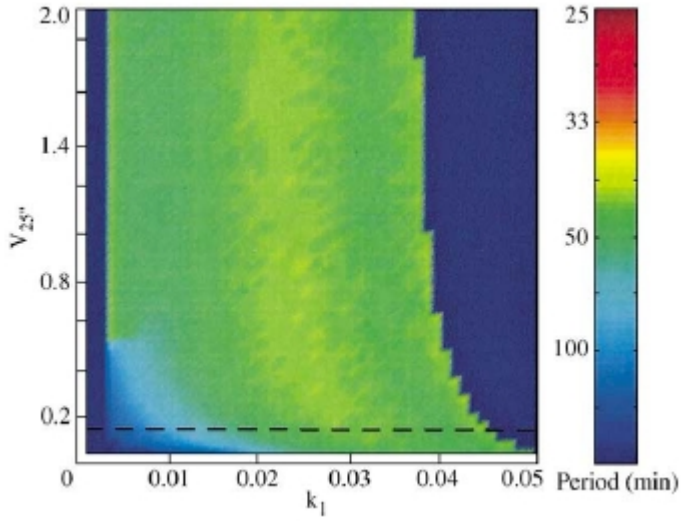


Figure 14: Cleavage frequency contour plot.
The parameter was set using the Marlovits et al. parameter values for the 1998 model. Note that the oscillatory region is very narrow, but has the advantage of a much more stable oscillation period range (45-50 min) in most of the oscillatory region. The dashed horizontal line indicates the experimentally derived value of $V_{25}''$ used by Marlovits et al. In this region of the parameter space, the oscillation period ranges from 50 down to 10 min; thus negating the apparent greater stability of the 1998 model.

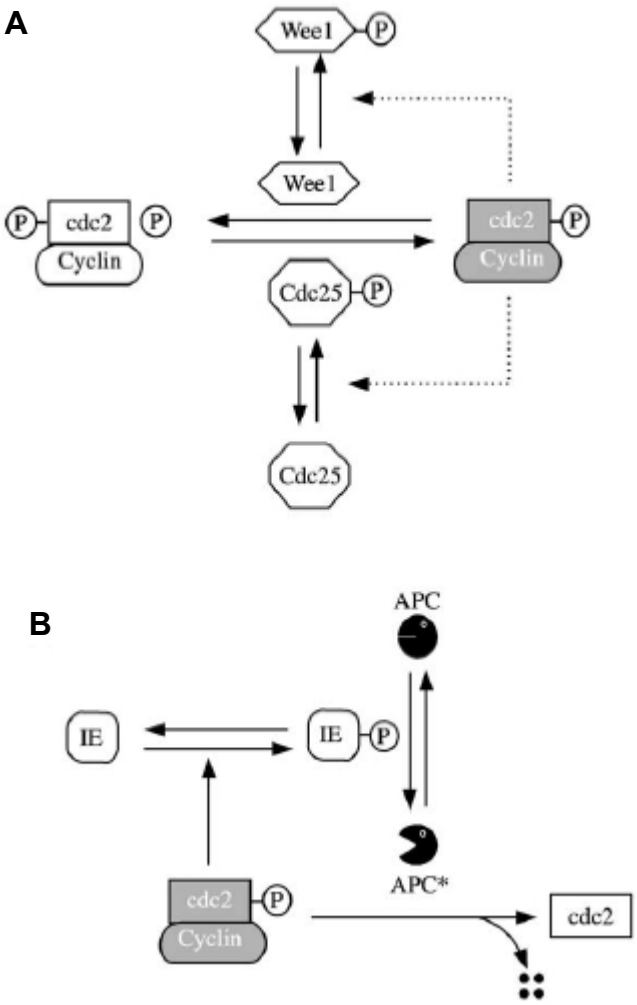


Figure 15: Details of the additional reactions included in the 1998 model.
(A) The push – pull positive feedback mechanisms replacing the simple phenomenological feedback of active-MPF on itself in the 1991 model. The CDC25 path enhances the rate of active-MPF production while the wee1 path reduces the rate of return of active-MPF (gray filled dimer) to inactive form. (B) The feedback mechanism of active-MPF on its own degradation. APC is the anaphase promoting factor, and its role in active MPF degradation has been experimentally verified. But intermediate enzyme (IE) has not been experimentally identified and its interactions represent only an abstract path.

Although we have shown that the structure of the 1998 model is capable of providing highly robust oscillatory behavior in a manner far exceeding the capabilities of the 1991 model, it cannot be assumed that the 1998 model is a complete representation of all the pertinent interactions constituting the *Xenopus* cell cycle oscillator. The

structure of the 1998 model is clearly more plausible than the 1991 model, but it could be further optimized. For example, Figure 17 shows that increasing the autocatalytic rate of active MPF degradation can enlarge the oscillatory region of the model considerably. The structure of the 1998 model ensures that the expanded oscillatory region has a very stable period (in Figure 17 optimized to lie in the range 28 to 30 minutes corresponding to in vivo oscillations). Moreover, there is no co-dependence on parameters other than $k_1$: Experimental data only put a lower bound on the value of the parameter optimized here ($V_2''$). The exact in vivo value is not known. Nor is it significant for our purposes. The important observation here is that more detailed modeling of this particular part of the model may be illuminating.

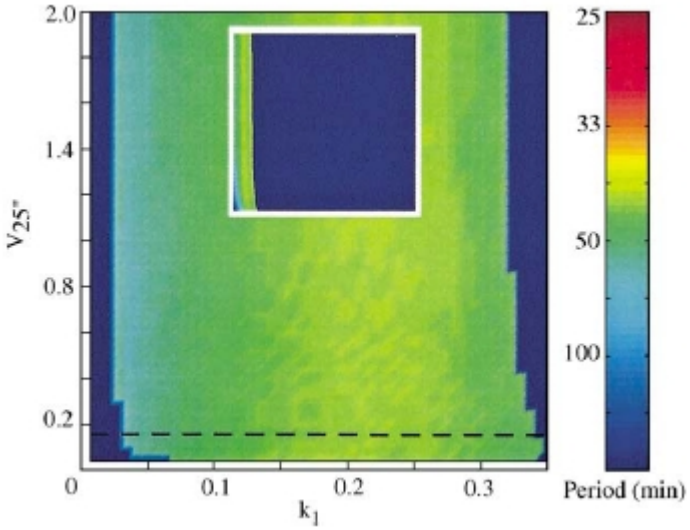


Figure 16: The 1998 model optimized to give in vitro like oscillations.
The period is highly stable across the whole region, ranging between 45 and 65 min: Note also the nearly vertical boundaries of the oscillatory region: $k_1$ can control state transition without codependence on other parameters ($V_{25}''$ plotted, but similar for others). The width of the oscillatory region (and hence the operational stability margin) is also considerably wider than for the 1991 model. The IE-related parameters for which experimental data were not available and have been optimized here are: $k_{le} = 1.2$; $k_{mie} = 0.006$; $k_{ier} = 0.7$; $k_{mier} = 0.001$; $k_{map} = 1$; $k_{apr} = 0.11$; $k_{mapr} = 4$ (symbols correspond to the notation of (Marlovits et al. 1998)). The inset represents 1998 model using original paramter values.

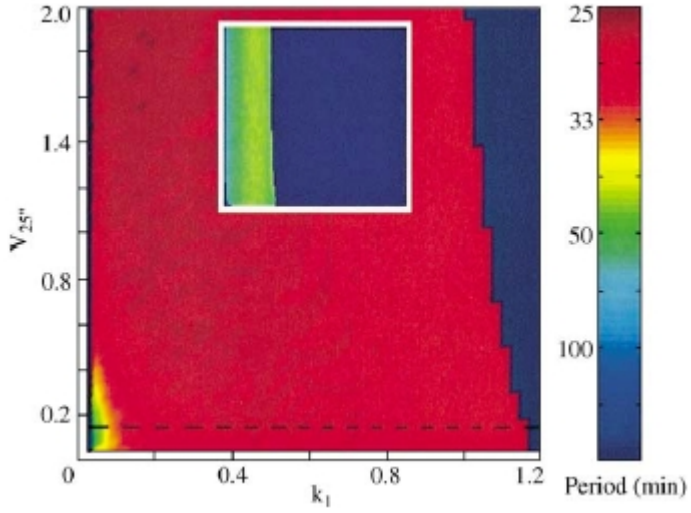


Figure 17: The 1998 model optimized to give *in vivo* like oscillations.
This particular plot was obtained by simply increasing $V_2''$ — the fast rate of degradation of active-MPF by APC – to 1.5 $min^{-1}$: $V_2''$ corresponds to $k_6$ in the 1991 model. As shown in Figure 9(D), increasing $k_6$ has a similar effect on the 1991 model. But whereas in the 1991 model the period of oscillations varies widely across the region, in the optimized 1998 model the period is highly stable in the range 28–30 min: Note that the Marlovits et al. choice of $V_2''$ . 0.25 $min^{-1}$ was based on experimental evidence that suggests a lower limit on $V_2''$ but no specific upper limit. The inset represents the 1998 model using paramter values in Figure 16 (x-axis is adjusted).

# DISCUSSION AND CONCLUSIONS

Model building necessarily involves making choices between alternative explanations with apparently equivalent behaviors. We put forward an argument from first principles suggesting that robustness analysis can help distinguish between more and less plausible models, and pinpoint structural weaknesses in models. The proposal is predicated on the expectation that essential cellular processes that are conserved across multiple species must be functionally robust to mutational variations. Our analysis of two models of the *Xenopus* cell cycle oscillator confirms this theoretical expectation, but further examples are needed.

Our choice of models for this paper was highly serendipitous. The parameter space of the more complex 1998 model had already been mapped in great detail by Borisuk (Borisuk 1997). We could thus concentrate on comparing the two models rather than characterizing each in detail first. Efficient characterization of the parameter space of models with tens of parameters is a significant remaining challenge. Recently, fairly general relaxation methods that exploit linear matrix inequalities to simplify the searching of multi-dimensional spaces have been developed (Parrilo 2000). We hope to exploit these developments to facilitate characterization and parameter searching in future applications of our approach to model building and validation.

# CHAPTER 4: DEVELOPMENT OF FILTERING METHOD FOR CE-MS BASED METABOLOMICS

*Man errs so long as he strives.*

— Johann Wolfgang von Goethe

Metabolomics is expected to boost data driven research. In biomarker discovery, powerful filtering methods to remove noise and outliers are essential for screening significant candidates from the huge volume of omic data. In this chapter, we propose a post-measurement peak filtering method (named P-BOSS) for CE electrospray ionization time-of-flight MS (CE-TOFMS) data, where we leave as many peaks as possible. Combining outlier detection method functions in parallel, we applied P-BOSS to the data using *Escherichia coli* knockout mutants of the tryptophan and purine biosynthesis pathways. As the result, P-BOSS showed remarkably superior performance, reducing 65% of all peaks, while leaving significant peaks.

# INTRODUCTION

Metabolomics is a relatively new discipline for high-throughput metabolic profiling (Fiehn 2002). One of the major challenges in metabolomics is to quantitatively characterize metabolome data simultaneously for system-level understanding of biological systems. Recently a wide variety of metabolome analysis technologies have emerged, including GC-MS (Fiehn et al. 2000; Fiehn et al. 2000), NMR (Reo 2002), FT IR spectroscopy (Harrigan et al. 2004), and CE-MS (Soga et al. 2002; Soga et al. 2002; Soga et al. 2003; Soga et al. 2006). CE-MS has recently been demonstrated as a powerful tool for the analysis of charged species. Its major advantages are that it exhibits extremely high resolution and almost any charged species can be infused into the mass spectrometer. We have shown that CE-MS techniques are quite useful for the global analysis of charged metabolites (Soga et al. 2002; Soga et al. 2002).

Most intracellular metabolites have a charge, and thus CE-MS is particularly useful for revealing those metabolites. Consequently, CE-MS enables us to obtain a large amount of information on metabolites, which can be helpful for profiling the dynamics of metabolic pathways or for biomarker discovery. The data obtained by CE-MS have large numbers of peaks in each sample — typically > 2,000 peaks with CE time-of-flight MS (CE-TOFMS) — as well as other omic data. Therefore, how to obtain useful and significant peaks remains a major challenge (Kell 2004; Jarvis and Goodacre 2005; Broadhurst 2006). Screening large numbers of peaks in advance should allow us to focus on succeeding data analyses. Though many statistical methods have been proposed (Raamsdonk et al. 2001; Taylor et al. 2002; Hirai et al. 2004; Weckwerth and Morgenthal 2005), they are likely to be sensitive to data containing noise, resulting in significant compounds being overlooked.

Thousands of peaks can be detected by CE-MS and data analyses are processed either manually or automatically. Regarding the peaks in metabolome analyses, the peaks can scarcely be identified due to lack of metabolite standard data. They have

significantly different characteristics from those of transcriptome or genome. Moreover, due to the data characteristics of CE-MS, the peak shapes of many compounds are aberrant or are relatively small, and thus they can hardly be distinguished from noise peaks, which easily leads to false-positive peaks.

In the present study, we propose a powerful filtering method by which potentially false-positive peaks are removed, and reproducible peaks are retained. Our filtering method consists of two filters functioning in parallel. One of the filters automatically determines the threshold values of parameters. It tends to remove non-reproducible peaks, potentially noise peaks, while leaving reproducible peaks. The other filter is applied to reproducible peaks to detect and remove outliers. We performed preprocessing after extracting peaks from each data, thereby reducing the data size and calculation cost enormously.

To verify our method, experiments were conducted using tryptophan and purine biosynthesis-relevant knockout mutant data from *Escherichia coli*. Using the obtained data, we confirmed that our method has powerful filtering functions which are widely applicable to peak screening.

## MATERIALS AND METHODS

### BACTERIAL STRAINS, GROWTH CONDITIONS, AND METABOLITE EXTRACTION

The *E. coli* strains JWK1253 (Δ*trpB*), JWK2461 (Δ*purC*), JWK0512 (Δ*purE*), JWK3970 (Δ*purH*), JWK2541 (Δ*purL*), and JWK2484 (Δ*purM*) were used (Baba 2006). The strains are derivatives of BW25113 (Datsenko and Wanner 2000). Cells grown on LB plates were inoculated in a M9 minimal medium supplemented with 5

mg/ml of L-tryptophane (adenine and guanine for purine-related mutants) and incubated at 37°C with shaking. Growth was monitored by measuring optical density at 600 nm ($OD_{600}$). When cell density reached $OD_{600}$= approx. 0.8, cells were collected by brief centrifugation and re-suspended in the same volume of M9 medium without L-tryptophan. The cells were collected at $T_0$, $T_{15}$, $T_{30}$, and $T_{60}$ ($T_x$ indicates the time x in minutes after amino acid shift-down). The metabolites were extracted immediately at each time point as previously described (Soga et al. 2003).

### INSTRUMENTATION

All CE-TOFMS experiments were performed using an Agilent CE Capillary Electrophoresis System G1600A (Agilent Technologies, CA, USA), and an Agilent TOFMS System G1969A. For system control and data acquisition, we used the G2201AA Agilent ChemStation software for CE and the Analyst QS for Agilent TOFMS software.

### CE-TOFMS CONDITIONS FOR CATIONIC METABOLITES

Separations were carried out on a fused silica capillary (50 μm i.d. x 100 cm total length) using 1M formic acid for cationic metabolites. Sample was injected with a pressure injection of 50 mbar for 3 sec. The applied voltage was set at +30 kV. Sheath liquid was prepared as 50% MeOH/Water. For TOFMS, ions were examined successively to cover the whole range of m/z values from 50 through 1000. Fragmentor voltage was set to 75 V. Skimmer voltage was set to 50 V. Oct RFV was set to 125 V. Capillary voltage was set to 4000 V.

### CE-TOFMS CONDITIONS FOR ANIONIC METABOLITES

Separations were carried out on SMILE(+) using 50 mM ammonium acetate (pH 8.5). Sample was injected with a pressure injection of 50 mbar for 30 sec. The applied voltage was set at -30 kV. Sheath liquid was prepared as 5 mM ammonium acetate 50% MeOH/Water. For TOFMS, ions were examined successively to cover the whole range of m/z values from 50 through 1000. Fragmentor voltage was set to 100 V. Skimmer voltage was set to 50 V. Oct RFV was set to 200 V. Capillary voltage was set to 3500 V.

### DATA PROCESSING

Peak extraction was carried out using Human Metabolome Technologies' proprietary software. Statistical analyses were performed using MATLAB R2006b (Mathworks, MA, USA). All other data processing was carried out using Excel 2003 (Microsoft, WA, USA), and Perl script.

## RESULTS AND DISCUSSION

### STRATEGY AND TACTICS FOR EXTRACTING SIGNIFICANT PEAKS

A simple schematic representation of the CE-TOFMS-based analytical workflow, particularly applicable to biomarker discovery, is illustrated in Figure 18(A). Sample

was appropriately prepared to infuse into CE-TOFMS. After measurement by CE-TOFMS, total ion chromatography was performed with a large amount of noise. The noise could be attributed to isotopic compounds, ringing, spikes, and so forth. The peak data set was then compared across sample profiles (or repetitive experiments) and aligned, according to various indices (e.g., *m/z*, migration time). Statistical analyses were then performed to elicit significant peaks (considered as potential biomarkers). In this study, we focus on improving the performance of peak filtering (shadowed box) in Figure 18(A), where detail is shown in Figure 18(B). Basically, noise removal, outlier removal, and missing value imputation should be carried out. This part of the process is very important, because unless noise peaks are removed in this process, the following statistical analyses may suffer critical damage, producing different outcomes due to the noisy peak data. For noise removal, we developed a filtering method that we have named "P-BOSS". For outlier removal, we employed AIC. As shown in Figure 18(B), our strategy employs each function in parallel, in which the procedure is selected according to the number of missing values, thereby avoiding elimination of significant peaks.

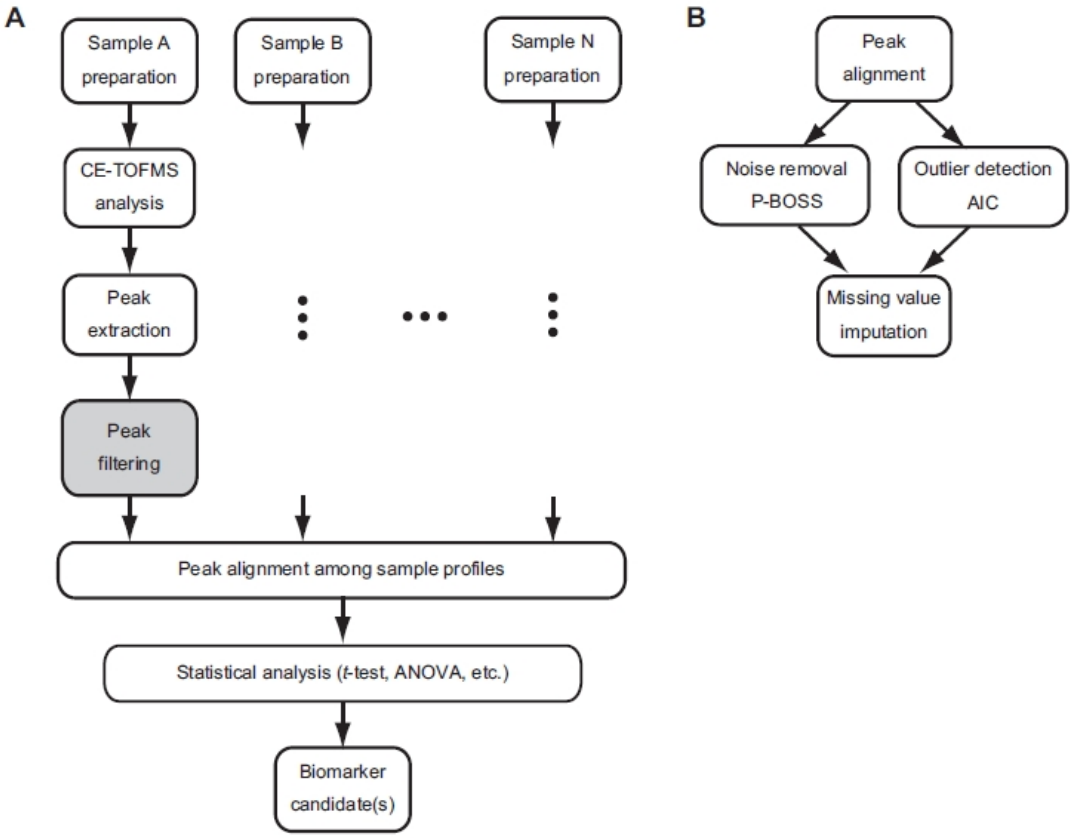


Figure 18: Schematic representation of basic strategy for biomarker search.
(A) Workflow of metabolome analysis from sample preparation to biomarker discovery; (B) Detailed representation of peak filtering in (A).

THRESHOLD COULD NOT BE DETERMINED FOR NOISE REMOVAL

Since the data from the CE-TOFMS analysis include a huge number of noise signals, as well as compound corresponding peaks, the number of peaks increases unexpectedly. The noise could be attributed to isotopic peaks, ringing, or spike peaks. In order to appropriately extract plausible signals, we first characterized the peak data according to the number of missing values across repetitive analysis data. We designated the subset "orphanX" based on the number of missing values as depicted in Figure 19. For example, orphan0 is a subset of peaks detected across every experiment. Next, we categorized those peaks, which are detected in more than half of

experiments, as "non-orphan peaks", indicated as ‘*a*’ in Figure 19. Thus, "orphan peaks" indicate minority peaks across multiple experiments (indicated as ‘*b*’ in Figure 19).

| | Experiment | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | category |
| Metabolite 1 | a | a | a | a | a | orphan0 |
| Metabolite 2 | m | a | a | a | a | orphan1 |
| Metabolite 3 | m | m | a | a | a | orphan2 |
| Metabolite 4 | m | m | m | b | b | orphan3 |
| Metabolite 5 | m | m | m | m | b | orphan4 |

Figure 19: Definition of "orphan" categories.
Suffix depends only on the number of missing values, and not on position. ‘m’ indicates the missing datum. Peak subsets including ‘a’ and ‘b’ are categorized as "non-orphan peak" and "orphan peak", respectively.

The percentile rank for four parameters in CE-TOFMS signals was analyzed using the data set for JWK1253 (Δ*trpB* mutant). The data are shown in Figure 20 as a percentile rank graph. In traditional filtering methods, threshold values are determined such that aberrant peaks are removed before succeeding processes. However, we could not find a threshold value that removes noise peaks without affecting orphan0. Empirical threshold values resulted in the removal of approximately 30% of orphan0 category peaks (data not shown). We assumed that any peak that is dominantly detected over repetitive experiments corresponds to a compound (and not a noise peak). Given this fact, simply determining the threshold values is not appropriate, because it also removes significant peaks.

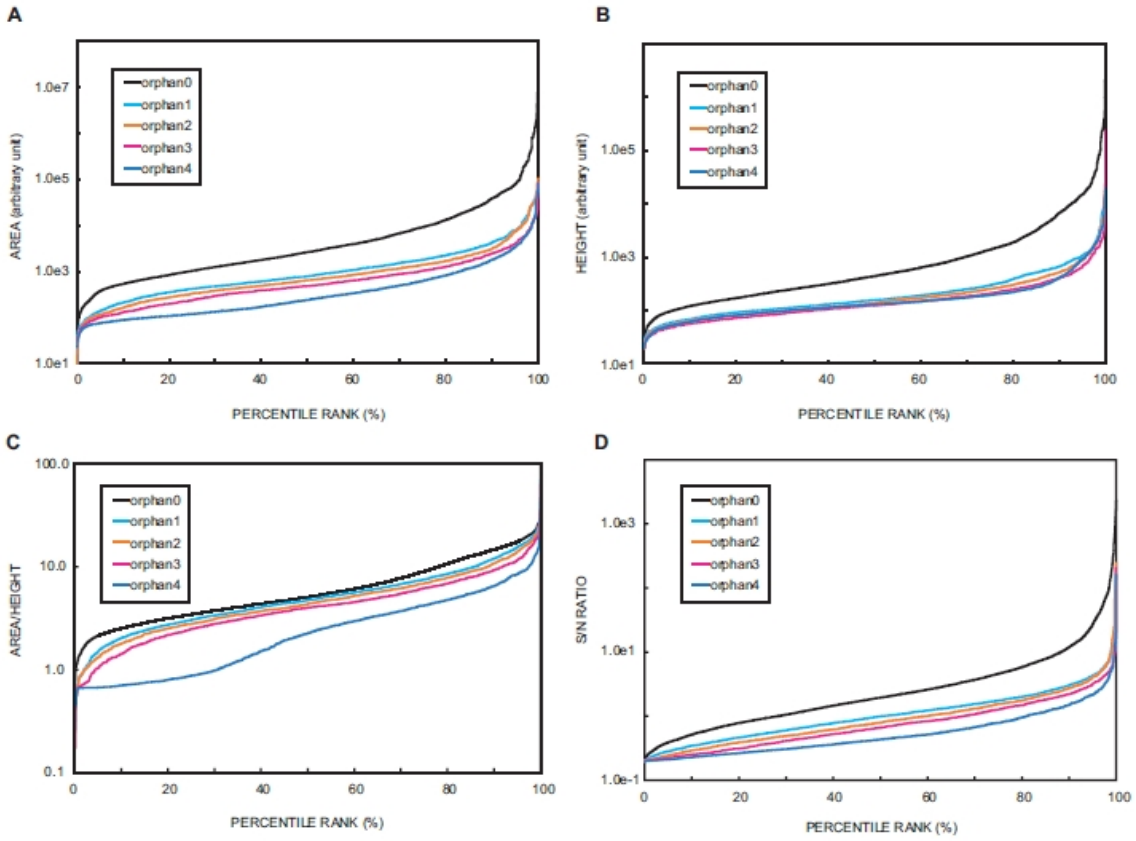


Figure 20: Percentile rank of four parameters in CE-TOFMS signals.
Since repetitive experiments were conducted five times in this analysis, orphan0 is a peak subset detected at all times throughout the experiments. Panels (A) to (D) indicate the percentile rank of analytical parameters of peak area, peak height, area to height ratio, and signal to noise ratio, respectively.

## P-BOSS FILTERING METHOD

Instead of dispensable peak filtering, we developed a powerful filtering method as the procedure summarized in Figure 21. Peaks detected in more than half of the repetitive experiments are excluded from the filtering process, which are then, in turn, processed for outlier detection to flatten peak data across repetitive experiments. Other peaks are regarded as potential noise peaks, and thus the filtering method is being processed. We refer to the filtering method as P-BOSS (Peak filter Based on non-Orphan Survival Strategy).

P-BOSS is superior to the traditional filtering method from two perspectives: (1) it non-heuristically determines a threshold value according to data characteristics, and (2) it applies a filtering process not to all peak data but only to peaks that are probably noise peaks, thus avoiding removal of dispensable peaks. False discovery of peaks is avoided by this filtering process.

Recently, Kadota and colleagues proposed an algorithm to determine threshold values of microarray-based profiles (Kadota et al. 2001). According to their algorithm, two criteria were used to determine appropriate threshold conditions, and their strategy of using two conflicting but important criteria to determine moderate values of filtering parameters is reasonable.

In our study, threshold values of the parameters were inevitably determined according to following procedure. Two indices were employed to determine the values, $R_o$ and $R_n$. $R_o$ represents the proportion of remaining orphan peaks; reduced $R_o$ means that the noise peaks are removed efficiently. $R_n$, on the other hand, represents the proportion of remaining non-orphan peaks. The more $R_n$ increases, the more peaks (which plausibly correspond to compounds) are retained. These two indices are in a trade-off relationship; both are significant to filter peaks. We thus decided to leave $R_n$ as large as possible and $R_o$ as small as possible, by obtaining the product of the two values as an objective function. The objective function *f(x)* for each parameter can be formulated as follows:

$$\text{maximize } f(\mathbf{x}) = R_O(\mathbf{x}) \cdot R_n(\mathbf{x}) \qquad (1)$$

where x represents the vector of areas of peak data.

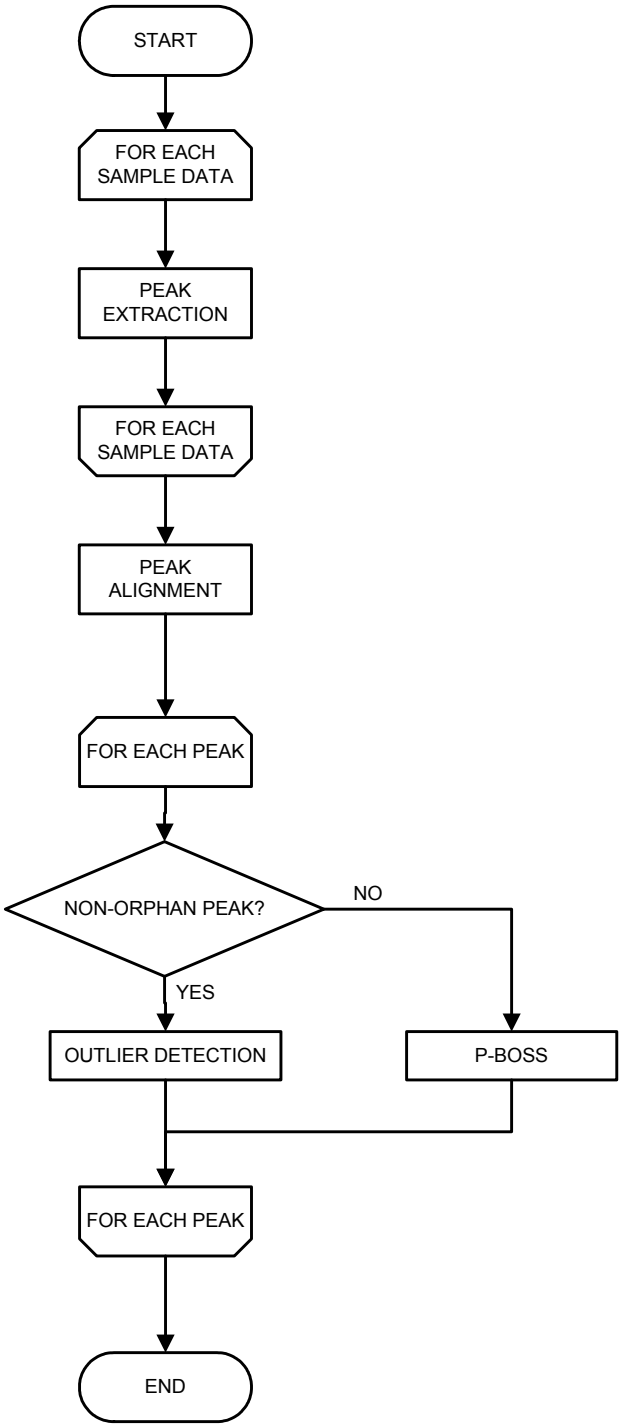


Figure 21: Schematic representation of filtering process with P-BOSS/AIC.

## P-BOSS APPROPRIATELY REMOVES NOISE PEAKS WHILE LEAVING SIGNIFICANT PEAKS

In this experiment, we employed four parameters; area, height, area to height ratio, and signal to noise ratio of peaks. $R_o$ was determined as the ratio of removed peaks in the orphan4 category, whereas $R_n$ was determined as the ratio of left peaks in the orphan0 category. The transition of $f(x)$ over the values of the four parameters is shown in Figure 22. According to the analysis, appropriate threshold values were one-sidedly determined for each parameter, because the shape of $f(x)$ was shown to be parabolic. Absolute values to maximize $f(x)$ are summarized in Table 2. These values are approximately identical to our empirical knowledge, but provide objective evidence to determine the threshold values. To confirm that no significant peaks are removed by the P-BOSS filter, 44 identified peaks corresponding to standard compounds (Table 1, commercially available reagents) were investigated. No peak was removed by the filter, suggesting that P-BOSS functions efficiently to remove noisy peaks only, without removing significant (compound) peaks. Even if peaks were non-orphan (no missing values), they varied in area (mean coefficient of variation was 32%). This implied that outliers are exactly contained in the experimental data, which cause a large variance of scattering data. Tiny and/or deformed peaks are most likely to yield a variance. Eliminating the variance should yield a difference of biomarkers in the chaotic raw data obtained by metabolome analysis.

Table 1: Identified standard compound peaks.

| |
|---|
| Adenine |
| Adenosine |
| Anthranilate |
| Arg |
| Asn |
| Asp |
| Carnosine |
| Citrulline |
| Cys |
| Cytidine |
| Cytosine |
| DOPA |
| GABA |
| Gln |
| Glu |
| Glutathione (ox) |
| Glutathione (red) |
| Gly |
| Guanine |
| Guanosine |
| His |
| Homoserine |
| Hydroxyproline |
| Hypoxanthine |
| Ile; Leu |
| Inosine |
| Lys |
| Met |
| Ornithine |
| Phe |
| Pro |
| SAM |
| Ser |
| Spermidine |
| Spermine |
| Trp |
| Tyr |
| Tyramine |
| Uridine |
| Val |
| β-Ala |

Table 2: Threshold values determined according to the max value of *f*(x).

| Parameter | Threshold | *f(x)* |
|---|---|---|
| Area[a] | 1000 | 0.673 |
| Height[a] | 200 | 0.508 |
| Area/height | 2 | 0.403 |
| S/N ratio | 7 | 0.527 |

a: arbitrary units

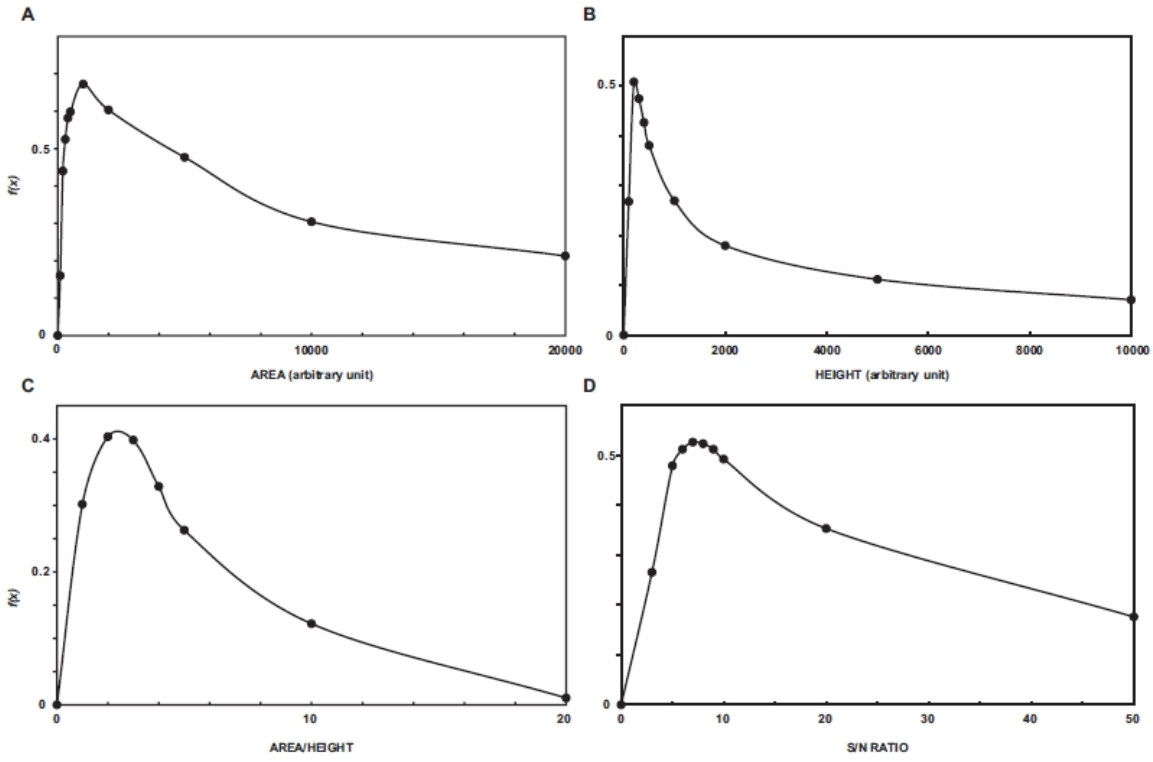


Figure 22: Transition of f(x) according to each parameter.
Panels A to D indicate the f(x) values for the analytical parameters of peak area, peak height, area to height ratio, and signal to noise ratio, respectively.

## INTRODUCTION OF AKAIKE'S INFORMATION CRITERION TO P-BOSS

In order to smooth the values (i.e., detecting outliers), we employed a method based on Akaike's Information Criterion (AIC). The main advantage of the method is objectiveness of decision, because the procedure is independent of a level of significance (i.e., *p* values). In addition, the method can detect multiple outliers without setting the number of outliers; that is, the number of outliers varies depending upon the data set. AIC was used to evaluate mouse DNA microarray expression profiles, and proved its efficiency (Kadota et al. 2003). We applied the method to the data above in parallel with the noise filter. The data set was first categorized into two groups (i.e., orphan peaks and non-orphan peaks), according to the number of missing values. P-BOSS was applied to the orphan peaks, while AIC was applied to the non-orphan peaks. For calculating *f(x)*, orphan4 was used to determine $R_o$, and orphan0 to determine $R_n$. The results are shown in Table 3. The data showed that upon applying P-BOSS/AIC, the number of peaks in the raw data (JWK1253 at $T_0$, n=5) was reduced to 54% (due to P-BOSS), and the coefficient of variation to 1/4 (due to AIC). These results reveal that the combination of AIC and P-BOSS successfully eliminates the noise peaks and leaves the significant peaks in variable repetitive data.

Table 3: Results between before and after applying P-BOSS

| | Number of peaks | RSD (%) |
|---|---|---|
| Non-treated | 5356 | 32 |
| P-BOSS treated | 2918 | 8 |

## DIFFERENTIAL ANALYSIS USING P-BOSS/AIC

Next, we compared two associated metabolome data to investigate the ability of the P-BOSS/AIC combination. The metabolome data used here are provided from *E. coli* JWK1253 (Δ*trpB)* at $T_0$ and $T_{15}$ (n=5 each). The data were processed through P-BOSS/AIC as mentioned above, and representative values for all peaks were extracted. The values were then matched according to *m/z* (determined by MS) and migration time (determined by CE). Table 4 shows the number of matched/unmatched peaks in orphan0 and orphan4 categories between the samples at $T_0$ and $T_{15}$. The portion of matched peaks remained 1.7-fold more than that of unmatched peaks, suggesting that the peaks detected in either sample, corresponding to unmatched peaks, are aberrant/noise peaks. The remaining orphan peaks after P-BOSS/AIC treatment are potential candidates corresponding to substances with altered levels (so-called biomarkers). For instance, given two datasets to compare (e.g., wild-type and knockout samples), certain compounds might exist in either sample. In the other case, when measuring time-series data, certain compounds might be accumulated gradually. Regarding biomarker search, as many peaks should be left as possible (even tiny or weak peaks). P-BOSS is particularly suitable for these purposes, because our method leaves peaks as much as possible while removing the minimum number of possibly noisy peaks. This should function as a primary screening method, enabling reliable statistical analyses to be performed thereafter.

Table 4: Matching ratio of peaks (orphan0 and orphan4 categories only)

| | Non-treated | P-BOSS treated | Remaining proportion (%) |
|---|---|---|---|
| Matched peaks | 3569 | 2311 | 64.8 |
| Unmatched peaks | 3983 | 1510 | 37.9 |
| Total | 7552 | 3821 | 50.6 |

The resolution power of P-BOSS was examined. Metabolite groups in which m/z values and migration times are similar (i.e., structural isomers) are sometimes found in metabolome samples. For example, L-Leucine and L-isoleucine are structural isomers with the same molecular weight and similar migration time (the difference is less than 10 sec). Such a situation makes it difficult to identify which of these peaks is the candidate, especially unidentified peaks. Actually, we found two candidate peaks around the *m/z* value of L-glutamine, ca. 147.0764, before applying our method, as shown in Table 5 (upper). The gray area corresponds to the peaks where eliminated by P-BOSS. After applying P-BOSS, however, one of the peaks was eliminated, and only one peak successfully remained. In this case, we confirmed that the remaining peak corresponds to L-glutamine, by spiking authentic L-glutamine. Another example, that of L-tyrosine, is shown in Table 5 (lower). This is also an identified peak, and the corresponding peak in the sample at $T_0$ was eliminated (second row in the lower table in Table 5 due to a tiny and/or aberrant peak shape. These results indicate that P-BOSS processing realizes accurate matching of multiple metabolome data and provides a rapid differential analysis of metabolome profiles for biomarker discovery. When handling data without applying filters such as P-BOSS/AIC, identification of peaks is quite difficult, since more than one peak with the same *m/z* value can potentially be detected. Our method demonstrated validity for peak identification by removing adjacent peaks.

Table 5: Removal of ambiguous peaks adjacent to objective peaks.

**L-glutamine**

| Non-treated | | | | P-BOSS treated | | | |
|---|---|---|---|---|---|---|---|
| $T_0$ | | $T_{15}$ | | $T_0$ | | $T_{15}$ | |
| *m/z* | MT(min) | *m/z* | MT(min) | *m/z* | MT(min) | *m/z* | MT(min) |
| 147.0755 | 12.50 | 147.0752 | 12.49 | 147.0755 | 12.51 | 147.0752 | 12.51 |
| 147.0756 | 11.51 | 147.0761 | 11.52 | | | | |

**L-tyrosine**

| Non-treated | | | | P-BOSS treated | | | |
|---|---|---|---|---|---|---|---|
| $T_0$ | | $T_{15}$ | | $T_0$ | | $T_{15}$ | |
| *m/z* | MT(min) | *m/z* | MT(min) | *m/z* | MT(min) | *m/z* | MT(min) |
| ND | ND | 182.0664 | 18.93 | | | | |
| 182.0802 | 15.74 | 182.0811 | 15.85 | | | 182.0811 | 15.85 |

abbreviation: MT; migration time, ND; not detected

## ACCURATE FILTERING BY P-BOSS/AIC

We next performed a differential analysis upon multiple metabolome profiles to extract significant compounds whose levels are statistically significant (we refer to "significant compounds" as those which show similar behavior irrespective of a different environment). In this study, five E. coli mutants were examined: JWK2461 (Δ*purC*), JWK0512 (Δ*purE*), JWK3970 (Δ*purH*), JWK2541 (Δ*purL*), and JWK2484 (Δ*purM*). These mutations are located in genes associated with the purine biosynthesis pathway. We analyzed metabolome samples prepared from cells at $T_0$, $T_{15}$, $T_{30}$, and $T_{60}$ after guanine and adenine shift-down (see Materials and Methods). P-BOSS/AIC was applied to the data of each mutant. To evaluate the performance of P-BOSS/AIC, we also examined the original profiles (neither P-BOSS or AIC was

applied). For metabolome data on each mutant, P-BOSS/AIC (or none for original profiles) was applied, and then the number of extracted candidates was examined by differential analysis by matching each peak. To find a significant difference among the profiles, ANOVA (analysis of variance) was performed to see which peak shows a significant difference among the profiles. The significance level was set within 5%. As the result, we found striking difference between the profiles. P-BOSS reduced 65% of peaks (20651 to 7305 peaks), while leaving all significant compounds (934 peaks). Further, P-BOSS resulted in adding 5% more significant peaks (48 peaks), due to removing ambiguous peaks adjacent to significant ones. This result clearly indicates that P-BOSS has outstanding advantages in screening significant peaks. While there need to develop additional filtering method for further screening the peaks, P-BOSS can be a powerful filter for initial screening.

## CONCLUSION

We have proposed here a powerful filtering method for the preprocessing of metabolome data measured by a CE-TOFMS system. Employing an appropriate filtering method, such as P-BOSS/AIC, should thus enable us to narrow down potential compound-associated peaks. Our strategy is widely applicable to omics-based biomarker discovery.

# CHAPTER 5: METABOLOMICS AND SIMULATIONS UPON *BACILLUS SUBTILIS*

*The fact of evolution is the backbone of biology, and biology is thus in the peculiar position of being a science founded on an improved theory, is it then a science or faith?*

— Charles Darwin

When environmental conditions fluctuate unexpectedly, the choice by an organism of a pure strategy increases its extinction risk, thus a mixed strategy is the evolutionarily stable strategy in natural environments. The time-dependent mixed strategy is seen in many insects and higher plants, however, its molecular mechanism is still unclear. The soil bacterium *Bacillus subtilis* forms dormant, robust spores as a strategy to ensure its survival under conditions of starvation. Recent studies suggest that polyphenism, whether to initiate sporulation or not, are their tactics for their adaptive response to starvation. We show here that polyphenism during sporulation is primarily modulated by negative feedback circuits in the signaling pathway, resulting in generating a bistable response within the sporulating culture. We predict this phenomenon by building a simple mathematical model for signal transduction of the sporulation cue by wild type and mutants involving both positive and negative feedback. We confirmed our models experimentally by using mutants virtually inhibiting and activating the negative feedback. Besides, metabolome analysis was conducted to see the dynamics of metabolic pathway upon sporulation. As the results, we found that metabolism at sporulation starting stage is significantly different, depending on *spo0E* mutants.

# INTRODUCTION

Survival strategies affect the fate of organisms living in highly fluctuating environments. In general, organisms have two or more strategies to assure survivorship; they hedge the extinction risk by emerging these strategies simultaneously (Cohen 1967; Hopper 1999; Clauss and Venable 2000). For example, in annual plants, a delay in seed germination reduces the temporal variance in individual fitness and minimizes the risk of extinction (Clauss and Venable 2000). This ability is known as the bet-hedging strategy (Hopper 1999) and is thought to be programmed elaborately in the genome. However, how clonal cells express the strategies simultaneously in response to changes in the environments remains to be elucidated.

Phenotypic heterogeneity in clonal population has been found in bacteria as well under some circumstances, e.g., the persister production in antibiotic-treated *Escherichia coli* populations (Balaban et al. 2004), the lactose utilization of *E. coli* (Ozbudak et al. 2004), competence development in *Bacillus subtilis* (Maamar and Dubnau 2005), and sporulation (Chung et al. 1994; Chung and Stephanopoulos 1995). However, the ecological, evolutional significance of phenotypic heterogeneity is not clear in bacteria.

Sporulation of the soil bacterium *B. subtilis* is initiated to ensure survivorship under conditions of starvation (Grossman 1995). In cells receiving a sporulation signal, the phosphorylation of kinases (such as KinA) is stimulated and the phosphate group is transferred to Spo0A through phosphorelay (Burbulys et al. 1991). Phosphorylated Spo0A (Spo0A~P) is the master regulator for sporulation; it acts as a transcriptional factor for sporulation-associated genes. This signal transduction system is regulated by complex mechanism including multiple positive- and negative-feedback loops (de Jong et al. 2004).

A primary effect of positive feedback on gene expression is amplification of its own expression rate. Another function is the generation of a bistable state in the population; this results in two distinct cell subpopulations with different gene expression levels (Ferrell and Machleder 1998; Becskei et al. 2001; Ozbudak et al. 2004). It has been suggested that *kinA* mutation, lack of a positive-feedback loop in phosphorelay, affects the bistable state of Spo0A~P in sporulating cells (Chung et al. 1994). The negative-feedback loop acts as a noise-reduction system of gene expression (Becskei and Serrano 2000). Otherwise, dephosphorylation of Spo0A~P by Spo0E, or autostimulation of Spo0A~P contributes to bistable state of sporulating cells (Veening et al. 2005). Though the behavior of genetic network is considered to be effective for cellular functions such as differentiation and adaptation to the environment, no direct observation is yet reported.

Recent theoretical and experimental studies suggested intrinsic characterstics of system to generate population heterogeneity (see (Smits et al. 2006) for review). Voigt and colleagues investigated dynamics of sin operon with a mathematical model (Voigt et al. 2005), and showed that combining genes from a regulatory protein and its antagonist within the same operon could lead to diverse regulatory function, such as bistability, oscillation, and pulse generator. Iber and colleagues, on the other hand, used spoIIA operon as an example to show similar results (Iber 2006; Iber et al. 2006), and confirmed by experiments. De Jong and colleagues performed a qualitative simulation (de Jong et al. 2004), by which qualitative characterstics, consistent with experimental results, could be reproduced, yet the model involves too many factors to extract essential functions out of it.

Here we employed a simple mathematical model to elucidate dynamics of Spo0A~P involving both positive and negative feedback. Although experimental data cannot be obtained for quantitative simulation, the model allows us to investigate the essential functions of sporulation relevant signal transduction pathway. Given variation of signal intensity for sporulation, Spo0A~P exhibited unstable dynamics under specific condition, corresponding to show either sporulation or non-sporulation subpopulation. The results were verified experimentally using wildtype and various mutants of *B.*

*subtilis*. In addition, metabolome analysis was performed to see the dynamics of metabolites during sporulation. The results confirmed that energy metabolism is significantly affected by *spo0E* mutants, suggesting its link to sporulation.

## MOLECULAR AND BIOCHEMICAL FEATURES OF SPORULATION IN BACILLUS SUBTILIS

The process of sporulation in *Bacillus subtilis* proceeds through a well-defined series of morphological stages that involve the conversion of a growing cell into a two-cell-chamber sporangium within which a spore is produced (Stragier and Losick 1996). While the sporulation process involves over 125 genes, availability of huge database of genetic and genomic information and advanced technologies have enabled us to investigate and manipulate the organisms quite a detail. Completion of gene sequencing of *B. subtilis* further facilitated understanding role of individual genes (Kunst et al. 1997).

The successive morphological stages of sporulation are shown in Figure 23 (see (Stragier and Losick 1996) for detail). There are seven stages, with each stage designated by Roman numerals. Entry into sporulation is characterized by the formation of axial filament in which two chromosomes from the last round of DNA replication become aligned across the long axis of the cell. As a next stage, a septum is formed at an extreme polar position.

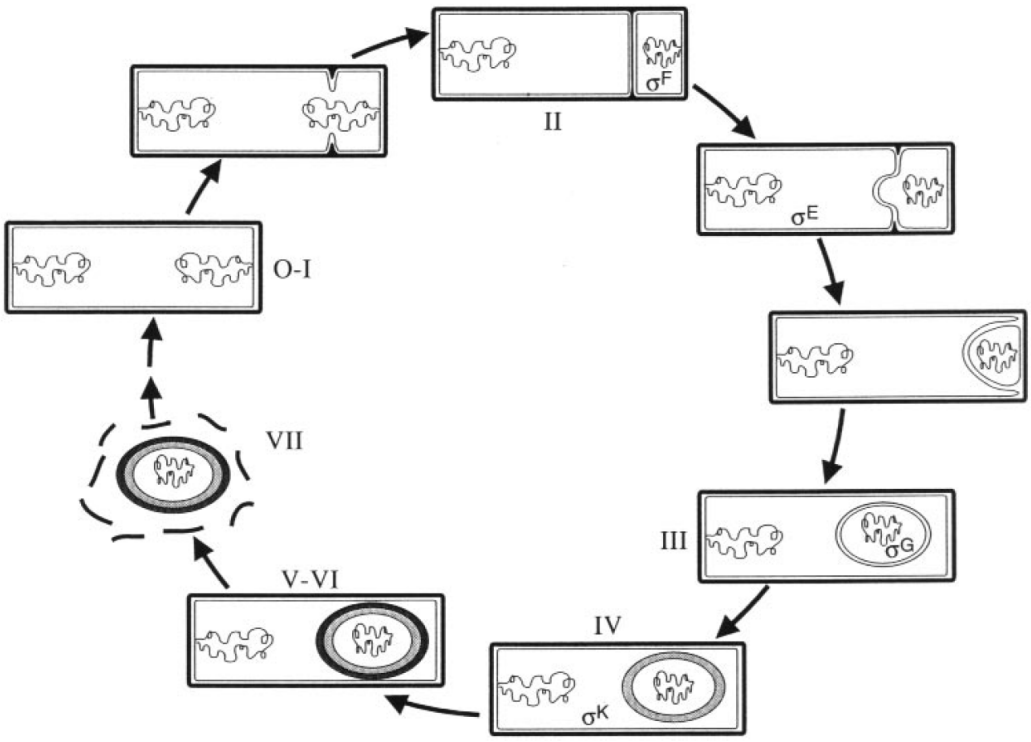


Figure 23: The morphological stages of sporulation.
Image is reprinted from (Stragier and Losick 1996).

The stages are designated by Roman numerals. The wavy lines are chromosomes. The sporangia are surrounded by a cytoplasmic membrane (*thin line*) and a cell wall (*thick line*). The developing spore (stage IV-VI) is encased in a layer of cortex (*light stippling*) and a coat layer (*dark stippling*). The four specific sporulation sigma factors are shown in the cells where they become active.

Following is the brief description of each stage.

Stage 0

This stage represents cells that have not entered the sporulation pathway. Upon starvation of nutrition, they slow down their growth rate, and shift to sporulation.

Stage I

This stage represents cells that have entered the pathway and formed an axial filament. Chromosomes are replicated, and they become aligned along the long axis of the cell

in a structure known as the axial filament. Levin and Losick (Levin and Losick 1996) found, with Spo0H mutant, that cells stop at the stage where specific asymmetric FtsZ ring is formed. Levin thus proposed to define Stage 1 as when the ring is formed, and to change spo0H to spoIH (Levin and Losick 1996). CitC gene (structural gene of isocitric acid dehydrogenase, an enzyme of citric acid pathway) has only been reported (Jin et al. 1997), which fulfills the definition.

Stage II

This stage represents sporangia that have reached the stages of polar septation. Septum is formed from one side of asymmetric Z ring. The other Z ring, in turn, vanishes (Levin and Losick 1996; Pogliano et al. 1999). This asymmetric separating membrane makes cells segregated into two compartments; smaller one is called prespore, and becomes spore in future.

Stage III

This stage represents sporangia that have reached the stages of engulfment. The membrane is invaginated, and forms a forespore surrounded by mother cell. At this point, cell membrane is two-fold.

Stage IV

In the intermembrane space between the forespore and mother cell, a thin layer of peptidoglycan known as the germ cell wall is produced on the surface of the forespore membrane. This is followed by the synthesis of a thick layer of peptidoglycan known as the cortex, which is thought to be involved in attaining or maintaining the dehydrated and heat-resistant state of the spore.

Stage V

This stage represents cells in which spore coat deposits on the surface of forespore. Most of the constituent part are called spore coat proteins (cot; spore coat), and massively expressed in mother cell. In spore cell, on the other hand, various genes

such as ssp (Small acid Soluble Protein) are expressed specifically. This gene is to give spore resistant to ultra violet, and cells obtain the resistance at this stage.

Stage VI

In this stage, spores come to maturity, and obtain resistance to heat.

Stage VII

In this stage, spores separate from mother cell.

From molecular viewpoint, on the other hand, there are number of factors involved during the stage transition, as summarized in Figure 24. To initiate sporulation, *B. subtilis* uses a phosphorelay system, which involves five sensory histidine kinases (KinA – KinE) that respond to various extracellular and intracellular signals (Piggot and Hilbert 2004). These kinases phosphorylated the sporulation initiation phosphotransferase Spo0F and the single-domain response regulator Spo0B, and the phosphoryl group is then passed on to the master regulator Spo0A. Spo0A, in turn, regulates large variety of factors, including sigma cascade – regulon of Spo0A was extensively investigated (Molle et al. 2003), by which totally 121 genes (organized as 30 single-gene units and 24 operons) are likely to be under the direct control.

To survive in rapidly changing environmental conditions, bacteria have evolved a diverse set of regulatory pathways that govern various adaptive responses (See (Smits et al. 2006) for review). Among them, Spo0A plays a key role, which subject to several autostimulatory loops, both at the transcriptional level and at the level of activation (Piggot and Hilbert 2004).

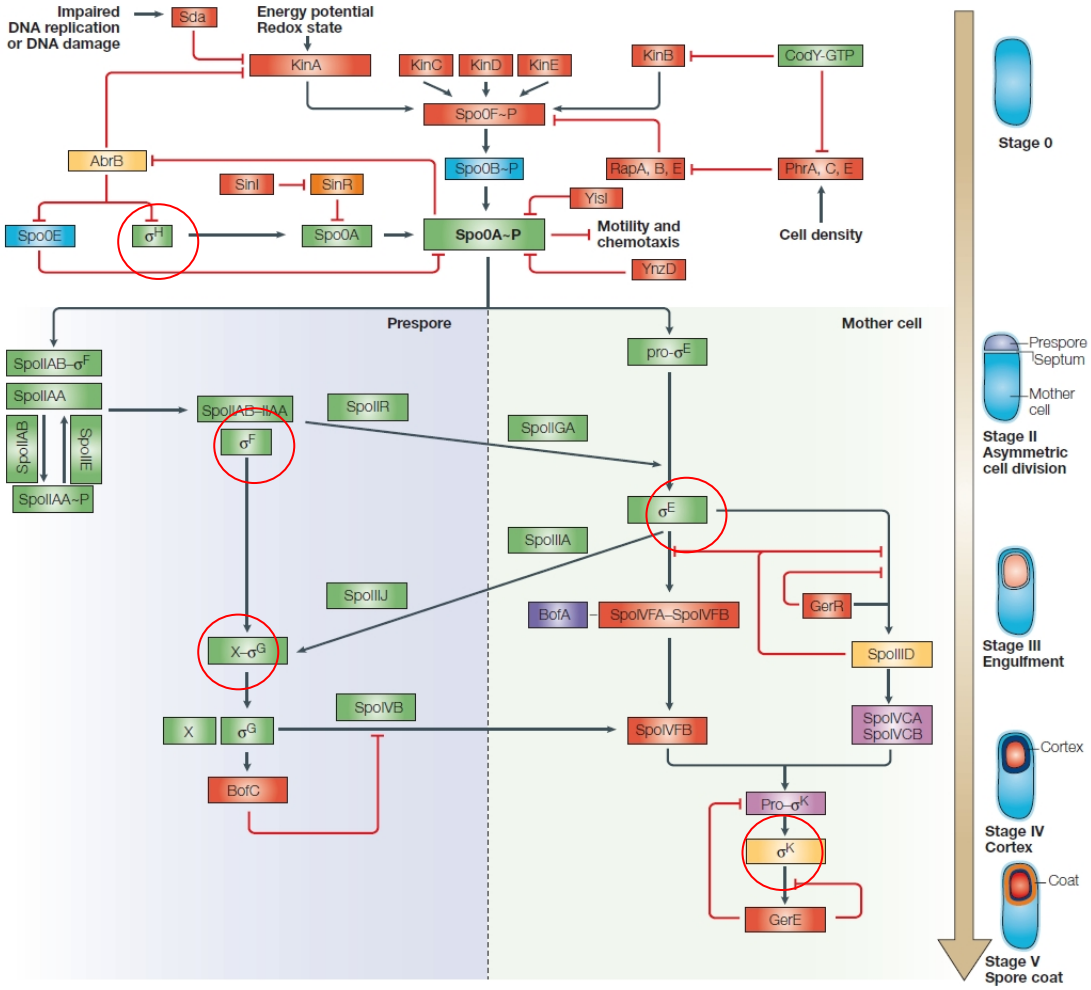


Figure 24: The sporulation cascade in *Bacillus subtilis* and selected clostridia.

(image reprinted from (Paredes et al. 2005))

# MATERIALS AND METHODS

## PLASMIDS AND BACTERIAL STRAINS

Bacterial strains used in this study are summarized in Table 7. The $P_{spoVG}$-*GFPuv* reporter gene (encoding green fluorescent protein) was constructed as follows. The polymerase chain reaction (PCR) product containing the *GFPuv* structural gene was introduced into the *Eco*RV site of pHASH103 (Ohashi et al. 2003) using the TA-cloning method; this generated pSHINE2192. Next, the PCR fragment containing the promoter sequence of the *spoVG* gene was amplified with the specific primer pair PspoVG-F (5'-GCCCGAAATGAAAGCTTTATGA-3') and PspoVG-R (5'-GCATTAGTGTATCAATTCCACG-3'). Genomic DNA of *B. subtilis* 168 (laboratory stock) was used as a template. The PCR fragment was introduced into the *Sma*I site of pSHINE2192 by the TA-cloning method (Ohashi et al. 2003). The generated pSHINE2172 was introduced into BEST2131 (*leuB*::pBRTc) (Itaya 1993), generating BEST12008. BEST12007 and BEST12005 were constructed by transforming UOT1317 (*sof1 spo0F*Δ) (Hoch et al. 1985) and RIK3 (*spo0H*::pJ0H7d) (Jaacks et al. 1989; Ohashi et al. 1999), respectively, using pSHINE2172. BEST12013 was constructed as follows. The *cat* gene of pHASH102 (Ohashi et al. 2003) was PCR-amplified using the specific primer pair cat-F (5'-CAGTAATATTGACTTTTAAAAAAGGATTG-3') and cat-R (5'-GAAACCATTATTATCATGACATTAACC-3'), and introduced into the *Sma*I site of pRIK0Ed (Nanamiya et al. 2000). This yielded plasmid pRIK0Ed-cat, which was introduced into the genomic DNA of 168, yielding BEST12013. To obtain BEST12014, BEST12013 was transformed using BEST12008 genomic DNA. BEST12025 was constructed as follows. The promoter of the *spo0E* gene was PCR-amplified with the specific primer pair Pspo0EF (5'-CCTGGGTATTGTTCTTCTAATCCTATC-3') and Pspo0ER (5'-GCTAAGAAATAGGAAACAAGTTTGATTGGG-3') and introduced into the *Sma*I site of pHASH103 (Ohashi et al. 2003), generating pST0E1. Next, the *spo0E* gene

was amplified by PCR using primer pair spo0ErandomF (5'-ATGNNNNNNNNNNNGGCGGTTCTTCTGAACAAGAAAGATTG-3') and spo0ER (5'-GGCCGCTATTTATTTGCATCATATGC-3') to attach the potential downstream elements that enhance translation efficiency (Ohashi et al. 2005). The fragment was introduced into the *Eco*RV site of pST0E1 and then introduced directly into the genomic DNA of BEST2136 (Itaya 1993). Transformants with various sporulation frequencies appeared on plates containing chloramphenicol, erythromycin, and tetracycline. A sporulation- defective mutant BEST12026 was isolated, and the mutant *spo0E* gene was designated *spo0E102*. To generate BEST12022, BEST12006 was transformed using the genomic DNA of BEST12008.

### SPORULATION CONDITION

*B. subtilis* cells were grown at 37°C in 2X SG medium containing 0.1% (w/v) D-glucose (Leighton and Doi 1971). To synchronize the growth phase of the cells, the culture was diluted 10-fold when optical density at 660 nm ($OD_{660}$) reached 0.5. The end of the logarithmic growth phase ($T_0$) was defined as the point at which the culture reached the $OD_{660}$ of 1.5. The sporulation fraction was defined in terms of colony-forming units (CFU)/ml.

### MICROSCOPY AND DATA PROCESSING

An aliquot (approximately 20 μl) of the culture at the sporulation phase was briefly centrifuged and the supernatant was removed. Cells were washed once in MilliQ water and resuspended in 2 μl of component A of SlowFade-Antifade Kits (Molecular Probes, Inc., OR). A 1 μl aliquot of the cell suspension was inoculated onto an agarose layer on the glass slide and covered with a coverslip. Microscopic analyses were done using an AxioskopMOT 2 microscope (Carl Zeiss, Göttingen, Germany) and a CoolSNAP fx CCD camera (Roper Scientific, Inc., Arizona, USA). To detect the fluorescence of GFPuv, Filter Set 17 (Carl Zeiss) was used. The images were

taken 40 seconds after excitation by UV. The fluorescence intensity of individual cells was calculated using MetaMorph Ver. 4.6 software (Universal Imaging, Co., PA).

## INSTRUMENTATION

All CE-TOFMS experiments were performed using an Agilent CE Capillary Electrophoresis System G1600A (Agilent Technologies, CA), and an Agilent TOFMS System G1969A. For system control and data acquisition we used the G2201AA Agilent ChemStation software for CE and the Analyst QS for Agilent TOFMS software.

## CE-TOFMS CONDITIONS FOR CATION

Separations were carried out on a fused silica capillary (50 μm i.d. x 100cm total length) using 1M formic acids. Sample was injected with a pressure injection of 50 mbar for 3 sec. The applied voltage was set at +30 kV. Sheath Liquid was prepared as 50% MeOH/Water. For TOFMS, ions were examined successively to cover the whole range of m/z values from 50 through 1000. Fragmentor voltage was set to 75 V. Skimmer voltage was set to 50 V. Oct RFV was set to 125 V. Capillary voltage was set to 4000 V.

## CE-TOFMS CONDITIONS FOR ANION

Separations were carried out on a fused silica capillary (50 μm i.d. x 100cm total length) using 50 mM ammonium acetate (pH8.5). Sample was injected with a pressure injection of 50 mbar for 30 sec. The applied voltage was set at -30 kV.

Sheath Liquid was prepared as 5 mM ammonium acetate 50% MeOH/Water. For TOFMS, ions were examined successively to cover the whole range of m/z values from 50 through 1000. Fragmentor voltage was set to 100 V. Skimmer voltage was set to 50 V. Oct RFV was set to 200 V. Capillary voltage was set to 3500 V.

### DATA PROCESSING

Peak extraction was carried out using our proprietary software. Peak preprocessing was performed using P-BOSS/AIC (Morohashi et al. 2007), using Excel 2003 (Microsoft, WA, U.S.A.). Mathematical simulation was conducted using XPP-AUTO (Ermentrout 2002). Statistical analyses were performed via MATLAB (Mathworks, MA, U.S.A.).

## RESULTS AND DISCUSSION

### FEEDBACK COEFFICIENTS MODULATE THRESHOLD OF SPORULATION SWITCH

In cells initiating sporulation, the expression of *spo0H* encoding sporulation-specific $\sigma^H$ is induced by a reduction in the AbrB level (Figure 25). The RNA polymerase that contains $\sigma^H$ (RNAP-$\sigma^H$) stimulates the expression of phosphorelay components, *kinA*, *spo0F*, and *spo0A*, which constitute multiple points of positive-feedback loops. Negative-feedback regulation is also found in *B. subtilis* phosphorelay. The expression of the *spo0E* gene encoding the Spo0A~P-specific phosphatase is induced by a reduction in the AbrB level at sporulation onset (Perego and Hoch 1991). It has been suggested that this is a solo system that negatively regulates phosphorelay as a feedback loop (Perego and Hoch 2002).

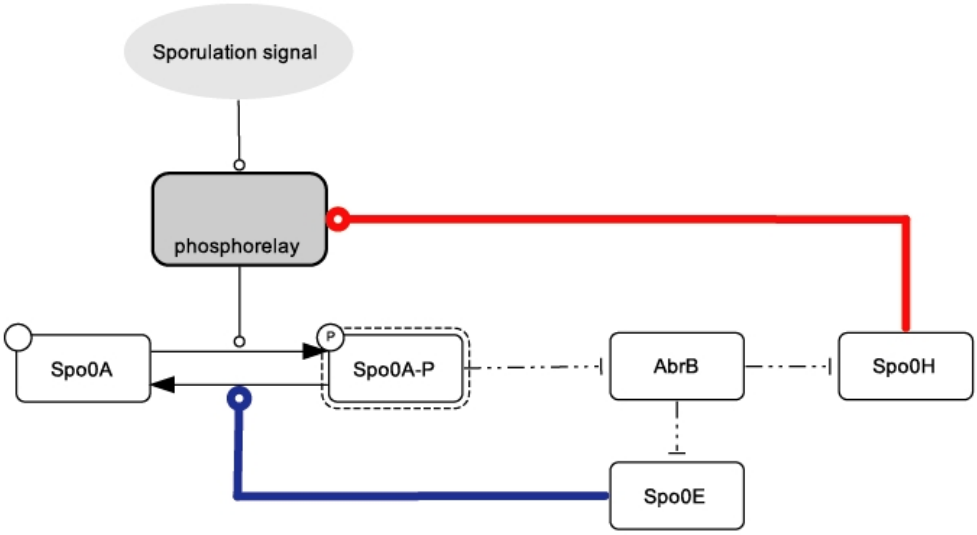


Figure 25: Schematic representation of the phosphorelay network in *B. subtilis*.
The diagram is illustrated by CellDesigner 3.5.1 (The Systems Biology Institute, http://celldesigner.org, (Funahashi et al. 2003)), and the notation follows that proposed by Kitano (Kitano et al. 2005). The networks downstream of AbrB are simply categorized into positive and negative feedback loops, where red line represent positive feedback regulation, and blue line represents negative feedback regulation.

To see the impact of balance of positive- and negative-feedback loops to the system, we built a mathematical model. This process is thought to be driven stochastically and activated in cells in which the level of Spo0A~P exceeds the threshold (Chung et al. 1994; Chung and Stephanopoulos 1995; Maughan and Nicholson 2004; Fujita et al. 2005). Sporulation stimuli induced by nutrient starvation is substantially unidentified, but the amount of stimulus is defined by the level of Spo0A~P. The behavior of the bistable memory module is mathematically explained by introducing feedback coefficients (Xiong and Ferrell 2003). This introduction allows representation of the accumulation of Spo0A~P via the phosphorelay of *B. subtilis* by the equation:

$$\frac{d[A]}{dt} = \varphi\left([A]_{\text{int}} - [A] + f_P \frac{[A]^n}{{k_P}^n + [A]^n}\right) - f_N \frac{[A]^m}{(k_N + [A])^m} - k_{inact}[A] \qquad (1)$$

where [*A*] denotes the concentration of Spo0A~P, $\varphi$ is the sporulation signal intensity that cells can receive from the environment, and $f_P$ and $f_N$ represent the feedback coefficients for positive and negative, respectively. $k_P$ and $k_N$ are the concentration of Spo0A~P for a half-maximum response for the feedbacks. $n$ denotes the Hill coefficient for positive feedback, and $m$ is the order of negative feedback. Since the magnitude of the negative feedback is represented by the combination of the enzymatic activity of Spo0E and the concentration of Spo0E (both terms are represented by Michaelian hyperbola), we assumed the order $m = 2$. $k_{inact}$ is the coefficient of spontaneous dephosphorylation of Spo0A~P. The first term of equation 1 represents the positive-feedback loop of Spo0A~P synthesis that is cued by sporulation signal $\varphi$. Second and third terms represent the negative-feedback loop of the system and spontaneous dephosphorylation of Spo0~P, respectively.

To see the dynamics of the model, we first derived nominal values of each parameter based on subpopulation ratio. By assuming that the intensity of the sporulation signal reflects a Gaussian distribution, the relationship between $r$ and the expected scale of the sporulating subpopulation can be computed. The relationship between the sporulation signal intensity and the cell number, $Y_\varphi$, is computed as Gaussian and is represented as

$$Y_\varphi = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\varphi-\bar{\varphi})^2}{2\sigma^2}\right\} \qquad (2)$$

where $\varphi$ is mean, and $\sigma$ is the standard deviation. The mean intensity of sporulation signal is assumed to be 0.5, and $\sigma$ to vary between 0.1 to 10. For each sigma value, transition of subpopulation of non-sporulating was calculated. The results are shown in Figure 26, clearly indicating that the feedback coefficient $r = f_N/f_P$ significantly affects the sporulation ratio. This result suggests that subtle feedback regulation may be critical to generation of both subpopulations; otherwise either a sporulating or non-sporulating population only would be generated. From our observation of the fraction

of sporulating individuals (53%, see Figure 28(A) and (D)), we suggest that $r = 0.5$ is the nominal value of the feedback coefficient.

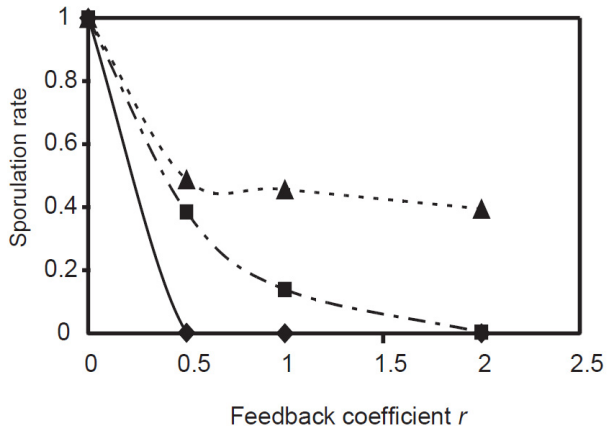


Figure 26: Dependency of sporulation rate upon the feedback coefficients. Triangles: $r = 0.1$, squares: $r = 1$, diamonds: $r = 10$.

Table 6: Parameter values used in this study.

| Symbol | Description | Value |
|---|---|---|
| $[A]_{int}$ | Initial concentration of Spo0A~P | 40 nM |
| $f_P$ | Positive feedback coefficient | 60 nM/s |
| $f_N$ | Negative feedback coefficient | 30 nM/s |
| $k_P$ | Conc. of Spo0A~P required for a half-maximum response of positive feedback | 20 nM |
| $k_N$ | Conc. of Spo0A~P required for a half-maximum response of negative feedback | 1 nM |
| $k_{inact}$ | Spontaneous dephosphorylation of Spo0A~P | 0.001 nM |
| $n$ | Hill coefficient of positive feedback | 2 |
| $m$ | Hill coefficient of negative feedback | 5 |

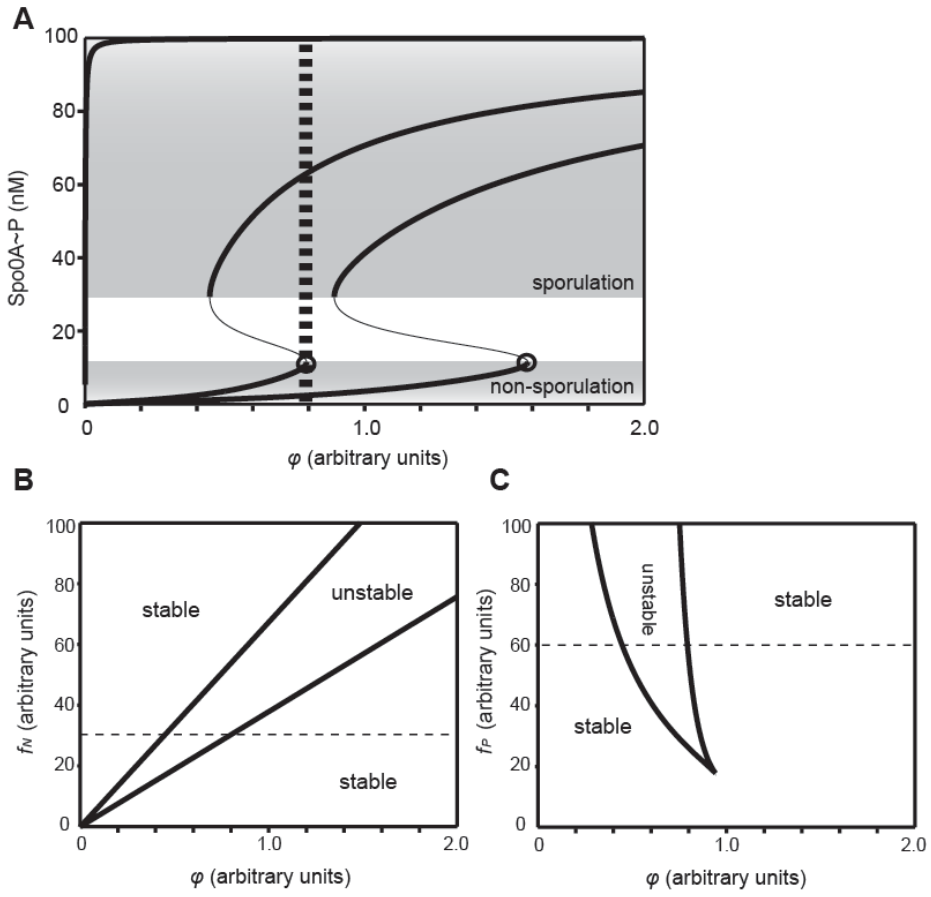


Figure 27: Behavior of the sporulation-decision system upon simulation.

(A) Bifurcation diagram of the concentration of Spo0A~P against the sporulation signal. The curve in the middle represents characteristics under nominal parameter values, corresponding to the wild-type. The left side curve represents that obtained when negative feedback was completely removed; that is, $r = 0$. The right side curve represents that when the negative feedback coefficient was doubled. The circle represents the threshold point of signal intensity and the two shaded regions the parameter space where cells underwent sporulation (upper region) or non-sporulation (lower region), which is dependent on the level of Spo0A~P. Given that the sporulation signal intensity is represented by a dotted line in the center, the wild-type could yield either a low or high level of Spo0A~P, presumably via stochastic fluctuation of the signal (Maughan and Nicholson 2004). Removing negative feedback results in a consistently high level of Spo0A~P, whereas doubling the feedback results in a consistently lower level, yielding only sporulating/non-sporulating subpopulations, respectively. (B and C) A bifurcation diagram obtained by varying the sporulation signal and either feedback signal. Dotted lines indicate the parameter space of nominal values: (B) negative and (C) positive coefficient.

The other parameter values are listed in Table 6. The sporulation signal intensity $\varphi$ is changed between 0-1. The dynamics of the model was examined, and the result is shown in Figure 27. The amount of stimulus $\varphi$ required for switching increases as the ratio of the strength of the negative- and positive-feedback loops $r = f_N/f_P$ increases. Comparing the system characteristics by varying the feedback coefficients ($f_N$ and $f_P$) revealed that as the value of $f_N$ increased, the bistability region shifted its operating region dramatically toward a larger region against the sporulation signal (Figure 27(B)), while $f_P$ did not change its operating region sufficiently (Figure 27(C)). These findings indicate that negative feedback, which is achieved by expression of the spo0E gene, primarily modulates bistability behavior.

While our model is simple and rather phenomenological one, it is suggested that the whole population of *B. subtilis* is divided into two subpopulations (i.e., sporulating or non-sporulating). Improvement of model structure will enable us to analyze the system dynamics further in detail. It is interesting to see that whether adding detail mechanisms on feedback system enhances robustness of the system to generate bistability, as suggested in previous chapter (Morohashi et al. 2002). In particular, Veening and colleagues (Veening et al. 2005) reported that autostimulation of Spo0A should be critical to generate bistability of sporulation, elucidating the dynamics by embedding the autostimulation loop will provide us further insight into the sporulation mechanisms.

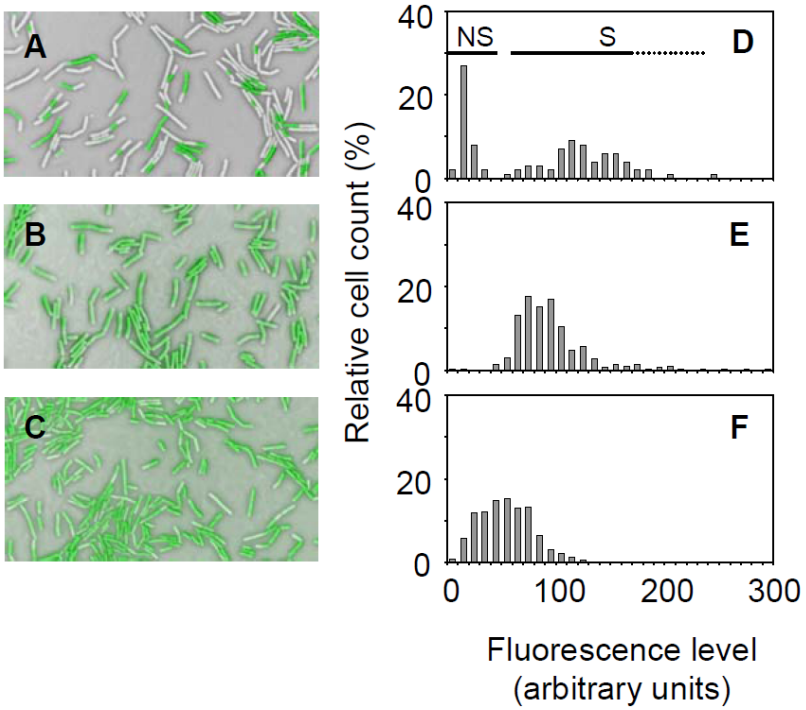


Figure 28: Effects of phosphorelay-associated mutations at sporulation onset. (A to C) Microscopic images of the population distribution of phosphorelay-associated mutants. Wild-type and mutants harboring the P*spoVG-GFPuv* reporter gene were grown under sporulation conditions and observed under a microscope at $T_3$. Green fluorescence and phase-contrast images were inverted and overlaid. (D to F) Distribution of sporulating and non-sporulating subpopulations. The fluorescence intensity of single cells was measured under a fluorescence microscope. Histograms show fluorescence data obtained from more than 300 cells. (A and D) Wild-type (BEST12008), (B and E) *sof1 spo0F*Δ (BEST12007), and (C and F) P*spac-spo0H* (BEST12005). For BEST12005 only we added 1 mM IPTG to the medium at $T_0$. NS and S indicate non-sporulating and sporulating subpopulations, respectively.

## FORMATION OF PHENOTYPIC HETEROGENEITY IN STARVED B. SUBTILIS POPULATION

Here we demonstrate experimentally that wild-type cells at the sporulation phase are distributed into two subpopulations (Figure 28(A), (D)). To detect the sporulating subpopulation we employed $P_{spoVG}$-*GFPuv* transcriptional fusion read by RNAP-$\sigma^H$ at the onset of sporulation. Under that condition, approximately 50% of the wild-type

cells sporulated at 3 hours after the end of the vegetative phase ($T_3$). The decision to sporulate is intrinsic to individuals and not affected by other surrounding cells (Figure 28(A)). These results indicate that the non-sporulating subpopulation emerged upon receiving the fluctuated signal input and then was actively regulated by congenital systems. This is known as polyphenism (Nijhout 2003). We also investigated the function of the phosphorelay pathway on behavior by using strains that harbor *sof1 spo0F*Δ (BEST12007) and $P_{spac}$-*spo0H* (BEST12005). In BEST12005, the expression of *spo0H* is regulated by the $P_{spac}$ promoter and induced by the addition of isopropyl-β-D-galactopyranoside (IPTG) to the medium (Jaacks et al. 1989). As expected, a monostable, Gaussian-like distribution of the population was observed in cultures of BEST12005 (Figure 28(C), (F)). This finding excludes the possibility that the bistable behavior of wild-type populations is generated in unsynchronized cultures. In BEST12007, the *sof1* mutation in *spo0A* accelerates the phosphorylation of mutated Spo0A protein directed by KinC (Quisel et al. 2001) and yields tolerance for dephosphorylation by Spo0E (Stephenson and Perego 2002), resulting in a curtailment of the phosphorelay pathway. This strain also showed a Gaussian-like distribution of the sporulating population (Figure 28(B), (E)), indicating that the phosphorelay pathway is the one critical for the bimodal distribution of sporulating cell populations, yielding sporulating subpopulation in a genetically identical *B. subtilis* society.

The function of Spo0E in polyphenism, which is suggested in our mathematical model (Figure 27), was demonstrated by using BEST12014 (*spo0E*::*cat*) in which the negative-feedback loop by *spo0E* is destroyed ($r = 0$ in our model). In that strain, distribution is excessively biased toward the sporulating subpopulation at $T_3$ (Figure 29(A), (C)), resulting in sporulation of more than 95% of the cells. This was consistent with the sporulation frequency at $T_{24}$. Next we constructed a strain that can highly produce the Spo0E protein. Strain BEST12026 harbors the *spo0E102* mutation that contains an ideal ribosome-binding site and a strong downstream element (Ohashi et al. 2005). In BEST12026, the scale of the sporulating subpopulation at $T_3$ was less than 5% (Figure 29(B), (D)). These results were anticipated by our mathematical

model for phosphorelay (Figure 27) and indicate that the negative-feedback regulation by Spo0E is a gearbox for modulating the sporulating subpopulation.

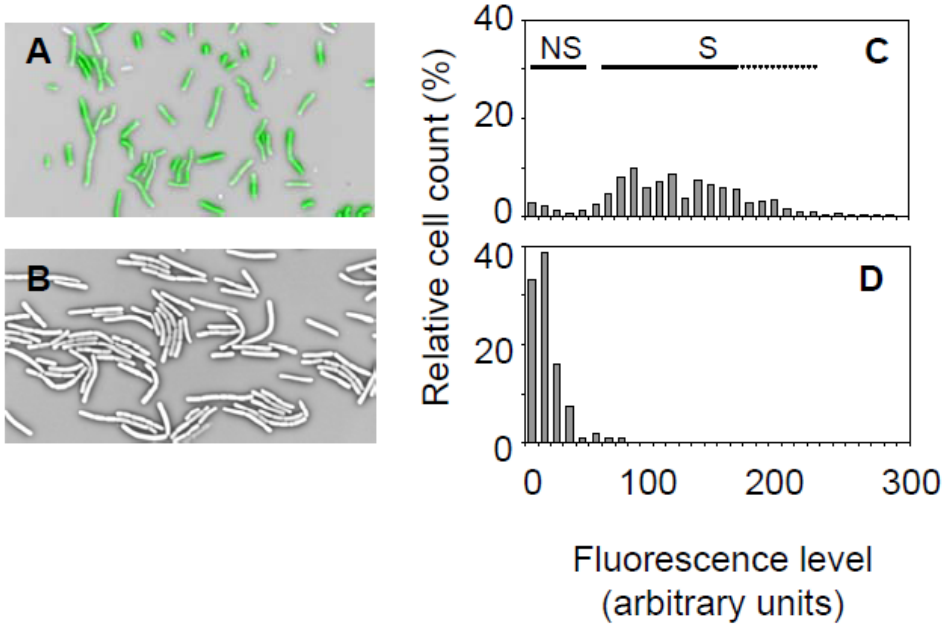


Figure 29: Effects of phosphorelay-associated mutations at sporulation onset.

(A and B) Microscopic images of the population-distribution of phosphorelay-associated mutants. Wild-type and mutants harboring the $P_{spoVG}$-*GFPuv* reporter gene were grown under sporulation conditions and observed under a microscope at $T_3$. Green fluorescence and phase-contrast images were inverted and overlaid. (C and D) Distribution of sporulating and non-sporulating subpopulations. The fluorescence intensity of single cells was measured under a fluorescence microscope. Histograms show fluorescence data obtained from more than 300 cells. (A and C) *spo0E*::*cat* (BEST12014), (B and D) *spo0E102* (BEST12022). NS and S indicate non-sporulating and sporulating subpopulations, respectively.

Table 7: Bacterial strains used in this study.

| Strain | Relevant genotype | Reference and notes |
|---|---|---|
| BEST2131 | *spo0A+ spo0E+ spo0F+ spo0H+ leuB*::pBRTc | (Soga et al. 2003) |
| BEST12008 | *spo0A+ spo0E+ spo0F+ spo0H+ leuB*::pBR::*erm*-P*spoVG-GFPuv* | This study |
| BEST12007 | *sof1 spo0E+ spo0F*Δ *spo0H+ leuB*::pBR::*erm*-P*spoVG-GFPuv* | This study |
| BEST12005 | *spo0A+ spo0E+ spo0F+ spo0H*::pJ0H7d *leuB*::pBR::*erm*-P*spoVG-GFPuv* | This study |
| BEST12013 | *spo0A+ spo0E*::*cat spo0F+ spo0H+* | This study |
| BEST12014 | *spo0A+ spo0E*::*cat spo0F+ spo0H+ leuB*::pBR::*erm*-P*spoVG-GFPuv* | This study |
| BEST12022 | *spo0A+ spo0E+ spo0F+ spo0H+ leuB*::pBR::*cat*-P*spo0E-spo0E102* | This study |
| BEST12026 | *spo0A+ spo0E+ spo0F+ spo0H+ metB::pBR::cat-Pspo0E-spo0E102 leuB*::*pBR*::*erm-PspoVG-GFPuv proB*::*pBRBS* | This study |
| BEST12033 | *spo0A+ spo0E*::*cat spo0F+ spo0H+ leuB*::pBRTc | This study |

DISTINCT METABOLOME PROFILES AMONG SPORULATION STAGES.

To investigate the effects of population heterogeneity on omics data, we demonstrated the variation of metabolome data in wild type (*spo0E*$^{+}$), *spo0E*::*cat* (*spo0E*$^{-}$), and *spo0E102* (*spo0E*$^{++}$) strains. The detailed mechanisms and association between activities of the metabolic pathway and sporulation remain relatively unknown (Dworkin and Losick 2001), compared to genetic and protein level analyses. All inactivating mutations in *B. subtilis* Krebs cycle genes cause a defect, and terminate at certain stages during sporulation (see (Sonenshein 2002) for a review). Phosphorelay of Spo0A~P is thought to be controlled by the Krebs cycle, but a detailed metabolomics approach has yet to be performed. By comparing wild-type and *sdpC* knockout strains, Gonzalez-Pastor *et al.* reported that ATP synthase is strongly expressed during sporulation (Gonzalez-Pastor et al. 2003). This suggests a link between energy metabolism and the sporulation signaling pathway. We previously conducted metabolome analysis in wild-type *B. subtilis*, confirming that most glycolysis metabolites are markedly decreased in the early stage of sporulation (Soga et al. 2003). However, we did not compare the metabolic profiles of sporulation-deficient strains, thus making it difficult to determine which pathways are critical or which metabolites are strongly correlated to sporulation activities. Recently, the relationships between branched-chain amino acids and CodY were revealed (Ratnayake-Lecamwasam et al. 2001; Shivers and Sonenshein 2004; Sonenshein 2005). CodY protein controls more than one hundred genes that are induced when cells experience nutrient deprivation. GTP and isoleucine independently and additively increase the affinity of CodY toward its target sites, resulting in activation of its repressor function. To obtain more details on the metabolic profiles during sporulation, *i.e.*, to obtain a better understanding of the metabolic pathway as a whole, we therefore conducted metabolome analysis based on CE-TOFMS (Soga et al. 2003).

We employed three strains carrying various spo0E alleles in their genomes: BEST2131 (wildtype), BEST12022 (*spo0E102*), and BEST12033 (*spo0E*::*cat*). The genetic background of these strains is *trpC2* and *leuB*::*pBR*, resulting in L-tryptophan

and L-leucine auxotrophy. The growth characteristics of the strains were equivalent in 2x SG sporulation medium including 0.1% (w/v) of D-glucose (Figure 30). Each strain was sampled at four time points, $T_{-0.5}$, $T_0$, $T_{1.5}$, and $T_3$, relative to the end of the logarithmic growth phase. These time points approximately correspond to the middle-logarithmic phase, transition phase, the time when the final symmetric septation is completed, and the time when sporulation-specific asymmetric septation is completed in wild-type cells, respectively. We prepared metabolome extracts and performed CE-TOFMS analysis (see Materials and Methods).
To demonstrate whether each strain can be characterized independently by its metabolic profile, we performed PCA using the metabolome data, as is frequently conducted against omic data (Raamsdonk et al. 2001). We first selected the signals that were significantly different among the three strains by employing one-way ANOVA under a 5% significant level; it resulted in extracting 94 metabolites, which were then used for PCA. Due to aberrant peak shape, in some cases, our software cannot detect peaks, causing to have missing values. In order to avoid contamination of the missing peaks, the 94 peaks were extracted from the peaks which are automatically detected over all experiments, resulting in selection of peaks in very strict manner. Improvement of software and measurement technology will have higher accuracy of peak extraction, and will lead to characterize metabolic profiles in more detail. As shown in Figure 31(A), the phase transitions could be clearly traced within the three-dimensional principal component spaces. Intriguingly, we found that BEST12033 (*spo0E*$^-$) at $T_{-0.5}$ was not explicitly discriminated from that at $T_0$. This result indicates that the metabolic state during the logarithmic growth phase is similar to that at transition phase in BEST12033 (*spo0E*$^-$). This may be due to the function of Spo0E phosphatase not only at sporulation onset, but also in the logarithmic phase, i.e., Spo0A~P, which is slightly generated during the vegetative phase, is canceled by the positively-regulated spo0E product, inhibiting the initiation of sporulation. However, at present this hypothesis is no more than speculation. Because the cumulative contribution up to the third principal component was approx. 70%, up to the third principal component may not be still enough to capture the full characteristics of the sample. However, all other samples except BEST12033 (*spo0E*$^-$)

exhibited distinct results at $T_{-0.5}$. As depicted in Figure 31 (B) and (C), the three samples were clearly discriminated at $T_{1.5}$ and $T_3$, indicating that their metabolic states are clearly different at the onset of sporulation. We concluded that this was due to the population heterogeneity of sporulating wild-type cells; the metabolic profiles of wild-type cells may mislead our interpretation. To investigate the sporulating/non-sporulating subpopulations in a more distinct manner, we compare extreme cases hereafter (*i.e.*, BEST12022 and BEST12033).

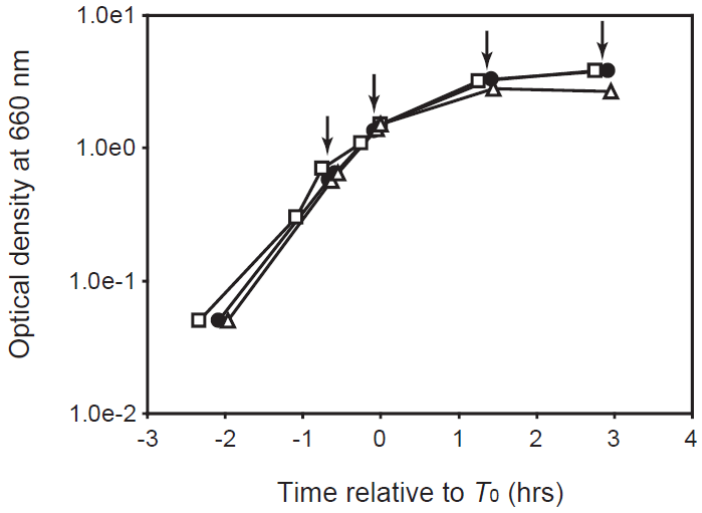


Figure 30: Growth curve of examined strains.

Representative data from 3 samples is shown. *spo0E*$^+$ (closed circles), *spo0E*::*cat* (open squares), and *spo0E102* (open triangles). Sampling points ($T_{-0.5}$, $T_0$, $T_{1.5}$, and $T_3$) are indicated by arrows.

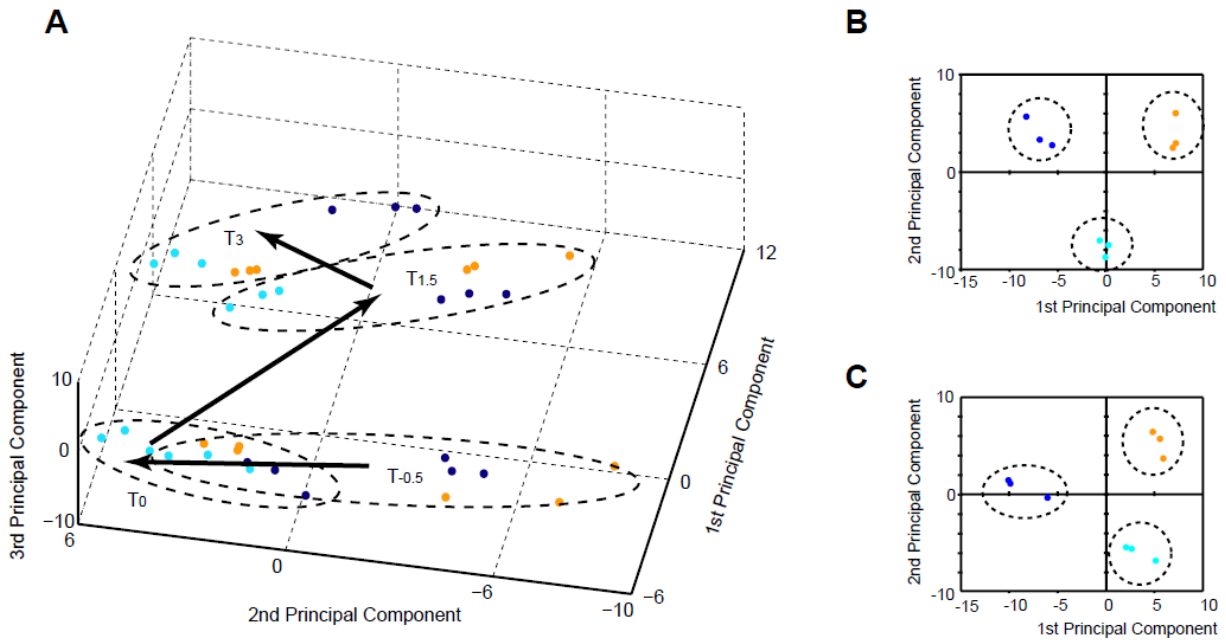


Figure 31: The metabolic state of sporulating *B. subtilis*.

(A) A three-dimensional principal component plot of the metabolome profiles of the sporulating spo0E variants (n = 3). Transition of the metabolic profiles of the three examined strains is indicated by arrows. (B and C) Two-dimensional principal component plots at $T_{1.5}$ (B) and $T_3$ (C). Dark blue, BEST2131 (*spo0E*$^+$); light blue, BEST12033 (*spo0E*::*cat*); orange, BEST12022 (*spo0E102*).

ENERGY METABOLISM EXHIBITS DISTINCT FEATURES UPON SPORULATION

Sporulation is initiated by deprivation of nutrients such as carbon, nitrogen, and phosphate sources. Further, under our experimental conditions, the glucose level in the medium was considered an important factor controlling the initiation of sporulation, since the addition of more than 0.5% (w/v) D-glucose efficiently inhibited sporulation (data not shown). We therefore expected glucose utilization pathways including glycolysis, the Krebs cycle, and the pentose phosphate pathway to drastically fluctuate at sporulation onset. The overall metabolome profiles are depicted in Figure 33. Intracellular levels of metabolic intermediates, especially fructose-1,6-bisphosphate (F1,6P) and acetyl CoA, were reduced upon sporulation onset (Figure 33). This suggests that the glycolysis pathway is activated during growth phase by the aggressive use of glucose. This is consistent with the data obtained in our previous report (Soga et al. 2003). Significant differences were observed in lactic acid levels between the two spo0E variants. In BEST12033 (*spo0E*$^{-}$), the intracellular level was higher than that in BEST12022 (*spo0E*$^{+++}$) throughout cultivation, suggesting that trace amounts of Spo0A~P accumulated in vegetative BEST12033 (*spo0E*$^{-}$) cells likely activate the glycolysis pathway. In the Krebs cycle, on the other hand, different profiles were obtained for intermediate metabolites (Figure 33). In particular, the metabolism pathway seemed significantly altered after the initiation of sporulation. In both spo0E variants, citrate levels were drastically increased after $T_0$, but not generated until $T_0$. This observation suggests that the Krebs pathway, which is dormant during the logarithmic growth phase, is activated at stationary phase when environmental glucose is spent. Behavior of the downstream pathway of citrate was also different between the two strains; 2-oxoglutararic acid was highly accumulated in sporulating BEST12033 (*spo0E*$^{-}$). This result suggests that metabolism, at least that involving 2-oxoglutaric acid in the Krebs cycle, is required for metabolic differentiation towards sporulation. Inactivation of citrate synthase and aconitase, which catalyze this metabolism, results in a sporulation-deficient phenotype at early stages of sporulation (Sonenshein

2002). Nitrogen metabolism is also considered a key factor in sporulation, because the pathway from 2-oxoglutaric acid to glutamine via glutamic acid is the solo acceptor of ammonium ions. Metabolite levels of glutamine, as well as that of purines, were also shown to increase after $T_{1.5}$. The downstream metabolites of 2-oxoglutaric acid, succinic acid, fumaric acid, and malic acid, were highly accumulated in stationary BEST12022 (*spo0E*$^{++}$). This observation suggests that the latter pathway of the Krebs cycle is blocked or rapidly circulated in sporulating BEST12033 (*spo0E*$^{-}$).

## CHANGES IN THE ENERGY CHARGE STATE DURING SPORULATION

In sporulating BEST12033 (*spo0E*$^{-}$), we could not decide whether or not the latter steps of the Krebs cycle were activated. If activated, ATP should be generated by an electron transfer system using NAD(P)H. We therefore determined the levels of phosphorylated adenosine (AXP) during sporulation (Figure 32). Unexpectedly, sporulating cells (BEST12033, *spo0E*$^{-}$) exhibited a dramatic increase in ATP levels, while non-sporulating cells (BEST12022, *spo0E*$^{++}$) decreased its level after $T_{1.5}$. These results strongly support the data of Gonzalez-Pastor *et al.* (Gonzalez-Pastor et al. 2003) who showed, from a metabolite viewpoint, that ATP synthase is strongly expressed upon sporulation. It was previously reported that ATP remains approximately constant during the sporulation process; however, this conclusion was based on the outcomes of observations using a wild-type strain, which generates heterologous culture of both sporulating and non-sporulating cells. If these distinct features discriminate between sporulating/non-sporulating cells, our results suggest that this past study presented the average results of both subpopulations, thus canceling out the sporulation/non-sporulation effect.

Levels of phosphorylated guanosines (GXP) were also determined (Figure 32) since a decrease in GTP is thought to trigger sporulation (Beaman et al. 1983). We observed a decrease in GTP at $T_0$, as well as a decrease in inosine-5'-phosphate (IMP) at the same time. The GTP level increased at $T_3$ only in BEST12033 (*spo0E*$^-$), supporting ATP production at the onset of sporulation. This result indicates that sporulating cells accumulate energy to prepare for the energy-consuming process of spore formation. Intriguingly, several nucleotides (AMP, ADP, GMP, and GDP) increased transiently at $T_0$ in both strains. It would therefore be interesting to examine whether their transient activities, in addition to that of GTP, are linked to the initiation of sporulation.

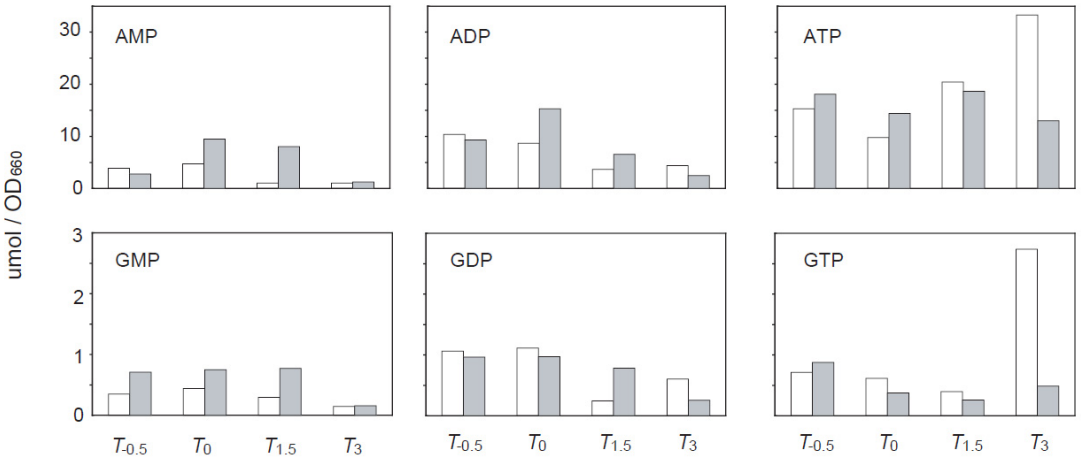


Figure 32: Metabolic profiles of nucleotides.

White bars, BEST12033. Grey bars, BEST12022.

## MOST AMINO ACIDS ARE INDEPENDENT OF THE SPORULATION PROCESS

All amino acids except for cysteine, which was under the detection limit, were categorized according to their transition characteristics. The results are summarized in Table 8. Most amino acids exhibited similar characteristics when comparing the two strains. An association between branched-chain amino acids (e.g., isoleucine, leucine and valine) and CodY, a GTP-binding protein, has been shown (Molle et al. 2003). In BEST12033 (*spo0E*$^{-}$), valine (but not isoleucine) exhibited distinct characteristics compared with BEST12022 (*spo0E*$^{++}$). That is, its level increased transiently at $T_0$ then dropped again, which is consistent with the finding that CodY directly targets its biosynthesis (Ratnayake-Lecamwasam et al. 2001; Shivers and Sonenshein 2004; Sonenshein 2005). Asparagine and glutamine are nitrogen-rich amino acids and central components of nucleic acid production. It is suggested that in combination with the nucleotide results, *i.e.*, the massive generation and thus accumulation, the increased levels of these amino acids may be attributed to sporulation preparation.

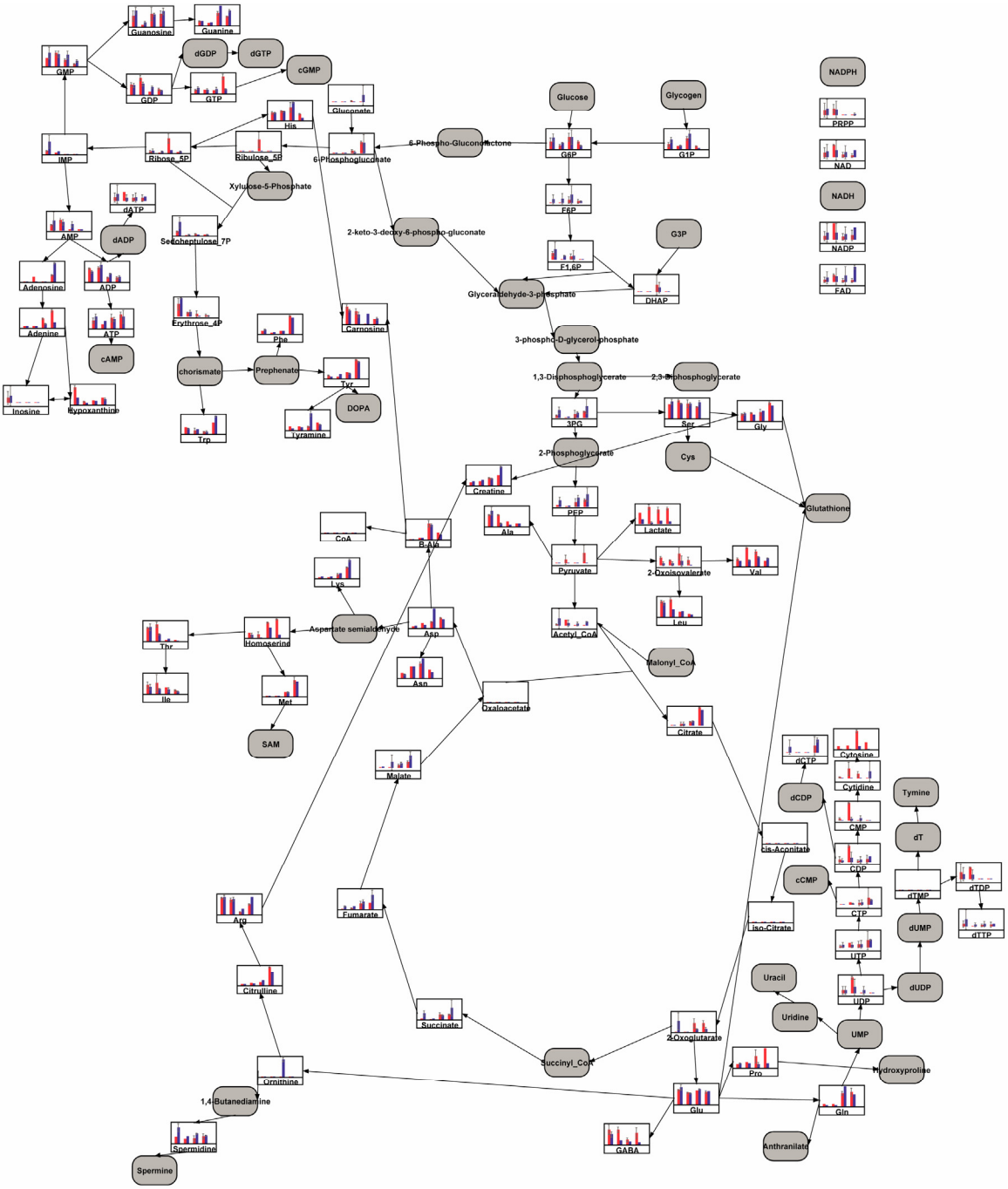


Figure 33. Metabolic profiling of *B. subtilis*.

Average (+ standard deviation) were represented. Red; *spo0E*::*cat*. Blue; *spo0E102*.

Table 8: Clustering of amino acids.

Each amino acid is categorized into 5 classes, depending on tendency of time series profiles. I: Increase during sporulation transition and decrease; II: Decrease; III: Decrease during sporulation transition and increase; IV: Increase; and V: No change. Amino acids exhibiting different profiles were highlighted as grey area.

| Amino acid | Spo0Ed | SpoEh | Pathway |
|---|---|---|---|
| **Ser** | I | V | Glycolysis |
| **Val** | I | IV | Glycolysis |
| Ala | II | II | Glycolysis |
| Gly | IV | IV | Glycolysis |
| Leu | II | II | Glycolysis |
| Trp | III | III | Pentose |
| His | I | I | Pentose |
| Phe | IV | IV | Pentose |
| Tyr | IV | IV | Pentose |
| **Asp** | IV | I | Krebs |
| **Gln** | IV | I | Krebs |
| **Pro** | IV | V | Krebs |
| Arg | III | III | Krebs |
| Asn | I | I | Krebs |
| Ile | II | II | Krebs |
| Lys | IV | IV | Krebs |
| Met | IV | IV | Krebs |
| Thr | II | II | Krebs |
| Glu | V | V | Krebs |

# CONCLUSION

We investigated the mechanism of heterogeneity during sporulation in wild-type *B. subtilis*. Using simulation and modeling techniques, we found that negative feedback is a primary modulating factor in the bistability of sporulation, which is directly affected by the *spo0E* gene. We then confirmed these results experimentally by deleting/overexpressing *spo0E*. The findings mathematically support the proposals suggested in previous reports (Veening et al. 2005), and suggest that population heterogeneity should be considered in omics studies including transcriptomics, proteomics, and metabolomics.

In addition, we also examined the sporulation stages by metabolome analysis. To investigate the sporulating and non-sporulating stages in a distinct manner, we used the mutants mentioned above (i.e., BEST12033 and BEST12022) and compared transition of metabolite levels. As a result, we found that metabolism was significantly different among each stage regardless of whether the population was sporulating or non-sporulating. Although we need to further investigate each metabolite in detail, this study provides enormous information suggesting links between metabolome activities and spore formation. The inclusion of additional mutants (such as spo0A) will provide further insight into the molecular mechanisms of sporulation as well as cell differentiation.

# CHAPTER 6: CONCLUSION

*We ourselves feel that what we are doing is just a drop in the ocean. But the ocean would be less because of that missing drop.*

— Mother Teresa

In this thesis, we developed methods for various analyses toward system level understanding of life. Owing to recent advances of numbers of measurement technologies, we are now encountering unexpected and astounding results from biological systems. To fully utilize those cutting edge technologies, however, analysis workflow and infrastructure must be appropriately developed. We developed methods and tools, and applied to simulation analyses. Furthermore, combination of omics approach and simulation approach enabled us to get further insight into biological mechanisms.

# SUMMARY OF RESULTS

## DEVELOPMENT OF ANALYSIS TOOLS AND METHODS

In order to utilize top-down/bottom-up approaches in efficient manner, various methods and tools were developed.

First, CellDesigner was designed and developed for modeling and simulation purpose. As general purpose drawing tools (such as Adobe Illustrator) achieved, basic biological models could be drawn easily without having any mathematical knowledge. Further enhancement of the models is possible by embedding reaction rules, mathematical equations, which are then saved as a file in SBML format. SBML files can be imported to over 110 software tools for further analyses, but CellDesigner has now capability to directly invoke simulation tools. CellDesigner thus could be a suitable medium to work on simulation and other analyses. In addition, the employed notation in CellDesigner is rigidly defined ((Kitano et al. 2005), now underway to be defined as SBGN (Systems Biology Graphical Notation)), models themselves could be used as presentation materials or visual models as well.

Second, to efficiently process metabolome data, a filtering method P-BOSS was developed. The method was developed to minimize unnecessary filtering so that as much as peaks can be left for further analyses. Combining with outlier detection method, P-BOSS showed superior performance to traditional method. The method is now used as a default method along analysis workflow in Human Metabolome Technologies.

## APPLICATION TO BIOLOGICAL MODELS

The above methods and tools were applied to investigate two models; that is, 1) cell cycle model in *Xenopus*, and 2) sporulation mechanisms in *B. subtilis.*

The former research compared two models, representing cell cycle of *Xenopus*. On the assumption that biological models have gained their evolvability toward getting robustness against various environmental stimuli, robustness was used as a measure to validate plausibility between models. The two models were different in their abstraction level, feedback mechanisms were phenomenologically modeled at one hand, and detail molecular level mechanisms were embedded at the other hand. As the results, the latter model showed more robustness against various kinetic parameters, suggesting that as the model is refined as more detail of molecular mechanisms are known, which leads to obtain robustness, which is analogous to how biological systems evolved.

On the other hand, the latter approach analyzed sporulation mechanism in *B. subtilis* by building a mathematical model and experimentally verified. Besides, metabolome analyses were performed to see the mechanism from omics standpoint. As the results, energy metabolism exhibited significant difference between sporulating and non-sporulating cells, suggesting close link between metabolism and sporulation. This approach enabled us to capture fundamental mechanism of sporulation, as well as showing dynamics of metabolic pathway to see phenomena from other direction. The omics data could thus be used not only from information retrieval viewpoint, but also from complement understanding of phenomena from systems biology perspectives.

The former analysis, using cell cycle model in *Xenopus*, is rather methodology research, and thus could be applied to the latter model, although we have not performed the analysis due to lack of detail data.

# FUTURE DIRECTIONS

## ISSUES IN SYSTEMS BIOLOGY

As Kitano mentioned in (Kitano 2002), a system-level understanding of a biological system can be derived from insight into four key properties. To quote them,

1. *System structures*. These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures.

2. *System dynamics*. How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved.

3. *The control method*. Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.

4. *The design method*. Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

Our approach in this thesis is to focus on 1 and 2 among four properties. Even microorganisms, such as *E. coli* and *B. subtilis*, have still huge numbers of black box

in regulation mechanisms, although they have been extensively investigated from molecular to physiology level.

The questions in our research, for example in sporulation mechanism, are followings:

1. Which factors are the keys to determine if they should proceed to sporulate or not?
2. Which network is the significant module to generate bistability?
3. How did the bistability evolve? Do we human have similar strategy to survive?
4. How can we control the system to tightly and quantitatively determine the sporulation ratio (e.g., decrease sporulation rate to 7.5%)?
5. How can we design a system, having robust bistable characteristics?

The questions above are just few examples of what has come to our mind at the moment. As the network and relevant components are revealed, we may be able to answer the questions completely. It should be matter of time.

## SYSTEMS BIOLOGY IN INDUSTRIES

Systems biology approach is getting paid attention from wide spectrum of fields, including pharmaceutical and medical fields. For example, simulation approach has now reflected by founding companies, particularly in the United States region. To cite few of them:

- Gene Network Sciences, Inc. (http://gnsbiotech.com)
- Entelos, Inc. (http://entelos.com)
- Genomatica, Inc. (http://genomatica.com)

All companies above are working on simulation based approach (although Gene Network Sciences is now shifting toward knowledge inferring methodology from large scale data). They have already started collaborating research with various pharmaceutical companies, or food production, fermentation-based companies. It should be note that there are demands from such industries, suggesting that they are looking for another approach other than traditional approaches.

Among them, Genomatica is focusing on metabolomics research. Their basic business model is to apply metabolic analysis (e.g., flux balance analysis (see (Schilling et al. 1999; Schilling et al. 2000) for example). Although their premise is limited at some point, theory itself is rigidly established and have a long history in the field, thus their business may interests companies utilizing metabolome data.

While simulation based approach has emerged recently, and still at the dawn of systems biology, omics approach is now widely applied not only in academic field, but also in medical, or pharmaceutical fields. Genome based drug discovery is the perfect example. In contrast to past traditional methodology, with almost random and extensive screening of potential targets, or looking various receptors drug discovery starts from looking into genome. The functions in the genome can be obtained from other omics data such as transcriptome, proteome, and metabolome data. Once potential functions of targets are found, next stage comes up using combinatorial chemistry, or high throughput screening. These processes enable researches to be rationale and cost effective, and are expected to boost personalized medicine.

# FINAL REMARKS

It is said for many years recently that “understanding life” is a grand challenge (or could be ultimate goal!) for us human, and cannot be achieved without involving wide variety of background professionals; those could be biologist, mathematicians, computer scientist, electrical engineers, analytical chemists, and so forth.

We believe that this work significantly contributes to the research in system biology field, enabling researchers to handle data and analyze in efficient manner, or even provide insight into analyzing approach.

# BIBLIOGRAPHY


Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397(6715): 168-171.

Alves R, Savageau MA (2000) Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. Bioinformatics 16(6): 534-547.

Arita M (2004) The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A 101(6): 1543-1547.

Arita M (2005) Scale-freeness and biological networks. J Biochem (Tokyo) 138(1): 1-4.

Baba T, Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology.

Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. Science 305(5690): 1622-1625.

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439): 509-512.

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2): 101-113.

Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387(6636): 913-917.

Beaman TC, Hitchins AD, Ochi K, Vasantha N, Endo T et al. (1983) Specificity and control of uptake of purines and other compounds in Bacillus subtilis. J Bacteriol 156(3): 1107-1117.

Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405(6786): 590-593.

Becskei A, Seraphin B, Serrano L (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. Embo J 20(10): 2528-2535.

Borisuk MT (1997) Ph.D. thesis. Virginia Polytechnic Institute and State University.

Borisuk MT, Tyson JJ (1998) Bifurcation analysis of a model of mitotic control in frog eggs. J Theor Biol 195(1): 69-85.

Broadhurst DI, and Kell, D. B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2: 171-196.

Burbulys D, Trach KA, Hoch JA (1991) Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay. Cell 64(3): 545-552.

Carlson JM, Doyle J (2000) Highly optimized tolerance: robustness and design in complex systems. Phys Rev Lett 84(11): 2529-2532.

Chung JD, Stephanopoulos G (1995) Studies of transcriptional state heterogeneity in sporulating cultures of *Bacillus subtilis*. Biotechnology and Bioengineering 47: 234-242.
Chung JD, Stephanopoulos G, Ireton K, Grossman AD (1994) Gene expression in single cells of Bacillus subtilis: evidence that a threshold mechanism controls the initiation of sporulation. J Bacteriol 176(7): 1977-1984.
Clarke BL (1980) Stability of complex reaction networks. Advances in Chemical Physics. New York: John Wiley. pp. 1-215.
Clarke BL (1994) Steady state bifurcation hypersurfaces of chemical mechanisms. Comparison Methods and Stability Theory, Lecture Notes in Pure and Applied Mathematics. New York: Marcel Dekker. pp. 67-85.
Clauss MJ, Venable DL (2000) Seed Germination in Desert Annuals: An Empirical Test of Adaptive Bet Hedging. Am Nat 155(2): 168-186.
Cohen D (1967) Optimizing reproduction in a randomly varying environment when a correlation may exist between the conditions at the time a choice has to be made and the subsequent outcome. J Theor Biol 16(1): 1-14.
Cook DL, Farley JF, Tapscott SJ (2001) A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. Genome Biol 2(4): RESEARCH0012.
Csete ME, Doyle JC (2002) Reverse engineering of biological complexity. Science 295(5560): 1664-1669.
Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A 97(12): 6640-6645.
de Jong H, Geiselmann J, Batt G, Hernandez C, Page M (2004) Qualitative simulation of the initiation of sporulation in *Bacillus subtilis*. Bulletin of Mathematical Biology 66: 261-299.
Dearden P, Akam M (2000) Segmentation in silico. Nature 406(6792): 131-132.
Doedel EJ. AUTO: a program for the automatic bifurcation analysis of autonomous systems; 1981. pp. 265-284.
Dworkin J, Losick R (2001) Linking nutritional status to gene activation and development. Genes Dev 15(9): 1051-1054.
Eanes WF (1999) Analysis of selection on enzyme polymorphisms. Annu Rev Ecol Syst 30: 301-326.
El-Samad H, Kurata H, Doyle JC, Gross CA, Khammash M (2005) Surviving heat shock: control strategies for robustness and performance. Proc Natl Acad Sci U S A 102(8): 2736-2741.
Ermentrout B (2002) Simulating, Analyzing, and Animating Dynamical Systems. Philadelphia: SIAM Press.
Fell D (1997) Understanding the control of metabolism. London: Portland Press.
Ferrell JE, Jr., Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280(5365): 895-898.
Ferrell JE, Jr., Wu M, Gerhart JC, Martin GS (1991) Cell cycle tyrosine phosphorylation of p34cdc2 and a microtubule-associated protein kinase homolog in Xenopus oocytes and eggs. Mol Cell Biol 11(4): 1965-1971.
Fiehn O (2002) Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol 48(1-2): 155-171.

Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal Chem 72(15): 3573-3580.
Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN et al. (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18(11): 1157-1161.
Freeman M (2000) Feedback control of intercellular signalling in development. Nature 408(6810): 313-319.
Fujita M, Gonzalez-Pastor JE, Losick R (2005) High- and low-threshold genes in the Spo0A regulon of Bacillus subtilis. J Bacteriol 187(4): 1357-1368.
Funahashi A, Morohashi M, Tanimura N, Kitano H (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 1: 159-162.
Gonzalez-Pastor JE, Hobbs EC, Losick R (2003) Cannibalism by sporulating bacteria. Science 301(5632): 510-513.
Groisman I, Huang YS, Mendez R, Cao Q, Theurkauf W et al. (2000) CPEB, maskin, and cyclin B1 mRNA at the mitotic apparatus: implications for local translational control of cell division. Cell 103(3): 435-447.
Grossman AD (1995) Genetic networks controlling the initiation of sporulation and the development of genetic competence in Bacillus subtilis. Annu Rev Genet 29: 477-508.
Harrigan GG, LaPlante RH, Cosma GN, Cockerell G, Goodacre R et al. (2004) Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. Toxicol Lett 146(3): 197-205.
Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(6761 Suppl): C47-52.
Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T et al. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proc Natl Acad Sci U S A 101(27): 10205-10210.
Hoch JA, Trach K, Kawamura F, Saito H (1985) Identification of the transcriptional suppressor sof-1 as an alteration in the spo0A protein. J Bacteriol 161(2): 552-555.
Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696): 640-643.
Hopper KR (1999) Risk-spreading and bet-hedging in insect population biology. Annu Rev Entomol 44: 535-560.
Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4): 524-531.
Iber D (2006) A quantitative study of the benefits of co-regulation using the spoIIA operon as an example. Mol Syst Biol 2: 43.
Iber D, Clarkson J, Yudkin MD, Campbell ID (2006) The mechanism of cell differentiation in Bacillus subtilis. Nature 441(7091): 371-374.

Ideker TE, Thorsson V, Karp RM (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. Pac Symp Biocomput: 305-316.
Ishii N, Nakahigashi K, Baba T, Robert M, Soga T et al. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 316(5824): 593-597.
Itaya M (1993) Integration of repeated sequences (pBR322) in the Bacillus subtilis 168 chromosome without affecting the genome structure. Mol Gen Genet 241(3-4): 287-297.
Jaacks KJ, Healy J, Losick R, Grossman AD (1989) Identification and characterization of genes controlled by the sporulation-regulatory gene spo0H in Bacillus subtilis. J Bacteriol 171(8): 4121-4129.
Jarvis RM, Goodacre R (2005) Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. Bioinformatics 21(7): 860-868.
Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407(6804): 651-654.
Jin S, Levin PA, Matsuno K, Grossman AD, Sonenshein AL (1997) Deletion of the Bacillus subtilis isocitrate dehydrogenase gene causes a block at stage I of sporulation. J Bacteriol 179(15): 4725-4732.
Kacser H, Burns JA (1973) The control of flux. Symp Soc Exp Biol 27: 65-104.
Kadota K, Miki R, Bono H, Shimizu K, Okazaki Y et al. (2001) Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. Physiol Genomics 4(3): 183-188.
Kadota K, Nishimura S, Bono H, Nakamura S, Hayashizaki Y et al. (2003) Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. Physiol Genomics 12(3): 251-259.
Kell DB (2004) Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol 7(3): 296-307.
Kitano H (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet 41(1): 1-10.
Kitano H (2002) Systems biology: a brief overview. Science 295(5560): 1662-1664.
Kitano H (2004) Biological robustness. Nat Rev Genet 5(11): 826-837.
Kitano H (2007) A robustness-based approach to systems-oriented drug design. Nat Rev Drug Discov.
Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23(8): 961-966.
Kohn KW (2001) Molecular interaction maps as information organizers and simulation guides. Chaos 11(1): 84-97.
Kohn KW, Aladjem MI, Weinstein JN, Pommier Y (2006) Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. Mol Biol Cell 17(1): 1-13.
Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G et al. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390(6657): 249-256.
Kyoda K, Baba K, Onami S, Kitano H (2004) DBRF-MEGN method: an algorithm for deducing minimum equivalent gene networks from large-scale gene

expression profiles of gene deletion mutants. Bioinformatics 20(16): 2662-2675.
Kyoda KM, Morohashi M, Onami S, Kitano H (2000) A gene network inference method from continuous-value gene expression data of wild-type and mutants. Genome Inform Ser Workshop Genome Inform 11: 196-204.
Leighton TJ, Doi RH (1971) The stability of messenger ribonucleic acid during sporulation in Bacillus subtilis. J Biol Chem 246(10): 3189-3195.
Levin PA, Losick R (1996) Transcription factor Spo0A switches the localization of the cell division protein FtsZ from a medial to a bipolar pattern in Bacillus subtilis. Genes Dev 10(4): 478-488.
Liang S, Fuhrman S, Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput: 18-29.
Ma L, Iglesias PA (2002) Quantifying robustness of biochemical network models. BMC Bioinformatics 3: 38.
Maamar H, Dubnau D (2005) Bistability in the Bacillus subtilis K-state (competence) system requires a positive feedback loop. Mol Microbiol 56(3): 615-624.
Maimon R, Browing S. Notation and computational structure of gene networks; 2001; Pasadena, U.S.A. pp. 311-317.
Marlovits G, Tyson CJ, Novak B, Tyson JJ (1998) Modeling M-phase control in Xenopus oocyte extracts: the surveillance mechanism for unreplicated DNA. Biophys Chem 72(1-2): 169-184.
Masui Y, Wang P (1998) Cell cycle transition in early embryonic development of Xenopus laevis. Biol Cell 90(8): 537-548.
Maughan H, Nicholson WL (2004) Stochastic processes influence stationary-phase decisions in Bacillus subtilis. J Bacteriol 186(7): 2212-2214.
Molle V, Fujita M, Jensen ST, Eichenberger P, Gonzalez-Pastor JE et al. (2003) The Spo0A regulon of Bacillus subtilis. Mol Microbiol 50(5): 1683-1701.
Molle V, Nakaura Y, Shivers RP, Yamaguchi H, Losick R et al. (2003) Additional targets of the Bacillus subtilis global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis. J Bacteriol 185(6): 1911-1922.
Morohashi M, Kitano H. Identifying gene regulatory networks from time series expression data by in silico Sampling and Screening; 1999; Lausanne, Switzerland. pp. 477-486.
Morohashi M, Winn AE, Borisuk MT, Bolouri H, Doyle J et al. (2002) Robustness as a measure of plausibility in models of biochemical networks. J Theor Biol 216(1): 19-30.
Morohashi M, Shimizu K, Ohashi Y, Abe J, Mori H et al. (2007) P-BOSS: A new filtering method for treasure hunting in metabolomics. J Chromatography A: (in press).
Morohashi M, Ohashi Y, Tani S, Ishii K, Itaya M et al. (2007) Model based definition of population heterogeneity and its effects on metabolism in sporulating Bacillus subtilis. J Biochem (Tokyo).
Murray A, Hunt T (1993) The Cell Cycle. London: The Oxford University Press.
Murray AW, Kirschner MW (1989) Cyclin synthesis drives the early embryonic cell cycle. Nature 339(6222): 275-280.

Nanamiya H, Fugono N, Asai K, Doi RH, Kawamura F (2000) Suppression of temperature-sensitive sporulation mutation in the Bacillus subtilis sigA gene by rpoB mutation. FEMS Microbiol Lett 192(2): 237-241.
Nijhout HF (2003) Development and evolution of adaptive polyphenisms. Evol Dev 5(1): 9-18.
Ohashi Y, Ohshima H, Tsuge K, Itaya M (2003) Far different levels of gene expression provided by an oriented cloning system in Bacillus subtilis and Escherichia coli. FEMS Microbiol Lett 221(1): 125-130.
Ohashi Y, Sugimaru K, Nanamiya H, Sebata T, Asai K et al. (1999) Thermo-labile stability of sigmaH (Spo0H) in temperature-sensitive spo0H mutants of Bacillus subtilis can be suppressed by mutations in RNA polymerase beta subunit. Gene 229(1-2): 117-124.
Ohashi Y, Yamashiro A, Washio T, Ishii N, Ohshima H et al. (2005) In silico diagnosis of inherently inhibited gene expression focusing on initial codon combinations. Gene 347(1): 11-19.
Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. Nature 427(6976): 737-740.
Paredes CJ, Alsaker KV, Papoutsakis ET (2005) A comparative genomic view of clostridial sporulation and physiology. Nat Rev Microbiol 3(12): 969-978.
Parrilo PA (2000) Ph.D. thesis. Control and Dynamical Systems. California Institute of Technology.
Perego M, Hoch JA (1991) Negative regulation of Bacillus subtilis sporulation by the spo0E gene product. J Bacteriol 173(8): 2514-2520.
Perego M, Hoch JA (2002) Two-component systems, phosphorelays, and regulation of their activities by phosphatases. In: Sonenshein AL, Hoch JA, Losick R, editors. *Bacillus subtilis* and its closest relatives: from genes to cells. Washington DC: ASM Press. pp. 473-481.
Piggot PJ, Hilbert DW (2004) Sporulation of Bacillus subtilis. Curr Opin Microbiol 7(6): 579-586.
Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE et al. (2000) The visual display of regulatory information and networks. Trends Cell Biol 10(10): 404-408.
Pogliano J, Osborne N, Sharp MD, Abanes-De Mello A, Perez A et al. (1999) A vital stain for studying membrane dynamics in bacteria: a novel mechanism controlling septation during Bacillus subtilis sporulation. Mol Microbiol 31(4): 1149-1159.
Quisel JD, Burkholder WF, Grossman AD (2001) In vivo effects of sporulation kinases on mutant Spo0A proteins in Bacillus subtilis. J Bacteriol 183(22): 6573-6578.
Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A et al. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19(1): 45-50.
Ratnayake-Lecamwasam M, Serror P, Wong KW, Sonenshein AL (2001) Bacillus subtilis CodY represses early-stationary-phase genes by sensing GTP levels. Genes Dev 15(9): 1093-1103.
Reo NV (2002) NMR-based metabolomics. Drug Chem Toxicol 25(4): 375-382.

Ringland J (1991) Rapid reconnaissance of a model of a chemical oscillator by numerical continuation of a bifurcation feature of codimension 2. J Chem Phys 95(1): 555-562.
Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H et al. (2003) Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. Omics 7(4): 355-372.
Savageau MA, Kotre AM, Sakamoto N (1972) A possible role in the regulation of primary animation for a complex of glutamine: -ketoglutarate amidotransferase and glutamate dehydrogenase in Escherichia coli. Biochem Biophys Res Commun 48(1): 41-47.
Schilling CH, Schuster S, Palsson BO, Heinrich R (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol Prog 15(3): 296-303.
Schilling CH, Edwards JS, Letscher D, Palsson BO (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnol Bioeng 71(4): 286-306.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11): 2498-2504.
Shivers RP, Sonenshein AL (2004) Activation of the Bacillus subtilis global regulator CodY by direct interaction with branched-chain amino acids. Mol Microbiol 53(2): 599-611.
Smits WK, Kuipers OP, Veening JW (2006) Phenotypic variation in bacteria: the role of feedback regulation. Nat Rev Microbiol 4(4): 259-271.
Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M et al. (2002) Simultaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. Anal Chem 74(10): 2233-2239.
Soga T, Ueno Y, Naraoka H, Matsuda K, Tomita M et al. (2002) Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. Anal Chem 74(24): 6224-6229.
Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M et al. (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. J Proteome Res 2(5): 488-494.
Soga T, Baran R, Suematsu M, Ueno Y, Ikeda S et al. (2006) Differential metabolomics reveals ophthalmic acid as an oxidative stress biomarker indicating hepatic glutathione consumption. J Biol Chem.
Sonenshein AL (2002) The Krebs citric acid cycle. In: Sonenshein AL, Hoch JA, Losick R, editors. *Bacillus subtilis* and its closest relatives: from genes to cells. Washington D.C.: ASM Press.
Sonenshein AL (2005) CodY, a global regulator of stationary phase and virulence in Gram-positive bacteria. Curr Opin Microbiol 8(2): 203-207.
Stephenson SJ, Perego M (2002) Interaction surface of the Spo0A response regulator with the Spo0E phosphatase. Mol Microbiol 44(6): 1455-1467.
Stragier P, Losick R (1996) Molecular genetics of sporulation in Bacillus subtilis. Annu Rev Genet 30: 297-241.

Tanaka R, Yi TM, Doyle J (2005) Some protein interaction data do not exhibit power law statistics. FEBS Lett 579(23): 5140-5144.
Tanaka R, Csete M, Doyle J (2005) Highly optimised global organisation of metabolic networks. Syst Biol (Stevenage) 152(4): 179-184.
Taylor J, King RD, Altmann T, Fiehn O (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics 18 Suppl 2: S241-248.
Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y et al. (1999) E-CELL: software environment for whole-cell simulation. Bioinformatics 15(1): 72-84.
Tyson JJ (1991) Modeling the cell division cycle: cdc2 and cyclin interactions. Proc Natl Acad Sci U S A 88(16): 7328-7332.
Veening JW, Hamoen LW, Kuipers OP (2005) Phosphatases modulate the bistable sporulation gene expression pattern in Bacillus subtilis. Mol Microbiol 56(6): 1481-1494.
Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA et al. (1997) Characterization of the yeast transcriptome. Cell 88(2): 243-251.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. (2001) The sequence of the human genome. Science 291(5507): 1304-1351.
Voigt CA, Wolf DM, Arkin AP (2005) The Bacillus subtilis sin operon: an evolvable network motif. Genetics 169(3): 1187-1202.
von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406(6792): 188-192.
Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356): 737-738.
Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. Drug Discov Today 10(22): 1551-1558.
Xiong W, Ferrell JE, Jr. (2003) A positive-feedback-based bistable 'memory module' that governs a cell fate decision. Nature 426(6965): 460-465.
Yi TM, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc Natl Acad Sci U S A 97(9): 4649-4653.