Università degli Studi di Pisa
Dipartimento di Informatica
Dottorato di Ricerca in Informatica

PhD. Thesis

Modelling Biological Systems From Molecular Interactions to Population Dynamics

Thomas Anung Basuki

SUPERVISORS
Dr. Antonio Cerone
Prof. Andrea Maggiolo-Schettini

November 27, 2010

ACKNOWLEDGEMENT

Many people have helped me through discussions, suggestions and critics during the process of writing my thesis. I would like to express my gratitude to them.

First of all I want to thank my supervisors Antonio Cerone and Andrea Maggiolo-Schettini, who have continuously guided, criticised, and encouraged me during my research.

Several people have also helped me with discussions, suggestions and critics. Especially I would like to thank Paolo Milazzo, who has always agreed to spend some time to discuss with me, Roberto Barbutti, Elisabetta Rossi, Giovanni Pardini, Zhiming Liu, Rafael Viana Carvalho, Andreas Griesmayer, Rudolf Schlatte. I would also thank all my reviewers: Gabriel Ciobanu, Jean Louis Giavitto and Mario Perèz Jimènez, for commenting and suggesting improvements for my thesis.

Finally I want to thank my wife Steffani and all my children, Ivan, theo, Marcel and Gabi. Thank you for supporting and motivating me throughout my works.

Contents

List of Figures

List of Tables

Abstract

Biological systems are examples of complex systems, which consist of several interacting components. Understanding the behaviour of such systems requires a multidisciplinary approach that encompasses biology, mathematics, computer science, physiscs and chemistry. New research areas are emerging as the result of this multidisciplinarity, such as bioinformatics, systems biology and computational biology. Computer science plays an important role in the newly emerging research areas by offerring techniques, algorithms, languages and software to solve research problems efficiently. On the other hand, the efforts to solve these research problems stimulate the development of new and better computer science techniques, algorithms, languages and software.

This thesis describes our approach in modelling biological systems as a way to better understand their complex behaviours. Our approach is based on the Calculi of Looping Sequences, a class of formalisms originally developed to model biological systems involving cells and their membrane-based structures. We choose Stochastic CLS and Spatial CLS, two variants of the calculi that support quantitative analysis of the model, and define an approach that support simulation, statistical model-checking and visualisation as analysis techniques. Moreover, we found out that this class of formalisms can be easily extended to model population dynamics of animals, a kind of biological systems that does not involve membrane-based structures.

Chapter 1

Introduction

Cell biology, also called *cytology*, is an academic discipline that studies cells. In the past few decades, there has been an explosion of knowledge about the contents of living cells. Biologists realise now that, despite of its microscopic size, a cell is full of proteins. Each protein has specific function. Frequent interactions between proteins also occur in cells. A great variety in protein structure, function and interaction has made us realise that a cell is a very complex system. This urged biologists to shift from the traditional reductionist paradigm, which studies individual components of a living system separately, to the integrative paradigm, which studies the living system as a whole [80]. Recently, a new field of biology that studies complex interactions in biological systems has emerged: *systems biology*.

Even with the abundant amount of knowledge about cell contents, biologists still have a difficult job in understanding many cellular processes. In fact, a cellular process is usually a combination of many chained and concurrent subprocesses, which results in an emergent complex behaviour. Computer science can play an important role by supporting biologists in understanding how such a high complexity emerges out of interacting cellular processes. In particular, formal methods and concurrency theory provide powerful techniques that can be applied to the modelling of cellular processes. Bray [21] suggested the need for a novel language to model the universe of cells. Inspired by the success in using computer graphics tools to visualise protein structure, he suggested to use the same approach to understand cell universe. Church, Apagyi and Fisher [31] described some important aspects and challenges in developing the language for

biological models, which include:

- modelling and analysis capability of the language to handle complex and diverse problems in biology,
- ability to support the communication of the models to other biologists,
- ease of use for biologists with limited computer science background,
- ability to deal and reason with incomplete information.

The increased use of computer science techniques to solve biological problems introduces the new terminology *in silico*, as an analogy to the Latin phrases *in vivo* and *in vitro* which are commonly used in biology. In vivo and in vitro refer to experiments done in living organisms and outside of living organisms respectively, while in silico refers to the use of computer to perform biological studies [80].

1.1 Applications of Formal Methods to Systems Biology

Computer scientists find similarities between concurrency theory and systems biology. Both disciplines deal with systems consisting of smaller elements interacting with each other. Following this view, process algebras have been used to model cells and interactions occurring inside cells. A cell consists of many smaller elements that can be modelled as algebraic processes. The interactions between these elements can be modelled as interactions between processes.

Degano and Priami [37] claimed that both systems biology and formal methods for concurrency can cross-fertilize each other. Being based on sound and deep mathematics, concurrency theories may offer solid ways to describe biological systems and safely reason upon them. On the other hand, systems biology studies many complex biological phenomena. Modelling and reasoning about these complex phenomena may require techniques that are more efficient and reliable than existing techniques. It is expected that the effort to understand biological mechanisms in terms of computer technology will possibly lead to new techniques that are more robust, efficient and reliable to model and analyse complex systems.

Regev, Silverman and Shapiro [92] observe that a formalism to model cells should fulfill four goals:

- provide a unifying view of both the molecular data and the dynamic behaviour it underlies,
- formally represent data to be used for computer execution and analysis,
- facilitate comparative studies of structures, dynamics and functions within and between species,
- be scalable and modularised to higher levels of organisations.

In the attempt to achieve these goals, several formalisms have been proposed for modelling biological phenomena, such as Petri Nets [89, 90, 53], Brane Calculi [23, 36], P Systems [84, 83], the π-Calculus [30, 88, 100, 101], CCS-R [35], which is a variant of Milner's CCS (Calculus of Communicating Systems), and Calculi of Looping Sequences [76]. Petri Nets, CCS and the π-Calculus are general formalisms used to model concurrent processes, thus they were not originally designed to model cellular systems. Brane Calculi and Calculi of Looping Sequences are new classes of formalisms developed specifically for modelling cells. P Systems, the class of formalisms inspired by membrane systems in Cell Biology, were not initially intended to model cellular systems. In P Systems, the computational nature of various features of membranes is explored and investigated in order to be used in a model of computation. However P Systems have been later also used to model cells as computing processes.

Reddy, Liebman and Mavrovouniotis were the first to apply formal methods in modelling biological systems. In their work [90, 89], they proposed a method for qualitative analysis of biochemical pathways using Petri Nets. They used several Petri Net properties, such as boundedness, liveness, and invariants to identify properties in the biological system. For example, the accumulation of some toxic intermediates in a biological system can be identified using the boundedness of a Petri Net. Barjis and Barjis also used Petri Nets to model protein production process in details and showed how to convert the model into an executable program [11]. A limitation of these approaches is that the use of ordinary Petri Nets for modelling biological systems is limited to the analysis of qualitative properties of the systems.

In order to overcome this limitation, extended versions of Petri Nets were also applied in Systems Biology. The first extension was Functional Petri Nets. In Functional Petri Nets it is possible to assign to Petri Nets arcs equations using marking variables instead of natural numbers as in standard Petri Nets [53]. Hofestädt and Thelen [59] used Functional Petri Nets to do quantitative modelling of biochemical networks. They extended the model developed by Reddy, Liebman and Mavrovouniotis by allowing dynamic representation of concentration of metabolites participating in a chemical reaction.

Hardy and Robillard surveyed the use of three other extensions of Petri Nets in modelling biological systems: Coloured Petri Nets (CPN), Hybrid Petri Nets (HPN) and Stochastic Petri Nets (SPN) [53]. Goss and Peccoud [50, 51] modelled quantitative aspects of molecular interactions. Using a Stochastic Petri Nets tool called UltraSAN, they were able to model and perform quantitative analysis of two case studies: protein synthesis and plasmid ColE1 replication.

Matsuno, Doi, Nagasaki and Miyano [73, 77, 78] proposed a method for modelling biological systems using HPN. HPN extend Petri Nets by allowing places and transitions to have real numbers as values of tokens. HPN enable modelling complex and more realistic biological systems. Real numbers in places and transitions can be used to represent various features. For instance, they can be used to represent concentrations and speed of reactions [73]. Just like SPN, HPN support probabilistic analysis of the system. The choice between SPN and HPN depends on the nature of the system to be modelled. If the model deals with a big number of molecules, HPN is usually preferred since it enables using real numbers to represent big numbers. If the model deals with a small number of molecules, SPN is usually preferred.

CPN also support quantitative modelling and analysis of biological systems. Genrich, Küffner and Voss [45] use colours of tokens to represent time-related information needed for simulation of the model. Voss, Heiner and Koch use colours to distinguish the origin of molecules in a place [102]. Marking the origin of molecules is useful in the completeness and feasibility analysis of the model.

Overall, Petri Nets and their extensions support modelling chemical reactions in biological systems as well as qualitative and quantitative analysis. Petri Nets are equipped with graphical representations, which make them favourable to be used for modelling and analysis of biological systems. However no classes of Petri Nets support modelling the structural details of biological systems.

Regev, Silverman and Shapiro were the first to use the π-calculus to model biological systems [92, 93]. The π-calculus enables modelling interaction between processes using complementary channels, and communicating names of channels via channels. This feature allows network structure to change with interaction (mobility). Regev, Silverman and Shapiro developed a piece of software called PiFCP to simulate their model. They also suggested the use of bisimulation as a way to check equivalence between two processes. In the context of cell biology, this facilitates comparative studies of several biological systems. Comparison of similar pathways is the first step to study cell evolution.

Priami, Regev, Silverman and Shapiro used stochastic π-calculus to model biological systems [87]. Their approach supports qualitative modelling of biological systems. They also developed a piece of software called BioSPI to simulate the model.

Formalisms that originated from concurrency theory are capable of modelling activities inside cells (chemical reactions), but are not capable to model compartments. A molecule or an object in a cell can do its function only when it is on the right location. Compartments play an essential role by organising biological systems hierarchically and also introduce the notion of object's location. This was first realised by Regev, Panina, Silverman, Cardelli and Shapiro [91]. They developed the π-calculus into BioAmbients in order to model compartments. BioAmbients models both membrane-bound compartments, in which boundaries are clearly defined, and molecular compartments, in which the boundaries are not clearly defined.

Cardelli was the first to define specific formalisms for describing cellular systems. He named such a class of formalisms Brane Calculi [23]. In Brane Calculi, cellular systems are modelled as membrane systems that can perform computations. Computations are performed on membranes rather than inside them. A membrane system may consist of several membranes. Computations on membranes are defined using actions.

The simplest version of Brane Calculi is the PEP Calculus. In the PEP Calculus, there are only three kinds of actions: Phago, Exo and Pino. Phago (*phagocytosis* or literally "cellular process of eating") and Pino (*pinocytosis* or "cell drinking") are *endocytosis*, cellular processes of engulfing solid particles by the cell membrane. In Phago the process engulfs one external membrane while in Pino no external membrane is engulfed. Exo (*exocytosis*) is the reverse process of endo-

cytosis, where the membrane ejects objects to the extracellular environment.

Cardelli also extended the PEP Calculus by adding more actions to be performed on membranes, such as Mate, Bud ad Drip [23]. He also defined molecules which may interact with the membranes, and defined complexation of molecules to model protein complexes. Other extensions in his work are communications between membranes, choice operations and atonal transport. Danos and Pradalier also extended PEP Calculus by replacing actions with directed actions [36]. They defined two kinds of directed actions: inward and outward actions.

P Systems were first defined by Păun [82] as a model of computation inspired by membrane systems. In P Systems, computation occurs inside the membrane. The membrane only functions as a boundary, so that the system can be modelled hierarchically. The computation performed by the system is modelled using a set of rewrite rules.

More recently P Systems have been also used for modelling biological systems. For example, Bernardini [15] and Bianco [17] in their theses showed how to use variants of P Systems to model biological systems. Bernardini's work focuses on two main issues in P Systems, exploiting a variant of P Systems called Population P Systems, and investigating bio-inspired communication mechanisms in P Systems. He showed his approach by modelling quorum-sensing in bacteria, a communication strategy among many bacteria to coordinate gene expression according to the local density of bacteria producing signalling molecules. Bianco's work focuses on dynamics of signal transduction networks. Perez-Jimenez and Romero-Campero modelled the epidermal growth factor receptor signalling cascade [85]. P Systems models are always analysed through simulation, followed by a comparison of the in silico result with the in vitro result.

There are also works on comparing Brane Calculi and P Systems. Busi and Zandron compared Brane Calculi and P Systems [22] by modelling the LDL cholesterol degradation pathway in both Brane Calculi and P Systems. Cardelli and Păun [24] showed the expressiveness of P Systems by emulating Brane Calculi actions Pino, Exo, Bud and Mate using P Systems. Krishna extended this work by adding actions Phago and Drip to the emulation of Brane Calculi using P Systems [65].

In our work, we will mainly deal with Calculi of Looping Sequences. In his thesis [76], Milazzo has defined the Calculus of Looping Sequences (CLS) and four additional variants of it. He has used these formalisms to model several

biological phenomena. Milazzo has also contributed to develop a Stochastic CLS simulator [96]. He has used the simulator to simulate gene regulation in *E. coli* and compared the result of the simulation with the real experiment.

Biologists already know that biological systems can be modelled as stochastic systems. Parameter values of the model are taken from the results of in vitro/in vivo experiments. Therefore simulation can be performed to analyse the model quantitatively. Another kind of analysis that can be done is model-checking. Model-checking is useful for both qualitative and quantitative analysis. There have been several attempts to apply model-checking to biological systems. Chabrier-Rivier, Chiaverini, Danos, Fages and Schachter developed a formalism to model biological phenomena [27] and proposed the use of Computation Tree Logic (CTL) to query the model. Later Fages, Soliman and Chabrier-Rivier [40] developed this modelling formalism into a modelling environment called Biochemical Abstract Machine (BIOCHAM). They showed two examples of modelling biological systems, the mammalian cell cycle control and the regulation of gene expression [26, 27]. They also showed some possible properties about the model and how to model-check such properties. Bernot, Comet, Richard and Guespin formally modelled biological regulatory networks using graphs and analysed CTL properties of the model using the SMV model-checker [16].

David Harel et al. [63, 95] used statecharts and LSC (Live Sequence Charts) to model and verify biological models. Harel proposed the task of fully modelling a muti-cellular animal as a grand challenge in computing [54]. He proposed to model *Caenorhabditis elegans* nematode worm (*C. elegans*) as grand challenge case study. *C. elegans* is very well-defined in terms of anatomy and genetics. In his long-term proposal, Harel suggested to employ formal verification technique to compute ways to satisfy a desired scenario. Harel et al. started with modelling *C. elegans* vulval development, which occurs during egg development [63, 64]. By combining the state-based approach of statecharts and scenario-based approach in LSC, they performed in silico experiments with their model and compared the result with data observed from in vivo experiments [95, 41, 42].

Bodei, Bracciali and Chiarugi [18] offered a simple formalism to model-check causality in biological systems. Their formalism was implemented in Prolog, using Horn-clauses to represent chemical reactions in biological systems. They modelled the metabolic network of *E. coli* K-12 genes. They simulated gene knock-out and compared the result with the in vitro experiment. Gene knock-out is important to find out which genes are essential to produce a specific metabolite. Some

genes are even essential for the life and death of the cell. Since Bodei, Bracciali and Chiarugi only focus on causality, many details are abstracted away in their approach.

To perform quantitative model-checking on a system Kwiatkowska, Norman and Parker developed a probabilistic model checker called PRISM [67]. In PRISM, models can be defined by using either discrete-time Markov chains (DTMCs), Markov decision processes (MDP) or continuous-time Markov chains (DTMCs). PRISM has been successfully used to model and analyse several biological case studies: FGF (Fibroblast Growth Factor) signalling pathway [57], 3-way biochemical oscillator [6], MAPK cascade [66] and mRNA translation [19]. Although all these case studies show the success of modelling and analysing quantitative aspects of biological systems, there is one common limitation of the approach. In order to avoid the state explosion problem, the state space is reduced by limiting the number of molecular species or the number of molecules for each molecular species involved in the system.

1.2 Modelling Biological Systems at Different Levels of Representation

Most approaches described in Section 1.1 use texts and plots to show results. Although texts and plots provide detailed information on specific aspects of the analysed biological system, they are often inadequate when the aim is to acquire global knowledge about the high-level organisation and dynamics of the biological system. For example, in most experiments the analyst can only vary molecular concentrations in the environment and within cells, whereas the aim of the experiment or simulation may be to observe the resultant behaviour of cells or even the whole organ or organism. Such high-level behaviours can be better described through two or three dimensional visualisation/animation rather than using texts and plots.

The work on visualisation of biological systems started in 1968, when Aristid Lindenmayer defined his famous formal model of plant development which is called L-systems (or Lindenmayer systems) [70, 71]. L-systems have successfully been used to model the growth of parts of a plant, without dealing with the molecular interactions triggering the growth. Only interactions between the plants cells and

external factors in the environment are taken into account [52, 86, 44].

Slepchenko, Schaff, Macara and Loew developed a tool to visualise cellular systems according to a model of the system at the molecular level [98]. This tool, which is called Virtual Cell, is based on a deterministic numerical simulation of the model, which is defined by using differential equations.

David Harel and his group developed an approach in modelling at different levels of representation [55]. They used object oriented approach and defined the cell as the basic building block of their approach. Their approach uses scenario to define system behaviour and uses animation on a 2-dimensional grid [2]. Scenarios define cell behaviour related with interactions between molecules in the environment and their receptors on cell membranes. Another interesting application of their approach is the modelling of pancreatic organogenesis [97]. In this application they show how molecular interactions affect cell growth and, in the end, affect the growth of mammalian pancreas. A three dimensional visualisation is used to visualise the pancreatic organogenesis process.

In both Virtual Cell and the works of Harel and his group visualisation of the higher level behaviour of a biological system is triggered by the behaviour of the system at the lower level. This kind of research brings in silico biology closer to in vivo and in vitro biology. However, those two approaches are deterministic, whereas real biological systems are stochastic.

Michel, Spicher and Giavitto use rule-based programming language MGS to model and simulate the λ phage genetic switch [75]. They present a multilevel model of the system; a molecular level defined by using built-in Gillespie's algorithm (which is a stochastic algorithm) and a population of cells level defined by using GBF (Group Based Field) and Delaunay topological collections. The result of a simulation can be printed into a file, which later can be visualised by using another tool.

1.3 Modelling Population Dynamics of Biological Systems

Research in population biology aims to study factors that affect the dynamics of a population of individuals and how to regulate the population size. Mathematical

models are often used as tools to predict population dynamics. Population dynamics modelling and analysis are usually important for the following purposes:

- controlling the spread of a disease,
- conservation of endangered species,
- predicting the economical impacts of the population dynamics.

Diseases like Malaria and Dengue are rapidly spread by mosquitoes. Female mosquitoes need blood to oviposit and humans are the main source of blood for these mosquitoes. A few models of mosquito population have been proposed [1, 68]. Some works have also connected the population models to disease spread [60, 43].

In the case of insect population models scientists have to deal with more uncertainties than in the models of bigger animals. For instance, mosquitoes usually spend their immature stages in water and field data about their immature stages are not available. The result of simulating these models can only be compared with real data for adult mosquitoes.

Armstrong and his group monitored the population of some birds in New Zealand and proposed a population model based on the data [5, 4, 3]. By comparing the simulation result of their model with the real data they calibrate the parameters of the model. They considered factors such as population density, age, and gender. Jenouvrier, Barbraud, Cazelles and Weimerskirch proposed a model of seabird population by considering the stages in the bird's development and climate factors in the model [61].

Economy also affects population of some animals. Many animal products are traded legally or illegally. Some research have started to model animal population by considering economical factors. Examples of this kind of research are modelling the effect of legalising markets to the population of some animals [58, 94].

All approaches described above are deterministic.There are also some approaches that model nondeterminism in the systems using formal methods. Barbuti et al. [10] extend P Systems with features typical of timed automata with the purpose of describing periodic environmental events such as changes of seasons. Cardona et al proposed a formal modelling approach based on P systems and applied it to model the population dynamics of bearded vulture in the Pyrenees [25].

McCaig, Norman and Shankland [74] present a process algebraic approach to the modelling of population dynamics.

1.4 Contributions of the Thesis

This thesis focuses on the use of Calculi of Looping Sequences to quantitatively model and analyse biological systems. Simulation is still used in our approach, but we also propose the use of model checking [14]. In Chapter 3, to deal with the large size of the state space we propose the use of statistical model checking. This requires a compact representation of system states and the definition of an operational semantics for Stochastic CLS based on such compact representation.

The second contribution of this thesis is the definition of an approach based on Spatial CLS to model a biological system at different levels of representation [13]. This approach, presented in Chapter 4, is stochastic and supports visualisation of the system. In this way we provide a modelling approach that makes the presentation of experimental results close to what we observe with in vitro and in vivo experiments.

In Chapter 5, we propose an approach to model population dynamics of animals. We define a general approach to deal with environmental factors, which are not controlled by the simulation algorithm but affect the rates of events in the model. To handle this new context we extend Stochastic CLS with a list of external events modelling environmental factors. In this way we can use real environmental data such as temperature and rainfall to calibrate the parameters in our model and then compare the simulation results with real data [12].