UNIVERSITE D'EVRY
VAL D'ESSONNE

LaMI
Laboratoire de Méthodes Informatiques

Modeling, Observability and experiment : a case study

- Positive feedback loop in a genetic regulatory network -

G. Bernot, J. Guespin-Michel, A. Zemerline, J.-P. Comet,
P. Amar, F. Delaplace, P. Ballet

email(s) : `{bernot,comet,delapla,pa}@lami.univ-evry.fr,`
`janine.guespin@univ-rouen.fr,`
`{Abdallah.Zemirline,pascal.ballet}@univ-brest.fr`

Rapport de Recherche n^{o} **82-2002**

Octobre 2002

CNRS – Université d'Evry Val d'Essonne
523, Place des Terrasses
F–91000 Evry France

Modelling, observability and experiment : a case study.

- Positive feedback loop in a genetic regulatory network -

G. Bernot[1], J. Guespin-Michel[2], A. Zemirline[3], J.P. Comet[1], P. Amar[1,4], F. Delaplace[1], P. Ballet[3]

Abstract

We propose an interdisciplinary methodology for biological modelling inspired by the design and validation of large computing systems.

To know if a model for biology can be satisfactorily validated by a set of experiments seems to be a natural and necessary constraint for its definition. Defining a model should go with experimental methods and conditions able to validate or invalidate it. As in the design of large sized softwares, we will distinguish two activities : first to build an accurate model specifying the observed behaviour, second to design plans of experiments to verify *a posteriori* the model predictions.

We wish to experiment, through the case of the modelling of the mucus production by the bacterium *Pseudomonas aeruginosa*, the application of this working methodology.

1 Introduction

Biologists put a large number of meanings in the term "model". Even when precised as "mathematical model" we are far from the unicity of meaning. One of the difficulties comes from the interdisciplinarity already contained in the expression "a mathematical model in biology". Who makes the model, who is using it, and first, what is it utility ? Most of the mathematical models in biology are made by mathematicians (biomathematicians), physicists, or computer scientists (bioinformaticians). They use data imported from the literature (which have been obtained for another usage). The resulting model, published in journals which are not read by biologists, is in general made to point or explain some known biological phenomenons. At worst, they replace a phenomenological description by a mathematical expression ; but they can also give (or else suggest) a new explicative framework for the biological process studied. If some biologists hear something about that kind of model, they may be interested in, but more often, they do not see in which way they can be commited. From their point of view, the model appears to be often useless, and may involve a large range of reactions, from violent reject to polite interest.

Conversely, and also revealing the lack of interdisciplinary knowledge, some biologists think that by giving to a model maker data they have obtained in some particular context, this one will put it into his magic box and get the explication, or even better, a predictive tool for therapeutic issues. . .Interdisciplinarity needs not only learning how to work together, but also the common design of usable tools. It demands moreover that each contributor finds a scientific interest working together, in other words, that the collaboration will profit to both disciplines.

This is where analogy between computing systems specification and modelling in biology is involved (cf. annex). In computer science, the design of systems requires to :

- specify, i.e. build a rigorous model of the desired behaviour of the future computing system ;
- verify *in fine* if a system corresponds to its specification, i.e. to the desired behaviour as described by the theoretical model previously built.

This last activity is mainly based on sophisticated software test methods, based on test generation from model theories. The goal is then to propose a set of experimentations on the delivered software which is sufficient to establish, by extrapolations, that the tested software will have a behaviour compatible with its model.

[1] Laboratoire des Méthodes Informatiques, CNRS UMR 8042, Univ. Evry

[2] Laboratoire de Microbiologie du Froid et GR en Biologie Intégrative et Modélisation, Univ. de Rouen

[3] EA 2215, Dept Informatique, Univ. de Bretagne Occidentale

[4] Laboratoire de Recherche en Informatique, CNRS UMR 8623, Univ. Paris-Sud, Orsay

Within this framework, the notions of operability and observability constitute a major issue :

- the *operability* is the capability to make a program run some choosen pieces of its internal code (in order to test them), sometime activated in rare, complex or very specific configurations. It is also the capability to make a program modify the value of variables hidden in the very large set of data managed by the program. These actions have to be done by only using the often limited user interface of the tested program.
- the *observability* is the capability to make the effects produced by the previous manipulations visible, in order to verify their correctness according to the desired model of behaviour.

Some models or softwares are not testable, either because of a lack of operability or a lack of observability. A necessary step to design a software is to know if a model can be validated by a reasonnable sized set of tests (experiments). The reader can easily transcribe this argumentation to the case of biological modelling :

- *operability :* what would be the utility of a too much detailed model of some biological entity if no biological experiment of those details can be done ?
- *observability :* what would be the utility of an experiment which cannot let us observe a revealing behaviour ?

Some mathematical models for biology are not very helpful because of a lack of operability or observability. A necessary first step to propose a model for biology is to know if it can be validated by a set of biological experiments at a reasonnable cost.

Hence a good scientific approach requires that a bio-informatics model must be systematically delivered with a set of experimental methods/conditions able to validated or invalidate it. By using the same kind of theories developed in computer science for the validation and the verification of softwares, we wish to experiment, through the case studied here, a new interdisciplinary working method for the modelling in biology.

The case we will study here involves a tight coupling between two mathematical theories and one biological process. A first modelling step (already published), based on the multistationnarity theory, makes an innovating hypothesis plausible (section 2). A second mathematical step, based on the formal logic in computer science, allowed us to determine the biological experiments sufficient to validate or invalidate the hypothesis (section 3). The long term goal is, by extrapolating the results to some other systems, *to create a new tool imported from computer science* allowing in the one hand to better commit biologists in the modelling process by giving them a model validation tool, and in the other hand, to increase the scope of an already usable tool in computer science. Of course this approach needs a tight collaboration between biologists and model makers. The model designed this way, is not only an *a posteriori* explaination attempt of results from biology, but a guide for biological experimentation, which will be *in fine* the determining criterion.

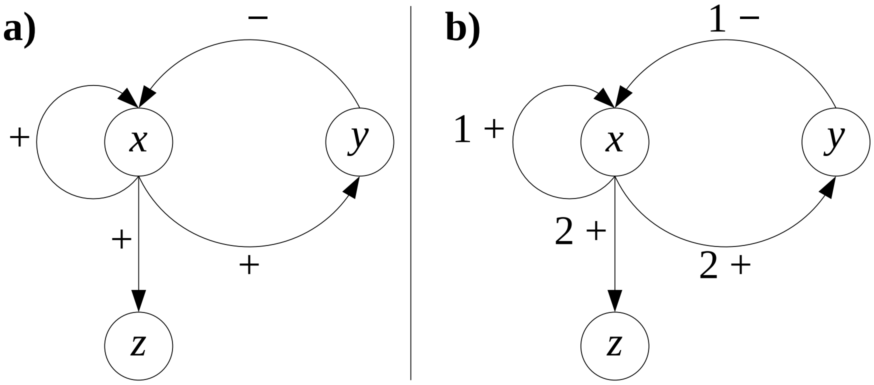
2 The chosen case

The biological system chosen is the production of mucus (alginate) by the bacterium *Pseudomonas aeruginosa*. Bacteria of this specie do not generally produce this mucus if they have not experienced a sejourn inside the lungs of patients suffering from cystis fibrosis (production which is the main cause of lethality in this disease). Not only do these bacteria produce alginate in the patients' lungs, but they continue doing so, more or less stably, once extracted from these lungs and cultivated in the laboratory. It is generally admitted that mutations arising inside these lungs cause the ability of these bacteria to produce this mucus in other conditions (cf. [1]).

But an other hypothesis has been put forward, according to which the ability to produce or not alginate are two stable states that arise from each other by an epigenetic modification, prior to the selection of mutants (cf. [2]).

A very simplified model of the regulatory network has thus been constructed (cf. [2]) as depicted in figure 1. The 3 variables are x for the AlgU protein, y for the AlgU inhibitors, and z for the alginate production. The 4 arcs[5] represent : the self-regulation of variable x (arc $x \to x$), the transcription of the genes encoding the antisigma factors (arc $y \to x$), the transcription of the genes involved in alginate production (arc $x \to z$)), and finally the inhibition of AlgU by the antisigma factors (arc $x \to y$). Two feedback circuits control AlgU, a positive feedback loop at the transcriptional level, and a negative feedback circuit involving the activity of the AlgU sigma factor. The extreme simplification of this model is directly related to the theory that supports it, which stipulates that feedback circuits are the only elements that are determinants for the emergence of epigenesist (cf. [3]). It is therefore stipulated that the other known regulatory interactions are of minor importance with regard to the question of the existence of an epigenetic modification.

[5] the arrows in figure 1



> Note : the left side of the picture (a) shows the activation ($+$) or the inhibition ($-$) between the genes (resp. proteins) of the network. The right side (b), each arc is labelled with the minimal value of the threshold for which this action (activation or inhibition) is triggered.
> The variables means : x AlgU protein, y antisigma factor, z mucus synthesis.

FIG. 1: Genetic network regulating the production of mucus (alginate) by the bacterium *Pseudomonas aeruginosa*(simplified model)

This model can be studied by a system of differential equations or by generalised logical analysis (cf. [4]), which was chosen because of the lack of known values for the interaction parameters. To summarise, when variable x interacts with variable y, the curve that represents y as a function of the level of x is a sigmoid. This sigmoid defines a threshold $S_{(x,y)}$ (cf. figure 2-a). Similarly the influence of x on an other variable z defines an other threshold $S_{(x,z)}$ (cf. figure 2-b). The two thresholds are generally different and lead to three different possible behaviours of variable x depending whether it is below both thresholds, between them or above them (cf. figure 2-c). Thus it is possible to ascribe discrete values to the different levels of variable x. Then the thresholds correspond to interaction values between the variables. In order to describe that AlgU must be present above threshold 2 to trigger the expression of the alginate genes, it will be noted in the graph $x \xrightarrow{2+} z$ (cf. figure 1-b). When an arc can be skipped, i.e. when the conditions on the level of the variable are set, the evolution of the system must be described. In other words, we have specify the level reached by a variable v as a function of the levels of the other variables that influence it (cf. figure 1). These values are represented by function K.

In our model, two feedback circuits co-exist. The first one (positive feedback loop $x \to x$) is a necessary condition for the existence of two stable states : if x is high, it is self-maintained, if it is below the first threshold, it remains so. The negative feedback circuit can switch the system toward one or the other of these stable states depending on its strength (i.e depending on function K).

A mathematical study of this model (cf. [2]) has shown that the epigenetic hypothesis (the possibility that two stable states may exist depending on the previous history of the system) is coherent and that biologically consistent values of Ks can lead to properties that are precisely those of the system. But there is more to it. For instance it can be predicted that, if the hypothesis is true, a pulse of AlgU will suffice to switch the bacteria to a mucoid state. The model is thus predictive as well as explanatory.

The question that we will address here is : if such an experiment succeeds, if a pulse of AlgU is able to induce mucus production, or at least the expression of the first genes involved in mucus production, will this be sufficient to prove the underlying hypothesis of epigenetic modification ? Inversely, if this experiment fails, will it prove the unreliability of the epigenetic hypothesis in this case ?

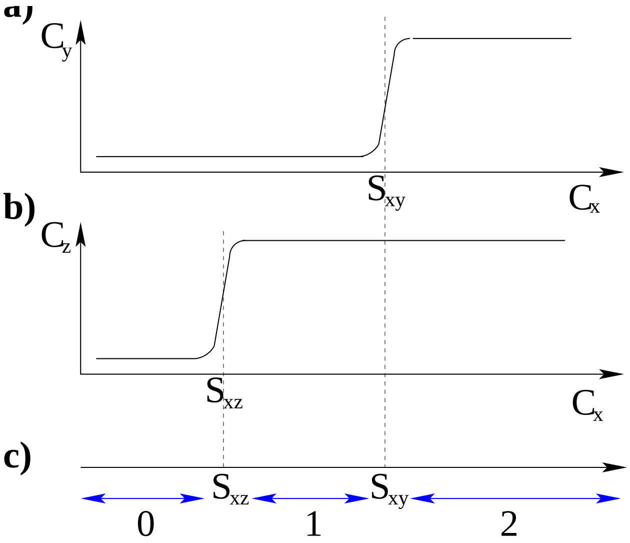


FIG. 2: Influence of the variable x on the variables y, z

3 Formal logic to propose experiments

In this part, we outline the general methodology through the example of the production of mucus in the bacterium *Pseudomonas aeruginosa*described in section 2. Although the mathematical proofs are informally presented, they can all be formally performed on a computer. Indeed a model is used to establish properties on a system, to express and handle these properties in order to extract some non trivial other ones. It is thus necessary to formalise the properties in such a manner that they are easy to handle by a computer. The objective laid down here relates to the generation of scenarii of experiments, in which time plays a central part. We are necessarily confronted with the concept of time when we want to express properties of the system in the future. These constraints lead naturally to temporal logics (cf. [5], [6] for a general description of temporal logics).

More precisely, we want to prove that, *in the presence of* y, it is possible to have a recurrent state in which mucus is produced. By construction of the graph, y is present and the topology of the graph was biologically validated as well as the signs of interaction. Only the thresholds and the values of the function K can involve several different behaviours. Thus we only have to show that the values of the thresholds and function K for the studied organism are such that the bacterium can pass in a state where z is expressed in a recurrent way.

The language of temporal logics offers the traditional connectors such as for example, the “or”, noted $\vee$, the “and”, noted $\wedge$, the implication noted $\longrightarrow$. It also offers connectors particular to this type of logic which relate to time. We can for example create the connector $\mathcal{F}_s$, which means that the formula which follows the connector is true in the “strict future”. We call here “strict future” the future starting after a certain amount of time. This amount of time must be choosen according to biological considerations on the studied system.

Once the model has been mathematically defined, it is necessary to establish the formulae to be proven with this formalism. We want to prove that, in the model which we consider, if at a given time the bacterium in a mucous state, then later (in a strict future) it will be again in a mucous state. From previous experiments we know that the threshold associated with the interaction $x \rightarrow z$ is equal to 2, and we know by construction that the one associated to $y \rightarrow x$ is 1 (y has influence only on x, therefore there is only one threshold for y). On the other hand we do not know the thresholds of the arcs $x \rightarrow x$ and $x \rightarrow y$. In other words, we do not know the relative quantities of the variable x necessary to obtain a self-induction effect, an effect on y, or a combined effect. So, we want to prove with experiments that the relative forces of these two circuits, used by the bacterium are such that it is possible to make recurrent the state $(x = 2)$, which is written as :

$$(x = 2) \Longrightarrow \mathcal{F}_s(x = 2)$$

Any scenario which tests this formula must starts by assigning (artificially if necessary) 2 to x. In the opposite case, the first part of the formula is false, which involves that all the formula is true whatever the value of the second part

$\Rightarrow$	0	1
0	1	1
1	0	1

TAB. 1: Truth table of $p \Rightarrow q$

of the formula is. This is easy to deduce from table 1 by a simple algorithm. The scenario of experiments is thus the following :

1. Start by imposing $(x = 2)$.
2. Wait a lapse of time, then test the mucous state.
   - If the bacterium is not in a mucous state, then the experiment *a priori* fails.
   - If the bacterium is in a mucous state (thus $x = 2$) stop the experiment, because the preceding formula is reentrant, which means that if is true at a given time, then it implies that it will be true again in the future.

More precisely, at time t_0, we impose $x = 2$, by the formula, the experiment tells us that there exists $t_1 = t_0 + d$ (where d is an amount of time), such that $x = 2$. By making a shift of the origin of time, we can then affirm that there exists $t_2 = t_1 + d$ such that $x = 2 \ldots$

By iterating this formula, we show that the mucous state reached by the bacterium is recurrent. The scenario, sufficient to prove the theorem, is thus the following :

1. boost x by an external intervention until it reaches its maximal value (2),
2. wait long enough (to be sure that we are not any more in the initial condition, i.e. to be sure that the transient phase due to the initial boost is past) and test if we meet a mucous state again, i.e. $x = 2$.

If step 2 is successful then the experiment prove that the mucous state can be a steady state in the presence of y. Thus it can be an epigenetic phenomenon.

Operability and observability - The next question is then, is this prediction amenable to experimentation, is it both operable and observable ? In other words, is it possible to raise x up to 2, then quit the conditions that have allowed this, and observe the production of mucus in the "strict future" ?

Indeed, there are several ways to increase x without introducing the bacteria inside the lungs of a cystic fibrotic patient. For instance, one can introduce into wild type cells of *Pseudomonas aeruginosa*, a plasmid where gene *algU* would be under the control of an artificially inducible promoter. A short pulse of expression of this gene would lead to an artificial increase of the amount of protein AlgU inside the cell.

To observe the results of this experiment, again several experimental devices are currently available, either by measuring the mucus produced, or, more easily by measuring expression of the first gene of the alginate biosynthesis chain (gene *algD*).

Limits of the approach - The graph on which the model is based (figure 1) is actually only a subgraph of a more general graph showing all the variables of the organism. So it would be necessary to consider all the interactions with the neglected part of the general graph. Having neglected the outgoing arcs of the graph does not have any consequences since we are only interested in the subsystem involving the production of mucus. On the other hand, having neglected the arcs entering this subsystem can have an important impact. By construction of the graph, some situations can be eliminated.

1. By definition of z, the only arc controlling z is the one we take into account : $x \rightarrow z$, and thus it does not exist any other entering arc on z.
2. All the arcs which were not considered in the model but which control x or y are not involved in a circuit. The number of steady states does not change (cf. [7]).
3. If there are arcs entering on x and y whose influence does not vary, the only consequence of having extracted a subgraph is to shift the various thresholds associated with the variables x and y. The system will have other values for the thresholds and possibly for the function K, but the variables will always be discretised the same way. Thus, the satisfiability of the formula remains the same.

Only one case remains awkward : when regulators external to this subgraph (on x and y) have an influence which varies in time. The study presented here makes the assumption that these influences are negligible. This work remains valid under the assumption that a merge of the subgraph into the global graph have constant influence on the variables x and y.

Lastly, let us recall that the amount of time mentioned above between step 1 and step 2 of the experiment remains empirical.

4 Conclusion

The interdisciplinary work undertaken by our working group in genopole® gives a methodological framework to define models including a tool kit for experimental validation/refutation. This way, our work resolutely reinforces the modelling activity. It increases its credibility with respect to the mistrust which it causes in biology. Indeed, following a Popperian approach, this methodology offers the opportunity to strongly and properly link the modelling activity and the experimental activity, which is central in biology.

To establish such an approach requires a theory which fixes the rules allowing to reason from a model. The theory introduced here is *temporal logic*, usually employed for the logical analysis of the discrete dynamic systems in computer science. According to this theory, our case study proves that a discrete qualitative model of gene expression based on the work of René Thomas fulfills the methodological requirement mentioned. It makes it possible to determine, in a computer aided manner, a protocol of experimentation to prove or refute the epigenetic assumption described by this model (section 3).

Because our approach is inspired by the software engineering testing methods, this suggests that we can automate it, in other words, that we can provide software assistants for the design of biological models. We have shown the feasibility of the approach on the *Pseudomonas aeruginosa*example. Realizing these software assistants in a more general setting requires to continue our investigations on the application of formal methods from computer science to life science.

Annex : modelling and observability

The activity of modelling in life sciences, as in other sciences, has to extract from a necessarily *finished* set of biological observations, a mathematical representation expressing a generally *infinite* set of behaviours. The *a priori* infinite set of behaviors captured by a model rely for instance on the infinity of the possible values of the parameters (reflecting in particular the possible conditions of experiment), the infinity of the possible scenarii of simulation, etc. To be fruitful, the activity of modelling should answer several difficult questions such as : Is such or such model of good quality ? according to which precise criterion is it better than another ? Given two models which do not contradict any effective observation but which have different internal behaviours, which one provides the better explanatory view ? Modelling a phenomenon, in order to understand it, necessarily requires several abstractions. Therefore it requires the approximation of details and the reduction of the number of interacting objects. Consequently, it is not reasonable to imagine that a model can exactly represent a reality.

Biological complexity is far too rich, internal mechanisms are too badly known, and internal and external interactions, at all the scales, are nonforeseeable. In addition, the majority of the elementary laws chosen for the model, and their parameters, are not directly established *in vivo*. They can be extrapolated from an organism to another, obtained from *in vitro* constants, and sometimes even roughly fixed. Thus, every model of a biological phenomenon is false by construction.

So, what is then the interest to model a complex phenomenon when the capacity of prediction is so questionable, according to the simplest elementary scientific doubt ? Even badly, if a model answers the behaviours for which it has been built then it does not prove anything. The model can fail *a priori* in relation to any future biological experimentation which would reveal a new behaviour. In practice, such cases have often the advantage to destroy a dogma, but this leads to this question : How to prevent a new model to implicitly reinforce the tendency to replace a dogma by another ?

These interrogations show that the activity of modelling requires to evaluate the limits of any model. Nevertheless, they should not dissuade the biologist to make use of this tool which proved reliable, and acquired a major notoriety and usefulness in engineerings. The construction of a model becomes well founded if it follows the approach of Karl

Popper, according to which science progresses by conjectures and refutations [8]. A major criterion for an assumption to be in the scope of science is its falsifiability. If one detects an internal behaviour of the model which is not consistant, or an external behaviour which contradicts a biological experiment, then one can review the model and improve it.

Some behaviours of a model are direct logical consequences of known phenomena biologically validated. Some others are extrapolations introduced by the model maker with the aim of building a general model which is not only an enumeration of known particular cases, or with the aim of testing a scientific assumption nonaccessible in an intuitive way. In all cases, modelling implies to make assumptions and conjectures for the behaviour of the model. These assumptions are the weakness of the models which one should systematically try to validate by attacking them by biological experiments and well choosen simulations. But they can also be the source of major projections, making it possible to better understand the process and predict original properties. Thus, even if the results producing scientific advances are negative results, highlighting errors of a model, it remains that the more a model resists to attacks which try to refute it, the more it is interesting and valuable.

A good modelling must thus not only clarify its assumptions but also the conditions and the protocols used to observe the behaviours. Every model would prove to be false as soon as the capacity of observation is increased (e.g. when one can observe small particles at high speed, the mechanics of Newton has to be replaced by the Einstein theory of relativity). Indeed, many false properties appear intrinsically irrefutable for lack of observability ! The observation of the behaviours is limited, even more in biology where certain properties, having however a strong explanatory importance, cannot be observed. Consequently, for the assumptions and the conjectures related to the model, we must systematically try to produce judicious observable consequences of them to make experiments. We have to choose the observable consequences which optimize the chances of refutation.

An effective observability must aim at refuting a model. It must try to reveal its internal inconsistencies, find contradictions with other models supposed to be compatible with it, or exhibit predicted behaviours which differ from biological reality.

Once a model has been formalized, its related assumptions specified and its intrinsically observable properties defined, one theoretically can use a data-processing program based on formal logic to determine the choice of the experiments of refutation, to evaluate the level of testability of certain assumptions, to ensure a certain level of covering of the set of refutations, to point out non covered generic cases, etc. Indeed, these computer aided tools already exists in the software testing activity. A strong similarity exists between software testing and the refutation of biological models :

1. The activity of modelling extrapolates a model from a reasonable number of biological experiments and tries to give a certain confidence that this model is conform to the reality. The activity of testing a software tries to extrapolate from a reasonable number of tests a certain confidence in its conformity with the specifications.
2. The assumptions of modelling mainly rely on certain regularities or uniformities of the real behaviours. The assumptions of software testing consist in considering that each test represents a whole set of tests uniformly sharing its behaviour.
3. Biological observability is limited by instrumentation, ethics, etc. The observability of a software is limited by the input/output peripherals (e.g. screen) and to the memory states which can be accessed during the testing activity.

A test of software proceeds by, first, selecting test scenarii from the specification of what the software is supposed to do, second, carrying out on the computer the selected tests, third, analyse the results of these tests to determine the parts of the software which fail, and finally, make corrections to the software. Modelling in biology will find an interest to operate in a way similar to software testing. By analogy, what computer scientists call “the level of specification” is similar to the biological level of description. The notion of “software testability” can help to install suitable conditions of experiment and/or simulation. This analogy enforces the importance of a well choosen level of description and abstraction, according to the level of possible observations. In the analysis and the comprehension of a biological phenomenon, a careful definition of what is observable could be the key to properly define the level of description.

Références

[1] J. Govan and V. Deretic, “Microbial pathogenesis in cystic fibrosis : Mucoid pseudomonas aeruginosa and burkholderia cepacia,” *Microbiol Rev*, vol. 60, pp. 539–574, 1996.

[2] J. Guespin-Michel and M. Kaufman, "Positive feedback circuits and adaptive regulations in bacteria," *Acta biotheoretica*, vol. 49, pp. 207–218, 2001.

[3] R. Thomas, "On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations," *Springer Series in Synergies*, vol. 9, pp. 180–193, 1980.

[4] R. Thomas and M. Kaufman, "Multistationarity, the basis of cell differentiation and memory. ii. logical analysis of regulatory networks in terms of feedback circuits," *Chaos*, vol. 11, pp. 3375–3382, 2001.

[5] M. R. Huth and M. D. Ryan, *Logic in Computer Science : Modelling and Reasoning about Systems.* Cambridge University Press, 2000.

[6] R. Lalement, *Logique, réduction, résolution*. Masson, 1990.

[7] E. Snoussi, "Qualitative dynamics of a piecewise-linear differential equations : a discrete mapping approach," *Dynamical Stab. System*, vol. 4, pp. 189–207, 1989.

[8] K. Popper, *conjectures et réfutations*. Taylor and Francis Books, 1969.