Modelling, observability and experiment: a case study
- Positive feedback loop in a genetic regulatory network -

Gilles Bernot[1], Janine Guespin-Michel[2], Jean-Paul Comet[1], Patrick Amar[1,3], Abdallah Zemirline[4], Frank Delaplace[1], Pascal Ballet[4], Adrien Richard[1]

[1]Laboratoire de Méthodes Informatiques, Université d'Évry Val d'Essonne & CNRS UMR 8042, Tour Évry 2, 523 Place des terrasses de l'agora, F-91000 Évry
[2]Laboratoire des Processus Intégratifs Cellulaires, Université de Rouen & UPRESA CNRS 6037, F-76821 Mont-Saint-Aignan Cedex
[3]Laboratoire de Recherche en Informatique, Université Paris Sud & CNRS UMR 8623, 15 avenue George Clémenceau, F-91405 Orsay Cedex
[4]Departement Informatique, EA 2215, Université de Bretagne Occidentale, BP 809, F-29285 Brest Cedex

Abstract

We propose an new methodology for modelling of biological regulatory networks inspired by the design and validation of large computing systems. We take into account the capability to validate a model by a set of biological experiments. So defining a model goes with experimental methods and conditions 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 assumed 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? Although some biological scientific communities already use modelling routinely, other are still very reluctant to that usage, and molecular biologists most often think modelling useless for their field because it does not contribute to experimental knowledge. This may involve a large range of reactions, from violent reject to polite interest This leads often model makers to use data from the literature, which may lead to very interesting models, but without enough feedback from and toward experimentation.

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 benefit to both disciplines. This is where analogy between computing systems specification and modelling in biology is involved. 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 testing methods *via* 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 software under test will have a behaviour compatible with its model.

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

- the *operability* is the capability to make a program run some chosen 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 limited user interface of the 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 model sould be 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.

In the case we will study here a first modelling step (already published), based on the multistationnarity theory, makes an innovating hypothesis plausible (section 2). A second mathematical step, based on formal logics in computer science, allowed us to determine the biological experiments sufficient to validate or invalidate the hypothesis (section 3).

Far further this biological example we have created a new software environment which integrate sophisticated algorithms from computer science (`SMbioNet`). This software environment allows 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 existent tools in computer science.

Of course such an approach needs a tight collaboration between biologists and model makers. The models designed this way, are not only an attempt for *a posteriori* explanation 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 sojourn 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 [1].

But another 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 [2].

A very simplified model of the regulatory network has been constructed [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[1] represent: the self-regulation of variable x (arc $x \to x$), the transcription of the genes encoding the anti-AlgU (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 anti-AlgU (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 protein. It has been observed that in the mutant bacteria y is removed from the corresponding graph. The negative action of y on x disappears.

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 epigenesis [3]. It is proved in this framework that the other known regulatory interactions are of minor importance with regard to the question of the existence of an epigenetic modification.

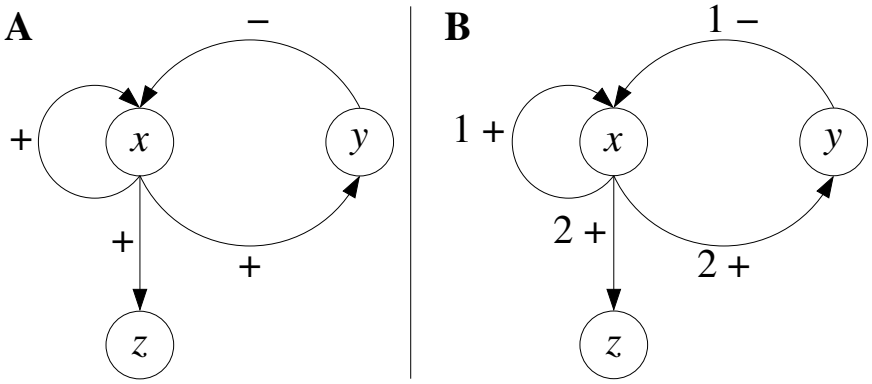

Figure 1: Genetic network regulating the production of mucus (alginate) by the bacterium *Pseudomonas aeruginosa* (simplified model). **A** shows the activation ($+$) or the inhibition ($-$) between the genes (resp. proteins) of the network. In **B** each arc is labelled with the minimal value of the threshold for which this action (activation or inhibition) is triggered. The variables means: x AlgU protein, y anti-AlgU, z mucus synthesis.

This model can be studied by a system of differential equations or by generalised logical analysis [4]. The generalised logical approach has been proved to be a correct approximation of the differential approach. To summarise, when variable x interacts with variable y, the curve that represents the level of y as a function of the level of x is a sigmoid. This sigmoid defines a threshold $S_{(x,y)}$ (Figure 2-a). Similarly the influence of x on another variable z defines another threshold $S_{(x,z)}$ (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 (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 $x \xrightarrow{2+} z$ (Figure 1-B). In order to describe the dynamic evolution of the system, we have to specify the level reached by a variable v as a function of the levels of the other variables that influence it (Figure 1). These values are represented by function K.

[1] the arrows in Figure 1

In our model, two feedback circuits co-exist. The functionnality of the positive feedback loop $x \rightarrow x$ is a necessary condition for the existence of two stable states: if x is high, it is self-maintained, else, if it is below the first threshold, it remains so. The negative feedback circuit may attract 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 [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.

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

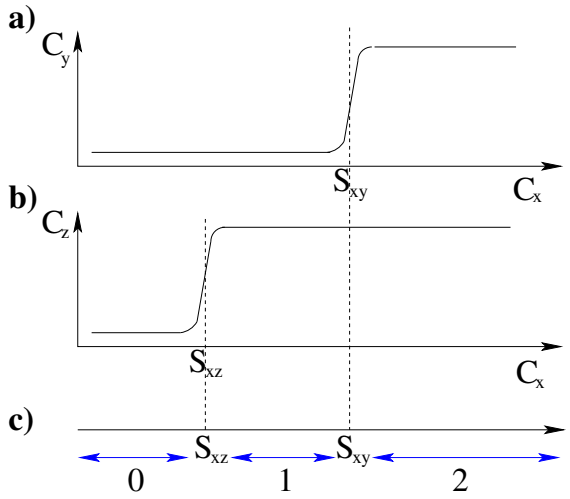

Figure 2: Influence of the variable x on the variables y and z

3 Formal logic to propose experiments

In this section, 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. This leads naturally to temporal logics (see [5], [6] for a general description of temporal logics).

For a given graph, there is a large number of models depending on the values of the tresholds and on all the possible functions K. Each one of these models defines a specific temporal behaviour. A temporal logic formula expresses a property which can be used to split the set of models into two parts: the models which satisfy the property and those which do not. Formalising biological knowledge into temporal formulae will allow us to consider only sensible models.

Our software environment `SMbioNet` allows the user to draw the regulatory network as in Figure 1, assign the thresholds when they are known, and then generate all the compatible models. `SMbioNet` allows the user to enter temporal logic formulae to keep only the models which satisfy them.

In our case study the epigenetic hypothesis means to prove that *in the presence of* y in the graph it is possible to have a recurrent state in which mucus is produced. By construction of the graph for the wild bacterium, y is present and the topology of the graph has been biologically validated as well as the signs of interactions. We have entered this graph in `SMbioNet` which has generated more than 600 models each one with its own behaviour corresponding to different thresholds and function K.

The epigenetic hypothesis is *consistent* if and only if there exist at least one model such that the behaviour can reach a state where z (mucus production) is expressed in a recurrent way. The only work we have to do is to write a temporal logic formula which expresses the recurrent expression of z. By giving this formula to `SMbioNet` we will automatically know how many models satisfy the formula.

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 modalities particular to this type of logic related to time. We can for example create the modality $\mathcal{F}_s$, which means that the formula which follows the modality is true in the "strict future". We call here "strict future" the future starting after a certain amount of time (i.e. excluding the current state). This amount of time must be chosen according to biological considerations on the studied system.

We want to verify if there is a model such that, if at a given time the bacterium is in a mucous state, then later (in a strict future) it will be again in a mucous state. This can be formalised as:

$$(z = 1) \Longrightarrow \mathcal{F}_s(z = 1)$$

Indeed we know enough about the graph structure to deduce that $z = 1$ is equivalent to $x = 2$. We know experimentally that the threshold associated with the interaction $x \to z$ is the maximal value of x (equal to 2), and we know by construction that the threshold associated to $y \to x$ 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 $x \to x$ and $x \to y$. 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, the epigenetic hypothesis means that the relative forces of these two circuits are such that it is possible to make recurrent the state $(x = 2)$. Consequently the epigenetic hypothesis can be written as:

$$(x = 2) \Longrightarrow \mathcal{F}_s(z = 1)$$

Note that this formula expresses that $z = 1$ in a recurrent way because when $z = 1$ at a given time, $x = 2$ at the same time, which in turn implies that $z = 1$ will be true again in the future, and so on.

`SMbioNet` has shown that several models satisfy this formula, which means that the epigenetic hypothesis is consistent.

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

Table 1: Truth table of $p \Rightarrow q$

The consistency of the epigenetic hypothesis being established, it remains to find experiments which prove it *in vivo*. As we can see on table 1, when the left part of the formula $(x = 2)$ is false,

the whole formula is true regardless the value of the right part. So the only revealing experiments always start by assigning (artificially if necessary) x to 2. The scenario of experiments is thus the following:

1. Start by imposing ($x = 2$).
2. Wait a lapse of time (of course the length of it remains empirical) then stop imposing (x =2) and test the phenotype for as many subsequent generations as possible.
    - If the bacterium has not changed it's phenotype, then the experiment *a priori* fails.
    - If the bacterium has become mucoid ($z = 1$) and remains so for several generations after the external signal has been removed, then, epigenesis is proved.

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. The most rigorous one would be to introduce into wild type cells of *Pseudomonas aeruginosa*, a plasmid carrying gene *algU* 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 models are 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: $x \to z$, 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 [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 *Observability* working group in Genopole® gives a methodological framework to define models including a tool kit for experimental validation or refutation. This way, our work resolutely enforces the modelling activity. It increases its credibility 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.

Establishing 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. SMbioNet is a software assistant for the design of biological models which fully handles temporal logic. More than this, this approach can be used to generate and optimise biological experiment scenarii. We have shown the feasibility of the approach on the *Pseudomonas aeruginosa* example. Extending SMbioNet to also suggest experiment scenarii requires to continue our investigations on the application of formal methods from computer science to life science.

Acknowledgements

We thank Christelle Koundibia, Catherine Meignen and Hélène Pollard for organisational help and Genopole Recherche for supporting our interdisciplinary work groups.

References

[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.