Towards Executable Biology

Nicola Bonzanni

Towards Executable Biology

Ph.D. Thesis

Nicola Bonzanni

VU University Amsterdam, 2012

vrije Universiteit amsterdam

This research was founded by the VU University Amsterdam.

Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Cover: design by Nicola Bonzanni. The nucleotides sequence on the back cover is part of the first chromosome of the *Felis catus* (cat) genome.

ISBN 978-94-6203-156-2

Printed by Wöhrmann Print Service, Zutphen

VRIJE UNIVERSITEIT

Towards Executable Biology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op maandag 5 november 2012 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Nicola Bonzanni

geboren te Ponte San Pietro, Italië

promotoren:	prof.dr. J. Heringa
	prof.dr. W. Fokkink
copromotor:	dr.ir. K.A. Feenstra

To my family

Contents

CHAPTER 1

INTRODUCTION

1.1 A language for systems biology

On 28 February 1953, Francis Crick walked into the Eagle pub in Cambridge and announced: "We have found the secret of life". At least that is what the legend says. Actually, that very morning James D. Watson and Francis Crick had figured out the double helix structure of DNA. Some 50 years later, on 26 June 2000, the U.S. President Bill Clinton and the British Prime Minister Tony Blair jointly announced - to a much larger audience than the Eagle crowd - the completion of the first survey of the entire human genome by the Human Genome Project. The White House press release reads: *"Today's announcement represents the starting point for a new era of genetic medicine"*. The promise of personalized medicine has yet to be fulfilled but the genomic revolution has catapulted molecular biology to the realm of systems biology (Westerhoff & Palsson, 2004). Indeed, the exceptional breakthroughs of the last century and the advent of high-throughput technologies in the early 2000s has crucially changed the way in which biology is practiced today.

Although it has been noted (Westerhoff & Palsson, 2004) that the roots of systems biology can be traced back to the history of molecular biology and control theory, the new *Zeitgeist* perfusing life science in the post-genomic era suggests a widely accepted change in the way biologists view their science. It is still unclear whether the emergence of systems biology as a field in its own right is a Kuhnian 'paradigm shift' to the study of integrated systems from the reductionism that characterized biochemistry and molecular biology. However, if we were facing a new paradigm, we should be facing a profound and lasting change in the language of biology since, in Kuhn's theory, competing paradigms are incommensurable (Kuhn, 1996). Regardless of whether we accept Kuhn's epistemological ideas, the necessity of a new language for biology is clearly perceived by the biological community itself (Moya *et al.,* 2009). The experimentalist Yuri Lazebnik, in his famous paper "Can a biologist fix a radio?" (Lazebnik, 2002), writes:

> “The language used by biologists for verbal communications is not better and is not unlike that used by stock market analysts. Both are vague (*e.g.*, “a balance between pro- and antiapoptotic Bcl-2 proteins appears to control the cell viability, and seems to correlate in the long term with the ability to form tumor”) and avoid clear predictions. [...] However, I hope that it is only a question of time before a user-friendly and flexible formal language will be taught to biology students, as it is taught to engineers, as a basic requirement for their future studies. My advice to experimental biologists is to be prepared.”

Systems biology inherited the denotational language proper of control theory. Denotational languages, such as differential equations, express mathematical relationships between quantities and how they change over time. Although ordinary differential equations have been successfully applied in different systems, the two key assumptions of this approach, continuity and determinism, are not always fulfilled (Moya *et al.,* 2009). Metaphorically one can ask the question whether molecules in a cell, or cells themselves, solve differential equations to decide what to do in a particular situation, or rather follow simple sets of rules when they encounter one another derived from their physical interactions (Bonzanni *et al.,* 2009b). Furthermore, the flourishing of ontologies (Ashburner *et al.,* 2000) and ‘executable’ formal languages (Fisher & Henzinger, 2007) suggests that a denotational language alone might not be enough to satisfactorily replace, or at least fill the formal void, of the current systems of notation, diagrams, and descriptive knowledge that faithfully have served biology for the last couple of centuries.

Many formal languages have been set forth for consideration, yet the scientific community has not reached an agreement on the most suitable one to fully capture biological systems. Therefore, the question “what is *the* best formal language to write the secret of life?” is still without an answer and unlikely to ever have one. In fact, talking about languages for biology without specifying the abstraction level at which they are applied is a pointless exercise. The real question should rather be at which level of abstraction we want to understand living systems and what are the best languages to capture such abstractions. In 1976, the same year of publication as his most famous book “The Selfish Gene”, but in a different essay (Dawkins, 1976), Richard Dawkins writes:

> “If a computer is doing something clever and life-like, say playing chess, and we ask how it is doing it, we do not want to hear about transistors, we simply accept them. The useful answer to the question is purely in terms of software; indeed the programme is likely to be written in such a way that it could easily be run with completely different hardware.
>
> We need software explanations of behaviour. I do not mean that animals necessarily work like computers. They may be very different. But just as the lowest level of explanation is not always the most appropriate for a computer, no more it is for an animal. Animals and computers are both

> so complex that something on the level of software explanation must be appropriate for both of them.”

Interestingly, it was during the same period in which Dawkins wrote the above paragraphs that many of the major language paradigms now in use were invented (*e.g.* C between 1969 and 1973, Prolog in 1972). Some 40 years and several programming languages later, it is possible to push his metaphor even further. Nowadays, it is common knowledge that a programming language suitable to encode a software application, might not be appropriate to encode a different one, although both applications will possibly run using the same hardware. Modern programming languages differ from one another; they are often designed to operate at different abstraction levels, and in different application domains, but all of them can eventually be reduced into the same “ground form” which specifies a set of basic operations. It is possible and common to develop a single computer application employing multiple programming languages, where each component is written using the most suitable language. Similarly, a (small) set of formal languages for biology should be defined, such that it is possible to translate the languages into one another, or at least to reduce each of them to a common “ground form” that allows for compositionality, paving the way towards multi-scale modelling, and, eventually, the systematic engineering of new living organisms.

1.2 Thesis outline

The need to define a framework built upon complementary and yet coherent formal languages for systems biology repeatedly emerged during my studies. Instead of a systematic implementation of a framework defined *a priori*, this thesis should be regarded as a preliminary investigation of the features and requirements of such a framework, as they surfaced from the direct application of formal methods to concrete biological case studies.

Each chapter of this thesis tackles a different biological process (*e.g.* signalling, gene regulation) in a different organism (*e.g.* yeast, *C. elegans*). Notwithstanding the diverse biological applications, all case studies share a common formal method: Petri nets. The choice to use Petri net, was initially dictated by my own familiarity with this modelling technique, and subsequently by its intuitive graphical representation reminiscent of the traditional cartoons used in biology. As stated above, a single method can hardly cover all the facets of biology at all levels of abstraction. However, Petri nets seemed a suitable language to bridge the communication gap between biologists and computational scientists. Furthermore, exploring different processes using the same technique provided the opportunity to better assess strengths and weaknesses of Petri nets in a biological context.

Chapter 2 extends the concepts enucleated in this introduction and argues that operational modelling approaches from the formal methods community can fruitfully be applied within the systems biology domain. Chapter 3 demonstrates a large scale application of formal methods to multi-cellular pattern formation. We applied our model-

ling approach to the well-studied process of *C. elegans* vulval development, showing that our model correctly reproduces a large set of *in vivo* experiments with statistical accuracy. Chapter 4 focuses on signalling networks. We investigated the effect of proteolysis after nutrient starvation in *S. cervisiae*. Particularly, we showed how computational models, bioinformatics analyses and *in vivo* observations can be integrated in order to formulate and validate novel biological hypotheses. The last case study is presented in Chapter 5. We constructed a regulatory network model based on the functionality of cis-regulatory elements in order to generate fundamental insights into cellular fate differentiation during haematopoiesis. Finally, in Chapter 6 I elaborate on the results of this PhD work and discuss possible future directions towards the systematic and successful adoption of formal languages in systems biology.

CHAPTER 2

WHAT CAN FORMAL METHODS BRING TO SYSTEMS BIOLOGY?

Partially adapted from:

What can formal methods bring to systems biology?
Nicola Bonzanni[1 2] , K. Anton Feenstra[1 2] , Wan Fokkink[2] , and Elzbieta Krepska[2]
Lecture Notes in Computer Science [FM09] 5850/2009:16–22 (2009)

Design issues for qualitative modelling of biological cells with Petri nets
Elzbieta Krepska[2] , Nicola Bonzanni[1 2] , K. Anton Feenstra[1 2] , Wan Fokkink[2] , Thilo Kielmann[2] , Henri Bal[2] , and Jaap Heringa[1 2]
Lecture Notes in Bioinformatics [FMSB08] 5054/2008:48–62 (2008)

[1]Centre for Integrative Bioinformatics, VU University Amsterdam, The Netherlands
[2]Department of Computer Science, VU University Amsterdam, The Netherlands

Systems biology studies complex interactions in biological systems, with the aim to understand better the entirety of processes that happen in such a system, as well as to grasp the emergent properties of such a system as a whole. This can for instance be at the level of metabolic or interaction networks, signal transduction, genetic regulatory networks, multi-cellular development, but also at higher levels such as social behaviour of insects.

Biological systems are reactive systems, as they continuously interact with their environment. In November 2002, David Harel (2004) put forward a grand challenge to computer science, to build a fully animated model of a multi-cellular organism as a reactive system; specifically, he suggested to build such a model of the *C. elegans* nematode worm, which serves as a model organism in developmental biology.

Open questions in biology that could be addressed in such a modelling framework include the following, listed in order from a detailed, molecular viewpoint to a more global view of whole organisms: How complete is our knowledge of metabolic, signaling and regulatory processes at a molecular level? How is the interplay between different pathways or network modules organized and regulated? How is the interaction between intra-cellular processes and inter/extra-cellular processes organized? How do cells self-organize? How do cells differentiate? How are self-organization and differentiation of cells connected? How does self-organization and differentiation lead to the formation of complex structures like organs (*e.g.* the eye, brain, kidney)?

One grand open question that pervades the whole of biological research is, how could all of this *evolve*? This is exemplified by the 1973 essay by Theodosius Dobzhansky (1973) entitled “Nothing in biology makes sense except in the light of evolution”. Some recent theoretical work (Crombach & Hogeweg, 2008) highlights an interesting possibility, that flexibility in regulation is a necessary component of evolution, but has itself evolved in biological systems.

2.1 Formal models of biological systems

A formal model is mathematical model of a process or system at some chosen level of abstraction. Why, then, would a biologist want to use *formal* models? They can be an excellent way to store and share knowledge on biological systems. Furthermore, *in vivo* experiments in the lab tend to take an awfully long time, and are labour intensive. In comparison, *in silico* experiments on a computer can take relatively little time and effort. For instance genetic perturbations can be difficult (or unethical) to perform in the lab, while they may require trivial adaptations of a formal model.

The time is ripe for exploiting the synergy between (systems) biology and formal methods. First of all we have reached the point where biological knowledge of, for instance, signal transduction has become detailed enough to start building sensible formal models. Second, the development of analysis techniques for formal methods, and the power of the underlying computer hardware, has made it possible to apply formal methods to very complex systems. Although we are certainly not (and possibly never will be) at a level where a full-fledged formal analysis of the entire genetic reg-

ulatory network of one cell is within reach, we can definitely already study interesting, and challenging, fragments of such networks.

It is important to realise that biology (like physics, chemistry, sociology, and economics) is an empirical science. This is basically orthogonal to the standard application of formal methods in computer science, where a formal analysis is used to design and prove properties of a computer system. If a desired property of a computer system turns out to fail, then we can in principle adapt the system at hand. In contrast, biological systems are simply (and quite literally) a fact of life, and formal models 'only' serve to better understand the inner workings and emergent properties of such systems. While in computer science, model validation typically leads to a redesign of the corresponding computer system or implementation, in systems biology it leads to a redesign of the model itself. A nice comparison between these two approaches can be found in the introduction of Sadot *et al.* (2008).

Fisher & Henzinger (2007) distinguish two kinds of formal models for biological systems: operational versus denotational (or, as they phrase it, computational versus mathematical). On the one hand, operational models (such as Petri nets) are executable and mimic biological processes. On the other hand, denotational models (such as differential equations) express mathematical relationships between quantities and how they change over time. Denotational models are in general quantitative, and in systems biology tend to require a lot of computation power to simulate, let alone to solve mathematically. Also it is often practically impossible to obtain the precise quantitative information needed for such models. Operational models are in general qualitative, and are thus at a higher abstraction level and easier to analyse. Moreover, Fisher and Henzinger, as well as Regev & Shapiro (2002), make a convincing case that a good operational model may explain the mechanisms behind a biological system in a more intuitive fashion than a denotational model.

An operational model progresses from state to state, where an event at a local component gives rise to a state transition at the global system level. Fisher *et al.* (2008) argue that (unbounded) asynchrony does not mimic real-life biological behaviour properly. Typically, asynchrony allows that one component keeps on executing events, while another component is frozen out, or executes only few events. While in real life, all components are able to execute at a certain rate. Bounded asynchrony, a phrase coined by Fisher *et al.* (2008), lets components proceed in an asynchronous fashion, while making sure that they all can proceed at their own rate. A good example of bounded asynchrony is the maximally parallel execution semantics of Petri nets (Burhard, 1983); we will return to this semantics in the next section.

We briefly mention the three modelling paradigms from the formal methods community that are used most frequently for building operational models of biological systems.

Petri nets are well-suited for modelling biochemical networks such as genetic regulatory networks. The places in a Petri net can represent genes, protein species and complexes. Transitions represent reactions or transfer of a signal. Arcs represent reaction substrates and products. Firing of a transition is execution

of a reaction: consuming substrates and creating products. Cell Illustrator (Nagasaki *et al.,* 2009) is an example of a Petri net tool that targets biological mechanisms and pathways.

Process calculi, such as process algebra and the π-calculus, provides algebraic laws to manipulate agents and processes in concurrent systems. When extended with probabilities or stochastics, it can be used to model the interaction between organisms. Early ground-breaking work in this direction was done by Tofts (1992), who used process algebra to simulate behavioral patterns in ants. The Bioambients calculus (Regev *et al.,* 2004), which is based on the π-calculus, targets various aspects of molecular localisation and compartmentalization.

Live sequence charts are an extension of the graphical specification language of the message sequence charts; interaction diagrams originally intended for modelling communication behaviour in real-time systems. Notably, they allow a distinction between mandatory and possible behaviour. They have been used successfully by Harel and his co-workers to build visual models of reactive biological systems, see *e.g.* Kam *et al.* (2003).

Model checking is in principle an excellent methodology to verify interesting properties of specifications in any of these three formalisms. In practice, abstraction techniques and distributed model checking (see *e.g.* Barnat *et al.* 2008) allows to verify global properties of non-trivial systems. However, in view of the very large scale and complexity of biological systems, so far even these optimisation techniques cannot push model checking applications in this area beyond toy examples. Simulation methods are commonly used to evaluate complex and high-dimensional models, and are applicable in principle to both operational and denotational models. Well-known drawbacks, compared to model checking, are that this approach can suffer from limited sampling due to the high-dimensional state space, and that there may be corners of the state space that have a biological relevance but that are very hard to reach with simulations. Still, in spite of these drawbacks, Monte Carlo simulations are currently the best method to analyse formal specifications of real-life biological systems.

2.2 Petri nets

In our view, for the successful application of formal methods in the systems biology domain, it is expedient to use a simple modelling framework, and analysis techniques that take relatively little computation power. This may at first sound paradoxical, but simplicity in modelling and analysis methods will make it easier to master the enormous complexity of real-life biological systems. Moreover, it will help to communicate with biologists on the basis of formal models, and in the hopefully not too far future will make it attractive for biologists to start using formal modelling tools.

Based on these considerations, we built our modelling frameworks around three concepts: (i) biological significance, (ii) communicative power, and (iii) simplicity. In

their basic formulation, Petri nets have a simple definition, an intuitive graphic representation that resemble biological cartoons, and they can be easily adapted to model biological processes. Hence, we chose to ground our modelling work upon Petri net theory.

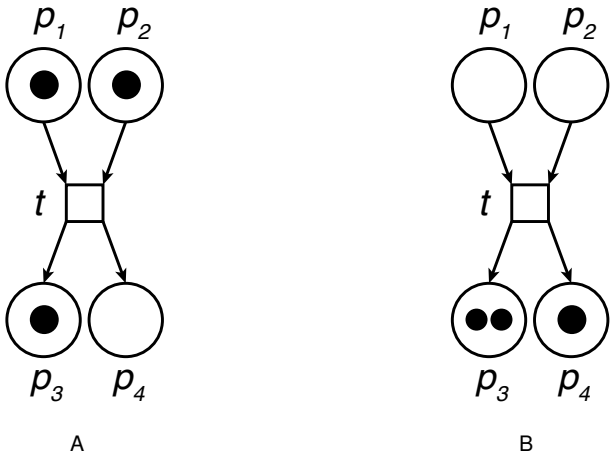


Figure 2.1: An example of a place transition network marked by two different markings. Places $p_1, p_2, \ldots$ are depicted as circles, the transition t as a square, and the flow relations as directed arcs. The number of black dots (*i.e.* tokens) in each place represent their marking. The marking shown on the right (B) is reached by the occurrence of t in the network shown on the left (A).

Intuitively, a Petri net (Petri, 1962; Reisig & Rozenberg, 1998) is a bipartite directed graph consisting of two kinds of nodes: places that indicate the local availability of resources, and transitions which are active components that can change the state of the resources. Each place can hold one or more tokens. Arcs connect places and transitions. Tokens can flow from place to place following the arcs. More formally, a place-transition (PT) Petri net is a tuple $N = \langle P, T, F \rangle$ where P is a set of places and T a set of transitions with $P \cap T = \emptyset$. $F \subset (P \times T) \cup (T \times P)$ is a set of flow relations that defines the arcs. Given a transition $t \in T$, the pre-set of t is the set of its input places ${}^\bullet t = \{p \in P \mid (p,t) \in F\}$. The post-set of t is the set of its output places $t^\bullet = \{p \in P \mid (t,p) \in F\}$. A state (or marking) of a PT net is determined by a distribution of tokens on its places and is defined by a mapping $\mathrm{m} : P \to \mathbb{Z}_{\geq 0}$. A place s is marked by a marking m if $\mathrm{m}(s) > 0$. A transition t is enabled by a marking m if m marks all the places in ${}^\bullet t$. If t is enabled, then it can fire. The occurrence of t transforms the marking m into a mapping m' defined as

$$\mathrm{m'}(p) = \begin{cases} \mathrm{m}(p) - 1 & \text{if } p \in {}^\bullet t - t^\bullet, \\ \mathrm{m}(p) + 1 & \text{if } p \in t^\bullet - {}^\bullet t, \\ \mathrm{m}(p) & \text{otherwise.} \end{cases} \tag{2.1}$$

Petri nets can be represented as an integer matrix, *i.e.* the incident matrix, or as a graph where places are depicted as circles, transitions as rectangles, arcs as arrows, and tokens as dots (see Figure 2.1). Instead of further enriching this formalism to extend its expressiveness (but also its complexity), we focused on preserving the simplicity of the formalism, and developing an execution semantics which resembles biology.

2.2.1 Petri net framework for modelling signal transduction

The first formalism we developed was built to model signalling networks. Specifically, the *C. elegans* vulva development process (see Chapter 3). In this framework, the translation of places and transitions into biological entities is straightforward. Places represent genes, protein species, and complexes. However, we have encountered many cases when we had to represent a single entity with various characteristics as multiple places. For example, to differentiate between active and inactive molecules conformations of the same protein. Transitions represent reactions or transfer of a signal. Arcs represent reaction substrates and products. Firing of a transition is execution of a reaction: consuming substrates and creating products. A marking in our model does not represent directly the number of molecules or a fixed molar concentration as in Gilbert *et al.,* 2007. We interpret this number in two ways. For genes as a boolean value: 0 means not present and 1 present. For proteins, we use abstract concentration levels $0-6$: going from not expressed, via low, medium, and high concentration to saturated level. The rationale behind this approach is to abstract away from often unknown absolute molecule concentration levels, as we intend to represent relative concentrations. We choose to use seven concentration levels in order to stay in between a simple boolean level and a complex ODE model, and because seven concentration levels sufficed to express the biological knowledge from the literature on *C. elegans* vulva development in a satisfactory fashion. If desired, a modeller could fine-tune the granularity of the model by adjusting the number of available concentration levels.

The interleaving semantics of Petri nets describes an asynchronous behaviour. A fully asynchronous approach would allow for prolonged activity in only one part of the network while another part shows no activity at all. However, biological systems are highly concurrent, as in cells all reactions can happen in parallel and most are independent of each other. The high level of compartmentalisation in cells, for example, guarantee that reactions can take place in each compartment, independently. Since each compartment can progress simultaneously, in a parallel fashion, we apply a principle of maximal parallelism (Burhard, 1983).

The maximal parallel execution semantics can be summarised informally as *execute greedily as many transitions as possible in one step*. A step $\mathcal{S} : T \to \mathbb{N}^0$ is a multi-set of transitions, *i.e.* a transition can occur multiple times in $\mathcal{S}$. A maximally parallel step is a step that leaves no enabled transitions in the net, and in principle should be modelled in such a way that it corresponds to one time step in the evolution of the biological system. This is possible because the modeller can capture relative

speeds using appropriate weights on arcs, where weights are defined by a mapping $w : (P \times T) \cup (T \times P) \to \mathbb{Z}_{\geq 0}$ such that $w(a) = 0$ if $a \notin F$. Then Equation 2.1 can be written as

$$m'(p) = m(p) + w(t, p) - w(p, t). \tag{2.2}$$

Typically, if in one time unit a protein A is produced four times more than a protein B, then the transition that captures production of A should have a weight that is four times as large as the weight of the one that captures B production (see Figure 2.2). Similarly, it is possible to model different consumption speeds.

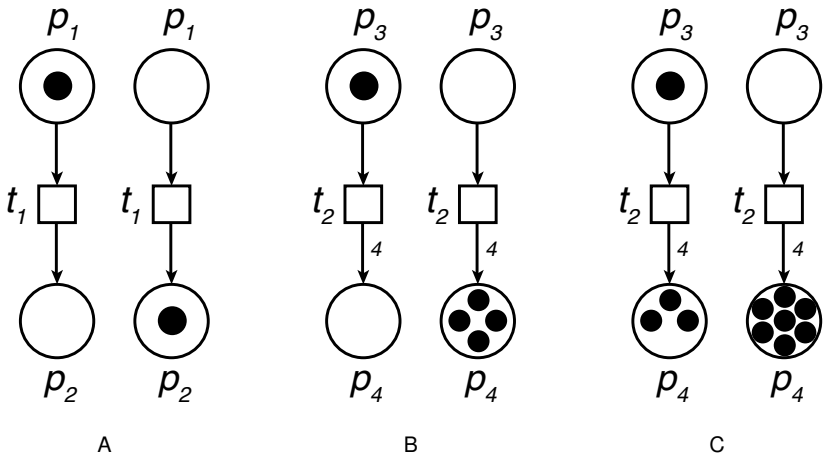


Figure 2.2: The tokens production of transition t_2 (B) is four time faster than the production of transition t_1 (A). On the right (C), transition t_2 overshoots place p_4. Arcs without a weight label have weight 1.

Implementing a pure maximally parallel semantics requires to generate all possible partitions of independent transition occurrences, and select one randomly, uniformly. However, with the growth of the network, this procedure becomes prohibitively slow. Since all transitions in a maximally parallel step must be independent, while building the maximally parallel step $\mathcal{S}$, we must ensure that all transitions $t \in \mathcal{S}$ can be fired $\mathcal{S}(t)$ times without regard to the order. Hence, the computational cost of generating the next maximally parallel step is at least $\Omega(\exp(|T|))$ as it involves verifying all subsets of T in the worst case, when all transitions are in conflict. Indeed, if two transitions t and t' are in conflict, *i.e.* ${}^\bullet t \cap {}^\bullet t' \neq \emptyset$, and both are enabled by the same marking, then firing one might disable the other and, therefore, they can't fire independently.

Thus, we approximated the maximally parallel semantics by building a maximally parallel step incrementally, selecting one transition after another, randomly, until all enabled transitions have been exhausted as shown by the Algorithm 2.1.

Furthermore, unrestricted production of proteins is usually not realistic, as in

Algorithm 2.1 computeStep(N, m), where N is a Petri net marked by a vector m

```
1:  S ← ∅                          ▷ initialise the maximally parallel step S as an empty set
2:  m' ← m                         ▷ make a copy of the current marking
3:  E ← T                          ▷ assume all transitions are enabled
4:  while E ≠ ∅ do                 ▷ loop until there are no more enabled transitions
5:      t ← selectRandom(E)
6:      if t is enabled then
7:          for all p ∈ •t do      ▷ consume all required tokens in t pre-set
8:              m'[p] ← m'[p] − w(p, t)
9:          end for
10:         S ← S ∪ {t}            ▷ add t to the maximally parallel step
11:     else
12:         E ← E − {t}            ▷ remove t from the set of enabled transitions
13:     end if
14: end while
15: for all t ∈ S do
16:     for all p ∈ t• do          ▷ produce all required tokens in t post-set
17:         m'[p] ← m'[p] + w(t, p) · S(t)
18:     end for
19: end for
20: return m'                      ▷ return the marking produced by the firing of S
```

nature the cell would saturate with the product, and the reaction would slow down or stop. Therefore, to mimic this behaviour, each place has a predefined maximum capacity $\mathcal{N} = 6$. To guarantee that the highest concentration level can be freely attained, we introduced bounded execution with overshooting. Instead of redefining Equation 2.2 we changed the enabling condition, *i.e.* a transition can only fire if each place $p \in t^\bullet$ holds fewer than $\mathcal{N}$ tokens. Since each transition can possibly move more than one token at once into its output places, each transition can overshoot the pre-given capacity $\mathcal{N}$ at most once (Figure 2.2c). Therefore, the network is bounded with a finite bound $k \geq \mathcal{N}$. In Chapter 3 we discuss further strategies and patterns to pragmatically build models based on this formalism.

2.2.2 Petri net framework for modelling gene regulatory networks

In Chapter 5 we focus on modelling gene regulatory networks (GRN). These models are usually grounded on different assumptions with respect to signalling networks. GRNs are based on two elements: a set of genes and a set of interactions among them. In turn, interactions can be either positive or negative, when a gene product (*e.g.* mRNAs, proteins) has an enhancing or repressive effect, respectively, on the expression of another gene. GRNs are traditionally represented using directed graphs. This representation, although intuitive, can be ambiguous. A directed graph can not

capture the cooperative interactions that are essential to correctly reproduce the behaviour observed during *in vivo* experiments. Intuitively, we want to be able to express, without ambiguity, whether a *single* gene (*de facto* its gene product) or *multiple* genes are required for a specific interaction.

An elegant way to avoid ambiguities is to use Petri nets to encode gene regulatory networks. In this framework, places represent genes, transitions represent interactions, and the marking of a place models the level of gene expression. Unfortunately, this simple construction, based on the standard definition of PT nets, conflicts with three main assumption of GRNs; in a GRN (i) the gene products (tokens) are not consumed by the interactions (transitions); (ii) interactions might have negative effects on the gene products (tokens can be removed from post-set places); (iii) the absence of a gene product (a place not marked) can be a prerequisite for an interaction. Instead of redefining enabling conditions and Equation 2.1, we decided to build simple Petri nets modules with a topology that satisfy these assumptions by construction, and then use these modules to build larger GRNs.

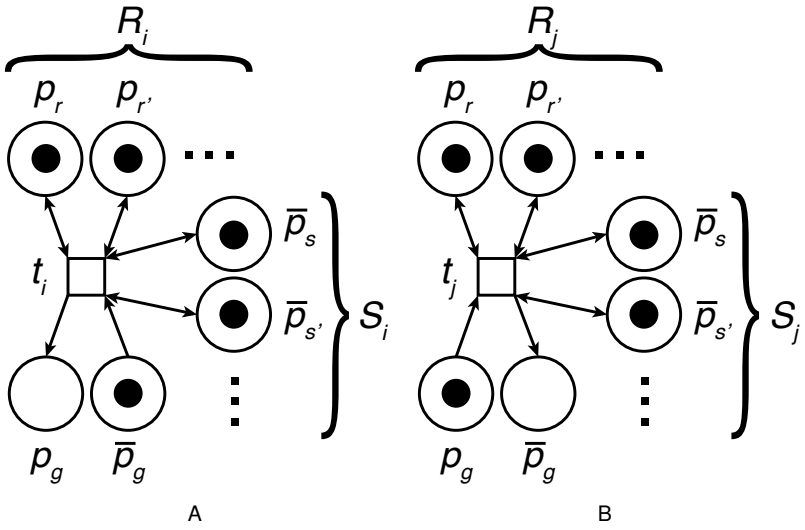


Figure 2.3: Petri net modules used to model positive (A) and negative (B) gene interactions in gene regulatory networks. R is the set of genes required for the occurrence of the interaction. S is the set of genes that block the interaction. Places p represent genes in expressed state while places $\bar{p}$ represent genes in repressed state. Arcs with a double arrow head denote arcs in both directions.

Based on the above PT network definition, and building on previous work by Chaouiya *et al.* (2004), we represent each gene g in the GRN using two complementary places p_g and $\bar{p}_g \in P$ such that the sum of tokens in p_g and $\bar{p}_g$ always equals $\mathcal{N} \in \mathbb{Z}_{>0}$. We use p_g to represent g being expressed, and $\bar{p}_g$ for g being repressed. $\mathcal{N}$ is the maximum gene expression level. Each interaction is modelled

by a transition. Let i be a positive interaction of the GRN. The set of genes R_i defines the gene products necessary for the occurrence of i, S_i defines the set of genes that block the occurrence of i, and g is the gene activated by the occurrence of i. Thus, it is possible to define a transition t_i modelling i such that

$$^\bullet t_i = \{p_r \in P \mid r \in R_i\} \cup \{\bar{p}_s \in P \mid s \in S_i\} \cup \{\bar{p}_g\},$$

and

$$t_i^\bullet = \{p_r \in P \mid r \in R_i\} \cup \{\bar{p}_s \in P \mid s \in S_i\} \cup \{p_g\}.$$

Intuitively, we want to enforce that t_i can be enabled if all the required gene products are available and all gene products blocking interaction i are absent. As a result of an occurrence of t_i, a token is moved from $\bar{p}_g$ to p_g, while all the tokens consumed in the places belonging to $^\bullet t_i$ are replaced by new ones. This construction, depicted in Figure 2.3a, complies with assumption (i) and (iii). Similarly, we can define a transition t_j that represent a negative interaction j on a gene g by moving a token from p_g to $\bar{p}_g$. Thus, the pre-set of t is

$$^\bullet t_j = \{p_r \in P \mid r \in R_j\} \cup \{\bar{p}_s \in P \mid s \in S_j\} \cup \{p_g\},$$

and the post-set of t is

$$t_j^\bullet = \{p_r \in P \mid r \in R_j\} \cup \{\bar{p}_s \in P \mid s \in S_j\} \cup \{\bar{p}_g\}.$$

This construction, depicted in Figure 2.3b, models the negative effect of j on g gene products (see assumption ii), and also complies with assumptions (i) and (iii). Therefore, by combining these constructions, and inferring F from the pre- and post-sets, it is possible to build a full GRN that satisfies all three assumptions using the basic PT net formalism.

One limitation of this approach is that the graphical representation of a non-trivial GRNs loses its intuitiveness. Indeed, the number of places and arcs necessary to model an interesting GRN explodes. Since an intuitive graphical representation is one of our goals, we tried to compress our construction by enriching the formalism. A minimal enrichment is sufficient to achieve an elegant formalism and an intuitive graphical representation. The extended formalism includes just an additional distinction between "positive" and "negative" arcs. Hence, our new framework is defined as a tuple $B = \langle \Pi, T, F, A, I \rangle$. Π is a set of places such that there exists a single place in Π for every gene in the GRN. T is the set of transitions. F is the set of flow relations as defined for PT nets. $A \subseteq F$ and $I \subseteq F$ are disjoint sets of positive and negative arcs respectively such that $F = A \cup I$. Given a transition $t \in T$, we call

$$\begin{array}{ll}
^\bullet t = \{\pi \in \Pi \mid (\pi, t) \in A\} & \text{positive pre-set of } t, \\
^\bullet \bar{t} = \{\pi \in \Pi \mid (\pi, t) \in I\} & \text{negative pre-set of } t, \\
t^\bullet = \{\pi \in \Pi \mid (t, \pi) \in A\} & \text{positive post-set of } t, \text{ and} \\
\bar{t}^\bullet = \{\pi \in \Pi \mid (t, \pi) \in I\} & \text{negative post-set of } t.
\end{array}$$

Note that, by definition, there exists a surjective function $\gamma : P \to \Pi$ that associates to each place $p \in P$ the place $\pi \in \Pi$ that corresponds to the same gene. Now, it is

possible to fold the constructions of Figure 2.3 into our new Petri net definition using Algorithm 2.2. This algorithm has two steps. The first step, intuitively, generates the set of places Π by compressing each couple of complementary places of P into a single place. The second step generates the set of arcs A and I. For each bidirectional arc from a positive place p to a transition t we create the arc $(\gamma(p), t)$ in A. For each bidirectional arc from a negative place $\bar{p}$ to a transition t we create the arc $(\gamma(\bar{p}), t)$ in I. Finally, given the effect, positive or negative, of the transition on a gene g we create an arc $(t, \gamma(p_g))$ in A or I, respectively. The unfolding procedure is intuitively similar. Each network $B = \langle \Pi, T, F, A, I \rangle$ can be graphically represented with great parsimony of elements as shown in Figure 2.4. This representation is intuitive but formally rigorous.

Algorithm 2.2 fold(N, m), where $N = \langle P, T, F \rangle$ is a PT Petri net marked by a vector m

$\Pi \leftarrow \emptyset$, $A \leftarrow \emptyset$, $I \leftarrow \emptyset$, m' $\leftarrow \emptyset$
for all pairs of complementary places $p_g, \bar{p}_{g'} \in P$ such that $g = g'$ **do**
    $\Pi \leftarrow \Pi \cup$ new(π) ▷ where π is a new place corresponding to gene g
    m'$[\pi] \leftarrow$ m$[p_g]$
end for
for all $t \in T$ **do** ▷ see Figure 2.3 notation
    for all $p_r \in P$ such that $r \in R_t$ **do**
        $A \leftarrow A \cup \{(\gamma(p_r), t)\}$
    end for
    for all $\bar{p}_s \in P$ such that $s \in S_t$ **do**
        $I \leftarrow I \cup \{(\gamma(\bar{p}_s), t)\}$
    end for
    if t models a positive interaction **then**
        $A \leftarrow A \cup \{(t, \gamma(p_g))\}$
    else
        $I \leftarrow I \cup \{(t, \gamma(p_g))\}$
    end if
end for
return $B = \langle \Pi, T, A, I, m' \rangle$

Although it is possible to unfold a network in its basic PT form where the traditional enabling conditions and marking transformations apply, it is convenient to lift these conditions and transformations. Thus, given a Petri net in the form $B = \langle \Pi, T, F, A, I \rangle$, a transition $t \in T$ is said to be pre-enabled by a marking m if every place in ${}^\bullet t$ is marked by m, and every place in ${}^\bullet \bar{t}$ is *not* marked by m; it is said to be post-enabled if m$(\pi) < \mathcal{N}$ for each place $\pi \in t^\bullet$, and m$(\pi) > 0$ for each place for each place $\pi \in \bar{t}^\bullet$. Finally, t is said to be enabled by a marking m if it is pre- and post-enabled by m. The

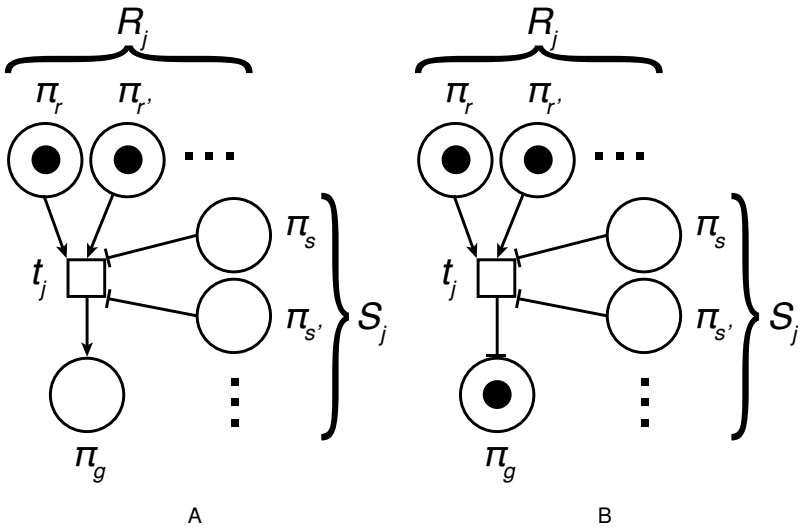


Figure 2.4: Graphical representation of the interactions of Figure 2.3 using the folded network definition $B = \langle \Pi, T, F, A, I \rangle$. Negative arcs belonging to I have a flat arrowhead. Both transitions are enabled by the depicted markings. In practice, R and S are small, further simplifying the representation. A realistic network model based on this formalism is shown in Figure 5.4

occurrence of t transforms the marking m into a mapping m' defined as

$$\text{m'}(\pi) = \begin{cases} \text{m}(\pi) - 1 & \text{if } \pi \in \bar{t}^{\bullet}, \\ \text{m}(\pi) + 1 & \text{if } \pi \in t^{\bullet}, \\ \text{m}(\pi) & \text{otherwise.} \end{cases} \tag{2.3}$$

This formalisation can be extended to use arc weights as shown before. However, by preserving this simple definition and assigning $\mathcal{N} = 1$, it is still possible to generate biologically meaningful networks as demonstrated in Chapter 5. These networks behave in a boolean fashion, *i.e.* each gene can either be expressed or repressed.

In order to build biologically faithful networks, an additional fourth assumption must hold: if transcription is suspended, all gene products should, eventually, degrade over time. In Petri nets terms, although tokens are not consumed by gene interactions, they must be consumed whenever the conditions that enable their production cease to hold. Intuitively, if a gene product g was produced in a previous step, but currently there are no more pre-enabled transitions that could have a positive effect on π_g, then gene product g should be degraded. This behaviour is achieved by adding new *ad hoc* transitions enabled when the conditions for gene activation are not met. Fortunately, this new set of transitions D and their pre- and post-sets can be inferred from the initial network topology, dispensing the user with the task of manually specifying them.

Formally, given a place $\pi \in \Pi$ it is possible to define

$$T_A^\pi = \{t \mid t \in T \wedge (t, \pi) \in A\} \subseteq T,$$

as the set of transitions that have a positive effect on π. Then, we represent T_A^π as a boolean formula in disjunctive normal form (DNF). Each conjunctive clause of the DNF defines a transition $t \in T_A^\pi$. More specifically, each conjunctive clause uses as variables the places in ${}^\bullet t$, and the places in ${}^\bullet \bar{t}$ as negated variables. For instance, the network in Figure 2.4a correspond to the formula $(\pi_r \wedge \pi_{r'} \wedge \ldots \wedge \overline{\pi_s} \wedge \overline{\pi_{s'}} \wedge \ldots)$. Thus, the boolean formula corresponding to a generic $T_A^\pi = \{t_1, t_2, \ldots, t_k\}$ is

$$\begin{aligned} \mathcal{B}(T_A^\pi) = &(\pi_{r_{t_1}} \wedge \pi_{r'_{t_1}} \wedge \ldots \wedge \overline{\pi_{s_{t_1}}} \wedge \overline{\pi_{s'_{t_1}}} \wedge \ldots) \\ &\vee (\pi_{r_{t_2}} \wedge \pi_{r'_{t_2}} \wedge \ldots \wedge \overline{\pi_{s_{t_2}}} \wedge \overline{\pi_{s'_{t_2}}} \wedge \ldots) \\ &\vee \ldots \vee (\pi_{r_{t_k}} \wedge \pi_{r'_{t_k}} \wedge \ldots \wedge \overline{\pi_{s_{t_k}}} \wedge \overline{\pi_{s'_{t_k}}} \wedge \ldots). \end{aligned} \tag{2.4}$$

Similarly, it is possible to reconstruct the network topology starting from a formula in DNF. A variable of the Formula 2.4 is *true* if the corresponding place is marked, *false* otherwise. Hence, if all preconditions of a transition are met, *i.e.* it is pre-enabled, then all the literals of the corresponding conjunctive clause, and the whole formula itself, evaluate as *true*. Therefore, for a generic place π, Formula 2.4 evaluates as *true* if there exists a pre-enabled transition that has a positive effect on π.

$\overline{\mathcal{B}(T_A^\pi)}$, the negation of Formula 2.4 in DNF, is *true* if at least one of its conjunctive clauses evaluates to *true*. By constructing D_π, the set of degradations of π, from the conjunctive clauses of $\overline{\mathcal{B}(T_A^\pi)}$, it is guaranteed that at least one of the transitions in D_π will be enabled if and only if *none* of the transitions with a positive effect on π is pre-enabled, satisfying the fourth assumption. For example, given the simple network of Figure 2.5a, $\mathcal{B}(T_A^{\pi_g})$ equals $\pi_a \wedge \pi_b \wedge \overline{\pi_c}$; therefore, $\overline{\mathcal{B}(T_A^{\pi_g})}$ is $\overline{\pi_a} \vee \overline{\pi_b} \vee \pi_c$. Figure 2.5b shows the degradation transitions built from the conjunctive clauses of $\overline{\mathcal{B}(T_A^{\pi_g})}$.

2.2.3 Petri net framework analysis

The analysis of the framework used in Section 2.2.1 is based on simulations (for more details see Chapter 3). In the case of gene regulatory networks, however, we focus on the attractors of the model state space. The state space (or marking graph) is a directed multigraph where each node identifies a marking, and each arc represents the occurrence of a transition. An attractor is a forward invariant subset of the state space, *i.e.* a set such that if a marking m belongs to it, then each markings reachable from m also belong to this set. In a biologically faithful model, each attractor should correspond to an observable biological steady state. Intuitively, as in a biological steady state the recently observed behaviour of the system will continue into the future, likewise, in the state space the model execution will keep cycling between the same states in the attractor set. Since we are particularly interested in steady states, we can abstract from time. Therefore, we can compute the state space using

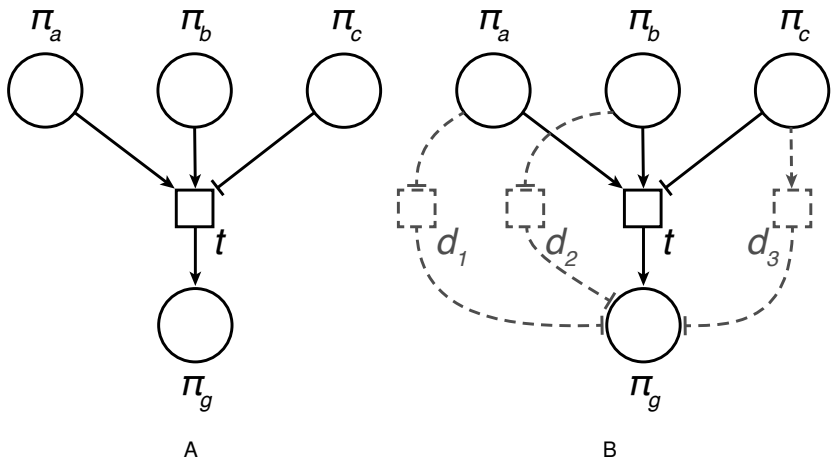


Figure 2.5: The network on the right (B) shows the set of transitions D (dashed lines) necessary to model the degradation of π_g. Notice that at least one transition in D is enabled if t is not pre-enabled. Since D can be automatically inferred from the gene interactions in the initial topology (A), it is safe and usually convenient to omit the transitions of D from the graphical representation to avoid cluttering.

Algorithm 2.3 computeStateSpace(B), where $B = \langle \Pi, T, F, A, I \rangle$

$E \leftarrow \emptyset$
$D \leftarrow$ computeDegradations(B)
$T \leftarrow T \cup D$
$V \leftarrow$ computeAllMarkings($|\mathcal{N}|^{\Pi}$) ▷ compute all possible bit vectors of size $|\mathcal{N}|^{\Pi}$
for all $v \in V$ **do**
    $\mathcal{E} \leftarrow$ computeAllEnabledTransitions(T, v)
    for all $e \in \mathcal{E}$ **do**
        $v' \leftarrow$ fireTransition(v, e) ▷ compute the marking obtained by firing e in marking v
        $E \leftarrow E \cup (v, v')$ ▷ add the new edge (v, v')
    end for
end for
return $G = \langle V, E \rangle$

a fully asynchronous semantics as showed in Algorithm 2.3. The algorithm works in two steps. First, it generates all vertices of the state space graph by computing the ensemble of all possible initial markings, *i.e.* all bit vector of size $|\mathcal{N}|^{\Pi}$. Secondly, for each bit vector the algorithm fires all enabled transitions one by one. For each transition fired, an arc from the current bit vector to the vector generated by the occurrence of the transition is added to the state space.

The strategy we choose to identify the attractors is to compute the terminal strongly connected components (TSCC) of the state space. TSCCs are a particular class of strongly connected components (SCC). The SCCs of a directed graph are its maximal strongly connected subgraphs, *i.e.* the induced subgraphs formed by the equivalence classes defined on the vertices by the relation of *mutual reachability*. Two vertices u and v are said to be mutually reachable if and only if there exist a path from u to v and from v to u. A TSCC T is a SCC such that if $u \in T$, then $v \in T$ for each directed arc (u, v). Thus, a TSCC is a SCC that does not have outgoing arcs to other SCCs. Therefore, by trapping the execution in a subset of states, each TSCC is an attractor of the the dynamical system. Given a graph $G = \langle V, E \rangle$, the TSCCs can be easily computed by the Tarjan algorithm (Tarjan, 1975) in $O(|V| + |E|)$. The complexity of this analysis lies in the generation of the state space. The size of the multigraph G is exponential in the number of places; $|V| = \mathcal{N}^{|\Pi|}$, where we recall that $\mathcal{N}$ is the maximal gene expression level. For our case studies, based on the boolean-like approach explained above, $|V|$ is 2^{11}, therefore, still tractable. Despite the complexity, building and exploring the whole state space instead of the single attractors may be valuable since the state space graph contains more information about the model dynamics. This information can lead to significative biological discoveries, as shown in Chapter 5. However, more efficient strategies should be devised to further extend the formalism to multiple gene expression levels. Some strategies to cope with state space graphs with millions of nodes are presented in Krepska (2012).

CHAPTER 3

EXECUTING MULTICELLULAR DIFFERENTIATION IN *Caenorhabditis elegans*

Published as:

Executing Multicellular Differentiation: Quantitative Predictive Modelling of C. elegans Vulval Development

Nicola Bonzanni[1 2] , Elzbieta Krepska[2] , K. Anton Feenstra[1 2] , Wan Fokkink[2] , Thilo Kielmann[2] , Henri Bal[2] , and Jaap Heringa[1 2] . Bioinformatics, 25(16):2049–2056 (2009).

[1]Centre for Integrative Bioinformatics, VU University Amsterdam, The Netherlands
[2]Department of Computer Science, VU University Amsterdam, The Netherlands

Abstract

Motivation: Understanding the processes involved in multi-cellular pattern formation is a central problem of developmental biology, hopefully leading to many new insights, *e.g.*, in the treatment of various diseases. Defining suitable computational techniques for development modelling, able to perform *in silico* simulation experiments, is an open and challenging problem.
Results: Previously, we proposed a coarse-grained, quantitative approach based on the basic Petri net formalism, to mimic the behaviour of the biological processes during multicellular differentiation. Here we apply our modelling approach to the well-studied process of *C. elegans* vulval development. We show that our model correctly reproduces a large set of *in vivo* experiments with statistical accuracy. It also generates gene expression time series in accordance with recent biological evidence. Finally, we modelled the role of microRNA mir-61 during vulval development and predict its contribution in stabilising cell pattern formation.

3.1 Introduction

Many efforts have been undertaken to elucidate how cells are able to coordinate different and sometimes conflicting signals, producing a precise phenotype during the animal organogenesis (Sternberg, 2005). *C. elegans* vulval development provides an elegant and relatively well-charted model to study how multiple pathways, in multiple cells, interact to produce developmental patterns.

The *C. elegans* hermaphrodite vulva develops from three of the six vulval precursor cells (VPCs), consecutively numbered P3.p to P8.p in Figure 3.1. Each VPC is competent to respond to intercellular signals, and is potentially able to adopt either of the three cell fates: 1°, 2°, or 3°. Each fate corresponds to a specific cell division pattern. The 1° and 2° fate cell lineages constitute the vulva, generating eight and seven progeny cells, respectively. The 3° fate lineage becomes a constituent of the hyp7 hypodermal syncytium, a large cell-like structure with many nuclei enveloping the developing nematode. In the wild-type hermaphrodite, the six VPCs adopt an invariant 3°-3°-2°-1°-2°-3° pattern (Sternberg & Horvitz, 1986), shown in Figure 3.1. This precise fate distribution is the result of the interplay between two competing signals: the spatially graded inductive signal produced by the anchor cell (AC), and the lateral signal originating from a presumptive 1° fate cell.

During this cell-cell interaction, the inductive epidermal growth-factor signal is produced by the AC and transported to the three nearest precursor cells. The signal is encoded by the protein LIN-3 and transduced by the receptor LET-23 into the Ras/MAPK pathway. This has the direct effect of up-regulating MPK-1, and of promoting 1° fate in P6.p. Further downstream the Ras/MAPK pathway, LIN-12 is down-regulated (Shaye & Greenwald, 2002) to suppress the promotion of 2° fate, while production of the lateral signal is stimulated (Chen & Greenwald, 2004). This signal promotes 2° fate in the neighbouring cells P5.p and P7.p (Sundaram, 2004), and inhibits Ras signalling

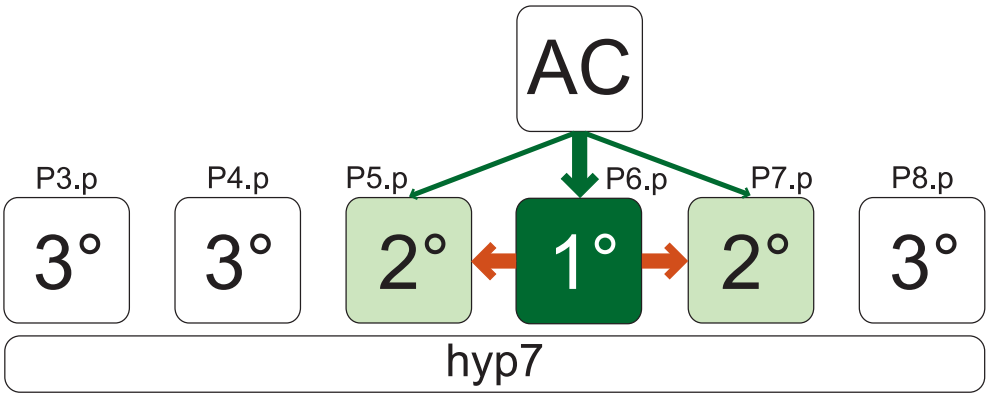


Figure 3.1: Vulval development in the wild-type *C. elegans*, showing the AC, the VPCs (*P3.p*-*P8.p*), and the hyp7. The inductive signal from the AC promotes the 1° fate in P6.p, and stimulates the production of the lateral signal near the flanking cells, promoting the 2° fate in P5.p and P7.p. The 3° fate lineage becomes a constituent of the hyp7.

to block transduction of the inductive signal through the Ras/MAPK pathway (Yoo *et al.,* 2004). This negative feedback helps maximise LIN-12 activity in the presumptive 2° fate cells (Yoo & Greenwald, 2005).

The first diagrammatic model, describing the regulatory network underlying VPC determination, was proposed by Sternberg & Horvitz, 1989. Since then, global understanding of the biological network has improved greatly. The first computational model, proposed by Kam *et al.,* 2003, combined multiple experimental "scenarios" from Sternberg & Horvitz, 1986 into a single model, using Live Sequence Charts (LSCs). Afterwards, in two landmark papers, Fisher *et al.,* 2005, 2007 suggested two state-based mechanistic models. The first (Fisher *et al.,* 2005) used statecharts to represent internal states of components, and LSCs to execute actions between them. They formalised Sternberg's model (Sternberg & Horvitz, 1989) but did not incorporate any additional data. A more recent approach (Fisher *et al.,* 2007) was based on Reactive Modules, with modelling principles akin to the previous paper. In contrast to the model presented in the current paper, the three listed models build on representing rules that the system adheres to, rather than modelling the underlying biological processes. Two other insightful models of *C. elegans* vulval development have been published. Giurumescu *et al.,* 2006 proposed a partial model based on ODEs, while Sun & Hong, 2007 developed a model based on automatically learned dynamic Bayesian networks with discrete states. Independent from us, Li *et al.,* 2009 recently modeled part of *C. elegans* vulval development using hybrid functional Petri nets with extensions. While they focused on model validation, we additionally generated new insightful predictions.

In this paper, we apply our approach (Krepska *et al.,* 2008), which is discrete, nondeterministic, and based on Petri nets to *C. elegans* vulval development. Petri nets

are a convenient formalism to represent biological networks. This formalism models process synchronisation, asynchronous events, conflicts, and in general concurrent systems in a natural way. Moreover Petri nets offer direct insights into causal relationships, and allow a graphical visualisation that resembles the diagrams used to describe biological knowledge. The reader may find recent survey papers concerning modelling of biological systems with Petri nets in Chaouiya, 2007; Koch & Heiner, 2008; Matsuno *et al.,* 2006; Peleg *et al.,* 2005.

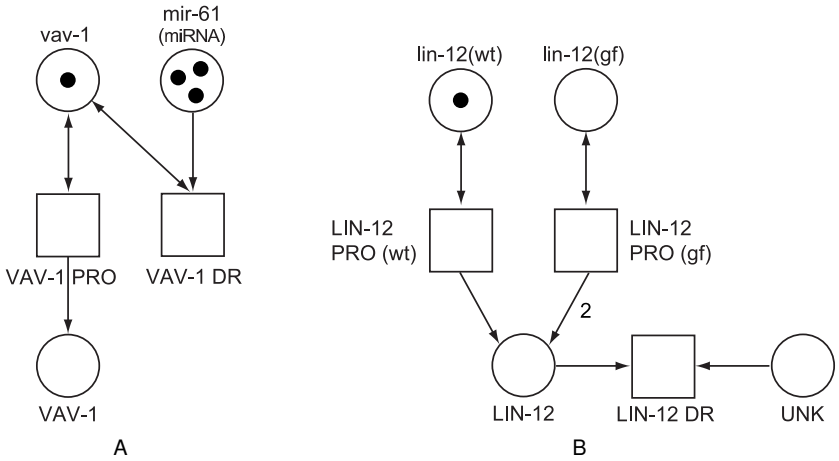


Figure 3.2: (a) The presence of the microRNA mir-61 down-regulates VAV-1, by enabling the transition VAV-1 DR, which is in conflict with VAV-1 PRO. (b) Two example modules, gene production (*vertical*) and endocytosis (*horizontal*), interact in the Notch/LIN-12 pathway.

Several adaptations of the Petri net formalism have been introduced in the context of modelling biological systems. On the one hand, qualitative Petri nets (Gilbert *et al.,* 2007) can be used for structural and invariant analysis, but they greatly abstract from the biological system. On the other hand, Stochastic Petri nets (Goss & Peccoud, 1998) incorporate kinetic constants, but these are mostly unknown or approximate. Hybrid Petri nets (Matsuno *et al.,* 2000) and their extensions on which Cell Illustrator (Matsuno *et al.,* 2006) is based, are rich and expressive, but model understanding and causal backtracking are impeded by the complexity of the formalism.

In our model we have chosen to preserve the simplicity of the original Petri net formalism. Our modelling approach is aimed to mimic the underlying biological mechanisms as much as possible, and not only to reproduce the expected phenotype according to a specific set of mutations. To achieve this, we apply a principle of maximal parallelism (Burhard, 1983), and bounded execution with overshooting (Krepska *et al.,* 2008). Using this simple framework, we identified different modules, each corresponding to different biological functions. Thus, combining functional modules into cells, and joining such cells together, we iteratively developed the whole network. Unlike the aforementioned works on formal modelling of *C. elegans* vulval development,

the ability of our model to capture biological functions into small building blocks allows these to be reused in new case studies on multi-cellular signalling and regulation modelling.

We show that our model, encoding biological hypotheses from the literature (Shaye & Greenwald, 2002; Yoo & Greenwald, 2005), is able to reproduce *in silico* a set of *in vivo* experiments, providing the necessary statistical data to establish a more detailed comparison with biological observations than was previously possible. To the best of our knowledge, we are the first to model microRNA interactions during *C. elegans* vulval development. Furthermore, we predict a possible "tuning" role played by the *mir-61* microRNA gene, in ensuring stability of the fate pattern.

3.2 Methods

3.2.1 Model construction

We developed an executable Petri net model for cell fate determination during *C. elegans* vulval induction (see Chapter 2). This large network can be visualised on the web page of our project (`http://http://www.cs.vu.nl/concell`); a schematic representation is given in Figure 3.5. The entire network comprises 600 nodes (places and transitions) and 1000 arcs. Nevertheless the simplicity of the formalism, and its graphical representation, helps us to identify different modules. These correspond to different biological functions, such as gene expression, protein activation, and protein degradation. It is possible to reuse modules corresponding to a function, like small building blocks, to compose more complex modules, and eventually build a full cell. The cell itself is a module that can be reused. Applying these principles, we have built the VPC network out of six interconnected cells as identical modules of a multi-potent cell. We also built a separate block for the AC (producing the inductive signal) and for the hyp7. The possibility to divide the entire graph into simple, small, and meaningful modules has three main advantages: (i) the modelling process becomes easier, (ii) the resulting network is homogeneous, and (iii) modules (at different levels) can be reused throughout the model, or for modelling other organisms.

Figure 3.2 shows selected examples of how to represent biological modules as a Petri net. Figure 3.2a illustrates VAV-1 down-regulation by decreasing the translation rate of the gene vav-1. In fact, if mir-61 is not present, the reaction VAV-1 PRO is enabled and produces the protein. However, when mir-61 is present, the reaction VAV-1 DR is enabled and has 0.5 chance of firing compared to VAV-1 PRO, thus the production of VAV-1 will halve. Figure 3.2b depicts two connected basic modules, a gene expression and the endocytosis mediated down-regulation of LIN-12. In this example, activation of the Ras/MAPK cascade leads to the transcription of a hitherto unknown gene that enhances the LIN-12 endocytosis, as hypothesised by Shaye & Greenwald, 2002. Note that here the produced LIN-12 is removed, while in Figure 3.2a the gene production was reduced. An alternative way to represent down-regulation using transitions has been proposed by Grunwald *et al.,* 2008.

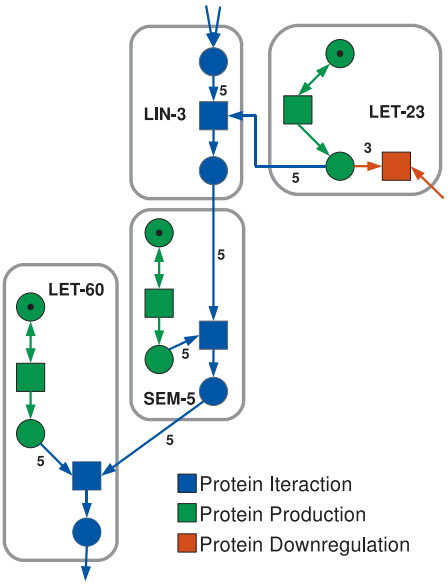


Figure 3.3: Basic biological functions used in more complex modules.

At the system level, a module can be viewed as a "meta-transition", a Petri net with specified inputs (places which can receive tokens) and outputs (arcs outgoing of the module). Instead of constructing the Petri net model using boolean functions as basic network components (Sackmann *et al.,* 2006) and subsequently check the biological meaning of each subnetwork extracted by invariant analysis (Grunwald *et al.,* 2008; Sackmann *et al.,* 2006), we focused on using basic biological functions as network building blocks. In our experience, the procedure of building a modular biological Petri net can be split into five phases:

Level 1: Basic biological functions. We created six basic modules representing the basic biological functions used to encode the relations described in the literature related to *C. elegans* vulva development: protein production, protein activation, down-regulation, up-regulation, signalling and constitutive degradation.

Level 2: Protein interactions. Combining basic modules, we built more complex blocks, each modelling the interactions of one protein. The division into protein interaction modules is presented in Table 3.1. Figure 3.3 shows an example of how basic biological functions are combined to build protein interaction modules.

Level 3: Pathways. In Figure 3.4, modules LIN-3, LET-23, SEM-5, LET-60, MPK-1, and DSL constitute the Ras/MAPK pathway, and modules LIN-12, VAV-1, MIR-61, DPY-23, LST constitute the competing Notch/LIN-12 pathway.

Level 4: Cells. Figure 3.4 presents the Petri net model of a single VPC cell with four links to the environment.

Level 5: Multi-cellular interactions. In Figure 3.5 we show how the six VPCs, the AC and the hyp7 modules are connected. Adjacent cells are linked with each other, the hyp7 connects to all six cells, and the AC can directly influence cells P5.p, P6.p, and P7.p.

Table 3.1: Description of modules constituting the model of *C. elegans* vulval development depicted in Figure 3.4 and Figure 3.5.

Module	Function
SEM-5	Production and activation of SEM-5 from gene sem-5(wt).
LET-60	Production and activation of LET-60 from gene let-60(wt).
LIN-3	Reception of LIN-3 from AC and hyp7.
LET-23	Production and activation of LET-23 from gene let-23(wt). Down-regulation of LET-23 by DPY-23.
LST	Production and activation of LSTs from lst-1(wt), lst-2(wt), and lst-4(wt) genes. Down-regulation of LSTs by MPK-1. Up-regulation of LSTs promoted by LIN-12*
MPK-1	Production and activation of MPK-1 from mpk-1(wt) gene. Down-regulation of MPK-1 by LST.
DSL	Production of DSL signal.
LATERAL	Transport of lateral signal (DSL) to adjacent cells.
DPY-23	Production of DPY-23 from dpy-23(wt) gene, promoted by LIN-12*.
LIN-12	Production of LIN-12 from lin-12(wt) gene. Activation of LIN-12 by binding to DSL. Endocytotic down-regulation of LIN-12 mediated by VAV-1 and promoted by the Ras/MAPK pathway.
VAV-1	Production of VAV-1 from vav-1(wt) gene. Down-regulation of VAV-1 by microRNA mir-61.
MIR-61	Production of miR-61 microRNA.
AC	Production of LIN-3 and diffusion in a graded fashion to P6.p and the two adjacent cells P5.p and P7.p.
hyp7	Production of LIN-3 and diffusion to all VPCs.
Not shown	Constitutive degradation of various proteins.

* Protein names followed by a star (*) stand for the active proteins

Figure 3.3 highlights the top-left portion of the VPC model depicted in Figure 3.4. One can see how basic biological functions are reused in different protein interaction modules, where the links describe the interactions between different modules. For instance, in Figure 3.3 the LET-23 module is connected to LIN-3, which is connected to SEM-5, which in turn interacts with LET-60. The biological mechanisms underlying these interactions are found in the literature and encoded by the basic biological functions mentioned. The network shown in Figure 3.3 models the first steps during signal

transduction within the Ras/MAPK cascade, where the transmembrane receptor LET-23 is activated by the ligand LIN-3. The resulting activated complex then activates the core Ras protein LET-60, by signalling through SEM-5.

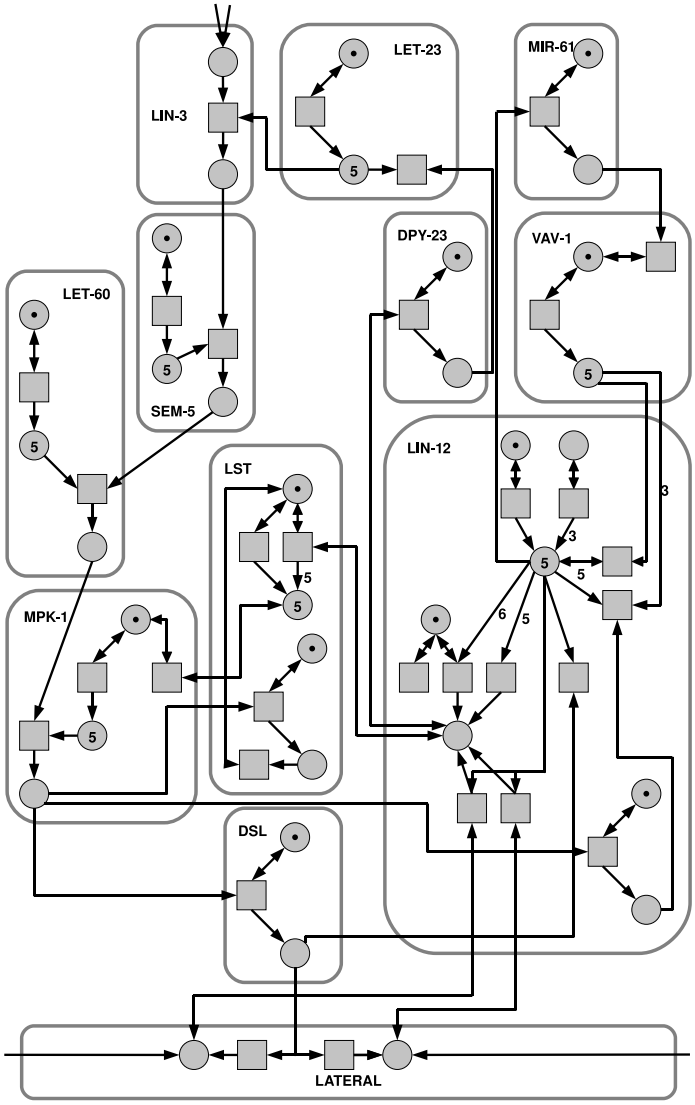


Figure 3.4: Schematic representation of a VPC in our Petri net model. Each rounded box is a module. Note that LIN-3 and LATERAL are connected to the environment.

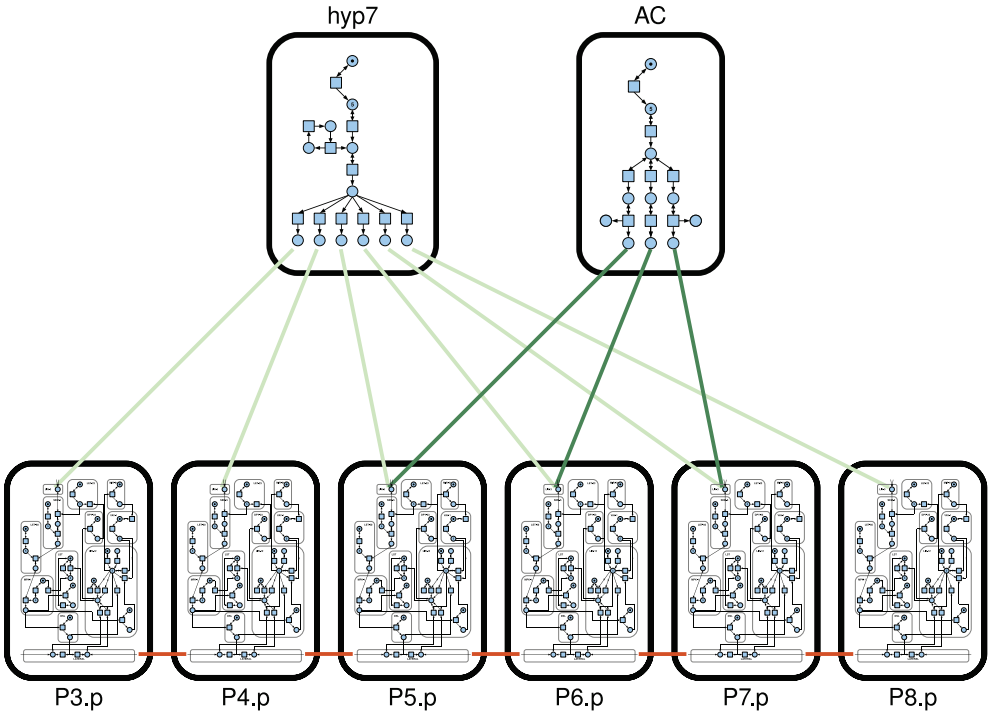


Figure 3.5: Schematic representation of the whole system. The VPCs are connected with the AC, the hyp7, and their adjacent cells.

3.2.2 Modelling genetic perturbations

For each genetic background, each gene can be in wild-type form (*wt*, *i.e.* the most common form of a gene as it occurs in nature) or in one of the following mutated forms: loss-of-function (*lf*, *e.g.* the gene is deleted or dysfunctional) or gain-of-function (*gf*, *e.g.* the gene transcription is over-stimulated). It is possible to derive an initial configuration corresponding to a given genetic perturbation placing a token in one of the two different places used to represent gain-of-function and wild-type for each gene in the genetic background. Loss-of-function mutation is represented by token removal. It is therefore possible to initiate the network in an appropriate initial configuration by simply placing or removing tokens in opportune places.

Figure 3.6a depicts an example of a typical gene transcription. The transition LIN-12 PRO(wt) produces LIN-12 proteins when the wild-type gene lin-12(wt) is present. When the gene is not present (*i.e.* lin-12(wt) holds no token), the event does not take place. Figure 3.6b and Figure 3.6c depict two different genetic backgrounds, corresponding respectively to the loss-of-function and gain-of-function of the lin-12 gene.

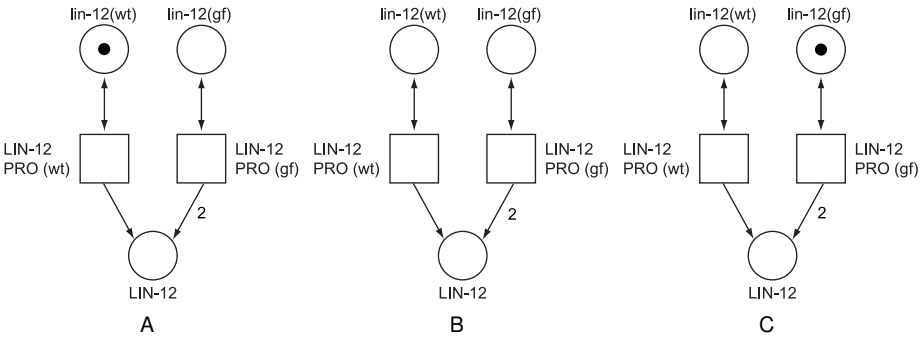


Figure 3.6: Gene expression with different initial conditions, corresponding to different genetic backgrounds. (a) corresponds to lin12 wild-type, (b) corresponds to the *lin12(lf)* mutant, and (c) corresponds to the *lin12(gf)* mutant. Note that arcs with two heads represent two arcs, one either way.

3.2.3 Model calibration

We started off with assuming that initially proteins are expressed at low basal levels and reactions require high protein concentration levels. Therefore, we set the initial concentration levels for proteins to zero and we assigned high requirements and low level production to all transitions, respectively arc weights five and one (see the Supplementary information for an example).

We subsequently simulated the 22 *in vivo* experiments in our calibration set (Table 3.2). We identified mismatches between the simulation results and the expected phenotypes, and back-tracked the problem (*e.g.* an overly strong or weak down-regulation) following the causal chain from one module to the other (*i.e.* from products, to transitions, up to their requirements). For selected modules, arc weights and occasionally initial protein concentration levels were fine-tuned to recover the expected behaviour.

This manual calibration process iteratively converged upon a stable and fixed set of parameters that we used for all further simulations. During the process we noticed that only in very few cases single parameter adjustments were able to sensibly change the simulation results, whereas more often combinations of parameters were changed to approach the expected behaviour. This suggests a “spectrum of sensitivities” as discussed in Gutenkunst *et al.,* 2007 that should allow the modellers to focus on predictions rather than on parameters.

3.2.4 Simulation procedure

In experimental biology, experimental replicates are necessary to overcome the variability intrinsic to biological systems. In our modelling approach, which is non-deterministic, we interpret the outcome of a simulation run as the phenotype of an

individual worm. Thus, to reproduce a worm population, we performed 5000 simulation runs for each genetic background with different random seeds, each for 1000 maximally parallel steps.

Based on the current experimental knowledge (Shaye & Greenwald, 2002, 2005), we determine the fate adopted by each cell by measuring and correlating the concentration levels of MPK-1* and LIN-12*. Specifically, 1° fate is induced by a high level of MPK-1* and is refractory to LIN-12*. 2° fate is induced by a low level of MPK-1* and a high level of LIN-12*. Low levels of both MPK-1* and LIN-12* lead to 3° fate. From the simulation we calculate the LIN-12* and MPK-1* concentration levels as the average number of tokens over the final 50 steps to avoid unnecessary noise from the continual movement of tokens. Predicted cell fates are not influenced significantly by the length of the averaging because the protein concentration levels at the very end of the simulation are generally in a steady state. Assuming that a *high* concentration level corresponds to more than three tokens in a place and a *low* level corresponds three tokens or less, it is possible to determine the adopted fate. The corresponding piecewise function is formalized in the Supplementary information.

To facilitate parameter adjustments during calibration we also implemented three scoring functions (one for each cell fate) as sigmoids, in order to obtain a continuous curve instead of the discrete and discontinuous profile of a piecewise function. Such a continuous score was very useful during calibration to guide recovery of the expected behaviour by comparing slight changes in the scores produced by different adjustments.

Each scoring function rewards (*i.e.* score tends to 1) concentration levels that match the corresponding description. In the scoring functions, more than four tokens corresponds to *high* and less than two corresponds to *low*, while intermediate numbers of tokens (in between 4 and 2) produce the S-shaped gradient peculiar to sigmoids. In our experience slight changes in the shape of the functions (*e.g.* steepness) do not significantly change the results. Consequently, each scoring function, using the simulated LIN-12* and MPK-1* concentration levels as variables, computes a score in the interval $[0, 1]$ that measures how closely a simulated cell reproduces the fate description captured by the scoring function. For each cell we calculate three scores (one for each function), and assign to the cell the fate corresponding to the function that returns the highest score. The analytic form of these functions can be found in the Supplementary information.

The intersection of the three scoring functions generates a landscape (Figure 3.7) that can be compared to the discrete representation of the piecewise function and resembles the fate plane proposed by Giurumescu *et al.,* 2006, in which the quadrants identify cell fates.

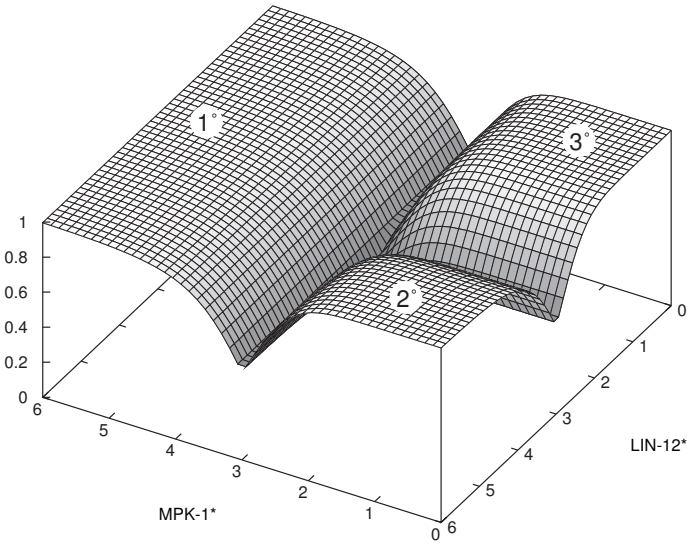


Figure 3.7: Landscape produced intersecting the three scoring functions. The plateaus are labelled with the corresponding cell fates.

3.3 Results

3.3.1 Model validation

To determine the capability of our model to reproduce and predict the biological behaviour, we simulated 64 different experimental conditions.Twenty-two experiments (Table 3.2) previously selected in Fisher *et al.,* 2005 were used for model calibration. Thirty perturbations were used for validation: 26 (see the Supplementary information) from Fisher *et al.,* 2005, three (Table 3.4) from Sternberg, 2005, and one (Exp. 52, Table 3.5) from Yoo & Greenwald, 2005. Particularly, experiment 51 (Table 3.4) was never simulated in any previous work that we are aware of. The remaining twelve simulations constitute new predictions. Of these, the most remarkable (Table 3.5) are discussed in Section 3.3.2. Statistical details for all simulations and a short animation displaying a typical single run are available in the Supplementary information.

Our model reliably reproduces all the mutant combinations, except for the double mutant *lin-12(gf);lin-15(lf)* (Table 3.2, Exp 21 and 45), even if in these cases, a fraction of the predictions matches the expected pattern. The noticeable differences of biological observations from different labs, and the few worms examined *in vivo*, do not help to establish a trustworthy expected outcome.

Of the 22 experiments in Table 3.2, particularly interesting are the experimental conditions that lead to unstable fate patterns. These results were already discussed in Fisher *et al.,* 2007 and Sun & Hong, 2007, but these discussions lacked statistical detail about the possible outcomes. In fact, Sun & Hong, 2007 observed that the

Table 3.2: *in vivo* experiments selected in Fisher *et al.,* 2005, and used by us for model calibration. In the *AC* column, – stands for no AC, while + means that the AC is present. In the *Genotype* column, for each gene a loss of function (*lf* or knock-out) or gain of function (*gf* or overexpression) mutation is indicated. *lst* is the group of lst-1, lst-2, lst-3, lst-4, dpy-23. *Vul* is the group of let-23, sem-5, let-60, mpk-1. In the *Fate Pattern* column, 1 indicates 1° cell fate, 2, 2° fate, 3, 3° fate, and 1\2 either 1° or 2° fate.

Exp.	AC	Genotype				Fate Pattern					
		lst	*Vul*	*lin-15*	*lin-12*	*P3.p*	*P4.p*	*P5.p*	*P6.p*	*P7.p*	*P8.p*
1[a]	+					3	3	2	1	2	3
2[b]	+	lf				3	3	1	1	1	3
3[c]	+		lf			3	3	3	3	3	3
5[c]	+			lf		1\2	1\2	2	1	2	1\2
6[d]	+	lf		lf		1	1	1	1	1	1
7[e]	+		lf	lf		3	3	3	3	3	3
9[f]	+				lf	3	3	1	1	1	3
10[g]	+	lf			lf	3	3	1	1	1	3
11[c]	+		lf		lf	3	3	3	3	3	3
13[c]	+			lf	lf	1	1	1	1	1	1
17[c]	+				gf	2	2	2	1	2	2
19[c]	+		lf		gf	2	2	2	2	2	2
21[c]	+			lf	gf	1\2	1\2	2	1	2	1\2
25[h]	–					3	3	3	3	3	3
26[d]	–	lf				3	3	3	3	3	3
29[c]	–			lf		1\2	1\2	1\2	1\2	1\2	1\2
33[c]	–				lf	3	3	3	3	3	3
37[c]	–			lf	lf	1	1	1	1	1	1
41[c]	–				gf	2	2	2	2	2	2
42[d]	–	lf			gf	2	2	2	2	2	2
43[c]	–		lf		gf	2	2	2	2	2	2
45[c]	–			lf	gf	1\2	1\2	1\2	1\2	1\2	1\2

[a] Sulston & Horvitz, 1977b
[b] Berset *et al.,* 2005; Yoo *et al.,* 2004
[c] Sternberg & Horvitz, 1989
[d] Berset *et al.,* 2001
[e] Cui *et al.,* 2006; Sternberg & Horvitz, 1989; Sulston & Horvitz, 1977a
[f] Greenwald *et al.,* 1983; Sternberg & Horvitz, 1989
[g] Berset and Hajnal, unpublished data
[h] Kimble, 1981

Table 3.3: Detailed statistical results for the 5000 simulation of *in vivo* experiment 5, Table 3.2 (*lin-15(lf)*).

Exp.	Fate Pattern *P3.p P4.p P5.p P6.p P7.p P8.p*						Occurences	Percentage
5	*Combinations matching the commonly observed pattern:*							*86.2%*
	1	2	2	1	2	1	1348	27.0%
	2	1	2	1	2	1	1180	23.6%
	2	1	2	1	2	2	946	19.0%
	1	2	2	1	2	2	830	16.6%
	Three or more adjacent 2° fate cells:							*4.5%*
	2	2	2	1	2	1	132	2.6%
	2	2	2	1	2	2	93	1.9%
	Two adjacent 1° fate cells:							*2.7%*
	1	1	2	1	2	1	88	1.8%
	1	1	2	1	2	2	46	0.9%

Table 3.4: *in vivo* experiments not used for the model construction.

Exp.	AC	Genotype		Fate Pattern					
		let-60	*lin-3*	*P3.p*	*P4.p*	*P5.p*	*P6.p*	*P7.p*	*P8.p*
49[i]	+	lf		3	3	3	3	3	3
50[j]	+		lf	3	3	3	3	3	3
51[j]	+		gf	3	2	1	1	1	2

[i] Beitel *et al.,* 1990
[j] Sternberg, 2005

statecharts model of Fisher *et al.,* 2005 often produces two adjacent 1° fate cells, which they claim is rarely observed in experiments, but they also do not provide supplementary statistical details.

In Table 3.3 we provide statistical details for experiment 5 from Table 3.2. More than 93.4% of the predicted patterns match one of the expected biological 1\2°-1\2°-2°-1°-2°-1\2° combinations. Of all matching patterns, only 4.5% contain three or more adjacent 2° fate cells, while just 2.7% have two or more adjacent 1° fate cells. These quantities correspond to the biological evidence that in these experiments three adjacent 2° fate, or two adjacent 1° fate cells are very unlikely. In the remaining 6.8% (not included in Table 3.3) one or more cells adopt 3° fate, and we interpret these outcomes as the “rare phenotypes” in which uninterpretable lineages are observed (*i.e.* in between 2° and 3°), as noted for instance in Sternberg & Horvitz, 1989.

In our approach, each maximally parallel step corresponds to a time step in the

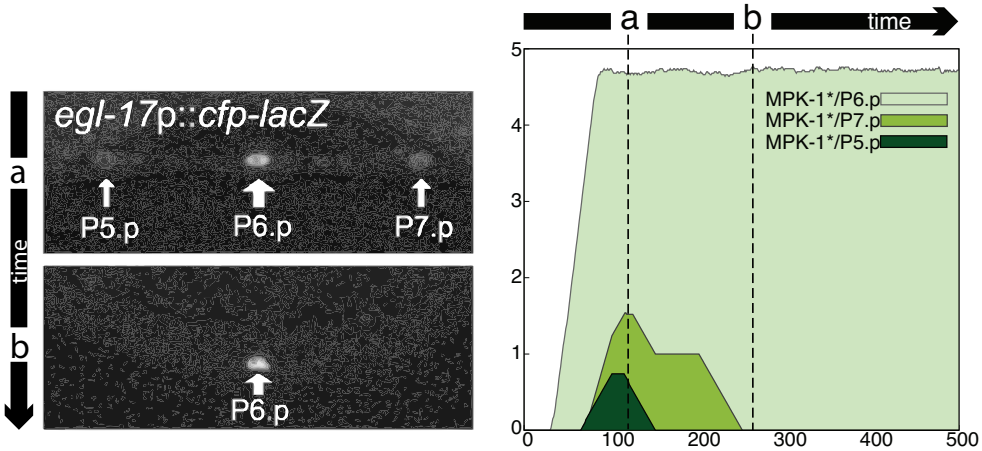


Figure 3.8: Comparison between photomicrographs of gene activity by fluorescently labelled gene products, and simulation results. (a) Photomicrographs of the graded expression of the inductive signal adapted from Yoo *et al.,* 2004, Science Magazine. © 2004, AAAS. (b) Time series plot generated by our model, showing the graded expression of the inductive signal, initially faintly present in P5.p and P6.p. A running average over 50 steps is used for clarity of presentation. Concentration levels are on the vertical axis while maximally-parallel steps on the horizontal. One can correlate photomicrographs *a* and *b* with point *a* and *b* in the time series.

ontogeny of the biological system. Thus our simulations can also be interpreted as time courses of gene regulation in vulval development. In Figure 3.8, the gene expression time series generated by our model are compared with the fluorescent photomicrographs published by Yoo *et al.,* 2004. They show evidence of the graded expression of the egl-17p::cfp-lacZ reporter that responds to the Ras/MAPK pathway. Figure 3.8b depicts the time series generated by our model from the simulation results of a wild-type animal. Initially MPK-1* (downstream product of the Ras/MAPK pathway as EGL-17) is faintly expressed in P5.p and P7.p. Subsequently, expression in P5.p and P7.p disappears, and MPK-1* remains at a high level only in P6.p, in accordance with the fluorescent photomicrographs of Figure 3.8a. We note that the concentration levels at the end of the simulation are approximately constant, indicating a steady state. In a related experiment, Yoo *et al.,* 2004 divided lst genes into two groups: pattern A which contains dpy-23 and lst-3, and pattern B to which lst-1, lst-2 and lst-4 belong. Each group has its own characteristic temporal expression pattern that corresponds closely to the time series generated by our simulation (see the Supplementary information).

3.3.2 mir-61: developmental switch and modulator

Our computational model, besides reproducing well-known biological experiments, encodes and unifies different published hypotheses and conjectures, shedding light on the vulval development process. The two hypotheses described next are related to LIN-12 down-regulation, which is essential during vulval organogenesis (Shaye & Greenwald, 2002; Yoo *et al.,* 2004), and link the microRNA mir-61 to the vulva development process.

Shaye & Greenwald, 2002 propose that, besides the degree of constitutive internalisation displayed by LIN-12, Ras activation leads to transcription of an unknown factor that enhances the rate of internalisation, promoting the endocytic routing of LIN-12. In Figure 3.9 one can see how we captured this hypothesis in our model. Activation of Ras enables the transcription of the unknown gene, which down-regulates LIN-12 post-translationally. Notably, changing the model of LIN-12 down-regulation from post- to pre-translation disrupts this behaviour and significantly alters our results.

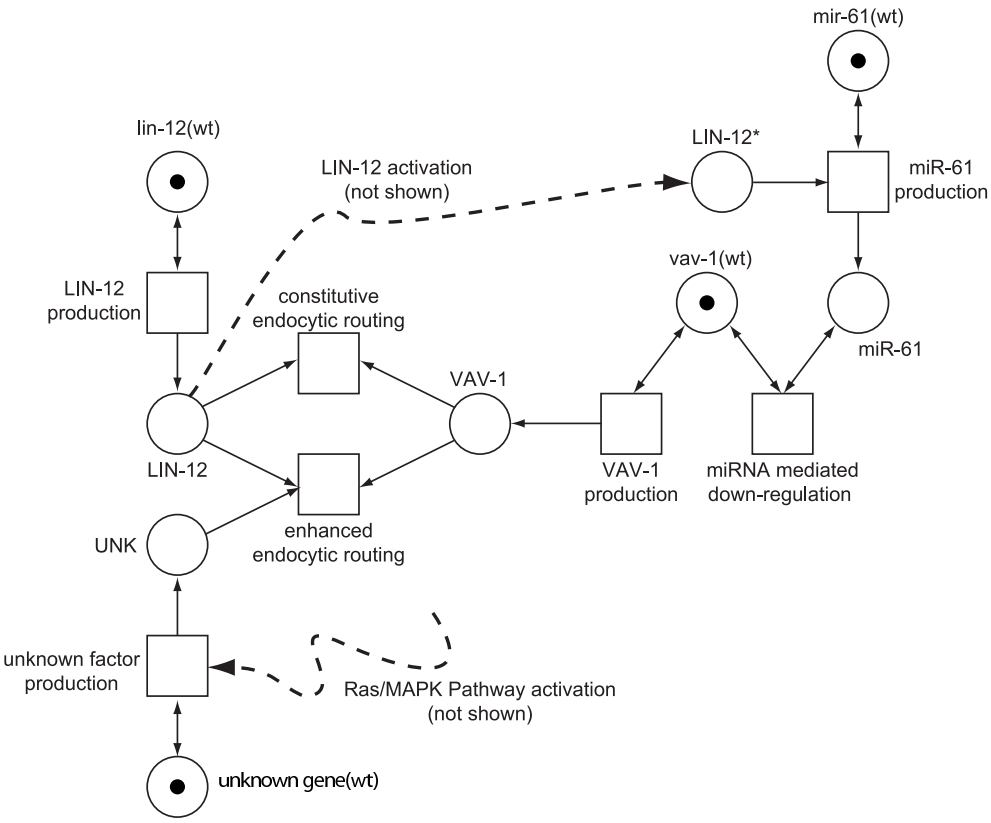


Figure 3.9: Single model capturing different biological suggestions as explained in Section 3.3.2.

Table 3.5: Selection of microRNA experiment outcomes predicted by our model. *mir-61(ce)* stands for constitutive expression of mir-61.

Exp.	AC	Genotype			Fate Pattern					
		mir-61	*Vul*	*lst*	*P3.p*	*P4.p*	*P5.p*	*P6.p*	*P7.p*	*P8.p*
52[k]	+	ce			2	2	2	2	2	2
53	–	ce			2	2	2	2	2	2
54	+	ce	lf		2	2	2	2	2	2
55	+	ce		lf	2	2	1	1	1	2
56	+	lf		lf	3	2	1	1	1	2

[k] Yoo & Greenwald, 2005

Table 3.6: Detailed statistics for the simulation of experiment 2 (*lst(lf)*) Table 3.2 and 56 (*mir-61(lf);lst(lf)*) Table 3.5. Outcomes below 0.1% are omitted.

Exp.	Fate Pattern						Occurences	Percentage
	P3.p	*P4.p*	*P5.p*	*P6.p*	*P7.p*	*P8.p*		
2	3	3	1	1	1	3	4800	96.0%
	3	3	1	1	1	2	199	4.0%
56	3	2	1	1	1	2	1594	31.9%
	3	3	2	1	1	2	1399	28.0%
	3	3	2	1	2	3	1000	20.0%
	3	2	1	1	2	3	998	20.0%

Yoo & Greenwald, 2005 identified mir-61 as direct transcriptional target of the LIN-12/Notch pathway. The gene mir-61 encodes a microRNA which blocks expression of the mRNA encoding VAV-1, a protein involved in LIN-12 down-regulation, possibly promoting LIN-12 endocytosis. They therefore proposed that activation of mir-61 by LIN-12 and the consequent down-regulation of VAV-1 constitute a positive-feedback loop that promotes LIN-12 activity in presumptive 2° fate VPCs. Although the unknown factor conjectured by Shaye and Greenwald does not seem to be required for the initial internalisation of LIN-12, VAV-1 is necessary for the constitutive internalisation of LIN-12. Notice that VAV-1 is invfolved in both constitutive and enhanced post-translation (endocytosis mediated) down-regulation of LIN-12.

Modelling these hypotheses (Figure 3.9) and capturing their behaviour has proven to be necessary to obtain the expected results during *in silico* experiments. Moreover, we simulated several perturbations of the mir-61 microRNA gene, obtaining the outcomes shown in Table 3.5. This nicely confirms the role of the positive-feedback loop proposed by Yoo & Greenwald, 2005. All experiments of Table 3.5, as far as we know, have not been tested *in vivo* (with the exception of experiment 52, which is described in Yoo & Greenwald, 2005).

Experiments 52, 53, 54, and 55 confirm the specific role of mir-61 in influencing the

cell fate decision, as determined by Yoo and Greenwald. Experiment 56 suggests a possible secondary role. This is a double mutant *mir-61(lf);lst(lf)* variation of the *lst(lf)* experiment 2, Table 3.2. Although the single mutant *lst(lf)* expresses a stable VPC fate pattern, the loss-of-function of mir-61 in the double mutant disrupts the stability of the pattern, as can be seen in the statistical breakdown of Table 3.6. Based on this observation, we suggest that besides acting as developmental switch, mir-61 plays a "tuning" role (Karp & Ambros, 2005) to ensure the stability of the cell fate pattern formation.

To the best of our knowledge, we are the first to model *in silico* microRNA interactions during *C. elegans* vulval induction, supporting the conjecture formulated in Yoo & Greenwald, 2005 that lin-12, mir-61, and vav-1 form a feedback loop that helps maximise lin-12 activity in the presumptive 2° VPCs.

3.4 Discussion

Modelling and analysing developmental processes is a challenging task, as these biological processes often encompass several cells and evolve over the course of several hours. Moreover, the current lack of precise quantitative parameters at molecular level and the descriptive form of this biological knowledge welcome research on different modelling approaches able to reach the sweet spot in between abstraction and biological significance. In the work presented here, we abstracted the descriptive knowledge into a simple formal model that suitably mimics the underlying biological mechanisms and retains an adequate predictive power.

The Petri net used in our approach has a rather simple formalism, but the network designed by us is fairly large. Although several tools able to build extensive Petri nets with modular support exist (Peccoud *et al.,* 2007; Ratzer *et al.,* 2003), they are often quite complex in order to support much richer formalisms than the one we used, or they do not scale to the size of our Petri net model. Furthermore, the lack of a Petri net tool with a robust and efficient implementation of the maximal parallel execution semantics led us to build our own simulation tool (available on the web page of our project).

In conclusion, we applied our Petri net approach to *C. elegans* vulval development, reproducing several *in vivo* experiments. We generated insightful and testable predictions involving the microRNA mir-61. Our model is a suitable but partial representation of the whole intricate developmental process that leads to the formation of the *C. elegans* vulva. New understanding of the process, supported by further experimental analysis, can be conveniently integrated in our model taking advantage of its modular fashion.

3.5 Acknowledgements

Funding: This work was supported in part by ENFIN, a Network of Excellence funded by the European Commission within its FP6 Program, under the thematic area "Life

Sciences, genomics and biotechnology for health", contract number LSHG-CT-2005-518254.

Chapter 4

Formalising nutrient starvation response in *Saccharomyces cerevisiae*

Published as:

The role of proteosome-mediated proteolysis in modulating potentially harmful transcription factor activity in Saccharomyces cerevisiae

Nicola Bonzanni[1 2] , Nianshu Zhang[3] , Stephen G. Oliver[3] , and Jasmin Fisher[4]

Bioinformatics [ISMB/ECCB] 27(13): 283–287 (2011)

[1]Centre for Integrative Bioinformatics, VU University Amsterdam, The Netherlands
[2]Department of Computer Science, VU University Amsterdam, The Netherlands
[3]Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, UK
[4]Microsoft Research Cambridge, UK

Abstract

Motivation: The appropriate modulation of the stress response to variable environmental conditions is necessary to maintain sustained viability in *Saccharomyces cerevisiae*. Particularly, controlling the abundance of proteins that may have detrimental effects on cell growth is crucial for rapid recovery from stress-induced quiescence.
Results: Prompted by qualitative modeling of the nutrient starvation response in yeast, we investigated *in vivo* the effect of proteolysis after nutrient starvation showing that, for the Gis1 transcription factor at least, proteasome-mediated control is crucial for a rapid return to growth. Additional bioinformatics analyses show that potentially toxic transcriptional regulators have a significantly lower protein half-life, a higher fraction of unstructured regions, and more potential PEST motifs than the non-detrimental ones. Furthermore, inhibiting proteasome activity tends to increase the expression of genes induced during the Environmental Stress Response more than those in the rest of the genome. Our combined results suggest that proteasome-mediated proteolysis of potentially toxic transcription factors tightly modulates the stress response in yeast.

4.1 Introduction

A prompt and appropriate response to abrupt fluctuations in external conditions is crucial to survive stressful environmental changes, especially in unicellular organisms such as the yeast *Saccharomyces cerevisiae*. During nutrient starvation, in order to ensure extended survival, *S. cerevisiae* cells exit the cell cycle at G_1 and enter the quiescent state (called G_0), but rapidly resume growth and proliferation when nutrient conditions turn favourable. Two conserved signaling pathways Ras/cAMP and TOR are known to coordinate the entry into and exit from the quiescent phase (Wilson & Roach, 2002). These two pathways regulate the entry into the stationary phase, converging on the protein kinase Rim15 (Pedruzzi *et al.,* 2003) and downstream transcriptional activators, including the stress response (STRE) transcription factors (TFs) Msn2/Msn4 and the post-diauxic shift (PDS) transcription factor Gis1 (Zhang *et al.,* 2009). The nutrient starvation response is an intensively studied process, but the exact molecular mechanisms involved have not yet been fully elucidated. On the one hand, the scarcity of quantitative data poses a problem for the construction of quantitative models; on the other hand, the current understanding of the causal regulatory wiring encourages the use of qualitative computational models to gain new insights.

Executable Biology (Fisher & Henzinger, 2007; Fisher & Piterman, 2010) is an evolving paradigm that focuses on the design of executable computer algorithms that mimic biological phenomena through the use of formal methods from engineering and computer science. Biological knowledge can be captured in mathematically sound formalisms and then easily translated into executable algorithms for dynamical analysis and automatic reasoning. Here, we show that formalizing the available know-

ledge on the nutrient starvation response as a qualitative model highlighted the different modulation of Gis1 availability, encouraging further *in vivo* investigations on the role of proteosome-mediated proteolysis.

Proteasome-mediated proteolysis is essential for many cellular processes in yeast and other eukaryotes, including regulation of protein concentrations and degradation of misfolded proteins. Integrating our computational insights and the *in vivo* experiments with genome-wide bioinformatics analyses leads us to suggest that proteasome-mediated proteolysis of potentially toxic transcription factors tightly modulates the stress response in yeast.

4.2 Methods

4.2.1 Petri nets

We have built a qualitative logical model of nutrient starvation based on Petri nets. Petri nets are mathematically sound formalisms that can be graphically represented (Reisig & Rozenberg, 1998). Recently, Petri nets have been used in systems biology to build and analyze coarse-grained models of complex processes (Bonzanni *et al.,* 2009a), taking advantage of the intuitiveness of their representation and the soundness of their foundation. The Petri net modeling framework used in this work has been derived from the seminal work of Chaouiya and colleagues (Chaouiya *et al.,* 2006) as explained in Chapter 2, Section 2.2.2 and Section 2.2.3. The states predicted by the model can be found in Supplementary Information. Statistical analyses of bioinformatics data were performed using R.

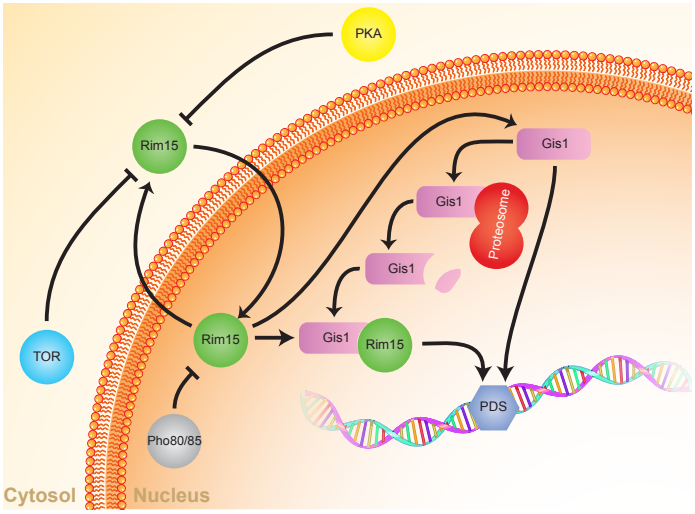
4.2.2 Gis1 overexpression at the transition phase

Wild-type (BY4742) cells were transformed with pCM190 (Garí *et al.,* 1997) and pCM190-*GIS1* (Zhang & Oliver, 2010). Transformants were grown on SMM (Amberg *et al.,* 2005) containing 20 µg/ml of doxycycline (Sigma-Aldrich) and 2% glucose to glucose starvation. Cells were harvest-ed, washed once in sterile water and resuspended in SMM medium containing no doxycycline or glucose for 36 hours to allow Gis1 over-expression. Growth was resumed by adding 2% glucose and doxycycline. Cell viability was checked by staining cells with phloxine B (Sigma-Aldrich).

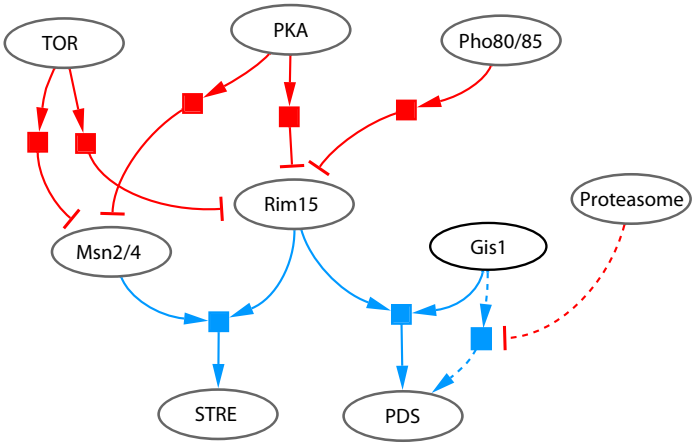
4.3 Results

4.3.1 Model construction and analysis

In order to investigate the consistency and explanatory power of the available knowledge about the nutrient starvation response in yeast, we have constructed a dynamic computational model based on Petri nets (Reisig & Rozenberg, 1998). Petri nets



A



B

Figure 4.1: Model of nutrient starvation response in yeast. (a) Diagrammatic model depicting the proteolytic control over Gis1 and the regulation of Rim15 by TOR, PKA and Pho80/Pho85. (b) Partial formal model of nutrient starvation response. Ovals=nodes that represent 'places' – proteins (e.g. PKA, Rim15, Gis1) and genes (PDS and STRE); colored squares=interactions. Arcs ending with an arrowhead (in blue) represent positive interactions (*e.g.* activations), while arcs ending with bars (in red) represent negative interactions (*e.g.* inhibitions). Note that if multiple arrows target the same square, all the sources are required at the same time. Dashed lines represent the interaction responsible for the discrepancy between the modeled and observed behaviors.

can be depicted as graphs that contain two kinds of nodes: *places*, which represent resources and correspond to proteins and genes, and *transitions*, which represent interactions between places. Interactions can be either activations or inhibitions (Figure 4.1b) and, during the course of the execution, each resource can change its state (in a Boolean fashion) from active to inactive (and vice versa) based on the surrounding interactions. Given a network topology, it is possible to execute the model and compare its behaviour with the one observed empirically. Due to the lack of fine-grained quantitative data, we captured the coarse-grained descriptive knowledge available in the form of a qualitative model firmly based on published experimental evidence. This model includes the inhibition of Msn2/4 activity by TOR and PKA (Beck & Hall, 1999; Görner *et al.,* 1998), which is represented in Figure 4.1b by the red transitions connecting the TOR and PKA nodes to Msn2/4. Notice that the arc connecting TOR to the transition ends with an arrowhead, while the arc connecting the transition to Msn2/4 ends with a bar. This means that the availability of TOR is a necessary precondition for Msn2/4 repression. Similarly, we have represented Rim15 inhibition by TOR, PKA, and Pho80/85 (Pedruzzi *et al.,* 2003; Wanke *et al.,* 2005), the expression of STRE and the PDS genes upon Rim15 activation of Msn2/4 and Gis1, as well as the recently discovered proteolytic control over Gis1 (Zhang & Oliver, 2010).

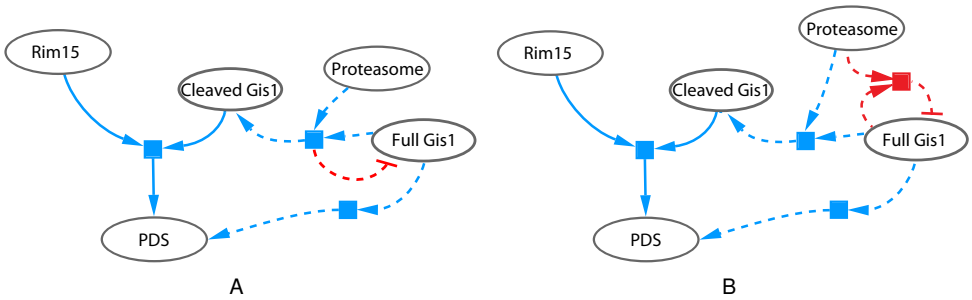


Figure 4.2: Multiple possible wiring choices allow refinement of the model. Fragment of the model under refinement. The dashed interactions in this figure are more accurate alternatives than the ones in Figure 4.1b. Two alternative options are presented: (a) proteolytic activity induces complete degradation of the full-length Gis1 protein and simultaneous availability of cleaved Gis1 fragments. (b) Decoupling the production of cleaved Gis1 fragments and degradation of full-length protein allows partial depletion of the full-length Gis1.

After the construction of the network model, we analyzed its dynamics. By comparing our model with the experimental observations (Zhang & Oliver, 2010), we discovered a significant discrepancy in the behavior of Gis1 reproduced by the model. Our computational results (see Supplementary Information) suggested that only the full-length Gis1 was necessary for the activation of PDS genes. However, upon nutrient starvation or TORC1 inhibition, the abundance of full-length Gis1 decreases, which does not correspond to the increase of transcription activation of PDS genes

(Zhang & Oliver, 2010). Moreover, although full-length Gis1 is essential for PDS gene expression, the smaller Gis1 fragments, resulting from constitutive proteolysis by the proteasome, are also able to initiate transcription upon Rim15 activation (Zhang & Oliver, 2010). These data suggested that full-length Gis1 and its smaller variants activate the transcription of PDS genes cooperatively. Therefore, we concluded that our model needed to be refined by including the full-length protein and the smaller fragments separately, in order to fully capture the biological observations and increase the model's accuracy. Different wiring choices were possible. One possibility, shown in Figure 4.2a, is to allow proteolytic activity to induce *complete* degradation of full-length Gis1. This is the behavior observed during nutrient starvation; however, Gis1 is also subject to a constitutive, but *partial*, degradation by the proteasome (Zhang & Oliver, 2010) during exponential growth. Therefore, an alternative modeling choice is to allow partial depletion of full-length Gis1. This can be accomplished by decoupling the availability of the cleaved Gis1 fragments from the complete degradation of the full-length protein (Figure 4.2b). By refining our model as shown in Figure 4.2b, it qualitatively reproduced (see Supplementary Information) the behaviour observed in Zhang & Oliver, 2010.

4.3.2 Proteolytic control over Gis1 allows fast recovery from lag phase

The different causal wirings imply differences in the model behavior and may therefore suggest different roles for the proteolytic control. In order to understand the evolutionary advantages of the different proteolytic controls over Gis1 in the context of nutrient response, we were prompted to investigate its physiological role. *GIS1* overexpression leads to accumulation of the full-length protein and is toxic to cell growth (Pedruzzi *et al.,* 2000; Zhang & Oliver, 2010). Inhibition of the proteasome function results in hyperactivation of PDS genes in nutrient-starved conditions (Zhang & Oliver, 2010). Knowing that growth and budding are suspended in stationary phase, we performed an experiment to determine whether the proteolytic control over Gis1 is necessary for survival of cells entering stationary phase, the recovery of cells from glucose starvation, or both. Wild-type yeast cells were transformed with plasmid pCM190 or the same plasmid bearing the *GIS1* gene under the control of the repressible promoter, tetO. Cells were grown in the presence of doxycycline to early stationary phase, washed, and resuspended in medium with no glucose or doxycycline for 36 hours. There is no difference in viability between cells bearing the empty plasmid and those carrying the tetO-*GIS1* plasmid (data not shown). Glucose and doxycycline were added to allow cells to resume growth. As shown in Figure 4.3, cells harboring the tetO-*GIS1* plasmid display a 15% longer lag phase than those bearing the empty plasmid, suggesting that *GIS1* overexpression during the transition to quiescence delays the subsequent resumption of exponential growth on readdition of nutrients. These data indicate that proteolytic degradation of Gis1 by the proteasome may provide cells with an important evolutionary advantage, since periods of nutrient availability and starvation are commonly experienced by microorganisms (Gasch & Werner-Washburne, 2002).

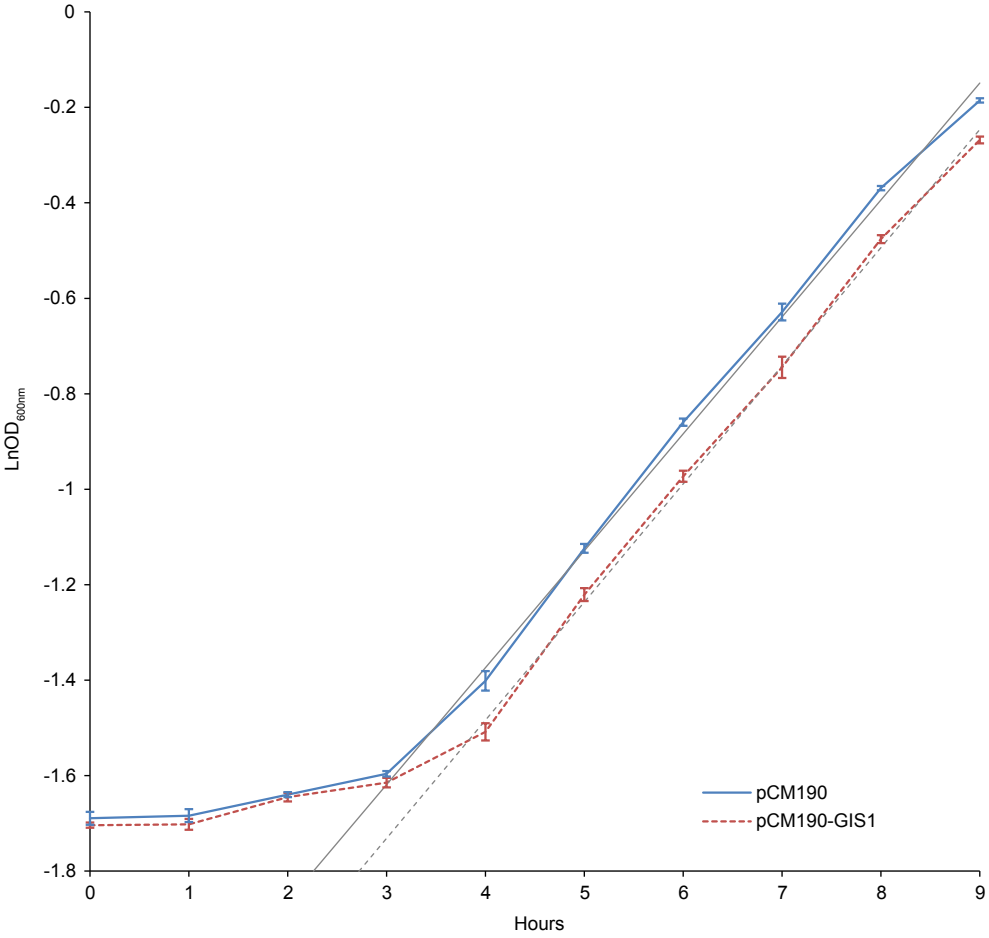


Figure 4.3: Cells over-expressing Gis1 at starvation have a longer lag phase than wild-type cells. Cells bearing either pCM190-*GIS1* or the empty vector, pCM190, were grown (in the presence of doxycycline) for 24 h to glucose starvation, at which point >90% of cells have no buds. Aliquots (2 ml) of cell cultures were washed twice in SMM medium without glucose or doxycycline, resuspended in 40 ml of SMM, and incubated for 36 h to allow GIS1 expression. At this point, glucose (2%) and doxycycline (20µg/ml) were added to the cultures. Growth was monitored as OD_{600nm}.

4.3.3 Predicting that toxic transcriptional regulators are subject to tighter proteolytic control

Prompted by the proteolytic regulation of Gis1 and its physiological implications, we went on to inquire if, in general, the stress response is restrained by the proteasome. We adopted two strategies: the first to discover whether toxic transcription factors are likely to be controlled post-translationally by the proteasome, and the second to find out whether proteasome inhibition allows transcription factors normally targeted by the proteasome to elicit a stress response.

Toxic transcriptional regulators have lower half-life

To monitor the validity of our hypothesis, we performed a sequence of bioinformatics analyses. First, we partitioned the known yeast transcriptional regulators into two disjoint sets. The first set contained 75 potentially toxic regulators and was created by filtering the set of 796 genes whose overexpression was found to be detrimental for cell growth (Sopko *et al.,* 2006) using the GO annotation "transcription regulator activity" (GO:0030528). The second set contained 251 non-toxic regulators and was built by filtering the whole yeast genome with the same GO annotation after removing the toxic genes contained in the first set. Detailed data are available as Supplementary Information. With our first analysis, we assessed whether the protein half-lives of toxic regulators are shorter than those of non-toxic regulators, using the protein half-life measurements of Belle *et al.,* 2006. Since the measurements are not normally distributed ($P < 10^{-15}$; Shapiro-Wilk test), we computed the Wilcoxon rank sum test under the null hypothesis that the median difference between the two measurement sets is zero and the alternative hypothesis that the median half-life of the toxic transcription factors is less than that of the non-toxic ones. The null hypothesis has been discarded with the statistically significant value of $P = 5.54 \times 10^{-3}$ (Figure 4.4a). Note that it was not possible to find measurements for all the proteins in the two sets. We also analyzed the mRNA half-life data (Wang *et al.,* 2002) for the transcripts of the toxic and the non-toxic TFs and found no significant difference between the two ($P = 0.256$; Wilcoxon test), supporting the hypothesis that a significant portion of the control over the toxic TFs is exerted post-transcriptionally (Figure 4.4b).

Toxic transcriptional regulators have a higher fraction of unstructured regions

The availability of many intrinsically unstructured proteins (IUPs) is regulated via proteolytic degradation (Gsponer *et al.,* 2008). Therefore, for both the toxic and non-toxic regulators, we computed (using Disopred2; Ward *et al.,* 2004 the fraction of the amino acids in each protein that are predicted to lie within an unstructured regions. We found (Figure 4.4c) that the median content of unstructured regions is higher for toxic transcription factors than that for non-toxic regulators ($P = 2.48 \times 10^{-4}$, Wilcoxon test), supporting the hypothesis that proteasome-mediated degradation plays a significant role in the regulation of the activity of potentially detrimental TFs.

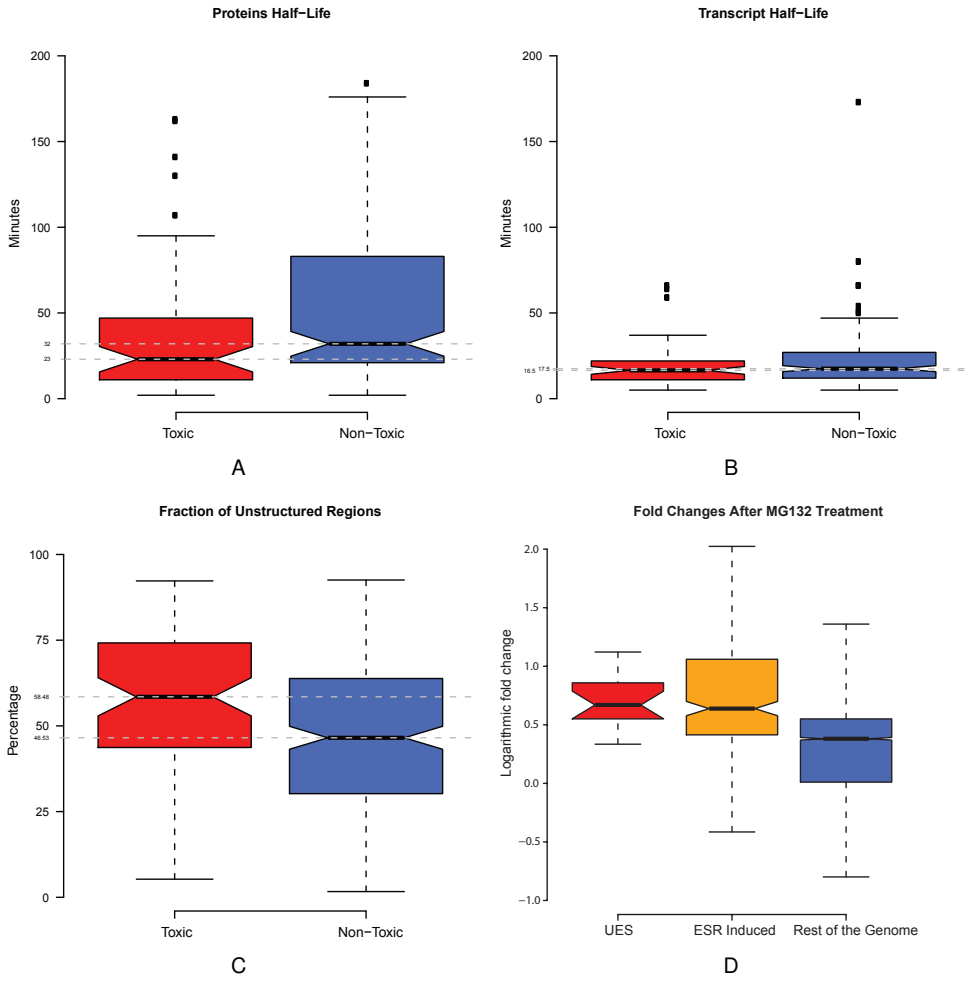


Figure 4.4: Comparison between toxic and non-toxic regulators. (a) The half-lives of toxic regulators (red) are significantly lower ($P = 5.54 \times 10^{-3}$; Wilcoxon test) than those of non-toxic ones (blue), while (b) the median half-life of the transcripts of toxic regulators is not significantly different from that of the non-toxic ones ($P = 0.256$; Wilcoxon test). (c) The fraction of amino acids predicted to form unstructured regions is significantly higher in toxic than in non-toxic proteins ($P = 2.48 \times 10^{-4}$; Wilcoxon test). (d) After 120 min of proteasome inhibition by MG132, transcription rates of UES genes ($P = 8.26 \times 10^{-5}$; Wilcoxon test) and ESR induced genes ($P < 2.2 \times 10^{-16}$; Wilcoxon test) tend to be higher than those for the rest of the genome.

Toxic TFs contain more potential PEST motifs

Sequence regions rich in proline (P), glutamic acid (E), serine (S), and threonine (T) are found in many rapidly degraded proteins and have been suggested to serve as signals for proteolysis (Rogers *et al.,* 1986). We analyzed the number of potential PEST motifs in the protein sequences of the two classes. Using the epestfind algorithm from the EMBOSS package (Rice *et al.,* 2000), we predicted the number of potential PEST motifs for both sets of proteins. While 44/75 (59%) toxic regulators contain at least one PEST motif, the ratio is 109/251 (43%) for the non-toxic ones (P value of 3.7×10^{-2}, Fisher's exact test). This, again, provides some support for our hypothesis on the role of proteolysis in regulating the activity of potentially toxic TFs.

The proteosome modulates the expression of a significant fraction of genes induced by environmental stress

Finally, we investigated whether proteolytic control could contribute to modulating the stress response by checking transcriptional changes after proteasome inhibition. A previous study has shown that 23% of all yeast genes (1386 mRNAs) increase their rate of transcription by a factor of 1.5 or more (6% increase more than 2 times) after 120 min treatment with the proteosome inhibitor MG132 (Dembla-Rajpal *et al.,* 2004). We extracted the data for the Universally Expressed at Starvation (UES) genes (Wu *et al.,* 2004); these genes are controlled by Gis1 and Msn2 – two TFs known to be under proteolytic control. We found that the fold changes of the UES genes tend to be higher than for the rest of the genome ($P = 8.26 \times 10^{-5}$; Wilcoxon test). More interestingly, we observe a significant fold increase with respect to the rest of the genome ($P < 2.2 \times 10^{-16}$; Wilcoxon test), further extending the analysis of the effect of inhibiting proteasome activity on the induction of gene transcription in the Environmental Stress Response (ESR; Gasch *et al.,* 2000), see Figure 4.4d.

To summarize, our work suggests that proteasome-mediated proteolysis of TFs tightly modulates the stress response in yeast. This hypothesis is the result of the integration of computational and *in vivo* analysis. Our computational model highlighted the particular behaviour of the proteolytic control, suggesting further *in vivo* investigations. Our *in vivo* experiments showed that, for the Gis1 transcription factor at least, proteasome-mediated control is crucial for a rapid return to growth after nutrient starvation, which may give yeast cells an important selective advantage over their competitors. Finally, our bioinformatics analyses generalized our *in vivo* observations to the class of potentially toxic transcription factors that control the stress response in yeast.

4.4 Acknowledgements

We would like to thank A. Feenstra, W. Fokkink and J. Heringa for helpful discussions. Part of this work was done, while NB was an intern at Microsoft Research Cambridge. *Funding*: Work in the Cambridge Systems Biology Centre was supported by BBSRC

(Grant BB/C505140/2 awarded to S.G.O.); Work in the Centre for Integrative Bioinformatics VU was supported in part by ENFIN; a Network of Excellence funded by the European Commission within its FP6 Program, under the thematic area 'Life Sciences, genomics and biotechnology for health', contract number LSHG-CT-2005-518254.

Chapter 5

Formal model reveals hard-wired heterogeneity in blood stem cells

Submitted as:

Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model[a]

Nicola Bonzanni[1 2] , Abishek Garg[3] , Samuel D. Foster[4] , Nicola K. Wilson[4] , Sarah Kinston[4] , Diego Miranda-Saavedra[4] , Jaap Heringa[1 2] , Anton Feenstra[1 2] , Ioannis Xenarios[3] , Berthold Göttgens[4]

[a]Supplementary data for this chapter are available at http://www.bonzanni.com/phd/ch5-sup.pdf

[1]Centre for Integrative Bioinformatics, VU University Amsterdam, The Netherlands
[2]Department of Computer Science, VU University Amsterdam, The Netherlands
[3]Swiss Institute of Bioinformatics, Switzerland
[4]Cambridge Institute for Medical Research, UK

Abstract

Combinatorial interactions of transcription factors with cis-regulatory elements control the dynamic progression through successive cellular states and thus underpin all metazoan development. The construction of regulatory network models based on the functionality of cis-regulatory elements therefore has the potential to generate fundamental insights into cellular fate and differentiation. Haematopoiesis has long served as a model system to study mammalian differentiation, yet modelling based on experimentally informed cis-regulatory interactions has so far been restricted to pairs of interacting factors. Here we have generated a network model based on detailed cis-regulatory functional data connecting 11 haematopoietic stem cell (HSC) regulators. Dynamic analysis of our model predicts that HSCs display heterogeneous expression patterns and possess many intermediate states that appear to act as "stepping stones" for the HSC to achieve a final differentiated state. By focussing on intermediate states occurring during erythrocyte differentiation, we predicted a novel negative regulation of Fli1 by Gata1 which we confirmed experimentally thus validating our model. In conclusion, we present the most advanced mammalian regulatory network model based on experimentally validated cis-regulatory interactions to date. This model has allowed us to make novel, experimentally testable hypotheses about transcriptional mechanisms that control differentiation of mammalian stem cells.

5.1 Introduction

The remarkable power of small combinations of transcription factors to program and reprogram cellular phenotypes is exerted through their ability to modulate the expression levels of their target genes, typically in the range of a few hundred to a few thousand genes. Despite the power of single transcription factors to influence cell fate decisions, it is clear that the transcriptional state of any given cell type is the result of interactions within wider transcriptional regulatory networks. These regulatory networks are composed of both the transcription factors (TFs) and the cis-regulatory elements they are bound to (Davidson, 2006). Regulatory network reconstruction therefore requires the identification of cis-regulatory elements as well as the upstream factors which bind them.

Haematopoiesis (blood formation) has long served as a model process for studying stem cells and represents the best characterised adult stem cell system with sophisticated purification strategies and functional stem cell assays. Transcriptional regulation is a key factor controlling haematopoiesis (Miranda-Saavedra & Göttgens, 2008), a fact underlined by the large number of TF genes that play key roles in normal haematopoiesis and/or the development of leukaemia (Göttgens, 2004). However, relatively little is known about the way key regulators interact with each other in forming the transcriptional networks controlling haematopoiesis.

Identification and subsequent characterisation of gene regulatory elements is central to the reconstruction of transcriptional regulatory networks because these

elements dictate the connectivity and topology of transcriptional regulatory networks (Davidson, 2006). Regulatory elements can be analysed using a variety of assays such as transfection assays of luciferase reporter constructs or chromatin immunoprecipitation (ChIP) analysis to identify upstream regulators. However, the identification of true in vivo activities of mammalian regulatory elements requires the use of transgenic mouse systems. Regulatory elements from 11 gene loci active in haematopoietic stem/progenitor cells (HSPCs) have been validated using all the abovementioned assays including transgenic mice (Donaldson *et al.,* 2005; Göttgens *et al.,* 2002, 2004; Kobayashi-Osaki *et al.,* 2005; Landry *et al.,* 2008; Nottingham *et al.,* 2007; Pimanda *et al.,* 2007; Vyas *et al.,* 1999; Wilson *et al.,* 2009). This wealth of data therefore represents a unique opportunity to (re)construct transcriptional network models for developing blood stem cells.

Network modelling is increasingly recognised as a powerful approach to deal with the complexity of biological processes including the intricate interactions between TFs (Georgescu *et al.,* 2008; Hu *et al.,* 2007; Karlebach & Shamir, 2008; Spooner *et al.,* 2009). Most of the current experimental data describing the function of haematopoietic TFs are of a qualitative nature (e.g. Gata1 and Scl together activate Scl expression) which limits the choice of possible modelling approaches. However, the accumulated knowledge of regulatory interactions (Foster *et al.,* 2009; Swiers *et al.,* 2006) contains experimentally validated information on the topology of regulatory subcircuits, including positive and negative feedback loops which are important for maintenance of both the stem cell phenotype (Pimanda *et al.,* 2007) and differentiation into different mature blood cell types (Sieweke & Graf, 1998). An important challenge for regulatory network reconstruction is to devise models that can represent the dynamic interactions between important subcircuits and represent the changes in gene expression when cells are undergoing differentiation.

Importantly, experimentally defined regulatory hierarchies based on regulatory elements up to now largely represent a static view, which, in the case of blood stem cell formation is centred on a single time-point in transgenic mouse assays (activity within the dorsal aorta region and fetal liver of the mid-gestation mouse embryo). Here, we have generated a network model based on extensive experimental data with the goal to better understand how core stem cell network circuits are incorporated into the wider dynamic system of blood stem cell development and differentiation. Through the modelling of steady states and dynamic network behaviour, we were able to identify specific genes and feedback loops within the network that are likely key players in cellular decision making such as the dynamic processes of stem cell maintenance and/or differentiation. Moreover, this analysis revealed heterogeneous gene expression states within undifferentiated blood stem cells as well as perturbations required to push the network out of the “stem cell state”. Importantly, a new hypothesis on negative regulation of Fli1 by Gata1 was validated experimentally using transcriptional assays thus providing new insights into the dynamic nature of regulatory networks controlling erythroid differentiation.

5.2 Results

5.2.1 Functional analysis of an Erg enhancer element identifies novel regulatory interactions

An enhancer element +85 kb into the mouse Erg gene locus has recently been identified as a region bound by Scl *in vivo* (Wilson *et al.,* 2009). Together with Fli1 and Pu.1, Erg is a key member of the Ets family of transcription factors important for regulating gene expression in HSPCs (Chan *et al.,* 2007; Kruse *et al.,* 2009; Pimanda *et al.,* 2006; Rainis *et al.,* 2005). We had therefore previously assessed binding of these three Ets factors to key HSPC regulatory elements but a regulatory element anchoring Erg itself into the emerging network had remained elusive until our recent discovery of the Erg +85 region. The Erg +85 region is highly conserved with candidate binding sites for important HSPC regulators including E-boxes and Ets sites representing the consensus binding sites for Scl and Erg/Fli1/Pu.1 respectively (see Figure 5.1). When tested in transgenic mice this region drove reporter gene expression to endothelial cells as well as blood progenitor cells in the dorsal aorta and fetal liver (see Figure 5.2). To assess the contribution of E-box and Ets motifs to the overall activity of the Erg +85 element, mutations were introduced into the two E-boxes and 3 Ets sites and the resulting mutant constructs assayed by stable transfection of luciferase constructs. These assays revealed that mutation of the two E-boxes resulted in a significant increase of enhancer activity whereas mutation of the three Ets sites reduced enhancer activity to baseline levels (Figure 5.3a).

Given that we had previously shown Scl to be a positive regulator of E-box containing elements within our network (Ogilvy *et al.,* 2007; Pimanda *et al.,* 2007), our finding that mutation of the E-box increased activity was surprising. However, Scl has been reported to form repressive complexes together with the co-repressor Eto2

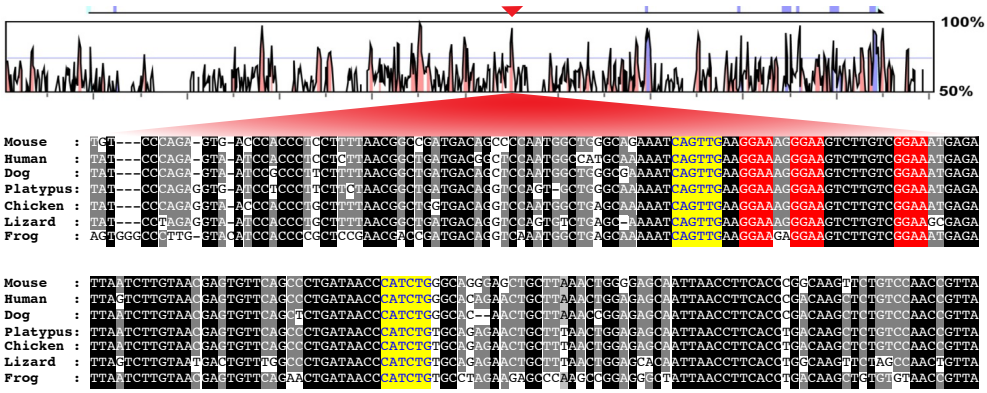


Figure 5.1: Schematic diagram of the mouse Erg locus with mouse/human sequence homology plot and nucleotide sequence alignment with conserved E-box and Ets sites marked in yellow and red respectively.

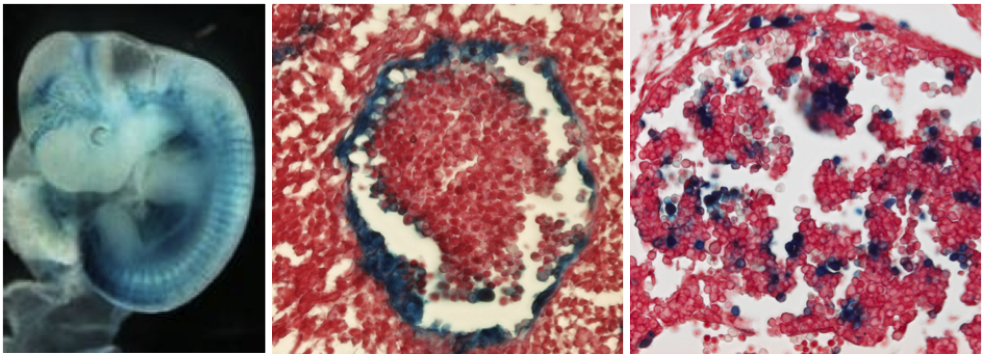


Figure 5.2: SV/lacZ/Erg+85 E11.5 transgenic embryo with transgene activity (blue staining) in blood vessels. Histological section of the dorsal aorta (middle panel) shows staining in dorsal aorta endothelium and clusters attached to it, the presumptive site of blood stem cell emergence in the mouse embryo. Fetal liver section (right panel) illustrates staining in fetal liver blood cells.

(Schuh *et al.,* 2005) and we therefore investigated binding of not only the Ets factors Erg, Fli1 and Pu.1 but also Scl and Eto2 by ChIP assays. We also included Gata2 in this analysis as it is the common binding partner of Scl in HSPCs. As shown in Figure 5.3b, all six transcription factors showed significant binding to the Erg+85 region. Functional analysis of the Erg+85 element therefore allowed us to identify upstream regulators and importantly determine whether they function as activators or repressors of the overall activity of this element. Of note, Eto2 itself was recently identified as a target of Scl and Gata2 (Wilson *et al.,* 2009). Taken together therefore, these transcriptional assays allowed us to proceed with full integration of Erg into an emerging transcriptional network and also include the newly discovered repressive link between the Scl/Eto2 complex and Erg.

5.2.2 Building of a transcriptional regulatory network model for blood stem cells

By combining the functional data generated for the Erg+85 enhancer with previously published results on haematopoietic regulatory elements, we were in a position to construct a regulatory network model based on 11 fully validated regulatory elements linking together 11 transcription factors all of which are active in early HSPCs. Figure 5.4 shows the resulting 11-gene regulatory network. Importantly, since all 11 elements have been studied extensively using DNA/protein binding assays as well as reporter gene assays of wild type and mutant elements, both the direction and value of regulatory interactions are known. Moreover, protein-protein interactions curated from the literature were included such as the well characterised Gata1-Pu.1 interactions whenever their value (activatory/inhibitory) was known.

The resulting network was modelled as logical interactions encoding the activating and/or inhibitory links, including the specific combinations in which particular

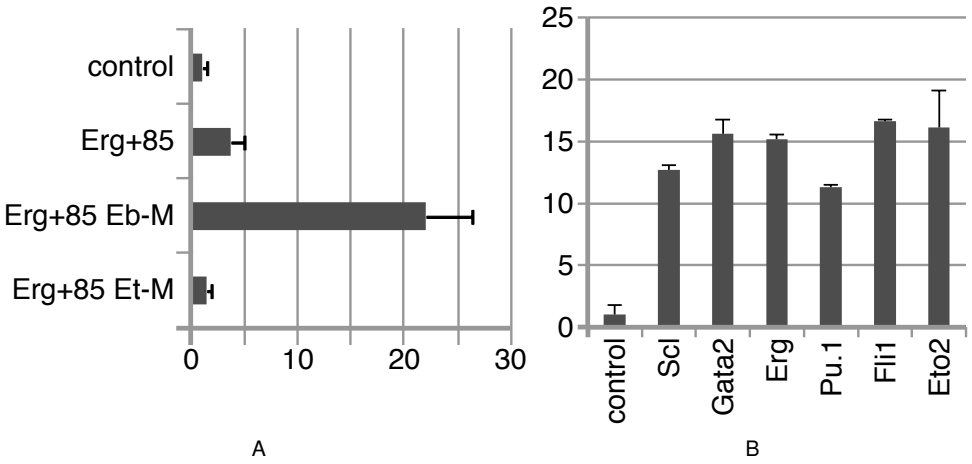


Figure 5.3: Scl is a negative regulator of the Erg +85 enhancer. (A) Results of stable transfection assays in 416B cells corresponding to the wild type and E-box as well as Ets mutant versions for the Erg +85 enhancer. The luciferase activities are given as the fold-increase over the activity of the negative control vector (pGL2-promoter) alone. (B) Real-time PCR analysis of ChIP assays in progenitor cells (HPC-7 cell line) performed with antibodies against the factors indicated. Levels of enrichment were normalized to IgG and compared with a negative control region as described (Wilson *et al.,* 2009).

interactions occur (e.g. Gata2 and Scl together activate Eto2). This logical model was implemented in Boolean notation (see supplementary data for full network description). Several observations are noteworthy: (i) a network of 11 genes with three types of possible interactions (activatory, inhibitory, none) could adopt in excess of 10^{50} possible network topologies. It would therefore simply be unfeasible to perform modelling analysis using all possible topologies and then work backwards to identify the likely correct topology (hence the need for experimental data). (ii) at the heart of the network lies the triad of Scl, Gata2 and Fli1 which is characterised by extensive positive feedback loops but negative regulatory interactions are common outside this central triad. (iii) we have 11 genes but 47 links (an average degree of 4.3) forming a densely connected network. Within this network we can identify an even more densely connected core consisting of Erg, Gata2, Scl and Fli1 with an average degree of 8.5. Furthermore Gata2 and Scl connect out to most other genes, and nearly always operate together as a dimer.

5.2.3 Network genes are expressed dynamically during haematopoiesis

In order for a network model to be useable as a predictive tool, the behaviour of its component genes needs to be assessed using available experimental data. We therefore explored the expression patterns of the 11 component genes in primary

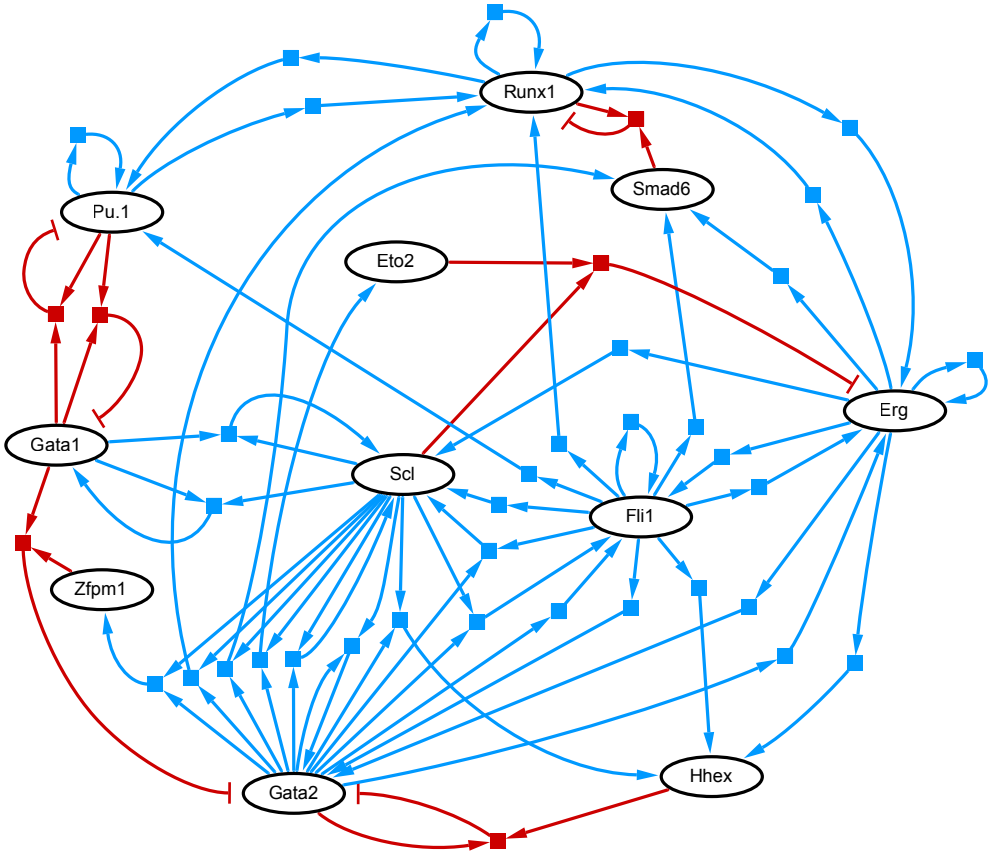


Figure 5.4: A blood stem cell regulatory network model grounded on comprehensive cisregulatory information. Petri net model of the haematopoietic gene regulatory network. Interactions between proteins are displayed as squares. Blue lines represent positive interactions (*i.e.* gwhile red arrows represent negative interactions (*i.e.* repressions).

haematopoietic cell types. To this end, we took advantage of two recently published datasets: a single cell gene expression profiling study comparing haematopoietic stem with progenitor cells (Ramos *et al.,* 2006) and the haematopoietic fingerprints database, a collection of expression profiling data for HSCs as well as 9 differentiated lineages (Chambers *et al.,* 2007). Based on the available literature, all our HSC network genes except Gata1 should be expressed in the most immature stem cell population which is precisely what we found when interrogating the two expression profiling datasets. Moreover, Gata1 expression was found in the immediate progeny of the most immature progenitors, e.g. the multipotent progenitor population. In contrast to the ubiquitous expression of our 11 genes in the stem/progenitor compartment, mature blood lineages only express subsets of the 11 genes that make up the HSC network ranging from 2 out of 11 in activated CD8 T-cells to 7 out of 11 in granulocytes. Of note, different mature cell types express different subsets of genes which prompted us to investigate whether this variability would be sufficient to at least partially reconstruct a haematopoietic differentiation tree. Indeed, clustering based on expression of these 11 genes was sufficient to capture key aspects of the haematopoietic differentiation tree (Supplementary Figure 1). Our HSC network model may therefore not only reveal properties of the stem cell state but also allow us to interrogate potential mechanisms and external stimuli that direct stem cell differentiation into specific mature lineages.

5.2.4 Dynamic modelling of the network predicts heterogeneous HSC expression states

Having generated a complex vertebrate transcriptional regulatory network model based on state of the art experimental evidence, we next performed dynamic modelling analysis to explore whether any predicted network behaviour would allow us to gain new insights into blood stem cell biology. Dynamic modelling revealed that the experimentally validated network topology allows for three stable states (Table 5.1): (i) all genes are off (S-3-1), (ii) only Gata1 and Scl are expressed (S-2-1) and (iii) an interconnected set of 32 expression states with multiple genes active but Gata1 always repressed (S-1-1 to 32). In order to explore if these steady states matched observed cell states, we next performed clustering of expression patterns from our stable states together with the expression patterns in the 10 haematopoietic cell types (see supplementary Figure 2). Steady state S-3-1 corresponds to a non-haematopoietic cell and S-2-1 closely resembles a mature erythrocyte. Most interesting however is stable state S-1 which is composed of 32 interconnected internal states including a state that matches the expected pattern for HSCs. This suggests that the precursor HSC is not a homogeneous cell population, but rather is composed of cells in different stages of activation. Furthermore, there is a striking correlation between gene expression profiling results from single HSCs (Ramos *et al.* (2006), summarised in Table 5.2) and the heterogeneous states predicted by our network since those genes predicted by our model to be stably present were consistently found expressed in a high proportion of single cell profiling experiments whereas genes predicted to be “oscillating” by our

Table 5.1: Stable states reveled by dynamic modelling.

Cell Type	ID	Erg	Eto2	Fli1	Gata1	Gata2	Hhex	Pu.1	Runx1	Scl	Smad6	Zfpm1
Non-Haematopoietic Cell	S-3-1	A	A	A	A	A	A	A	A	A	A	A
Erythrocyte-Like Cell	S-2-1	A	A	A	P	A	A	A	A	P	A	A
Haematopoietic Stem Cell	S-1-1*	P	P	P	A	P	P	P	P	P	P	P
	S-1-2	A	P	P	A	P	P	P	P	P	P	P
	S-1-3	P	P	P	A	A	P	P	P	P	P	P
	S-1-4	A	P	P	A	A	P	P	P	P	P	P
	S-1-5	P	A	P	A	P	P	P	P	P	P	P
	S-1-6	A	A	P	A	P	P	P	P	P	P	P
	S-1-7	P	A	P	A	A	P	P	P	P	P	P
	S-1-8	A	A	P	A	A	P	P	P	P	P	P
	S-1-9	P	P	P	A	P	P	P	A	P	P	P
	S-1-10	A	P	P	A	P	P	P	A	P	P	P
	S-1-11	P	P	P	A	A	P	P	A	P	P	P
	S-1-12	A	P	P	A	A	P	P	A	P	P	P
	S-1-13	P	A	P	A	P	P	P	A	P	P	P
	S-1-14	A	A	P	A	P	P	P	A	P	P	P
	S-1-15	P	A	P	A	A	P	P	A	P	P	P
	S-1-16	A	A	P	A	A	P	P	A	P	P	P
	S-1-17	P	P	P	A	P	P	P	P	P	P	A
	S-1-18	A	P	P	A	P	P	P	P	P	P	A
	S-1-19	P	P	P	A	A	P	P	P	P	P	A
	S-1-20	A	P	P	A	A	P	P	P	P	P	A
	S-1-21	P	A	P	A	P	P	P	P	P	P	A
	S-1-22	A	A	P	A	P	P	P	P	P	P	A
	S-1-23	P	A	P	A	A	P	P	P	P	P	A
	S-1-24	A	A	P	A	A	P	P	P	P	P	A
	S-1-25	P	P	P	A	P	P	P	A	P	P	A
	S-1-26	A	P	P	A	P	P	P	A	P	P	A
	S-1-27	P	P	P	A	A	P	P	A	P	P	A
	S-1-28	A	P	P	A	A	P	P	A	P	P	A
	S-1-29	P	A	P	A	P	P	P	A	P	P	A
	S-1-30	A	A	P	A	P	P	P	A	P	P	A
	S-1-31	P	A	P	A	A	P	P	A	P	P	A
	S-1-32	A	A	P	A	A	P	P	A	P	P	A

* This state matches the expected pattern for HSCs

Table 5.2: Heterogeneous gene expression observed in single-cell microarray experiments of 12 individual haematopoietic stem cells (columns) from Ramos *et al.* (2006).

Eto2	P	A	A	A	A	A	A	A	A	A	P	A
Fli1	P	M	A	A	A	A	A	A	A	P	P	P
Gata1	A	A	A	A	A	A	A	A	A	A	A	A
Gata2	M	P	A	A	A	A	A	A	P	A	P	A
Hhex	A	A	P	P	P	A	A	A	P	A	P	A
Pu.1	A	A	P	P	A	P	A	P	A	A	A	A
Runx1	A	A	A	A	A	A	A	P	A	A	A	A
Scl	P	P	P	A	P	A	A	A	P	A	P	A
Smad6	A	A	A	A	M	A	P	A	P	A	P	P
Zfpm1	P	A	A	A	P	A	A	A	A	A	A	A

model were consistently found expressed in fewer single cells (see Figure 5.5). This analysis therefore not only demonstrates that our knowledge-driven network topology is compatible with expression patterns observed in HSCs in vivo, but also suggests that expression of Gata2, Zfpm1, Erg, Eto2 and Runx1 is heterogeneous in HSCs, and may define intermediate states within this cell population.

5.2.5 Prediction and experimental validation of a novel regulatory mechanism in erythroid differentiation

Analysis of transitions between different steady states in the model can be useful to predict experimental conditions for cells to differentiate out of the HSC state. We analysed the state transitions required for differentiation from HSC to different cell types in our model (see "Analysis of state transitions" in supplementary data). We chose the differentiation pathway towards erythroid cells for further investigation because (i) the pathway is well characterised at the experimental level, (ii) it has been the subject of modelling approaches based on simple 2-gene interactions (Chickarmane *et al.,* 2009; Roeder & Glauche, 2006) and (iii) it connects the two stable states reproduced by our 11-gene network. Of note, experimental evidence suggests that a single "trigger" or "push" (e.g. ectopic expression of Gata1) would be sufficient to drive immature blood progenitors towards an erythroid fate (Heyworth *et al.,* 2002; Kulessa *et al.,* 1995). However, our results (see supplementary Figure 3) suggested that HSC cells need to undergo two state changes or "pushes" as a trigger to differentiate into erythroid cells. We considered that there might be two possible explanations for this: (i) Gata1 regulates a protein not present in our network and this can generate this second "push" or (ii) there is a missing link in our wiring diagram which when introduced would increase the "power" of Gata1 so that its ectopic expression would become a single push differentiation trigger. Interrogating the first of these two possibilities is potentially rather speculative, but the second could be readily explored.

We therefore considered potentially missing network links from our current topo-

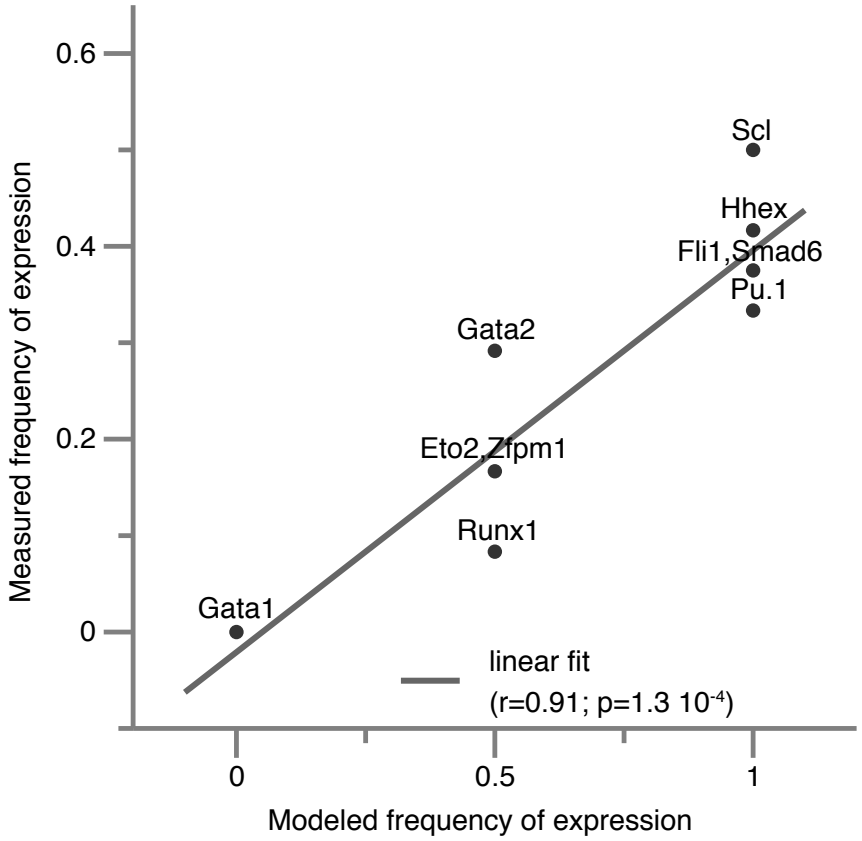


Figure 5.5: A near linear correlation of averaged gene expression activity from the 12 single cell profiles from Table 5.2 compared to average gene activity from the modelled HSC steady state. Circle size represents the number of genes (1 or 2) with the particular combination of experimental and modelled activity.

logy. In particular, we extended our model by introducing the possible repression of Fli1 by Gata1 based on the rationale that the Fli1 regulatory element is structurally similar to the Gata2 element, which is known to be repressed by Gata1 (Grass *et al.,* 2003). Interestingly, just introducing this single additional repressive link elevated Gata1 to a "single push" trigger for erythroid differentiation. Following on from this modelling result, we investigated whether Gata1 was indeed able to repress activity of the Fli1 enhancer in blood stem/progenitor cells. To test this, the haematopoietic progenitor cell line HPC7 was electroporated with a luciferase reporter construct containing the Fli1 enhancer together with either an empty control plasmid or a Gata1 overexpression construct. As shown in Figure 5.6, co-transfection of the Gata1 expression plasmid resulted in significant repression of the activity of the Fli1 enhancer

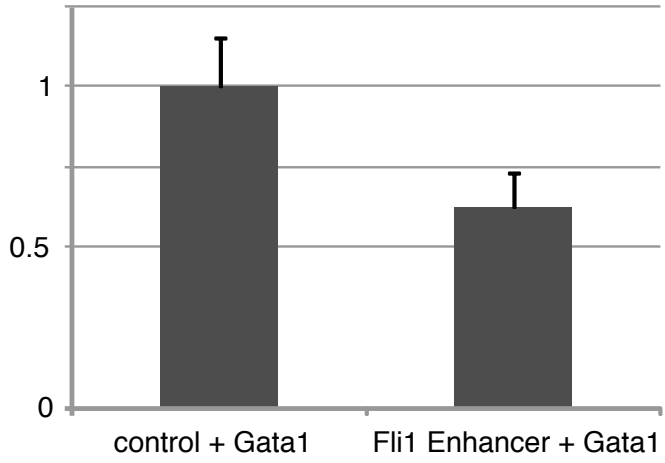


Figure 5.6: Co-transfection of the Fli1enhancer construct with a Gata1 expression vector results in significant reduction of the enhancer activity. Co-transfection studies were performed in the HSPC cell line HPC7. The data shown represent the average of 4 individual experiments, each performed in triplicate.

construct, thus demonstrating that Gata1 is indeed able to negatively regulate expression of Fli1. Network transition modelling therefore allowed us to predict a previously unrecognised network link which we were able to validate experimentally. The revised network diagram is shown in Figure 5.7 with the new repressive link indicated by a dashed line. Interestingly, including repression of Fli1 by Gata1 did not alter the steady states of our model, illustrating how some network links specifically influence transitions between states rather than the states themselves.

5.3 Discussion

The construction of accurate regulatory network models is an essential prerequisite towards gaining a systems level understanding of the transcriptional control of complex cellular behaviour. Here we have generated a regulatory network model for HSPCs based on comprehensive experimental data, which represents the most complex mammalian network model to date anchored on cis-regulatory functional data. This experimentally validated network topology generated three stable states, one of which was composed of 32 internal states including one that matched the stem cell expression pattern and the others oscillating around it. Analysis of state space transitions identified potential triggers that might mediate exit from the stem cell state and highlighted a previously unrecognised inhibition of Fli1 by Gata1, which was subsequently validated experimentally.

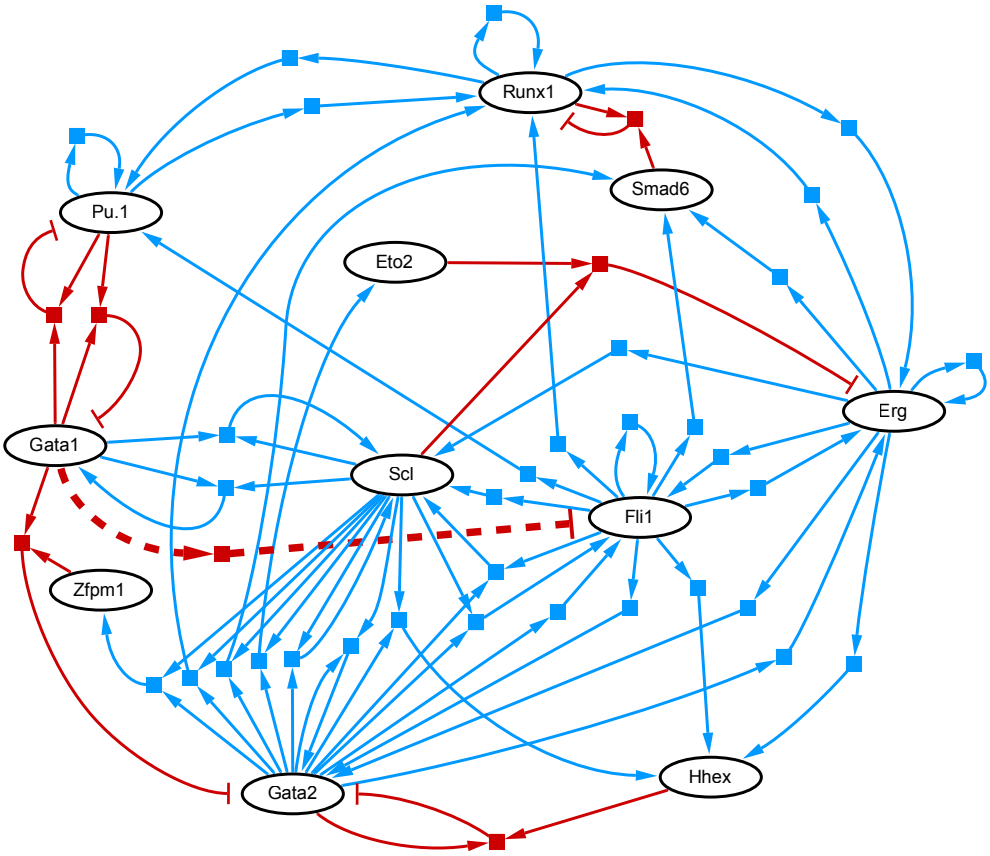


Figure 5.7: Diagram of the gene regulatory network, cf. Figure 5.4 showing the predicted and experimentally validated inhibition of Fli1 by Gata1 (indicated by a dashed line)

5.3.1 Experimentally validated network models – insights and open questions

Regulatory network topology determines the nature of possible regulatory states as well as the possible transitions paths between them. Full experimental validation of all interactions within our network model not only provides high confidence in the simulations/modelling, but also offers an opportunity to consider the possible consequences if our experimental knowledge was more limited. For example, without the repression of Erg by Scl validated in the current study, there would only be 16 rather than 32 internal substates in steady state 1. Importantly, introducing the novel interaction generates internal states that are closer to some of the differentiated states. Consequently, the number of internal states that a stem cell can "explore" increases with a concurrent decrease in the number of external triggers required to move out of the HSC state in order to differentiate.

Another notable observation is that most repressive interactions in the network (Figure 5.4) arise from pairs of genes. A common theme here is that co-regulators such as Eto2 and Zfpm1 are thought to bind DNA indirectly through interactions with conventional transcription factors such as Scl and Gata1, and by doing so convert the latter from activators to repressors. Interestingly, in our network these negative co-regulators are themselves activated by the conventional TFs thus generating an abundance of incoherent feed-forward loops within the wider network. Simple negative feedback loops have previously been proposed to result in oscillatory expression of important cell fate regulators (Hirata *et al.,* 2002; Lahav *et al.,* 2004). In order to better understand the potential for oscillatory behaviour in increasingly complex networks, future developments might need to include building more fine-grained models such as the use of Petri nets, which can be readily adapted to move from a Boolean range of values towards discrete multi-valued expression levels.

Within the context of our 11-gene HSPC network topology, several expression states that correspond to the differentiated cell types shown in supplementary Figure 3 can automatically revert to the stem cell state suggesting a potential for spontaneous reversion of differentiated cells to the immature stem cell phenotype. In a sense, this may merely be a reflection of the fact that our experimentally informed HSPC network topology generated a very stable HSPC attractor. However, it also suggests that "commitment features", that would block these regressions, may be missing from our network. It is likely that these commitment events will involve epigenetic processes that regulate the availability of regulatory regions for factor binding. For example, epigenetic silencing of a given regulatory element could prevent access of upstream factors with the consequence of "locking in" the differentiated state.

5.3.2 The "stem cell state" – a moving target?

Comprehensive exploration of the state space dictated by our experimentally validated HSPC network topology resulted in a set of 32 interconnected states which together constitute a stable state with a gene expression pattern consistent with HSPCs. However, only a single internal state in the HSC attractor matched expression levels of all HSPC associated genes whereas all others expressed different subsets of genes suggesting possible heterogeneity between discrete expression states. The heterogeneous steady state predicted by our model might at first have been considered an artefact due to either the unavoidably partial knowledge we have about the system, or introduced by the high level of discretization used (i.e. from potentially continuous expression levels to Boolean values). However, we believe that on the contrary our results may provide potentially important new insights into the nature of transcriptional control of stem cells and differentiation as outlined below: Firstly, the striking correlation between gene expression profiling results from single HSCs and the heterogeneous states predicted by our network (Figure 5.5). Moreover, single cell analysis of highly purified murine HSCs using digital PCR assays (Warren *et al.,* 2006) also showed heterogeneous transcription factor expression in individual HSCs. Taken together, these observations suggest that the stem cell state is composed of a dis-

crete set of substates with a substantial degree of oscillations in gene expression, which includes genes thought of as central regulators of stem cell fate. Of note, this concept is largely consistent with the recently introduced theory of non-genetic micro-heterogeneity in multipotential stem cell populations (Huang, 2009).

It might at first glance appear difficult to reconcile such oscillations and the resultant transcriptional heterogeneity with the model of multi-lineage priming. This latter concept was founded on the observation that some HSPCs display low-level co-expression of cytokine receptor genes affiliated with divergent differentiation pathways (Hu *et al.,* 1997). Consequently, HSCs have widely been thought of as highly promiscuous with widespread co-expression rather than only expressing subsets of genes. However, in addition to demonstrating the potential for multi-lineage priming, the original paper Hu *et al.* (1997) also found heterogeneous expression of stem cell affiliated genes when analysed at the single cell level. Both multi-lineage priming of cytokine receptor genes and expression of HSPC affiliated transcription factors therefore show cellular heterogeneity consistent with oscillating expression in individual HSPCs. Based on the results presented in this paper, cellular heterogeneity of multilineage priming may therefore be hard-wired into HSPC regulatory networks rather than being a consequence of low-level, non-specific gene expression noise as had been speculated previously. This in turn would suggest that characterisation of the underlying mechanisms will provide novel insights into the functional role of multi-lineage priming as a key mediator of differentiation. Rather than there being a "stem cell continuum", the regulatory space within which a stem cell can move may be constrained where a given differentiation trajectory requires passage through a number of specific intermediate states.

5.3.3 Discrete stem cell states and differentiation triggers

Since the stem cell state space is composed of a set of regulatory states with interconversions between them dictated by the network topology, the question arises to what extent knowledge of network wiring may increase our ability to manipulate stem cell fate choices. In this study we show that specific differentiation triggers can be modelled successfully and inform specific hypotheses for subsequent experimental testing. Importantly, specific substates within the stem cell state are closer to certain downstream cellular fates than others. This in turn suggests that the distribution of stem cell internal states has the potential to influence the propensity of a stem cell to choose between divergent differentiation choices. A mechanistic understanding of the underlying processes would have important scientific and clinical implications. For example, altering the levels of Gata2 has recently been shown to affect the ratio between cycling and quiescent HSCs (Tipping *et al.,* 2009) providing direct experimental evidence that levels for one of the factors shown to be oscillating in our network model are associated with phenotypically identifiable substates of HSCs. From a translational point of view, in vitro production of specific blood cell types from HSPCs has the potential to provide safer and cheaper alternatives to blood transfusions. However, directed differentiation in vitro remains disappointingly inefficient

suggesting that knowledge of the underlying regulatory networks is critical for the development of new protocols. Finally, treatment responses for patients carrying the same leukaemogenic mutations can be very different. Since many leukaemia oncogenes cause a differentiation block of early progenitors, it is possible that this block may occur in different "substates" of the stem cell compartment in different patients suggesting that a deeper understanding of these substates may provide novel treatment options.

5.4 Materials and methods

5.4.1 Experimental

The Erg +85 region (Wilson *et al.,* 2009) was inserted into lacZ and luciferase reporter constructs using standard techniques (details available on request). F0 transgenic mouse embryos were generated and analyzed as described (Sinclair *et al.,* 1999). All animal studies were performed with UK Home Office approval. Luciferase assays were carried out following stable transfection of reporter constructs in the progenitor cell line 416B as described (Göttgens *et al.,* 1997). Individual experiments were performed in triplicates on at least two different days. ChIP assays were performed as described previously (Wilson *et al.,* 2009) using the following antibodies: Scl - Santa Cruz sc-12954x, Gata2 Santa Cruz sc-9008x, Erg Santa Cruz sc-354x, PU.1 Santa Cruz sc-352x, Fli-1 abcam ab15289-500, Eto2 Santa Cruz sc-9739x. Primers used for real time Q-PCR analysis are available on request.

5.4.2 Boolean modelling

We used Boolean logic functions AND, OR, BUFF, IAND and NOT as described (Garg *et al.,* 2008; Kauffman *et al.,* 2003; Klamt *et al.,* 2006; Mendoza & Xenarios, 2006) to model the GRN where the state of a node i at time t is represented by a Boolean variable x_i^t. The Boolean functions were mathematically defined by Equations 1–5 (see supplementary). The expression of each gene i at time $t + 1$ can be written as a function $x_i(t + 1)$ of the state of the genes acting as its input at time t as illustrated for node B in Equation 6 (see supplementary data). A snapshot of the activity level of all the genes in the network at time t is called the state of the network represented by a Boolean vector, x^t, of size N (number of genes in the network) and is called the present state vector. Another Boolean vector, x^{t+1}, of size N is used to represent the state of the network in the next step and is called the next state vector. Assuming a synchronous model of transition, the transition function from the present state to the next state of the network is given by the Boolean function $T_i(x^t, x^{t+1})$ in Equations 7 and 8 (see supplementary data)

5.4.3 Petri nets

The network and the attractors have been computed using the Petri net approach and algorithms presented in Chapter 2, Section 2.2.2 and Section 2.2.3.

5.5 Acknowledgements

This study was funded by the European Commission FP6 project ENFIN (Experimental Network for Functional INtegration). Computations were performed at the Vital-IT Center of the Swiss Institute of Bioinformatics. We are grateful to Mark Ibberson for helpful comments. Experimental work was supported by Leukaemia & Lymphoma Research (UK), the Leukemia & Lymphoma Fund (USA) and the Medical Research Council (UK).

Chapter 6

Discussion

In this dissertation I showed several examples of how formal model definitions of biological processes ensure a consistent interpretation and help to clearly define problems and hypotheses. In Chapter 3 we demonstrated that it is possible to reason about large models of complex processes by exploiting modularity and Monte Carlo model checking. In Chapter 4 and Chapter 5 we took advantage of state space analysis techniques to guide *in vivo* validation of biological hypotheses. However, this work just touches the outer shell of a more ambitious challenge: to define a sound formal framework to represent and "execute" biological knowledge *purely in terms of software*, using Dawkins' words (see Section 1.1 of the Introduction). To better understand the relevance of this challenge, let's consider a statement written in a high-level programming language. For instance the following Python statement:

$a = 1 + 2.$

The result of the evaluation of the this statement is immediately clear; the variable a will eventually contain the value 3. However the low-level operations needed to implement such an evaluation (*e.g.* binary arithmetics, CPU registers manipulations) are much less intuitive and quite specific to the hardware used. Furthermore, the hardware implementation of the low-level software directives requires the involvement of complex electronic systems. Understanding the meaning of the simple statement above, by solely observing the electronic circuitry that computes the evaluation, would be a very difficult task even for an accomplished engineer. Similarly, an accomplished biologist, when interested in the comprehensive behavior of biological systems, should have an adequately abstract language to express and reason about the processes that govern living organisms.

Most of the modelling strategies currently employed to model biological systems (*e.g.* ordinary differential equations, stochastic simulations, Petri nets, boolean networks), without distinction between continuous or discrete, deterministic or stochastic, fine- or coarse-grained, all focus on reproducing molecular state changes. Hence,

they infer the biological effects on the living organism from the distributions of molecules in the simulated underlying chemical reaction system. I argue that, in order to reach a software-like description of biological behaviors, we should be able to abstract from the physical effects of chemical reactions towards the functions accomplished by such chemical changes in a systemic perspective. For instance, from a purely chemical point of view, the binding of a protein in a larger scaffolding protein is a locally defined event that changes the state of these two molecules. However, from a systemic point of view, the same binding event accomplishes several functions, *e.g.* bring the scaffold substrates in close proximity and sequester the scaffold substrates from the cytosol. Continuing on the same example, it is possible to further abstract the low-level systemic functions into higher-level directives; since sequestration reduces the total amount of free scaffold substrates, we could think about the sequestration function as a lower-level implementation of a more abstract "down-regulation" directive. In the same way it would be possible to climb the abstraction stack to the top and most abstract biological function "being alive" (Figure 6.1) or, in more appropriate evolutionary terms, "being reproductive". On this construction of abstraction layers, evolutionary pressure is exercised from the top to the bottom; higher-level functional abstractions tend to be conserved more than lower-level functional abstractions, which in turn tend to be preserved more than their chemical implementations.

Therefore, in order to build a useful abstraction stack for biology, it is necessary to create (i) a set of low-level functions that can capture the dynamics occurring in the "biological hardware", and (ii) a set of high-level languages grounded on the defined low-level function that will allow biologists to comfortably reason on complex system behaviors. Notably,

i biological hardware operates on a chemical level, and it seems possible to define a set of low-level primitives in terms of chemical interactions (*e.g.* phosphorilation and phosphatase, binding and unbinding). Unfortunately, while the exact specifications of the electronic hardware are known in advance, only a subset of all biological primitives is known at this time. However, new processors often implement new primitives without breaking, at most extending, current higher-level abstractions. In the same way, it is reasonable to think that, when new molecular mechanisms are being discovered, they will extend the initial set of biological primitives without necessarily making backward incompatible changes at higher-level of abstractions, preserving the compatibility of already built models.

ii Intuitively, the modelling frameworks presented in Chapter 2 pertain to the high-level layer of the biology abstraction pyramid of Figure 6.1, and the results presented in the following chapters show that existing formal methods can be used, or extended, to capture biological behaviour at a high but useful abstraction level. However, the exact placement of these formalisms in the pyramid is problematic since the mapping of the different layers on each other and of low-level software primitives onto biological hardware is not yet rigorously defined. This is a fundamental and still missing link for the construction of a consistent

chain of abstractions and the development of novel *ad hoc* biological languages.

Furthermore, the design of low-level chemical functions and higher-level languages should not be a simple ontology specification, but it should have an operational semantics attached that allows the execution of models expressed with such languages.

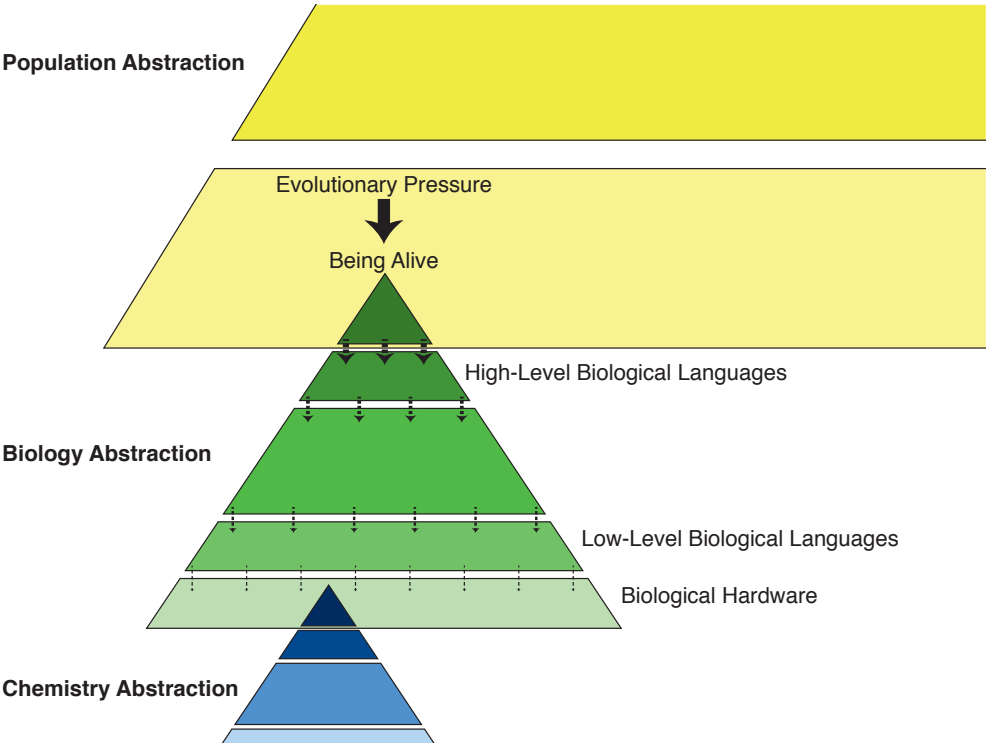


Figure 6.1: Layers of biologically oriented languages are built atop biological hardware. Biological hardware is described using the lower-level chemical abstraction stack. The biological abstraction stack creates a solid ground for reasoning about population dynamics. Evolutionary pressure (*black arrows*) is exercised from the top of the stack and propagate decaying towards the bottom. Similarly, in computer science, low-level electronic abstraction grounds the lowest layer of the computer languages, which in turn build computer networks protocols.

Obviously, the creation of formal and executable languages for biology is a complex task that requires to advance from both directions, bottom-level chemistry knowledge and top-level understanding of systems dynamics. Nevertheless, the construction of a sound chain of abstraction is essential to build composable multi-scale models. It is more and more evident that understanding complex processes such as cancer and differentiation requires extending the focus from the core pathway controlling the process of interest to the impinging neighboring pathways. Hence, the possibility to compose models is the key to build comprehensive descriptions of multiple path-

ways and possibly a whole organism. However, nowadays, each biological model is defined at an arbitrary abstraction level which varies between different models. It is possible, and often necessary, to abstract on multiple axes, *e.g.* time, space, topology, but I reckon that the number of different abstractions is not *per se* a problem. Instead, the problem lies in the difficulty to reconciliate them in a coherent abstraction scheme. Indeed, the current lack of a sound chain of abstraction imposes severe limits to compositionality. A limited compositionality hinders, among others, the possibility to integrate models published by other researchers and, hence, limits the efficiency of systems biology as a research paradigm. Furthermore, it is already necessary, and it will be even more in the future, given the incredible extent and speed of knowledge gain, to reduce the vast amount of complex molecular and systemic interactions to human understandable terms and, therefore, to create a sound and solid abstraction framework for the construction of biological models.

References

Alur, R. & T. A. Henzinger (**1999**). Reactive modules. *Formal Methods in System Design* 15(1), 7–48.

Amberg, D. C., D. J. Burke & J. N. E. T. Strathern (**2005**). *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual* 230. ISBN: 978-087969728-0 (Cold Spring Harbor Laboratory Press) (cit. on p. 51).

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock (**2000**). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1), 25–29. ISSN: 10614036 (cit. on p. 10).

Barnat, J., L. Brim, I. Černá, S. Dražan & D. Šafránek (**2008**). Parallel Model Checking Large-Scale Genetic Regulatory Networks with DiVinE. *Electronic Notes in Theoretical Computer Science* 194(3), 35–50. ISSN: 1571-0661 (cit. on p. 16).

Beck, T. & M. N. Hall (**1999**). The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature* 402(6762), 689–692 (cit. on p. 53).

Beitel, G. J., S. G. Clark & H. R. Horvitz (**1990**). Caenorhabditis elegans ras gene let-60 acts as a switch in the pathway of vulval induction. *Nature* 348, 503–509 (cit. on p. 42).

Belle, A., A. Tanay, L. Bitincka, R. Shamir & E. K. O'Shea (**2006**). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* 103(35), 13004–13009 (cit. on p. 56).

Bernardinello, L., N. Bonzanni, M. Mascheroni & L. Pomello (**2007**). *Modeling Symport/Antiport P Systems with a Class of Hierarchical Petri Nets* in *Membrane Computing* vol. 4860 (Springer-Verlag, Berlin / Heidelberg), 124–137.

Berset, T., E. F. Hoier, G. Battu, S. Canevascini & A. Hajnal (**2001**). Notch inhibition of RAS signaling through MAP kinase phosphatase LIP-1 during C. elegans vulval development. *Science* 291(5506), 1055 (cit. on p. 41).

Berset, T., E. F. Hoier & A. Hajnal (**2005**). The C. elegans homolog of the mammalian tumor suppressor Dep-1/Scc1 inhibits EGFR signaling to regulate binary cell fate decisions. *Genes & Development* 19(11), 1328 (cit. on p. 41).

Berthomieu, B., P.-O. Ribet & F. Vernadat (**2004**). The tool TINA – Construction of abstract state spaces for Petri nets and time Petri nets. *International Journal of Production Research* 42(14), 2741–2756.

Bonzanni, N., E. Krepska, K. A. Feenstra, W. Fokkink, T. Kielmann, H. Bal & J. Heringa (**2009a**). Executing multicellular differentiation: quantitative predictive modelling of C. elegans vulval development. *Bioinformatics* 25(16), 2049–2056 (cit. on p. 51).

Bonzanni, N., K. A. Feenstra, W. Fokkink & E. Krepska (**2009b**). *What Can Formal Methods Bring to Systems Biology?* in *Formal Methods* vol. 5850 (Springer-Verlag, Berlin / Heidelberg), 16–22 (cit. on p. 10).

Bonzanni, N., N. Zhang, S. G. Oliver & J. Fisher (**2011**). The role of proteosome-mediated proteolysis in modulating potentially harmful transcription factor activity in Saccharomyces cerevisiae. *Bioinformatics* 27(13), I283–I287.

Bonzanni, N., A. Garg, S. D. Foster, N. K. Wilson, S. Kinston, D. Miranda-Saavedra, J. Heringa, A. Feenstra, I. Xenarios & B. Göttgens (**2012**). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Submitted.*

Burhard, H. D. (**1983**). *On priorities of parallelism: Petri nets under the maximum firing strategy* in *Logics of Programs and Their Applications* vol. 148 (Springer-Verlag, Berlin / Heidelberg), 86–97. ISBN: 3-540-11981-7 (cit. on pp. 15, 18, 32).

Calder, M., S. Gilmore, J. Hillston & V. Vyshemirsky (**2010**). *Formal methods for biochemical signalling pathways* in *Formal Methods: State of the Art and New Directions* (eds Boca, P., J. P. Bowen & J. Siddiqi) (Springer-Verlag), 185–215. ISBN: 978-1-84882-735-6.

Chambers, S. M., N. C. Boles, K. Y. Lin, M. P. Tierney, T. V. Bowman, S. B. Bradfute, A. J. Chen, A. A. Merchant, O. Sirin, D. C. Weksberg, M. G. Merchant, C. J. Fisk, C. A. Shaw & M. A. Goodell (**2007**). Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* 1(5), 578–91 (cit. on p. 68).

Chan, W. Y., G. A. Follows, G. Lacaud, J. E. Pimanda, J. R. Landry, S. Kinston, K. Knezevic, S. Piltz, I. J. Donaldson, L. Gambardella, F. Sablitzky, A. R. Green, V. Kouskoff & B. Göttgens (**2007**). The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. *Blood* 109(5), 1908–16 (cit. on p. 64).

Chaouiya, C. (**2007**). Petri net modelling of biological networks. *Briefings in Bioinformatics* 8(4), 210 (cit. on p. 32).

Chaouiya, C., E. Remy, P. Ruet & D. Thieffry (**2004**). *Qualitative Modelling of Genetic Networks: From Logical Regulatory Graphs to Standard Petri Nets* in *Applications and Theory of Petri Nets 2004* (eds Cortadella, J. & W. Reisig) 137–156 (Springer-Verlag, Berlin / Heidelberg). ISBN: 978-3-540-22236-1 (cit. on p. 21).

Chaouiya, C., E. Remy & D. Thieffry (**2006**). Qualitative Petri net modelling of genetic networks. *Transactions on Computational Systems Biology VI,* 95–112 (cit. on p. 51).

Chen, L, G Qi-Wei, M. Nakata & H. Matsuno (**2007**). Modelling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets. *Journal of Biosciences* 32(1), 113–127.

Chen, N. & I. Greenwald (**2004**). The lateral signal for LIN-12/Notch in C. elegans vulval development comprises redundant secreted and transmembrane DSL proteins. *Developmental Cell* 6(2), 183–192 (cit. on p. 30).

Chickarmane, V., T. Enver & C. Peterson (**2009**). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol* 5(1), e1000268 (cit. on p. 70).

Clarke, D., D. Costa & F. Arbab (**2006**). *Modelling coordination in biological systems* in *Leveraging Applications of Formal Methods* vol. 4313 (Springer-Verlag), 9–25.

Crombach, A. & P. Hogeweg (**2008**). Evolution of Evolvability in Gene Regulatory Networks. *PLoS Computational Biology* 4(7), 13 (cit. on p. 14).

Cui, M., J. Chen, T. R. Myers, B. J. Hwang, P. W. Sternberg, I. Greenwald & M. Han (**2006**). SynMuv genes redundantly inhibit lin-3/EGF expression to prevent inappropriate vulval induction in C. elegans. *Developmental Cell* 10(5), 667–672 (cit. on p. 41).

DAS-3 Steering Group (**2007**). *The Distributed ASCI Supercomputer 3* <http://www.cs.vu.nl/das3/> (2007).

Davidson, E. H. (**2006**). *The regulatory genome: gene regulatory networks in development and evolution* **2,** 304 (Academic Press) (cit. on pp. 62, 63).

Dawkins, R. (**1976**). in (eds Bateson, P. P. G. & R. A. Hinde) 1, 548 (Cambridge University Press, Cambridge, UK) (cit. on p. 10).

Dematté, L., C. Priami, A. Romanel & O. Soyer (**2007**). *A formal and integrated framework to simulate evolution of biological pathways* in *Computational Methods in Systems Biology* vol. 4695 (Springer-Verlag), 106–120.

Dembla-Rajpal, N., R. Seipelt, Q. Wang & B. C. Rymond (**2004**). Proteasome inhibition alters the transcription of multiple yeast genes. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1680(1), 34–45 (cit. on p. 58).

Dobzhansky, T. (**1973**). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* 35(3), 125–129 (cit. on p. 14).

Donaldson, I. J., M. Chapman, S. Kinston, J. R. Landry, K. Knezevic, S. Piltz, N. Buckley, A. R. Green & B. Göttgens (**2005**). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Human Molecular Genetics* 14(5), 595–601 (cit. on p. 63).

Durchschlag, E., W. Reiter, G. Ammerer & C. Schüller (**2004**). Nuclear Localization Destabilizes the Stress-regulated Transcription Factor Msn2. *Journal of Biological Chemistry* 279(53), 55425–55432.

Fisher, J. & T. A. Henzinger (**2007**). Executable cell biology. *Nature Biotechnology* 25(11), 1239–1249 (cit. on pp. 10, 15, 50, 97, 99).

Fisher, J. & N. Piterman (**2010**). The executable pathway to biological networks. *Briefings in Functional Genomics* 9(1), 79–92 (cit. on p. 50).

Fisher, J., N. Piterman, E Hubbard, M. J. Stern & D. Harel (**2005**). Computational insights into Caenorhabditis elegans vulval development. *Proceedings of the National Academy of Sciences of the United States of America* 102(6), 1951 (cit. on pp. 31, 40–42).

Fisher, J., N. Piterman, A. Hajnal & T. A. Henzinger (**2007**). Predictive modeling of signaling crosstalk during C. elegans vulval development. *PLoS Computational Biology* 3(5), e92 (cit. on pp. 31, 40).

Fisher, J., T. A. Henzinger, M. Mateescu & N. Piterman (**2008**). *Bounded asynchrony: Concurrency for modeling cell-cell interactions* in *Formal Methods in Systems Biology* (ed Fisher, J.) vol. 5054 (Springer-Verlag), 17–32 (cit. on p. 15).

Foster, S. D., S. H. Oram, N. K. Wilson & B. Göttgens (**2009**). From genes to cells to tissues–modelling the haematopoietic system. *Molecular Biosystems* 5(12), 1413–20 (cit. on p. 63).

Garg, A., A. Di Cara, I. Xenarios, L. Mendoza & G. De Micheli (**2008**). Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* 24(17), 1917–25 (cit. on p. 76).

Garí, E., L. Piedrafita, M. Aldea & E. Herrero (**1997**). A Set of Vectors with a Tetracycline-Regulatable Promoter System for Modulated Gene Expression in Saccharomyces cerevisiae. *Yeast* 13(9), 837–848 (cit. on p. 51).

Gasch, A. & M. Werner-Washburne (**2002**). The genomics of yeast responses to environmental stress and starvation. *Functional & Integrative Genomics* 2(4), 181–192 (cit. on p. 54).

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein & P. O. Brown (**2000**). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell* 11(12), 4241–4257 (cit. on p. 58).

Genrich, H., R Küffner & K. Voss (**2001**). Executable Petri net models for the analysis of metabolic pathways. *International Journal on Software Tools for Technology Transfer* 3(4), 394–404.

Georgescu, C., W. J. Longabaugh, D. D. Scripture-Adams, E. S. David-Fung, M. A. Yui, M. A. Zarnegar, H. Bolouri & E. V. Rothenberg (**2008**). A gene regulatory network armature for T lymphocyte specification. *Proceedings of the National Academy of Sciences of the United States of America* 105(51), 20100–5 (cit. on p. 63).

Gilbert, D. & M. Heiner (**2006**). *From Petri nets to differential equations-an integrative approach for biochemical network analysis* in *International Conference on Applications and Theory of Petri Nets and Other Models of Concurrency* vol. 4024 (Springer-Verlag), 181–200.

Gilbert, D., M. Heiner & S. Lehrack (**2007**). *A unifying framework for modelling and analysing biochemical pathways using Petri nets* in *Computational Methods in Systems Biology* vol. 4695 (Springer-Verlag), 200–216 (cit. on pp. 18, 32).

Gillespie, D. T. (**1977**). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81(25), 2340–2361.

Giurumescu, C. A., P. W. Sternberg & A. R. Asthagiri (**2006**). Intercellular coupling amplifies fate segregation during Caenorhabditis elegans vulval development. *Proceedings of the National Academy of Sciences of the United States of America* 103(5), 1331 (cit. on pp. 31, 39).

Görner, W., E. Durchschlag, M. T. Martinez-Pastor, F. Estruch, G. Ammerer, B. Hamilton, H. Ruis & C. Schüller (**1998**). Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes & Development* 12(4), 586–597 (cit. on p. 53).

Goss, P. J. E. & J. Peccoud (**1998**). Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proceedings of the National Academy of Sciences of the United States of America* 95(12), 6750 (cit. on p. 32).

Göttgens, B. (**2004**). Transcriptional regulation of haematopoiesis. *Vox Sanguinis* 87 Suppl1, 15–9 (cit. on p. 62).

Göttgens, B., F. McLaughlin, E. O. Bockamp, J. L. Fordham, C. G. Begley, K. Kosmopoulos, A. G. Elefanty & A. R. Green (**1997**). Transcription of the SCL gene in erythroid and CD34 positive primitive myeloid cells is controlled by a complex network of lineage-restricted chromatin-dependent and chromatin-independent regulatory elements. *Oncogene* 15(20), 2419–28 (cit. on p. 76).

Göttgens, B., A. Nastos, S. Kinston, S. Piltz, E. C. Delabesse, M. Stanley, M. J. Sanchez, A. Ciau-Uitz, R. Patient & A. R. Green (**2002**). Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO Journal* 21(12), 3039–50 (cit. on p. 63).

Göttgens, B., C. Broccardo, M. J. Sanchez, S. Deveaux, G. Murphy, J. R. Gothert, E. Kotsopoulou, S. Kinston, L. Delaney, S. Piltz, L. M. Barton, K. Knezevic, W. N. Erber, C. G. Begley, J. Frampton & A. R. Green (**2004**). The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. *Molecular and Cellular Biology* 24(5), 1870–83 (cit. on p. 63).

Grass, J. A., M. E. Boyer, S. Pal, J. Wu, M. J. Weiss & E. H. Bresnick (**2003**). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proceedings of the National Academy of Sciences of the United States of America* 100(15), 8811–6 (cit. on p. 71).

Greenwald, I., P. W. Sternberg & H. R. Horvitz (**1983**). The lin-12 locus specifies cell fates in Caenorhabditis elegans. *Cell* 34(2), 435–444 (cit. on p. 41).

Grunwald, S., A. Speer, J. Ackermann & I. Koch (**2008**). Petri net modelling of gene regulation of the Duchenne muscular dystrophy. *Biosystems* 92(2), 189–205. ISSN: 0303-2647 (cit. on pp. 33, 34).

Gsponer, J., M. E. Futschik, S. A. Teichmann & M. M. Babu (**2008**). Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science* 322(5906), 1365–1368 (cit. on p. 56).

Gutenkunst, R. N., J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers & J. P. Sethna (**2007**). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology* 3(10), e189 (cit. on p. 38).

Harel, D. (**2004**). *A Grand Challenge for Computing: Towards Full Reactive Modeling of a Multi-cellular Animal* in *Verification, Model Checking, and Abstract Interpretation* (eds Steffen, B. & G. Levi) 39–60 (Springer-Verlag, Berlin / Heidelberg). ISBN: 978-3-540-20803-7 (cit. on p. 14).

Heyworth, C., S. Pearson, G. May & T. Enver (**2002**). Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *EMBO Journal* 21(14), 3770–81 (cit. on p. 70).

Hirata, H., S. Yoshiura, T. Ohtsuka, Y. Bessho, T. Harada, K. Yoshikawa & R. Kageyama (**2002**). Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science* 298(5594), 840–3 (cit. on p. 74).

Hofestädt, R & S Thelen (**1998**). Quantitative modeling of biochemical networks. *In Silico Biology* 1(1), 39–53.

Hu, M., D. Krause, M. Greaves, S. Sharkis, M. Dexter, C. Heyworth & T. Enver (**1997**). Multilineage gene expression precedes commitment in the hemopoietic system. *Genes & Development* 11(6), 774–85 (cit. on p. 75).

Hu, Z., P. J. Killion & V. R. Iyer (**2007**). Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics* 39(5), 683–7 (cit. on p. 63).

Huang, S. (**2009**). Non-genetic heterogeneity of cells in development: more than just noise. *Development* 136(23), 3853–62 (cit. on p. 75).

Kam, N., D. Harel, H. Kugler, R. Marelly, A. Pnueli, E. Hubbard & M. J. Stern (**2003**). *Formal modeling of C. elegans development: a scenario-based approach* in *Computational Methods in Systems Biology* vol. 2602 (Springer-Verlag), 4–20 (cit. on pp. 16, 31).

Kanehisa, M. & S. Goto (**2000**). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1), 27–30.

Karlebach, G. & R. Shamir (**2008**). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* 9(10), 770–80 (cit. on p. 63).

Karp, X. & V. Ambros (**2005**). Encountering microRNAs in cell fate signaling. *Science* 310(5752), 1288 (cit. on p. 46).

Kauffman, S., C. Peterson, B. Samuelsson & C. Troein (**2003**). Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America* 100(25), 14796–9 (cit. on p. 76).

Kimble, J. (**1981**). Alterations in cell lineage following laser ablation of cells in the somatic gonad of Caenorhabditis elegans. *Developmental Biology* 87(2), 286–300 (cit. on p. 41).

Klamt, S., J. Saez-Rodriguez, J. A. Lindquist, L. Simeoni & E. D. Gilles (**2006**). A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics* 7, 56 (cit. on p. 76).

Kobayashi-Osaki, M., O. Ohneda, N. Suzuki, N. Minegishi, T. Yokomizo, S. Takahashi, K.-C. Lim, J. D. Engel & M. Yamamoto (**2005**). GATA Motifs Regulate Early Hema-

topoietic Lineage-Specific Expression of the Gata2 Gene. *Molecular and Cellular Biology* 25(16), 7005–7020 (cit. on p. 63).

Koch, I. & M. Heiner (**2008**). in. *Analysis of biological networks* (eds Junker, B. & F. Schreiber) 139–180 (Wiley Online Library). ISBN: 0470041447, 9780470041444 (cit. on p. 32).

Koch, I., B. H. Junker & M. Heiner (**2005**). Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics* 21(7), 1219–1226. ISSN: 1367-4803.

Krepska, E. (**2012**). *Towards Big Biology: High-Performance Verification of Large Concurrent Systems* PhD thesis (VU University Amsterdam) (cit. on p. 27).

Krepska, E., N. Bonzanni, A. Feenstra, W. Fokkink, T. Kielmann, H. Bal & J. Heringa (**2008**). *Design issues for qualitative modelling of biological cells with Petri nets* in *Formal Methods in Systems Biology* vol. 5054 (Springer-Verlag), 48–62 (cit. on pp. 31, 32).

Kruse, E. A., S. J. Loughran, T. M. Baldwin, E. C. Josefsson, S. Ellis, D. K. Watson, P. Nurden, D. Metcalf, D. J. Hilton, W. S. Alexander & B. T. Kile (**2009**). Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. *Proc Natl Acad Sci U S A* 106(33), 13814–9 (cit. on p. 64).

Kuhn, T. S. (**1996**). *The Structure of Scientific Revolutions* (ed Nickles, T.) **3,** 212. ISBN: 0226458083 (University of Chicago Press) (cit. on p. 9).

Kulessa, H., J. Frampton & T. Graf (**1995**). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes & Development* 9(10), 1250–62 (cit. on p. 70).

Lahav, G., N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A. J. Levine, M. B. Elowitz & U. Alon (**2004**). Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nature Genetics* 36(2), 147–50 (cit. on p. 74).

Landry, J.-R., S. Kinston, K. Knezevic, M. F. T. R. de Bruijn, N. Wilson, W. T. Nottingham, M. Peitz, F. Edenhofer, J. E. Pimanda, K. Ottersbach & B. Göttgens (**2008**). Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver. *Blood* 111(6), 3005–3014 (cit. on p. 63).

Lazebnik, Y. (**2002**). Can a biologist fix a radio? – Or, what I learned while studying apoptosis. *Cancer Cell* 2(3), 179–182. ISSN: 15356108 (cit. on p. 9).

Li, C., M. Nagasaki & K. Ueno (**2009**). Simulation-based model checking approach to cell fate specification during Caenorhabditis elegans vulval development by hybrid functional Petri net with. *BMC Systems Biology* 3(1), 42. ISSN: 1752-0509 (cit. on p. 31).

Mar, J. C. & J. Quackenbush (**2009**). Decomposition of gene expression state space trajectories. *PLoS Comput Biol* 5(12), e1000626.

Martínez-Pastor, M. T., G Marchler, C. Schüller, A Marchler-Bauer, H Ruis & F. Estruch (**1996**). The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *The EMBO Journal* 15(9), 2227–2235.

Matsuno, H., A. Doi, M. Nagasaki & S. Miyano (**2000**). *Hybrid Petri net representation of gene regulatory network* in *Pacific Symposium on Biocomputing* vol. 5 (World Scientific Publishing), 338–349 (cit. on p. 32).

Matsuno, H., C. Li & S. Miyano (**2006**). Petri net based descriptions for systematic understanding of biological pathways. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E89-A(11), 3166–3174. ISSN: 0916-8508 (cit. on p. 32).

Mendoza, L. & I. Xenarios (**2006**). A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theoretical Biology and Medical Modelling* 3, 13 (cit. on p. 76).

Mersman, D. P., H.-N. Du, I. M. Fingerman, P. F. South & S. D. Briggs (**2009**). Polyubiquitination of the demethylase Jhd2 controls histone methylation and gene expression. *Genes & Development* 23(8), 951–962.

Miranda-Saavedra, D. & B. Göttgens (**2008**). Transcriptional regulatory networks in haematopoiesis. *Curr Opin Genet Dev* 18(6), 530–5 (cit. on p. 62).

Moya, A., N. Krasnogor, J. Peretó & A. Latorre (**2009**). Goethe's dream. Challenges and opportunities for synthetic biology. *EMBO Reports* 10(S1), S28–S32 (cit. on pp. 9, 10).

Nagasaki, M, A Saito, A Doi, H Matsuno & S Miyano (**2009**). *Foundations of Systems Biology. Using Cell Illustrator® and Pathway Databases* ISBN: 978-1-84882-023-4 (Springer-Verlag, London) (cit. on p. 16).

Noble, D. (**2006**). *The Music of Life: Biology Beyond Genes* 176. ISBN: 0199295735 (Oxford University Press).

Nottingham, W. T., A. Jarratt, M. Burgess, C. L. Speck, J.-F. Cheng, S. Prabhakar, E. M. Rubin, P.-S. Li, J. Sloane-Stanley, J. Kong-a San & M. F. T. R. de Bruijn (**2007**). Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. *Blood* 110(13), 4188–4197 (cit. on p. 63).

Ogilvy, S., R. Ferreira, S. G. Piltz, J. M. Bowen, B. Göttgens & A. R. Green (**2007**). The SCL +40 enhancer targets the midbrain together with primitive and definitive hematopoiesis and is regulated by SCL and GATA proteins. *Molecular and Cellular Biology* 27(20), 7206–19 (cit. on p. 64).

Okuno, Y., G. Huang, F. Rosenbauer, E. K. Evans, H. S. Radomska, H. Iwasaki, K. Akashi, F. Moreau-Gachelin, Y. Li, P. Zhang, B. Göttgens & D. G. Tenen (**2005**). Potential Autoregulation of Transcription Factor PU.1 by an Upstream Regulatory Element. *Molecular and Cellular Biology* 25(7), 2832–2845.

Peccoud, J., T. Courtney & W. H. Sanders (**2007**). Möbius: an integrated discrete-event modeling environment. *Bioinformatics* 23(24), 3412–3414 (cit. on p. 46).

Pedruzzi, I., N. Burckert, P. Egger & C. De Virgilio (**2000**). Saccharomyces cerevisiae Ras/cAMP pathway controls post-diauxic shift element-dependent transcription through the zinc finger protein Gis1. *The EMBO Journal* 19(11), 2569–2579 (cit. on p. 54).

Pedruzzi, I., F. Dubouloz, E. Cameroni, V. Wanke, J. Roosen, J. Winderickx & C. De Virgilio (**2003**). TOR and PKA Signaling Pathways Converge on the Protein Kinase

Rim15 to Control Entry into G0. *Molecular Cell* 12(6), 1607–1613 (cit. on pp. 50, 53).

Peleg, M., D. Rubin & R. B. Altman (**2005**). Using Petri net tools to study properties and dynamics of biological systems. *Journal of the American Medical* 12(2), 181–199. ISSN: 1067-5027 (cit. on p. 32).

Petri, C. A. (**1962**). *Kommunikation mit Automaten* PhD thesis (Technische Universität Darmstadt) (cit. on p. 17).

Phillips, A. & L. Cardelli (**2007**). *Efficient, correct simulation of biological processes in the stochastic pi-calculus* in *Computational Methods in Systems Biology* vol. 4695 (Springer-Verlag), 184–199.

Pimanda, J. E., W. Y. Chan, I. J. Donaldson, M. Bowen, A. R. Green & B. Göttgens (**2006**). Endoglin expression in the endothelium is regulated by Fli-1, Erg, and Elf-1 acting on the promoter and a -8-kb enhancer. *Blood* 107(12), 4737–45 (cit. on p. 64).

Pimanda, J. E., K. Ottersbach, K. Knezevic, S. Kinston, W. Y. Chan, N. K. Wilson, J. R. Landry, A. D. Wood, A. Kolb-Kokocinski, A. R. Green, D. Tannahill, G. Lacaud, V. Kouskoff & B. Göttgens (**2007**). Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proceedings of the National Academy of Sciences of the United States of America* 104(45), 17692–7 (cit. on pp. 63, 64).

Rainis, L., T. Toki, J. E. Pimanda, E. Rosenthal, K. Machol, S. Strehl, B. Göttgens, E. Ito & S. Izraeli (**2005**). The proto-oncogene ERG in megakaryoblastic leukemias. *Cancer Res* 65(17), 7596–602 (cit. on p. 64).

Ramos, C. A., T. A. Bowman, N. C. Boles, A. A. Merchant, Y. Zheng, I. Parra, S. A. Fuqua, C. A. Shaw & M. A. Goodell (**2006**). Evidence for diversity in transcriptional profiles of single hematopoietic stem cells. *PLoS Genetics* 2(9), e159 (cit. on pp. 68, 70).

Ratzer, A. V., L. Wells, H. M. Lassen, M. Laursen, J. F. Qvortrup, M. S. Stissing, M. Westergaard, S. Christensen & K. Jensen (**2003**). *CPN tools for editing, simulating, and analysing coloured Petri nets* in *Proceedings of the 24th international conference on Applications and theory of Petri nets* (Springer-Verlag, Berlin, Heidelberg), 450–462 (cit. on p. 46).

Reddy, V. N., M. N. Liebman & M. L. Mavrovouniotis (**1996**). Qualitative analysis of biochemical reaction systems. *Computers in biology and medicine* 26(1), 9–24.

Regev, A. & E. Shapiro (**2002**). Cellular abstractions: Cells as computation. *Nature* 419(6905), 343 (cit. on p. 15).

Regev, A., W. Silverman & E. Shapiro (**2001**). *Representation and simulation of biochemical processes using the pi-calculus process algebra* in *Pacific Symposium of Biocomputing* (PSB), 459–470.

Regev, A., E. Panina, W. Silverman, L. Cardelli & E. Shapiro (**2004**). BioAmbients: an abstraction for biological compartments. *Theoretical Computer Science* 325(1), 141–167 (cit. on p. 16).

Lectures on Petri Nets I: Basic Models (**1998**) (eds Reisig, W. & G. Rozenberg) 683. ISBN: 978-3-540-65306-6 (Springer-Verlag) (cit. on pp. 17, 51).

Rice, P., I. Longden & A. Bleasby (**2000**). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genetics* 16(6), 276–7 (cit. on p. 58).

Roeder, I. & I. Glauche (**2006**). Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. *Journal of Theoretical Biology* 241(4), 852–65 (cit. on p. 70).

Rogers, S, R Wells & M Rechsteiner (**1986**). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* 234(4774), 364–8 (cit. on p. 58).

Sackmann, A., M. Heiner & I. Koch (**1998**). Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics* 1491(1), 683–683.

— (**2006**). Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics* 7(1), 482. ISSN: 1471-2105 (cit. on p. 34).

Sadot, A., J. Fisher, D. Barak, Y. Admanit, M. Stern, J. A. Hubbard & D. Harel (**2008**). Toward Verified Biological Models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5(2), 223–234 (cit. on p. 15).

Schuh, A. H., A. J. Tipping, A. J. Clark, I. Hamlett, B. Guyot, F. J. Iborra, P. Rodriguez, J. Strouboulis, T. Enver, P. Vyas & C. Porcher (**2005**). ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Molecular and Cellular Biology* 25(23), 10235–50 (cit. on p. 65).

Shaye, D. D. & I. Greenwald (**2002**). Endocytosis-mediated downregulation of LIN-12/Notch upon Ras activation in Caenorhabditis elegans. *Nature* 420(6916), 686–690 (cit. on pp. 30, 33, 39, 44).

— (**2005**). LIN-12/Notch trafficking and regulation of DSL ligand activity during vulval induction in Caenorhabditis elegans. *Development* 132(22), 5081 (cit. on p. 39).

Sieweke, M. H. & T. Graf (**1998**). A transcription factor party during blood cell differentiation. *Current Opinion in Genetics & Development* 8(5), 545–51 (cit. on p. 63).

Simao, E, E Remy, D Thieffry & C. Chaouiya (**2005**). Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in *E. coli*. *Bioinformatics* 21(2), 190–196.

Sinclair, A. M., B. Göttgens, L. M. Barton, M. L. Stanley, L. Pardanaud, M. Klaine, M. Gering, S. Bahn, M. Sanchez, A. J. Bench, J. L. Fordham, E. Bockamp & A. R. Green (**1999**). Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Developmental Biology* 209(1), 128–42 (cit. on p. 76).

Sopko, R., D. Huang, N. Preston, G. Chua, B. Papp, K. Kafadar, M. Snyder, S. G. Oliver, M. Cyert, T. R. Hughes, C. Boone & B. Andrews (**2006**). Mapping Pathways and Phenotypes by Systematic Gene Overexpression. *Molecular Cell* 21(3), 319–330 (cit. on p. 56).

Spooner, C. J., J. X. Cheng, E. Pujadas, P. Laslo & H. Singh (**2009**). A recurrent network involving the transcription factors PU.1 and Gfi1 orchestrates innate and adaptive immune cell fates. *Immunity* 31(4), 576–86 (cit. on p. 63).

Srivastava, R, M. S. Peterson & W. E. Bentley (**2001**). Stochastic kinetic analysis of the Escherichia coli stress circuit using σ-32-targeted antisense. *Biotechnology and Bioengineering* 75(1), 120–129.

Srivastava, R, L You, J Summers & J Yin (**2002**). Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology* 218(3), 309–321.

Steggles, L, R. Banks & A. Wipat (**2006**). *Modelling and analysing genetic networks: From Boolean networks to Petri nets* in *Computational Methods in Systems Biology* vol. 4210 (Springer-Verlag), 127–141.

Sternberg, P. W. (**2005**). in. *WormBook* (ed The C. elegans Research Community) () (cit. on pp. 30, 40, 42).

Sternberg, P. W. & M. Han (**1998**). Genetics of RAS signaling in C. elegans. *Trends in Genetics* 14(11), 466–472.

Sternberg, P. W. & H. R. Horvitz (**1986**). Pattern formation during vulval development in C. elegans. *Cell* 44(5), 761–772 (cit. on pp. 30, 31).

— (**1989**). The combined action of two intercellular signaling pathways specifies three cell fates during vulval induction in C. elegans. *Cell* 58(4), 679–693 (cit. on pp. 31, 41, 42).

Sulston, J. E. & H. R. Horvitz (**1977a**). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology* 56(1), 110–156 (cit. on p. 41).

— (**1977b**). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology* 56(1), 110–56. ISSN: 00121606 (cit. on p. 41).

Sun, X. & P. Hong (**2007**). Computational modeling of Caenorhabditis elegans vulval induction. *Bioinformatics* 23(13), i499–i507 (cit. on pp. 31, 40).

Sundaram, M. V. (**2004**). Vulval Development: The Battle between Ras and Notch. *Current Biology* 14(8), R311–R313 (cit. on p. 30).

Swiers, G., R. Patient & M. Loose (**2006**). Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. *Developmental Biology* 294(2), 525–40 (cit. on p. 63).

Tarjan, R. (**1975**). Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing* 1(2), 146–160 (cit. on p. 27).

The Concell Group (**2009**). *The Concell Website* <http://www.cs.vu.nl/concell/celegans> (2009).

Tipping, A. J., C. Pina, A. Castor, D. Hong, N. P. Rodrigues, L. Lazzari, G. E. May, S. E. Jacobsen & T. Enver (**2009**). High GATA-2 expression inhibits human hematopoietic stem and progenitor cell function by effects on cell cycle. *Blood* 113(12), 2661–72 (cit. on p. 75).

Tofts, C. (**1992**). Describing social insect behaviour using process algebra. *Transactions of the Society for Computer Simulation* 9(4), 227–283 (cit. on p. 16).

Urban, J., A. Soulard, A. Huber, S. Lippman, D. Mukhopadhyay, O. Deloche, V. Wanke, D. Anrather, G. Ammerer, H. Riezman, J. R. Broach, C. De Virgilio, M. N. Hall & R. Loewith (**2007**). Sch9 Is a Major Target of TORC1 in Saccharomyces cerevisiae. *Molecular Cell* 26(5), 663–674.

Vyas, P., M. A. McDevitt, A. B. Cantor, S. G. Katz, Y. Fujiwara & S. H. Orkin (**1999**). Different sequence requirements for expression in erythroid and megakaryocytic

cells within a regulatory element upstream of the GATA-1 gene. *Development* 126(12), 2799–2811 (cit. on p. 63).

Wang, Y., C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag & P. O. Brown (**2002**). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* 99(9), 5860–5865 (cit. on p. 56).

Wanke, V., I. Pedruzzi, E. Cameroni, F. Dubouloz & C. De Virgilio (**2005**). Regulation of G0 entry by the Pho80-Pho85 cyclin-CDK complex. *The EMBO Journal* 24(24), 4271–4278 (cit. on p. 53).

Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton & D. T. Jones (**2004**). Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* 337(3), 635–645 (cit. on p. 56).

Warren, L., D. Bryder, I. L. Weissman & S. R. Quake (**2006**). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences of the United States of America* 103(47), 17807–12 (cit. on p. 74).

Wei, M., P. Fabrizio, J. Hu, H. Ge, C. Cheng, L. Li & V. D. Longo (**2008**). Life Span Extension by Calorie Restriction Depends on Rim15 and Transcription Factors Downstream of Ras/PKA, Tor, and Sch9. *PLoS Genetics* 4(1), e13.

Westerhoff, H. V. & B. Palsson (**2004**). The evolution of molecular biology into systems biology. *Nature Biotechnology* 22(10), 1249–1252. ISSN: 10870156 (cit. on p. 9).

Wilson, N. K., D. Miranda-Saavedra, S. Kinston, N. Bonadies, S. D. Foster, F. Calero-Nieto, M. A. Dawson, I. J. Donaldson, S. Dumon, J. Frampton, R. Janky, X. H. Sun, S. A. Teichmann, A. J. Bannister & B. Göttgens (**2009**). The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood* 113(22), 5456–65 (cit. on pp. 63–66, 76).

Wilson, W. A. & P. J. Roach (**2002**). Nutrient-regulated protein kinases in budding yeast. *Cell* 111(2), 155–158 (cit. on p. 50).

Wu, J., N. Zhang, A. Hayes, K. Panoutsopoulou & S. G. Oliver (**2004**). Global analysis of nutrient control of gene expression in Saccharomyces cerevisiae during growth and starvation. *Proceedings of the National Academy of Sciences of the United States of America* 101(9), 3148–3153 (cit. on p. 58).

Yoo, A. S. & I. Greenwald (**2005**). LIN-12/Notch Activation Leads to MicroRNA-Mediated Down-Regulation of Vav in C. elegans. *Science* 310(5752), 1330–1333 (cit. on pp. 31, 33, 40, 45, 46).

Yoo, A. S., C. Bais & I. Greenwald (**2004**). Crosstalk Between the EGFR and LIN-12/Notch Pathways in C. elegans Vulval Development. *Science* 303(5658), 663–666 (cit. on pp. 31, 41, 43, 44).

Zhang, N. & S. G. Oliver (**2010**). The Transcription Activity of Gis1 Is Negatively Modulated by Proteasome-mediated Limited Proteolysis. *Journal of Biological Chemistry* 285(9), 6465–6476 (cit. on pp. 51, 53, 54).

Zhang, N., J. Wu & S. G. Oliver (**2009**). Gis1 is required for transcriptional reprogramming of carbon metabolism and the stress response during transition into stationary phase in yeast. *Microbiology* 155(5), 1690–1698 (cit. on p. 50).

Summary

Towards Executable Biology

Exceptional scientific breakthroughs of the last century and the advent of high-throughput technologies in the early 2000s have catapulted molecular biology in the realm of systems biology. Systems biology inherited the denotational language proper of control theory. However, the necessity of a new language for biology, able to capture and express biological processes on the level of *software explanation*, is clearly perceived by the biological community.

The need to define a framework built upon complementary and yet coherent formal languages for systems biology repeatedly emerged during my studies. In this dissertation we showed several examples of how formal model definitions of biological processes ensure a consistent interpretation and help to clearly define problems and hypotheses. In Chapter 2, we explain the notion of executable models for biological processes as introduced by Fisher et al. and we present two formalism based on Petri nets. The first formalism was built to model signalling networks which analysis is based on simulations and Monte Carlo model checking. The second formalism focuses on gene regulatory networks and exploring the model state space in search of attractors.

In Chapter 3, we demonstrate a large scale application of our Petri net formalism for signal transduction to multi-cellular pattern formation. Our modelling approach to the well-studied process of *C. elegans* vulval development, showing that our model correctly reproduced a large set of *in vivo* experiments with statistical accuracy. Also Chapter 4 focuses on signalling networks. We investigated the effect of proteolysis after nutrient starvation in *S. cervisiae*. Particularly, we showed how computational models, bioinformatics analyses, and *in vivo* observations can be integrated in order to formulate and validate novel biological hypotheses.

The last case study is presented in Chapter 5. We constructed a regulatory network model based on the functionality of cis-regulatory elements in order to generate

fundamental insights into cellular fate differentiation during haematopoiesis. Particularly, we took advantage of state space analysis techniques, explained in Chapter 2, to guide *in vivo* validation of the novel inhibitory link between the proteins Gata1 and Fli1.

Finally, in Chapter 6, I argue that in order to reach a software-like description of biological behaviors, we should be able to abstract from the physical effects of chemical reactions towards the functions accomplished by such chemical changes in a systemic perspective. It is already necessary, and it will be even more in the future, given the incredible extent and speed of knowledge gain, to reduce the vast amount of complex molecular and systemic interactions to human understandable terms and, therefore, to create a sound and solid abstraction framework for the construction of biological models.

Samenvatting

Naar Executeerbare Biologie

Buitengewone wetenschappelijke doorbraken in de vorige eeuw en de opkomst van *high-throughput* technieken in de afgelopen tien jaar hebben de moleculaire biologie binnen het bereik van de systeembiologie gebracht. Systeembiologie heeft de essentie van de denotationele taal voor controletheorie overgenomen. De biologische gemeenschap is zich echter sterk bewust van de noodzaak voor een nieuwe taal om biologische processen te beschrijven op het nivo van *software explanation*.

Gedurende mijn onderzoek had ik herhaaldelijk een raamwerk nodig dat voortbouwt op de complementaire maar toch coherente formele talen van de systeembiologie. Dit proefschrift bevat verscheidene voorbeelden van biologische processen waarbij formele modellen een consistente interpretatie waarborgen, en helpen om problemen en hypotheses helder te definiëren. Hoofdstuk 2 presenteert de notie van executeerbare modellen voor biologische processen zoals geïntroduceerd door Fisher & Henzinger, en twee formalismen gebaseerd op Petri-netten. Het eerste formalisme heeft tot doel netwerken van signaaltransducties te modelleren en te analyseren met Monte Carlo *model checking*. Het tweede formalisme is gericht op genetische regulerende netwerken waarbij de analyse is gebaseerd op de exploratie van de toestandsruimte op zoek naar attractoren.

Hoofdstuk 3 bevat een grootschalige toepassing, van het formalisme gebaseerd op Petri-netten, op signaaltransductie bij multicellulaire patroonformatie. Ons *in silico* model van de al intensief bestudeerde ontwikkeling van de vulva bij *C. elegans*, reproduceert een aanzienlijk aantal *in vivo* experimenten op statistisch significante wijze. Ook Hoofdstuk 4 richt zich op signaaltransducties. We onderzochten het effect van proteolyse bij gebrek aan voedingsstoffen in *S. cervisiae*. Met name hebben we laten zien hoe computationele modellen, analyses met behulp van methoden uit de bioinformatica, en *in vivo* observaties kunnen worden geïntegreerd om nieuwe biologische hypotheses te formuleren en valideren.

De laatste toepassing wordt gepresenteerd in Hoofdstuk 5. We construeerden een model van een netwerk gebaseerd op de functionaliteit van cis-regulerende elementen om fundamentele inzichten te verkrijgen in celdifferentiatie bij hematopoiese. In het bijzonder maakten we gebruik van technieken voor de analyse van toestandsruimtes, zoals beschreven in Hoofdstuk 2, om de *in vivo* zoektocht naar een inhibitor tussen de eiwitten Gata1 en Fli1 toe te spitsen op een beperkt aantal mogelijk links.

In Hoofdstuk 6 beargumenteer ik tenslotte dat om een software-achtige beschrijving van biologisch gedrag te verkrijgen, we in staat zouden moeten zijn om te abstraheren van de fysieke effecten van chemische reacties, naar het nivo van functies die door zulke chemische veranderingen worden bewerkstelligd in een systemisch perspectief. Het is nu al noodzakelijk, en dit zal in de toekomst alleen maar toenemen vanwege de ongelofelijke mate en snelheid van kennisvergaring, om de enorme hoeveelheid complexe moleculaire en systemische interacties te reduceren tot voor mensen begrijpelijke proporties. Dit vergt de creatie van een effectief abstractieraamwerk voor de constructie van biologische modellen.

Sommario

Verso una Formulazione Eseguibile della Biologia

Le eccezionali scoperte dell'ultimo secolo e l'avvento delle tecnologie *high-throughput* agli inizi degli anni 2000 hanno catapultato la biologia molecolare nel regno della systems biology. La systems biology ha ereditato il linguaggio denotazionale proprio della teoria del controllo. Tuttavia, è chiaramente sentita dagli stessi biologi la necessità di un nuovo linguaggio capace di catturare ed esprimere processi biologici in modo intuitivo.

Durante i miei studi è emerso ripetutamente il bisogno di definire un sistema di linguaggi formali per la biologia che fossero complementari ma coerenti. In questa tesi, ho mostrato attraverso alcuni esempi, come la definizione di processi biologici tramite modelli formali assicuri una interpretazione consistente e aiuti a definire chiaramente sia problemi che ipotesi. Nel Capitolo 2 ho spiegato la nozione di modelli eseguibili per processi biologici così come è stata introdotta da Fisher & Henzinger e ho presentato due formalismi basati sulle reti di Petri. Il primo formalismo, la cui analisi è fondata su simulazioni e Monte Carlo *model checking*, è stato costruito per modellare *signaling networks*. Il secondo formalismo si concentra sulle reti di regolazione genica e la sua analisi si basa sulla ricerca di attrattori nello spazio degli stati generato dall'esecuzione del modello biologico.

Nel Capitolo 3 abbiamo dimostrato un'applicazione su larga scala del nostro formalismo basato sulle reti di Petri per la trasduzione del segnale nell'ambito della formazione di motivi multi cellulari. Abbiamo applicato il nostro metodo al processo di sviluppo dell'organo riproduttivo del nematode *C. elegans* mostrando che il modello da noi costruito riproduce un largo campione di esperimenti *in vivo* in modo statisticamente accurato. Anche il Capitolo 4 si concentra sulle reti di trasmissione del segnale. In questo caso, abbiamo investigato l'effetto della proteolisi dopo la privazione di nutrienti nel fungo *S. cervisiae*. In particolar modo abbiamo mostrato come modelli computazionali, analisi bioinformatiche e osservazioni *in vivo* possano

essere integrate in modo da formulare e validare nuove ipotesi biologiche.

L'ultimo caso di studio è presentato nel Capitolo 5. Abbiamo costruito un modello di rete di regolazione genica basato sulla funzionalità di elementi cis-regolatori, in modo da generare intuizioni fondamentali nella differenziazione cellulare durante il processo ematopoietico. In particolare, ci siamo avvantaggiati delle tecniche di analisi dello spazio degli stati spiegate nel Capitolo 2, per guidare la validazione *in vivo* della nuova relazione inibitoria tra le proteine Gata1 e Fli1.

Infine, nel Capitolo 6, discuto della necessità di astrarre dalle singole interazioni a livello molecolare la funzione svolta da tali interazioni a livello sistemico, in modo da poter fornire una spiegazione a livello sotfware dei processi biologici. Inoltre, è necessario già ora, ma lo sarà ancor più in futuro, data l'incredibile quantità di informazioni e velocità nell'accumularne di nuove, ridurre il vasto ammontare di complesse interazione molecolari e sistemiche in termini umanamente comprensibili e quindi, creare un valido e solido sistema di astrazioni per la costruzione di processi biologici.

Acknowledgments

It is only towards the end of your Ph.D. quest, when you can abstract from the daily steps and look back at the journey in its entirety, that you fully appreciate how much your (co)promotors contributed to your success. Therefore, I owe my deepest gratitude to them. In particular, I want to thank Jaap Heringa for the freedom, trust, and support that he granted me to pursue my research projects. His vast experience as bioinformatician is an invaluable source of inspiration. Likewise, I want to thank Wan Fokkink for the countless advices he gave me, and for honing my ideas with his sharp observations. His ability to spot flaws in my train of thoughts while talking about science or over the (chess) board is formidable. Finally, I thank Anton Feenstra for his constant presence and his unrelenting enthusiasm in doing science. He was always keen on experimenting with my new ideas and pushing them forward.

I'm also very grateful to Henri Bal, Thilo Kielmann, and, especially, Elzbieta Krepska who co-authored part of the work presented in this thesis. A number of other collaborations have been instrumental in the successful completion of this doctoral dissertation. Especially the collaborations with Berthold Göttgens (Cambridge Institute for Medical Research, UK), Jasmin Fisher (Microsoft Research Cambridge, UK), Steve Oliver (University of Cambridge, UK), and Ioannis Xenarios (University of Lausanne, Switzerland). In particular, I want to express my gratitude to Jasmin Fisher for inviting my twice at Microsoft Research Cambridge as an intern and for being an exceptionally kind host during that time.

I would like to extend my gratitude to Marcel Reinders (Delft University of Technology, Netherlands), Jasmin Fisher, Ioannis Xenarios, and Sanne Abeln (VU University Amsterdam, Netherlands) for the effort to serve on my reading committee, and Bas Teusink (VU University Amsterdam, Netherlands) for his willingness to take part in the opposition.

My life as a Ph.D. student at the VU University Amsterdam would not have been so enjoyable without all the past (Pavol, Sandra, Walter, Thomas, René and Bernd) and current members (Hans, Bart, Hannes, Sanne, Mohammed, Punto, Erik, Qingzhen, and Ali) of the Centre for Integrative Bioinformatics VU. I was fortunate to have the

opportunity to share the IBIVU roof with you.

It is also a pleasure to thank all my new colleagues on the floor B7 at the Nederlands Kanker Instituut. Especially Lodewyk Wessels for allowing me to complete this dissertation during the last few months.

During the last five years I had great time in Amsterdam, not only because it is, indeed, a small great city, but also because I had the privilege to meet brilliant people. Especially, I want to thank for various reasons Nienke, Roser, and all the fellow Ph.D. students of the math department Michelangelo, Thomas, Alvise, Simone, Blaz, Robert, Martijn, François, Bogdan, and Pia (who gave me the inspiration for the cover design). We had a great time at the *Tegenstelling* discussing about everything between devoted science and t-shirts designs.

Θα ήθελα επίοης να ευχαριοτήσω όλη την ελληνική συμμορία, ιδιαίτερα την Τίνα και τη Χρυσούλα. Μαζί μοιραστήκαμε αξέχαστες στιγμές χαράς, διασκέδασης και γαστριγαργικής απόλαυσης.

Non potrei mai dimenticarmi degli italiani della VU: Cristian, Alvise, Desy, Stefania, Simone, Gaia e Silvia. Compagni non solo di pranzi e caffè, ma di un pezzo di vita. Inoltre, voglio ringraziare tutti gli amici e parenti che dall'Italia, e in Italia, non mi hanno fatto mai sentire uno straniero. Specialmente Francesco e Tommaso.

L'ultimo ringraziamento, il più sentito, va alla mia famiglia: Pia, Ivano e Irene. A loro è dedicata questa tesi di dottorato.

Publications

Bernardinello, L., N. Bonzanni, M. Mascheroni & L. Pomello (**2007**). *Modeling Symport/Antiport P Systems with a Class of Hierarchical Petri Nets* in *Membrane Computing* vol. 4860 (Springer-Verlag, Berlin / Heidelberg), 124–137.

Bonzanni, N., E. Krepska, K. A. Feenstra, W. Fokkink, T. Kielmann, H. Bal & J. Heringa (**2009a**). Executing multicellular differentiation: quantitative predictive modelling of C. elegans vulval development. *Bioinformatics* 25(16), 2049–2056 (cit. on p. 51).

Bonzanni, N., K. A. Feenstra, W. Fokkink & E. Krepska (**2009b**). *What Can Formal Methods Bring to Systems Biology?* in *Formal Methods* vol. 5850 (Springer-Verlag, Berlin / Heidelberg), 16–22 (cit. on p. 10).

Bonzanni, N., N. Zhang, S. G. Oliver & J. Fisher (**2011**). The role of proteosome-mediated proteolysis in modulating potentially harmful transcription factor activity in Saccharomyces cerevisiae. *Bioinformatics* 27(13), I283–I287.

Bonzanni, N., A. Garg, S. D. Foster, N. K. Wilson, S. Kinston, D. Miranda-Saavedra, J. Heringa, A. Feenstra, I. Xenarios & B. Göttgens (**2012**). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Submitted.*

Krepska, E., N. Bonzanni, A. Feenstra, W. Fokkink, T. Kielmann, H. Bal & J. Heringa (**2008**). *Design issues for qualitative modelling of biological cells with Petri nets* in *Formal Methods in Systems Biology* vol. 5054 (Springer-Verlag), 48–62 (cit. on pp. 31, 32).

Curriculum Vitae

Nicola Bonzanni was born on December 15, 1980. His education started at the Liceo Scientifico G. Maironi da Ponte where he graduated in 1999. Next he attended the Università degli Studi di Milano-Bicocca to study Computer Science. His MSc project in the group of Prof.dr. Lucia Pomello resulted in a peer-reviewed publication that appeared in Lecture Notes in Bioinformatics. He obtained his MSc degree in 2007. Few months after his graduation, Nicola was accepted as a PhD Student at the VU University Amsterdam under the supervision of Prof.dr. Jaap Heringa, Prof.dr. Wan Fokkink, dr.ir. K. Anton Feenstra. During his PhD he collaborated in the FP6 project *ENFIN*. Furthermore, in 2009 and 2010, he was an intern at Microsoft Reasearch in Cambridge under the supervision of dr. Jasmin Fisher. Currently he is working as a Postdoctoral Researcher at the Centre for Integrative Bioinformatics VU and at the Nederlands Kanker Instituut.