1

Inference of Gene Regulatory Networks from Multi-scale Dynamic Data

Arnaud BONNAFFOUX
Vidium Solutions SAS, Lyon, France

We attribute the ability to "make decisions" to living systems, and in particular to cells, to modify their behavior following variations in their environment. This property is essential for the proper functioning of a multicellular or unicellular organism. The behavior of a cell can be defined by the adaptation of its functional capacities, following modifications of its environment or its internal state. Since these functions are mostly carried out by proteins, it is appropriate for the cell to regulate gene expression, coding for the appropriate proteins. This is what is done by the cell in the case of differentiation. This regulation is done partly by a series of interactions between genes that activate or inhibit each other. Knowledge of these gene regulatory networks (GRNs) would theoretically make it possible to understand, predict and potentially influence the behavior of a cell and thus propose new treatments in the case of pathologies involving deregulation of the GRNs. A lot of work has been carried out over the past 20 years to confront this challenge, and despite some success, the problem of GRN inference remains largely open. What these approaches have in common is the analysis of expression data using statistical or mechanistic models of GRN to infer correlation or causality relationships.

In this chapter, we will begin by presenting the notions of GRN and differentiation, and then we will review the different approaches for inferring GRNs from population or single cell data, as well as their alternatives. We will then review limitations of

Symbolic Approaches to Modeling and Analysis of Biological Systems,
coordinated by Cédric LHOUSSAINE and Élisabeth REMY.

existing approaches. Finally, in the last section we will cover an original approach to GRN inference based on the study of expression waves that we have developed to overcome current limitations.

1.1. GRN and differentiation

In this section, we present the mechanisms and roles of GRNs before detailing the specific case of differentiation. After an overview of the classic view of the differentiation process, we will see how GRNs are involved in a more modern view.

1.1.1. *The coordination of gene expression by GRNs*

1.1.1.1. *Controlling gene expression*

All the cells of a multicellular organism contain the same genetic code contained in their DNA molecules (with rare exceptions, such as immune cells or gametes), and yet they exhibit extremely varied phenotypes. This property is explained by the differential control of the genes expressed. At this stage, it is useful to define more precisely the notion of "gene expression", which will be used in this text. We assume that a gene corresponds to a DNA sequence coding for a protein. We effectively exclude all genes that do not produce mRNAs, although non-coding RNAs also play an important role as shown for miRNAs, siRNAs, lncRNAs, etc. He and Hannon (2004); Mercer et al. (2009). When a gene is expressed, it is assumed that its mRNA is significantly transcribed and in a way that is possible to detect. We agree, however, that this definition is very inappropriate at the single-cell level because of the stochasticity of expression, as we will see later.

A cell can therefore control the expression of its genes and produce the proteins necessary for its functioning. It is estimated that at any time a human cell expresses between 30% and 60% of its 25,000 genes (Davidson 2010). However, many processes are common to all cell types. Thus, only a part of genes is specific to a cell type. For example, chromatin proteins (histones), RNA polymerases, important metabolic enzymes or even cytoskeletal proteins are found in all cells, even if their level of expression may vary. On the other hand, hemoglobin is only detectable in erythrocytes, so it is specific to this cell type and contributes to its definition.

Control of gene expression in eukaryotes can occur at different levels ranging from DNA to protein. The most advanced and most studied level is that of transcription. Transcription of mRNA from DNA involves a series of actors, such as general and specific transcription factors, which must coordinate in time and space to initiate and maintain the transcription process. However, regulation of expression can also take place later during the maturing of the mRNA, its transport into the cytoplasm, its translation or its degradation (Valencia-Sanchez et al. 2006). At the

protein level, there may also be so-called post-translational regulations that alter its activity as well as its stability (Olsen and Mann 2013; Manning and Cooper 2017). As well as these differences in levels of control, we must add the differences in dynamics. There are stable and persistent controls, as in some cases of epigenetic regulations that induce X chromosome inactivation (Panning and Jaenisch 1998), unlike more dynamic controls by transcription factors that require regulatory proteins to be present at all times.

Although most studies on GRN inference are limited to transcriptional regulation, we draw attention to the fact that in this work we consider several levels of regulation. However, we will not consider persistent-type regulations as epigenetic mechanisms.

1.1.1.2. *Definition of GRNs*

When a cell modifies the expression of its genes following a change of environment or following a stimulus, it is very rare that a single gene is impacted. It is a set of functionally linked genes, which is regulated, as in the case of liver cells that respond to glucocorticoids by over-expressing a series of proteins specialized in the production of glucose (Rousseau 1975). Therefore, there is a coordination of the control of gene expression, which is itself dependent on the genes expressed. Only a few cell types are able to respond to glucocorticoids, unlike the other cell types that are also shown.

We then focus on GRN as the set of gene interactions that control the expression of other genes at all levels (transcription, translation, etc.). We will now detail important concepts of structures and states of GRNs to define them exactly.

The set of possible interactions between genes is called the “structure” of the GRN. What we call a “state”: for a given cell at a time T, the set of expression levels of the mRNAs and proteins of all the genes. The structure of a GRN is constant over time and does not depend on cell type, unlike the state of a cell. It is this fundamental difference that defines these two notions. The structure of a GRN includes the conditions for an interaction to be effective. In the previous example, for the target genes of glucocorticoids to be induced, you need the presence of the hormone, and the presence of proteins specific to the hepatic cells which authorize the response of the target genes. In the case of a skin cell, which has the same GRN structure for the response to glucocorticoids as the liver cell, its state is different and does not allow the activation of this interaction.

The GRN inference task can thus be defined as the identification of the structure of the GRN from experimental measurements of the state of the GRN under different conditions or during kinetics. It is useful to state here that only a small part of the structure of the GRN will be explicit and a significant part will be hidden as illustrated in Figure 1.1. To understand this limit, let us take the example of liver cells. If we seek to find the series of interactions that lead to the activation of the target genes following the presence of the hormone, it will not be possible to know

all the conditions necessary so this GRN is operational. If a gene A activates a gene B, it is very likely that this interaction requires the presence of genes constantly and specifically expressed in this cell type. The notion of GRN is therefore relative and restricted to genes whose expression varies during the process studied.

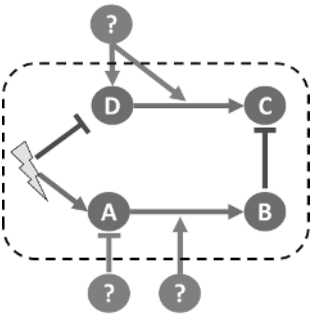

Figure 1.1. *Schematic representation of a GRN. An example of GRN involving four genes A, B, C and D regulated following a stimulus (yellow flash). For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.1.– *Interactions between stimulus/gene where gene/gene are represented by green arrows (activation) or red "blocks" (inhibition). These interactions constitute the visible structure of the GRN. The genes may also be influenced by other unidentified genes (gray with "?") but not impacted by the stimulus and therefore constant during the stimulation. These constant genes can also regulate the interactions as for the interaction between genes A and B. These interactions define the hidden structure of the GRN. The GRN will be defined as the entire structure visible in the dotted block. However, note that this GRN is highly dependent on the cellular context. Its state will be defined by the level of mRNAs and proteins of genes A, B, C, D during stimulation.*

1.1.1.3. *Approximation of gene interactions by separation of time scales*

We will now define more precisely the notion of interaction between genes, which is in itself an abstraction and hides complex mechanisms. As mentioned earlier in this introduction, the regulation of gene expression can occur at multiple levels. Each level contributes to the regulation, which, at first glance, complicates the overall regulation of the GRN. However, we will see that by applying a time scale separation hypothesis, we can reduce the overall regulation of the GRN to its slowest processes, which we will assume to be transcription and translation.

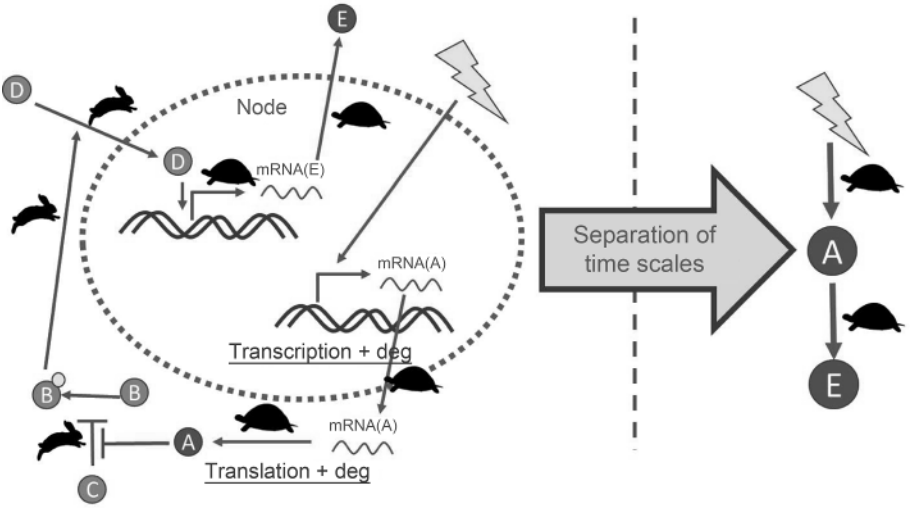

Figure 1.2. *Separation of time scales. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.2.– *On the left, a GRN involving genes A, B, C, D, E. Only genes A and E undergo regulation of their expression. A is directly under the control of a stimulus. D is a transcription factor sequestered in the cytoplasm that can be released in the nucleus because of the effect of protein B phosphorylated indirectly by the action of A, which inhibits protein C that prevents the phosphorylation of B. Transcription processes and translation are typically slow (hours, represented by turtles) compared to other phosphorylation or translocation processes (seconds or minutes, represented by rabbits). The time scale separation approximation (right) considers only slow processes and reduces this stream of reactions to the activation of A by the stimulus and then the activation of E by A.*

In our study, we will assume that interactions between genes occur via proteins. In fact, for the sake of simplification, we deliberately neglect the regulation carried out by non-coding RNAs, in particular at the level of degradation. Once the regulatory protein is produced in the cytoplasm, it can undergo post-translational modification cascades, such as phosphorylations. It can also in turn induce a cascade of phosphorylation on other proteins, which *in fine* will lead to the activation of a transcription factor that will regulate the expression of a target gene as shown in Figure 1.2. We predict that the characteristic reaction times that take place in the cytoplasm are very short compared to the dynamics of production/degradation of RNA and proteins, which are several hours. In fact, once the regulatory protein has been produced, all these intermediate reactions, which also act as filters, very quickly reach a steady state and the overall transfer is limited to the static gain generated by these reactions.

We will concentrate on the study of the slow dynamics of the GRN and neglect the faster ones. This assumption of separation of time scales is justified by the following points. Differentiation processes typically last several days, which underlies equivalent dynamics for the GRN to stabilize. The other reason is related to experimental constraints. Current techniques for measuring the observable phenomena of the GRN system are limited to RNA and proteins. Given the dynamics of these molecules, there is no point in measuring them every minute. Finally, the number of experimental points is, in practice, limited by their cost.

GRNs are involved in many stimulus response processes as we have just seen. They can also serve as a logic circuit to process information (Benenson 2012), or be used as oscillators as for the circadian clock (Tyson et al. 1999). But the most common example is the involvement of GRNs in the differentiation processe, which will be our biological model in the context of this chapter, and which we will introduce in the next section.

1.1.2. *The process of differentiation*

GRNs have historically been studied in eukaryotes, in the case of development and are therefore strongly linked to the process of differentiation. We will first present the process of differentiation in its classic view, then we will go into detail into the particular case of erythrocyte differentiation, which is the model studied in our team, and finally we will study the link between GRNs and the new view that results from it.

1.1.2.1. *The historical view of the cell differentiation process*

Differentiation is the process by which a cell will change cell type to specialize. Differentiation occurs during development to create an organism from a single cell, that is, the egg, as well as in adult life to regenerate tissues. Differentiation usually occurs during cell division, which can be symmetric or asymmetric, implying, respectively, that the daughter cells are differentiated and identical or that one daughter cell is identical to the mother and the other differentiated. From the classic viewpoint, differentiation is an irreversible process that leads an immature cell to specialize toward a more mature cell. In the last stage of differentiation, the cell is fully mature and cannot divide or differentiate. This irreversibility requires the cell to memorize the stages of differentiation through which it passed. The different stages of differentiation from the egg can then be seen as a family tree with directed descent links. A cell type can generate several cell types, so we then talk about multi-lineage. This view is a perfectly deterministic approach, and generally involves so-called "master" genes whose sole activation during differentiation allows the activation, or inhibition, of all its target genes specific to the new cell type (Mattick et al. 2010). These "master" genes are very often transcription factors or tumor suppressor genes. However, this classic vision is now widely challenged as we will see later.

1.1.2.2. *Features of the process of erythrocyte differentiation*

The biological model that we will study in this book concerns hematopoiesis, and more particularly erythropoiesis, in chickens (Gandrillon et al. 1999). Hematopoiesis consists of the generation of all blood cells from a single type of blood stem cell through a process of hierarchical differentiation, which has two main branches: lymphoid and myeloid (Orkin and Zon 2008). The myeloid progenitors will gradually differentiate and give rise, among other things, to erythrocyte progenitors which, in short, will only differentiate into an erythrocyte. Erythrocytes, commonly known as red blood cells, are specialized blood cells that serve to transport oxygen to all organs via the blood network. Oxygen is binded by hemoglobin, a protein exclusively and abundantly produced in this cell type. At the last stage of erythrocyte differentiation, the nucleus is removed and the mitochondria are eliminated by autophagy (Orkin and Zon 2008). This drastic step of irreversible differentiation can be understood in terms of the cell being nothing more than a "bag" of hemoglobin, the mitochondria which consume oxygen would decrease the efficiency of oxygen transport. However, these characteristics do not apply to birds. The nucleus is kept, but condensed to such a level that it is non-functional and any transcription activity is undetectable. Mitochondria are not phagocytosed and remain functional. The T2EC erythrocyte progenitors (Gandrillon et al. 1999) that we are studying are primary cultured cells from chicken embryos. They can be maintained in a state of self-renewal or induced in differentiation in 4 or 5 days.

1.1.2.3. *Differentiation and GRNs from the modern viewpoint*

The regulation of gene expression was initially studied in bacteria (Jacob and Monod 1961). In eukaryotes, where regulation is more complex, expression control was initially studied in the development of sea urchin and Drosophila (Davidson 2010). This made it possible to establish complex GRNs as logic circuits of interconnected genetic switches (Levine and Davidson 2005) mechanistically explaining the properties of differentiation. The "master" genes are activated sequentially according to external signals. Furthermore, the positive loops present in the GRNs make it possible to explain the notion of memory necessary for the irreversibility of the process. Within this context, the process of differentiation is the transitory phase of a GRN, which passes from a stable state to another following a stimulation.

However, we know today that the dogma of the process of irreversible hierarchical differentiation no longer holds, following the experiments of Yamanaka's team on the reprogramming of differentiated cells (Takahashi and Yamanaka 2006). In fact, we now know how to reprogram many differentiated cells into pluripotent cells (Yamanaka 2012) capable of generating all lineages. There are also many examples of transdifferentiation without going through the IPS step (Jopling et al. 2011). The framework, proposed by GRNs with the idea of invariant

structure and variable state, is perfectly compatible with these observations and makes it possible to propose mechanisms to explain them.

Despite the success of GRNs in explaining complex differentiation processes, some observations such as spontaneous differentiation or rebel cells are not explained by the deterministic view of logic circuits (Mojtahedi et al. 2016). As we will see later in this introduction, single cell measurements have highlighted the extreme stochasticity of gene expression. GRNs then appear more like a network of constraints that channel the stochasticity of gene expression. If the constraint exerted by the GRN is very strong, the probability associated with an end state is very large, and the response of the cell can appear deterministic. On the contrary, if the constraints are weaker and there are several stable states, the cell can then randomly explore this space and switch from one state to another (Kupiec 1997).

In the following section, we are going to present different methods of GRN inference for which the majority are based on the classical view of differentiation. However, we will see later how the new insight can be useful for inferring GRNs from single-cell measurements.

1.2. Inference of GRN from population data

The idea of inferring GRNs from expression data began in the late 1990s with the arrival of the first DNA expression microarrays (Schena et al. 1995; Lockhart et al. 1996). Inference algorithms then followed the evolution of transcriptomics technologies to the very latest single-cell data. To have the most complete and precise picture possible of the technologies, it is better to start by presenting the different families of algorithms developed for population data analysis, which have been adapted for single cell data analysis.

1.2.1. *Population expression data*

DNA microarray technology has made it possible to measure the level of expression at the level of the genome of a population of cells for the first time. In fact, rather than being interested in a group of genes in particular, we can observe all the genes impacted by a stimulation. Biologists understood the need to study the response of cells at a systemic level rather than a reductionist one. The concept of a genetic network then became important. Network inference algorithms were therefore developed at this time, as traditional approaches to manual analysis of expression data could no longer be done on thousands of genes.

Expression data is classified into two broad categories: perturbations and kinetics. The aim of the perturbation experiments is to impose an under-expression or

over-expression of certain GRN genes and to measure the impact on the other genes generated by the interactions. The idea is to find the interactions by *retro-engineering* measures. In practice, most of these experiments are carried out in unicellular organisms such as the bacterium *Escherichia coli* (Gardner et al. 2003) or the yeast *Saccharomyces cerevisiae* (Hughes et al. 2000) that are genetically manipulated easily. This technique makes it possible to carry out high-throughput screening, for example, in Hughes et al. (2000) where 300 perturbations were carried out. This type of experience is also found in human cells (Basso et al. 2005).

Kinetic data make it possible to follow the dynamic evolution of gene expression during physiological processes or after a disruption. The underlying idea is that the structure of the network imposes the dynamics observed. In this way, yeasts were synchronized to show oscillations due to the cell cycle in the expression of 800 genes (i.e. one-third of the genome) with the aim of finding the GRN involved in the cell cycle (Spellman et al. 1998). In the *Drosophila melanogaster* fly (Arbeitman et al. 2002), 4,028 genes were measured at 66 sequential stages of development from fertilization to the adult stage.

Systemic expression data can also be used for classification purposes, as in the case of tumors (Bittner et al. 2000). The objective is to establish expression profiles and match them with the tumor phenotype. Much more data have been produced, and to improve access to these, computer repositories have been created, such as NCBI's *Gene Expression Omnibus* to keep thousands of data from expression chips. RNA sequencing data came later and was ultimately used very little in GRN inference (Morin et al. 2008). All the approaches presented below only use chip data.

1.2.2. *Bayesian approaches*

Microarray expression data are snapshots of the expression level of several hundred genes and are considered noisy and semi-quantitative. Bayesian approaches, which model GRNs by a stochastic process, are particularly suitable for the analysis of these data and were used very early on Friedman et al. (2000).

1.2.2.1. *Bayesian networks*

In a Bayesian network, the GRN is represented by a directed acyclic graph, denoted as G (Figure 1.3) where the nuclei are random variables, denoted as X_i, which depend on the state of the parents, denoted as $Pa(X_i)$. A directed interaction in the graph represents the dependency of the child on its parents, which results in a conditional distribution of the state of the child depending on the state of the parents. In the Markovian hypothesis, this relationship means saying that the son is independent of his non-descendants, given his direct parents. It is this relationship that is at the heart of the causality relationship in a Bayesian network. The whole network can be defined by a joint distribution between all the network variables.

By using Bayes' theorem, associated with the Markov hypothesis of conditional independence with respect to parents, we can reduce the writing of this joint distribution to the following form:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \quad [1.1]$$

The graph G makes it possible to identify the parents of each nucleus, but to completely define the joint distribution of the Bayesian network we need to specify each conditional probability $P(X_i|Pa(X_i))$. These conditional probabilities can be defined using discrete (Pe'er et al. 2001; Yu et al. 2004; Dojer et al. 2006; Chai et al. 2012) or continuous functions (often Gaussians), which are specified by a set of parameters noted Θ. A Bayesian network is therefore defined by its graph G and its parameters Θ.

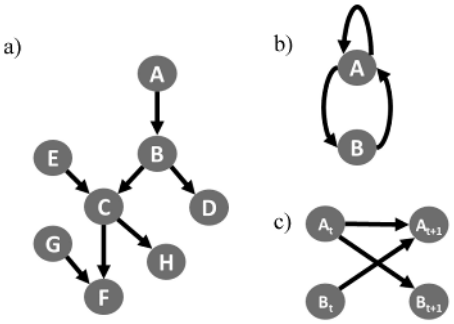

Figure 1.3. *Bayesian networks. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.3.– *(a) A Bayesian network is represented by a directed acyclic graph. A nucleus represents a gene defined by a random variable whose distribution density depends on its regulators or parents. In this example, gene C has a distribution that is linked to that of his parents E and B, but also to his grandparent A, his cousin D and its descendants, F and H. The distribution of C is independent of G because they are not related. With the Markovian hypothesis, the distribution of C conditioned to all genes except its descendants returns to the distribution of C conditioned in its parents. So we have* $P(C|A, B, D, E, G) = P(C|B, E)$. *(b and c) The graph in (b) illustrates a graph structure with loops that cannot be represented by a directed acyclic graph. The graph (c) represents the network (b) by a dynamic Bayesian network using an acyclic graph. nuclei A and B are represented at time* t *and* $t + 1$. *An interaction always takes place* t *to* $t + 1$, *which guarantees the absence of loops in this acyclic graph.*

1.2.2.2. *Bayesian inference*

To infer the GRN from a dataset D, the idea is to find the graph G that best explains D. To do this, we generally define the posterior probability $P(G|D)$ of having a graph G knowing D and we seek the graph G, which maximizes this probability. We can once again use Bayes' theorem to decompose $P(G|D)$, which can be written in the form:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \quad [1.2]$$

The term $P(D|G)$ is called the probability and $P(G)$ is the prior probability. $P(D)$ here corresponds to a constant. To calculate the probability, we often use the *Bayesian information criteria* (BIC) or *Bayesian Dirichlet equivalence* (BDe) scores. In particular, these scores integrate the complexity of the graphs to avoid problems of over-fitting. The inference is done in two steps; we first propose a graph structure G to test, then we look for the most appropriate set of parameters Θ to reproduce the data. It is at this last stage that a score is assigned to the applicant network. The search for the most representative graphs in the first step is a very difficult NP-complete problem which requires the use of heuristic search algorithms, like *Markov Chain Monte Carlo* (MCMC) or *Greedy-Hill climbing* (Wu and Liu 2008; Kunga and Mohamada 2012). With the inference problem being indefinite given the low number of data compared to the complexity of the problem, several networks are often returned. The prior term $P(G)$ can be used to integrate knowledge about genes from literature or gene annotation databases (Werhli and Husmeier 2007; Yavari et al. 2008) or by grouping them into a *cluster*, in particular using temporal correlations (Yavari et al. 2008; Vinh et al. 2012). Predictions of "rarity" interactions are also sometimes used to reduce the number of possible interactions (Yang et al. 2011; Chai et al. 2012).

1.2.2.3. *Dynamic Bayesian networks*

As we will see in section 1.2.2.4, one of the major limitations of Bayesian networks is that they are static and do not allow loops in GRNs. An alternative is possible with dynamic Bayesian networks (Yu et al. 2004; Dojer et al. 2006; Wu and Liu 2008; Grzegorczyk and Husmeier 2010; Chai et al. 2012; Vinh et al. 2012), which are a variant of Bayesian networks.

The idea is to duplicate the graph with a graph G_t and another graph G_{t+1} which represent the influence of the network at time t on the network at time $t + 1$, respectively. In fact, we introduce the notion of dynamics since the state of the past network impacts the present state, and we authorize loops as illustrated in Figures 1.3(b) and (c). The inference method remains unchanged and authorizes the use of discrete temporal data which will be used 2 by 2 for two consecutive times. In this formulation, We find the idea that information takes a certain time to spread in the network.

1.2.2.4. *Advantages and limitations*

Bayesian approaches have the advantage of identifying causalities according to the hypothesis of the independence of probabilities, and they are compatible with the hypothesis of the separation of time scales. Bayesian models have few parameters, which is an advantage in terms of identifiability and the necessary statistical power. These methods are also strong against noisy data.

However, in practice, it remains quite difficult to identify the regulatory gene and the regulated gene, even if the use of dynamic Bayesian networks facilitates the exercise. The discretization of the data quantitatively degrades the accuracy of the approach and adds many parameters depending on the number of discrete states. One solution is the continuous flow with linear Gaussian assumptions that simplify the model. Combinatorics very quickly becomes important (NP-hard type problem) which limits inference to small networks typically less than 20 genes (Dojer et al. 2006; Wu and Liu 2008; Grzegorczyk and Husmeier 2010; Chai et al. 2012). Hence, the use of heuristic methods and arbitrary hypotheses of the rarity of interactions. The use of *a priori* knowledge is practiced to get around these difficulties, but very often it involves *co-clustering* from gene annotation databases whose direct use is questionable. Non-dynamic Bayesian methods are too limited in practice because they do not consider temporal evolutions and cannot infer interaction loops.

1.2.3. *Information theory approaches*

In the family of statistical approaches, there are also those based on information theory. The hypothesis is based on correlation: if two genes share information, there is certainly a more or less direct interaction. As in the Bayesian case, this approach relies on the search for statistical dependence, but in this case the nature of the inferred relationships between genes is limited to correlation. It is in fact impossible to identify causal links contrary to Bayesian approaches. However, the efficiency of this method allows processing large volumes of data to analyze genome-scale data, unlike its Bayesian counterpart.

1.2.3.1. *Mutual information*

The mutual information $IM_{i,j}$ between two genes i and j is a generalization of the correlation. It is based on entropy in the sense of Shannon, noted H_i for the gene i, and is defined as follows:

$$IM_{i,j} = H_i + H_j - H_{i,j} \qquad [1.3]$$

The entropy is given by the following formula in the case where the measurement of the expression of a gene can take n finite values $(x_1, ..., x_n)$ with the associated probabilities $(p(x_1), ..., p(x_n)$:

$$H_i = -\sum_{k=1}^{n} p(x_k) log2(p(x_k)) \quad [1.4]$$

If $IM_{i,j} = 0$, it means that the joint distribution of the two genes does not contain more information than the two genes taken separately. If $IM_{i,j} > 0$, it means that the two genes are linked and that their joint distribution contains less information than the genes taken separately. Part of the information between the two genes is redundant. The idea is that the greater the mutual information between two genes, the greater the probability of a direct interaction.

1.2.3.2. *The principle of inference*

The first application of information theory to infer GRNs was described by Butte and Kohane (1999) and applied to genomic data in yeast. The aim is to calculate the mutual information for all the pairs of genes. There are several (Steuer et al. 2002) methods for this. They are then compared to a fixed threshold and only the interactions with mutual information above the threshold are kept. Note that this method requires the data to be statistically independent, otherwise experimental correlations would be misinterpreted. This is why steady-state data are generally used (Butte and Kohane 1999; Basso et al. 2005; Margolin et al. 2006).

The most popular algorithm using mutual information is ARACNE (Basso et al. 2005; Margolin et al. 2006). It uses the method of Gaussian kernels to calculate mutual information. It also uses *Data Processing Inequality* to avoid identifying indirect interactions. Let us take the example of three genes A, B and C in cascade with the interactions $A- > B- > C$. There is a correlation between A and C and therefore $IM_{A,C} > 0$. However, the idea is that the correlation resulting from an indirect interaction is weaker than the direct interactions and, in fact, we have the inequality $IM_{A,C} < min(IM_{A,B}, IM_{B,C})$. By comparing the $IM_{i,j}$ between them, we can then eliminate indirect interactions.

1.2.3.3. *Advantages and limitations*

The main advantage of these methods over others is the processing of large volumes of data. The inferred large GRNs can then be analyzed to identify *hubs*, which can be useful for medical research (Basso et al. 2005). However, the predictive capacity of these GRNs is limited, as it is difficult to predict the consequences in the event of disruption of the *hubs*. Another important limitation to consider is the inability of this method to infer the feedback loops, which are supposed to be present in great numbers and play an important role in GRNs. Finally, these methods make

use of Gaussian linear assumptions that impact the biological relevance of the inferred GRNs. In reality, mutual information and a simple linear Pearson correlation give approximately the same results (Steuer et al. 2002).

1.2.4. *Boolean approaches*

The use of Boolean networks to represent GRNs, and even metabolic networks, dates back from 1969 (Kauffman 1969). It is interesting to note that in this theoretical study, the objective is not to find interactions within a GRN, but to study the emerging properties of these networks built randomly as evolution would have done. Kauffman showed that these random Boolean networks were stable, possessed cyclic modes consistent with the cell cycle or even exhibited multi-stationary states comparable to different cell types. This study perfectly illustrates the interest of so-called mechanistic models, as opposed to the statistical models that we have described so far.

1.2.4.1. *Boolean networks*

A Boolean network consists of a set of variables $(x_1, ..., x_n)$, which represent the genes that can take the values $(0, 1)$ representing the inactive or active state respectively. A gene is therefore seen as a switch. We find here the idea of a strong parallel between a biological circuit and a computer circuit. Here, we find the idea of a strong parallel between a biological circuit and a computer circuit. The state of a variable at time $t + 1$ is defined by a Boolean equation, which depends on the other states at time t or even at previous times (Silvescu and Honavar 2001). This model is the general framework, but there are several variants grouped into three main families (Hickman and Hodgman 2009): random networks (*random Boolean network*), asynchronous networks and probabilistic networks.

Random Boolean networks were the first studied (Kauffman 1969) because of their simplicity, which requires low computing power. These are synchronous, perfectly deterministic networks where all states are updated simultaneously (Albert 2004). The qualifier of randomness comes from their large dimension which, during the inference, forces us to randomly test different network structures in reality.

In asynchronous networks (Harvey and Bossomaier 1997), the nuclei are updated randomly in a series. They were developed to correct the deterministic and perfectly synchronous aspect of random Boolean networks to be more representative of biological observations. On the other hand, asynchronous networks require more computing power. The characteristics of these different non-deterministic models are reviewed in Apostel (2003). Note that the same network can have different properties if synchronous or asynchronous behavior is considered. For example, cyclic attractors disappear as soon as asynchronism is considered.

The first probabilistic Boolean network was described by Shmulevich et al. (2002). In this model, it is the very structure of the network which is random and which changes over time. This model responds to the observation that, for a given gene, several rules make it possible to match the experimental data. Rather than choosing a single rule, the model considers the most probable rules for the same gene according to a Markov chain (Ching et al. 2007b; Marshall et al. 2007). In fact, these models are more robust and are a compromise between Boolean networks, based on totally deterministic rules, and Bayesian networks which reproduce the stochastic behaviors observed without mechanistic description.

1.2.4.2. *Boolean network inferences*

Boolean network inference began with the arrival of DNA chips. The first step common to all these inference algorithms is the binarization of the data required by the formalism of the model. The first REVEAL algorithm (Liang et al. 1998) considers synchronous networks. It uses the first step of identifying regulators for each gene based on a measure of mutual information. This is an example of a mixed approach that combines several concepts for GRN inference. The second step involves comparing the level of expression between two times to study the variations. A "brute force" approach tests all possible rules with up to three regulators per gene to reproduce the data. This technique is limited by combinatorics and is sensitive to noise. Akutsu et al. (1999, 2000) family of BOOL methods uses the same principle, but is more suitable for noisy data. This type of method has been successfully applied for the inference of the network involved in the yeast cell cycle with a limited number of genes (Hakamada et al. 2004). Boolean probabilistic networks have also been used for GRN inference (Shmulevich et al. 2002; Zhao et al. 2006; Ching et al. 2007a). The algorithm described in Zhao et al. (2006) is particularly interesting, because it uses the kinetics of the data by seeking the mutual information between a regulatory potential at a time t and the gene regulated at $t + 1$.

Moreover, it uses a *minimum description length* (MDL) metric, which jointly considers a distance of overlap with the data and a distance of complexity of the model. The data cross-check distance is based on a notion of conditional probability close to dynamic Bayesian networks. This approach therefore combines the principles of information theory, Bayesian theory and Boolean networks.

1.2.4.3. *Advantages and limitations*

Boolean models have the advantage of being simple to implement and easy to run. Their mechanistic and dynamic properties also make it possible to reproduce behaviors observed in experimental GRNs and make predictions on steady states, but not on transients. On the other hand, they are limited to the GRN inference with a small number of genes due to a large combinatorial and the large number of parameters. In practice, deterministic Boolean networks are not very robust to noisy data. The necessary binarization of the data is also problematic since a significant part of the

information is ignored. Finally, methods like REVEAL (Liang et al. 1998) and BOOL (Akutsu et al. 1999, 2000) do not seem to be better than sheer coincidence (Wimberly et al. 2003).

1.2.5. *ODE approaches*

1.2.5.1. *ODE models*

Ordinary differential equation (ODE) models are perfectly mechanistic, since they are, by definition, continuous temporally and quantitatively, like biological systems. In this formalism, we find the notions of states and structure of GRNs presented in section 1.2.1.2. Thus, the GRN is modeled by a function which integrates external signals (vector denoted U) to modify its internal state (vector denoted X) and generate observable outputs (vector denoted Y). The function that defines the dynamics of this system is denoted F, and the one which defines these observables, G. The dynamics of this system is described by an ODE:

$$\begin{cases} \dfrac{dX(t)}{dt} = F(X(t), U(t)) \\ Y(t) = G(X(t), U(t)) \end{cases} \quad [1.5]$$

The use of a differential equation as a function of time shows the importance of the history perceived by the system to define its state at time T. In fact, to know the state of a system at a time T, it is a temporal integral of the stimuli and its state must be done.

$$X(T) = \int_0^T F(X(t), U(t))dt \quad [1.6]$$

If we continue the comparison with differentiation, we find the concept of memory in the historical view. We also note that the dynamics of the GRN (defined by the functions F and G) are not limited to the stimulus, but takes into account the internal state of the system. We find the idea that two cells of different types, and so of a different state, with the same structure react differently to the same stimulus.

Like Boolean models, ODEs reproduce dynamic behavior (oscillations, multi-stationarity, robustness, etc.), but they are also able to reproduce temporal variations of the measured variables. The equivalent of this representativeness lies in the high number of parameters, which define these models that, in practice, are not able to be identified. There is a range of models ranging from the simplest totally linear model (Gardner et al. 2003; di Bernardo et al. 2005; Bansal et al. 2006) to the most complex, with nonlinear (Ando et al. 2002) and stochastic (Chen et al. 2005) functions. The complexity of the model has consequences on its properties (Polynikis et al. 2009), and in fact the choice of model must be a compromise between its biological representativeness and its use.

1.2.5.2. *ODE model inference*

ODE models can be inferred from perturbation experiments with steady-state data or with kinetics. In the former, the assumption of stationarity makes it possible to simplify the differential equations (see equation [1.5]) by canceling the derived terms. Furthermore, as the stimulus is clearly represented, if we know the perturbation targets (which can be a drug, a K.O. or an over-expression), we can deduce the coefficients of the Jacobian matrix of the linear model by the multiple linear regression method. This method was applied and validated in Gardner et al. (2003) on a known network in *E. coli* containing nine genes for which nine over-expression experiments were carried out. The same method was applied in di Bernardo et al. (2005) on a genomic-scale network on a set of 515 perturbation experiments in yeast, all with steady-state measurements. In this case, the perturbation targets were not known but inferred by a statistical approach. In these two studies, the generated models made it possible to predict the direct targets of a new drug that were not used in the initial data set.

In the second case, the model is inferred from kinetic data following a perturbation (Ando et al. 2002; Chen et al. 2005; Bansal et al. 2006). In Bansal et al. (2006), a linear model is used to infer the same GRN with the same data as in Gardner et al. (2003). First, the data are approximated by smooth functions and then made discrete in order to estimate the derivatives of the states at each measurement point. In addition, the dimension of the data is reduced via a principal component analysis to identify the problem. The inference of the Jacobian matrix is then done by simple matrix inversion, given the estimates of the derivatives and the states. For approaches using nonlinear models (Ando et al. 2002; Chen et al. 2005), inference uses heuristic methods for parameter estimation, such as genetic algorithms (Ando et al. 2002) or *Maximum Likelihood Estimation* in Chen et al. (2005), who also use knowledge from this book to define the list of potential regulators. In both cases, the aim is to minimize the difference between the simulated and experimental data by using the least squares method, for example.

1.2.5.3. *Advantages and limitations*

Just like other mechanistic and dynamic Boolean approaches, ODE methods have the advantage of reproducing observed behaviors and being predictive, even from a quantitative point of view. This property is due to the biological relevance of these models which clearly represent the stimulus and include nonlinear processes. However, in practice, they are limited by the large number of parameters that makes them difficult to identify. This is why methods based on ODEs resort to arbitrary assumptions of the rarity of interactions or to linearizations. Finally, these models, which are meant to be mechanistic, are completely deterministic, even if a Gaussian noise can be added which, as we will explain next, does not make it possible to reproduce the observed behavior at the cell level.

1.3. Inferring GRNs from single-cell data

1.3.1. *Single cell expression data*

Molecular biology techniques have experienced remarkable growth over the past 15 years, particularly in the field of studies at the single cell level. It has been known for a long time to observe cells individually in microscopy and to even quantify the level of some proteins at the scale of the cell because of flow cytometry. However, there has not yet been a high-throughput technique to access cell-scale RNA measurement. Since then, several single-cell techniques have been developed based on RNASeq or qPCR (Warren et al. 2006; Kolodziejczyk et al. 2015). At the same time, other epigenomic and genomic techniques have been developed at the cell level (Clark et al. 2016), but these have rarely been used for GRN inference, as we will see in the study of alternative inference methods. Single-cell data offers multiple benefits. We can eliminate the average effect and observe the heterogeneity of a population, like in blood. This also gives access to the internal content of a cell, which makes it possible to get rid of the contradiction of biologists who describe the behavior of a cell from measuring a population (Levsky and Singer 2003). Finally, the statistical power is important because these techniques make it possible to measure hundreds or even thousands of cells individually, which represents so many independent statistical achievements. This aspect is the reason for the popularity of research on GRN inference for the analysis of this new data as we will see.

However, the first results raised surprise and some doubts, about the credibility of the data. The measurements showed a large inter-cellular heterogeneity, which could be expected, but not at this level, as well as a significant number of bad measurements, which led the community to characterize these RNA distributions to be *zero inflated* (Pierson and Yau 2015). These observations were interpreted in two very different ways, which created two distinct strategies in the use of these data for the GRN inference. The first is to consider the variability as mainly technical noise and inter-cell desynchronization. The second, which will be our approach, is to consider stochasticity as a biological reality and integrate it into the inference. We will now explain these two approaches.

1.3.2. *Adaptation of GRN inference algorithms for single-cell data analysis*

For a very large proportion, the population-based GRN inference approaches that we have just shown have been adapted for single-cell transcriptome analysis. Detailed reviews can be found in Babtie et al. (2017); Fiers et al. (2018) and a comparative inference performance test in Chen and Mar (2018). The adaptation logic of the algorithms is based on the idea that the observed heterogeneity is nothing but noise that masks a completely deterministic and continuous process of

expression, in line with the classical view of the process of differentiation. The noise may be a technical problem, called a *dropout* (Brennecke et al. 2013), which explains the large number of zeros, or is due to the temporal desynchronization between cells during differentiation. Measurement noise is already present in population data and many existing algorithms have been developed to be robust against it. Managing the zeros is thus only a particular case of noise. In the same way, the desynchronization hypothesis has led to the reconstruction of "pseudo-time" trajectories (Cannoodt et al. 2016) to sort the cells chronologically and show a continuous deterministic process. This temporal ordering is based on the hypothesis that two cells close that are temporally close are close quantitatively at the molecular level, since they follow the same process. Consequently, all the biological assumptions on which the inference algorithms developed on population data are based are still valid. In theory, this allows direct use of these tools on these data to benefit from their statistical power, in terms of repetition and joint distribution. Thus, SingleCellNet (Chen et al. 2015) and BoolTraineR (Lim et al. 2016) are algorithms based on Boolean networks that integrate the first step of sorting cells temporally. Models of asynchronous Boolean networks have also been successfully applied on this hypothesis (Moignard et al. 2015). In the group of statistical approaches by mutual information or Bayesian inference, we can cite SCOUP (Matsumoto and Kiryu 2016), SCIMITAR (Cordero and Stuart 2016) or AR1MA1-VBEM (Sanchez-Castillo et al. 2017), which also use the first step of temporal sorting of cells. ODE approaches still retain the same models used for population data analysis, as in SCODE and InferenceSnapshot. These algorithms do not require kinetic data since time is assumed to be reconstructed by cell scheduling.

1.3.3. *Using single-cell stochastic models for GRN inference*

Although there is still some doubt about the quality of transcript quantification in a cell, there have long been biological observations that show biological stochasticity. We have already mentioned the case of non-deterministic behaviors during differentiation. More directly, at the molecular level, flow cytometry data showed scattered distributions of proteins over a population of cells. More recently, in 2002, Elowitz (Elowitz et al. 2002) showed in *E. coli* that gene expression was intrinsically random. This observation has since been confirmed in other organisms including humans (Suter et al. 2011; Corre et al. 2014) and has also shown that the dynamic expression regime was a succession of RNA *bursts*. These observations of *bursts*, followed by long periods of RNA-free relaxation over several hours, are consistent with the patterns of RNA distribution given by single-cell measurements. The large number of zeros would correspond to measurements during relaxation periods, and the high inter-cellular variability would come from the random nature of these *bursts*. However, we should not believe that everything would be pure chance and anarchy in the cell. It has been shown that the frequency and size of *bursts* are correlated with the activity of the (Kim and Marioni 2013) gene. So, a weakly expressed gene will

have a lower probability of generating *bursts*, compared to a strongly expressed gene. In this probabilistic view, the regulation of gene expression must therefore be seen as the modulation of the probabilities of *bursts*.

The molecular mechanisms responsible for these *bursts* are still not understood well. However, observations at the level of nascent RNAs (Larson et al. 2009) show that this stochasticity is present at the time of transcription. One of the most popular hypotheses is that promoter activity is the main source of stochasticity, due to the very limited number of promoters per gene in a cell (usually equal to the chromosome copy number). The *random telegraph*, or two-state model, proposed in 2005, models the mechanism of gene expression by summarizing it to two states, "on" and "off", between which the promoter oscillates randomly with an average frequency k_{on} and k_{off}, leading to the activation (on) or inhibition (off) of the gene, respectively (Golding et al. 2005; Pedraza and Paulsson 2008; Suter et al. 2011) . It is the random switching between these two conditions that generates the *bursts*. Their frequency and their average duration are dependent on the parameters k_{on} and k_{off}.

For GRN inference, if we consider that the stochasticity observed in the data does correspond to a biological reality, it is important to integrate it because it carries information and plays a role in the behavior of the GRN. Very few approaches have followed this idea. As far as we know, only the SINCERITIES (Papili Gao et al. 2017) algorithm integrates the two-state model into its algorithm. However, the *random telegraph* model is not in itself a GRN model, since no coupling is defined. A preliminary step would be the definition of a stochastic GRN model, as proposed in Herbach et al. (2017) and that we explain in section 1.3.3.1.

1.3.3.1. *A mechanistic approach based on the stochastic two-state model*

From the two-state stochastic expression model, we derived a simpler hybrid model called *piecewise deterministic Markov process* (PDMP)

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{\text{on}}} 1, \;\; 1 \xrightarrow{k_{\text{off}}} 0 \\ M_i'(t) = s_{0,i} E_i(t) - d_{0,i} M_i(t) \\ P_i'(t) = s_{1,i} M_i(t) - d_{1,i} P_i(t) \end{cases} \quad [1.7]$$

In this model, only the promoter (E) is defined by a stochastic process. RNA (M) and protein (P) are continuous variables whose evolution is defined by differential equations, which depend on the state of the promoter. Then, in order to model a GRN with causal interactions, we defined an interaction function based on mechanistic assumptions, which expresses the transition frequencies k_{on} and k_{off} of the regulated gene according to the regulatory gene protein concentrations. It is this model, called coupled PDMP, which will be used in (Bonnaffoux et al. 2019), presented in section 1.6.3 for inference based on the concept of wave of expression.

With a view to a mathematical approach to the inference problem (Herbach et al. 2017), we have derived a statistical model capable of giving an approximation of the joint distribution of RNAs in its steady state. This approximation does well to reproduce the stationary distributions of the coupled PDMP model, in the case of a *toggle switch* GRN, that is, two genes which self-inhibit creating a bistability, as shown in Figure 1.4.

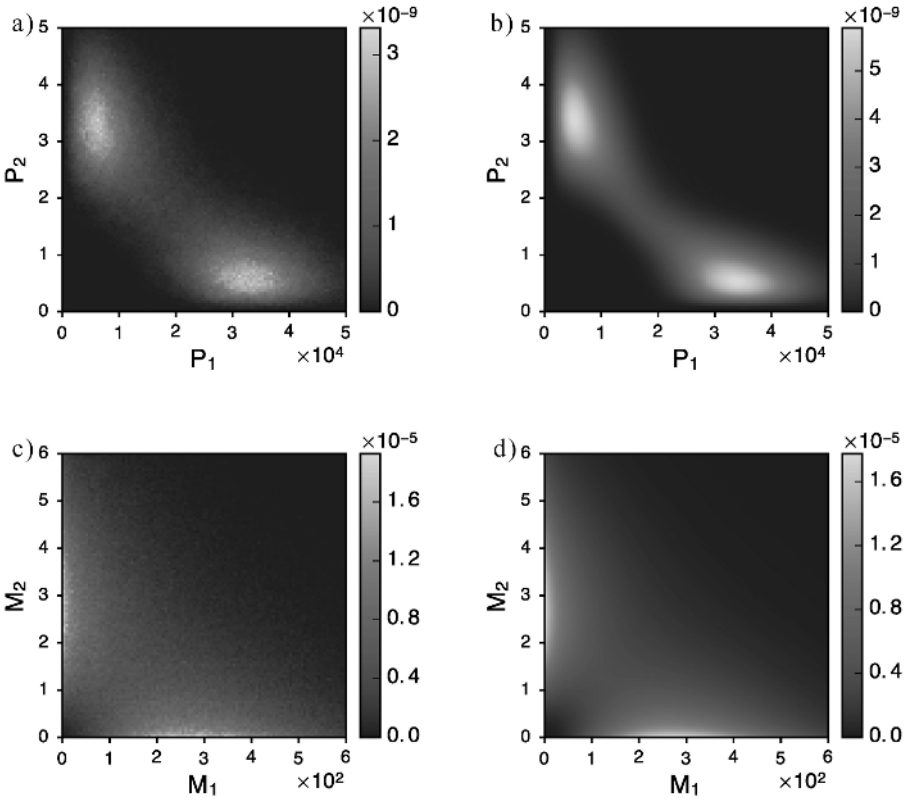

Figure 1.4. *Exact and approximate stationary distributions in the "toggle switch" example. The exact distributions (a, c) are estimated by numerical simulations. The approximate distributions (b, d) have an explicit formula. (a) Exact distribution of proteins. (b) Approximate distribution of proteins. (c) Exact distribution of RNAs. (d) Approximate distribution of RNAs. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

This statistical model also matches (Figure 1.5) the stationary marginal distribution in vitro of a gene well, measured in the context of the article (Richard et al. 2016).

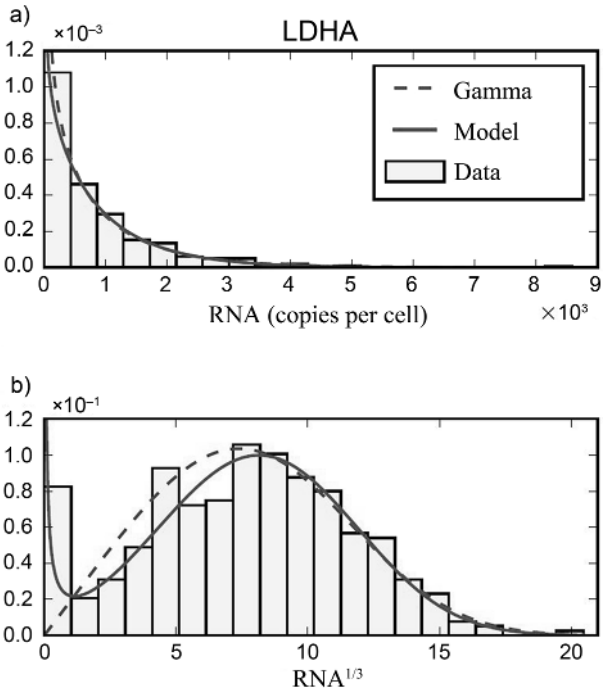

Figure 1.5. *Cross-checking marginal distributions from experimental data in single cells: example on the LDHA gene. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.5.– *The red curve is the stationary distribution associated with the simple network of a self-activated gene. The dotted blue line corresponds to the simple two-state model in the* bursty *regime (Gamma distribution). (a) The experimental data seem to reproduce a Gamma type distribution, which is close to our model. (b) We apply the transformation to the data* $x \mapsto x^{1/3}$. *The data appears bi-modal and the overlap is much better with the self-activation model.*

The statistical model was then used in a likelihood-maximization approach to infer small two-gene GRNs. The in silico experimental data were generated from the coupled PDMP model. Ten small GRNs were all successfully inferred. Therefore, we obtained a mechanistic stochastic model of GRN in the form of coupled PDMPs, which is a good compromise between biological representativeness and simplicity. By adding a stationarity assumption, it is, in practice, possible to infer small GRNs by deriving a statistical model.

1.4. Alternative strategies for GRN inference

In the previous sections, we wanted to provide an overview of different strategies for inferring GRNs from population or single-cell-level data. We could also have explained the neural network approaches (Blasi et al. 2005; Xu et al. 2007; Lee and Yang 2008), which are part of the family of statistical approaches. They are based on a learning process and require a large amount of data. However, in practice, many algorithms combine the methods we have seen and so it is difficult to classify them in one of these categories. For example, Zhao et al. (2006) is a Boolean approach which proceeds to a preliminary step of Bayesian inference to reduce combinatorics.

Another interesting approach is the integration of heterogeneous data. All the methods mentioned so far are based on the analysis of a single type of data based on the quantification of RNAs. Yet there are other informative experimental measurements in genomics, epigenomics and proteomics. We now have at least 10 different techniques for studying the epigenetic regulations of individual cells (Clark et al. 2016). Among the most well-known ones, we can note ChiP-seq for DNA-protein interactions, DNase-seq and ATAC-seq for chromatid structure, and HiC for the three-dimensional organization of chromosomes (Clark et al. 2016). In proteomics, there are new methods, such as CITE-Seq, which seem promising and already allow access to a large part of the protein section of a single cell (Stoeckius et al. 2017). The idea is that these so-called *omic* data contain some information on the regulation of gene expression. But if analyzed separately, these data do not provide an interconnected view of the mechanisms. This is why it is necessary to integrate all this *multi-omic* data in an integrative analysis. The existing methods are based on data mining or Bayesian statistical methods. For more information on these integrative methods, we recommend this recent review (Wani and Raza 2018).

Another strategy is the adaptation of these methods to the significant parallel computing capabilities that are currently available. The idea is to divide up the problem in order to be able to parallelize it and then partly bypass the problem of combinatorics. One strategy is to form subgroups of genes with the assumption that they belong to a relatively interconnected subset of the network. The idea is to infer the sub-GRNs at the same time, and then possibly connect them later, as proposed in Abduallah and Wang (2017). This strategy has also been applied with neural networks (Noman et al. 2013).

1.5. Performance and limitations of GRN inference

Following the significant growth in the number of GRN inference algorithms, many comparative test studies have been carried out (Hecker et al. 2009; Chai et al. 2014; Chen and Mar 2018), and even an international competition *DREAM Challenge* has been dedicated to GRN inference (Marbach et al. 2012). All these

tests are based on the reconstruction of known GRNs from in silico models or in vitro models that are supposed to be experimentally validated. In some cases, there are even synthetic GRNs that have been genetically engineered into bacteria (di Bernardo et al. 2005). The inferred GRNs are then compared to the real GRNs using different indicators. The conclusions of these comparative studies vary in their enthusiasm for the performance of the algorithms, but overall we note that there is no algorithm that stands out significantly from the others and that the inference capacity is slightly better than by chance. This criticism is harsh and disappointing, especially since methods based on single-cell data analysis do not do much better than their predecessors (Chen and Mar 2018). These results do not mean that these approaches are useless for research, as shown by the successes (Moignard et al. 2015), which advance scientific knowledge. But the aim of *reverse-engineering* is still far from being achieved.

In addition, we must be careful about the credit given to these comparative studies. As the organizers of the DREAM challenge themselves point out in their introduction (Stolovitzky et al. 2007), the question of the evaluation of inference methods is a problem in itself. The in silico models used are certainly far from being representative of biological reality, especially since they do not integrate the *burst* regime of genetic expression. The definition of new standard in silico models based on the *random telegraph* model would be a step forward for the generation of single-cell data. For the supposed known and experimentally validated biological GRNs, there is a fundamental contradiction. If there were an experimental GRN inference method, even a very expensive and long one, the problem of inference would be solved, which is obviously not the case. These studies are only based on non-formal interpretations of experiments that have consensus in the scientific community. When it comes to synthetic GRNs, we cannot ignore the interactions with endogenous GRNs, which must disrupt them.

We could also address the issue of commonly used evaluation metrics, like *receiver operating characteristic* (ROCs) that focus on the number of interactions identified, but not on the overall topology of the GRN. We will come back to this point later. The only acceptable method of validating a GRN inference algorithm is through repeated experimental validation of predictions generated by the inferred GRNs. A model is like a theory in physics, we cannot prove it to be true, but we can just test its predictions to try and reject it. This iterative method is precisely one of the major aims of so-called systems biology that we wish to apply in the medium term. We will return to this point in section 1.7 when we examine the perspectives.

Although the critical analysis of existing algorithms is not simple, as we have just shown, we can nevertheless summarize a list of their main limitations that we have discussed in this introduction:

– The use of correlations for the inference of interactions is problematic and they should not be considered as causalities. Correlations can only reproduce what has

previously been observed. Therefore, the making predicitions using GRNs in response to a new stimulus or a modification of its structure is impossible.

– Making predictions using simulation can only be done if the topology of the network is clearly defined. Algorithms in previous studies (Cordero and Stuart 2016; Matsumoto et al. 2017; Papili Gao et al. 2017; Sanchez-Castillo et al. 2017) propose interactions associated with independent confidence indices, usually in the form of an interaction matrix. The network topology combinatorial obtained from these matrices is far too large to be able to simulate them all.

– Very often, the regulatory proteins considered in GRNs are limited to transcription factors as in Ocone et al. (2015); Cordero and Stuart (2016); Lim et al. (2016); Matsumoto and Kiryu (2016); Matsumoto et al. (2017); Sanchez-Castillo et al. (2017). Indirect interactions, as illustrated in Figure 1.2, are completely ignored while they play such an essential role as transcription factors.

– Most algorithms use only one type of data, in this case mRNA measurements, assuming that the protein level is correlated. There are many examples of post-translational regulation that contradict this approximation (Gallego and Virshup 2007) on the circadian clock.

– The choice of biological hypotheses that are too simplistic is also a fundamental limitation of the performance of GRN inference. They are often justified by the use of powerful statistical tools that are capable of analyzing data from thousands of genes in thousands of cells, but the price to pay in terms of biological representativeness is definitely too high.

1.6. Inference based on the wave of expression concept

In order to overcome a number of the limitations of GRN inference discussed above, we adopted a mechanistic and numerical approach, done frequently in engineering. It involves analyzing its dynamics, that is, its transitory regime during a transition from one stationary state to another after stimulation.

By taking up the PDMP mechanistic model previously introduced in Herbach et al. (2017), we approached the problem of inference from a different angle, using dynamic signal processing concepts from engineering, with the aim of overcoming several current limitations of GRN inference.

We will begin by presenting some assumptions of signal processing that will be useful for understanding the concept of “waves” of expression.

1.6.1. *The differentiation process seen as a dynamic process of signal processing by GRNs*

This section is a brief introduction to the principles of signal processing applied to GRNs. For readers who wish to get more details on this subject, we recommend reading the book by Vecchio and Murray (2016), which is an excellent introduction.

In the field of signal processing, a system is defined by its ability to process signals (such as filtering), to analyze them (logical operations) and to interpret them (comparison with a context). Such system can be modeled by an ODE, as we have already presented in section 1.2.5.1. Using the notations of the equation [1.5], we are now going to show that the biochemical reactions of creation and degradation of molecules behave like a low-pass filter of the first order and induce a delay in the transmission of information, as summarized in Figure 1.6.

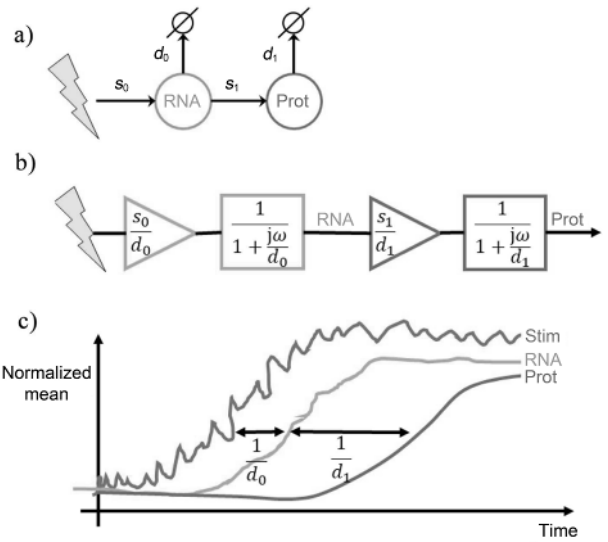

Figure 1.6. *Gene expression seen as a signal processing process. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.6.– *a) Classic representation of stimulus-induced population-level gene expression. The parameters s_0 and s_1 represent the transcription and translation rates, respectively. The parameters d_0 and d_1 represent the RNA and protein degradation rates, respectively. (b) Representation of the same process using the notions of signal processing. The signal passes through the successive stages of amplification and filtering. The triangles represent amplification*

with the corresponding gain. The squares represent first-order low-pass filters with the associated transfer function. (c) Example of a typical temporal trace of the average of RNAs and proteins after a noisy stimulation. The average of the RNAs is less noisy than the stimulus and delayed by $\frac{1}{d_0}$. Proteins filter the RNA signal with a delay of $\frac{1}{d_1}$. In our biological model, proteins generally have half-lives of around 20 h, unlike RNAs which have half-lives of around 4 h.

GRNs involve the production and degradation of mRNA (denoted M) and protein (denoted P). In the classic deterministic model where a stimulus (denoted U) induces the expression of a gene, we can write the dynamics of this system by the following differential equations (in this case the state vector is defined by $X = [M, P]$):

$$\begin{cases} \dfrac{dM}{dt} = s_0 U - d_0 M \\ \dfrac{dP}{dt} = s_1 M - d_1 P \end{cases} \quad [1.8]$$

To simplify the next part, only the equation relating to proteins is considered. Passing into the frequency domain and using the Laplace transform, where a derivation amounts to multiplying by $j\omega$ (with ω the pulsation), this equation becomes:

$$j\omega P = s_1 M - d_1 P \quad [1.9]$$

We then deduce the transfer function between the RNAs and the proteins:

$$\frac{P}{M} = \frac{s_1}{d_1} \frac{1}{1 + j\frac{\omega}{d_1}} \quad [1.10]$$

We find the typical form of the transfer function $H(\omega)$ of a low-pass filter of the first order of the cutoff frequency ω_c, which is recalled as follows (we also give the associated stage $\Phi(\omega)$ which corresponds to the stage delay generated by the filter):

$$H(\omega) = \frac{K}{1 + j\frac{\omega}{\omega_c}} \quad , \Phi(\omega) = -\arctan(\frac{\omega}{\omega_c}) \quad [1.11]$$

We deduce that the process of production/degradation of proteins from mRNAs is equivalent to a low-pass filter of the first order with gain of $K = \frac{s_1}{d_1}$ and cutoff frequency $\omega_c = d_1$. Low-pass filters have the property of strongly reducing the components of the signal with a frequency higher than the cutoff frequency. Another

property, which is more interesting, is the delay $\tau(\omega)$, induced by the filter on a pulse signal ω, which is given by:

$$\tau(\omega) = -\frac{\Phi(\omega)}{\omega} = \frac{\arctan(\frac{\omega}{\omega_c})}{\omega} \quad [1.12]$$

In the case where $\omega << \omega_c$, we can approximate the delay to:

$$\tau(\omega) = \frac{1}{\omega_c} = \frac{1}{d_1} \quad [1.13]$$

Therefore, the information contained by the mRNAs will be spread at the protein level, with a delay corresponding to the inverse of the rate of protein degradation. The same reasoning can be carried out between the stimulus and the mRNAs. In fact, the total delay between the stimulus and the proteins can be approximated by the sum of the inverses of the mRNA and protein degradation rates.

1.6.2. *Experimental demonstration of waves of expression*

First, we characterized the evolution of expression at the cell level on our biological model of differentiation, which is, in itself, a transitional regime. The initial objective was to study the evolution of the variability of gene expression, but we also took the opportunity to demonstrate the existence of delays in the transmission of information during interactions between genes. It is the principle of waves of expression that will be used in the next section for GRN inference.

As we saw in the introduction, different theoretical models suggest that the process of differentiation is accompanied by an increase in the variability of gene expression (Kupiec 1997; Kaern et al. 2005; Huang 2011). Through experimentation, our team analyzed the variability of gene expression during the process of erythrocyte differentiation, because of recent technologies allowing access to the molecular content of several cells, individually. This analysis was the subject of a publication (Richard et al. 2016).

First, the expression of 110 genes, selected on the basis of an RNASeq analysis previously carried out in the team, was measured by RTqPCR in populations of erythrocyte progenitors called T2EC (Gandrillon et al. 1999) in a state of self-renewal (0 h) or inferred in differentiation for 8 h, 24 h, 48 h and 72 h. In a population, the variability between cell populations is explained by the stage of differentiation. However, single-cell analyses on a subset of 92 genes measured on 96 cells at the same stages of differentiation show that intercellular variability is only explained by differentiation a little bit. Therefore, there is another very significant source of internal stochasticity.

In order to test the hypothesis that commitment to differentiation leads to a strong increase in the variability of gene expression, we used the measurement of entropy. For each differentiation time, we calculated an entropy value per gene and compared the distribution of these values during differentiation. Our results show that entropy increases significantly at 8 h, remains stable until 24 h, and gradually decreases until 72 h, and so suggests that erythrocyte differentiation is accompanied by a peak in the variability of intercellular gene expression at 8–24 h of the process.

We then looked for a cause for this increase in variability. We excluded a cell cycle variation, however, a volume variation was observed at 48 h, that is, 40 h after the increase in entropy, which suggests that the volume variation is rather a consequence of the variation entropy. Finally, we demonstrated, using the two-state model of gene expression, that the asynchrony between cells during differentiation cannot be the cause of the transitory increase in variability because the simulations show a decrease in the entropy peak instead when considering asynchrony.

The measurement of the peak of variability between 8 h and 24 h suggests the idea of a commitment of the cell. To test this hypothesis, T2ECs were inferred to differentiate for 24 h or 48 h, then returned to self-renewal environment. The results of this experiment show that after 24 h of differentiation, the cells are still capable of self-renewal, whereas after 48 h, they no longer multiply. The point of no return for the commitment of T2ECs in the differentiation process therefore seems to be between 24 h and 48 h.

This dynamic behavior then led us to propose the hypothesis that the information of the differentiation stimulus spreads in the network with a certain inertia. In order to validate this hypothesis, we compared the marginal distributions of each gene between the different times, which showed that some genes were regulated between 0 h and 8 h, then another group between 8 h and 24 h and so on. These results have not been published, but in order to identify the first group of regulated genes, called *early genes*, a kinetics on the initial response was specially carried out to measure the expression of the genes in single cells at 0 h, 2 h, 4 h and 8 h after differentiation. It made it possible to identify *early genes* that were activated between 0 h and 2 h, then between 2 h and 4 h, as shown in Figure 1.7.

Among these genes, we specifically found a cluster involved in the synthesis of sterols, which had already been observed on the kinetics in the population.

This work by Richard et al. (2016) shows very clearly that at the cell scale, stochasticity is very important from a quantitative point of view and it evolves dynamically with a transient peak that comes before cell engagement. On the other hand, the effect of differentiation on expression is obvious at the population level. These results clearly show the relevance of using a stochastic GRN model to analyze the expression data in a single cell in order to reproduce individual variability, but

also to constrain this stochasticity by the interactions to guide differentiation at the population level. The two-state model used for the in silico simulations clearly shows its relevance.

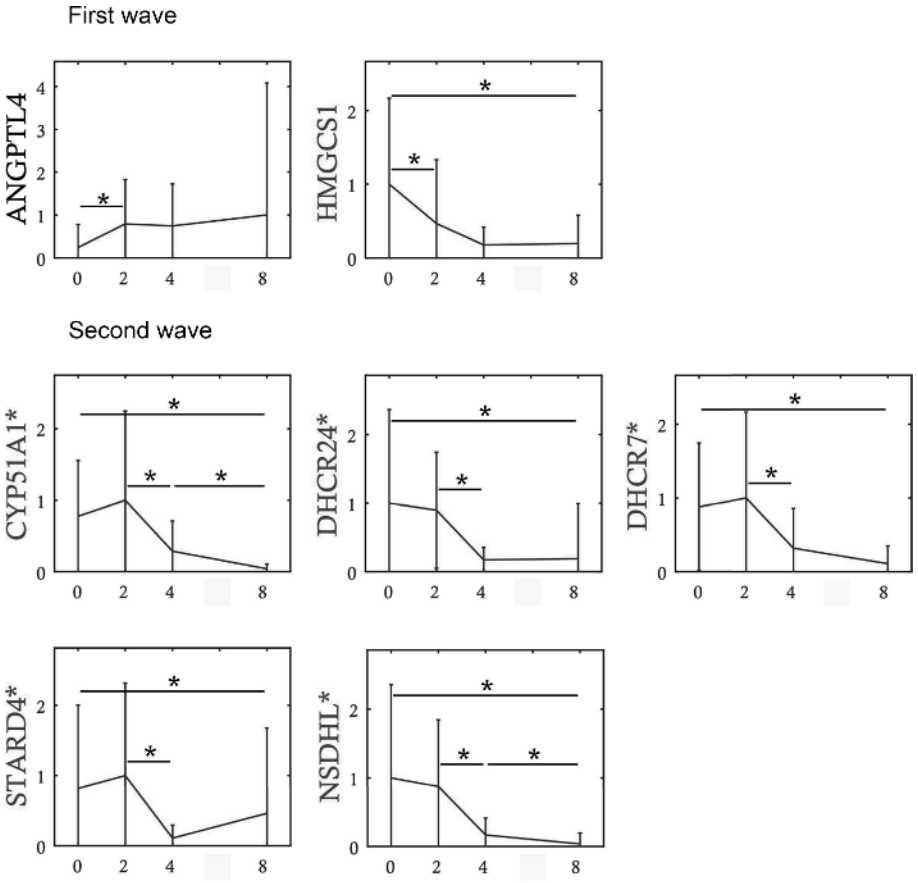

Figure 1.7. *Analysis of the first waves of expression. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.7.– *The genes are sorted according to the time of their first significant change in expression. The first wave corresponds to a detection between 0 h and 2 h after stimulation, the second to the genes presenting a significant variation detected during 2 h and 4 h, but without significant variation detected previously. The genes marked in red belong to the group of genes associated with sterol synthesis. Significant variations (-*-) are detected by the non-parametric Mann–Whitney test*

*(p-value < 0.05) if the test is positive in more than 90% of the 1,000 bootstrap samples. Genes prefixed with * have variation between 0 h and 8 h detected in both experiments (0–72 h, as well as 0–8 h). The probability of having six genes out of seven (in the first and second wave) belonging to the 10 genes of the sterol group among the 90 is estimated at* $p = 1.8 \times 10^{-6}$*, with the hypergeometric probability density function.*

The observation of "waves of expression" in the GRN made it possible to experimentally validate the concept of the wave presented above, which will be the central assumption of the GRN inference strategy in section 1.6.3.

1.6.3. *Using waves of expression for GRN inference*

From the concept of waves of expression validated in experiments in Richard et al. (2016), we have developed an algorithm called WASABI (*Waves Analysis Based Inference*). The principle described in Figure 1.8 is based on the fact that the genes are activated one after the other, according to the rule that the cause comes before the effect. We can then infer the causalities of the GRN in iterative steps by adding the genes one by one, following the chronological order of regulation. This iterative method allows us to cut and parallelize the problem of GRN inference. A preliminary step of estimating the individual parameters of each gene (such as the half-lives of RNAs and proteins) as well as their regulation time is necessary.

This algorithm is first applied and validated for in silico GRN inference, which is composed of six genes with different topologies. The experimental data are simulated using the coupled PDMP model from Herbach et al. (2017). After approximately 24 h of computation on 400 machines at the same time, WASABI returns several dozen candidates, including the real GRN. The self-activation loops and negative feedbacks present in GRNs were inferred well, showing the WASABI's ability to infer circular causalities.

The algorithm is then applied to in vitro single cell expression data from Richard et al. (2016), but also from population transcription inhibition kinetics data to estimate the half-lifetimes of RNAs, as well as kinetics in population proteomics (Leduc et al. 2018), to estimate the parameters relating to proteins. After 16 days of calculations on 400 machines, 364 candidate GRNs are generated with similar topologies (Figure 1.9) characterized by:

– a main role of the stimulus;

– a majority of inhibition by the stimulus on its targets;

– the absence of the *hub*;

– the presence of several *incoherent feedforward loops*;

– limited network depth;

– a high proportion of positive loops.

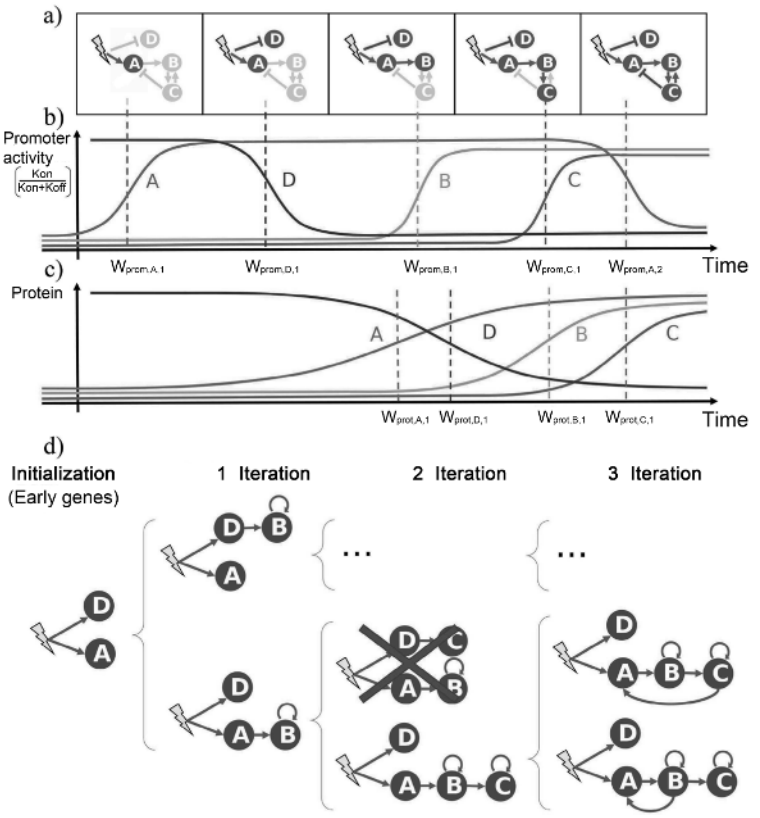

Figure 1.8. *Principle of wave inference: WASABI. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.8.– *(a) Schematic view of a GRN: the stimulus is represented by a yellow flash, the genes by blue circles and the interactions by green (activation) or red (inhibition) arrows. The spread of information induced by the stimulus is represented by blue arcs corresponding to wave times. Genes and interactions that are unaffected by information at a given wave time are shaded. At wave 5, the C gene feeds back information to the A and B genes using feedback, creating a feedback wave. (b) Promoter wave time: corresponds to the inflection points of the gene promoter activity, defined as the ratio $k_{\text{on}}/(k_{\text{on}} + k_{\text{off}})$. (c) Protein wave times: correspond to the inflection points of the average protein level. (d) Inference process: the blue arrows represent the interactions selected for*

calibration. Based on promoter waves, classification genes are iteratively added to subarrays that were deduced before to obtain new extended GRNs. The calibration is done by comparing the marginal RNA distributions between the in silico *and* in vitro *data. The inference is started by calibrating the early genes interacting with the stimulus, resulting in the initial subRNA. The last genes are added one by one to a subset of potential regulators, for which the wave time of the protein is sufficiently close to the wave time of the promoter of the added gene. Each resulting sub-GRN is selected based on its adjustment distance to* in vitro *data. If the adjustment distance is too large, the sub-GRNs can be eliminated (red cross). A significant advantage of this process is the possibility of making the sub-GRN calibrations parallel over several nuclei, which makes it possible to obtain a linear computation time, with respect to the number of genes. Only a fraction of all tested sub-GRNs are shown.*

These results show that the proposed WASABI algorithm overcomes some limitations, such as:

– causality inference, even circular causality;

– generating candidate GRNs with explicit topologies, that include the stimulus and that can be simulated;

– the ability to integrate post-translational regulation (observed for half of the genes) thanks to the analysis of proteomics data;

– the flexibility and the "curse of combinatorics" because of the iterative approach.

1.6.4. *Scaling up the distributed computing approach*

The WASABI strategy is based on the parallelization of inference, so its performance depends on the computing power used. Cloud-type solutions offer great computing power, but task flow management for scientific applications, such as WASABI, is a complex issue studied by the AVALON team at ENS de Lyon. We have collaborated on this study (Croubois et al. 2018) to apply and test their Cloud parallel computing optimization solution on a task flow simulation generated by WASABI.

The WASABI task flow has been tested and simulated with different types of task allocation, one static and the other automatic, with the DIET algorithm, developed by the AVALON team[1]. The main benefit comes from cost reduction. With static management on an Amazon cloud, the cost is just under $2,000 for WASABI inference, which is significant. Automatic management can greatly reduce this cost by 45% for the same work.

1. Available at: https://graal.ens-lyon.fr/DIET.

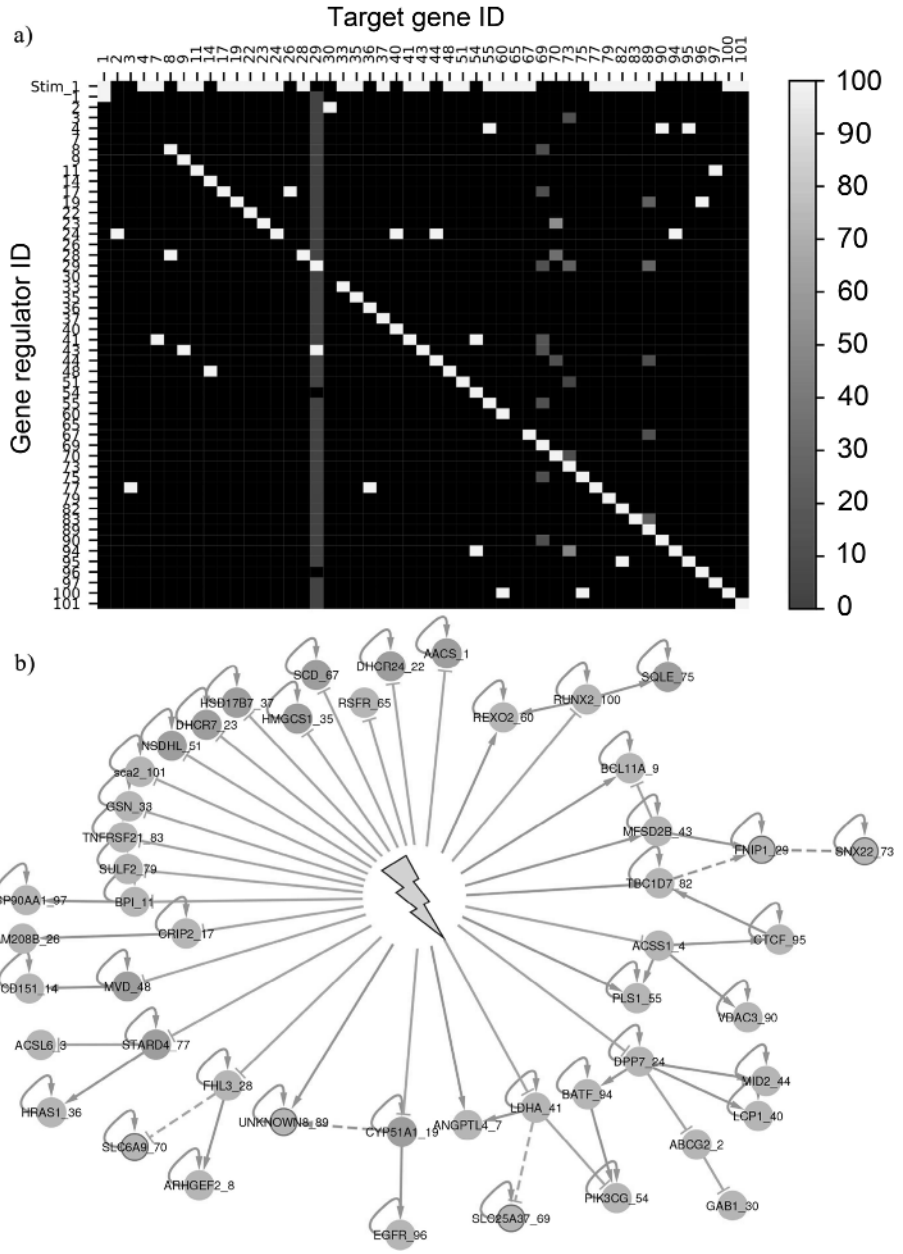

Figure 1.9. *Application of WASABI on in vitro data. For a color version of this figure, see www.iste.co.uk/lhoussaine/symbolic.zip*

COMMENT ON FIGURE 1.9.– *(a) in vitro Interaction consensus matrix. Each square in the matrix represents either the absence of any interaction, in black, or the presence of an interaction, the frequency of which is color-coded, between the ID of the regulator considered (line) and the ID of the regulated gene (column). The first line corresponds to the stimulus interactions. (b) Best candidate. Green: positive interaction; red: negative interaction; solid lines: interactions found in 100% of candidates; dotted lines: interaction found only in some of the candidates; orange: genes whose product contributes to the sterol synthesis approach; purple: last five genes added during the iterative inference.*

On the one hand, this work shows that it is possible to greatly reduce the cost of implementing an algorithm such as WASABI on commercial computing platforms, and, on the other hand, it shows that this implementation is adaptable and can be managed automatically on any platform to facilitate their use. These results confirm that the transition to a more industrial scale of WASABI is possible.

1.7. Conclusion

With this work, we hope to have contributed to a solution to the problem of GRN inference. The aim is to provide biologists with integrative and scalable tools based on engineering concepts. Even if WASABI still needs to be improved, it is part of a long-term process that relies on the exponential evolution of technologies in cellular and molecular biology. Only a cross-integration of current and future data and knowledge will make it possible to move forward step by step, especially our understanding and mastering complex biological systems.

1.8. References

Abduallah, Y. and Wang, J.T.L. (2017). A time-delayed information-theoretic approach to the reverse engineering of gene regulatory networks using apache spark. In *2017 IEEE 15th Int. Conf. on Dependable, Autonomic and Secure Comput., 15th Int. Conf. on Pervasive Intell. and Comput., 3rd Int. Conf. on Big Data Intell. and Comput. and Cyber Sc. and Techno*, Orlando.

Akutsu, T., Miyano, S., Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *World Scientific*, 99, 17–28.

Akutsu, T., Miyano, S., Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8), 727–734.

Albert, R. (2004). *Boolean Modeling of Genetic Regulatory Networks*. Springer, Berlin.

Ando, S., Sakamoto, E., Iba, H. (2002). Evolutionary modeling and inference of gene network. *Information Sciences*, 145(3–4), 237–259.

Apostel, J.L. (2003). Classification of random Boolean networks. *Artif. Life*, 8, 1–8.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., White, K.P. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297(5590), 2270–2275.

Babtie, A., Chan, T.E., Stumpf, M.P.H. (2017). Learning regulatory models for cell development from single cell transcriptomic data. *Curr. Opin. Syst. Biol.*, 5, 72D81.

Bansal, M., Gatta, G.D., Di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7), 815–822.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4), 382.

Benenson, Y. (2012). Biomolecular computing systems: Principles, progress and potential. *Nature Reviews Genetics*, 13, 455–468.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795), 536–540.

Blasi, M.F., Casorelli, I., Colosimo, A., Blasi, F.S., Bignami, M., Giuliani, A. (2005). A recursive network approach can identify constitutive regulatory circuits in gene expression data. *Physica A: Stat. Mechanics Its Applications*, 348, 349–370.

Bonnaffoux, A., Herbach, U., Richard, A., Guillemin, A., Gonin-Giraud, S., Gros, P.-A., Gandrillon, O. (2019). Wasabi: A dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics*, 20(1), 1–19.

Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10(11), 1093–1095.

Butte, A.J. and Kohane, I.S. (1999). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *World Scientific*, 2000, 418–429.

Cannoodt, R., Saelens, W., Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.*, 46(11), 2496–2506.

Chai, L.E., Mohamad, M.S., Deris, S., Chong, C.K., Choon, Y.W., Ibrahim, Z., Omatu, S. (2012). Inferring gene regulatory networks from gene expression data by a dynamic Bayesian network-based model. In *Distributed Computing and Artificial Intelligence*, Omato, S., Neves, J., Rodriguez, J.M.C., Santana, J.F.P., Gonzalez, S.R. (eds). Springer, Berlin.

Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S., Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.*, 48, 55–65.

Chen, S. and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1), 232.

Chen, K.-C., Wang, T.-Y., Tseng, H.-H., Huang, C.-Y.F., Kao, C.-Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics*, 21(12), 2883–2890.

Chen, H., Guo, J., Mishra, S.K., Robson, P., Niranjan, M., Zheng, J. (2015). Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. *Bioinformatics*, 31(7), 1060–1066.

Ching, W.-K., Zhang, S., Ng, M.K., Akutsu, T. (2007a). An approximation method for solving the steady-state probability distribution of probabilistic Boolean networks. *Bioinformatics*, 23(12), 1511–1518.

Ching, W.-K., Zhang, S.-Q., Jiao, Y., Akutsu, T., Wong, A. (2007b). Optimal finite-horizon control for probabilistic Boolean networks with hard constraints. *Lect. Notes Op. Res.*, 7, 21–28.

Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G., Reik, W. (2016). Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. *Genome Biology*, 17, 1–10.

Cordero, P. and Stuart, J.M. (2016). Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories. *World Scientific*, 576–587.

Corre, G., Stockholm, D., Arnaud, O., Kaneko, G., Viñuelas, G., Yamagata, Y., Neildez-Nguyen, T.M.A., Kupiec, J.-J., Beslon, G., Gandrillon, O., Paldi, A. (2014). Stochastic fluctuations and distributed control of gene expression impact cellular memory. *Plos ONE*, 0115574, 9(12).

Croubois, H., Caron, E., Bonnaffoux, A., Gandrillon, O. (2018). A cloud-aware autonomous workflow engine and its application to gene regulatory networks inference. In *8th International Conference on Cloud Computing and Service Science*, Funchal.

Davidson, E.H. (2010). *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Elsevier, Amsterdam.

Di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, 23(3), 377.

Dojer, N., Gambin, A., Mizera, A., Wilczyski, B., Tiuryn, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinf.*, 7(1), 249.

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z., Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief Funct Genomics*, 17(4), 246–254.

Friedman, N., Linial, M., Nachman, I., Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3–4), 601–620.

Gallego, M. and Virshup, D.M. (2007). Post-translational modifications regulate the ticking of the circadian clock. *Nat. Rev. Mol. Cell. Biol.*, 8(2), 139–148.

Gandrillon, O., Schmidt, U., Beug, H., Samarut, J. (1999). TGF-β cooperates with TGF-α to induce the self-renewal of normal erythrocytic progenitors: Evidence for an autocrine mechanism. *The EMBO Journal*, 18(10), 2764–2781.

Gardner, T.S., Di Bernardo, D., Lorenz, D., Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629), 102–105.

Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6), 1025–1036.

Grzegorczyk, M. and Husmeier, D. (2010). Improvements in the reconstruction of time-varying gene regulatory networks: Dynamic programming and regularization by information sharing among genes. *Bioinformatics*, 27(5), 693–699.

Hakamada, K., Hanai, T., Honda, H., Kobayashi, T. (2004). A preprocessing method for inferring genetic interaction from gene expression data using Boolean algorithm. *Journal of Bioscience and Bioengineering*, 98(6), 457–463.

Harvey, I. and Bossomaier, T. (1997). Time out of joint: Attractors in asynchronous random Boolean networks. In *Proceedings of the Fourth European Conference on Artificial Life*, Husbands, P. and Harvey, I. (eds). MIT Press, Cambridge.

He, L. and Hannon, G.J. (2004). Micrornas: Small rnas with a big role in gene regulation. *Nature Reviews Genetics*, 5(7), 522–531.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models – A review. *Biosystems*, 96(1), 86–103.

Herbach, U., Bonnaffoux, A., Espinasse, T., Gandrillon, O. (2017). Inferring gene regulatory networks from single-cell data: A mechanistic approach. *BMC Syst. Biology*, 11(1), 1–15.

Hickman, G.J. and Hodgman, T.C. (2009). Inference of gene regulatory networks using Boolean-network inference methods. *J. Bioinf. Comp. Biol.*, 7(06), 1013–1029.

Huang, S. (2011). Systems biology of stem cells: Three useful perspectives to help overcome the paradigm of linear pathways. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1575), 2247–2259.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1), 109–126.

Jacob, F. and Monod, J. (1961). On regulation of gene activity. *Cold Spring Harbor Symposia on Quantitative Biology*, 26, 193.

Jopling, C., Boue, S., Belmonte, J.C.I. (2011). Dedifferentiation, transdifferentiation and reprogramming: Three routes to regeneration. *Nat. Rev. Mol. Cell Biol.*, 12(2), 79.

Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.*, 6(6), 451–464.

Kauffman, S.A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3), 437–467.

Kim, J.K. and Marioni, J.C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, 14(1), 1–12.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. *Mol Cell*, 58(4), 610–620.

Kunga, T.A. and Mohamada, M.S. (2012). Using Bayesian networks to construct gene regulatory networks from microarray data. *Jurnal Teknologi*, 1–6.

Kupiec, J.J. (1997). A darwinian theory for the origin of cellular differentiation. *Molecular and General Genetics MGG*, 255(2), 201–208.

Larson, D.R., Singer, R.H., Zenklusen, D. (2009). A single molecule view of gene expression. *Trends in Cell Biology*, 19(11), 630–637.

Leduc, M., Gautier, E.-F., Guillemin, A., Broussard, C., Salnot, V., Lacombe, C., Gandrillon, O., Guillonneau, F., Mayeux, P. (2018). Deep proteomic analysis of chicken erythropoiesis. *BioRxiv*.

Lee, W.-P. and Yang, K.-C. (2008). A clustering-based approach for inferring recurrent neural networks as gene regulatory networks. *Neurocomputing*, 71(46), 600–610.

Levine, M. and Davidson, E.H. (2005). Gene regulatory networks for development. *Proc. Nat. Acad. Sci. USA*, 102(14), 4936–4942.

Levsky, J.M. and Singer, R.H. (2003). Gene expression and the myth of the average cell. *Trends in Cell Biology*, 13(1), 4–6.

Liang, S., Fuhrman, S., Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 3, 18–29.

Lim, C.Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., Gottgens, B. (2016). Btr: Training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics*, 17(1), 355.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13), 1675–1680.

Manning, K.S. and Cooper, T.A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nat. Rev. Mol. Cell. Biol.*, 18(2), 102–114.

Marbach, D., Costello, J.C., Kuffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Consortium, D., Kellis, M., Collins, J.J. et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods*, 9(8), 796–804.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 1–15.

Marshall, S., Yu, L., Xiao, Y., Dougherty, E.R. (2007). Inference of a probabilistic Boolean network from a single observed temporal sequence. *EURASIP Journal on Bioinformatics and Systems Biology*, 1–15.

Matsumoto, H. and Kiryu, H. (2016). Scoup: A probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*, 17(1), 232.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T., Nikaido, I. (2017). Scode: An efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics*, 33(15), 2314–2321.

Mattick, J.S., Taft, R.J., Faulkner, G.J. (2010). A global view of genomic information – Moving beyond the gene and the master regulator. *Trends in Genetics*, 26(1), 21–28.

Mercer, T.R., Dinger, M.E., Mattick, J.S. (2009). Long non-coding rnas: Insights into functions. *Nature Reviews Genetics*, 10(3), 155–159.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E. et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, 33(3), 269–276.

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I., Leong-Quong, R., Chang, H., Giuliani, A., Huang, S. (2016). Cell fate-decision as high-dimensional critical state transition. *BioRvix*.

Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J., Marra, M.A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45(1), 81–94.

Noman, N., Palafox, L., Iba, H. (2013). Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model. In *Natural Computing and Beyond*, Suzuki, Y., Nakagaki, T. (eds). Springer, Berlin.

Ocone, A., Haghverdi, L., Mueller, N.S., Theis, F.J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12), i89–i96.

Olsen, J.V. and Mann, M. (2013). Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics*, 12(12), 3444–3452.

Orkin, S.H. and Zon, L.I. (2008). Hematopoiesis: An evolving paradigm for stem cell biology. *Cell*, 132(4), 631–644.

Panning, B. and Jaenisch, R. (1998). Rna and the epigenetic regulation of x chromosome inactivation. *Cell*, 93(3), 305–308.

Papili Gao, N., Ud-Dean, S.M.M., Gandrillon, O., Gunawan, R. (2017). Sincerities: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2), 258–266.

Pe'er, D., Regev, A., Elidan, G., Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1), S215–S224.

Pedraza, J.M. and Paulsson, J. (2008). Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861), 339–343.

Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16, 241.

Polynikis, A., Hogan, S., Di Bernardo, M. (2009). Comparing different ODE modelling approaches for gene regulatory networks. *Journal of Theoretical Biology*, 261(4), 511–530.

Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J. et al. (2016). Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol*, 14(12), e1002585.

Rousseau, G.G. (1975). Interaction of steroids with hepatoma cells: Molecular mechanisms of glucocorticoid hormone action. *J. Steroid Biochem.*, 6(1), 75–89.

Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I.M., Carrion, M.C., Huang, Y. (2017). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, 34, 964–970.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–470.

Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W. (2002). Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2), 261–274.

Silvescu, A. and Honavar, V. (2001). Temporal Boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems*, 13(1), 61–78.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12), 3273–3297.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(2), S231–S240.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, 14(9), 865–868.

Stolovitzky, G., Monroe, D., Califano, A. (2007). Dialogue on reverse, engineering assessment and methods. *Annals of the NY Academy of Sciences*, 1115(1), 1–22.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028), 472–474.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4), 663–676.

Tyson, J.J., Hong, C.I., Thron, C.D., Novak, B. (1999). A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophysical Journal*, 77(5), 2411–2417.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.*, 20(5), 515–524.

Vecchio, D.D. and Murray, R.M. (2016). *Biomolecular Feedback System*. Princeton University Press, Princeton.

Vinh, N.X., Chetty, M., Coppel, R., Wangikar, P.P. (2012). Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network. *BMC Bioinformatics*, 13(1), 131.

Wani, N. and Raza, K. (2018). Integrative approaches to reconstruct regulatory networks from multi-omics data: A review of state-of-the-art methods. *Computational Biology and Chemistry*, 83, 107120.

Warren, L., Bryder, D., Weissman, I.L., Quake, S.R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Nat. Acad. Sci. USA*, 103(47), 17807–17812.

Werhli, A.V. and Husmeier, D. (2007). Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Wimberly, F.C., Heiman, T., Ramsey, J., Glymour, C. (2003). Experiments on the accuracy of algorithms for inferring the structure of genetic regulatory networks from microarray expression levels. In *Proc. IJCAI Bioinformatics Workshop*.

Wu, H. and Liu, X. (2008). Dynamic Bayesian networks modeling for inferring genetic regulatory networks by search strategy: Comparison between Greedy Hill climbing and MCMC methods. *Proc. World Acad. Sci., Engineer. and Technol.*, 34, 224–234.

Xu, R., Venayagamoorthy, G.K., Wunsch, D.C. (2007). Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. *Neural Networks*, 20(8), 917–927.

Yamanaka, S. (2012). Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell*, 10(6), 678–684.

Yang, B., Zhang, J., Shang, J., Li, A. (2011). A Bayesian network based algorithm for gene regulatory network reconstruction. In *2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, Xi'an.

Yavari, F., Towhidkhah, F., Gharibzadeh, S. (2008). Gene regulatory network modeling using Bayesian networks and cross correlation. In *2008 Cairo Intl Biomed. Engin. Conf.*, IEEE, Cairo.

Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18), 3594–3603.

Zhao, W., Serpedin, E., Dougherty, E.R. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, 22(17), 2129–2135.