Genetic regulatory networks inference : modeling, parameters estimation & model validation

Fomekong Nanfack, Y.

Publication date
2010

1

Introduction

Since it has become possible to identify the structure of Deoxyribonucleic acid (DNA) and, to sequence an organism's genome, one of the biggest challenges in today's biology is to decipher the genomic role in the structure, function development and, evolution of simple to complex organisms. Biology used to be perceived as a *data-poor* science where traditional approach mainly focused in understanding single element (genes, proteins, cells,...). The technical advances in molecular biology turned it into a *data-rich* field it became conceivable to conduct data-driven research. The enormous amount of data produced at different molecular levels grows too large to fully apprehend the ongoing process by traditional standard biological approaches or manual manipulations. New fields such as bioinformatics established themselves as keymethods in the management, computational statistical analysis of the data. Theoretical approaches such as mathematical formalisms also contributed in understanding the functional mechanism under biological systems. One emerging field that considerably gained attention the last decades is system biology. The main concern of systems biology is to study an ensemble of elements as an ensemble (or "a system"). The field aims at revealing the connections between elements, their dynamics, as well as the mechanism behind their evolution ultimately the precise role of each individual within the system as well as the entire system's functionality. A typical problem where studying a system, as a complete entity is essential is the mechanism of early development in multicellular organisms.

In many animals, morphogen gradients influence the movement organisation of cells that lead to the morphogenesis in early developmental embryogenesis. The morphogens provide spatial information by forming concentration gradients that subdivide the developing embryo into different regions. Distinct cell types and structures emerge because of the different combinations of mor-

phogen gradients. This is a general mechanism by which cells can generate type diversity and structures in body plan formation. Understanding the body plan formation also requires a comprehensive knowledge of the underlying biochemical process. This is the level at which genes influence the transcription of other genes. Genetic regulatory networks (GRNs) are an ensemble of interconnected genes that control the dynamic of gene expression level for each gene in the genome. Understanding how do GRNs control the mechanism that leads to a specific phenomena such as body patterning requires knowledge of the active genes their interconnection structure. By means of systems biology techniques, it is now possible to infer the GRNs involved in the mechanism of body plan formation in some organisms in early development.

In this chapter, we briefly discuss the mechanism of early development in Section 1.1. Section 1.2 discusses the basic principles of GRNs Section 1.3 reviews models methods for reverse engineering of regulatory networks from gene expression data.

1.1 Mechanism of early development

From the moment sperm fuses with an ovum in animals or, after the pistil is pollinated in plants, the process of fertilization starts. At this point, in most organisms, starting from a single cell, the complex biological mechanism behind the formation of an embryo begins will lead to a grown organism. The fundamental question is how do we go from fertilized egg to an organism?

Developmental biology has a long history, but nowadays, modern developmental biology studies the genetic control of cell growth [138, 216], differentiation "morphogenesis," [28, 32]. Theses three processes are the basic mechanisms that give rise to tissues, organs and anatomy. Growth is a consequence of cell division. Two forms of cell division development occur within an organism: at the early stage, successive series of cell division increases the number of cells without consequently increasing the cell mass and, at the latest stage the cell mass increases during the cell division [256, 292]. More than 200 different type of cells have been identified among humans and other vertebrates. Cell specialisation is conducted through a step called differentiation. During this late phase, cell fate is established through variation of the gene expression causing the cells to have different shapes, activities functions that will make them recognisable and specialised [292]. The combination and association of different cell's fate will determine the function of a part of the body.

1.1.1 Morphogenesis

Recent advances in biology have considerably increased the level of understanding of the mechanism of early development [27, 290]. At a very early

stage, some specific chemical regions are created in the extreme of the body plan. These gradients secreted by cells are proteins called morphogenes. By diffusion and degradation mechanisms, they will spread into the extracellular matrix forming chemical gradients that can span the whole embryo [83]. Cells can sense these gradients by means of specific receptors or, the gradients bind to specific sites on DNA. In response, the cells adjust their transcription rate of the targeted genes in a concentration-dependent manner. By this mechanism, a specific chemical body map can be generated that brings specific cellular regions of the developing embryo in a different chemical state. This mechanism enables these regions to develop along different developmental pathways as shown in Fig. 1.1

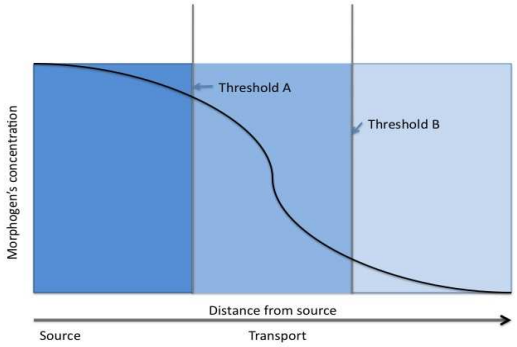


Figure 1.1: The concentration of the morphogen is produced by source cells at the left and transported along the body axis through diffusion and degradation. The two thresholds determine the level at which gene receptors (cytoplasmic receptors or membrane receptors) are sensitive to the morphogen. (After Wolpert [291])

The explanation of the early patterning and pattern formation by means of morphogen gradients has initiated the theory of positional information proposed by Wolpert [290]. However, many details about the molecular mechanisms of morphogen production, transport, or dynamics are still unclear [115].

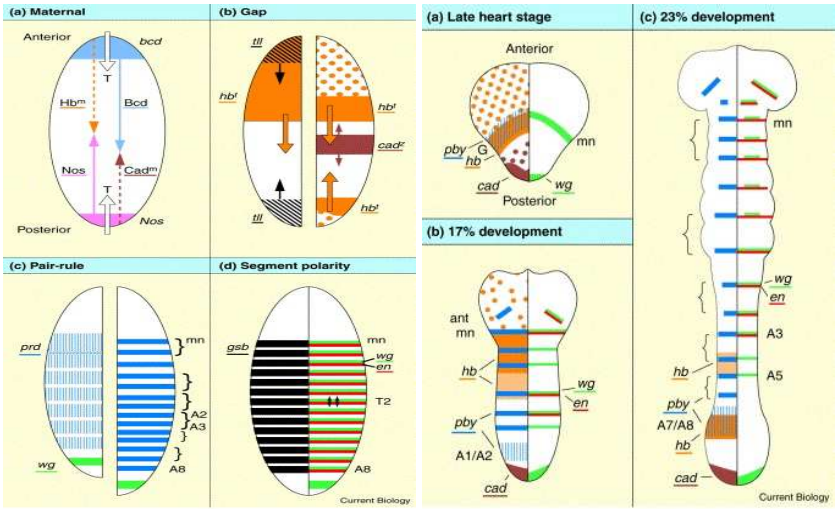
1.1.2 Basic principals of segmentation in animals

One of the striking example of organisms that have extensively been studied and has a clear segmentation mechanism is the fruit fly *Drosophila melanogaster* [38]. It is a little insect about 3mm long, of the kind that accumulates around spoiled fruit. It is also one of the most valuable of organisms in biological research, particularly in genetics and developmental biology. Its importance for

human health was recognized by the award of the Nobel prize in medicine and physiology to Edward B. Lewis, Christiane Nüsslein-Volhard and Eric Wieschaus in 1995 [199]. *Drosophila melanogaster* has been used as a model organism for research for almost a century, and today, several thousand scientists are working on many different aspects of the fruit fly. Body plan formation and pattern formation has been very well studied in *Drosophila melanogaster* [36,54,70,89,106,119,183,223,258,283]. Once morphogenes gradients are transported along a body axis, as for example the antero-posterior axis of the body (AP axis or head-tail) as shown in Fig. 1.1, and diffuse through the extracellular matrix, they will establish the location for some genes to be regulated. The precise location of the gene expression will consequently determine the beginning of the body segmentation. The pair-rule genes are the first set of gene expression that determine precise patterns showing future position of segments. Later on, they will regulate the segment polarity genes [38].

Although this feature is observed in many other organisms for which the body plan development is a consequence of multiple segmentation steps along the AP axis, the segmentation mechanism of *Drosophila melanogaster* does not necessary hold for all segmented animals. In vertebrates and also some insects, segmentation is formed sequentially from anterior to posterior during a long phase of growth and cell proliferation [9, 43, 84, 207] as show in Fig. 1.2 where parallel segmentation of *Drosophila melanogaster* and sequential segmentation of grasshopper are compared.

The process of segmentation along the AP axis is either by simultaneous or sequential formation of the segments. One of the fundamental questions is whether or not the organisms showing one or the other segmentation gene expression shares the same or a common molecular mechanism. Some early works compared the expression patterns of *Drosophila melanogaster* segmentation genes in other animals by means of comparative genomics. For instance, it was reported that similar mechanisms are present in insects showing simultaneous segmentation of the early embryo such as *Drosophila melanogaster* and insects having sequential subdivision such as beetles or grasshoppers [53, 84, 203, 204]. Also, it was shown that the vertebrate homolog's pair-rule gene *hairy* of *Drosophila melanogaster* is involved in zebrafish *Danio rerio* segmentation [188, 277] as well as the avian embryo chicken [202] suggesting that pair-rule patterning is an evolutionary process.



(a) *Drosophila melanogaster* (b) *Grasshopper Schistocerca*

Figure 1.2: Two extreme cases of segmentation gene expression in two different insects species. Embryos are shown in a ventral view with anterior at the top. (a) Sequential segmentation gene expression in *Drosophila melanogaster*. Developmental stages are shown. a.a:) maternal genes localisation during oogenesis. a.b:) Gap gene expression are expressed after maternal genes have diffused through the embryo. a.c:) Expression of the pair-rule gene. a.d:) segment polarity gene at stage 6/7. (b) Segmentation gene expression in the grasshopper Schistocerca. b.a:) late heart stage: hunchback (hb) and pairberry-1 (pby) are strongly expressed in arcs in the future gnathal region. b.b:) By 17% of development time, hunchback expression has resolved into high and lower level bands (indicated by colour intensity) and pairberry1 is expressed in thin stripes in antennal, gnathal and thoracic segments, and in a broad posterior stripe that will resolve later into thin stripes in the 1st and 2nd abdominal segments. b.c:) By 23% of development time, the anterior hb bands have faded, but there are two further bands of expression in the future abdomen. pby is now expressed in thin stripes in segments down to the 6th abdominal, and in a broad posterior stripe. Image from V. French [84].

1.2 Gene expression

In the previous section, we have briefly introduced the basic principle of body plan formation. It is important to mention that this mechanism is principally controled by a biochemical process. This is the level at which genes influence the transcription of other genes. A gene is the basic functional unit of the genome, which consists of long molecules of DNA made up of chains in a double-helix structure. The gene can be defined as the information stored in a sequence of a DNA region and it is required to transcribe the RNA into protein. Gene expression stands for the gene information translated into a particular protein. Proteins are the fundamental structure that fulfil functional units in

cells such as a structural element, enzyme catalyst or antibody. The translation of gene into proteins is carried out through two main steps: transcription and translation.

Transcription Within a gene, one region contains the information about the regulation time and another coding region specifies the shape and amount of proteins that will be produced after the gene has been activated. The proteins produced at these locations are known as transcription factors (TFs). This mechanism is slightly different between eukaryotes and prokaryotes [259]. In prokaryotes, the coding region is contiguous and the regulatory region is generally located directly upstream of the coding region while in eukaryotes elements of the regulatory region are located at a considerable distance both upstream and downstream from the coding region.

Translation Once the DNA has been transcribed into a complementary messenger RNA (mRNA), the molecules bind to another large molecule called a ribosome. The function of the ribosome is to read an mRNA molecule in triplets known as codons. The codons will then map to one 1-20 possible amino acids. After a very brief time, mRNA and proteins are broken down and their constituent nucleotides and amino acids are reused. They are degraded at different rates according to the presence or absence of chemical species present in the cell.

1.2.1 Genetic Regulatory Network

Spatiotemporal gene expression is the activation of genes within specific tissues of an organism at specific times during development. Gene activation patterns vary widely in complexity. In some cases, the pattern is expressed in all cells at all times of the organism life time. In other cases, it is extraordinarily intricate and difficult to analyse and predict, where the expression is fluctuating wildly from minute to minute or from cell to cell. Spatiotemporal variation plays a key role in generating the diversity of cell types found in developed organisms; since the identity of a cell is specified by the collection of genes actively expressed within that cell. In developmental biology, a fundamental question is "What causes spatial and temporal differences in the expression of a single gene?" The current expression pattern depend on the previous expression patterns. The inverse question is, how does the previous pattern form? By symmetry breaking mechanism, uniform gene expression becomes spatially and temporally differential. In the case of early embryonic *Drosophila melanogaster* development, the genes nanos and bicoid are asymmetrically expressed in extreme opposite locations of the embryo where maternal cells have deposited mRNA for these genes. (see left panel of Fig. 1.2)

There are many techniques to identify the expression pattern of a particular gene depending on the identification of the gene's promoter. If a gene's pro-

moter is known, a reporter gene downstream of its promoter is placed. The promoter gene will initiate the reporter gene to be expressed only where and when the gene of interest is expressed. The expression distribution of the reporter gene can be identified by visualizing it. If the promoter of the gene of interest is unknown, there are several ways to identify its spatiotemporal distribution. Immunohistochemistry involves preparing an antibody with specific affinity for the protein associated with the gene of interest. This distribution of this antibody can then be visualized by a technique such as fluorescent labelling shown in Fig. 1.3

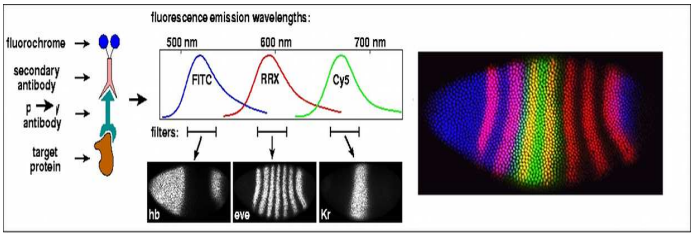


Figure 1.3: Three fluorochrome-tagged secondary antibodies label three primary antibodies, which in turn recognize three transcription factor proteins, hunchback, Kruppel, and eve (left). The fluorochromes indodicarbocyanine used emit light in different parts of the spectrum, so that three separate images of the embryo can be collected with the appropriate color filters in gray-scale mode. Each of these image represents the expression pattern of a single protein. The images can be color-coded and merged to visualise the spatial relationships between the patterns are easily perceived (right). Mixture of colours allows to distinguish overlapping gene patterns. Modified from Huges et al. [113]).

Once the spatiotemporal gene expression of some genes can be obtained from experiments, understanding this formation is not necessary trivial. This requires a clear knowledge of the gene regulatory network that controls the gene's regulation. A genetic regulatory network (GRN) is an ensemble of DNA segments present in a cell that interact with one another and other substances [52]. GRNs dynamically "manage" the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation. A simple GRN would consist of alternative input such as signalling pathways or regulatory proteins that regulate several target genes. The output or products of the target genes are RNA and proteins product as illustrated in Fig. 1.4. In addition, feedback loops are often included for further regulation of network architecture. Transcriptional control in the various differentiated cell types allows each type of cell to express different amounts of the possible proteins. Signal transduction pathways that will relay signals from outside of cells to the cell nucleus regulate TFs. Signal transduction pathways often involve receptors, receptor lig-

ands and enzymes such as protein kinases. One key class of genes that are differentially regulated by transcription factors in different cell types are genes for cell adhesion proteins. Cell adhesion proteins are among the key regulators of morphogenesis. Understanding the network of genes' interactions. Large networks of regulatory genes typically control the development of the body plan.

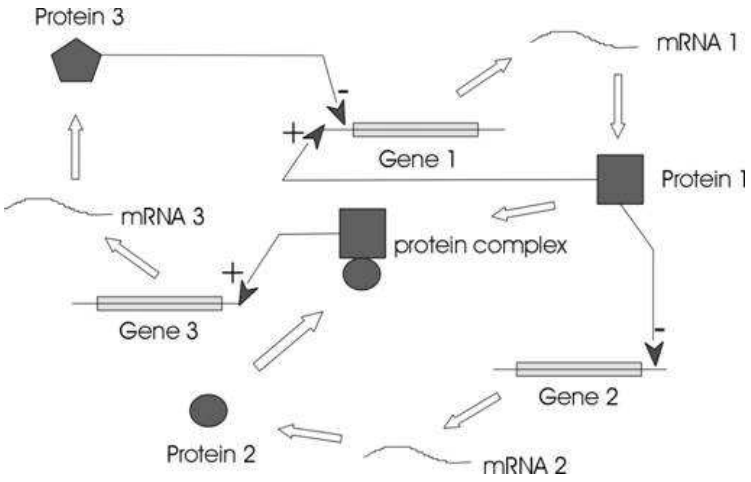


Figure 1.4: Simple representation of gene regulation.

Currently the genetic regulation of development in a number of model organisms is known in great detail, as for example: *Drosophila melanogaster*, sea urchins, *Caenorhabditis elegans*, ascidians. Recently much information has become available on the genetic regulation of growth and form in sponges [150] and cnidarians [180]. Within the metazoans, sponges and cnidarians represent the phyla with the simplest body plan and a relatively simple regulatory network controlling the development. This makes these organisms an excellent case study for understanding morphogenesis and the physical translation of the genetic information into a growth form, using a combination of a biomechanical model of cell aggregates and a model of the spatial and temporal expression of developmental genes.

1.2.2 Quantitative measurements of Gene expression

Many techniques were developed to measure gene expression in organisms [15] . One of the most prolific method to asses the expression of particular gene is DNA microarray technology [146]. This technique can provide a rough measure of the cellular concentration of different mRNAs; often thousands at a time. A DNA microarray is a collection of microscopic DNA spots attached

to a solid surface, such as glass, plastic or silicon chip forming an array. Usually, microarrays are used to quantify mRNAs transcribed from different genes which encode different proteins. Another important technique to identify gene expression is serial analysis of gene expression (SAGE) [172,280] tag sequencing. SAGE is a technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest.

All of these techniques generate extremely noisy data that are in most of the cases subject to bias in the biological measurement. A major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput (HT) gene expression studies.

A more sensitive and more accurate method of relative gene expression measurement is Real-Time PCR. The real-time PCR system is based on the detection and quantization of a fluorescent reporter [158,163]. This signal increases in direct proportion to the amount of PCR product in a reaction. With a carefully constructed standard curve it can even produce an absolute measurement (e.g., in number of copies of mRNA per nanolitre of homogenized tissue, or in number of copies of mRNA per total poly-A RNA).

Other techniques have been developed including, massively parallel signature sequencing (MPSS) [29,178], or by measuring protein concentrations with high-throughput mass spectroscopy. Expression data is also used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. Further analysis could take a variety of directions: for instance Beer et al. [16] analyzed the promoter sequences of co-expressed (clustered together) genes to find common regulatory elements and used machine learning techniques to identify the promoter elements involved in regulating each cluster.

1.3 Reverse-engineering gene regulatory systems

For many centuries, traditional reductionist biological sciences have focused their research in understanding organisms by studying theirs constituents as individual component. Since the important review of morphogenetic theories proposed by Ludwig von Bertalanffy [19], the idea of biological systems as a self-organization dynamics has opened doors to new concepts. One main obstacle in biology research used to be the limited quantity and poor quality of data produced, due to technology constraints. Consequently, biology was mainly a hypothesis-driven research field. The rapid technological ad-

vance that happened the last couple of decades has allowed the production of massive quantitative and qualitative data, leading to other research approach, mainly the data-driven research. One of the new foci is to find patterns and the underlying mechanisms behind their formation in the quantities of information produced from the molecular biology revolution trying to obtain a more system-level understanding. The discovery of the structure of the DNA [284] has allowed the identification of a large number of genes and their transcription factors in a large number of organisms. Later on, the technical advance leads to exact methods that facilitate the measurement of gene expression profiles giving information at the mRNA level, and also the potential interaction among genes [248].

1.3.1 Systems biology

The data-driven approach used to investigate biological mechanisms in organisms can be divided into two categories: the collection/analysis of large genomic, proteomic or metabolic data sets, and the mathematical modelling of complex biological systems. Systems biology is a relative new field that aims to combine these two fields for a complete integrated understanding of the biological process at different level (cellular, organism or evolutional). Systems biology is an interdisciplinary field that became an important assets in the study of biological systems as an entire "complex system" [251, 262]. The main aims are:

1. structure identification of the elements present in the system: the network of interaction.

2. elements 'dynamics: how does their state change in time and under which condition.

3. biological mechanism behind the elements' dynamics

Ultimately, system biology aims at understanding how these systems evolve and how to construct artificial systems [143].

Any attempt to infer gene regulatory networks requires at list a partial knowledge of the genes involved. Although some methods such as cluster and statistical analysis [60] allow to partially find the underlying network of gene-gene interactions from the measured dataset of gene expression, it is essential to determine the role and importance of each component. In addition to experimental observations, mathematical models and simulation models are an important option to obtain an understanding of regulatory networks. In an iterative process as shown in Fig. 1.5, they can be used to understand the function of the very complex processes involved in the development of organisms. This includes simulation of cell signalling, multicellular interactions and regulatory genomic networks in development of multicellular structures and

processes. Combined, mathematical modelling and quantitative experimentation has already provided useful information on the role of GRNs and/or how do the control systems based on molecular signals generate pattern and govern the timing of developmental embryos in organism such as *Drosophila melanogaster*, frog embryo by Bone Morphogenetic Protein signals, the auxin-mediated patterning of plant meristems, and the Notch-dependent somite segmentation (see [159,221] for reviews).

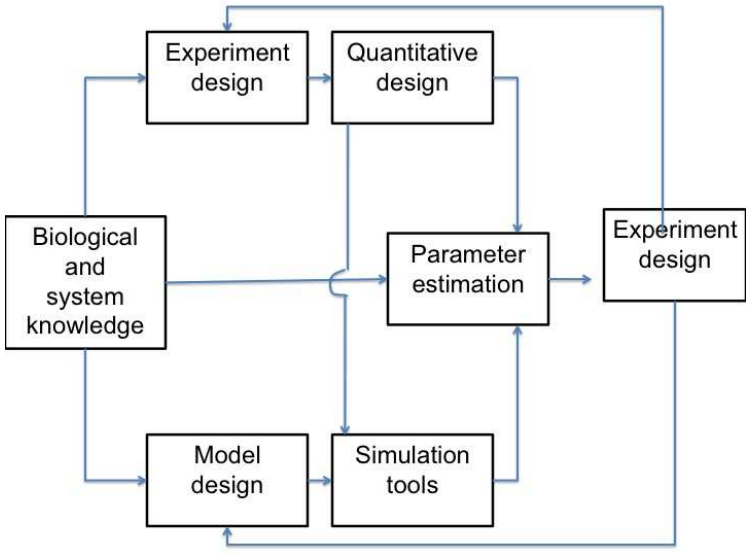


Figure 1.5: The diagram shows a typical systems biology scenario. First step is the collection of biological knowledge from literature. In parallel, experiments and models are designed to measure quantitative data and describe mathematically describe the system. By means of computational tools and parameter estimation, the system is simulated. Last step is the experiment design where analysis, interpretation and validation are performed. From this step, knowledge is gain and new experiments can be conducted.

1.3.2 Mathematical model

There are numerous techniques to model genetic regulatory networks (GRNs). The choice of modelling formalisms depends on the type of biological problem and properties of interest, the type and amount of the data, and also the prior information available [141]. Existing models can be broadly categorized as continuous vs. discrete and deterministic vs. stochastic, dynamic vs. static. Extensive reviews discuss the principles and motivations behind each of these modelling formalisms [55,74,87,239]

Dynamic vs. Static models When temporal aspect of the gene expression variation is investigated, dynamic models are used. In most cases, dynamic GRNS are modelled by means of differential equations or by boolean networks describing the gene concentration change over time. Static models do not handle the temporal component. Typical statistical models are graph theoretic models [185] or Bayesian networks [3]. Lately, Bayesian networks became more and more appealing for the inferring of regulatory network structure from gene expression data [85, 170]. Some of the limitations of these Bayesian networks are the inability to handle the direction of gene interactions and the acyclicity constraint, which rules out feedback loops.

Continuous vs. Discrete models When dynamical models are considered, the description of the time transition between two different states is important. This transition can be modelled continuously or through discrete transition. Discrete models such as Boolean network [26, 135] or logical formalisms [176, 270] discretise the time into a fix number of quantitative states. The temporal evolution of the values of the variables is defined by logical equations which is a combination of Boolean functions such as AND, OR or XOR. The values of each variable (presence or absence of a molecular species) depend on the preceding values of the variables of the system. The difficulty in reverse-engineering boolean network is to determine individual boolean functions for each "node" or gene. Methods [174] and algorithms such as REVEAL [161] are general reverse engineering methods that have been developed for this purpose. One of the main advantages of Boolean network models lies in its simplicity. When the elements of interest as well as their potential regulatory interactions are known, they are easy to interpret and to simulate gene regulatory events. As it is a discrete formalism, the analytical tractability can be defined and the simulation remains simple. This make them appropriate and useful to infer GRN when the input data are very noisy due to their weak sensitivity to measurement errors. The first and important limitation of boolean modelling is its perceived lack of applicability to biological systems. The "synchronism" does not reflect real biological situation. Another important disadvantage of this approach is the basis of these models: how do we validate the assumption. It is known that a gene can have different regulatory effect depending on their quantitative expression level. Also, as mentioned by Klipp et al. [144] it is very difficult to infer the "real" network from continuous data using Boolean networks. It generally leads to many "false" interactions

Continuous representation of the time dynamic can be formalised using differential equations such as ordinary differential equations (ODEs), partial differential equations (PDEs) or time delay different equations. There is a long history of using systems of differential equations to model the reaction kinetics of regulatory systems [42, 75, 104, 107, 296]. These approaches have several advantages. In principle, their more detailed representation of regulatory interactions provides a more accurate representation of the physical system under

investigation. In most of biological situation, the interactions considered are non-linear (effect of threshold, saturation, interaction synergistic, etc), which leads to models that are hardly solvable using analytical methods. It is then very impossible to derive the solutions from these systems and one must resort to numerical simulation.
Continuous models using partial differential equations (PDEs) can be used when attempting to model the dynamic of developmental phenomena involving GRNs and when considering the physical space in which gene regulation is occurring. They consist of chemical rate equations describing the regulation of gene within a cell and the diffusion of gene products between neighbouring cells. Recently, Salazar-Ciudad et al. [236] have used a continuum model to model regulatory networks to:

1. compare diffusion and direct-contact induction processes as mechanisms of pattern formation
2. identify the possible range of behaviour of real gene networks
3. suggest causal mechanisms to generate known patterns

Deterministic vs. Stochastic models Deterministic models such as ODEs do not describe the molecular fluctuations present in the system and assume that proteins are produced at a continuous rate. Biological systems must be relatively insensitive to variation or noise [4]. Biochemical reactions in many cases involve a low number of molecules [201, 266, 295]. Therefore, noise is an integral part of GRNs and deterministic assumptions may not hold or may be insufficient to capture all the dynamics [217]. Stochastic modelling approaches allow describing the stochastic events within the gene expression. Two main formalisms are commonly used: stochastic differential equations (SDEs) and probabilistic modelling. SDEs are an extension of ODEs, with an additional term describing the noise. Probabilistic model are based on master equations describing the time evolution of the system. Independent variables of master equations are time the population of reacting species. The master equation can be transformed into a partial differential equation by the use of a generating function. Although the master equations provide a clear description of the stochastic process ruling the dynamics of a regulatory system, it is still more difficult than ODEs and even impossible to find analytical solutions. Also, because of their stochastic nature, it is necessary to solve SDEs a large number of times for statistic purpose. Models using SDEs have been until now applied to small molecular networks [4, 186]. An interesting review summarizing the issues and technical aspects of stochastic model can be found in [217].

1.3.3 Inference by parameter estimation

A significant problem with the numerical approach is the lack of measurement of the various kinetic parameters in a system. The number of systems for which

detailed parameter values are known is very small, and the size of most systems makes it unfeasible to obtain in vitro or in vivo measurements of many parameter value of all parameters involved.

If the parameters and the initial conditions are known the time evolution of the gene expression patterns can be simulated by numerical integration of the partial differential equations and the properties of the model can be further studied. However, in most cases the precise parameter values are not known and methods are required to estimate the parameters. One way to obtain the parameters is by direct measurement, however in most cases this cannot be done experimentally because the individual regulatory processes are not easily isolated. An alternative approach is detection of cis-regulatory elements (interaction sites), which may be directly measured or predicted from the DNA sequences using bioinformatics techniques. Although this technique is useful to constrain the number of interactions, the precise parameter values cannot be estimated. Furthermore, in phenomenological models, which are used most of the time, the parameters often do not represent biophysically measurable quantities. Therefore a different approach is used; parameter inference is formulated as an inversion problem. Given a detailed set of experimentally measured expression patterns (observations) the parameters are varied such that the difference between the simulated and the experimentally observed expression patterns is minimised.

The standard optimisation approach minimizes a single cost function that represents the difference between the simulated and observed expression pattern. In most studies the least square method has been used, where parameters are estimated by minimizing the root mean square error (RMS). In these studies box constraints for the parameters are used to reduce the search space and also to ensure that the parameter values are within biophysical constraints. Because the response function possesses asymptotes at strong repression and strong activation a penalty function can be used to ensure that an inverse solution (i.e. the parameters) exists.

Non-linear inverse problems with many parameters are notoriously ill defined, many different solutions may be obtained that all fit the dataset equally well. This means that the objective function to be minimised possesses many local minima with similar "energy" values. There are two main sources for this. First, if the data does not contain enough information (missing genes, missing time points, only part of the space is modelled and there is also a limited time window) the parameters may not be identifiable, which can be observed as correlations between parameters that compensate each other because they act similarly in the model. Secondly, if the data contains features, e.g. artefacts, noise or certain processes, which are not represented in the model the optimisation strategy, will lead to spurious parameter values in the model because each parameter in the model is indiscriminately used to fit the data as good as possible. However, in the real system these parameters may not be related to

explaining these features at all.

Qualitative analysis of the model parameters is used to validate the estimated parameters. Precision and sensitivity of the parameters is an important aspect [51]. Also, parameter variances and covariances can be used to verify the overall validity of a model in representing the data and to ensure the significance and determinability of its parameters [124]. These different validations techniques also allow to compare different inferences approaches or for models discrimination [62].

When one is concerned with dynamical models, different additional aspects such as the dynamic stability (robustness to parameter variations) [92, 241], the structural stability (ability to maintain the system trajectory) are essential for the system behaviour analysis [142]. These analyses schemes can improve the understanding of the mechanism behind bistability, bifurcation or hysteresis.

Ultimately, model validation should be carried out *in vivo* by experimentally testing hypotheses generated by the model prediction. So far, few have carried this complete system biology approach [25, 72, 114, 249]

1.4 Research questions addressed in this thesis

In this thesis, we investigate several aspects of the inference of GRNs capable of simulating spatio-temporal pattern. The main focus of this thesis is the problem of parameter estimation of mathematical models describing biological systems and the reliability of the inference. As a case study we use a quantitative spatio-temporal model of the regulatory network for early development in *Drosophila melanogaster*. This model is capable of simulating pattern of the early development of the organism and serves as a basis to investigate several aspects of the robust inference of GRNs by means of reverse engineering.

We investigate the efficiency of an evolution strategy for the parameter estimation of GRN models capable of simulating spatio-temporal patterns. Our choice is inspired by [175,184] where the authors compared different global optimisation strategies and suggested that the evolution strategy is a very promising and competitive method for the problem of parameter estimation of biochemical networks. We combine this appoach with a local search strategy to further improve the quality of the global search. The motivation behind the implementation of such algorithm is the need for a fast and efficient parameter estimation method for the reverse engineering of complex spatio-temporal model of GRNs.

Once parameters have been estimated, it is essential to address their reliability as well as the robustness of the model. We investigated the sensitivity and robustness of circuits obtained from reverse engineering of the regulatory network for early development in *Drosophila melanogaster*. We analyse the uniqueness of the predicted network and the model stability. We aim to investigate whether the model shows signs of over-fitting (uniqueness) and if this over-fitting leads to variable circuit behavior (stability). The goal is to extract the best set of circuits that can simulate realistically the patterns, but also obtain the most plausible topology consistent with biological evidence. Ultimately, our aim is to understand the dynamics behind the gap gene mechanism.