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

Fomekong Nanfack, Y.

Publication date
2010

8

Discussion

In this dissertation, we have addressed the problem of inferring genetic regulatory network (GRN) capable of simulating spatio-temporal gene expressions. We have used as a case study a model capable of simulating pattern of the early development of *Drosophila melanogaster*. This model served as a basis to investigate several aspects of the robust inference of GRNs by means of reverse engineering. In particular, this work focused on the parameter estimation, the sensitivity and the robustness of inferred circuits. This chapter reviews the different problems and results presented in this thesis and summarises their implications. From these conclusions, directions for future research on the possible improvement of the reverse engineering approach are discussed.

8.1 Overview

8.1.1 Parameter estimation by Evolution Strategy

The main focus of this thesis was the problem of parameter estimation of mathematical models describing biological systems and the reliability of the inference. In Chapter 2, an overview of the most frequent used approaches to estimate unknown parameter is given. We discussed methods to fit the parameters of a mathematical model to experimental data and to analyze the results. Unfortunately, we cannot recommend one or the other algorithm as *the* method to search for parameters. An optimal use of a method, especially of the global ones, is problem-dependent. In practice, convergence to the minimum is not guaranteed and if it does, it is at a very high computational cost. Some studies (see Section 2.2.4) suggested that it is convenient to combine global and local methods to achieve satisfactory results and performance. In chapter 3, we

have presented a hybrid optimization method based on evolutionary strategy as global search followed by direct search algorithm as local optimizer. The global search is based on the stochastic ranking evolutionary strategy (SRES) introduced by Runarsson and Yao [234]. We have extended this method by proposing an island-based (μ, λ) – ES.

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. In order to understand the mechanism behind the dynamical control of the gap gene segmentation, Jaeger et al. [120, 121] inferred a GRN model of a dynamical system with 62 unknown parameters from quantitative spatio-temporal expression data [211]. Although the model could be spatially reduced to one dimension, the fitting procedure was highly computationally expensive. Using parallel simulated annealing (PLSA) [45, 155, 156], it took between 8 to 160 hours on ten 2.4-GHz Pentium P4 Xeon processors [120, 121]. The computational cost of the PLSA limited the selection to ten gap gene circuits out of 40 results on the basis of their RMS (≤ 12). Using the approach presented in Chapter 3, in Chapter 4, we have demonstrated that our method, (μ, λ)-ES followed by a Downhill simplex search, leads to similar qualitative and quantitative solutions. Three main important results were derived from this chapter:

1. **Reduction of the computational time.** The (μ, λ)-ES followed by a local search, requires less computational time (8-11 CPU-hours), and less resources (one 3.4-GHz processor) to achieve solutions as good as the one obtained with PLSA (between 8-160 CPU-hours using 10 parallel 2.4-GHz processors), making our method 5–140 times as fast.

2. **Higher percentage of "good solutions"**. Because of the stochastic nature of ES or PLSA, many trials are mandatory in order to reveal good guesses, and also gain confidence in the results. Approximately 50% of the trials have good score compared to the 25% performed by PLSA presented in [120, 121].

3. **Sequential Island-based (μ, λ)-ES is preferable to (μ, λ)-ES**. The island-based (μ, λ)-ES approach followed by DS even increases the ratio "good solutions" to 62% using the same amount of work.

4. **Easy parallelization of the island-based (μ, λ)-ES.** Recently, Jostins et al. [129] have implemented a parallel version of the algorithm and have demonstrated that the percentage of good solution was higher (about 75%) compare to a PLSA, both computed with same amount of resources within the same time. These new findings corroborate our conclusions presented in [78]

8.1.2 Circuit analysis and robustness

Jaeger et al. [121] suggested that the anterior shifts in the position of the gap domain are based on a regulatory mechanism that relies on the asymmetric gap-gap cross-repression (see Section. 4.1.1). Their conclusions were derived from a graphical analysis based on 10 gene circuits obtained using a PLSA. The number of circuits (81) obtained from our optimisation in Chapter 4 allows for a more extensive analysis.

Pattern validation The first validation criterion is the quality of the least square error presented in Equation 4.5. If a circuit has a low RMS, it is selected for further analysis. In Chapter 5, we have presented a simulated pattern analysis of our results [78, 80] and those obtained by Jaeger et al. [120, 121]. Overall, all the circuits present very well fit with respect to the observed data. This combination leads to 101 circuits patterns that were analysed in two manners. First, we have analysed the gastrulation (final time point) patterns of the circuits by means of clustering technique. The relative simplicity of this approach permits the identification of circuits showing minor patterns' defection. It also considerably highlights the regulatory mechanisms that lead to bad patterning if present. As shown in Fig. 5.2, it is possible to detect if the surplus expression of a gene A in a specific region is linked to the surplus expression of gene B in the same region. From this analysis, we see that the *hb* anterior deep is caused by Gt repression. Consequently, Hb activates *gt* and to maintain locally posterior *gt* at its normal level, there is a surproduction of *tll* causing a bump (to compensate for Hb activation of *gt*). This bad patterning suggests that Gt should not repress *hb*. Similar analysis on the bad patterning of *kr* allows to identify that in the absence of Gt repression, *kr* presents a bump at *gt* anterior-posterior domain, suggesting that Gt should repress *kr*. We did not performed any experimental validation of these predictions, but we believe that by small mutation in cis-regulatory regions, it might be plausible to check if these regulations are qualitatively precise.

From the clustering analysis, we have seen that although all the circuits show "good" patterns, they do not necessarily present the same topologies. It is therefore essential to identify circuits that are more plausible to have a network corresponding to the assumed biological correct one. One approach to discriminate among the circuits presented in Section 5.2 is to focus on the long term dynamic of their patterns. We have assumed that one of the requirements for gap gene pattern stability at later times may be motivated by its cascade-like hierarchical network: the gap genes activate and determine the periodic patterning of the pair rule genes. Therefore, we have assumed that gastrulation time corresponds to the time at which the pattern reaches its steady state.
Few different alternative models (logical model, regulator Hill function) have been proposed to describe the establishment of maternal and gap proteins segmental patterning along the antero-posterior axis of the Drosophila early em-

bryo [2, 18]. Sanchez et al. [237] have used a logical model in which predefined functional threshold concentrations for the proteins were used and (gastrulation) was assumed. Alves et al. [2] have proposed a ODEs based model with fixed parameters in which they have assumed that the expression of the proteins of maternal origin remains at steady state. In 2007, Bergmann et al. [18] proposed a model based on individual threshold using Hill function but they failed in simulating the gap gene domains and dynamic. The main difference and strength of the gene circuits approach is to not explicitly describe the system's steady state and to have a dynamic completely driven by the regulatory mechanism.
The long-term dynamics of the circuits showed that the patterns converge to four main attractors (two stable and two oscillators). We have observed that one of the patterns attractors corresponds to the gastrulation pattern. We have shown that all circuits having their long term dynamic patterns converging to this attractor are those with the regulatory mechanism consistent with experimental evidence. From this observation, we suggest that long term dynamic can contribute in identifying robust circuits in terms of structural stability and using the gene circuit approach, we don not have to explicitly describe the steady state to capture the mechanism that leads to this stationary point. However, biologically speaking, the expression of the gap gene starts to shut down at around the final time, which is not the case in our simulations. The model could not simulate the transient genes and as far as we know, there is no available literature explaining the causes of the gap genes disappearance. We observed that in the presence of strong autoactivation, the gap genes are maintaining their gene expression after gastrulation instead of allowing a progressive fade. The inability of the circuits to predict this disappearance suggests that either an additional mechanism not present in the model is provoking the gap gene vanishment or the model failled to captured the complete network, if we assume that the gap gene is an autonomous system.

Robustness of the circuits In Chapter 6, we have examined the identifiability of the parameters by analysing the circuits' parameters. We have shown that most parameters seem to be well defined, few of them are considerably scattered. It is essential to have a clear confidence in the parameters to deduce the network. Parameters showing alternative sign suggest alternative regulation (activation/repression). The scattering of some parameters suggests that either many alternative networks can predict the pattern or the model is subject to an over-fitting. Assuming the second hypothesis, we have analysed the correlation among parameters and showed that they are strongly correlated.

The reverse engineering of complete artificial data obtained from the simulation of a circuit lead to circuits with very low RMS. Although the parameter scattering is reduced, we see from Fig. 6.14 that the parameters having a very broad scattering in the original results given in Fig. 6.1 still show some scattering. The correlation matrix given in Fig. 6.15 also shows complex correlation.

We observe that some of the parameters are still not determinable and therefore, we conclude that the poor identifiability must principally emanate from the model instead of the data.

In Chapter 7, we have investigated the robustness of the circuits with respect to parameter perturbation and stochastic fluctuation. Intrinsic noise in gene expression is present at the cellular level [48,269] resulting from mRNA and protein fluctuation and environmental conditions [230]. By introducing noise by means of stochastic simulation of the gene expression level, we have showed that the robust circuits are those with very precise domain formation. The goal was to analyse the stability to stochastic gene expression fluctuation of each of all the circuits. Ideally we expected that a "good solution" has its mean overlapped with the deterministic solution and the gene profile would have a very small deviation from the mean. Unfortunately it was not the case. Stochastic simulation of the circuits showed that the genes have different domains and boundary precisions. Nevertheless, some circuits were identified as very robust, independently of their regulatory interactions. It is suggested that these interactions cause gap domain boundaries to shift and are required to convert noisy early patterns into precise late ones [122,168]. Manu et al. [264] suggested that gap-gap cross regulations increase spatial precision by means of feedback mechanism. How results show that although the circuits predict precise gap domain boundaries, in the presence of noise, the system breakdown. The simple stochastic formalism used here provides preliminary evidences that the gap-gap interactions are not sufficient to explain the robustness of the gap gene patterning. However, it is known that maternal gene such as Bcd considerably contributes in the robustness of the gap gene [2,18,108,168]. In the current model, the gap gene was not properly used and this fact may contribute in the poor robustness of the circuits.

Dynamics of the patterns The circuits have different asymptotical behaviour (stable or oscillatory) with very few close to the expected steady-state pattern as shown in Figs. 5.6 and 5.7. We observed that in the presence of motifs such as positive-negative feedback and autoactivation, the network tends to lead to oscillatory patterns. In [121], it was suggested that *kni* and posterior *gt* shift is caused by an asymmetry gap -gap repression. Here we suggest that in combination with this asymmetric repression, *hb* expansion is strongly related to the shift and might be the trigger of the shift. We also showed that alternative networks leading to the same pattern with the dynamical shift, could be considered. It would be interesting to experimentally verify this assertion by an artificial increase the level of Hb protein at the posterior of the embryo.

Over-fitting Beside Hb patterning showing two different late-anterior profiles as illustrated in Fig. 5.2, all the others gap genes have good profiles that

are consistent with the observation. Nevertheless, the parameter distribution show major scattering for many parameters (see Fig. 6.3). In an effort to understand the cause of their broad distribution, through correlation analysis, we have demonstrated that solutions with completely different topology could all lead to good simulations with realistic patterns and this suggests an overfitting problem. Although the nature of the overfitting is not well known, we believe it is a consequence of many reasons such as:

- missing data at some time points (late cad, early Tll) (see Fig. 5.1)
- incomplete description of the observation: the model does not explicitly incorporate all the data features such as late degradation of cad, hb.

The incomplete model leads to complex correlations between parameters that can be classified as:

- direct correlation describing compensation of an interaction on another one in an effort to maintain the simulated gene expression level at its normal rate (compared to the observation). This occurs when the gene change rate remains at the same level while changing different parameters. Typical example is the strong correlation of production rates and decay rates, or the compensation between positive input weights and negative input weights acting on the same domain.
- co-correlation describing the relation between two parameters acting on separate domain to control the boundary definition of a gene expressio. If gene "A" controls the left boundary domain of a gene "C" by repression and similary, if a gene "B" controls the right boundary domain of the same gene "C", there is a positive co-correlation between the two repressors: $A \dashv C$ and $C \dashv C$.
- indirect correlation resulting from variation in the gene concentration.

Typically weight compensation among regulatory parameters as well as non-regulatory parameters can occur if the time derivative, or gene change rate is remaining the same, while changing different parameters. Examples of these are the promoter rates R and the decay rates λ, which both scale the expression profile, but in different directions and in general show strong correlation patterns. Furthermore, the input weights on a single gene can also compensate each other. If a positive input on a gene becomes stronger, increasing negative weights or decreasing positive weights can adjust for the increased total input, such that the total input on that gene is not altered much. However, these correlation patterns are quite variable and difficult to predict and strongly depend on the precise spatial pattern.

Error with respect to the data Although some rigorous techniques (see [148]) have been used to extract quantitative information, one can not ignore the inaccuracy in measurement. This error might be caused by the type or amount

of antibody used for the fluorescent staining or the antigen purification process. An embryo is always stained simultaneously for three different genes. Each gene is stained using the same amount and type of antibody in each embryo. This could mean that the error measurement is different for two different genes. Also, usage of integrated data may not necessary reflect the expression of all the genes, but as mentioned by Surkova et al. [263], without the ability to simultaneously measure the expression of all modelled genes *in vivo*, it is not trivial to check how does the usage of average data will affect the fitting procedure. However, the average data seems to be a reasonable starting point. Comparison of individual gene expression least square error show that genes with missing data (*cad, tll*) have the highest error. However, These two genes are not well modelled since the gap genes do not up regulate *cad* and *tll* needs to have additional regulatory inputs. If we assume that the poor fit of these two genes is caused by missing data, incorporating more data would reduce their error. Comparison of individual gene expression score on complete synthetic data shown on Tab. 8.1 demonstrates that such an assertion does not necessary hold for *tll*. We have shown in Fig. 6.15 that fitting complete synthetic data does not reduce the over-fitting.

Exp/gen	*cad*	*hb*	*gt*	*Kr*	*kni*	*tll*
data	37.5057	15.9040	19.4748	17.4544	16.5046	19.1107
synthetic	2.8694	2.1713	1.3855	1.6247	2.1786	4.0380

Table 8.1: Average individual gene's score. For each gene j, the root mean score is given as: $\sqrt{\sum_i^t (g_i^j(t,\theta)_{model} - g_i^j(t)_{data})^2 / nbd_j}$ where i is the nucleus number and t the time. nbd is the number of data point available for the gene j. Contrarily to the total RMS given in Eq. 4.9, penalties are not included for simplicity. The second raw gives the average score (out of 101 circuits) and the third raw the average score out of 10 synthetic circuits (fitted against synthetic data). In the synthetic data, there is no missing data, consequently cad score is similar to the gap genes' scores, but not for tll.

8.2 Limitations of the reverse-engineering of gene regulatory network

8.2.1 Limitations of the parameter estimation

As dicussed in Chapter 3, setting a correct strategy to evolve the mutation strength is desirable. Although we have used an adaptive mutation strategy, we are still facing the stagnation problem that occurs after fast convergence in typical evolutionary algorithm. In Section 4.3.2, we discussed the convergence of the different settings employed in the ES. In Fig. 4.10, it is shown that (μ, λ)-ES fitness landscape always has a very fast convergence before being stabilized. Typically, the optimisation starts at a very high score and after very

few generations (approximatively $1/6$ of of the total generations), reaches a low score. From this point, very little improvement is achieved until completion of the generations. As suggested by Beyer et al. [20], tuning the mutation strategy is problem dependent and it would have been interesting and maybe an improvement if we had designed a specific mutation strategy for the current problem. Also, we have used as termination criteria the number of iterations that have to be reached by the ES. Since the convergence behaviour is typically observed in the initial phase of a (μ, λ)-ES, setting the termination criteria to the number of iterations should be avoid, but instead, a systematic "stagnation" detection could considerably reduce the computational time. In our case, the initial phase of (μ, λ)-ES took in average one to two hours out of the ten hours of the total computation on a serial 3.4-GHz "Intel Xeon" processor. Theoretically, no improvement can be obtained once the mutation strategy converges to a certain ϵ_σ (see Section 3.1.3), however, reaching this termination point would have required a very high computational time.

8.2.2 Limitations of the model

Disadvantages of the connectionist model The connectionist model is very useful for fitting real gene expression patterns and phenomenological modelling of networks. However, because the production terms are highly phenomenological, the predictive power of the model seems to be very limited. Some of the several disadvantages of the current description are the following:

- The threshold does not have a clear biological counterpart in the way genes are regulated in general. If a gene is constitutively active a more or less arbitrary value for h_a is chosen $h_a > -1$. If a gene is not active a rather negative values is chosen, typically between -3.5 and -2.5; however even for these values the gene is still fractionally active. The model behaves rather non-linear in the range where the gene is turned off, which seems to be rather artificial.

- Repression is only represented in the model by competition. There is no independent repression.

- The interactions between genes are described by linear terms represented by matrix products, which do not represent physical parameters in the real world other than connecting one gene to another. The values of the elements in the matrices do not have a clear biophysical meaning; hence they can take negative values and cannot be measured experimentally.

- The production term does only contain linear combinations describing the gene interactions, and does not contain concentration dependent activation of genes. Dimers may form or transcription factors may bind to other components in the system. This can partly be adjusted by using

tensor products. Also the same gene can only act as an activator or a repressor, and not both. If the DNA contains multiple binding sites with different affinity the same transcription factor may act as an activator at low concentrations but as a repressor at higher concentrations.

Failure to predict mutant patterns We have simulated four loss of function mutants by setting the promoter rate R_a and the initial gene concentration to zero [252]. The four loss of function mutants hb, Kr, gt and kni were simulated for each of the 101 gap gene circuits and the final gene expression pattern of hb, Kr, gt and kni were qualitatively compared with the reported expression patterns from literature [36,54,70,89,106,119,183,223,258,283]). In the experimental measurements of loss of function mutants it is observed that knocking out one of the genes in general only affects the adjacent domains in most cases, i.e. a local modular effect. Adjacent domains tend to expand into the region of the domain that is missing because of the mutation. We observe that all the 101 circuits failed in predicting the correct patterns for all the four loss of function mutant expression. In some circuits some domains (Kr, kni and gt) are correctly predicted, however we never observe that all these domains are correctly predicted in the same circuit. Furthermore, we observe that by knocking out one gene in the model, the model has a strong tendency to affect all genes in a rather dramatic non-local manner, yielding completely derailed patterns. This is in contrast with real measurements in the various mutants, which show effects by the mutations that are much more local of nature.

Possible reasons for weak robustness Although some circuits appear more robust than others, most of the circuits are extremely sensitive towards gene expression level fluctuations and also pattern formation is very sensitive with respect to perturbation of a large number of parameters. In most cases the parameters causing this are sometimes linked to Tll regulation, in some cases the parameters are linked to anterior gt and Kr regulation and finally the promoter threshold. There may be a multiple reasons for weak robustness. First, the incompleteness of the model may be the cause. In the real system it is known that the terminal pathway with Tll, Huckebein [285] and Torso also regulate gap genes, the latter two are missing in the current model. Hb regulation is not only zygotic but also has a maternal component at the anterior end of the embryo that is also regulated by nanos, nanos and a maternal description hb mRNA are missing in the model. Furthermore, only a part of the embryo is considered, which may lead to boundary effects at the anterior end. Secondly, the current reverse engineering approach using a simple RMS optimization may lead to sensitive circuits, not only some time points are missing in the data also features in the data that are not represented in the model may be over-fitted. Finally, the rather phenomenological approach of the connectionist model may also be a source of sensitivity because promoter threshold has a profound effect on robustness.

8.3 How to improve the reverse-engineering of the GRNs

In modelling processes from developmental biology especially the class of quantitative spatio-temporal, models of gene regulation is relevant. This type of models can potentially linked with three-dimensional biomechanical models of morphogenesis and provide new insights into developmental biology. Especially the quantitative spatio-temporal models of gene regulation are characterized by a large number of unknown parameters and an (infinite) class of potential solutions. So far, very few model-based analysis method have been proposed to validate or invalidate models, especially for nonlinear and spatially distributed models [69, 213, 221]. To our knowledge, the connectionist model has been the only model to provide spatio-temporal quantitative simulation of the gap gene based on parameter optimization techniques [78, 121, 206, 224]. Using this approach, some of the qualitative regulations in terms of gene-gene interactions could be captured and insights in the gap gene segmentation could be improved. The strength of this description resides in its non-requirement of any knowledge on the network architecture. Unfortunately, this also leads to an overfitting problem and conclusions concerning the mechanism governing the dynamic of the patterning may be spurious. Many improvements could be considered to have results that are more reliable. The following section briefly discusses the different possible improvements.

8.3.1 Possible improvements of the model

Hierarchical modular model One major improvement could be to take into account the cascade nature of the system. It might be necessary to model Bcd and Cad separately before addressing the gap genes [38, 52]. This would prevent cad to be repressed by the gap genes. Also Tll and maternal Hb could be modelled prior to the gap gene. Recently, in the PhD dissertation by Manu [167], a model using Bcd and Cad as external input was proposed. Ashyraliyev et al. [6] also suggested that using external Cad and Bcd reduced the parameters' indeterminability.

Decay term The decay term is represented by a simple linear decay. Many biological mechanisms exist that actively regulate the inactivation or removal of gene products. The connectionist model does not take into account these well-known mechanisms, e.g. binding of inhibitors, phosphorylation or dephosphorylation, ubiquitination leading to degradation of a protein or other mechanisms. Of course this may be adjusted in the above model by introducing a Φ-function in the decay term.

Hill-type model An obvious choice for describing interactions between genes is the use of Hill-type functions for the production term. The interactions be-

tween genes in these models are based on reversible binding reactions; hence they are biologically more realistic and the parameters represent apparent association constants. In the Hill type model we could explicitly separate the activation input a, and the repression input b, which replace the positive and negative inputs in the connection matrix. The activation input a is represented by molecules, like transcription factors that can bind to the DNA. Each transcription factor has a certain binding affinity K. Then the activation input a equals K_g, where g denotes the concentration of the transcription factor. If the transcription factor can form homodimers, and the homodimer binds to the DNA then a is represented by K_g^2. If the homodimer does not act as an activator then it will compete with the monomer binding, which leads to a concentration dependent activation. A similar argument is used for repressors, hence $b = K_g$. Repression could be described by an independent or competitive model.

8.3.2 Hierarchical parallel multi-objective island (μ, λ)-ES

In general the inverse problem that is only based on minimising the RMS leads to over-fitting, allowing for many solutions that all fit the dataset equally well but may not show correct behaviour or properties beyond the dataset. It is therefore difficult to know, which solution is most similar to the real system. The different post-optimisation analyses presented in this dissertation revealed some simple but efficient invalidation tests. In this thesis, we used the different tests on all the circuits. Ideally, we would like to funnel all circuits into a pipeline of tests where the circuits are filtered and those passing all the tests are the robust and stable solutions. It may however turn out that none of the solutions will pass all the tests. Therefore we propose not only to minimise the RMS but also to incorporate other objectives into the optimisation strategy. By introducing multiple objectives the solutions obtained should possess better properties and also have better defined parameters that show correlation patterns. The possible additional objectives for gene regulatory networks are:

- find networks that are robust towards parameter or input perturbation as found in Section. 7.1
- find networks that are robust towards fluctuations
- find networks with fewer connections (modular or a set of minimal networks)
- find networks with correct (stability) behaviour beyond the dataset as illustrated in Section. 5.2.

Sensitivity constraints Robustness towards parameter perturbations could simply be addressed by incorporating a sensitivity analysis within the optimisation procedure. Using levenberg-marquardt, Ashryalyev et al. [6] have investigated simultaneously while estimating the parameters, their identifiability. Rodriguez-Fernandez et al. [228] have presented a hybrid method that

can handle the sensitivity analysis while optimisation. Using such method, we suggest attributing to each circuit a sensitivity value that determines the overall sensitivity of its parameter. During the successive ES generations, circuits with the lowest sensitivity value should be prioritising.

Noise robustness constraint Another improvement would be to infer the network using stochastic models that include the fluctuations in the system. It is computationally expensive to numerically solve such systems and it will require efficient optimizations methods [273, 289]. Furthermore, it is known that patterning is insensitive to external fluctuations such as maternal gene expression. Consequently, an efficient model should also be able to simulate gap-gene expression given noisy external Bcd expression. An efficient way to distinguish all the solutions obtained from the parameter estimation would then be to change the BCD dosage during the optimisation. Good circuits should be able to reproduce the patterns as well as the shift without showing any major fluctuation of the gap gene expression.

Reducing network connectivity The long term dynamic revealed that in the presence of certain motifs, the circuits pattern converge to oscillatory attractors. This behaviour is not desirable as we expect the patterns to converge to a stationary point being the steady state. Therefore, we would ideally ignore or penalised the circuits having this behaviour. Also, we have shown that circuits with realistic topologies do not have these motifs and converge to the desired attractors. This could be obtained by either running the circuits beyond the data set to force asymptotic stability [101], or by adding an entropy function to prioritise circuits with the minimal connectivity.

Multi-objective The previous new constraints can be used within the optimisation framework using the stochastic ranking ES. In this dissertation, only one penalty function was used. Many multi-objective algorithms exist [46, 56] such as methods where multi-objective is transformed in one single objective [123, 128, 265], Pareto [17, 96] and non-Pareto dominance approaches. We have shown the sum of the weighted penalties. The implementation of the stochastic ranking ES is suitable for multi-penalties where the weighted sum of penalties $\mathbf{G}(g^+(\boldsymbol{\theta}))$ is used [184] as a multi-objective constraints. It is not trivial to determine the weight of each penalty. An alternative multi-objective would be to use a criterion-based approach where each penalty $g_j(\boldsymbol{\theta})$ is chosen with a certain probability and it is used as sorting criterion in a single ES. The strength of the ES is its intrinsic parallel nature. One alternative would be to use an island based ES where each the number of island is determined by the different objectives. Each island minimises the multi-objectives, but the distribution of the weight is different from weight to weight. Migration between islands

will spread individuals with a particular strong property in the other island. The main objective being the least square difference between the simulation and the data, it might be more practical to start with one population where the penalties have equal weights and switch to islands once the LSE is acceptable. This would guarantee that in all islands, the individuals have at least one good RMS. This sort of method would be a combination of Pareto and aggregate sum. Recent results obtained by Marcel Gokskun[1] tends to confirm this optimisation direction. The supplementary cost caused by the stochastic simulation should not be a limitation.

8.4 Conclusions

It will still take some time before scientists can give explicit explanations of the implication of genetic regulatory networks in the control of early development. The combination of biological experiments and mathematical models can already improve existent knowledge but there is still a need for more accurate data and more realistic models that cope different aspects of the mechanism. For instance, it would be interesting to infer GRN models for pattern formation in organisms where moving cells and deformable shape are essential features. Three-dimensional models will then be necessary and the number of parameters will increase substantially. The major contribution of this dissertation is the fast and efficient parameter estimation methods provided, in comparison to previous available methods, for the current available model. We have shown the limitations of the current gap gene model and suggested possible improvements. The different analyses conducted in this thesis suggest that it is essential to carefully analyse inferred GRNs in order to gain confidence. It also shows simple methods to perceive the distinguishing features of circuits that are more robust, although they are obtained from the same model description and the same quantitative data. To avoid multiple circuits topology and obtain more robust circuits, it is essential to incorporate more constraint in the optimisation procedure. Multi objective optimisation should be considered where, in addition to minimising the distance between simulated data and observation, more criteria should be included such as the sensitivity of the model to parameter variation or noise, the deterministic stability to ensure that the pattern converges to its biological attractor and ultimately the network connectivity.

[1]Personal communications.