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

Fomekong Nanfack, Y.

Publication date
2010

7

Gene circuit perturbation analysis[1]

Reverse engineering of GRN capable of simulating spatio-temporal gene expression mainly consists in revealing the GRN structure that leads to the observed pattern. Optimization is therefore important [5, 184] and it is used to infer from a gene network whether it is a transcriptional or protein interaction. A system identification approach is used to select a model structure and by means of parameter estimation, the network topology is estimated from experimental data. The optimization problem consists in minimizing the difference between simulated expression profiles and available data in order to estimate the best circuit that predicts with very good accuracy the spatio-temporal gene expression. Based on the resulting parameter-set, the network diagram is extracted and one tries to establish causal-relation of the dynamic mechanism that governs the pattern formation, using further analysis. The prediction is clearly linked to the quality and amount of the data; even with sufficient data, it is not guaranteed that the inferred network corresponds to the appropriate gene network that leads to the observed pattern [103, 152]. From a theoretical point of view, this question is a matter of the structural identifiability of the model [124]. Given a data set, is it possible to uniquely infer the network? Although there exist some theoretical methods [103] to investigate the identifiably before inferring the network, when confronted with a complex mechanism characterized by a multi-dimensional parameters space, the feasibility is still analytically complex. In many situations, one is confronted with an overfitting problem that leads to parameter-set with very different quantitative values and even worse, yielding different network topologies having similar good realistic patterns. To draw conclusions, one has to differentiate between circuits

[1] This chapter is partially based on the paper:
Yves Fomekong-Nanfack and Marten Postma and Jaap A. Kaandorp, "Inferring *Drosophila* gap gene regulatory network: a parameter sensitivity and perturbation analysis", BMC Syst Biol, 3:94, 2009. [79]

that are more likely to be biologically realistic.

It was shown that for certain experimental data that it is not possible to confirm whether the inferred model is really valid [210]. As long as the model is not contradicted by the given data, it is necessary to extend the validation test further. Especially when the model does not only predict a unique network, parameter-set discrimination can be addressed using diverse models validation approaches. The quality of the circuits can for example be quantified by measuring the parameter reliability and sensitivity, the uniqueness of the predicted network, the model robustness and the predictability of the model. The two previous chapters focussed on the other aspects that can be used to quantify the quality of the circuits: methods for exploring the solution space, parameter correlation and asymptotical stability analysis. In this chapter, we focus on the model robustness against perturbation.

By analyzing the robustness of the inferred network, we can test the quality of the circuits. The term robustness has various meanings, but in the current study, robustness is addressed as the ability of a system to maintain its mechanical and dynamical behaviour under perturbation. From a biological view, robustness is related to stability, homeostasis, canalization, redundancy, and plasticity [37, 125, 142, 282], and can also be applied to dynamical process in development [254]. In simulations of dynamical developmental processes such as pattern formation, one should also expect robust behaviour within a certain extent in the model [255]. In the current paper, model robustness is addressed in two different perspectives. First, we investigate the quantitative robustness of the model towards internal fluctuations in expression level. It is known that presence of noise in gene regulation can lead to phenotype variation [23, 269]. There are some studies on the robustness of GRNs under the influence of molecular fluctuation [50, 126] and show the importance of noise and stochastic events [217]. In most cases, deterministic models are used to infer GRNs from noisy data. It has been shown in several studies that the robustness to noise mainly depends on the network structure instead of the parameter setting [14, 92, 131]. Therefore, one way to discriminate between circuits having different gene network but exhibiting the same pattern is to analyze their behaviour under noisy conditions. Model-solutions showing more robustness or stability can be considered for further analysis. Second through simple parameter perturbation, we investigate how model-circuits behave distinctly. This analysis allows us to identify the parameters that have the most significant influence on the model and to distinguish circuits that are less sensitive to overall perturbation. The combination of these two analyses allows one to discriminate between sets of circuits that are more robust, although they are obtained from the same model description and quantitative data.

Using a local sensitivity analysis, determinability of some parameters were studied in [6], where it was suggested that one could confirm on the nature of some biological interactions for parameters that were shown to be identifi-

able. Using a large set of parameters as base value, it is possible to examine the robustness of the model. Zak et al. [50] showed how input perturbations and stochastic gene expression influence the identifiability of a specific regulatory network given gene expression profiles and prior structural knowledge. Using as starting point the parameters-set obtained in [78,121] as base value, we discuss how does input perturbation and stochastic simulation allow a model-based robustness analysis of a model that leads to multiple circuits.

7.1 Robustness towards fluctuations

Gene transcription, degradation and diffusion inherently are stochastic processes, which lead to fluctuations in gene expression levels. If expression levels are high, these fluctuations generally do not affect the time evolution of the system. In these cases the deterministic model and stochastic model will yield a similar result. However if the system is non-linear and the fluctuations occur at an early time in development, where levels are still low, fluctuations may lead to different patterns. From a biological point of view this may not be favourable because gene pattern defects can lead to aberrations in development, hence development should be robust towards modest transient disturbances [10, 98]. Here we investigate if the circuits obtained from the optimization, which are based on a deterministic model, are robust towards fluctuations. Analysis of fluctuations in the bicoid gradient [294] suggests that the expression level of bicoid is about five fold higher than the fluorescence unit. In our analysis we assume that the expression level of all genes is five fold higher than the fluorescence unit. We have used the Gillespie algorithm [94] to introduce fluctuations in the gene expression levels.

7.1.1 Stochastic model

We implemented a nonlinear stochastic deterministic differential equation model of the gap gene where the dynamic is driven by the Gillespie stochastic simulation algorithm [94]. The algorithm is intended for explicitly simulating individual reactions, and sampling the reaction probability density function for determining the time step until the next reaction. This algorithm essentially generates one possible, statistically correct, solution of a stochastic differential equation, the master equation. In the current case, the master equation is based on the differential equation describing the concentration change of a gene expression in a particular nucleus. Each equation describes three different reactions : protein production, protein decay and diffusion. Only one reaction can occur at a specific time. The Gillespie algorithm simulates the system by choosing first in a probabilistic manner which reaction occurs, and then estimates when does the next reaction will be realized. The all process is described as follows:

Algorithm 3 Gillespie

1: Initialize the protein quantities, and set the maximum running time
2: Determine the total reaction rate.
3: Determine the time for the next reaction to occur.
4: For each reaction determine the probability that it takes place.
5: Choose the reaction to occur during this time step. *production, decay, or diffusion*
6: Execute the reaction determined in the previous step.
7: Update all reaction rates.
8: If the maximum running time has not been exceeded, go to step 2.

At the beginning, only a finite number of proteins are initialized. The total reaction rate expressed at that time is determined based on the sum of all production, decay and diffusion terms in all reactions. The time step until the next reaction will occur is randomly determined as $t + \tau$ where τ is drawn from an exponential distribution. Consequently for each reaction in each differential equation the probability that this reaction has taken place during that time step is determined. This probability is defined by the corresponding reaction rate in the differential equation divided by the total reaction rate. To choose the single reaction that has taken place during that time step, a number from the uniform distribution on [0; 1) is randomly picked and the reaction is determined. To determine the corresponding reaction, the interval [0; 1) is partitioned based on the size of the previous probability. Each reaction is associated to a certain gene for a certain cell. The selected reaction is update by updating the number of protein for that gene in that cell. If the reaction is a protein production, one protein is added and if it is a decay, one protein is subtracted. For diffusion, a protein is subtracted in one nucleus and added in a, randomly chosen, neighboring nucleus.

Since the algorithm is computationally, to speed the process, we have used a method [93] that only calculates and updates reactions that have changed. The probabilistic nature of the algorithm imposed us to run 100 simulations for each of the 101 gap gene circuits. All simulations used the same initial condition and the number of molecules is 5 times the deterministic case (based on [294]). For simplicity, analyses are only made on the final pattern (gastrulation time) by comparing the deterministic simulation with the stochastic one.

7.1.2 Stochastic simulations

For each circuit we performed 100 stochastic simulations and analyzed the quality of the final expression pattern at $T = 68.1$ min. In each run, by comparing the deterministic model with the stochastic model using a RMS cut-off criterion, we counted how many of domains *cad*, *hb*-anterior (hba), *hb*-posterior

(hbp), *Kr*, *gt*-anterior (gta), *gt*-posterior (gtp), *kni* and *tll* were considered to be correct. For each circuit we also calculated an overall score by counting the runs where all expression domains formed correctly (see Tab. 7.1). On average the domains *cad* (99%) and *hb*-anterior (93%) show the best scores. The domains of *kni* (86%), *Kr* (80%) and *gt*-posterior (81%) have an intermediate score. The domains *gt*-anterior (66%), *hb*-posterior (70%) and particularly *tll* (47%) have a low score. Most circuits have a very low overall score; only the top 25% of the circuits have an overall score higher than 35%. Most of the circuits have a score lower than 20%.

nr	cad	hb-a	hb-p	Kr	gt-a	gt-p	kni	tll	all
7	100	99	94	93	88	92	96	94	73
12	100	100	98	84	67	100	100	100	66
80	100	100	98	94	72	99	100	83	63
48	100	95	100	100	98	96	100	67	62
26	100	100	95	87	77	97	100	80	62
36	100	100	99	76	61	100	99	95	61
87	100	100	97	97	60	100	100	100	59
35	100	98	100	95	97	82	92	75	56
8	100	100	92	82	71	96	98	92	55
47	100	100	93	78	70	80	99	98	54
22	100	100	95	80	56	100	98	99	54
4	100	99	96	79	65	88	97	96	54
14	100	100	89	84	68	100	100	80	52
20	100	100	100	80	52	100	100	95	51
42	99	99	89	84	60	99	97	79	46
46	100	68	100	100	97	100	100	65	45
72	100	99	92	68	61	77	87	100	41
24	100	97	83	96	92	64	93	71	40
81	100	99	84	81	72	85	95	75	39
71	100	100	62	81	67	100	98	97	39
25	100	100	78	84	62	83	96	86	39
38	100	94	74	87	78	59	95	100	38
40	100	100	96	81	47	97	90	81	37
75	100	100	63	78	72	100	98	58	36
11	100	100	100	44	35	100	100	100	34
13	100	97	70	91	80	89	93	59	33
30	100	99	76	99	94	75	97	56	30
6	100	95	100	75	89	42	50	82	30
61	100	71	100	100	99	98	99	39	29
37	99	99	74	82	62	90	100	51	29
74	100	100	70	78	64	97	99	46	27
10	100	100	63	96	67	98	100	43	27
78	100	100	82	95	66	99	100	46	26

9	100	99	82	64	47	89	95	55	25
64	100	100	75	89	62	70	88	39	24
43	100	100	50	99	78	98	62	65	24
69	100	100	55	81	63	100	97	45	23
55	94	80	99	100	79	98	99	40	23
85	76	82	100	99	100	86	92	30	21
67	100	83	88	92	59	99	100	46	21
91	93	84	100	100	100	92	95	27	20
39	100	98	50	93	86	52	93	61	20
21	100	99	46	85	66	92	99	34	19
49	100	97	58	97	59	97	100	36	18
98	100	100	100	54	41	91	98	41	17
90	100	100	73	46	34	100	100	52	17
84	100	100	58	96	75	83	97	52	17
82	100	100	51	82	56	100	100	36	17
17	100	89	96	100	78	99	99	27	17
62	100	100	41	92	69	69	97	50	16
77	100	93	54	73	55	79	91	38	14
60	97	87	100	96	78	77	89	30	14
31	100	99	78	69	61	82	64	22	14
27	97	100	57	91	56	89	54	71	14
95	100	98	98	46	39	88	92	27	13
94	98	92	72	81	48	65	91	48	13
92	100	98	46	90	69	74	94	31	13
73	100	86	77	83	92	66	73	21	13
68	100	97	33	93	76	83	94	33	13
29	100	98	58	89	63	87	99	25	13
65	100	95	28	96	83	65	98	51	12
2	100	100	37	98	98	94	70	20	12
101	100	100	85	54	68	100	100	76	11
59	100	70	99	77	96	71	76	27	11
50	100	99	69	84	67	58	85	38	10
5	100	90	82	96	64	100	100	20	10
83	100	97	78	59	61	34	62	64	9
76	100	98	100	70	100	46	72	21	9
54	100	83	86	76	83	53	78	33	9
18	99	88	68	99	75	89	97	15	9
1	100	94	72	68	50	91	70	32	9
99	100	82	66	33	42	49	49	42	8
28	99	99	46	63	57	69	79	34	8
63	99	73	42	99	65	87	95	13	7
45	100	100	42	93	58	90	79	33	7
3	69	97	59	63	72	66	85	13	7
44	100	98	26	85	71	84	70	46	6
93	100	99	29	70	43	68	80	17	5
57	100	99	40	76	68	73	86	26	5

34	84	80	100	100	72	94	96	12	5
32	100	99	33	96	58	73	90	10	5
23	99	75	38	83	62	75	94	12	5
15	99	97	25	83	54	81	88	20	5
96	100	85	60	45	52	42	38	37	4
58	98	83	97	80	71	56	76	14	4
52	100	94	98	32	28	69	67	26	4
51	100	99	74	58	37	97	94	23	4
33	99	80	26	72	73	33	54	18	4
16	98	91	33	56	54	74	92	24	4
88	100	95	24	91	56	94	65	57	3
70	100	98	34	79	61	78	87	7	3
66	99	97	15	86	63	66	80	32	3
100	100	99	57	67	84	52	22	13	2
89	100	83	49	76	61	72	71	24	2
79	78	100	30	45	33	85	86	6	2
19	100	100	31	90	70	90	44	38	2
53	100	99	78	32	24	85	52	7	1
97	100	86	25	37	19	50	13	12	0
86	97	32	99	72	60	59	76	4	0
56	100	59	16	71	60	46	67	14	0
41	100	78	55	59	80	67	48	10	0
avg	**99**	**93**	**70**	**80**	**66**	**81**	**86**	**47**	**22**

Table 7.1: For each circuits,100 stochastic simulations were run. The score is calculated based on the overall pattern, but also the individual gene score. Genes expressed in two domains (such as *hb*) have two different scores.

If all domains except one develop correctly the overall score is determined by that domain score, however if two or more independent domains have scores lower than 100%, the overall score will be lower than the individual domain scores. For example, if all domains would have a score of 99% and are independent the overall score would be $0.99^8 * 100\% = 92\%$. However we also observe interactions between domains, which leads to a higher overall score than would be expected if they behave independently. In these cases, if one domain yields a low score for a particular run, it is likely that another domain also yields a lower score for the same run. This is typically observed between adjacent domains. Notable examples are *Kr* and *gt*-anterior, *hb*-anterior and *gt*-anterior, *kni* and *gt*-posterior, and *gt*-posterior *hb*-posterior and *tll*.

In Fig. 7.1 the patterns of eight different circuits are shown. In these graphs the dashed lines represent the final profiles obtained from the deterministic model and the solid lines the average profiles obtained from the stochastic simulations calculated from the individual stochastic runs, which are shown in grey.

In Fig. 7.2 the time-evolution of the expression level of different gene combinations at a particular nucleus position are shown for the same circuits as in Fig. 7.1. For comparison, the deterministic model is plotted using a solid red line and all stochastic runs (100) using light coloured lines. At the end of these lines a dot is shown, which represents the final concentration at $T = 68.1min$. The end points were used to extract two clusters (with k-means clustering), which are represented by the blue and green colour.

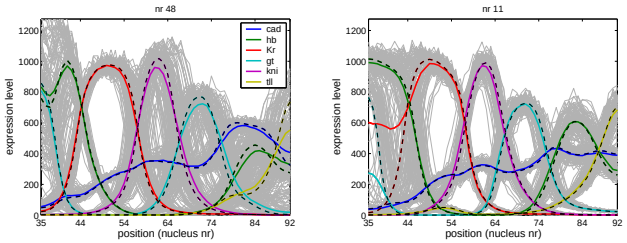
In Fig. 7.1A,B the final pattern formed during the stochastic simulations are shown for circuit 48 and 11. Circuit 48 has an overall pattern score of 62% and circuit 11 has an overall score of 34%. In circuit nr 11 all domains except *Kr* (44%) and *gt*-anterior (35%) score 100%. *Kr* and *gt*-anterior show a strong interaction, if *gt*-anterior disappears then *Kr* expands into the *gt*-anterior domain. In circuit 48 this interaction between *Kr* and *gt*-anterior is not present. In this circuit especially *tll* is not well defined. Fig. 7.2A-B shows the trajectory of *Kr* and *gt* at nucleus 35, for circuit 48 in almost all runs *gt* and *Kr* develop correctly, however in circuit 11 there are two possible outcomes of the stochastic run. Here *gt* and *Kr* can develop correctly, however *gt* may also disappear and completely be repressed by *Kr*. In this circuit there are two pathways for the system to evolve, which may lead to two stable points or a single stable point but two pathways. Jaeger et al. [120] suggested that non-overlapping gap gene are mutually exclusive and have mutual repression. This result was confirmed for *gt* and *Kr* by Ashyraliyev et al. [6]. Circuit 48 and 11 both show this strong mutual repression but we believe that the bad stochastic score of *gt*-anterior and *Kr* might be caused by the weak repression of *gt* by Tll.

In Fig. 7.1C,D the final pattern formed during the stochastic simulations are shown for circuit 2 and 41. Both circuits have a very low overall score of 12% and 0% respectively. In circuit 2, a very low score for *hb*-posterior, *gt*-posterior and *tll* mainly causes this and for circuit 41 all domains except *cad* are not well defined. In Fig. 7.2C,D the trajectories at nucleus 35 of *hb* and *gt* are shown. For circuit 2 the pathway is well defined, in most runs both *gt* and *hb* evolve similar to the deterministic model. However in circuit 41 the pathways are not well defined at all and show variability. In circuit 2, we see that Hb represses *gt* while Gt weakly activates *hb*, and inversely for circuit 41. It is suggested that overlapping gap genes do not activate each other [6, 120]. The poor robustness of these 2 circuits is therefore a consequence of the wrong connectivity of these 2 interactions.

In Fig. 7.1E,F the final pattern formed during the stochastic simulations are shown for circuit 5 and 24. Circuit 5 has a very low score of 10% and circuit 24 a score of 40%. Although circuit 5 has a low overall score both the *kni* and *gt*-posterior domain score 100%, circuit 24 has a lower score for these domains. In Fig. 7.2E,F the trajectories are shown for *kni* and *gt* at nucleus 69. In circuit 5 the final points of the stochastic runs are very close to the final point of the deterministic run. Although the final points are well defined, the pathway, which is

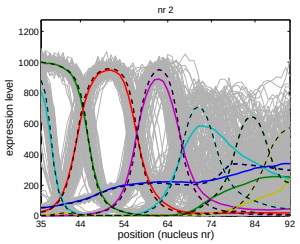
a shift of both domains, to these points is quite variable. In circuit 24 the same phenomenon is observed, only here Kni in some cases completely suppresses *gt*. Both circuits show *gt* activation by Kni (weak), which seems to be a wrong interaction [6,120]. Circuit 5 shows repression of *hb* by Gt and weak activation of *gt* by Hb while circuit 24 shows the inverse mechanism. Based on the identifiability analysis obtained by Ashyraliyev et al. [6], it is suggested that Gt does not repress *hb* and Hb does not regulate *gt*. However, the determinability of these parameters has a very poor confidence and qualitative conclusion on these interactions is still ambiguous.

In Fig. 7.1G,H the final pattern formed during the stochastic simulations are shown for circuit 20 and 79. Circuit 20 has an overall score of 50% and circuit 79 has a very low score of 2%. In circuit 20 all domains except *Kr* and *gt*-anterior are well defined. In circuit 79 all domains except *cad* are not well defined. In Fig. 7.2G,H the trajectory of *hb* and *tll* at nucleus 92 are shown. In circuit 20 most final points are very close to the deterministic model, however in circuit 79 almost all points are far away from the deterministic model. In this circuit the *tll* domain is repressed by Hb and completely disappears and the *hb* domain continuous to grow. In some runs in circuit 20 (blue trajectories) we see the same tendency of continuous *tll* repression combined wit *hb* increase. Circuit 79 shows very inconsistent regulatory mechanism with respect to available literature [36,54,70,268]. *gt* and *hb* show mutual activation, and Kni activates *kr*. These interactions seem to be the wrong regulatory mechanism, leading to stochastic instability and a very low pattern score.

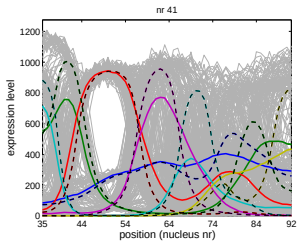


(a) solution number 48

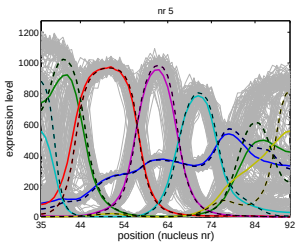
(b) solution number 11



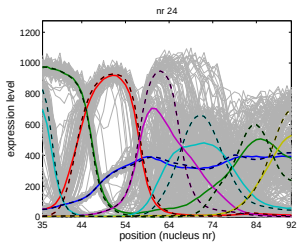
(c) solution number 02



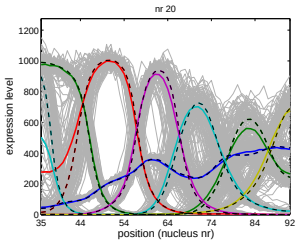
(d) solution number 41



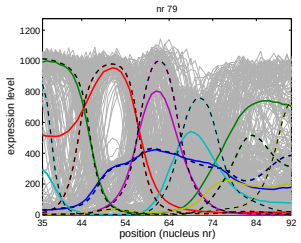
(e) solution number 05



(f) solution number 24

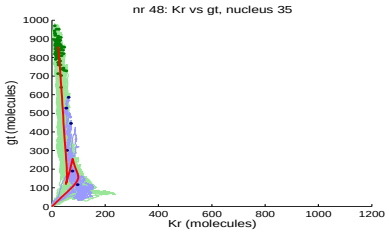


(g) solution number 20

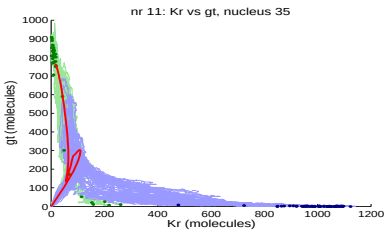


(h) solution number 79

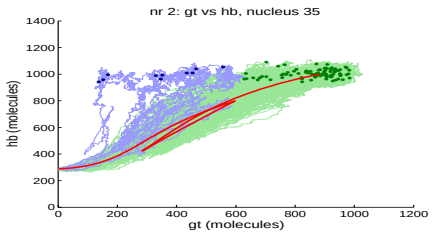
Figure 7.1: The patterns of eight different circuits. In these graphs the dashed lines represent the final profiles obtained from the deterministic model and the solid lines the average profiles obtained from the stochastic simulations calculated from the 100 individual stochastic runs, which are shown in grey.



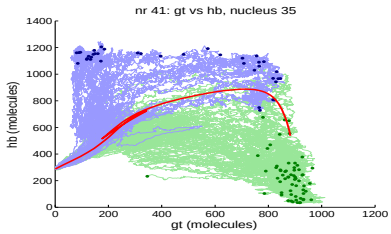
(a) Kr vs. Gt in nucleus 35 of solution 48



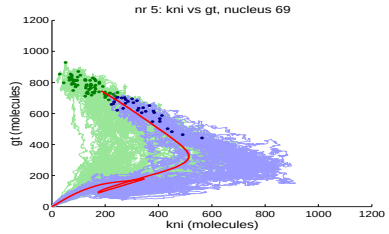
(b) Kr vs. Gt in nucleus 35 of solution 11



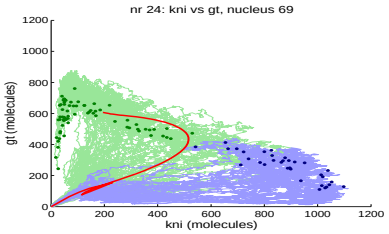
(c) Gt vs. Hb in nucleus 35 of solution 02



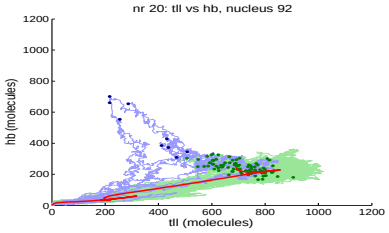
(d) Gt vs. Hb in nucleus 35 of solution 41



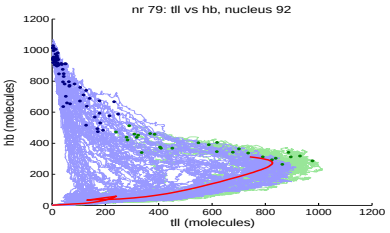
(e) Kni vs. Gt in nucleus 69 of solution 5



(f) Kni vs. Gt in nucleus 69 of solution 24



(g) Tll vs. Hb in nucleus 92 of solution 20



(h) Tll vs. Hb in nucleus 92 of solution 79

Figure 7.2: The time-evolution of the expression level of different gene combinations at a particular nuclear position are shown for the same circuits as in Fig. 7.1. For comparison, the deterministic model is plotted using a solid red line and all stochastic runs (100) with light coloured lines. At the end of these lines a dot is plotted, which represents the final concentrations at T = 68.1 min. These end points were used to extract two clusters (with k-means clustering), which are represented by blue and green colours. The left panels show circuits with final points and pathways very close to deterministic model and the right panels show less well defined circuits.

7.1.3 Correlation between robustness and parameters

In Fig. 7.3 the correlation between the eight different expression domains scores and all parameters are shown. The correlation pattern reveals that circuits with higher promoter rates, diffusion coefficients and higher degradation rates are more robust towards fluctuations. This suggests that higher rates tend to increase robustness of the circuits. Except for W_{cad}^{bcd} all maternal inputs correlate negatively with pattern robustness. This suggests that strong maternal inputs tend to reduce the robustness of the circuit. All negative inputs on *cad* show a positive correlation, which suggests that strong Cad input weights tend to reduce robustness. Furthermore, the inputs of Cad on all genes except *tll* show a strong negative correlation, hence strong weights reduce robustness. In Fig. 7.4 the correlations is shown separately for circuits with promoter threshold $H = -2.5$ and $H = -3.5$.

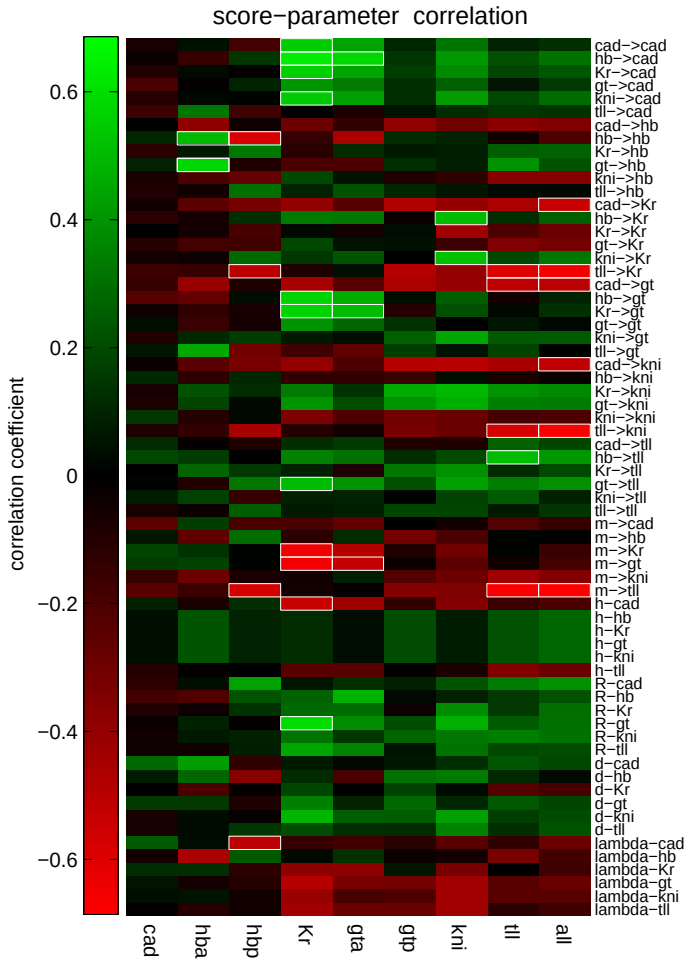


Figure 7.3: Correlation between the score of the eight different expression domains obtained from stochastic simulations and all parameters. Bright green represents a strong positive correlation and bright red a strong negative correlation. Squares in white borders are the most significant correlations.

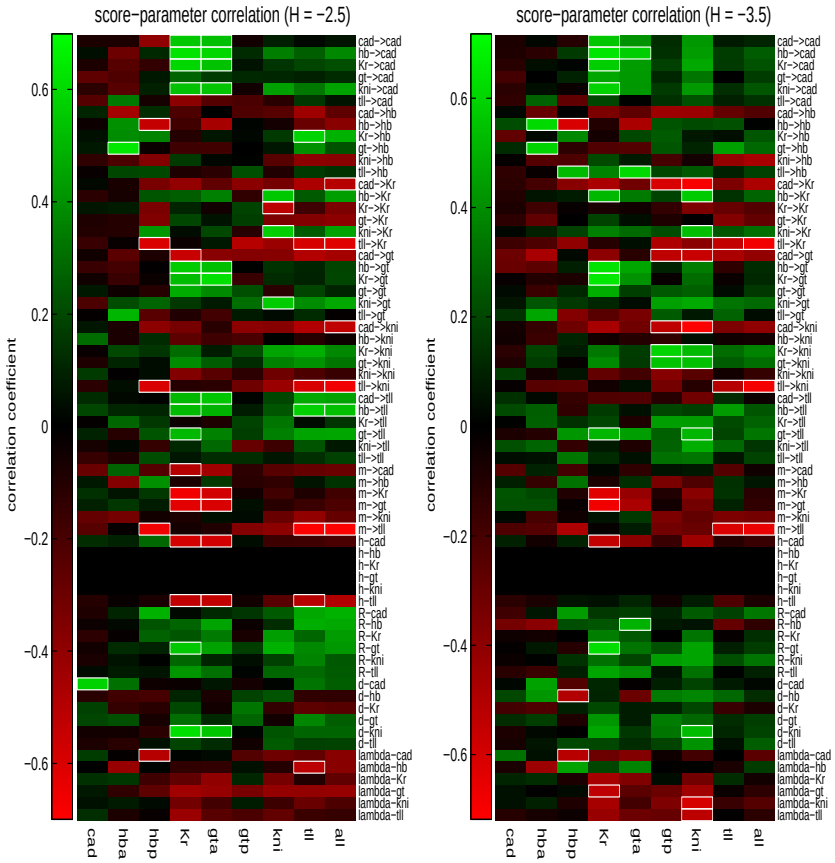
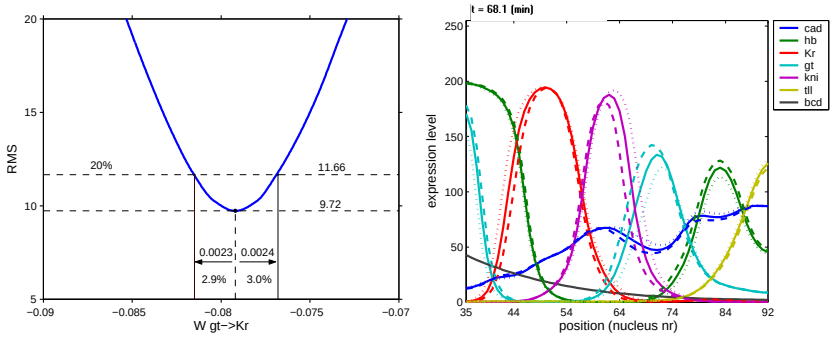


Figure 7.4: Correlation between the score of the eight different expression domains scores and all parameters obtained for solution with respectively $H = -2.5$ on the left panel A, $H = -3.5$ on the right panel B. Light green expresses a strong positive correlation and light red a strong negative correlation. Squares in white borders are the most significant correlations.

7.2 Robustness towards parameter perturbation

If a single parameter is slightly decreased or increased, the RMS of the simulated pattern will increase because the quality of the fit to the gene expression data reduces. The amount, or range, by which a parameter can be changed before the fit becomes significantly bad, is a measure for the sensitivity of the parameter for a particular solution. For each parameter within each solution we computed the lower and upper value of a parameter where the RMS increases by 20%, which corresponds to a situation where the gene pattern becomes significantly bad. Hence, the lower value $\theta_{n,i} - \Delta L_{n,i}$ is computed from $f(\theta_{n,i} - \Delta L_{n,i}) = 1.2 * RMS_n$ and the upper value $\theta_{n,i} + \Delta U_{n,i}$ is computed from $f(\theta_{n,i} + \Delta U_{n,i}) = 1.2 * RMS_n$, where RMS is the original RMS value of the fit obtained from the optimization and f denotes the cost function, n denotes the solution number and i the parameter index. We define the sensitivity interval as $[\theta_{n,i} - \Delta L_{n,i}, \quad \theta_{n,i} + \Delta U_{n,i}]$ and the sensitivity as $S_{n,i} = -\log(\frac{1}{2}\Delta L_{n,i} + \frac{1}{2}\Delta U_{n,i})$. The average sensitivity of a solution and a parameter are defined as $S_n = \sum_i^{np} S_{n,i}/N_{np}$ where np is the number of parameters and respectively $S_{\theta_i} = \sum_n^{ns} S_{n,i}/N_{ns}$ where ns is the number of circuit. The lower relative sensitivity is defined as $\theta_{n,i}/\Delta L_{,i} * 100\%$ and the upper relative sensitivity is defined as $\theta_{n,i}/\Delta U_{,i} * 100\%$. From a biological point of view pattern formation should be robust towards small perturbations in the parameters [275].



(a) (b)

Figure 7.5: Parameter perturbation. (A) perturbation of the parameter W_{Kr}^{gt} and in (B) simulated gene expression after parameter perturbation.

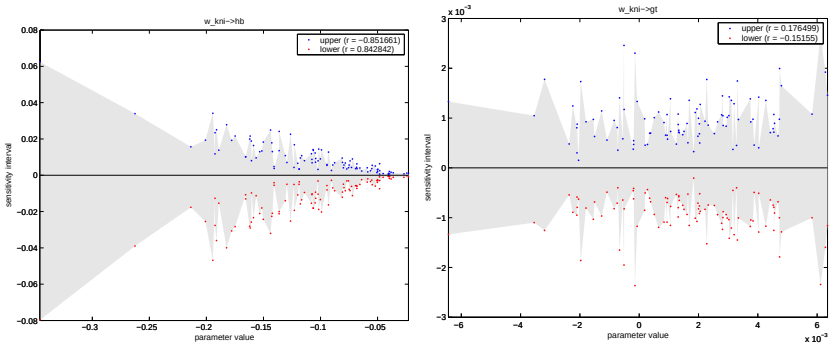
In Fig. 7.2, sensitivity interval versus parameter value for D_{gt}, W_{hb}^{kni} and W_{gt}^{kni} are plotted. For most circuits the lower value for the diffusion coefficient reaches the value zero, hence the lower bound of SI is equal to the parameter value, meaning that without diffusion the pattern still forms correctly. This phenomenon is also observed for all other diffusion coefficients. Furthermore, the

upper bound of SI does not scale with the parameter value, hence circuits with very similar diffusion coefficients can have very different SIs. Non-scaling behaviour is observed for most parameters, another example is W_{gt}^{kni} shown in Fig. 7.2B. For this parameter, both the upper and lower bound do not scale with parameter value, also here we observe that circuits with very similar parameter values can have very different SIs; this was observed for most parameters. For some parameters the SI does however scale with parameter value. For example all decay time constants show scaling, but also some weights. Fig. 7.2A shows the plot of W_{kni}^{hb}, here both the upper and lower bound of the SI decreases with smaller absolute parameter values. Hence the circuits become more sensitive towards perturbations when the parameter is smaller in magnitude.
In Fig. 7.7 the correlation matrix, calculated on the basis of the SIs is shown. The correlation pattern shows blocks on the diagonal, where circuit parameters that regulate one particular gene tend to cluster together. For example, in the case of Cad, if in a particular circuit a parameter that regulates *cad* has a large SI, it is likely that the other parameters that regulate *cad* also have a larger SI. This suggests that sensitivity appears to behave in a modular fashion, where the genes represent modules. One exception is Kr and Gt, which both are found in the same cluster, suggesting these two genes interact with each other. This correlation is caused by their strong mutual repression [6], forcing them to act in similar way to perturbation.

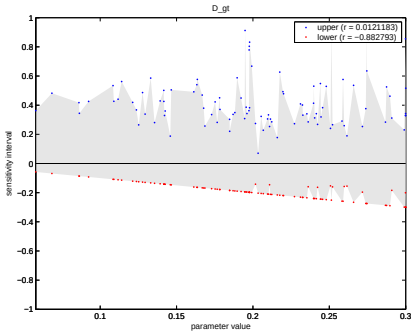
7.2.1 Circuit sensitivity

To investigate which parameters and which circuits are more sensitive than others we have calculated the average sensitivity on a logarithmic scale for each parameter and each circuit (see methods). Then the SI of each parameter value in each circuit was plotted using an intensity plot, where the colour corresponds to: $-\log SI$. The parameters and circuits were ordered in Fig. 7.8 according to their average sensitivity, showing the parameters with smallest SI and most sensitive circuit at the top right corner. From Fig. 7.8 it can be seen that the difference between the smallest and largest SIs are 4 log units. The circuits are less sensitive with respect to diffusion coefficients, promoter rates, decay time constants and promoter threshold. Circuits are most sensitive with respect to W_x^{cad} weights followed by the auto-regulation weights W_x^x.

Because the different parameter types (diffusion, decay, promoter rate, thresholds and weights) are not on the same scale we also calculated the average relative sensitivities of the circuit parameters (see Tab. 7.2. This approach can however be problematic, if the parameter values are close to zero, possibly yielding a very high relative sensitivity. Therefore these outliers were removed from the analysis. Using this measure, the least sensitive parameters are the diffusion coefficients with a relative sensitivity of about 100-200%. The thresholds are now more comparable with the weights and have a relative sensitivity of about 1%. Furthermore the W_x^{cad} weights are still amongst the most sensi-



(a) (b)



(c)

Figure 7.6: Distribution of sensitivity interval versus parameter value. (A) Plot of hb_{kni} sensitivity interval. Here both the upper and lower sensitivity interval scale with absolute parameter values. Hence the parameter becomes more sensitive when it is smaller in magnitude. (B) W^{gt}_{kni} sensitivity shows non-scaling behaviour, which is also observed for most other parameters. Furthermore, circuits with very similar parameter values can have very different sensitivities; this was also observed for most parameters. (C) Diffusion sensitivity illustrated by D_{gt}. For most circuits the lower value for the diffusion coefficient reaches zero, hence the lower sensitivity interval is equal to the parameter value, meaning that without diffusion the pattern still forms correctly.

tive parameters with a relative sensitivity of about 0.5-1%. From Fig. 7.9 we see that the average sensitivity of the circuits varies from one to another. Certain parameters in the least sensitive circuits, notably weights related to *tll* and *cad* tend to have lower sensitivities. By comparing the least sensitive circuits with the more sensitive circuits using a t-test several parameters where found that are significantly different in these groups. These are in fact the parameters that show scaling between sensitivity and parameter value W^{hb}_{Tll}, bcd_{Tll}, W^{gt}_{Tll}, W^{Tll}_{kni}, W^{Tll}_{Kr}, W^{cad}_{Kr}, W^{cad}_{gt} and W^{cad}_{kni}.

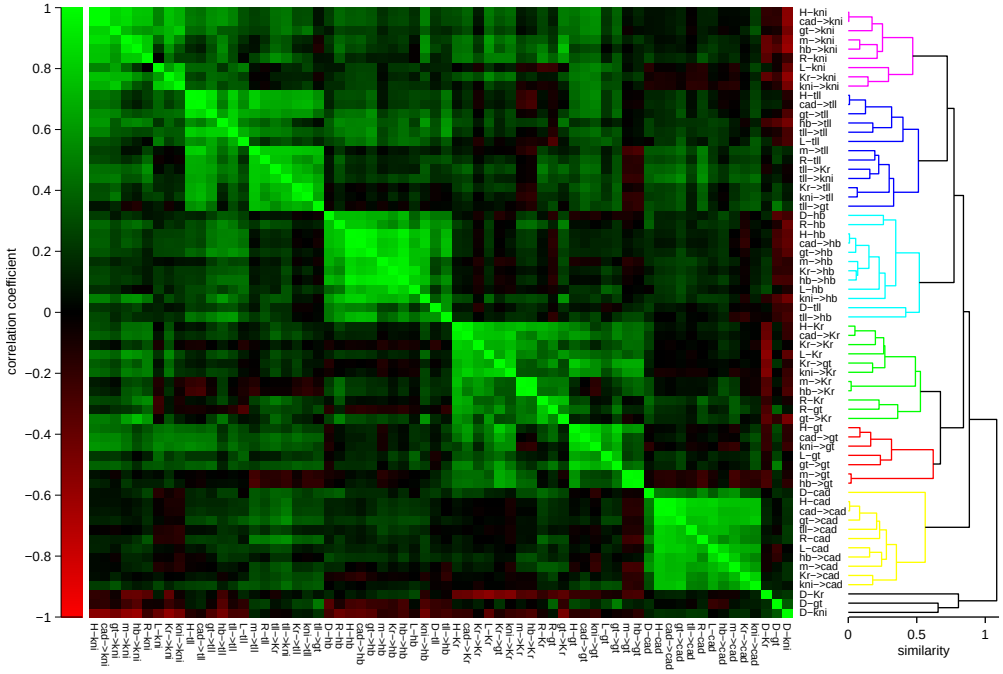


Figure 7.7: The correlation pattern calculated on the basis of sensitivity intervals. The correlation matrix shows blocks on the diagonal, where circuit parameters regulating one particular gene tend to cluster together.

The standard deviation of the circuit parameters can be quite large compared to the average parameter value (as shown in Fig. 6.3). And furthermore, average parameter value and standard deviation strongly correlate (r = 0.81). In Fig. 7.9 the average $-\log SIs$ versus standard deviation is plotted. Amongst similar parameter types (weights and non-weights) no correlation is observed between SI and standard deviation. This suggests that if a parameter is very variable across the circuits, i.e. it is not well-defined [261], it is not necessary less sensitive. A notable example is bcd_{cad} and W^{cad}_{Tll}, which both have a similar standard deviation but differ in sensitivity by 2 log units. Furthermore, we also observe that the standard deviation on average is much higher than the sensitivity interval.

7.2.2 Model sensitivity vs. pattern robustness

In Fig. 7.10 we have plotted the average sensitivity versus the overall score of the circuit. In this figure the circuits with promoter threshold $H_{hb,Kr,gt,kni}$ =

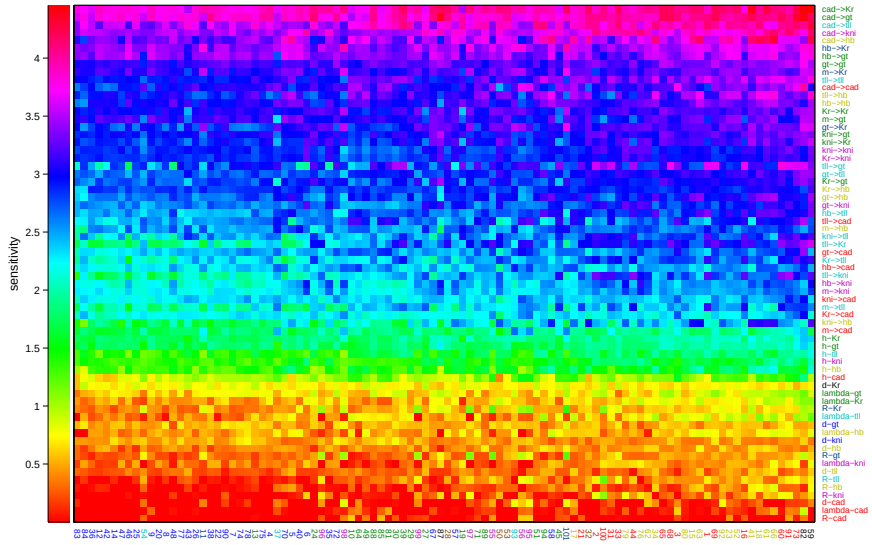


Figure 7.8: Intensity plot of the sensitivity of all parameters and all circuits sorted according to the average parameter sensitivity and average circuit sensitivity. The most sensitive parameter and most sensitive circuit is shown at the top right.

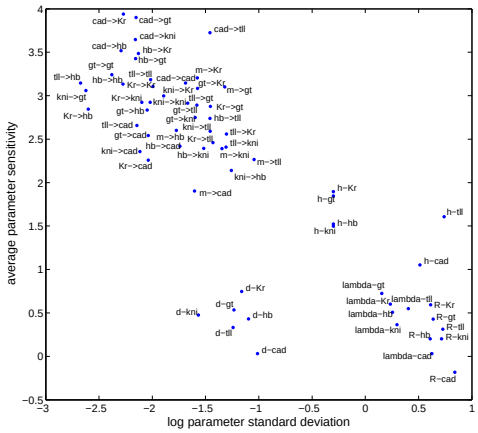


Figure 7.9: Relationship between average $-\log SIs$ (on the y-axis) and parameter standard deviation (on the x-axis). It shows that there is no obvious correlation between SI and STD. Parameters identifiability is not necessarily linked to SI.

parameter	mean parameter value	average lower value	average upper value	relative lower (%)	relative upper (%)
D_{Kr}	0.246	0.173	0.196	70.39	79.52
D_{cad}	0.242	0.242	2.174	100	898.07
D_{gt}	0.201	0.192	0.416	95.23	206.54
D_{hb}	0.238	0.23	0.548	96.8	230.29
D_{kni}	0.292	0.258	0.436	88.43	149.35
D_{tll}	0.278	0.278	0.698	100	250.82
H_{Kr}	-2.97	0.014	0.013	0.49	0.44
H_{cad}	5.961	0.112	0.098	1.87	1.64
H_{gt}	-2.97	0.015	0.016	0.51	0.52
H_{hb}	-2.97	0.041	0.036	1.38	1.21
H_{kni}	-2.97	0.033	0.033	1.13	1.13
H_{tll}	-1.613	0.028	0.03	1.76	1.89
L_{Kr}	7.649	0.275	0.284	3.6	3.72
L_{cad}	14.912	1.121	1.012	7.52	6.78
L_{gt}	5.768	0.2	0.212	3.47	3.67
L_{hb}	7.472	0.348	0.374	4.66	5
L_{kni}	7.252	0.457	0.522	6.31	7.2
L_{tll}	6.745	0.341	0.359	5.05	5.32
R_{Kr}	20.756	0.297	0.281	1.43	1.36
R_{cad}	24.235	2.13	2.154	8.79	8.89
R_{gt}	27.199	0.419	0.429	1.54	1.58
R_{hb}	21.926	0.731	0.696	3.33	3.18
R_{kni}	23.302	0.766	0.762	3.29	3.27
R_{tll}	24.464	0.58	0.607	2.37	2.48
bcd_{Kr}	4.49E-02	8.05E-04	5.57E-04	1.79	1.24
bcd_{cad}	-2.37E-02	1.39E-02	1.42E-02	58.53	59.76
bcd_{gt}	5.61E-02	7.09E-04	1.13E-03	1.26	2.02
bcd_{hb}	2.41E-02	2.68E-03	3.23E-03	11.12	13.38
bcd_{kni}	2.70E-02	4.48E-03	4.41E-03	16.59	16.34
bcd_{tll}	-8.01E-02	7.42E-03	7.08E-03	9.26	8.83
Kr→Kr	2.00E-02	9.30E-04	8.32E-04	4.64	4.16
Kr→cad	-2.51E-02	6.71E-03	5.17E-03	26.68	20.55
Kr→gt	-7.54E-02	1.49E-03	1.63E-03	1.98	2.17
Kr→hb	4.23E-05	1.61E-03	1.46E-03	3793.97	3440
Kr→kni	-4.51E-03	1.28E-03	1.25E-03	28.32	27.82
Kr→tll	-5.14E-02	4.83E-03	4.66E-03	9.39	9.07
cad→Kr	2.38E-02	1.27E-04	1.25E-04	0.53	0.53
cad→cad	-2.81E-02	9.13E-04	8.00E-04	3.25	2.85
cad→gt	2.12E-02	1.41E-04	1.34E-04	0.66	0.63
cad→hb	1.34E-02	4.85E-04	3.91E-04	2.91	2.3
cad→kni	2.52E-02	2.52E-04	2.55E-04	1	1.01
cad→tll	1.79E-02	2.17E-04	2.36E-04	1.21	1.32
gt→Kr	-4.19E-02	1.05E-03	9.60E-04	2.51	2.29
gt→cad	-3.25E-02	3.71E-03	2.84E-03	11.42	8.75
gt→gt	1.47E-02	6.33E-04	5.80E-04	4.32	3.96
gt→hb	6.95E-03	1.91E-03	1.62E-03	27.49	23.27
gt→kni	-1.85E-02	2.12E-03	1.98E-03	11.42	10.65
gt→tll	-1.80E-02	1.41E-03	1.54E-03	7.81	8.55
hb→Kr	-4.23E-03	4.22E-04	2.88E-04	9.99	6.8
hb→cad	-3.52E-02	4.83E-03	3.75E-03	13.73	10.68
hb→gt	-1.21E-03	3.27E-04	5.06E-04	26.95	41.74
hb→hb	1.98E-02	8.63E-04	7.94E-04	4.35	4
hb→kni	-5.61E-02	4.91E-03	4.79E-03	8.77	8.54
hb→tll	-2.15E-02	2.20E-03	2.29E-03	10.19	10.63
kni→Kr	-1.39E-02	1.20E-03	1.14E-03	8.29	8.82
kni→cad	-1.92E-02	4.96E-03	4.50E-03	25.81	23.41
kni→gt	1.51E-03	9.37E-04	9.81E-04	62.21	65.13
kni→hb	-1.08E-01	1.26E-02	9.90E-03	11.72	9.2
kni→kni	1.83E-02	1.29E-03	1.33E-03	7.08	7.3
kni→tll	-5.02E-02	3.41E-03	3.63E-03	6.79	7.23
tll→Kr	-9.06E-02	5.62E-03	5.19E-03	6.21	5.74
tll→cad	-1.97E-02	3.59E-03	2.09E-03	18.22	10.6
tll→gt	-2.44E-02	2.32E-03	2.05E-03	9.52	8.39
tll→hb	1.25E-03	1.23E-03	5.94E-04	98.45	47.61
tll→kni	-8.91E-02	6.32E-03	6.18E-03	7.1	6.93
tll→tll	1.67E-02	7.74E-04	7.85E-04	4.63	4.7

Table 7.2: Table shows the relative and absolute parameter sensitivity intervals. Colours indicate the level of sensitivity where green are the least sensitive (such as the diffusion coefficient) yellow intermediates and red the most sensitive's.

-2.5 are shown in blue and the circuits with $H_{hb,Kr,gt,kni} = -3.5$ in red. The circuits with an average sensitivity higher than 2.05 are significantly less robust towards fluctuations than the circuits with an average sensitivity lower than 2.05. The circuits with $H = -3.5$ are more sensitive than the circuits with $H = -2.5$. Furthermore the robust circuits have a better overall pattern score on average; however lower sensitivity does not necessarily yield a good pattern score. This suggests that the average sensitivity is not the only determinant for robustness towards fluctuations. In Fig. 7.11 the correlations between parameter SIs and stochastic simulation of individual gene domain robustness is shown separately for circuits with promoter threshold $H = -2.5$ and $H = -3.5$.

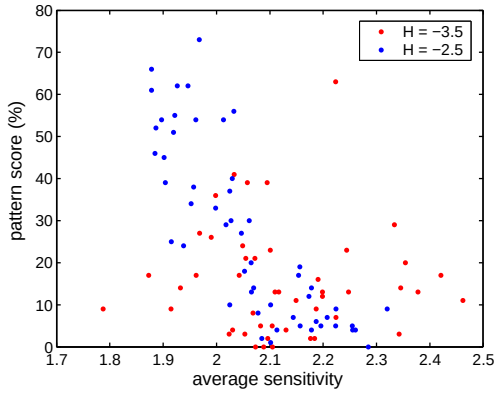


Figure 7.10: Scatter plot of pattern scores versus average circuit sensitivity. On the x-axis, average sensitivity of a circuit is plotted and on the y-axis the overall pattern score as percentage, which was obtained from 100 stochastic simulation runs. This figure shows that sensitive circuits are not robust towards fluctuations.

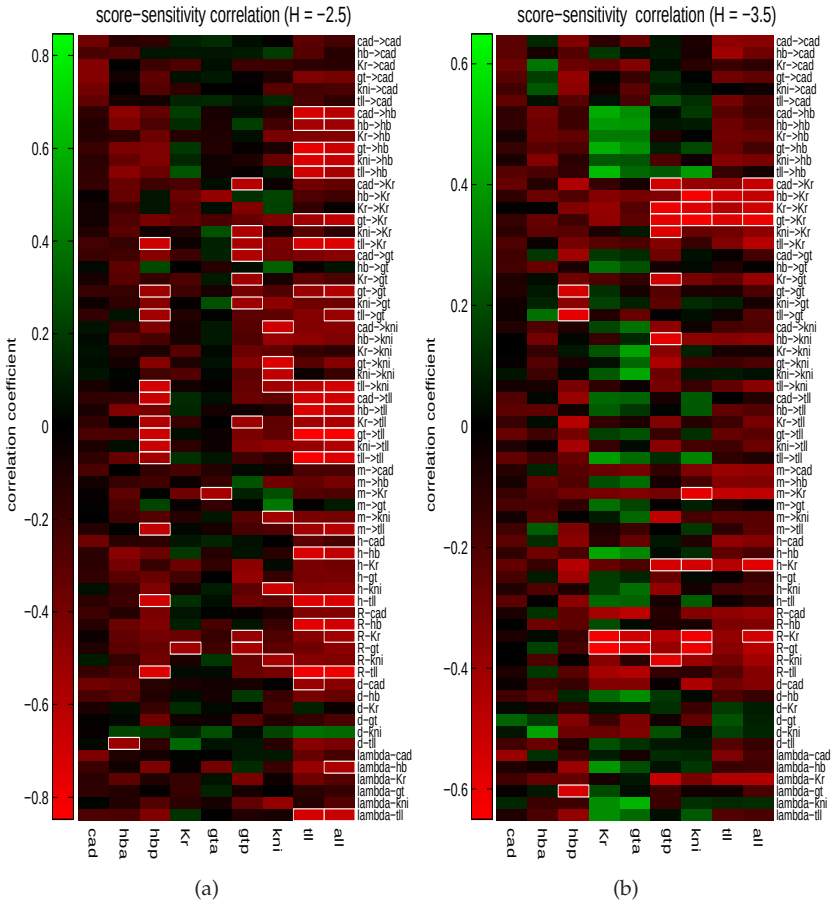


Figure 7.11: Correlation between the score of the eight different expression domains scores and all the parameter sensitivity intervals obtained for solution with respectively $H = -2.5$ on the left panel A, $H = -3.5$ on the right panel B. Light green expresses a strong positive correlation and light red a strong negative correlation. Squares in white borders are the most significant correlations.

7.3 Discussion

The 101 gap gene circuits were previously obtained with inverse modelling [78, 121] of a seven gene (*bcd, cad, hb, Kr, gt, kni* and *tll*) connectionist model using detailed spatio-temporal gene expression data [190]. All circuits were selected only on the basis of a low RMS value ($RMS \leq 12$), and in all cases the circuits were able to reproduce the dynamics and the patterns accurately. However, the parameter values of the circuits appeared not to be well defined, for some parameters multiple clusters were found, which represented multiple circuit topologies. Furthermore many parameters showed a single cluster with a high degree of scattering around the mean (see Fig. 6.3). To further analyze the properties of the model and to classify the quality of the circuits we conducted a perturbation analysis. We analyzed the robustness of 101 *Drosophila melanogaster* gap gene circuits using two different approaches, first the robustness of pattern formation towards intrinsic fluctuations in gene expression levels was investigated using a stochastic model and secondly the sensitivity of the circuits with respect to the parameters in the model was investigated using a simple parameter perturbation technique.

Robustness of domain formation Introducing fluctuations in the model had a profound effect on the formation of the expression patterns. Although some of the circuits showed robust formation for most domains, none of the circuits were sufficiently robust when looking at the formation of all the domains. We observe that fluctuations can lead to an increase or decrease of domain amplitudes (see Fig. 7.1). Furthermore we also observe posterior and anterior boundary shifts, which leads to domain expansion, domain contraction or domain shifts. We also observe that domains completely disappear or that domains appear in other regions, where they repress other domains. Especially in a significant number of circuits the anterior domains gt and Kr did not form robustly during the simulations. The least robust was the Tll domain at the posterior end, this domain interacted with gt and hb, which all showed defects in many circuits. We believe that *tll* weak robustness is caused by an incomplete model where gap gene are regulating *tll* which is inconsistent with experimental evidence [30, 120].

Looking at the evolution of the system in phase space, we found that the pathways in many cases were not well defined, which appeared to be linked to the existence of multiple attractors. The fluctuations allowed the system to jump from one attractor to another, causing the system to evolve into a pattern very different compared to the deterministic model. This suggests that by using the reverse engineering approach many circuits can be found, which are well defined for the deterministic case but not for the stochastic case. Multiple attractors can easily occur in a non-linear model with many parameters, where overfitting may lead to many connections in the circuits. Also the type

of feedback loops may be crucial, e.g. *cad* has a negative auto-regulation, which reduces the effect of fluctuations, however all other genes have strong positive auto-activation [49,65]; in these cases fluctuations are amplified if there are no negative feedback loops associated with the increase of the expression level.

The effect of certain parameters on robustness The connectionist model contains a promoter threshold for each gene, which determines if gene production is on or off when there are no other inputs. In the current gap gene model the production of hb, Kr, gt and kni can only be turned on by maternal inputs or by auto-activation, therefore the promoter threshold are set to a fixed negative value, which was either $H = -3.5$ or $H = -2.5$. Looking at the robustness of individual circuits, we find that circuits with promoter threshold set to $H = -2.5$ are more robust towards fluctuations compared to circuits with promoter threshold set to $H = -3.5$, and also their parameters are on average less sensitive. In general we find that weaker maternal inputs from bicoid and caudal increase the robustness with respect to fluctuations. Furthermore higher promoter rates, decay rates and diffusion coefficients improve the robustness towards fluctuations.

We have seen that circuits are relatively less sensitive with respect to diffusion coefficients, and with some extent, to decay time constant. Regarding the diffusion coefficient, this result confirms observations made elsewhere where it was suggested that the diffusion is not essential to explain precise gene expression pattern formation [2, 121], but the diffusion term does account for the boundary nuclei effect [101]. It was also shown by Nusslein-Volhard et al. [196] that the effect of diffusion is reduced with the exponential increase of the number of nuclei. Also, Gregor et al. [100] showed that diffusion coefficient does not scale with varying embryos length. In Tab. 7.2, we showed that the rest of the parameters have a mixed sensitivity behavior and do not fall into a specific category. Some of the weights are very sensitive and some are not, idem for the production rate. Although we do not have precise biological explanations regarding the difference of the sensitivity behaviour of the parameters, it might be interested to experimentally investigate if such difference is a property of the connectionist model or a characteristic of the regulatory mechanism. The sensitivity information can then guide the selection of the optimal mutation targets and thereby reduce the experimental effort. This validation could be done for example by measuring mRNA degradation rate of zygotic Hb in embryos with over expressed maternal *hb*, or by measuring the binding affinity in mutants. Also one could consider inducing genetic mutation to control kinetic parameters that can be measured [73].

We also find that certain parameters that regulate posterior Tll and also some parameters that regulate *gt*-anterior and Kr affect robustness of the corresponding domains. Furthermore we find that robustness and parameter sensitivity are linked. Especially for the group with a lower promoter threshold we find that circuits with a lower average sensitivity are more robust towards fluctu-

ations. These results show that in the current gap gene model not just the network topology but to a large extent the precise value of the parameters determines robustness. We noticed that inputs of *cad* to other genes correlate negatively with robustness, these parameters are amongst the most sensitive in the model, with a $-\log SI$ in the order of 4. If we assume that fluctuations are proportional to the square root of the concentration level (this is a reasonable assumption for steady production) the relative fluctuation level is about $1/\sqrt{n} * 100\%$, where n is the number of molecules. For n = 100, n = 1000 and n = 10000 the relative fluctuation level then is 10%, 3.2 % and 1%. The most sensitive parameters in the model have a relative sensitivity of about $0.5 - 2\%$ (see table 1 supplementary material). This holds assuming that 20% of the circuit RMS is a critical cut-off.

Attempt to deduce some of the regulatory interaction Based on the results obtained in the previous chapter (focussing only on circuits that have regulatory interactions consistent with experimental evidence) and the current correlation analysis, we ignore parameters that are still unclear being indeterminable, we focus on those having very weak cross-correltions, the parameter estimates suggest that the following regulatory interactions can be trusted:

1. All the gap gene are activated by Cad. (very weak correlation coefficient for any $W^{cad}_{Hb,gt,kr,kni}$)
2. All the gap gene but kni are activated by Bcd. (very weak correlation coefficient for any $bcd_{Hb,gt,kr}$)
3. Hb, Kr gene have an auto-activation. (very weak correlation coefficient for any W^a_a)
4. kni does not have a auto-repression ($W^{kni}_{kni} > 0$), but certitude on auto-activation can not be deduced (strong correlation coefficient with most parameters)
5. Mutual repression between Hb and Kni
6. Mutual repression between Gt and Kr
7. Kr does not regulate *hb* (no correlation with any other parameter)
8. Kni represses the central domain of *hb*
9. Asymmetric repression from right to left of Hb $\dashv$ Kr $\vdash$ kni $\vdash$ Gt $\vdash$ Hb.

These interactions are consistent with the regulatory mechanism proposed in [121] as well as those obtained in early literature [36, 54, 70, 89, 106, 119, 183, 223, 258, 283]. Also, these interactions are defined by the parameters that were reported to be identifiable in [6]. The alternative interactions proposed by the other circuits are consequence of over-fitting leading to parameters having wrong qualitative regulation compensate by other parameters.

7.4 Conclusions

In this chapter, we have provided a robust analysis of the model used to infer the gap gene regulatory network of *Drosophila* melanogaster. The model has been extensively used elsewhere [78,121,206] to simulate and provide some explanations concerning the regulatory mechanism that leads to precise pattern formation. Unfortunately, many assumptions were based on a limited number of circuits obtained using simulated annealing and previous researchers assumed that the model was correct with a correct topology. We have shown that the mathematical model leads to different circuits all capable of reproducing the quantitative spatio-temporal gene expression pattern. Consequently, it is difficult to decide based solely on the architecture which circuit is the correct one.

Robustness towards fluctuation has revealed that the overall gap gene domain tends to be poorly resistant to perturbation and this weak property could be related to some particular interactions predicted by the circuits. Furthermore, parameter perturbation analysis has shown that the circuits with lower sensitivity do not necessarily yield to robustness to fluctuation. The reason for these few exceptions is related to the promotor threshold and to local domain robustness, which can considerably affect the overall global robustness. Overall, the analysis shows that the network possesses modular robustness and some local properties may affect the robustness of a gene expression locally as show in Fig. 7.11. This feature is strongly related to the network topology as well as the interaction weights.

From a biological point of view, this paper has shown that it is difficult to relate the connectionist model with biological evidence. The model ability to simulate the gene expression does not necessarily provide meaningful information since alternative networks are predicted. Therefore, it would be interesting to test some of the different alternatives, especially when there is not yet any experimental evidence to invalidate. However, it was recently demonstrated that it is still possible to draw qualitative conclusions on the regulatory topology of the gap gene network [6,80]. The overall analysis has shown that based on the robustness toward gene expression fluctuations and parameter perturbations, it is possible to identify robust circuits as well as the parameter that are identifiable according to their sensitivity intervals. For a computational/system biologist, this shows that it is essential to further analyze a model prediction, when results are obtained from reverse engineering based on parameter estimation, since some of its properties may or may not invalidate the results.

This analysis revealed some of the limitations of the way reverse-engineering was conducted. The next chapter will provide some preliminary suggestions to efficiently improve the GRN inference to avoid reverse engineering that leads to circuits with different topologies.