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

Fomekong Nanfack, Y.

Publication date
2010

6

Analysis of the gene circuit parameters [1]

In Chap. 4, we have presented the gene circuits using an evolution strategy instead of parallel simulated annealing for the parameter estimation. A connectionist description was used to model the gap gene regulatory network in *Drosophila melanogaster*. We have shown that the (μ, λ)-ES could lead to circuits with the same qualitative score in a significant smaller computational time in comparison with circuits obtained from Jaeger et al. [121]. In the previous chapter, we have shown that compared to the variance in the experimental data, all solutions fitted the dataset spatially and temporally accurately. Furthermore, the model reproduced the experimentally observed shift behaviour of the expression profiles. In 2004 [121], Jaeger et al. suggested that the gap gene domain shift was a consequence of the asymmetric repression between the gap genes. This hypothesis was based on analysis of one circuit out of the 10 circuits obtained from the reverse engineering using parallel simulated annealing (PLSA) [44]. The limited number of circuits were obtained due to computational limitations. Also, analysis on the reliability of the parameters was not performed. In the previous chapter, we have seen that small defects in the gene expression patterns were caused by some inconsistencies within the regulating parameters, which is also causing the circuits long term dynamics to converge to different attractors. This shows that the regression re-

[1] This chapter is partially based on the paper:
Yves Fomekong-Nanfack and Jaap A. Kaandorp and Joke Blom, "Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of *Drosophila melanogaster*", Bioinformatics, Vol. 23, No. 24, pp. 3356-3363. September 2007. [78] and
Yves Fomekong-Nanfack and Marten Postma and Jaap A. Kaandorp "Inferring *Drosophila* gap gene regulatory network: pattern analysis of simulated gene expression profiles and stability analysis", submitted March 2009. [80]

sults could be inaccurate and subsequent the conclusions can be erroneous. In the current chapter, we analyse the reliability of the parameters by simple post-regression diagnostic tests for the evaluation of the overall model validity. First, we present the parameter estimates and their distributions. From this distribution, we derive the different regulatory networks obtained from the gene circuits, and in the last section we analyse the identifiability of the parameters.

6.1 Statistical analysis of parameters estimates

For networks the general tendency is to focus on the qualitative value (or sign) of the interaction. Given the nature of the connectionist model, we cannot just focus on the sign of the weight describing the interaction between two genes, but the strength of the weight might also have an importance justified by the sigmoid function used (see Chap. 4 Section 4.2.2).

Scatter plot of gap-circuits In Figs. 6.1 and 6.2 scatter plots are given comparing the parameters obtained by our simulation using ES and island-ES with those obtained by Jaeger et al. [120, 121]. The parameters are for the 62-dimensional problem with $h_{(hb,Kr,gt,kni)} = -2.5$. The parameters obtained are in most cases comparable for different optimisation runs and with the ones obtained in [121]. Incidentally our regulatory weight matrix entries differ from those obtained in [121], like W^{cad}_{tll} and W^{hb}_{tll}.

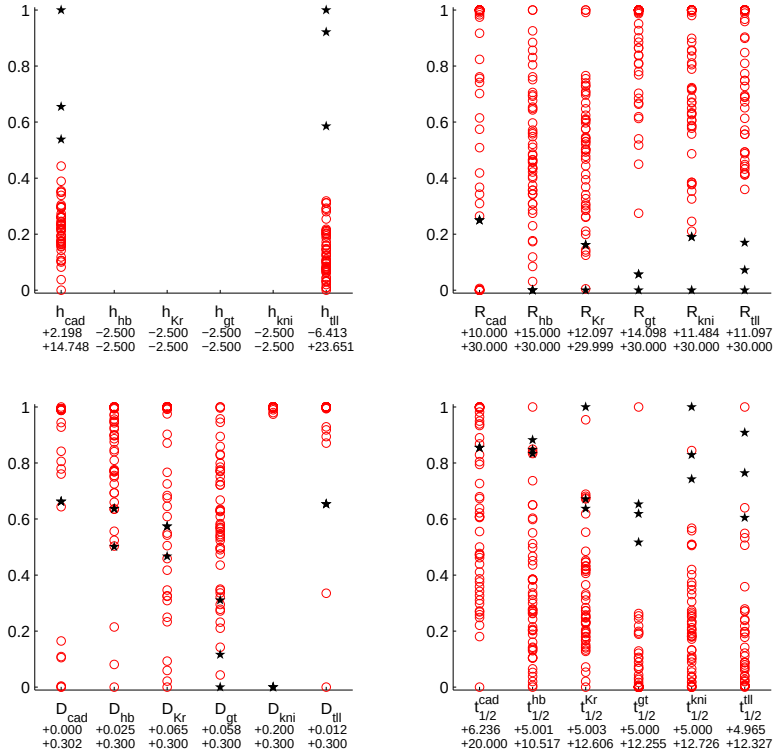


Figure 6.1: Scatter plot of parameters h, R, D and $t^a_{1/2} = \ln(2)/\lambda_a$. Each vertical line contains 53 different values for the same parameter. Each circle represents a parameter obtained from our simulations and the black stars represent parameters obtained by [120, 121] using parallel simulated annealing. All the gap-circuits resulted in a RMS ≤ 12.00. Our 50 results were obtained with the different ES and island-ES with fixed promoter thresholds $h_{(hb,Kr,gt,kni)} = -2.5$. The parameters are scaled to [0,1] using the minimum and maximum values given below the x-axis.

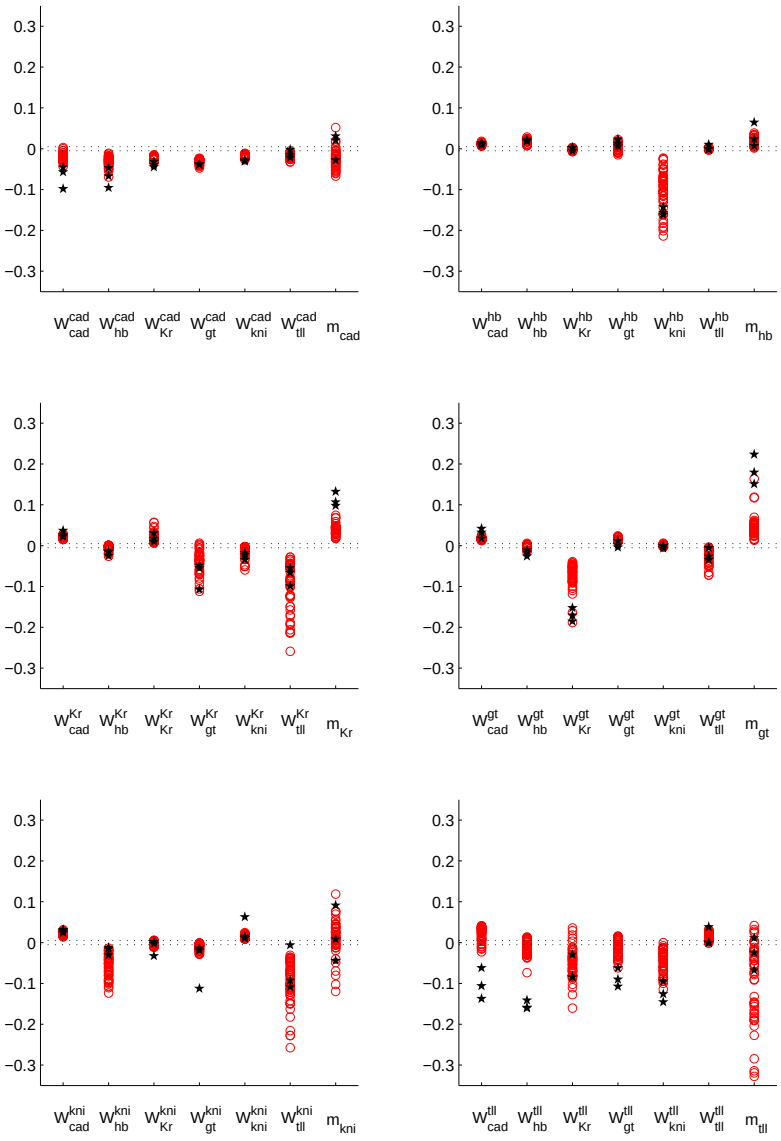


Figure 6.2: Scatter plot of the regulatory influences. W-entries and m-values above the horizontal line at 0.005 are considered to be activating, below -0.005 inhibiting (see also Tab. 6.1). Each vertical line contains 53 different values for the same parameter. Each circle represents a parameter obtained from our simulations and the black stars represent parameters obtained by [120,121] using parallel simulated annealing. All the gap-circuits resulted in a RMS ≤ 12.00. Our 50 results were obtained with the different ES and island-ES with promoter thresholds $h_{(hb,Kr,gt,kni)} = -2.5$.

Distribution of the parameters In Fig. 6.3, the distribution of the 66 parameters is shown. The grey line represents the different 101 values of each parameters and the blue line their average. From this figure, we see that the parameters distribution varies from circuit to circuit. None of the regulatory parameters such as the production rate, the diffusion or the decay seems to be very consistent from solution to solution. In many cases, the promotor rate and the diffusion hit the upper boundary. Many parameters show a strong tendency to converge to a specific value or location (specifically those around zero such as W_{gt}^{hb}, W_{hb}^{kr} or W_{gt}^{kni}). In some cases, we see a very broad distribution around the mean, especially for the parameters describing strong repression. For few cases, some regulatory parameters can have both types of interactions (bcd_{tll}, W_{tll}^{hb}, bcd_{kni}).

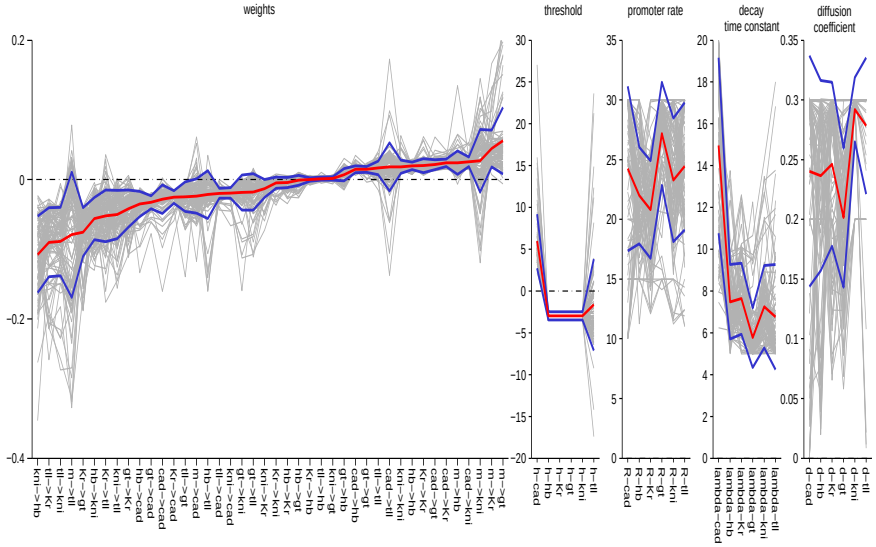


Figure 6.3: Distribution of the 66 parameters obtained from the 101 gene circuit. The solid red line represents the average value, the solid blue lines represent the standard deviations and the light gray lines represent the different individual circuit parameters. Parameters are sorted according to their mean value. Most parameters show a strong tendency to cluster around a particular value, defining the type of interaction. However, some of them have a very broad distribution around their mean and in a few cases, they show all different types of interactions i.e. activation, repression or no interaction.

Circuits distribution From the previous parameter distributions, we see that some of the weights are considerably scattered. In [121], the network was derived based on 10 circuits obtained from the same simulated annealing setting. Assuming that the spread distribution of some of the parameters might result from the various setting of ES, we present in Tab. 6.1 the different regulatory

combination obtained from the different setting. The different setting seems to lead to more or less the same network. Nevertheless, some parameters do not have a 100% confidence on the sign (describing the interaction) whereas to derive any network, it is necessary to gain confidence on the parameters. From the parameter distributions presented above, we see that some parameters have very weak regulatory relationships. It is difficult to determine if they are clearly describing an absence of interaction between two genes or if it is just a very weak regulation. Following Jaeger et al. [120], we have assumed that parameters comprise in the interval $[-0.005, 0.005]$ correspond to a non-interaction. In the previous chapter we have suggested that the model might be over-fitted (because of incomplete data or incomplete model). Some parameters have very small variance and it can be assumed that they represent the real interactions. However, it is not trivial to discern which parameters are describing biological interactions from those that are over-fitted. It must be confirmed that the available data uniquely determine the value of each parameters. From the available data and the current model, it is possible that the model is structurally identifiable yet some of the parameters can be undeterminable [6, 124]. This would suggest some hidden mechanisms such as missing genes or hidden dependecies within the data.

	bcd	*cad*	*hb*	*Kr*	*gt*	*kni*	*tll*
N=62 h=-2.5 Regulatory network							
cad	42/5/3	47/3/0	50/0/0	50/0/0	50/0/0	50/0/0	50/0/0
hb	0/2/48	0/0/50	0/0/50	3/47/0	6/9/35	50/0/0	0/50/0
Kr	0/0/50	0/0/50	19/31/0	0/0/50	46/3/1	39/11/0	50/0/0
gt	0/0/50	0/0/50	4/45/1	50/0/0	0/0/50	0/49/1	48/2/0
kni	12/5/33	0/0/50	50/0/0	12/36/2	36/14/0	0/0/50	50/0/0
tll	42/4/4	5/4/41	31/5/14	45/1/4	34/7/9	49/1/0	0/4/46
N=62 h=-3.5 Regulatory network							
cad	36/1/4	41/0/0	41/0/0	41/0/0	41/0/0	41/0/0	40/1/0
hb	1/0/40	0/0/41	0/0/41	2/38/1	5/12/24	41/0/0	0/36/5
Kr	0/0/41	0/0/41	2/38/1	0/0/41	41/0/0	29/12/0	41/0/0
gt	1/1/39	0/0/41	1/28/12	41/0/0	0/0/41	0/38/3	35/6/0
kni	2/2/37	0/0/41	40/1/0	15/25/1	33/8/0	0/0/41	41/0/0
tll	27/6/8	2/5/34	29/7/5	38/2/1	30/6/5	40/1/0	0/2/39
N=66 Regulatory network							
cad	20/2/12	34/0/0	34/0/0	34/0/0	32/2/0	34/0/0	30/2/2
hb	4/14/16	22/8/4	3/17/14	13/21/0	9/14/11	34/0/0	21/13/0
Kr	7/3/24	20/4/10	27/7/0	0/3/31	31/3/0	31/3/0	34/0/0
gt	0/0/34	5/5/24	17/13/4	34/0/0	0/2/32	5/16/13	26/7/1
kni	26/0/8	4/3/27	34/0/0	10/18/6	24/10/0	0/1/33	34/0/0
tll	31/1/2	3/4/27	21/7/6	20/3/11	26/4/4	23/7/4	0/2/32

Table 6.1: Distribution of the entries of the regulatory weight matrix W and the regulatory input of *bcd* to the zygotic system (m_a). The subtables give the type of interaction between two genes for the 3 different settings (the dimension of the optimization problem and the value of the fixed parameters) independently of the ES configuration. The triplet R/N/A represents the number of interactions of type: repression ($W_a^b < -0.005$), no interaction ($W_a^b \in [-0.005, 0.005]$) and activation ($W_a^b > 0.005$), idem for m_a. The sum of the triplet is the number of selected simulations for the corresponding experiment. Table rows represent target and columns represent regulators. Background color (green) represents activation, (light-blue) represents no interaction and (pink) represents repression. The choice of the type of interaction is based on the highest triplet value. The Reduced Search settings ($N = 62, h = -2.5$) and ($N = 62, h = -3.5$) have exactly the same regulatory network of interactions. This is very similar to the one obtained by [120] except for 3 parameters W_{Kr}^{hb}, W_{gt}^{hb} and W_{kni}^{Kr} for which we conclude that there is no interaction and [120] derived a repression.

6.2 Correlation analysis

One simple approach to explore the parameter determinability is to use cross-correlation between parameters [124]. A correlation matrix shows the degree of association between two parameters. The parameter values are centered on the mean and the normalised cross-correlation between -1 and $+1$ is computed using the Pearson correlation. From the inverse modelling paradigm, the correlations describe compensation that may arise from an incomplete or inaccurate data set, i.e. the data set does not contain enough information to cover all parameters. Compensation may however also arise from an incomplete model, i.e. the model does not sufficiently represent the underlying biological mechanism. Typically compensation 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.

Production rate and decay Systematically for all genes but *tll*, strong negative correlation is observed between all pairs of production rate and decay coefficients ($\mathrm{r}(R_a/\lambda_a) \geq 0.65$). The strong linear correlation represents the scaling of the expression profile. If one increases the production rate of a gene a and wants to keep the system in its normal expression level, one has to decrease the decay related to the protein half-life of the product of gene a. Fig. 6.5 illustrates the negative pairwise production/decay correlation of the genes *hb* and *kni*.

Gene regulatory parameters

In classical micro-array data analysis, a direct correlation exists between the regulator-regulatee relationships. This association is described if a set of genes (regulatees) increases or decreases their protein level with the increase or decrease of the expression of another group of genes (regulators). In the current context, we see the same behaviour at the parametric level. It is necessary to discriminate between interactions that are consequence of an over-fitting and parameters that might suggest a real interaction. First, we describe the different regulatory interactions predicted by the 101 gap gene circuits obtained from [78,121]. Based on the correlation-matrix, we identify the parameters that have a very large cross-correlation with all (or most) other parameters. This implies that it is not possible to trust their significance or ensure the regulatory interactions they predict. We do not discuss *cad* and *tll* regulation since the regulatory

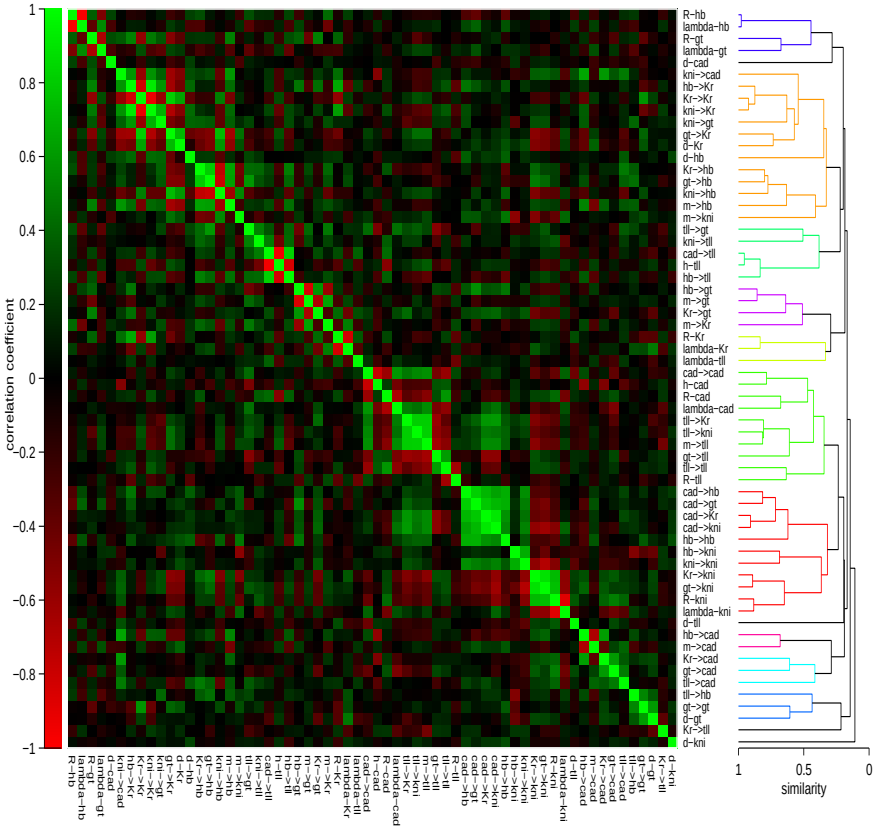
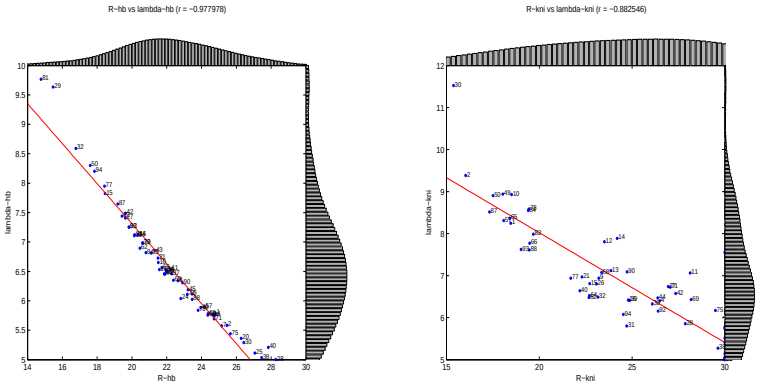


Figure 6.4: Parameter correlation matrix. Left: Matrix showing the pairwise correlation; the colour scale goes from intensive red (strong negative correlation) to bright green (positive correlation). The correlation matrix shows that there exist many pair wise correlations that tend to form clusters. Right: The absolute value of the correlation coefficients are used as a similarity measure to cluster the parameters, which is presented as a dendrogram. The parameters are sorted according to the dendrogram.



(a) R_{Hb}/λ_{Hb} (b) R_{kni}/λ_{kni}

Figure 6.5: Scatter plots of production and decay with regression lines and correlation coefficients. At the top and right the estimated parameter distribution is shown, which was calculated using using ksdensity.

interactions predicted by the circuits can not reflect biological reality since *cad* and *tll* are not regulated by other gap genes at this stage of development.

hunchback regulation obtained from the reverse engineering is mainly controlled by the following:

- activation by Bcd and Cad, confirming that they are both the primary activators of the gap domain, acting respectively on the anterior and the posterior.
- auto-activation.
- repression by Kr (weak), Tll (weak), Gt and Kni (strong)
- activation by Kr (weak) , gt and Tll (weak)

The typical correlations shown in Fig. 6.7 of the parameters regulating *hb* are:

1. negative correlation between opposite regulators (W_{hb}^{bcd} vs. W_{hb}^{kni}) and (positive W_{hb}^{gt} vs. W_{hb}^{kni})
2. positive correlation between regulators with opposite functionality on the same domain on a gene (W_{hb}^{gt} vs. bcd_{hb} and W_{hb}^{Kr} vs. bcd_{hb})
3. co-correlation caused by the domain geometry (W_{hb}^{cad} vs. W_{hb}^{hb})

Perkins et al. [206] suggested that the posterior of *hb* is activated by Tll while Jaeger et al. [121] found that posterior *hb* is activated by Cad. We found that Gt and Tll have both positive and negative regulatory parameters on *hb*. Assuming that posterior *hb* is also activated by Tll, we were expecting to see negative correlation between Cad and Tll regulation on *hb*. Surprisingly, it was not the case, and *hb* regulation by Tll did not show any particular correlation with any other parameter. In fact, it shows very weak correlation with most of the other parameters implying that this parameter is well determined and that Tll activates posterior domain of *hb*. It was shown that posterior *hb* is regulated by Tll and Hkb [169]. Since Hkb is not present in the model, we believe that Cad is taking over this role.

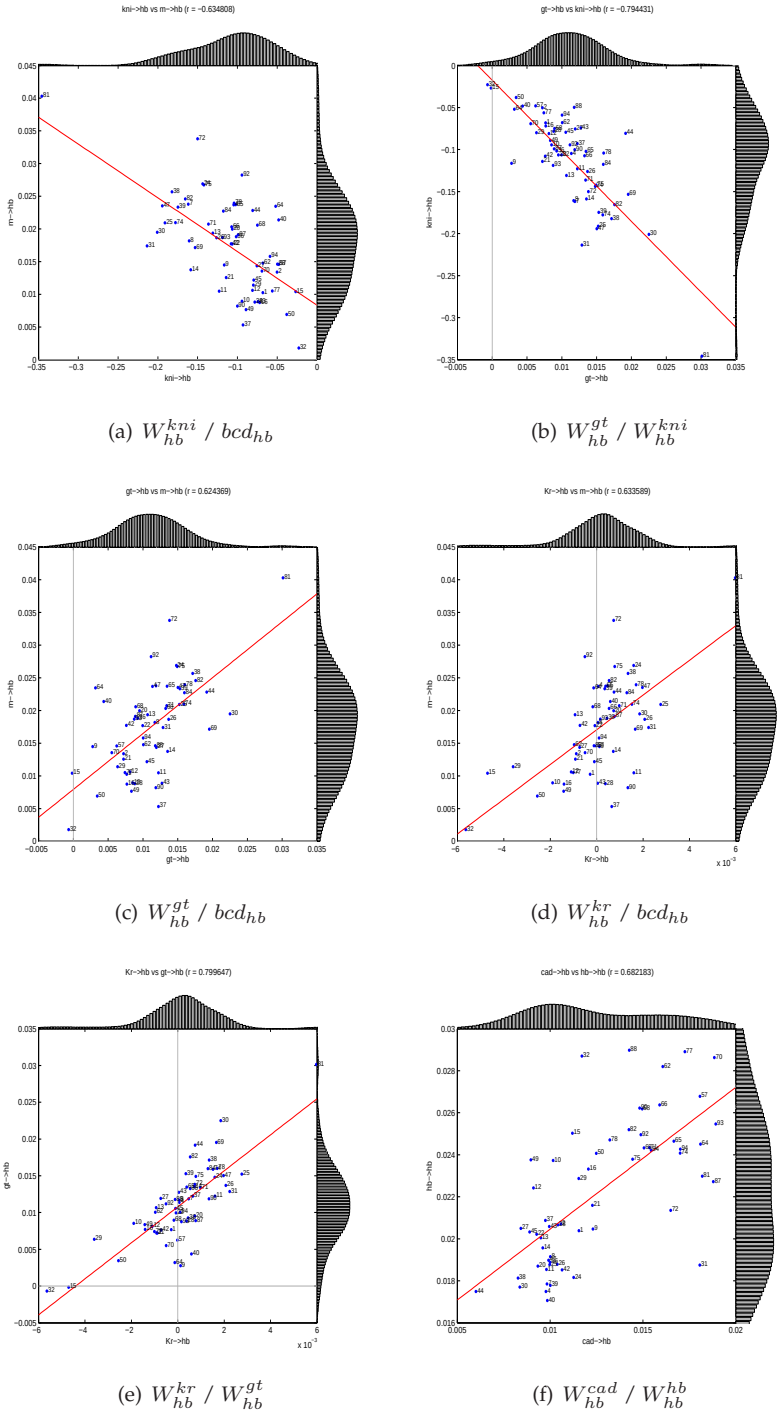
krüppel : from the parameter estimates, the different regulatory mechanisms that control *Kr* gene expression dynamic is defined by:

- maternal activation by Bcd and Cad.
- auto-activation.
- repression by Hb, Gt, Kni and Tll.
- activation by Hb and kni.

The correlations shown in Fig. 6.7 with a meaningful value are the following:

1. negative correlations between *Kr*'s repressors when their contribution is mostly on overlapping domain : W_{Kr}^{hb} vs. W_{Kr}^{gt} at the anterior domain and W_{Kr}^{gt} vs. W_{Kr}^{kni} at the posterior domain.
2. negative correlation between W_{Kr}^{kr} vs. W_{Kr}^{hb} and W_{kr}^{kni} (decrease repression weight if auto-activation is weaker)
3. positive correlation between activators vs. repressors when their contribution is mostly on the same domain : W_{Kr}^{hb} vs. R_{Kr} , W_{Kr}^{kni} vs. R_{Kr}, W_{Kr}^{Kr} vs. W_{Kr}^{gt} and W_{Kr}^{hb} vs. W_{Kr}^{kni} (*NB: kr auto-activation and production contribute in the entire domain*).
4. positive co-correlation caused by the domain geometry: W_{kr}^{hb} vs. W_{kr}^{kni}

Jaeger et al. [121] suggested stronger influence of Bcd than Cad and found $bcd_{kr} \geq W_{kr}^{cad}$. We find equivalent weight for W_{kr}^{cad} and bcd_{kr}. However, we did not estimate the total contribution of the gene's parameter and the gene's product. It is suggested that Hb activates anterior *Kr*. The resulting gap gene circuits found both role activation (very weak) and repression. The strong correlation between W_{kr}^{hb} and W_{kr}^{kni} suggests that if one repression increases, the other one also has to increases in order to maintain symmetry and to avoid domain expansion on one side. This result confirms Jaeger et al. [121] hypothesis suggesting that Hb and Kni contribute in the establishment of *Kr* border. Hb represses the anterior border while Kni represses the posterior border.



(a) W_{hb}^{kni} / bcd_{hb}

(b) W_{hb}^{gt} / W_{hb}^{kni}

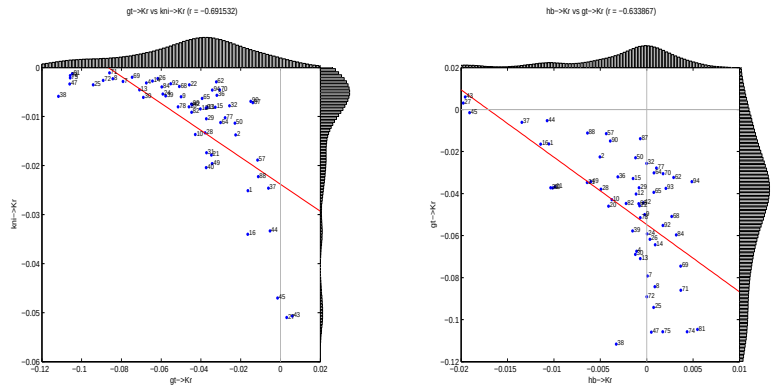
(c) W_{hb}^{gt} / bcd_{hb}

(d) W_{hb}^{kr} / bcd_{hb}

(e) W_{hb}^{kr} / W_{hb}^{gt}

(f) W_{hb}^{cad} / W_{hb}^{hb}

Figure 6.6: Scatter plots of parameters that regulate hunchback



(a) W_{Kr}^{gt} / W_{Kr}^{kni}

(b) W_{Kr}^{hb} / W_{Kr}^{gt}

(c) W_{Kr}^{hb} / W_{Kr}^{Kr}

(d) Kr_{Kr}^{Kr} / W_{Kr}^{kni}

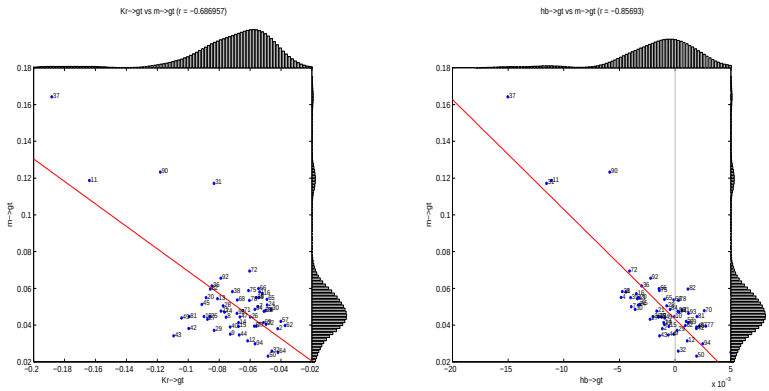
Figure 6.7: Scatter plots of parameters that regulate Kr.

giant : From the 101 circuits obtained, mechanism controlling gt is as follow:

- maternal activation by Bcd and Cad.
- auto-activation.
- repression by Hb, Kr, Kni and Tll.
- activation by Hb and Kni (very weak).

The circuits show that both Bcd and Cad contribute respectively in the expression of anterior and posterior gt. Only two significant correlations (shown in Fig. 6.8) were found: negative correlations between W_{gt}^{Kr} and bcd_{gt} and between W_{gt}^{hb} and bcd_{gt}. The central domain of gt regulation is mainly repressed

by Kr and the negative correlation translates the balance between decreasing repression and decreasing activation. Although Hb role on gt seems weak ($|W_{gt}^{hb}| \leq 0.005$), the correlation with bcd_{gt} shows that Hb represses anterior gt as suggested in earlier literature [70, 121]. When Hb positively regulates gt, Hb mainly contributes in the expression of posterior gt. This is observation is confirmed by the negative correlation between repression of gt by Tll and the regulation of gt by Hb for the case where $W_{gt}^{hb} \geq 0$.



(a) W_{gt}^{kr} /bcd_{gt} (b) W_{gt}^{hb} /bcd_{gt}

Figure 6.8: Scatter plots of parameters that regulate gt.

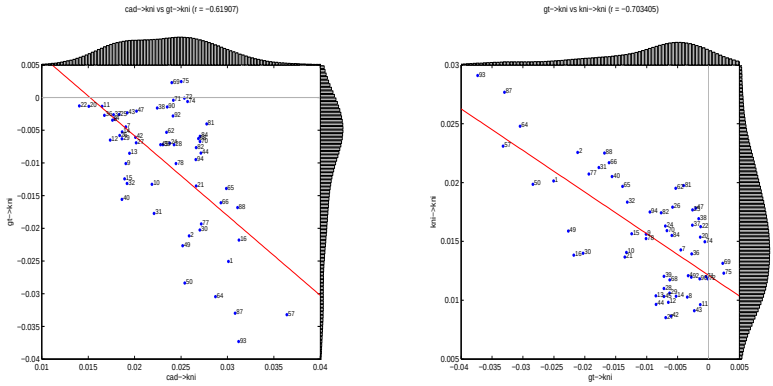
knirps regulation obtained from the gap gene circuits is described as follow:

- maternal activation by Cad and Bcd.
- maternal repression by Bcd.
- auto-activation.
- activation by Kr and Gt
- repression by Hb, Kr, Gt and Tll.

The main correlations with a significant Pearson value are the following:

1. negative correlations between W_{kni}^{cad} vs.W_{kni}^{gt}, and W_{kni}^{gt} vs. positive W_{kni}^{kni}.
2. positive correlation between W_{kni}^{hb} vs. W_{kni}^{kni}, W_{kni}^{gt} vs. R_{kni} and W_{kni}^{Kr} vs. R_{kni} .
3. co-correlation caused by the domain geometry between W_{kni}^{Kr} vs. W_{kni}^{gt}.

1 &2 are direct correlations related to compensation phenomena to maintain the expression level. Jaeger et al. [121] proposed that *kni* anterior border is set by repression by Hb and Kr, and posterior border is controlled by Gt and Tll. They also pointed that Kr might not be necessary in the regulation of *kni*. We found that in 100% of gap gene circuits, *kni* is repressed by Hb and Tll, but it is not systematically repressed by Gt and Kr. W_{kni}^{Gt} andW_{kni}^{Kr} have a similar distribution and seems to have the same role on *kni*. The very strong positive correlation between W_{kni}^{kr} and W_{kni}^{gt} confirms this hypothesis and indicates the role of both parameter in maintaining domain symmetry of *kni* to avoid domain expansion.



(a) W_{kni}^{cad} /W_{kni}^{gt}

(b) W_{kni}^{gt} /W_{kni}^{kni}

(c) W_{kni}^{hb} /W_{kni}^{kni}

(d) w_{kni}^{kr} /W_{kni}^{gt}

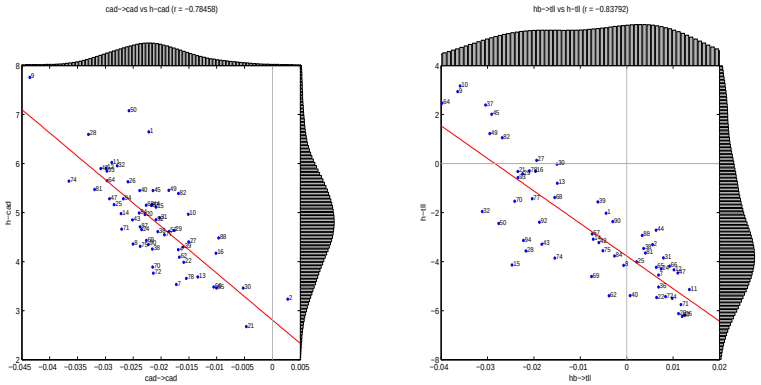
Figure 6.9: Scatter plots of parameters that regulate *kni*.

Diffusion In [121], Jaeger et al. have showed that diffusion does not consistently contributes in the expression of the shift domain. We did not find systematic strong correlation between diffusion and any other parameters beside *kr* and *gt*. Their auto-activation parameters are respectively positively correlated to the diffusion coefficient ($r(W_{gt}^{gt}/D_{gt}) = 0.605$ and $r(W_{Kr}^{Kr}/D_{Kr}) = 0.64$). If their gene concentration is increased by means of auto-regulation, the amount of protein diffusing should also increase. Although these correlations are obvious, we cannot explain why similar feature is not present for *hb, kni*, and *tll*. In contrary, the others diffusion parameters have very weak correlation with any other parameters, signifying that the diffusion coefficient can be determined from the current model with the available data.

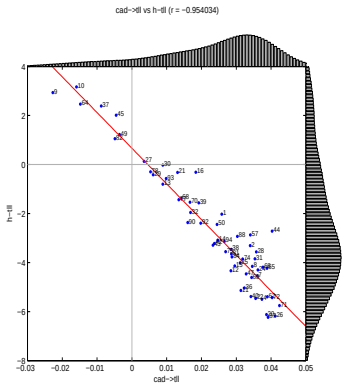
Geometry based co-correlations Clustering the parameters reveals a group composed of Cad activation on *hb, kr, gt* and *kni*. All these parameters are strongly positively correlated (W_{hb}^{cad} vs. W_{kni}^{cad}, W_{gt}^{cad} vs. W_{kni}^{cad}, W_{hb}^{cad} vs. W_{Kr}^{cad}, W_{Kr}^{cad} vs. W_{gt}^{cad}, W_{hb}^{cad} vs. W_{gt}^{cad} and W_{Kr}^{cad} vs. W_{kni}^{cad}). These correlations express the maintenance of gap gene profile proportional to each other on the action of Cad.

Indirect correlations Few indirect correlation of type W_b^a vs. W_d^c mainly caused by profile variation are present. These correlations are mainly related to Tll regulation such as: W_{kni}^{tll} vs. bcd_{tll}, $Wtll_{Kr}$ vs. bcd_{tll}, W_{Kr}^{tll} vs. W_{kni}^{tll}, W_{Kr}^{tll} vs. W_{kni}^{cad} and W_{Kr}^{tll} vs. W_{tll}^{gt}. Also there is a positive correlation between bcd_{Kr} and bcd_{gt}. This indirect correlation is caused by the mutual repression between of Kr and gt. The change in the repressive parameter is balanced by the Bcd. If one repressor increase/decrease, the mutual repressor acts in an similar manner. Consequently, the maternal influence is adjusted to keep the gene expression at the desire level.

Influence of the promoter threshold We also observe a strong negative correlation between W_{cad}^{cad} and its promotor threshold suggesting that the level of auto-activation or auto-repression is clearly linked to the threshold. Another interesting type of correlations is the one between Tll promoter threshold with some of the regulators (W_{tll}^{hb} vs. h_{tll} and W_{tll}^{cad} vs. h_{tll}). Therefore, one cannot conclude that it is a strong or weak action just by focusing on the weight of the parameter given that the level of production depends on the threshold [6].



(a) W_{cad}^{cad} /h_{cad} (b) W_{tll}^{hb} /h_{tll}



(c) W_{tll}^{cad} /h_{tll}

Figure 6.10: Scatter plots of parameters that regulate caudal. Only the scatter plots with pairwise correlations higher than 0.6 are shown.

6.3 Data vs. Model

It is not trivial to identify the principal reason of the strong correlation. The poor determinability might be caused by an incomplete model, or by insufficient data or noisy data. We enquire if the complex correlation is caused by the insufficient data or the noise present in the data. To explore our hypothesis, a gap-gene circuit is inferred from real data using ES followed by DS. This circuit, named gn62h35_Target, has a final RMS = 10.56 and exhibits a correct pattern without any visible defect. Firstly, "artificial simulated data" are generated by solving circuit gn62h35_Target. These data are used as target data for an inverse modelling problem. Twenty series $(40, 200)$-ES with 600000 generations

followed by 40000 iterations of Downhill simplex were performed. Sixteen of the simulations have a RMS $\leqslant 5.00$ with patterns very close to the artificial target. The RMS of the best solution is 0.73, this solution is shown in Fig. 6.11. Figs. 6.13 and 6.14 give scatter plots of all model parameters of five of these reversed-solutions (red circle) vs the target model parameters (black star).

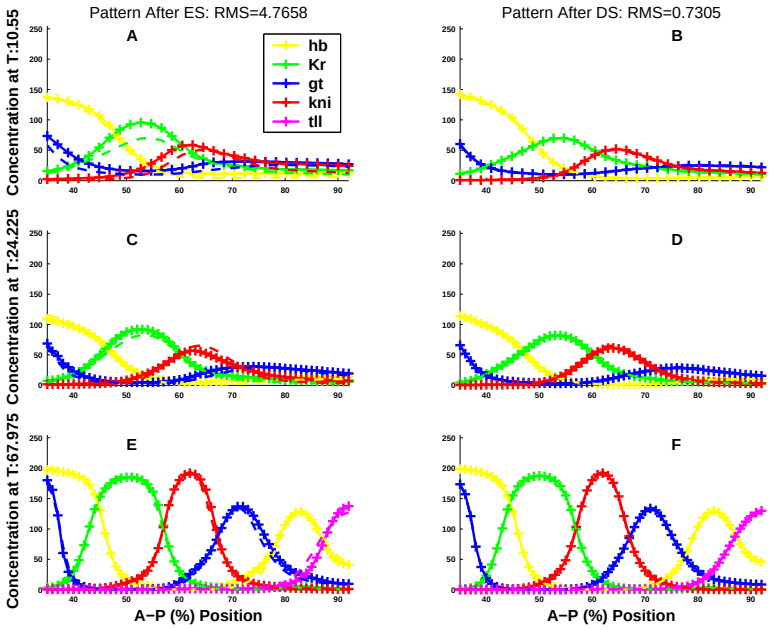


Figure 6.11: Illustration of the expression patterns obtained by reverse engineering of gap-gene circuit gn62h35_Target using (μ, λ)-ES followed by Downhill simplex search. The artificial target data used for the parameter estimation is obtained by simulating the model with the parameters obtained from the gn62h35_Target data. Left plots correspond to the simulation obtained after 60000 iterations of a $(40, 200)$-ES at time points: T = 10.550, $T1$ = 24.225 and $T8$ = 67.975 and the right plots are the final patterns obtained after 40000 iterations of Downhill simplex. The dashed lines are the artificial target data and the full-crossed lines correspond to the simulated pattern. The y-axis gives the relative protein concentration expression level and the x-axis corresponds to the anterior-posterior (A-P) axis of the embryo. The final RMS is 0.73. Optimization took approximatively 10 hours CPU time on a single 3.4-GHz processor.

The second test consist in inferring a GRN from noisy artificial data. We use gap-gene circuit gn62h35_Target as target data. We add to all data points a percentual perturbation, drawn from a uniform distribution between $[-RN, +RN]$. Simulations with 7 different RN = $[1, 2, 3, 4, 5, 10, 25]$ were performed and in each case, 2 simulations were run. The resulting RMS lie between 2.0 and 20.0. The expression patterns of circuits obtained from noisy data with random perturbations up to 10% are very similar to the original reversed one or to the

real gene expression data. Figs. 6.12 illustrates the target data (corresponding to gn62h35_Target) with a random perturbation $RN = 4$ versus the simulated data after re-inverse modelling. The final RMS = 5.05 after ESDS. For the case where we add a large random perturbation (25%) the best gap-gene circuits gave a RMS = 11.54, and, as to be expected, the simulated pattern is closer to the real data than to the noisy data.

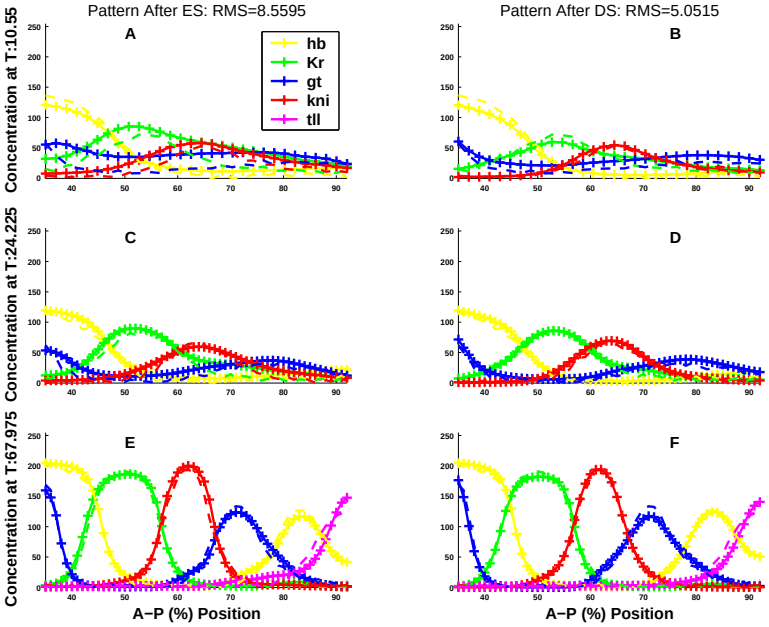


Figure 6.12: Illustration of the solution of a reverse engineering of gap-gene circuit gn62h35_Target with noise added, using (μ, λ)-ES followed by Downhill simplex search. The perturbed data are obtained by adding random noise between -4% and +4%. Left plots correspond to the simulation obtained after 10000 iterations of a $(100, 500)$-ES at time points: $T = 10.550$, $T1 = 24.225$ and $T8 = 67.975$ and the right plots are the final patterns obtained after a consequently 40000 iterations of Downhill simplex. The broken lines are the artificial target data and the full-crossed lines correspond to the simulated pattern. The y-axis gives the relative protein concentration expression level and the x-axis corresponds to the anterior-posterior (A-P) axis of the embryo. The final RMS is 5.05. Optimization took approximatively 10 hours CPU time on a single 3.4-GHz processor.

Figs. 6.13 and 6.14 give scatter plots of model parameters of seven of these perturbed reversed-circuits (blue diamond) vs the target model parameters (black star). Like in the real data fit, we see that non weight parameters (R, h, D, and λ) have a relative broader scattering in comparison to the regulatory interactions. The weight parameters are very preciely recovered with an accuracy of 10^{-3} in most of the cases, with exceptions. Comparing parameters recover from

noisy or non-noisy data, we observe that the data source seems not to matter in the regulatory parameters case, contrarily to the non-weigh. In that case, we see that for some parameters such as $h_{tll}, R_{cad}, R_{hb}, R_{Kr}$ and eventually the protein half life, the parameters represented with red circles (no noise) and those illustrated as blue diamonds (noisy data) do not have the same scattering or distribution, and fall in to different groups. We could not find a reasonable explanation for that, but we believe that all the non regulatory parameters R, D, h, λ are not identifiable. From the correlation matrix shown in Fig. 6.13, we see that the parameters are still strongly correlated. Nevertheless, from the individual pairwise correlation analysis, we observe that in very few cases, the pearson correlation value was reduced.

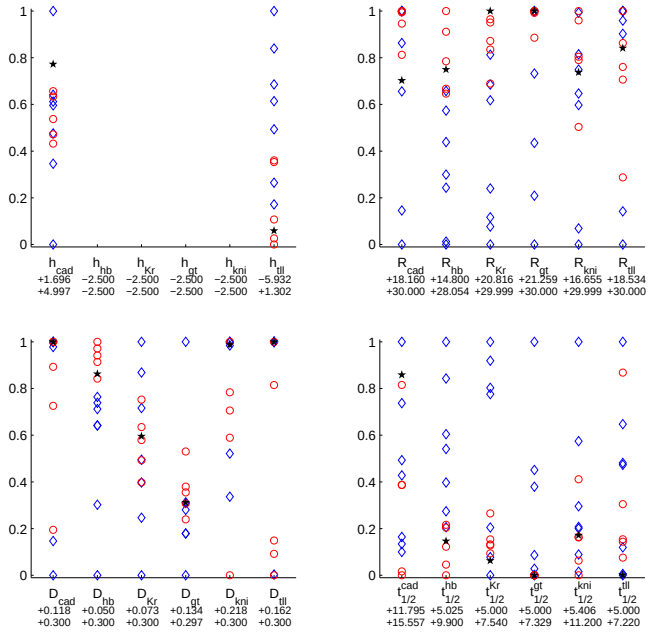


Figure 6.13: Scatter plot of parameters h, R, D and $t^a_{1/2} = \ln(2)/\lambda_a$. Each vertical line contains 13 different values for the same parameter. 5 simulations correspond to simple reversed-engineering (red circle) and 7 (blue diamond) to perturbed reversed-engineering with varying additive noise (1–5, 10 and 25%). The black star represents the target parameter. The parameter values are scaled to [0,1] using the minimum and maximum values given below the x-axis. In all simulations, including the target, the promoter thresholds $h_{(hb,Kr,gt,kni)}$ are fixed to -2.5 during the optimization.

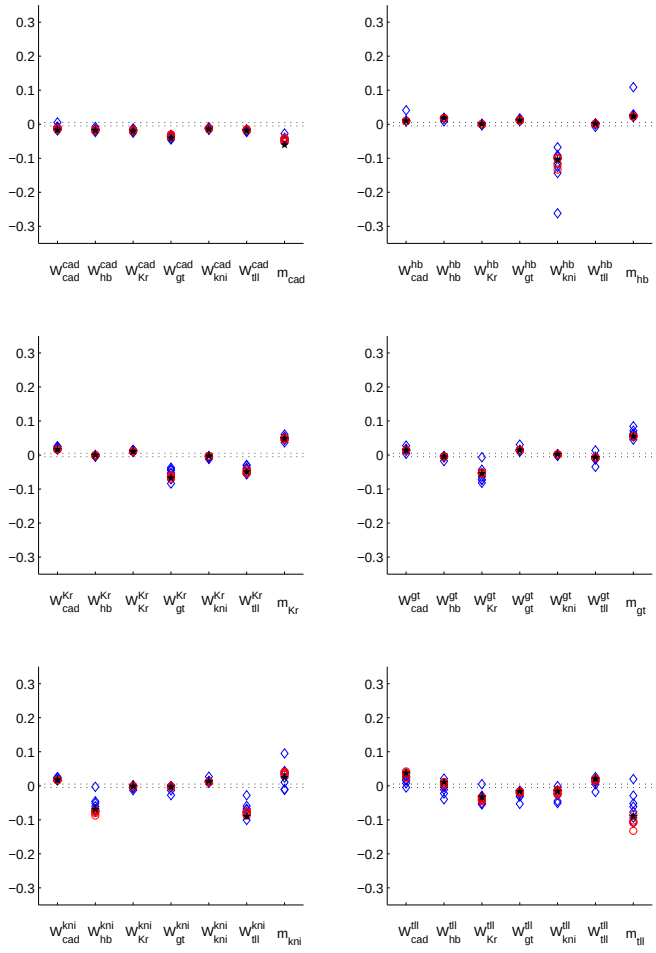


Figure 6.14: Scatter plot of the regulatory influences. W-entries and m-values above the horizontal line at 0.005 are considered to be activating, below -0.005 inhibiting. Each vertical line contains 13 different values for the same parameter. 5 simulations correspond to simple reversed-engineering (red circle) and 7 (blue diamond) to perturbed reversed-engineering with different additive noise. The black star represents the target parameter. In all simulations, including the target, the promoter thresholds $h_{(hb,Kr,gt,kni)}$ are fixed to -2.5 during the optimization.

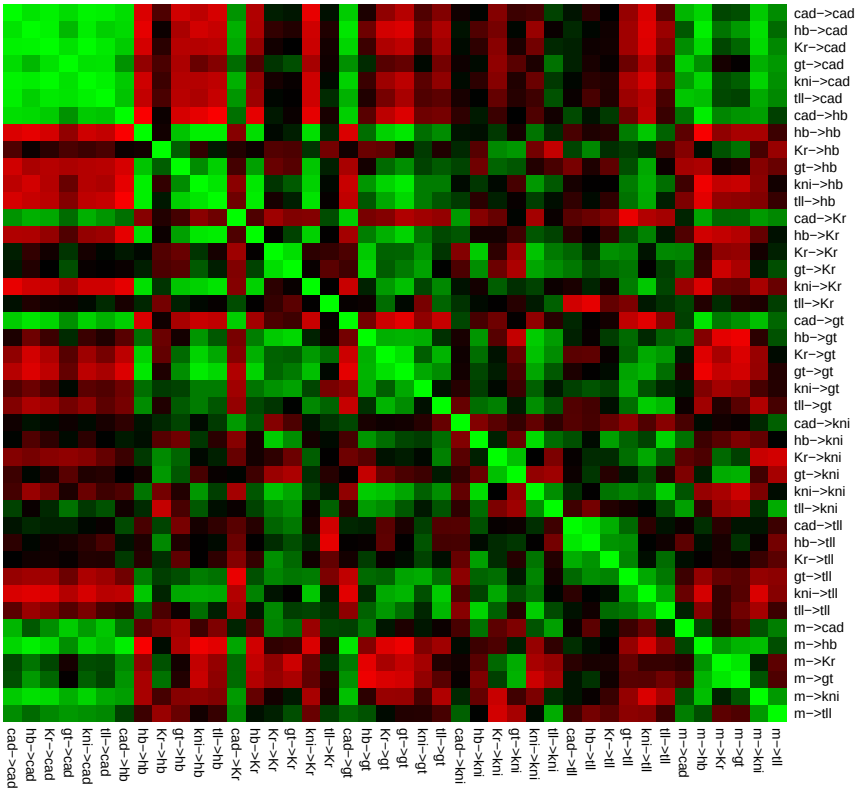


Figure 6.15: Correlation matrix of the parameters obtained from synthetic data. The matrix shows the pairwise correlation; the colour scale goes from intensive red (strong negative correlation) to bright green (positive correlation). The correlation matrix shows that there exist many pair wise correlations that tend to form clusters.

6.4 Conclusions

This chapter has addressed the question of the reliability of the parameter estimates. We have shown that some of the parameter estimates are quite precise and consistent while few of them are largely scattered. Solely based on the parameter values, we see that it is not possible to derive an unique network and one needs to examine the reliability of the parameters. We have used a very simple approach based on correlation analysis to determine the parameter identifiability. The correlation matrix obtained from the 101 gap gene circuits shown in Fig. 6.4 suggests that the parameters are strongly correlated. Dealing with the intricateness of the correlation patterns, from individual analyses on all the pairwise correlations, we derive the following explanatory rules:

1. direct correlations: (involving gene regulators)
   (a) negative correlations between a gene's activators when their contribution is partially on the same anterior-posterior (A-P) domain.
   (b) negative correlations between a gene's repressors when their contribution is partially on the same (A-P) domain.
   (c) positive correlation between a gene's activators vs. its repressor when their contribution is partially on the same (A-P) domain.
   (d) co-correlation caused by the domain geometry (boundary control mainly). Usually, it is regulatory interactions of two different parameters on a gene having the same function (activation or repression) but acting on non-overlapping (A-P) domain.
2. indirect correlation caused by the profile variation.

We believe that some of these correlations might reflect the biological reality such as compensation for redundant activity on a single gene or boundary determination. However, some of the correlations are much more difficult to relate to biology and might be caused by mathematical properties of the model or by indirect regulation through feedback loops mechanism. For instance, we see that parameters that are largely scattered or showing dual role (activation and repression) have very complex correlation. For instance, the negative correlation between activation of *hb* by Gt and repression of *hb* by Kni is only present when Gt activates *hb* (which is supposed to be a false positive).

It is possible to derive some of the qualitative interactions, mainly when the parameter estimates are very precise and do not show major correlations or unexplainable correlations with other parameters. Nevertheless, it is quite risky to define the precise value of these parameters or use the model for quantitative assertions. For instance, we see that quantitative value of some parameters strongly affect the correlation, even when the parameters are quite precise or very less scattered.

It is convenient to imagine that the poor determinability is caused by insufficient data. Our analysis on synthetic data (complete and/or noisy) demonstrated that although some parameters gained considerable precision or less scattering, the intricateness of the parameter correlations remains and we can not conclude that the poor determinability is mainly caused by missing data or hidden mechanism within the data.

It might be possible to reduce the correlation and improve the parameters determinability, by changing some models properties or reducing the number of parameter (for instance, by removing parameters that are known not to reflect any particular interaction). However, as suggested elsewhere [103], some of the correlations can not be reduced and 100 % determinability might not be possible in general, for models of biological systems.

The complexity of genetic network models (both in terms of number of parameters and non-linearities) is far from being matched by the available expermental data. It is thus highly desirable to detect poorly supported estimations, as well as to consider alternative sources of information to infer the parameters. In the next chapter, we will study the robustness of the circuits to parameter perturbation and fluctuation.