Centrum Wiskunde & Informatica

REPORT*RAPPORT*

MAS

Modelling, Analysis and Simulation

Modelling, Analysis and Simulation

Parameter estimation for a model of gap gene circuits with time-variable external inputs in Drosophila

M. Ashyraliyev

REPORT MAS-E0904 MAY 2009

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3703

Parameter estimation for a model of gap gene circuits with time-variable external inputs in Drosophila

ABSTRACT
We study a model for spatio-temporal pattern formation of gap gene products in the early development of Drosophila. In contrast to previous studies of gap gene circuits, our model incorporates a number of proteins as time-variable external inputs, including a protein Huckebein which is necessary for setting up the correct posterior domain boundary and its shift in time for the gap gene hunchback. Unknown model parameters are inferred by fitting the model outputs to the gap gene data and statistical analysis is applied to investigate the quality of the parameter estimates. Our results, while being consistent with previous findings, at the same time provide a number of improvements. Firstly, it takes into account correct regulation of hunchback at the posterior part of the embryo. Secondly, confidence interval analysis shows that the regulatory topology of the gene network in our model which consists of parameters representing the regulation between genes is more consistent with the experimental evidences. Our results also reveal that for data fitting the Weighted Least Squares sum is a more suitable measure than the Ordinary Least Squares sum which has been used in all previous studies. This is confirmed by a better fit of the boundaries of the gap gene expression domains and an absence of patterning defects in the model outputs, as well as by a correct prediction of mutant phenotypes.

2000 Mathematics Subject Classification: 92C15
Keywords and Phrases: parameter estimation; parameter determinability; gap gene circuits; Drosophila
Note: This work was supported from NWO's `Computational Life Science' program, projectnr. 635.100.010.

Parameter estimation for a Model of Gap Gene Circuits with Time-Variable External Inputs in *Drosophila* *

M. Ashyraliyev
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

We study a model for spatio-temporal pattern formation of gap gene products in the early development of *Drosophila*. In contrast to previous studies of gap gene circuits, our model incorporates a number of proteins as time-variable external inputs, including a protein Huckebein which is necessary for setting up the correct posterior domain boundary and its shift in time for the gap gene *hunchback*. Unknown model parameters are inferred by fitting the model outputs to the gap gene data and statistical analysis is applied to investigate the quality of the parameter estimates.

Our results, while being consistent with previous findings, at the same time provide a number of improvements. Firstly, it takes into account correct regulation of *hunchback* at the posterior part of the embryo. Secondly, confidence interval analysis shows that the regulatory topology of the gene network in our model which consists of parameters representing the regulation between genes is more consistent with the experimental evidences.

Our results also reveal that for data fitting the Weighted Least Squares sum is a more suitable measure than the Ordinary Least Squares sum which has been used in all previous studies. This is confirmed by a better fit of the boundaries of the gap gene expression domains and an absence of patterning defects in the model outputs, as well as by a correct prediction of mutant phenotypes.

1 Introduction

Gap genes constitute the first step in a regulatory cascade that leads to the determination of body segment positions along the major (or anterior-posterior, A–P) body axis during early *Drosophila* development [1]. They are involved in the regulation of pair-rule and segment-polarity genes, the latter of which establish a segmental pre-pattern of gene expression by the onset of gastrulation.

The gap gene system in the early *Drosophila melanogaster* is a well studied developmental gene network (see [2] and references therein). Initially the system is set up by spatial gradients of maternal proteins Bicoid (Bcd), Hunchback (Hb), and Caudal (Cad). Zygotic gap genes, such as *hunchback* (*hb*), *Krüppel* (*Kr*), *knirps* (*kni*), and *giant* (*gt*), are regulated by these maternal gradients, which establishes their expression in broad, overlapping regions

*The material presented in this note is used for a joint paper with J. Jaeger (EMBL/CRG, Barcelona).

of the embryo. These spatial domains of gap gene expression are stabilized and refined by gap-gap cross-repression and regulation by zygotic terminal gap genes *tailless* (*tll*) and *huckebein* (*hkb*).

The gap gene system has been studied extensively using a model of genetic regulatory networks described by a system of reaction-diffusion equations [2]. Quantitative expression data available for all relevant maternal coordinate and gap genes [3, 4] (except for *hkb*) have been used to infer regulatory interactions between gap genes using different global and local optimization strategies [2, 5, 6, 7, 8]. The gap gene system has been modeled by a 6-gene network, including *hb*, *Kr*, *kni*, *gt*, *tll*, and *caudal* (*cad*). The maternal protein Bcd has been incorporated as an external input constant in time. Although the obtained results have given significant insight into the underlying mechanism of the gap gene system, further investigation is needed for some important issues.

Results for the 6-gene model revealed a major patterning defect for the expression of gap gene *hb*. The posterior boundary of the posterior *hb* domain was not established correctly. Moreover, anterior shift of this boundary as well as the shift of the domain peak found in data was not reproduced by model outputs. This was explained by the absence of the terminal gap gene *hkb* in the 6-gene model. Huckebein (Hkb) is the main repressor of *hb* in that region of the embryo [9]. The missing *hkb* gene was also predicted to have an influence on the regulatory topology inferred from data. The model wrongly predicted that *hb* is not regulated by Tailless (Tll), while it is known that *hb* is activated by Tll [9]. This contradiction was explained by the ambiguous role of Tll in the regulation of *hb*. On the one hand, in the absence of the repressor Hkb, Tll has to take over its repressing function. On the other hand, Tll is an activator of *hb*. This dual role yields a cancellation effect and the model predicts that *hb* is not regulated by Tll.

Results of a parameter determinability analysis for the 6-gene model in [8] show that the parameter estimates corresponding to the regulation of *cad* and *tll* by gap genes are highly unreliable. The observed uncertainty was explained by the fact that the products of maternal genes (such as *cad*) and terminal gap genes (such as *tll*) regulate gap genes, but not vice versa. Despite the reasonable fit obtained for the expression of *cad* and *tll* in the 6-gene model, the unrealistic assumption that their dynamics is prescribed by the regulation by gap genes increases the level of uncertainty in the gap gene model. Due to the correlations between parameters in the model, this influences the determinability of other, biologically relevant, regulatory weights.

In this work we consider a reduced 4-gene model, including only gap genes *hb*, *Kr*, *gt*, and *kni*. In contrast to the 6-gene model, we now incorporate *cad* and *tll* as time-variable external inputs. Thereby, in our model the expression of *cad* and *tll* is obtained directly from data rather than being computed as state variables. A second important change is that data for gene *hkb* have become available [10]. Similar to *tll*, the terminal gap gene *hkb* is not regulated by other gap genes and therefore, it is also included in the model as time-variable external input. Finally, we incorporate the maternal gradient Bcd in our new model as external input, similar to previous studies. However, the data suggest that the protein Bcd varies with time rather than being constant. Therefore, contrary to the 6-gene model, we allow Bcd to be time-variable input.

Thus, we replace the previously studied 6-gene network by a more realistic, reduced 4-gene network with four external time-variable inputs. This significantly decreases the size of the problem, both with regard to the number of equations in the model to be solved and the number of unknown parameters to be estimated. We will infer the regulatory

topology and we will investigate the parameter determinability for the reduced model. Most importantly, we will show that despite the simplifications we made, the reduced model not only gives comparable results as the 6-gene model but also overcomes the above mentioned shortcomings. Note that the reduced gap gene network has also been recently investigated in [11], but in that study Bcd has been kept constant in time and *hkb* has not been used in the model.

Inference of the parameters is done by fitting model outputs to experimental data, i.e., by minimizing a cost function which measures the difference between them. The choice of the cost function is important for obtaining unbiased estimates and the computation of statistical quantities for parameter estimates (such as confidence intervals and correlations coefficients). It greatly depends on the nature of errors in the data. In all previous works [2, 5, 6, 7, 8, 11], the Ordinary Least Squares (OLS) measure has been used for the optimization and the statistical analysis. It is well known that OLS is suitable if the measurement errors are independent of each other and normally distributed with a constant standard deviation. However, the data for the gap genes suggest that the level of noise in the measurements varies both in space and in time. In such case, the Maximum Likelihood Estimates (MLE) can be obtained only if the Weighted Least Squares (WLS) sum is used as a distance measure with the weights chosen to be inversely proportional to standard deviations [12]. Since the standard deviations are available from [4], there is no additional computational work needed when the WLS sum is minimized in comparison with the OLS case. In this work we will obtain parameter estimates and study the parameter determinability using both the OLS and the WLS measures, and we will provide a detailed comparison between both results. We will demonstrate that for the problem under consideration, WLS gives indeed a more suitable measure than OLS.

The note is organized as follows. In Section 2 we describe the necessary materials and methods that are used. In Section 3 we give the results of our simulations. We conclude this note with a discussion in Section 4. In the Appendix we include all additional plots.

2 Materials and Methods

2.1 Gap Gene Circuits

Segment determination occurs during the blastoderm stage of *Drosophila* development, between 1.5 and 3 hours after egg laying [13]. During this stage, the embryo consists of a syncytium: there are no cell membranes between the nuclei. These nuclei constitute the basic objects of the model. They are arranged in a row along the A–P axis. Nuclei divide rapidly and synchronously [14]. Periods between mitotic divisions are called cleavage cycles, where cycle n occurs between mitoses $n-1$ and n. The models considered here run from early cycle 13 ($t = 0.0$ min) to the onset of gastrulation at the end of cycle 14A ($t = 71.1$ min). Mitosis occurs at the end of cycle 13, between $t = 16.0$ min and $t = 21.1$ min [14].

Gene circuit models describe the change in concentrations of each gap gene product in each nucleus over time by the following system of ODEs

$$\frac{dg_i^a}{dt} = R_a \Phi \left(\sum_{b=1}^{N_g} W_a^b g_i^b + \sum_{e=1}^{N_e} E_a^e g_i^e + h_a \right) - \lambda_a g_i^a + D_a \left(g_{i+1}^a - 2g_i^a + g_{i-1}^a \right), \qquad (2.1)$$

where a and b refer to regulated gap genes and regulators, respectively, and e refers to

external regulators. Here, a and b are integer indices representing *hb*, *Kr*, *kni*, and *gt*; e is an integer representing the regulators Bcd, Cad, Tll, and Hkb. The independent variable g_i^a denotes the concentration of the product of gene a in nucleus i; the input variable g_i^e denotes the concentration of the external protein e in nucleus i. $N_g = 4$ is the number of gap genes and $N_e = 4$ is the number of external proteins in the model. The function

$$\Phi(x) = \frac{1}{2}\left(\frac{x}{\sqrt{x^2+1}} + 1\right) \tag{2.2}$$

is a sigmoid regulation-expression function. The first term in the right hand side of (2.1) models the protein synthesis, while the second and third terms correspond to protein decay and protein diffusion, respectively.

During mitosis, protein synthesis is shut down. Nuclei divide instantaneously at the end of mitosis and the protein concentrations from each mother nucleus are copied to its daughter nuclei. The distance between nuclei is halved which is implemented in the model by reducing the diffusion coefficients D_a by the factor of 4. Gap gene circuits cover the region from 35% to 92% of the A–P axis, which includes $N_c = 30$ and $N_c = 58$ nuclei at cycles 13 and 14A, respectively. Therefore, system (2.1) consists of 120 and 232 ODEs during cycles 13 and 14A, respectively. At the boundary points $i = 1$ and $i = N_c$ we replace the diffusion term in right hand side of (2.1) by $D_a(g_{i+1}^a - g_i^a)$ and $D_a(g_{i-1}^a - g_i^a)$, respectively. This way we mimic the homogeneous Neumann (no flux) boundary conditions.

Gap genes *Kr*, *kni*, and *gt* are not expressed in the embryo before cycle 13. Therefore, zero initial conditions are taken for these. The initial condition for *hb* is prescribed by the maternal gradient of Hb shown in Figure 2.1. It is obtained by averaging the measurements from 18 individual embryos at cycle 12 ($t = -6.2$) and then using linear interpolation between this averaged pattern and *hb* data at cycle 13 ($t = 10.55$). Measurements for *hb* from individual embryos at cycle 12 (data without background) and the averaged *hb* pattern at cycle 13 are all available from [4].

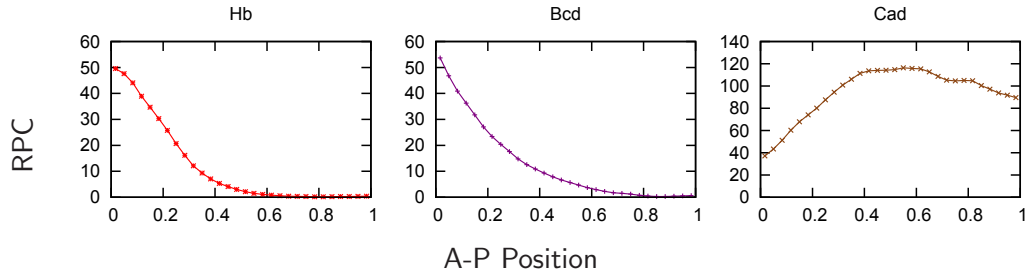

Figure 2.1: Quantitative gene expression for Hb, Bcd, and Cad at $t = 0$. Lines show the relative protein concentration (RPC) plotted against the position on the A–P axis (the trunk region of the embryo, from 35% to 92% A–P position is scaled to relative co-ordinates $[0, 1]$).

In system (2.1) there are $m = 48$ unknown parameters. These include the genetic interconnection or regulatory weight matrices W and E of size $N_g \times N_g$ and $N_g \times N_e$, respectively. The matrix elements W_a^b and E_a^e represent the regulation of gap gene a by gene b and gene e, respectively. Regulatory parameters represent repression (if < 0), activation (if > 0) or no interaction (if ≈ 0). The other parameters are promoter thresholds h_a, promoter strengths R_a, diffusion coefficients D_a, and decay rates λ_a.

Data The data set used for model fitting consists of $N = 1976$ measurements of protein concentrations (available from [4]). Measurements were taken at one time point during cycle 13 (T_0), and eight time points T_i ($1 \leq i \leq 8$) during cycle 14A (Figure 2.2). Measurements for the concentrations of all gap gene products represented in the model at all time points are available. Each data point represents concentration values which have been averaged over the bin (volume) from the measurements taken in individual embryos [3]. The number of embryos varies from 9 to 62 for different genes and different time points (with exception for *kni* at T_0 where only measurements from 4 embryos are available). Since from each embryo a few values per bin are available, the number of individual measurements used in the computation of the averaged value (sample mean) is much larger than the number of embryos. Using the Central Limit Theorem (CLT) we may assume that the experimental errors are approximately normally distributed [16]. Figure 2.3 shows the gap gene data at all time points (solid lines) and the standard deviations of the experimental errors (shaded areas).

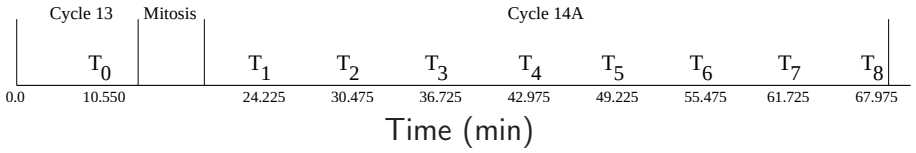

Figure 2.2: Time axis and the points when measurements were taken: one in cycle 13 and eight in cycle 14A; the duration of mitosis is also indicated.

To our knowledge, the presence of any hidden dependencies in the available dataset has not been investigated in literature. Although the measurements from different embryos are most likely to be uncorrelated (assuming that there was no systematic error in experiments), it is not known whether the gene expression data in the same embryo are correlated. In this work, we assume that the experimental errors are independent of each other.

External Inputs To solve (2.1), one needs the level of gene expression for external inputs at all $t \in [0, T]$, where $T = 71.1$. Measurements for Bcd, Cad, and Tll at all time points T_i ($0 \leq i \leq 8$) are available from [4] except for Bcd and Cad at T_7 and T_8. We obtain the patterns for Cad at those missing time points by integrating measurements from individual embryos (from [4]), 13 at T_7 and 12 at T_8. A similar procedure for Bcd however leads to an artificially high level of gene expression for Bcd at T_7 and T_8 and therefore they are not used here. Data for Hkb at all time points T_i are obtained from [10]. Figure 2.4 shows the relative protein concentration of external genes at all time points.

Genes Tll and Hkb are not expressed before cycle 13 and therefore we use a zero level for them at $t = 0$. Bcd and Cad at $t = 0$ have initial maternal gradients shown in Figure 2.1. We obtain them in the same way as the initial data for *hb*, i.e., by averaging the data from individual embryos at cycle 12 and then using linear interpolation between the patterns at cycles 12 and 13.

Now, the values of the external genes at any $t \in [0, T_8]$ can be linearly interpolated from the data at $t = 0, T_0, T_1, \ldots T_8$. The expression of Bcd for $t > T_6$ is linearly extrapolated from the values at T_5 and T_6, while the expression of other external inputs for $t > T_8$ is linearly extrapolated from corresponding values at T_7 and T_8. If the extrapolated value is

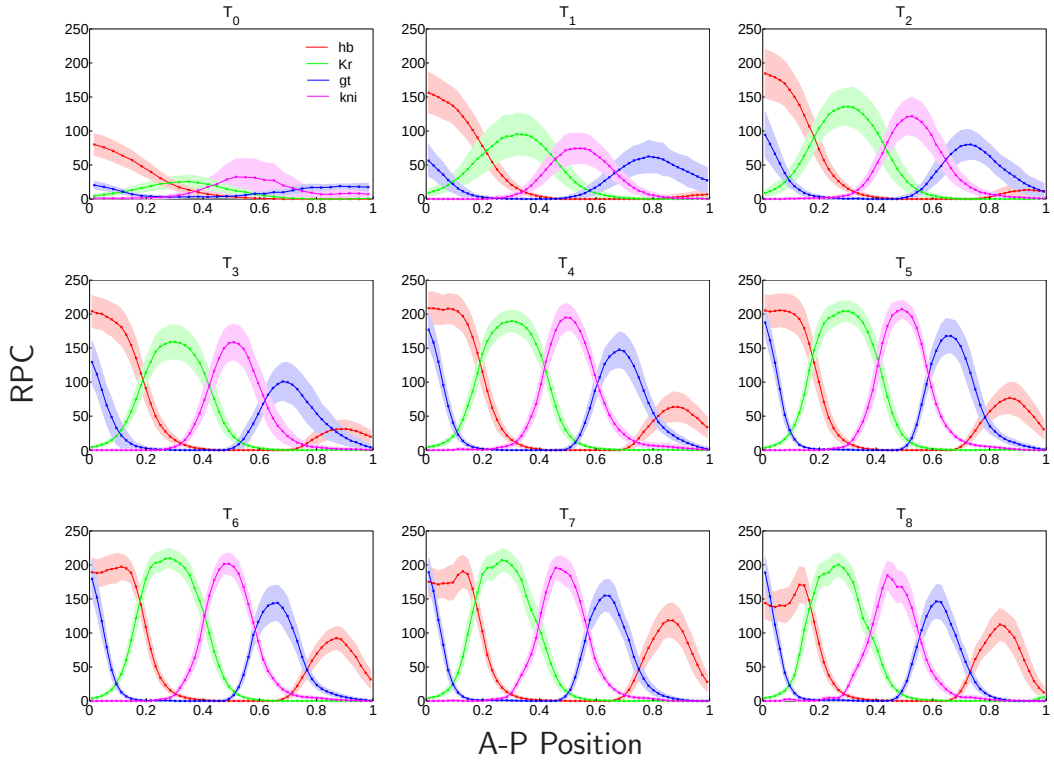

Figure 2.3: Quantitative gap gene expression data (solid lines) at the different time points. The shaded areas give the range of one standard deviation of the experimental error. Axes are as in Figure 2.1.

negative then we replace it with zero. Finally, we note that higher order interpolations give rise to artifacts from experimental noise [11] and therefore they are not used here.

Parameter inference We denote each measurement by $g_i^a(T_j)_{data}$, specified by the time T_j when the concentration of the gene product a in nucleus i is measured. The corresponding model value obtained from (2.1) is denoted by $g_i^a(T_j)_{model}$. The estimation of unknown parameters in (2.1) amounts to minimizing the cost function

$$CF = \sum_{a=1}^{N_g} \sum_{i=1}^{N_c} \sum_{j=0}^{N_t} v_{ij}^a \left(g_i^a(T_j)_{model} - g_i^a(T_j)_{data} \right)^2 , \tag{2.3}$$

where v_{ij}^a are positive weights, $N_g = 4$ is the number of gap genes, N_c is the number of nuclei (30 and 58 during cycles 13 and 14A, respectively), and $N_t = 8$ is the number of time classes. When all weights are equal to one, (2.3) is the OLS sum. Note that previously in the studies of gap gene circuits, only OLS is used as cost function to minimize. The quality

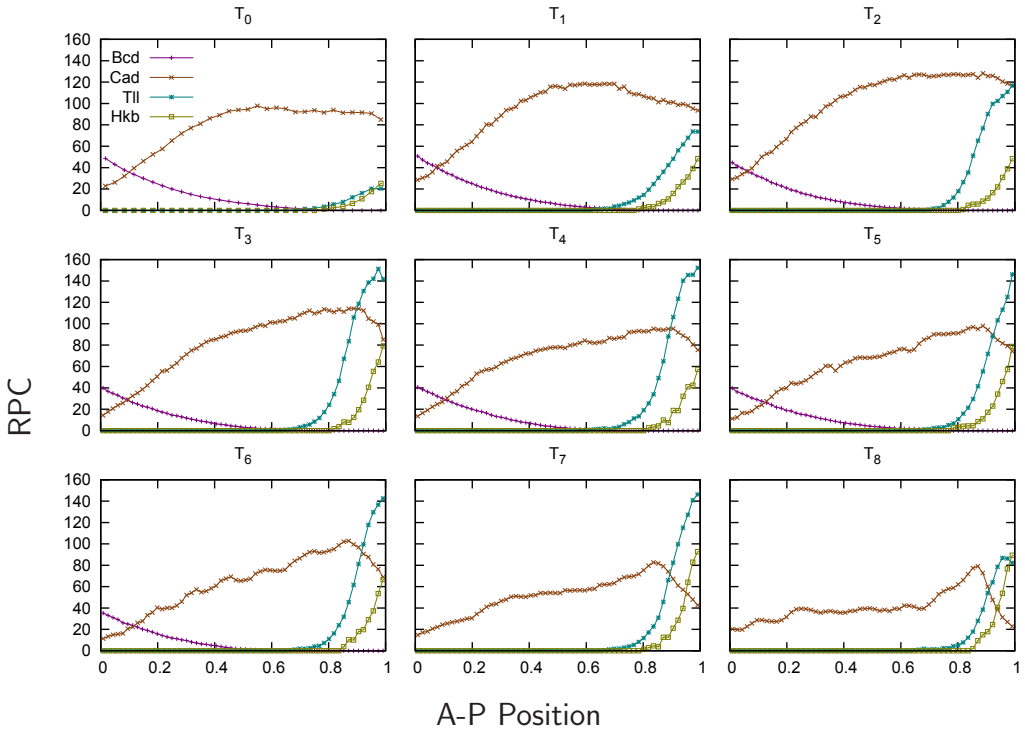

Figure 2.4: Quantitative data for external inputs at different time points. Note that Bcd at T_7 and T_8 is not available. Axes are as in Figure 2.1.

of the fit of the model to the data is measured by the root mean square (RMS) given by

$$RMS = \sqrt{\frac{1}{N}\sum_{a=1}^{N_g}\sum_{i=1}^{N_c}\sum_{j=0}^{N_t}\left(g_i^a(T_j)_{model} - g_i^a(T_j)_{data}\right)^2}, \tag{2.4}$$

where $N = 1976$ is the total number of all measurements. A solution is considered to be 'good' if $RMS < 12.0$ and if there are no visible pattern defects in the model response [2, 5, 6, 7].

We note that OLS is an appropriate measure under certain assumptions only. Namely, all measurement errors have to be independent of each other and be from a normal distribution with zero mean and constant standard deviation. The latter does not hold for our dataset [4]. The shaded areas in Figure 2.3 show how the standard deviation varies per gene and both in space and time. Note that the standard deviation (the level of noise) becomes smaller at late time points. Also important is that the standard deviation at the domain boundaries is relatively small and the level of noise in the non-expressing regions is almost negligible indicating that the stripe locations at the end of cycle 14A are determined with little variation [15].

When the weights v_{ij}^a in (2.3) are taken inversely proportional to the corresponding standard deviations, the cost function becomes the WLS distance. We emphasize here that

this is a theoretically more justified measure than the OLS measure due to the variation in the experimental errors. Since the standard deviations are available in [4], minimization of the WLS sum has no additional computational expenses compared to the corresponding procedure for the OLS sum.

In this note we use both the OLS sum and the WLS sum as the cost function to minimize. Our aim is to demonstrate that WLS is a more suitable measure than OLS not only in theory but also in practice. Throughout this note we will use the notations *OLS search* and *WLS search* meaning that the OLS and WLS sums, respectively, are minimized. Similarly, *OLS results* and *WLS results* indicate the parameter estimates obtained by minimizing OLS and WLS sums, respectively.

For practical reasons it is better to constrain the parameter space, especially for the global search optimization methods. Similar to previous studies of the gap gene system [2, 5, 6, 7, 8], we define the search space for the parameters by the linear constraints

$$10.0 \leq R_a \leq 30.0, \quad 0.0 < D_a \leq 0.3, \quad 5.0 \leq \frac{ln(2)}{\lambda_a} \leq 20.0, \quad a = 1, \ldots, N_g, \tag{2.5}$$

and by the nonlinear constraints

$$\sum_{b=1}^{N_g} \left(W_a^b \; g_{max}^b\right)^2 + \sum_{e=1}^{N_e} (E_a^e \; g_{max}^e)^2 + (h_a)^2 \leq 10^4, \quad a = 1, \ldots, N_g, \tag{2.6}$$

where g_{max}^b and g_{max}^e are the maximum values in the data set for proteins b and e, respectively. Note that in [2, 5, 6] the threshold parameters h_a for genes Kr, Kni, gt, and hb are fixed to negative values representing a constitutively repressed state for the corresponding genes [2, 5]. In [8] it is shown that fixing promoter thresholds improves the parameter determinability in comparison to the case when they are estimated along with other parameters. Therefore, we take $h_a = -2.5, \; a = 1, \ldots, N_g$ in all simulations, which leaves us with 44 unknown parameters in (2.1) to be estimated.

Mutation analysis The regulation of gene b on gene a is studied experimentally in the following way, called mutation: gene b is knocked out in the embryo and from the change in the expression of gene a the possible type of regulation is deduced. If the expression of gene a decreases (increases), then it is assumed that b is activator (repressor). If the mutation does not affect the expression of gene a then it means that b does not regulate gene a. Experiments with double mutants (when two different genes are knocked out) are also widely used. Similarly, mutation can also be done by over-expressing a certain gene to study its effect on the expression of the other genes. Although the conclusions based on mutant analysis can be ambiguous in some cases, such as indirect influence, still this method is a commonly applied approach in genetics.

Once the regulatory weights in the gap gene model (2.1) are estimated based on wild type data, mutation analysis can be easily conducted in *silico* [17]. Namely, b mutants can be modelled simply by setting W_a^b (or E_a^b) for all gap genes a to zero and leaving all other parameter estimates unchanged. It is an important issue whether the model with parameter estimates found using only wild type data can predict correct mutant phenotypes. Although the quantitative mutant data is not available, qualitative behaviour for mutant phenotypes of gap gene products in *Drosophila* is well studied. For instance, the posterior *hb* domain fails to retract from the posterior pole of the embryo in *hkb* mutant embryos [9], indicating that

Hkb represses *hb*. The posterior *hb* domain is absent in *tll* mutant embryos [9], indicating that Tll activates *hb*.

2.2 Methods

We consider a model given by the system of ODEs of the general form:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}, \theta), & 0 < t \leq T, \\ \mathbf{y}(t, \theta) = \mathbf{y}_0, & t = 0. \end{cases} \tag{2.7}$$

Here the m-dimensional vector θ contains all unknown parameters, $\mathbf{y}$ is an n-dimensional state vector, and $\mathbf{f}$ is a given vector function, differentiable with respect to t, $\mathbf{y}$ and θ.

Let us assume that for fitting (2.7) there are N measurements available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the state vector $\mathbf{y}$ is measured. The corresponding model value obtained from (2.7) is denoted by $y_{c_i}(t_i, \theta)$. We denote the vector of weighted discrepancies between the theoretical values and the measured values by $\mathrm{Y}(\theta)$. Then the least squares estimate $\hat{\theta}$ of the parameters is the value of θ that minimizes the sum of squares

$$S(\theta) = \sum_{i=1}^{N} w_i^2 (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta), \tag{2.8}$$

where w_i are positive weights. If the measurement errors in $\tilde{y}_i$ are independent of each other, normally distributed with standard deviations σ_i, and the weights w_i are proportional to $1/\sigma_i$, then $\hat{\theta}$ is a maximum likelihood estimate [12].

Parameter Estimation

In general, model (2.7)—being nonlinear in θ—leads to a least squares problem (2.8) that has several minima, first because the problem has more than one solution, and second because the fitness function (2.8) can have several stationary points that do not correspond to the lowest value of the cost function (so-called local minima). *Local search methods*, like Levenberg-Marquardt (LM) [18], easily get trapped in one of the local minima rather than finding the global minimum. To explore the whole search space one needs *global search methods*, like Evolution Strategy (ES) or Simulated Annealing (SA). Unfortunately, these methods converge very slowly once near a minimum. In contrast, gradient-based methods are efficient optimizers [19] for nonlinear least-squares problems once a sufficiently good initial guess for the parameter values is available. Therefore, for large scale problems, such as a gap gene system, it is efficient to use a global search method followed by a local gradient-based technique. In this way, the chance of missing the global minimum is reduced and the determination of the minima is precise and fast.

In this paper we use the LM method for local optimization. For the initial parameter values we use the parameter estimates obtained by Johannes Jaeger (EMBL/CRG, Barcelona) with SA global search.

Levenberg-Marquardt Method In general, any gradient-based optimization procedure seeks a correction $\delta\theta$ for the parameter vector, such that $S(\theta + \delta\theta) \leq S(\theta)$ holds. The LM method [18] determines the correction as the solution of the equations

$$\left(J^T(\theta)J(\theta) + \lambda I_m\right)\delta\theta = -J^T(\theta)\mathrm{Y}(\theta), \tag{2.9}$$

where $\lambda \geq 0$ is a control parameter (see below), I_m is the identity matrix of size m and the Jacobian $J(\theta) = \frac{\partial \mathrm{Y}(\theta)}{\partial \theta}$ is the so-called 'sensitivity' matrix of size $N \times m$. The entry $J_{i,j}$ in $J(\theta)$ shows how sensitive the model response is at the i-th data point for a change in the j-th parameter. The entries of J can be found by solving the system of variational equations

$$\begin{cases} \dfrac{\partial}{\partial t}\dfrac{\partial \mathbf{y}}{\partial \theta_i} = \dfrac{\partial \mathbf{f}}{\partial \theta_i} + \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \theta_i}, & 0 < t \leq T, \\ \dfrac{\partial \mathbf{y}}{\partial \theta_i}(t, \theta) = 0, & t = 0, \end{cases} \tag{2.10}$$

where $i = 1, 2, \ldots, m$, coupled to (2.7).

The LM method can be seen as the combination of two gradient-based approaches: Gauss-Newton and steepest descent [19]. If $\lambda = 0$ in (2.9), it coincides with the Gauss-Newton method. However, when the matrix $J^T(\theta)J(\theta)$ is (almost) singular, to solve (2.9), λ has to be positive and for large λ the LM method approaches the steepest descent method. During the optimization λ is adapted such that the algorithm strives to exploit the fast convergence of the Gauss-Newton method whenever this is possible [18, 20].

In order to solve (2.9), the singular value decomposition (SVD) [21] of the matrix $J(\theta)$ can be used, i.e.

$$J(\theta) = U(\theta)\ \Sigma(\theta)\ V^T(\theta), \tag{2.11}$$

where $U(\theta)$ is an orthogonal matrix of size $N \times m$, such that $U^T(\theta)U(\theta) = I_m$, $V(\theta)$ is an orthogonal matrix of size $m \times m$, such that $V^T(\theta)V(\theta) = V(\theta)V^T(\theta) = I_m$, and $\Sigma(\theta)$ is a diagonal matrix of size $m \times m$ which contains all singular values in non-increasing order. Then the correction $\delta\theta$ can be found as

$$\delta\theta = -V(\theta)\ \left(\Sigma^2(\theta) + \lambda I_m\right)^{-1}\Sigma(\theta)\ U^T(\theta)\ \mathrm{Y}(\theta). \tag{2.12}$$

Numerical integration of (2.7) and (2.10) requires a fast and reliable ODE solver. Searching in the parameter space may lead to some values of θ such that the systems of ODEs become stiff. It is well known that for stiff ODE systems explicit schemes can give rise to numerical instability or, alternatively, extremely small time steps. Therefore, an implicit scheme is the best choice for time integration for stability reasons. In our simulations we use implicit multistep Backward Differentiation Formulas (BDF) [26]. For numerical and implementational aspects of this method we refer the reader to [26] and [8] and references therein.

Statistical Analysis of Parameter Estimates

Once the parameter vector $\hat{\theta}$ minimizing (2.8) is found, it is important to know how reliable the obtained estimate is. This is the subject of a posteriori identifiability analysis [22, 23, 24]. The ellipsoidal region around $\hat{\theta}$ in which the 'true' parameter vector θ^* lies with a certain probability $1 - \alpha$ is defined by

$$(\theta^* - \hat{\theta})^T\left(J^T(\hat{\theta})J(\hat{\theta})\right)(\theta^* - \hat{\theta}) \leq \frac{m}{N-m}S(\hat{\theta})F_\alpha(m, N-m), \tag{2.13}$$

where $F_\alpha(m, N-m)$ is the upper α part of Fisher's distribution with m and $N-m$ degrees of freedom. To remind the reader, here m and N are the number of parameters and measurements, respectively. From (2.13) one can derive dependent and independent confidence intervals for parameter estimates $\hat{\theta}_i$ $(i = 1, 2, \ldots, m)$. These are, respectively,

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq \frac{r_\sigma}{\sqrt{\left(V(\hat{\theta})\Sigma^2(\hat{\theta})V^T(\hat{\theta})\right)_{ii}}} \right\} \tag{2.14}$$

and

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq r_\sigma \sqrt{\left(V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta})\right)_{ii}} \right\}. \tag{2.15}$$

Here $V(\hat{\theta})$ and $\Sigma(\hat{\theta})$ are obtained from (2.11), $r_\sigma^2 = \frac{m}{N-m} S(\hat{\theta}) F_\alpha(m, N-m)$.

Clearly, small confidence intervals for $\hat{\theta}_i$ indicate that it is well-determined. However, in some cases considering only individual confidence intervals can be misleading. For instance, in the presence of strong correlations between parameters, the dependent confidence intervals underestimate the confidence region while the independent confidence intervals overestimate it. For this reason, in addition to confidence intervals, it is essential to compute correlations between parameters. The correlation coefficient between $\hat{\theta}_i$ and $\hat{\theta}_j$ is given by

$$\rho_{ij} = \frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}. \tag{2.16}$$

where $B(\hat{\theta}) = V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta})$. For detailed explanations of these statistical quantities and their derivations we refer the reader to [8] and references therein.

3 Results

We estimated all 44 unknown parameters of the gap gene circuit model (2.1), such that the state variables fit the given data (Figure 2.3), subject to the constraints (2.5)-(2.6). We applied statistical analysis for the final parameter sets to assess the quality of the parameter estimates. Both OLS and WLS were used as a cost function in the data fitting procedure and the statistical analysis. We present here both results and give a detailed comparison between them.

3.1 OLS results

3.1.1 Selection of OLS gene circuits

The search with the OLS cost function leads to 740 parameter sets. About 80% of them have good-scoring RMS values, i.e., $RMS < 12.0$, which is below the level of experimental errors. However, a closer look at the model outputs for good-scoring sets reveals that most of them have a common patterning defect. Figure 3.1 shows the patterns obtained with one of those parameter sets (with $RMS = 9.21$) for the expression of gap gene Kr at time points T_3 and T_8 (green lines) compared to data (red lines). The model outputs have an artificial Kr hump in the region where no expression is detected for this gene in the data.

This hump arises at the beginning of cycle 14A and remains there until the end of cycle 14A. It is noteworthy that the gap gene network topology, i.e., the signs of regulatory weights in (2.1), in the parameter sets possessing such a patterning defect is in contradiction with known theoretical evidence. In other words, despite the overall reasonable fit to the data, model (2.1) predicts wrong regulations between genes. For instance, in the parameter set for which the patterns in Figure 3.1 are shown, *hb* is repressed by Tll and activated by Hkb, while it is known that Tll activates *hb* and Hkb represses it. We have found that the inferred network topology in good-scoring parameter sets producing an artificial Kr hump has some other artifacts as well (not shown here). Therefore, we exclude those parameter sets.

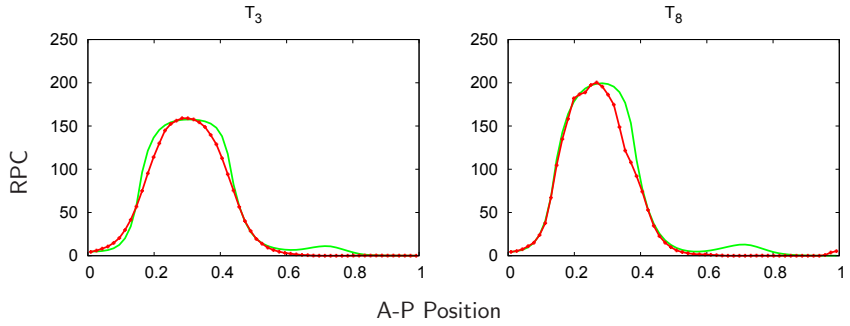

Figure 3.1: Comparison between data (red lines) and patterns obtained with a low-scoring parameter set yielded from the OLS search (green lines) for the expression of gap gene Kr at two different time points. Axes are as in Figure 2.1.

Although many good-scoring parameter sets obtained from the OLS search have the artificial Kr hump, there are still 39 parameter sets left which do not have that patterning defect. Their RMS values vary between 8.71 and 10.11. None of these parameter sets show any significant patterning defects (see Figure 4.1 in the Appendix). As we shall see, their network topology is in agreement with theoretical evidence. We consider only these 39 parameter sets in our analysis.

In conclusion, our selection of OLS parameter sets has been based on two criteria. Firstly, only parameter sets with low RMS values are taken into account. Secondly, only those sets which do not have the artificial Kr hump are manually selected. Importantly, both conditions are necessary and one does not imply the other. Many of the obtained low-scoring parameter sets give overall a reasonable fit but do possess the patterning defect for Kr. This underlines the main drawback of using the OLS measure. Extensive amounts of runs and additionally exhaustive manual work of inspection of patterns were needed in order to obtain the parameter sets which correctly describe the gap gene system.

3.1.2 Analysis of OLS gene circuits

Posterior *hb* domain Model outputs for the selected OLS parameter sets reveal the correct set up of the posterior boundary of the posterior *hb* domain by the end of cycle 14A (see Figure 4.1). Figure 3.2a shows the pattern generated with one of those parameter sets compared to the result obtained with the 6-gene gap system from [8]. Clearly, the result for the 4-gene model has a significantly improved fit of the posterior *hb* boundary. As we will

see, this is solely due to the inclusion of Hkb in the 4-gene model which is a main repressor of *hb* in that region.

Gap gene domains are established during cycle 13 and the beginning of cycle 14A. Afterwards, there is an anterior shift in the position of these domains. This shift mechanism has been investigated and well understood by using the 6-gene model [5]. It has been noticed that the domain shifts are based only on regulatory interactions between genes and diffusion plays no role in it. The model for the 6-gene network was able to reproduce most of the domain shifts observed in the data. However, for the posterior *hb* domain the shift of its peak and posterior boundary was not present in model outputs [5].

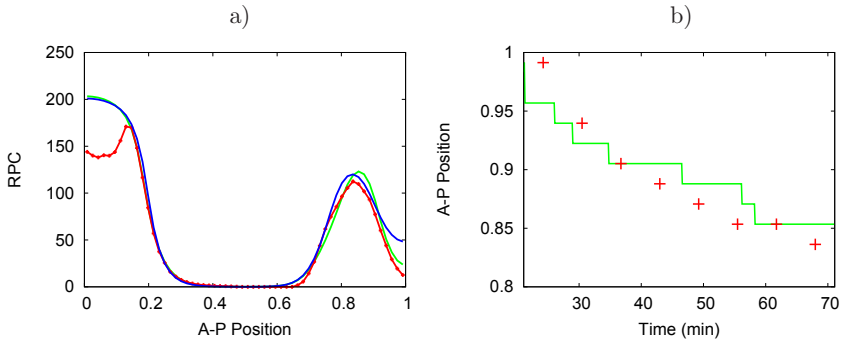

Figure 3.2: a) Comparison between data (red line), the pattern obtained by the parameter set for the 6-gene gap system from [8] (blue line), and the pattern obtained by one of the selected OLS parameter sets for our 4-gene network (green line) for the expression of gap gene *hb* at T_8. Axes are as in Figure 2.1. b) Anterior shift of the peak of the posterior *hb* domain during cycle 14A. Plot shows the A-P position of the peak in the model outputs for one of the selected OLS parameter sets (green line) compared to the corresponding shift observed in the data (+).

The model outputs with the selected OLS parameter sets for the 4-gene model show the shift in the peak of the posterior *hb* domain. We illustrate it in Figure 3.2b for one of the parameter sets, where the position of the peak on the A–P axis is plotted against time. Despite the slight difference with the corresponding shift in the data, the overall shift in the model output is visible.

Additionally, our results reveal the shift of the posterior boundary of the posterior *hb* domain. Similar to the approach in [5], we performed a graphical analysis of the *hb* regulation over time (cycle 14A) at three different nuclei which lie in the shift zone. Panels **a-c** of Figure 3.3 show a switch from protein synthesis (positive dg_i^{hb}/dt) to decay (negative dg_i^{hb}/dt) of *hb* at the end of cycle 14A. As we can see, diffusion plays no role in it. In fact, diffusion counteracts the boundary shift with an influx of protein into the region where *hb* decays. Note that a lack of smoothness in the protein synthesis term is a consequence of using linear interpolations for time-variable external inputs in the model. Panels **d-f** of Figure 3.3 reveal that the shift is solely driven by the temporal behaviour of the regulatory input for *hb* production (solid black lines). By plotting the individual contributions (coloured areas) we can analyse in detail the regulatory mechanism which underlines the shift. The areas below and above the black line represent the regulatory input from activators (Tll, Cad, Hb, and

Gt) and repressors (Kni and Hkb) of *hb*, respectively. Since the regulatory inputs from Kr and Bcd are negligible, they are not plotted here. As we can see, the activating contribution is mainly from Tll and less from Cad and autoactivation of *hb*. Note that insignificant activation by Gt is an artefact of the model. The repressing input from Kni is relatively small because of the low expression of *kni* in that region of the embryo. So, the shift is based on the regulatory input from Hkb, the main repressor of *hb*. This repression increases both in space (posteriorly) and in time.

In conclusions, our model predicts that *hb* in the posterior part of the embryo is mainly activated by Tll. However, this activation is suppressed by increasing repression of *hb* by Hkb yielding eventually the shift of the boundary domain. Contrary to the shifts of other boundaries of gap gene domains, this shift happens at late stages of cycle 14A.

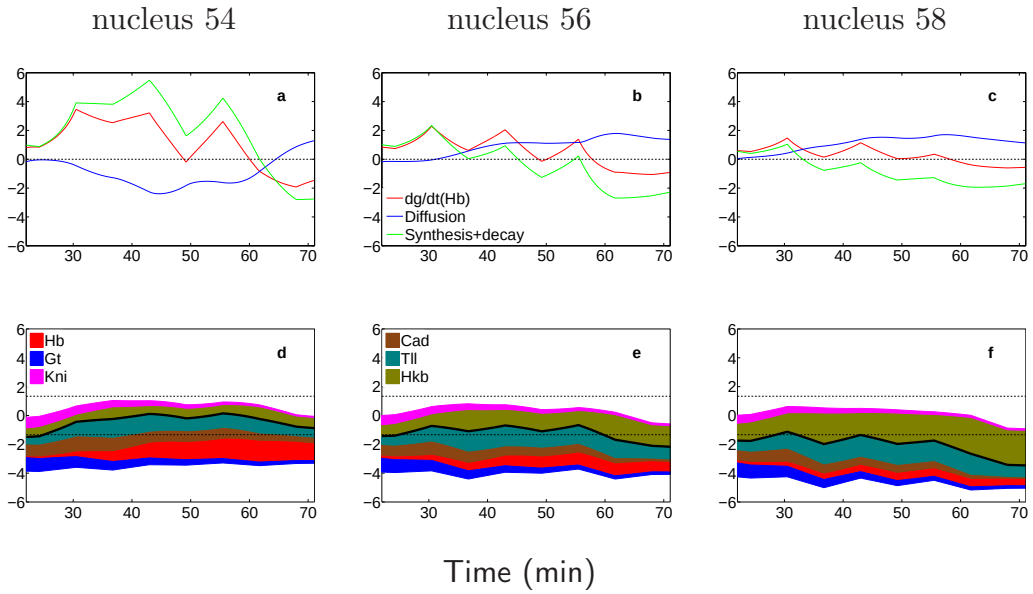

Figure 3.3: Graphical analysis of the shift of the posterior boundary of the posterior *hb* domain in the model outputs for one of the selected OLS parameter sets. Dynamic behaviour is illustrated at thee different nuclei within the shift zone over time (cycle 14A). Plots **a-c** show the rate of change in concentration of *hb* (dg_i^{hb}/dt), as well as individual contributions to it from diffusion and synthesis/decay terms in the right hand side of (2.1). Plots **d-f** show the temporal behaviour of the regulatory input for *hb* production (solid black lines), i.e., $u_i^{hb} = \sum_{b=1}^{N_g} W_{hb}^b g_i^b + \sum_{e=1}^{N_e} E_{hb}^e g_i^e + h_{hb}$. Upper and lower dashed lines indicate 90% and 10% of the maximum rate of protein synthesis, respectively. The sigmoid function (2.2) at those values is equal to 0.9 and 0.1, respectively. Coloured areas represent individual contributions to u_i^{hb} from repressors (above black lines) and activators (below black lines) of *hb*. The height of each coloured area is given by $|W_{hb}^b| g_i^b$ or $|E_{hb}^e| g_i^e$.

Network topology A classification of all estimates of the regulatory weights for all 39 parameter sets into 'activating', 'repressing' or 'no interaction' categories is shown in Figure 3.4.

The topology is mainly in agreement with the previous findings for the 6-gene model [8]. However, some ambiguities in the network are removed with these results. Namely, the repressive regulations of Hb on Kr and gt, Gt on kni, and Kni on Kr are present in all parameter sets, while previous results for the 6-gene case showed no regulation for these weights in many solutions. Importantly, the activation of hb by Tll is correctly predicted by our model in almost all sets. Note that previously it was found that there exists no regulation for this weight. Repression of gt by Tll is present in almost all parameter sets, while previously many circuits were found with no regulation for this weight. Another remarkable difference is that autoactivation of gt is much weaker than in the 6-gene case. To be more precise, its autoregulation is not required in most of the parameter sets. Finally, we note that the colours in Figure 3.4 do not change if we choose the threshold 0.01 instead of 0.005 for the classification of regulations, except in two regulatory weights. Specifically, the activation of hb by Cad and Tll changes to no regulation category, meaning that these activations in the network topology predicted by the model are weak (almost negligible).

	hb	*Kr*	*gt*	*kni*	*Bcd*	*Cad*	*Tll*	*Hkb*
hb	0/0/39	2/37/0	0/1/38	39/0/0	0/0/39	0/2/37	1/0/38	37/2/0
Kr	39/0/0	0/1/38	39/0/0	39/0/0	0/0/39	0/0/39	39/0/0	39/0/0
gt	39/0/0	39/0/0	0/35/4	0/0/39	0/0/39	0/0/39	38/1/0	2/2/35
kni	39/0/0	3/36/0	39/0/0	0/1/38	1/0/38	0/0/39	37/0/2	26/9/4

Figure 3.4: Gap gene network topology based on 39 selected OLS parameter sets. Each entry in the table corresponds to regulation of a gap gene given on a row by a gene given in a column. Triplets show the number of parameter sets in which a regulatory weight falls into one of the following categories: repression (values ≤ -0.005)/ no interaction (values between -0.005 and 0.005)/ activation (values ≥ 0.005). Colours: activation (green), no interaction (light-blue), repression (pink).

Confidence intervals The network topology shown in Figure 3.4 is based solely on the *values* of estimated parameters. To assess the *quality* of the parameter estimates, we computed dependent and independent confidence intervals for each parameter set using (2.14) and (2.15), respectively (see Figure 4.3 in the Appendix). We checked if the corresponding confidence intervals for regulatory weights fall entirely into the 'repression', 'no interaction', or 'activation' categories. Results in Figure 3.4 do not change when only dependent confidence intervals are taken into account. However, when including independent confidence intervals, one can no longer make similar qualitative conclusions about some entries in the regulatory weight matrix. For example, Figure 3.5 shows the confidence intervals for regulatory weights W_{Kr}^{gt} (a), E_{hb}^{Bcd} (b), and E_{kni}^{Tll} (c). The independent confidence intervals for W_{Kr}^{gt} lie in the negative part of the plane for almost all parameter estimates and therefore, repression predicted for this weight in Figure 3.4 is confirmed by statistical analysis. The independent confidence intervals for E_{hb}^{Bcd} slightly extend into the negative part of the plane. Therefore, one can make a qualitative conclusion for this weight: the model predicts that Bcd does not repress hb. Note that this is a weaker conclusion than predicting activation for this weight from Figure 3.4. In contrast, we cannot draw any qualitative conclusions about E_{kni}^{Tll}. Thus, statistical analysis does not confirm the repression of kni by Tll inferred from

Figure 3.4.

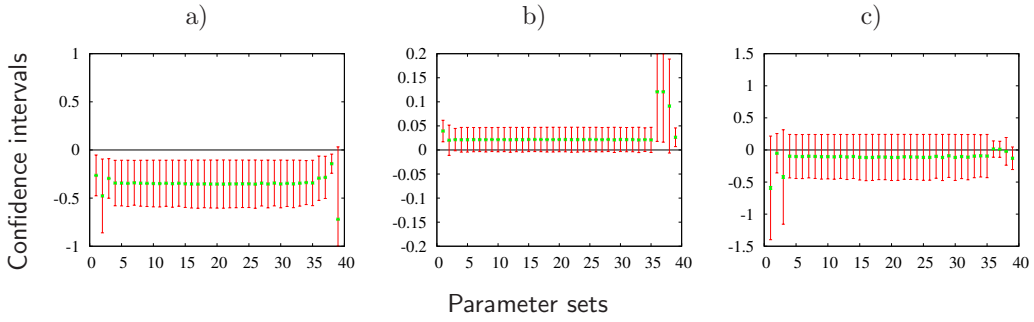

Figure 3.5: Dependent (green lines) and independent (red lines) confidence intervals for regulatory weights W_{Kr}^{gt} (a), E_{hb}^{Bcd} (b), and E_{kni}^{Tll} (c). Confidence intervals are plotted along the vertical axis for the 39 selected OLS parameter sets.

Based on the independent confidence intervals, Table 3.1 summarizes the qualitative conclusions for the regulatory weights in the gap gene model. These conclusions are weaker than those drawn from classifying the parameter values only. Only for 17 regulatory weights out of 32, the confidence intervals confirm the type of the regulation deduced from the network topology in Figure 3.4. For other 4 weights the conclusions in Figure 3.4 are confirmed weakly. Regulations for the remaining 11 weights cannot be qualitatively verified by the confidence interval analysis. However, the conclusions in Table 3.1 show qualitative improvement for a number of regulations in comparison with the corresponding results for the 6-gene gap model [8], where only for 9 regulatory weights the confidence intervals confirmed and for other 5 weights confirmed weakly the type of the regulation deduced from the corresponding network topology.

	hb	Kr	gt	kni	Bcd	Cad	Tll	Hkb
hb	$+$	$-=$	$\times$	$\times$	$+=$	$+$	$+=$	$\times$
Kr	$-$	$+$	$-$	$-$	$+$	$+$	$\times$	$\times$
gt	$-$	$\times$	$\times$	$\times$	$+$	$+$	$-$	$\times$
kni	$-$	$-=$	$-$	$+$	$+$	$+$	$\times$	$\times$

Table 3.1: Gap gene network topology based on independent confidence intervals of 39 selected OLS parameter sets. Each entry in the table corresponds to regulation of a gap gene indicated on a row by a gene indicated on a column. '+' ('−') indicates activation (repression) when the confidence intervals for the corresponding regulatory weight fall entirely into the positive (negative) part of the plane for a majority of parameter sets. Similarly, '+ =' ('− =') indicates no repression (no activation) when the confidence intervals for the corresponding regulatory weight fall into the positive (negative) part of the plane and slightly extend to negative (positive) part within 'no regulation' threshold range, i.e., ≥ -0.005 (≤ 0.005). If the confidence intervals significantly extend to both sides of the plane, then no conclusion can be made (denoted by '×').

Note that for all gap genes, promoter strengths R, diffusion coefficients D, and decay

rates λ have extremely large independent confidence intervals (not shown here) meaning that all these parameters are not determinable.

tll/hkb mutants The terminal gap genes *tll* and *hkb*, being expressed in the posterior region of the embryo, are responsible for setting up the posterior boundaries of the gap gene domains. In *tll* mutants the expression of *Kr* is normal, the *kni* domain expands posteriorly, the posterior *gt* domain does not retract from the posterior pole, and the posterior *hb* domain is absent (see [2] and references therein). In *hkb* mutants the posterior *hb* domain fails to retract from the posterior pole [9]. We shall investigate here if the gap gene model is capable of reproducing such behaviour in *tll*/*hkb* mutants.

We obtain the model outputs for *tll* mutants by setting $E_a^{Tll} = 0$ for all gap genes and leaving all other parameter estimates unchanged. Figure 3.6 shows the model outputs for *tll* mutants (first row) for the expression of gap genes at time point T_8 compared to wild type data. As we can see, OLS parameter sets mainly fail to produce correct mutant phenotypes. Most of the parameter sets have over-expression of the posterior *hb* domain which contradicts the experimental evidence. In most of the cases the posterior *gt* domain is expanded and only a few model outputs have the correct behaviour when the domain does not retract from the posterior pole. The expression of *kni* has not changed in some sets and an additional domain appears in others, failing to predict the expansion of the posterior boundary. The only consistent result can be stated for *Kr* which has a normal expression in all model outputs.

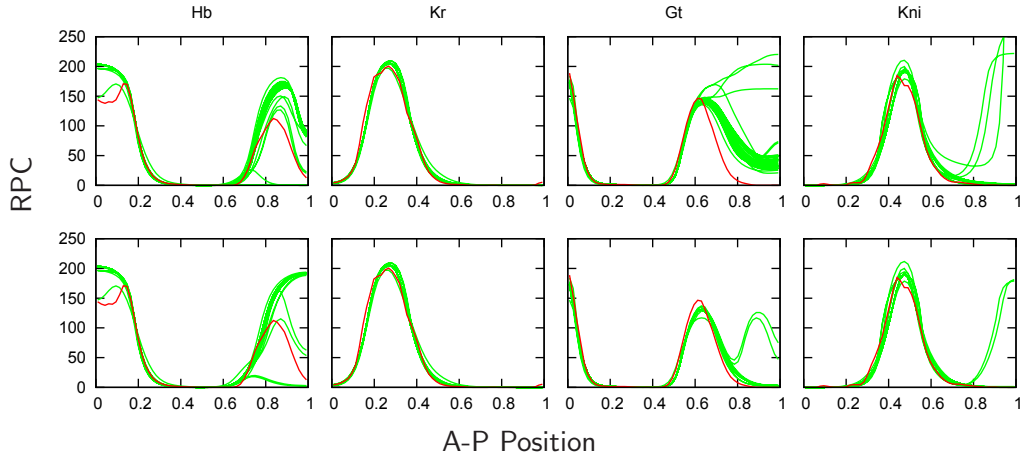

Figure 3.6: Comparison between data (red lines) and model outputs (green lines) with *tll* mutants (first row) and *hkb* mutants (second row) produced by 39 selected OLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 2.1.

Similarly, model outputs for *hkb* mutants are obtained by setting $E_a^{Hkb} = 0$ for all gap genes. Figure 3.6 (second row) shows the expression of gap genes at time point T_8 in *hkb* mutants. The posterior *hb* domain almost disappears in some circuits in contradiction to the experimental evidence. Additional expression domains appear for the gap genes *gt* and *kni*, while the expression of *Kr* has not altered.

In conclusion, OLS parameter sets fail to predict the correct behaviour when terminal genes *tll* and *hkb* are knocked out. Model outputs show both ambiguity and inconsistency, the only exception is the gap gene *Kr*.

3.2 WLS results

The LM search with the WLS cost function has been performed using as initial points the 39 selected OLS parameter estimates and also the 90 OLS sets with lowest RMS values possessing an artificial Kr hump. Additionally, we performed 80 runs starting with parameter estimates obtained from global WLS search (SA). From the obtained results we selected 117 parameter sets with WLS values varying uniformly between 1.08×10^3 and 1.13×10^3. For the comparison with the OLS results, we note that these WLS parameter sets have RMS values uniformly varying between 10.41 and 10.67. It suggests that the WLS search leads to less over-fitting compared to OLS search. None of these low-scoring parameter sets show any visible patterning defects (see Figure 4.2 in the Appendix), while the sets with larger WLS values do. As it is difficult to make a distinction between these 117 parameter sets based on WLS values and expression patterns only, we take all of them into consideration. We emphasize that with a significantly less number of WLS runs (209) compared to the OLS case (740) we have obtained three times more WLS parameter estimates than OLS ones. It is also important that the selection of WLS sets is only based on cost function values and the manual inspection of model outputs for patterning defects, as in the OLS case, is not required.

The most important difference between the model outputs generated by the OLS and WLS parameter sets is that the latter do not have a patterning defect for gap gene Kr (hump). This can be expected because the standard deviations in that region of the embryo are small and subsequently the corresponding weights in WLS are relatively large which prevents the rising of the Kr hump. We note that the model outputs generated by WLS parameter sets have one slight problem which does not show up in the OLS case. Model outputs for gap gene Kr at T_1 have a slight cavity next to the anterior boundary. However, this declination does not exceed the experimental error range and therefore is not considered to be significant.

Patterns for WLS parameter sets (Figure 4.2) at cycle 13 and late time points of cycle 14A show a better fit than the corresponding OLS patterns (Figure 4.1). Especially, the improvement can be seen at the boundaries of gap gene domains at the end of cycle 14A. This can be explained by a relatively small standard deviation of the experimental error at the domain boundaries at late time points (Figure 2.3).

Additionally, WLS model outputs have less variation than those produced by OLS parameter sets. Thereby, WLS model outputs are more consistent with each other while OLS model outputs reveal discrepancies.

Posterior *hb* domain Similar to OLS results, the posterior boundary of the posterior *hb* domain is set correctly (see Figure 4.2) and the anterior shift in the peak of the posterior *hb* domain can be detected in the model outputs (not shown here). The shift of the posterior boundary of the posterior *hb* domain is illustrated in Figure 3.7 by graphical analysis of the *hb* regulation over time (cycle 14A) at three different nuclei which lie in the shift zone. Similar to the OLS case (Figure 3.3), there is a switch from protein synthesis (positive dg_i^{hb}/dt) to decay (negative dg_i^{hb}/dt) of *hb* at the end of cycle 14A and the shift is solely based on

the regulatory mechanism rather than being driven by diffusion. Panels **d-f** of Figure 3.7 show that two major contributions to the regulatory input of *hb* are from the activator Tll and the repressor Hkb. Contrary to the OLS case, these inputs are more stronger than the inputs from other regulators. So, in WLS results *hb* in the posterior region of the embryo is superiorly regulated by terminal genes Tll and Hkb.

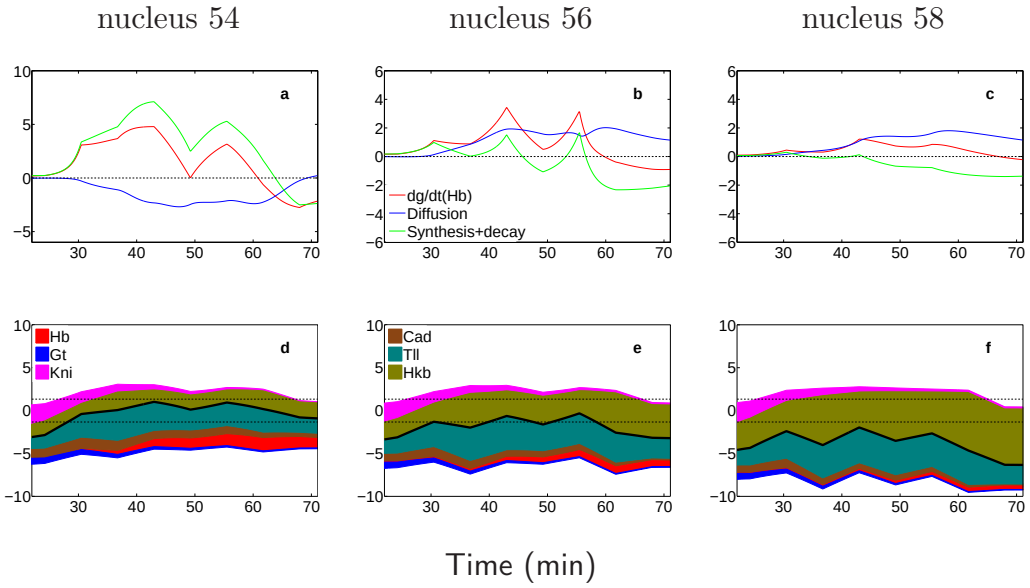

Figure 3.7: Graphical analysis of the shift of the posterior boundary of the posterior *hb* domain in the model outputs for one of the selected WLS parameter sets. Axes, lines, and coloured areas are as in Figure 3.3.

Network topology Classification of all parameter estimates for regulatory weights for 117 selected WLS parameter sets into 'activating', 'repressing' or 'no interaction' categories is given in Figure 3.8. There are only two differences in this topology in comparison with the OLS results in Figure 3.4. Hkb represses *gt* and activates *kni*, while in the OLS case it is the other way around. For other regulatory weights all conclusions agree. At the same time, numbers in Figure 3.8 indicate that WLS parameters estimates are more consistent than those obtained by OLS search. For instance, although activation is concluded for regulatory weights E_{hb}^{Tll} and E_{kni}^{Bcd} in Figure 3.4, still in each case there is one circuit showing repression. Those ambiguities are completely cleared in Figure 3.8 (the only exception is E_{kni}^{Hkb}).

Confidence intervals To assess the quality of the parameter estimates, we computed dependent and independent confidence intervals for each parameter set using (2.14) and (2.15), respectively (see Figure 4.4 in the Appendix). We checked if the corresponding confidence intervals for the regulatory weights fall entirely into the 'repression', 'no interaction', or 'activation' categories. Similar to the OLS case, dependent confidence intervals are small and cannot be trusted. Based on the independent confidence intervals, Table 3.2 summarizes the

	hb	*Kr*	*gt*	*kni*	*Bcd*	*Cad*	*Tll*	*Hkb*
hb	0/0/117	0/117/0	0/0/117	117/0/0	0/0/117	0/0/117	0/0/117	117/0/0
Kr	117/0/0	0/0/117	117/0/0	117/0/0	0/0/117	0/0/117	117/0/0	117/0/0
gt	117/0/0	117/0/0	0/117/0	0/0/117	0/0/117	0/0/117	117/0/0	117/0/0
kni	117/0/0	0/117/0	117/0/0	0/0/117	0/0/117	0/0/117	117/0/0	2/0/115

Figure 3.8: Gap gene network topology based on 117 selected WLS parameter sets. Numbers and colours are as in Figure 3.4.

qualitative conclusions for the regulatory weights in the gap gene model. The qualitative conclusions in Table 3.2 show no significant difference from the corresponding OLS results given in Table 3.1. For 17 regulatory weights the confidence intervals confirm and for another 3 weights they confirm weakly the type of the regulation deduced from the network topology in Figure 3.8.

	hb	*Kr*	*gt*	*kni*	*Bcd*	*Cad*	*Tll*	*Hkb*
hb	+	×	+ =	×	×	+ =	+	×
Kr	×	+	−	−	+	+	×	×
gt	−	−	×	+	+	+	−	×
kni	−	− =	−	+	×	+	×	×

Table 3.2: Gap gene network topology based on independent confidence intervals of 117 selected WLS parameter sets. Notations are as in Table 3.1.

In contrast to the OLS case, the confidence interval analysis for WLS solutions suggests that the number of unknown parameters can be reduced in the model. The dependent confidence intervals for all diffusion parameters in the WLS results have a non-empty intersection. This means that for practical reasons they can be fixed to any value in those intersections without giving a difference in the WLS sums. Since the main interest of the gap gene model lies in the inference of the regulatory network topology, the exact value of the diffusion parameters is not important. Correlation analysis shows that the diffusion coefficients are not strongly correlated to other parameters. Therefore, removing them from the parameter space will not change significantly the determinability of the remaining parameters but it will reduce the size of the problem.

tll/hkb mutants The model outputs for *tll* mutants are shown in Figure 3.9 (first row). The expression of posterior *hb* decreases compared to wild type data. Although it is not completely in agreement with experimental evidence (there is no posterior *hb* domain in such embryos), there is still an improvement in comparison with OLS results (Figure 3.6) where over-expression of *hb* is detected. Similar to OLS results, *Kr* has a normal expression which is in agreement with experiments. Expression of *gt* and *kni* at the posterior part of the embryo appears somewhat abnormal as in OLS outputs but they do not produce the behaviour observed in the experiments.

The model outputs for *hkb* mutants are shown in Figure 3.9 (second row). Contrary to the corresponding OLS results (Figure 3.6), they are more consistent with each other. The

posterior *hb* domain in all cases fails to retract from the posterior pole of the embryo which is in agreement with the experimental evidence [9]. This confirms again that Hkb is the main repressor of *hb* at the posterior part of the embryo. The expression of gap genes *Kr*, *gt*, and *kni* is not affected in *hkb* mutants. It suggests that Hkb does not regulate these genes exposing an unreliability of corresponding regulations in network topology in Figure 3.8. Thereby, we can conclude that in the WLS search the 3 regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb can be eliminated from the parameter search by setting up $E_{Kr}^{Hkb} = E_{gt}^{Hkb} = E_{kni}^{Hkb} = 0$. This is also confirmed by the statistical analysis, as their dependent confidence intervals include zero (see Figure 4.4).

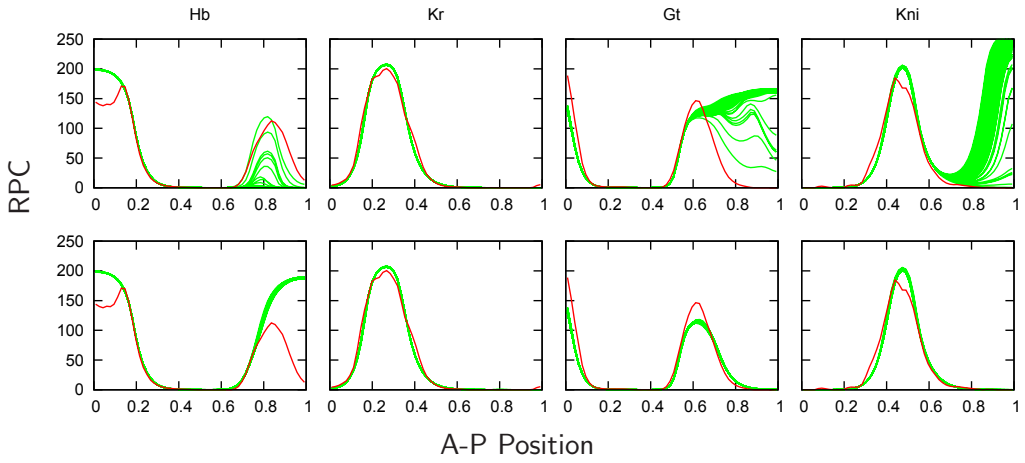

Figure 3.9: Comparison between data (red lines) and model outputs (green lines) with *tll* mutants (first row) and *hkb* mutants (second row) produced by 117 selected WLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 2.1.

3.2.1 WLS results with fixed parameters

We have found that for the WLS search it is possible to reduce the size of the parameter space by fixing all diffusion parameters and the regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb. For the diffusion coefficients we computed the averaged values based on the previously found estimates, $D_{hb} = 0.237$, $D_{Kr} = D_{kni} = 0.3$, and $D_{gt} = 0.115$. Note that these averaged values belong to the non-empty intersections of the dependent confidence intervals. So, it leaves us with 37 parameters in the model to be re-estimated. We used LM search with 60 initial parameter sets arbitrarily chosen from previously found 117 WLS parameter sets. Additionally, we performed 20 runs with initial parameter values obtained from global WLS search (SA) with those parameters fixed. From re-estimated parameter sets we select 66 circuits which have low WLS values (about 1.08×10^3). None of them reveals any visible patterning defects (not shown here). The network topology in Figure 3.8 remains unchanged with the new estimates except for the regulations of *Kr*, *gt*, and *kni* by Hkb which are set to zero. Table 3.3 presents the qualitative conclusions for the regulatory weights in the gap gene model based on the independent confidence intervals (Figure 4.5 in the Appendix). These results show an improvement in

comparison with Table 3.2. For 20 regulatory weights the confidence intervals confirm and for another 5 weights they confirm weakly the type of the regulation in the network topology and only 4 regulations still remain unclear.

	hb	Kr	gt	kni	Bcd	Cad	Tll	Hkb
hb	$+$	$\times$	$+$	$-$	$\times$	$+=$	$+$	$-=$
Kr	$-=$	$+$	$-$	$-$	$+$	$+$	$\times$	0
gt	$-$	$-$	$\times$	$+$	$+$	$+$	$-$	0
kni	$-$	$-=$	$-$	$+$	$+=$	$+$	$-$	0

Table 3.3: Gap gene network topology based on independent confidence intervals of re-estimated 66 WLS parameter sets. Notations are as in Table 3.1.

tll mutants The model outputs for *tll* mutants are shown in Figure 3.10. As we can see, there is a significant improvement in comparison with the OLS results (Figure 3.6) and preceding WLS results (Figure 3.9). Now, WLS circuits predict correct mutant phenotypes for all gap genes. Namely, the posterior *hb* domain is absent, the expression of *Kr* is normal, there is expansion of the posterior boundary of the *kni* domain, and the posterior *gt* domain does not retract from the posterior pole.

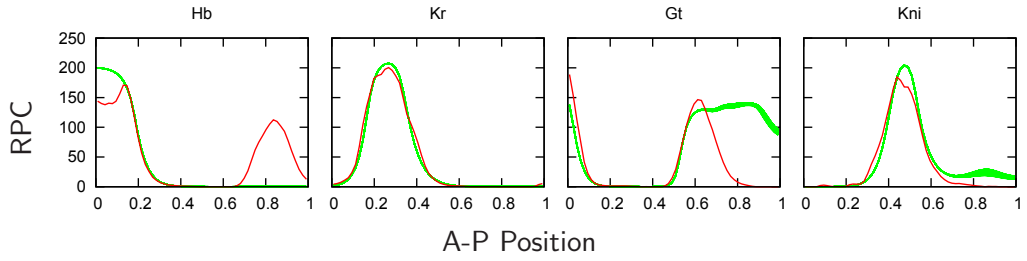

Figure 3.10: Comparison between data (red lines) and model outputs with *tll* mutants (green lines) produced by re-estimated WLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 2.1.

Correlations The qualitative conclusions from Table 3.3 are not completely consistent with the network topology obtained by only considering the value of the parameter estimates. The sizes of the independent confidence intervals (see Figure 4.5) give an indication about the determinability of the corresponding regulatory weights. Note the big difference between the size of the independent confidence intervals for the different regulatory weights indicating a different degree of determinability. The lack of determinability is due to the presence of correlations between parameter estimates indicated by the large difference between dependent and independent confidence intervals. Individual confidence intervals are not informative for understanding the reason of poor determinability of parameters when their estimates are correlated. Using (2.16), we find the correlation matrix for each parameter set. To detect the most significant correlations between parameters present in all

correlation matrices, we calculated an averaged matrix—which we call the mean correlation matrix—whose entries are the mean values of the corresponding correlation coefficients in the individual correlation matrices. The obtained mean correlation matrix has a block diagonal structure such that each block corresponds to a given gene and contains the correlation coefficients between parameters for the same gene (not shown here). This is mainly due to the nature of function (2.2) used in (2.1). The positive and negative inputs in its argument can compensate or complement each other. We identify the most significant parameter correlations which can be interpreted in biological terms with the emphasis on those for which the qualitative conclusions in Table 3.3 are weak or cannot be made at all:

- Activations of *hb* by Bcd and Cad are correlated;
- Activation of *hb* by Bcd is also correlated to its repression by Kni;
- Repression of *hb* by Hkb is correlated to activation by Tll;
- Repression of Kr by Hb is correlated to its activation by Bcd;
- Activation of *kni* by Bcd is strongly correlated to its repression by Hb;

The regulatory weights W_{hb}^{Kr}, W_{gt}^{gt}, and W_{kni}^{Kr} have relatively small independent confidence intervals. Results for these weights in Table 3.3 are based on the threshold 0.005 for classification of regulations. With a larger threshold, such as 0.01, for all 3 weights 'no regulation' type can be concluded confirming the corresponding predictions from Figure 3.8. Finally, we note that E_{Kr}^{Tll} is not correlated to any other weight. Posterior Kr is strongly repressed by Gt and somewhat weaker by Hb and Kni. Apparently, due to these interactions, repression of Kr by Tll is somewhat redundant in the model.

4 Conclusions

In this note we have investigated the model for spatio-temporal pattern formation of gap gene products (*hb*, *Kr*, *gt*, and *kni*) in early development of *Drosophila*. Previous studies of the gap gene system [2, 5, 6, 7, 8] along with these gap genes also included in the model the products of genes *cad* and *tll* as state variables. In our model we have included *cad* and *tll* as time-variable external inputs. This is a more natural way to model that they regulate gap genes, but not vice versa. Contrary to previous studies where protein Bcd was used as external input constant in time, we have incorporated its temporal behaviour in our model. Finally, new data for *hkb* [10] has allowed us to supplement the gap gene network by including *hkb* as time-variable external input. Note that *hkb*, which is absolutely necessary for correct regulation of posterior *hb* domain, was missing in previous studies of the gap gene system. Thereby, our model describes the spatio-temporal dynamics of 4 gap genes and includes 4 external inputs. It is noteworthy that with our model the complexity of the problem is reduced both with regard to the number of equations and the number of unknown parameters compared to previous models.

The model has a number of unknown parameters among which the most interesting are the regulatory weights, each one representing quantitatively the regulation of one gene by another gene. Following the common way, we have inferred the unknown parameters by fitting model outputs to gap gene data [3, 4]. As cost function to minimize in the parameter estimation procedure we have used both the Ordinary Least Squares (OLS) sum, similar

to all previous studies, and the Weighted Least Squares (WLS) sum with weights taken inversely proportional to the corresponding standard deviations of the experimental error distributions. Since the standard deviations are available from [4], the WLS method does not require additional computational work compared to the OLS search.

We have used the gradient-based Levenberg-Marquardt (LM) method in the optimization with the initial parameter values obtained from global search runs using Simulated Annealing (SA). A large amount of runs has been performed to obtain the parameter estimates, 740 and 209 with OLS and WLS search, respectively. From the obtained parameter sets we first selected the low-scoring sets based on the values of OLS and WLS sums only. It gave us 117 WLS and 589 OLS parameter sets. While the network topology based on the values of the estimated regulatory weights in the WLS case (Figure 3.8) shows an agreement with the known genetic evidences, corresponding OLS results reveal a number of contradictions. Interestingly, all those OLS sets which have disagreements with the theory, despite having an overall reasonable fit to the data, do possess one patterning defect (hump) in the expression of *Kr* in the region where this gene is not expressed in the data. By manual inspection of model outputs we have selected 39 OLS parameter sets which do not have that artefact. The network topology based on these sets (Figure 3.4) is in agreement with genetic evidence. Thus, the selection of parameter sets reveals the first drawback of using the OLS rather than the WLS measure. While the selection criterion based on the cost function value is sufficient for the WLS case, an additional check for patterning defects in the OLS model outputs has to be performed. Moreover, with WLS search we have done less estimation runs and still obtained more parameter sets than with OLS search.

The model outputs produced with the selected WLS parameter sets reveal a better fit at the boundaries of the gap gene domains at late stages of cycle 14A than the corresponding OLS patterns. Additionally, WLS patterns are more consistent with each other which is indicated by less variation in the model outputs.

Our results, both OLS and WLS, show a significant improvement in the regulation of the posterior *hb* domain compared to previous results. Namely, the posterior boundary of this domain is set up correctly and the anterior shift in the peak of the domain is present in the model outputs while previous models failed to reproduce such a shift. More importantly, with our network also the shift in time of the posterior boundary of the posterior *hb* domain is detected. We have shown that this shift is solely based on the regulatory mechanism rather than being forced by diffusion (Figures 3.3 and 3.7). Namely, the boundary shift is due to the suppressive repression of *hb* by Hkb. In previous studies, gap gene models failed to show this shift because *hkb* was missing in the network.

Confidence interval analysis for the selected OLS and WLS parameter estimates show no significant difference from each other in terms of their determinability. In both cases qualitative conclusions can be made only for 17 (out of 32) regulatory weights (Tables 3.1 and 3.2). Thus, the network topology based only on the values of parameter estimates is not entirely confirmed by confidence interval analysis. However, there is a significant improvement in comparison with the corresponding results in [8] where qualitative conclusions were deduced only for 9 weights. This improvement is most likely due to the change of genes *cad* and *tll* from state variables to external inputs in our model and decreasing by that the level of uncertainty in the model parameters.

We have used our OLS and WLS parameter sets for qualitative prediction of gap gene expression in *tll* and *hkb* mutants (Figures 3.6 and 3.9). In *tll* mutants both OLS and WLS sets fail to predict correctly the expression of gap genes, except for *Kr* which is not

altered. In *hkb* mutants the posterior *hb* domain in WLS outputs does not retract from the posterior pole in agreement with the experiments [9], while OLS results fail to reproduce such behaviour. Additionally, expression of other gap genes in WLS outputs is not changed suggesting that Hkb does not regulate those genes. In OLS results this is observed only for *Kr*.

The confidence intervals for WLS parameter sets show that all diffusion parameters and the regulatory weights corresponding to regulation of *Kr*, *gt*, and *kni* by Hkb can be eliminated from parameter space, i.e., they can be fixed during the search. We have performed additional WLS runs with those parameters fixed and selected 66 low-scoring sets from the obtained results. With the new parameter estimates, firstly, we have achieved an improvement in the qualitative conclusions for some of the regulatory weights (Table 3.3). Secondly, *tll* mutants with those sets give correct qualitative predictions for the expression of all gap genes (Figure 3.10).

To sum up, based on the results of our analysis, we conclude that the WLS sum is a more suitable measure for infering a gap gene circuit from the experimental data than the OLS sum.

Acknowledgement I acknowledge support from NWO's 'Computational Life Science' program, projectnr. 635.100.010. I would like to thank Johannes Jaeger for providing me the *hkb* data and the parameter estimates obtained from the global search. I am also grateful to Prof. dr. J.G. Verwer and J.G. Blom for their valuable comments and suggestions.

References

[1] M. Akam, (1987), *The molecular basis for metameric pattern in the Drosophila embryo*, Development 101, pp. 1-21.

[2] J. Jaeger, J. Reinitz (2004), *Dynamical analyses of regulatory interactions in the gap gene system of Drosophila melanogaster*, Genetics 167, pp. 1721-1737.

[3] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, J. Reinitz, (2004), *A database for management of gene expression data in situ*, Bioinformatics 20, pp. 2212-2221.

[4] FlyEx Database (http://flyex.ams.sunysb.edu/flyex).

[5] J. Jaeger, J. Reinitz (2004), *Dynamic control of positional information in the early Drosophila embryo*, Nature 430, pp. 368-371.

[6] T. J. Perkins, J. Jaeger, J. Reinitz, L. Glass, (2006), *Reverse Engineering the Gap Gene Network*, PLoS Computational Biology 2:e51.

[7] Y. F. Nanfack, J. A. Kaandorp, J. G. Blom (2007), *Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of Drosophila melanogaster*, Bioinformatics 23, pp. 3356-3363.

[8] M. Ashyraliyev, J. Jaeger, J. G. Blom (2008), *On Parameter Estimation and Determinability for Drosophila Gap Gene Circuits*, BMC Systems Biology 2:83.

[9] J. Casanova (1990), *Pattern formation under the control of the terminal system in the Drosophila embryo*, Development 110, pp. 621-628.

[10] J. Jaeger, K. Siggens (2009), unpublished results.

[11] Manu, S. Surkova, A. V. Spirov, V. V. Gursky, H. Janssens, A.-R. Kim, O. Radulescu, C. E. Vanario-Alonso, D. H. Sharp, M. Samsonova, J. Reinitz (2009), *Canalization of Gene Expression in the Drosophila Blastoderm by Gap Gene Cross Regulation*, PLoS Biology 7(3):e49.

[12] G. A. F. Seber, C. J. Wild (1988), *Nonlinear regression*, New York, John Wiley & Sons.

[13] A. A. Simcox, J. H. Sang (1983), *When does gastrulation occur in Drosophila embryos?*, Developmental Biology 97, pp. 212-221.

[14] V. E. Foe, B. M. Alberts (1983), *Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis*, J. Cell Science 61, pp. 31-70.

[15] E. Myasnikova, A. Samsonov, K. Kozlov, M. Samsonova, J. Reinitz (2001), *Registration of the expression patterns of Drosophila segmentation genes by two independent methods*, Bioinformatics 17, pp. 3-12.

[16] B. W. Lindgren, G. W. McElrath, D. A. Berry (1978), *Introduction to Probability and Statistics*, New York, Macmillan Publ.

[17] D. H. Sharp, J. Reinitz (1998), *Prediction of mutant expression patterns using gene circuits*, Biosystems 47, pp. 79-90.

[18] D. W. Marquardt (1963), *An algorithm for least-squares estimation of nonlinear parameters*, SIAM J. Appl. Math. 11, pp. 431-441.

[19] J. Nocedal, S. J. Wright (1999), *Numerical Optimization*, New York, Springer.

[20] J. C. P. Bus, B. Domselaar, J. Kok (1975), *Nonlinear least squares estimation*, CWI report, NW 17/75.

[21] G. H. Golub, C. F. Loan (1996), *Matrix computations*, Baltimore, Johns Hopkins UP.

[22] K. Jaqaman, G. Danuser (2006), *Linking data to models: data regression*, Nature Reviews Molecular Cell Biology 7, pp. 813-819.

[23] L. Ljung (1999), *System Identification Theory For the User*, New Jersey, Prentice Hall.

[24] R. C. Aster, B. Borchers, C. H. Thurber (2005), *Parameter Estimation and Inverse Problems*, USA, Elsevier.

[25] M. Ashyraliyev, Y. F. Nanfack, J. A. Kaandorp, J. G. Blom (2008), *Parameter estimation for biochemical models*, FEBS J 276(4), pp. 886-902.

[26] C. W. Gear (1971), *Numerical initial value problems in ordinary differential equation*, Englewood Cliff, Prentice Hall.

Appendix: Additional plots

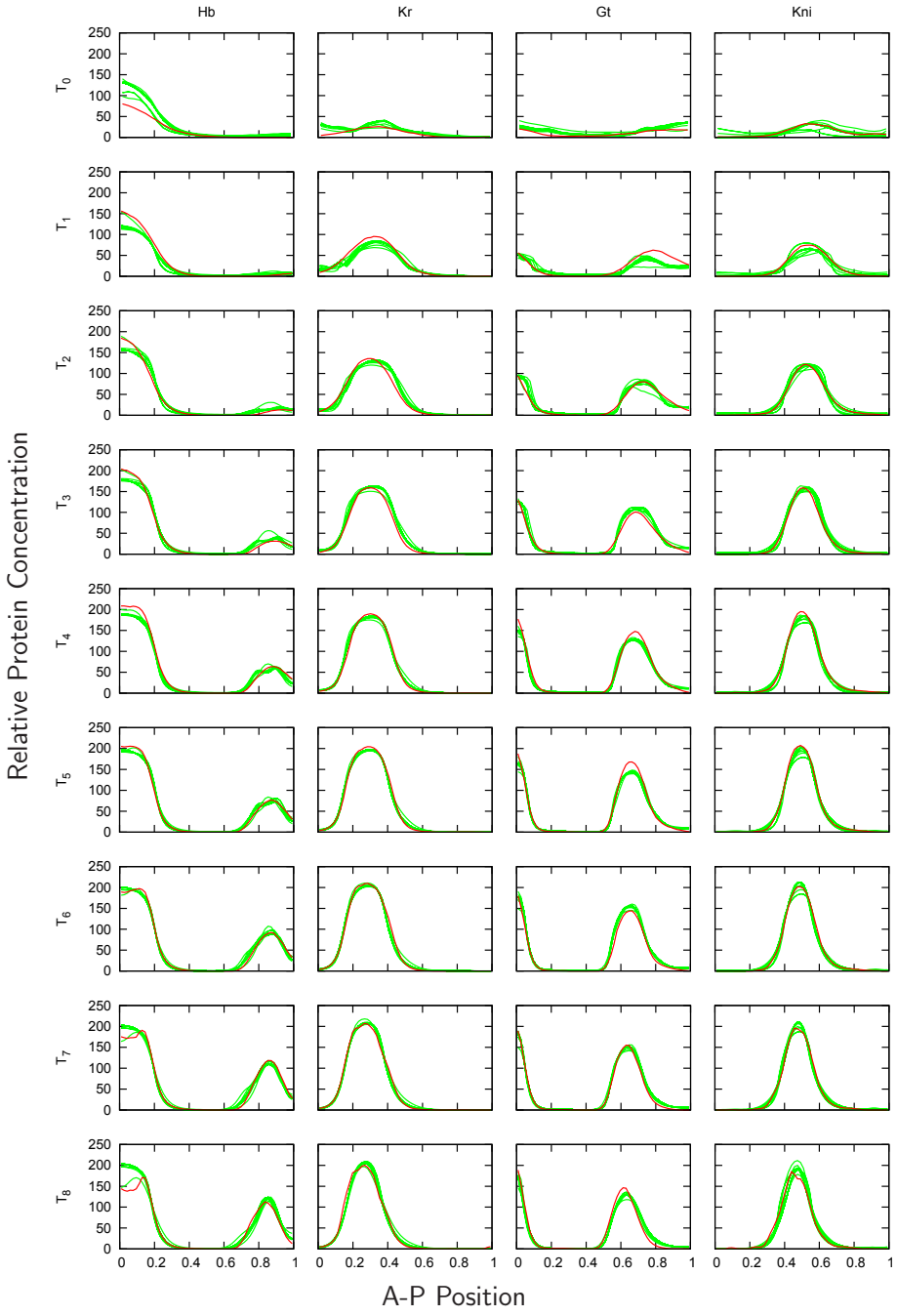

Figure 4.1: Model outputs for the 39 selected OLS parameter sets (green lines) vs data (red lines) for gap genes at all time points T_i $(i = 0, 1, \ldots, 8)$. Axes are as in Figure 2.1.

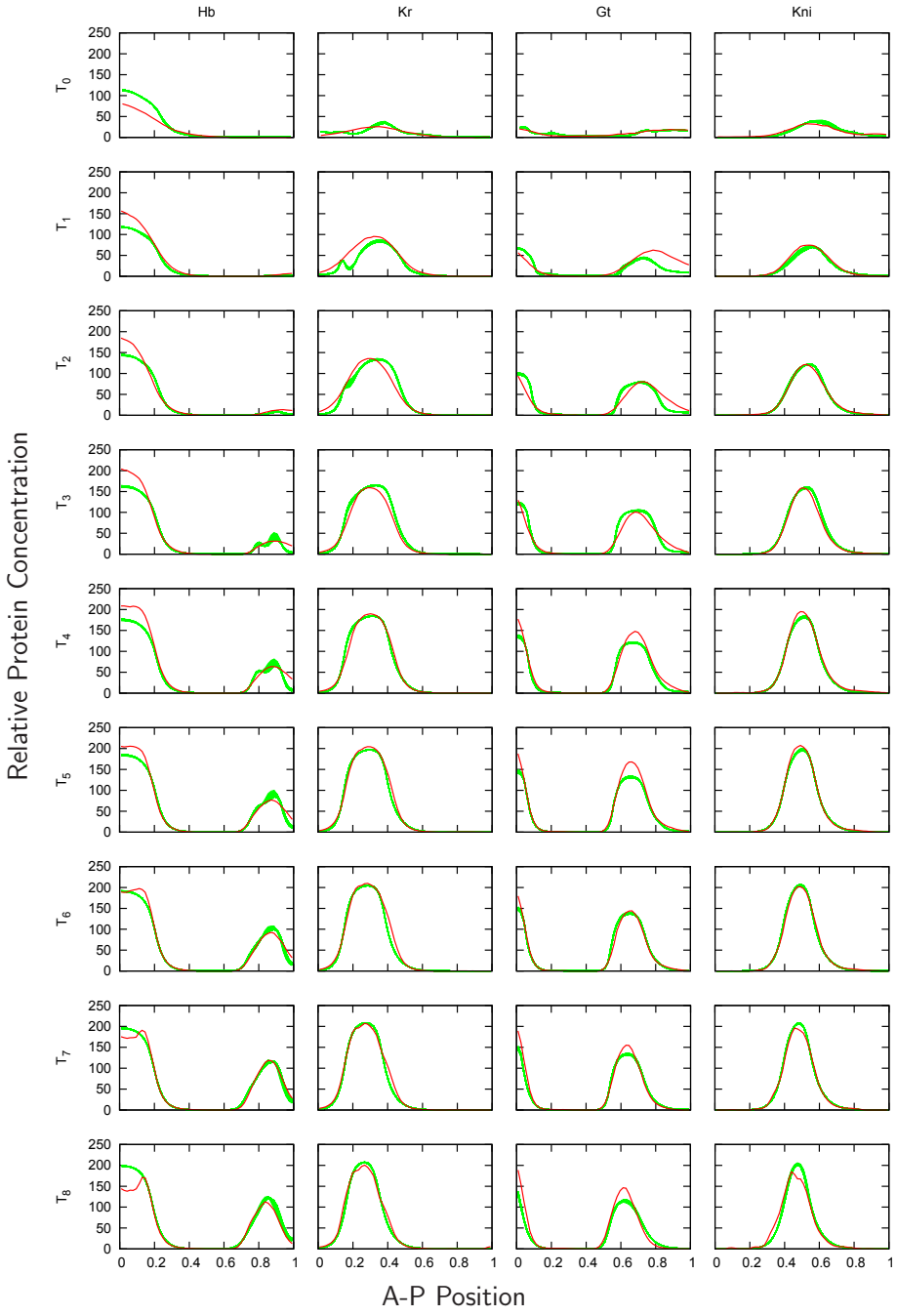

Figure 4.2: Model outputs for the 117 selected WLS parameter sets (green lines) vs data (red lines) for gap genes at all time points T_i ($i = 0, 1, \ldots, 8$). Axes are as in Figure 2.1.

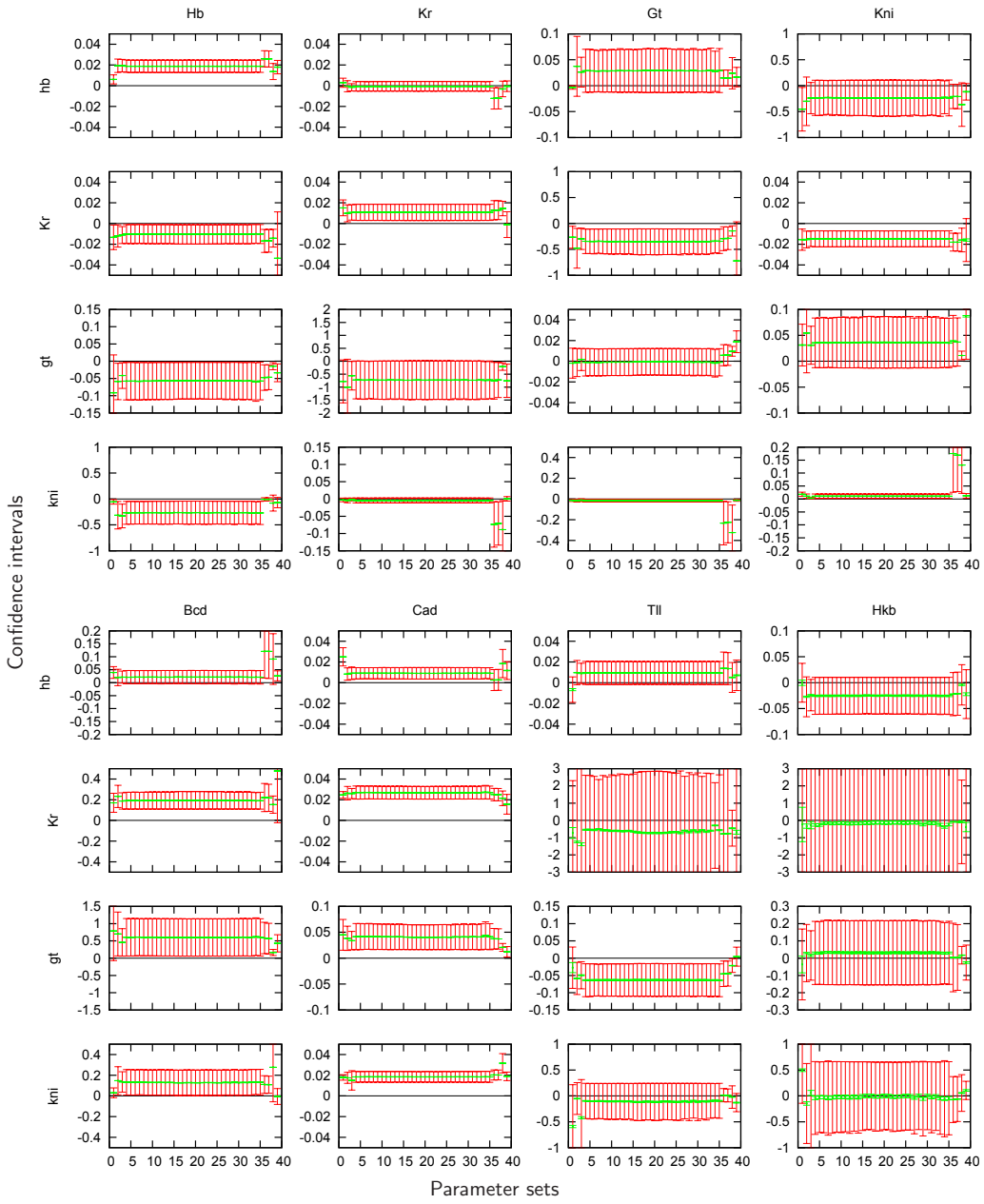

Figure 4.3: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 39 selected OLS parameter sets.

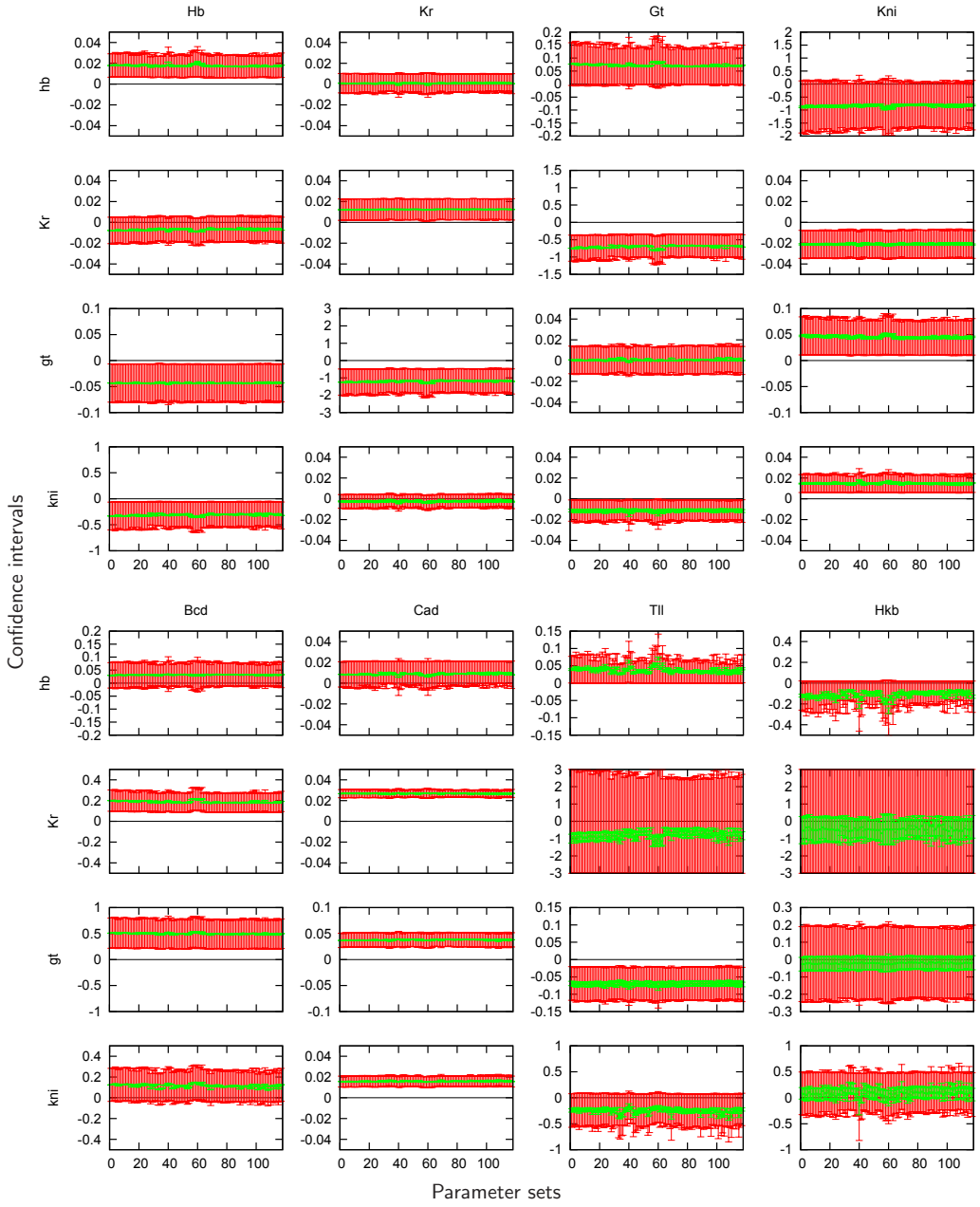

Figure 4.4: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 117 selected WLS parameter sets. Note the different scale in y-axis for some of the regulatory weights compared to the corresponding plot in Figure 4.3.

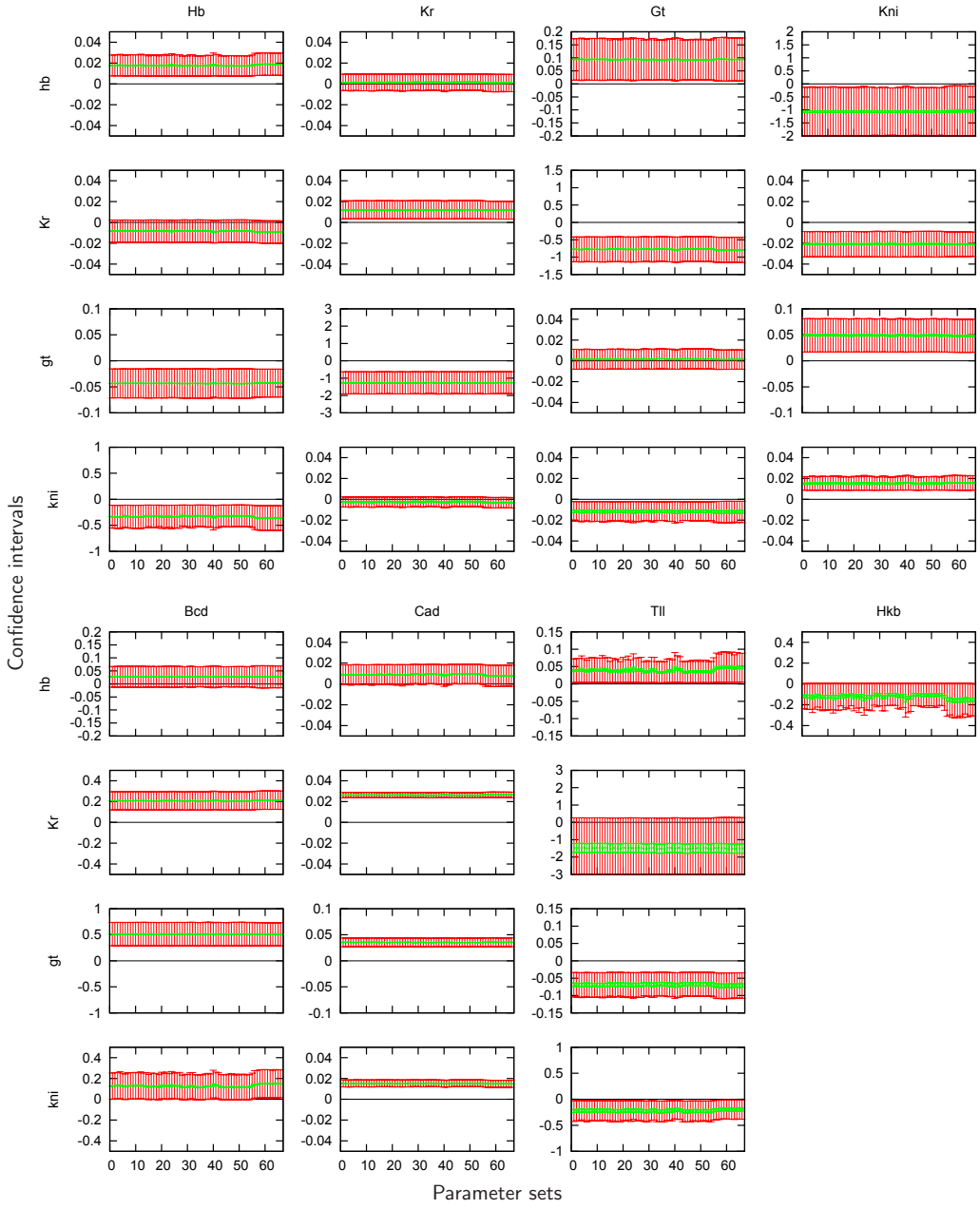

Figure 4.5: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 66 selected WLS parameter sets obtained with duffusion parameters and the regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb being fixed during the search.