Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Modeling the temporal evolution of the *Drosophila* gene expression

Alexandre Haye*, Yves Dehouck*, Jean Marc Kwasigroch*, Philippe Bogaerts, Marianne Rooman***

** Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, 1050 Brussels, Belgium (e-mail: ahaye/ydehouck/jkwasigr/mrooman@ulb.ac.be).*
*** Groupe de Biomodélisation et Bioprocédés, Université Libre de Bruxelles, 1050 Brussels, Belgium (e-mail: Philippe.Bogaerts@ulb.ac.be)*

Abstract: The evolution of the gene expression pattern of *Drosophila*, from the embryonic to adult development phases, was studied on the basis of a microarray time series involving the expression levels of 4028 genes over 67 time-points. The genes presenting a similar temporal evolution of their expression levels were clustered together, so as to define a small number of representative classes. To model the network interactions responsible for the dynamic behavior of gene expression, a system of linear differential equations with constant coefficients was used. The parametric estimation of this model was performed in two stages: a first stage of linear parameter identification allowing an analytical approach to the solution, and a second stage of nonlinear parametric estimation which refines this solution. This model is shown to reproduce the experimental gene expression profiles with a fairly good precision.

1. INTRODUCTION

The DNA microarray technology allows measuring simultaneously the expression levels of several thousands of genes in a cell sample. Time series, obtained by considering cells at different moments of their development or the development of their host organism, give the possibility to analyze the temporal evolution of gene expression.

Using this technology, Arbeitman *et al.* (2002) have measured the expression rates of 4028 genes of one of the model organisms of developmental biology, *Drosophila melanogaster*, across all its development stages. This temporal series is composed of 67 time points in total. It starts at the fertilization of the organism and follows its life cycle during 40 days, going through the embryonic, larval and pupal stages and the first 30 days of adulthood, where males and females were sampled separately. The cell samples were taken indistinguishably from any part of the organism and thus represent an average of the gene expression levels in the different tissues. The expression rates measured in these samples were compared with a common reference sample composed of mixed mRNA's from all development stages and tissues.

This time series was chosen for several reasons. We are interested in modeling the development stages of a multicellular organism in the absence of any external perturbation, and *Drosophila* is one of the simplest and most studied organisms of this type. Moreover, the *Drosophila* time series of Arbeitman *et al.* (2002) is currently one of the longest series available, which makes the modeling quite interesting and informative.

2. CLASSIFICATION OF GENE EXPRESSION PROFILES

2.1 Classification procedure

The Smoothing Spline Clustering (SSC) method has been specifically designed to classify DNA microarray time series, and has been validated on the *Drosophila* expression profiles (Ma *et al.*, 2006). It is freely available in the SSClust software (www.genemerge.bioteam.com/SSClust.html).

The SSC method groups in the same class expression profiles that present similar shapes, *i.e.* similar expression rates and similar time evolutions. Therefore, consider the expression rate $R_i(t)$ of a gene i at a given time t, normalized by its expression rate in the reference sample R_i^R: $y_i(t)=log_2(R_i(t)/R_i^R)$. The logarithm in basis 2 is considered for convenience, as genes can be taken as significantly more (or less) expressed than random when the expression rates are twice larger (or smaller) than in the reference sample, that is, when $y_i(t)>1$ (or $y_i(t)<(-1)$).

The expression profile y of a given gene i, belonging to cluster c, is decomposed as follows:

$$y_i(t) = x_c(t) + b_i + \varepsilon_i(t) \quad , \tag{1}$$

where t are the time-point indices, $x_c(t)$ represents the mean profile of the cluster c, b_i is the time-independent part of the deviation from the mean profile, and $\varepsilon_i(t)$ is the time

dependent part of this deviation. The b_i–parameters account for the constant difference in expression rates of genes whose expression profiles have similar shapes, and $\varepsilon_i(t)$ includes the measurement error. Note that the $\varepsilon_i(t)$-parameters may also include other effects, for example related to the fact that the cell samples are often extracted from different tissues, and that the measurements therefore mix the dependencies of the expression rates on the organism's development stage and on the cell's host tissue.

The SSC algorithm tends to find the optimal number of clusters and the optimal distribution of genes in these clusters, so as to ensure the smoothness of the mean curve $x_c(t)$ and to minimize the deviations of the expression profiles of the genes in the cluster with respect to this curve.

2.2 Classification results

The application of the SSC method to the *Drosophila* time series of Arbeitman *et al.* (2002) yields an optimal number of 17 clusters (Ma *et al.*, 2006), corresponding to a compromise between a sufficient population in the classes and the similarity of the profiles inside the classes. These profiles are depicted in Fig. 1, for the male flies. Ma *et al.* (2006) have moreover investigated the functional significance of these 17 clusters and have found in 12 of them a significant overrepresentation of genes involved in a well defined biological process. They also verified that the peaks in the expression profiles of these clusters correspond to development stages where this process occurs. This analysis tends to support the biological significance of the clustering.

3. MODELING GENE EXPRESSION EVOLUTION

3.1 Model structure

The formalism of differential equations is particularly well suited for modeling systems with explicit time evolution and for reproducing complex dynamic behaviors like oscillations or multistationarity, which can occur in biological systems. We thus chose this formalism to model the time evolution of the *Drosophila* gene regulatory network across its development stages. As a starting point, we assumed that the system is autonomous and used the simplest model where the differential equations are linear and have constant coefficients (Chen *et al.*, 1999). We thus supposed that the time evolution of the gene expression level x_c of cluster c only depends on the evolution of the gene expression levels $x_{c'}$ of all clusters c'. Defining the vector $\mathbf{x}=(x_1, x_2, \ldots, x_n)^T$ where n is the number of clusters, this corresponds to considering the set of coupled linear differential equations:

$$\frac{d\,\mathbf{x}}{d\,t} = \mathbf{M}\,\mathbf{x} \quad , \tag{2}$$

where t is the time and $\mathbf{M}$ a nxn matrix with constant coefficients.

The problem thus amounts to estimating the elements of the $\mathbf{M}$ matrix, that is, n^2 parameters, on the basis of the gene expression levels measured by DNA microarray techniques and clustered in a small number of groups. More precisely, n is here equal to 17, and the $x_c(t)$'s are the mean expression levels defined in eq.(1), taken at the 67 discrete time-points t. The parameter estimation is performed in two steps, a first step of linear parameter identification, which yields initial parameter values for the second, nonlinear, estimation.

3.2 Linear parameter estimation

To estimate the n^2 elements of $\mathbf{M}$ through eq. (2), we first need an estimation of the time derivatives $dx_c(t)/dt$ of the expression profiles. This is performed by a cubic smoothing spline using the *csaps* routine of the *Matlab* program. Defining the nxf matrix $\mathbf{X}=(\mathbf{x}(t_1), \mathbf{x}(t_2),\ldots, \mathbf{x}(t_f))$, where f=67 is the number of time-points, the matrix elements of $\mathbf{M}$ are estimated as:

$$\hat{\mathbf{M}}^{LS} = \frac{d\,\mathbf{X}}{d\,t} /_{LS}\, \mathbf{X} \quad , \tag{3}$$

where $/_{LS}$ means the right division in the least square sense (the *mrdivide* routine of *Matlab).*

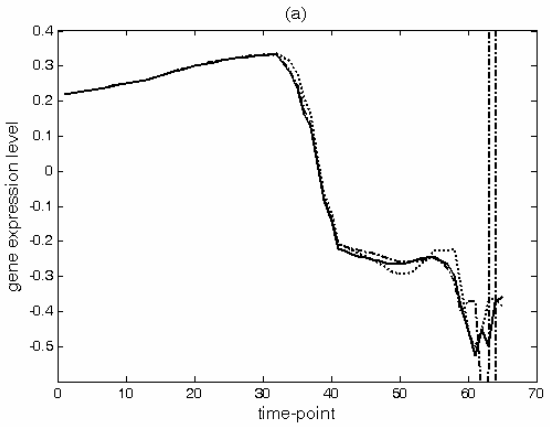

Figure 1. Experimental and modeled gene expression profiles of *Drosophila*, for each of the 17 gene clusters. The time spanned in the embryonic phase is 24 hours (time-points 1-31), 81 hours in the larval phase (time-points 32-41), 111 hours in the pupal phase (time-points 42-59), and 30 days in the adult phase (time-points 60-67). To improve the readability of the figure, the time axis was stretched so as to make each time-point equidistant of its neighbors. Solid lines correspond to the smoothed experimental profiles $x_c(t)$, dashed-dotted lines to profiles modeled using linear parameter estimation $\hat{x}_c^{LS}(t)$, and dotted lines to profiles modeled using nonlinear parameter estimation $\hat{x}_c^{Opt}(t)$. (a) cluster 1.

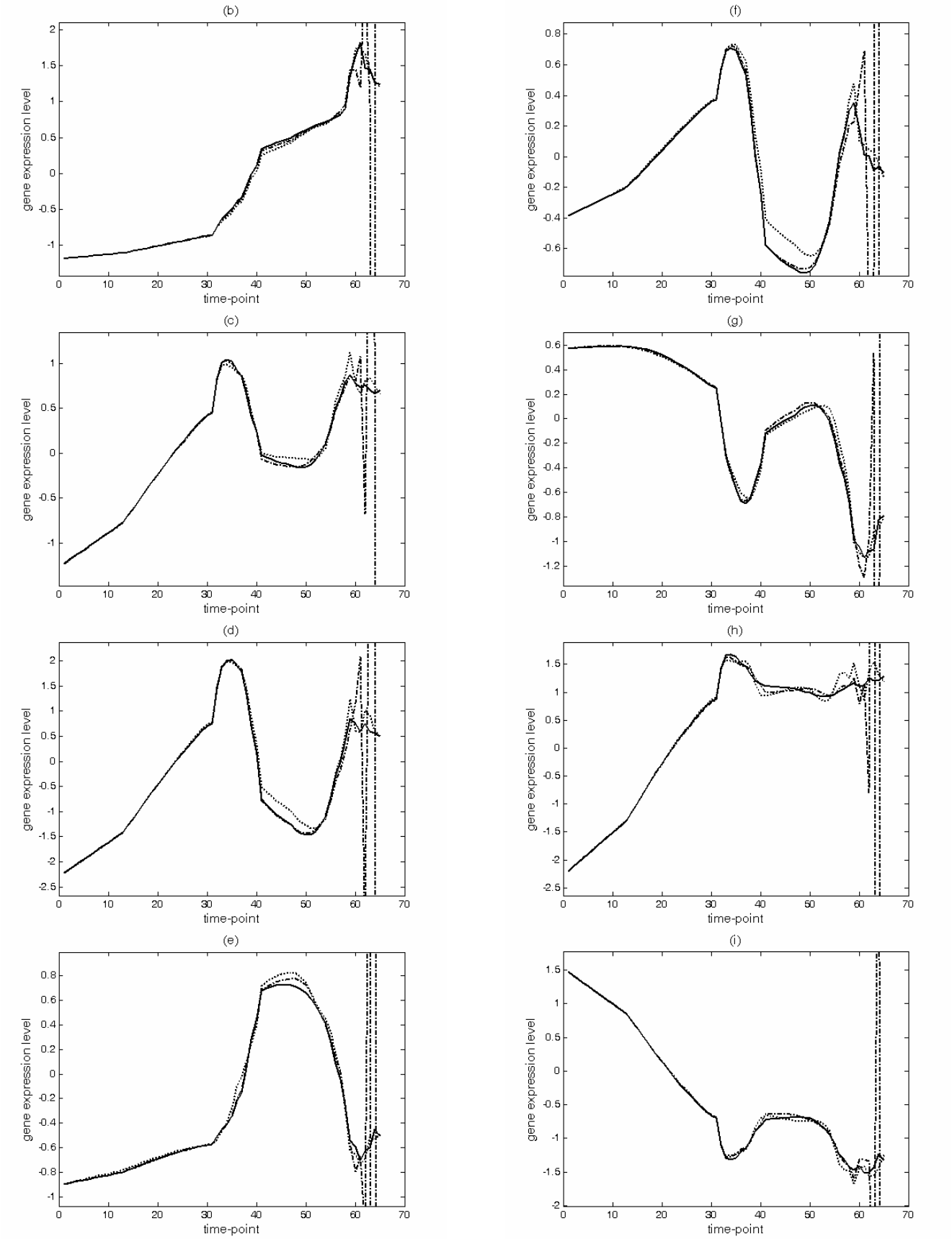

Figure 1 (continued). Experimental and modeled gene expression profiles of *Drosophila*. (b)-(i): clusters 2-9.

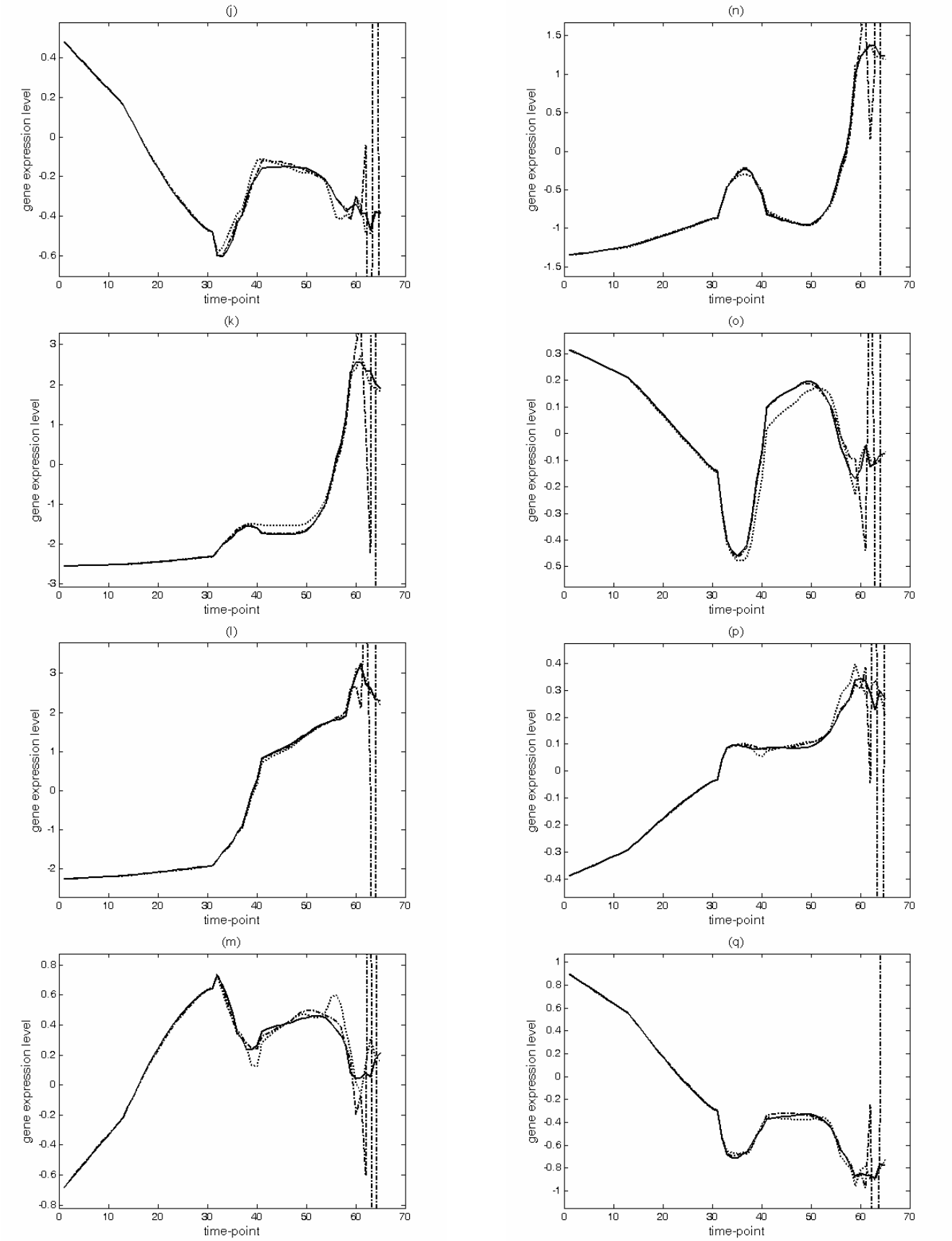

Figure 1 (continued). Experimental and modeled gene expression profiles of *Drosophila*. (j)-(q): clusters 10-17.

To compare the validity of this estimation, we integrate the differential equations (2) with the estimated matrix $\hat{\mathbf{M}}^{LS}$ instead of **M** and considering the first time-point $\mathbf{x}(t_1)$ as initial condition, using a classical Runge-Kutta algorithm (Forsythe *et al.*, 1977) (*ode45* routine of *Matlab*). This yields the estimated gene expression profiles $\hat{\mathbf{x}}^{LS}(t)$.

The experimental and estimated profiles, $\mathbf{x}(t)$ and $\hat{\mathbf{x}}^{LS}(t)$, are depicted in Fig. 1 for the 17 clusters. The modeled profiles reproduce perfectly well the measured ones for the first time-points, close to the initial conditions; the deviation increases with time and leads to divergent oscillations at the last time-points. This is due to the approximate estimation of the time derivative of **X**, which is not accurate enough in the region where moreover the information is sparse. Indeed, the time difference between successive time points in the adult phase is of 2 to 5 days, against 0.5 to 1 hour in the embryonic phase.

To have an objective quantification of the quality of the modeling of the experimental profiles, we compute the root mean square deviation $S(\hat{x}_c)$ for each cluster *c* as:

$$S(\hat{x}_c) = \sqrt{\sum_{k=1}^{f} \frac{\left(x_c(t_k) - \hat{x}_c(t_k)\right)^2}{f}} \quad . \qquad (4)$$

The values of the $S(\hat{x}_c^{LS})$ deviations are given in Table 1 for all the clusters. They are equal, on the average, to 10.47.

3.3 Nonlinear parameter estimation

The linear parameter identification uses an estimate of the derivatives that induces an error on the parameter identification. To reduce this error, we perform a nonlinear parameter estimation, using as initial parameter values those obtained by the linear identification procedure. More precisely, we search for the $\hat{\mathbf{M}}^{Opt}$ matrix that minimizes a cost function *J*:

$$\hat{\mathbf{M}}^{Opt} = \underset{\mathbf{M}^*}{ArgMin} \quad J(\mathbf{M}^*) \quad . \qquad (5)$$

The chosen cost function *J* corresponds to the quadratic sum of the differences between the experimental and modeled profiles, weighted by the inverse of the *nxn* diagonal matrix **V**(t), containing on its diagonal the variance of the experimental data in each cluster at a given time-point:

$$J(\mathbf{M}^*) = \sum_{k=1}^{f} \left[\left(\mathbf{x}(t_k) - \mathbf{x}^*(t_k)\right)^T \mathbf{V}(t_k)^{-1} \left(\mathbf{x}(t_k) - \mathbf{x}^*(t_k)\right) \right] \quad , \qquad (6)$$

where $\mathbf{x}^*(t)$ is the expression level vector associated to the $\mathbf{M}^*$-matrix whose cost function is evaluated. The division by the variance **V** in eq. (6) ensures that the lower the disparity of the data at a certain time-point, the larger the weight in the cost function and thus, the more important the goodness of the reproduction at that point.

To determine $\hat{\mathbf{M}}^{Opt}$, a local search is performed, in which all n^2 parameters are first set equal to the of the n^2 elements of $\hat{\mathbf{M}}^{LS}$ and are then released so as to minimize *J*. The algorithm used is a simplex search method of Lagarias *et al.* (1998) (*fminsearch* routine in *Matlab*). A total of 10,000 iterations are performed. The drawback of this method is the risk to be trapped into a local minimum of the cost function. However, thanks to the initialization to the linear parameter estimate, we may reasonably expect that the final solution will close to the absolute minimum.

To compare the expression profiles modeled by the linear and nonlinear parameter estimations, we integrate the differential equations (2) with the estimated matrix $\hat{\mathbf{M}}^{Opt}$, as we did for $\hat{\mathbf{M}}^{LS}$. We thus obtain the estimated gene expression profiles $\hat{\mathbf{x}}^{Opt}(t)$. As shown in Fig. 1, these profiles follow quite well the experimental profiles, much better than the $\hat{\mathbf{x}}^{LS}(t)$ profiles obtained by linear parameter identification. A quantitative evaluation of this improvement is obtained by computing the root mean square deviation $S(\hat{x}_c)$, defined in eq. (4), between the modeled and experimental profiles, which drops from 10.47 for $\hat{\mathbf{x}}^{LS}(t)$ to only 0.07 for $\hat{\mathbf{x}}^{Opt}(t)$.

Table 1. Root mean square deviation $S(\hat{x}_c)$ between experimental and modeled profiles

Cluster *c*	$S_c(\hat{\mathbf{x}}^{LS})$	$S_c(\hat{\mathbf{x}}^{Opt})$
1	3.88	0.03
2	10.47	0.05
3	5.34	0.05
4	28.17	0.13
5	7.78	0.05
6	17.02	0.08
7	6.75	0.05
8	9.19	0.11
9	2.77	0.06
10	0.47	0.03
11	29.92	0.11
12	17.25	0.05
13	8.04	0.07
14	0.14	0.03
15	8.18	0.04
16	0.36	0.02
17	2.12	0.03
$\langle S_c \rangle$	**10.47**	**0.07**

Finally, we computed the relative variation Δ of the parameters before and after the nonlinear estimation stage:

$$\Delta_{\alpha\beta} = \left| \frac{\hat{M}_{\alpha\beta}^{LS} - \hat{M}_{\alpha\beta}^{Opt}}{\hat{M}_{\alpha\beta}^{MC}} \right| \quad , \qquad (7)$$

where α and β are matrix indices. These variations range from 0 to 2%, with a mean value of 0.06%. This shows the high sensitivity of our model, where the small, but specific, parameter variations result in a clear improvement of the reproduction of the experimental expression profiles.

4. CONCLUSIONS AND PERSPECTIVES

The results of the dynamic modeling of the *Drosophila* gene expression profiles are quite encouraging. Indeed, with a simple model structure, where the time evolution of the expression level of one gene cluster is given as a linear combination with constant coefficients of the expression levels of all gene clusters, the scores are impressive: the deviation between the experimental and modeled expression curves is as low as 0.07 on the average.

The power of the 2-step parameter estimation is worth noting, where the first step is analytical and fast but suffers from errors due to time derivative estimations. The parameter values so estimated are used as initial values for the second estimation, nonlinear and more time-consuming, in which all parameters are freed and optimized. The significant improvement when going from the first to the second step is apparent in Fig. 1 and is monitored by a drop in root mean square deviation between experimental and modeled expression profiles from 10.47 to 0.07. Note, moreover, that the second step without the first, that is, without a reliable initial value estimation, is quite less effective and is likely to move the system into a local minimum.

The difference between the parameter values after the first and second estimation steps is very low, showing the large sensitivity of the model, where small parameter variations can provoke large modifications in the profiles.

The estimated matrix $\hat{\mathbf{M}}^{Opt}$ (eq. (5)) encoding the mutual influence of the gene clusters, has no vanishing parameters. This seems to indicate that the gene expression network is highly, and even totally, connected. However, this conclusion is premature, since other parameter sets, with some parameters kept to zero, could possibly model the expression profiles almost as well. This will be analyzed through parameter reduction techniques, with the aim of determining the minimal number of connections between the gene clusters to keep a sufficiently good profile modeling.

The next stage will be to analyze the biological meaning of the modeled network, and to study its transferability to other simple, or less simple, multicellular organisms.

ACKNOWLEGMENTS

We acknowledge support from the Belgian State Science Policy Office through an Interuniversity Attraction Poles Programme (DYSCO), and from the Belgian Fund for Scientific Research (FRS) through an FRFC project. AH benefits from a FRIA grant of the FRS, YD from a First-Postdoc grant of the Walloon region (PROMeTHe) and MR is Research Director at the FRS.

REFERENCES

Arbeitman M.N., Furlong E.M., Imam F., Johnson E., Null B.H., Baker B.S., Krasnow M.A., Scott M.P., Davis R.W. and White D. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270-2275.

Chen T., He H.L. and Church G.M. (1999) Modeling gene expression with differential equations, *Proc. Pacific Sympos. Biocomputing* 29-40.

Forsythe G., Malcolm M. and Moler C. (1977) Computer Methods for Mathematical Computations, Prentice-Hall, New Jersey, USA.

Lagarias J.C., Reeds J.A., Wright M.H. and Wright P.E. (1998) Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions, *SIAM Journal of Optimization* **9**, 112-147.

Ma P., Castillo-Davis C.I., Zhong W. and Liu J.S. (2006) A Data-Driven Clustering Method for Time Course Gene Expression Data, *Nucleic Acids Research* **34**, 1261-1269.