OPEN ACCESS Freely available online

PLOS ONE

Modeling the *Drosophila* Gene Cluster Regulation Network for Muscle Development

Alexandre Haye, Jaroslav Albert, Marianne Rooman*

BioModeling, BioInformatics & BioProcesses Department, Université Libre de Bruxelles, Bruxelles, Belgium

Abstract

The development of accurate and reliable dynamical modeling procedures that describe the time evolution of gene expression levels is a prerequisite to understanding and controlling the transcription process. We focused on data from DNA microarray time series for 20 *Drosophila* genes involved in muscle development during the embryonic stage. Genes with similar expression profiles were clustered on the basis of a translation-invariant and scale-invariant distance measure. The time evolution of these clusters was modeled using coupled differential equations. Three model structures involving a transcription term and a degradation term were tested. The parameters were identified in successive steps: network construction, parameter optimization, and parameter reduction. The solutions were evaluated on the basis of the data reproduction and the number of parameters, as well as on two biology-based requirements: the robustness with respect to parameter variations and the values of the expression levels not being unrealistically large upon extrapolation in time. Various solutions were obtained that satisfied all our evaluation criteria. The regulatory networks inferred from these solutions were compared with experimental data. The best solution has half of the experimental connections, which compares favorably with previous approaches. Biasing the network toward the experimental connections led to the identification of a model that is only slightly less good on the basis of the evaluation criteria. The non-uniqueness of the solutions and the variable agreement with experimental connections were discussed in the context of the different hypotheses underlying this type of approach.

Citation: Haye A, Albert J, Rooman M (2014) Modeling the *Drosophila* Gene Cluster Regulation Network for Muscle Development. PLoS ONE 9(3): e90285. doi:10.1371/journal.pone.0090285

Editor: Magnus Rattray, University of Manchester, United Kingdom

Received September 16, 2013; **Accepted** January 29, 2014; **Published** March 3, 2014

Funding: This research was supported by the Belgian Fund for Scientific Research (FNRS and FRIA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mrooman@ulb.ac.be

Introduction

Dynamical modeling of transcriptional regulation networks is an important goal of systems biology. It holds promise to understand the functioning of these networks as well as their malfunctioning, which can aid rational modification of some targeted properties. This goal is expected to be within reach due to the impressive amount of data generated during the last few years by powerful high-throughput technologies, such as DNA microarrays that provide the simultaneous expression levels of many or even all genes in a cell sample [1,2]. Moreover, time series of DNA microarray data yield information about the evolution of gene expression levels during, for example, the developmental stages of the host organism, the response to external perturbations, or the cell cycle. If these time-dependent data were accurate and numerous enough, they would, in principle, allow the reverse-engineering of the transcriptional regulation network (see e.g. [3–13]). However, the mathematical model structure to be used for that purpose is unknown. Additional issues are the non-uniqueness of the parameters of the model (see e.g. [14]), the usually high level of intrinsic noise of the microarray data, and the impurity of the samples that often contain mixtures of cell types. A possibility to handle the degeneracy of the solutions is to include biology-based constraints in the modeling procedure [13]. One constraint is the robustness of the solutions with respect to parameter variations (see e.g. [13,15–17]). It manages stochasticity and ensures that the overall behavior of biological systems does not vary with changes in the environment, except when large and specific perturbations come into play that lead the system to another state. The second biological constraint consists of requiring that the solutions be stable when extrapolated in time. It is indeed reasonable to assume that although the expression levels may drastically change in time, up to a few orders of magnitude, they do not become unreasonably large.

Another issue is the extremely large size of the transcription regulation network, where basically all genes of an organism are, directly or indirectly, connected. Even when large DNA microarray time series are available, the data are insufficient to identify all parameters of the model structures. Moreover, oftentimes different genes exhibit similar expression profiles, either because they are coregulated or because the noise level does not allow distinguishing them. To solve both of these problems, genes are often clustered into groups, and the modeling procedure is applied on these rather than the individual genes [9,11,13,18]. The disadvantage of this approach is the lack of straightforward physical interpretation of the resulting gene cluster networks. Another approach is to consider the full transcription network of an organism as separate subnetworks that are loosely connected and can be modeled separately to a good approximation [19–22].

We consider in this paper *Drosophila melanogaster* as our model organism, and focus on the subset of 20 genes that are involved in

Table 1. Effect of the preprocessing procedure and clustering algorithm on the quality of the clusters.

Preprocessing	Classification	Intraclass	Intraclass	Interclass	Interclass
procedure	**method**	$<D>$	$<D_{rep}>$	$<D>$	$<D_{rep}>$
Filtering	tree-like	0.43	0.39	1.05	1.06
	k-means	0.41	0.31	1.01	1.08
Smoothing	tree-like	0.33	0.29	1.00	1.00
	k-means	0.29	0.23	0.94	1.00
Filtering &	tree-like	0.31	0.29	0.94	0.99
Smoothing	**k-means**	**0.28**	**0.23**	**0.95**	**1.03**

The number of classes is set to 10. The optimal procedure is indicated in bold. Intraclass $<D>$: average distance between members of the classes; Intraclass $<D_{rep}>$: average distance between members of the classes and their representative member; Interclass $<D>$: average distance between members of different classes; Interclass $<D_{rep}>$: average distance between representative members of different classes.

doi:10.1371/journal.pone.0090285.t001

muscle development during the embryonic stage. This subset has the advantage of being well described and of having available experimental data about the transcriptional interactions. To tackle modeling, we use a combination of the approaches described in the previous paragraph: we disregard the connections with genes outside this network, and cluster the genes that have similar expression profiles into classes. We then proceed to model the dynamical behavior of gene cluster expression, using coupled differential equation. To reduce the number of solutions and select those that have a biological meaning, we impose the robustness and stability constraints described above. The resulting transcriptional networks are compared to the experimental information about the transcription factor-gene interactions.

Methods

Experimental data on *Drosophila* genes involved in muscle development

A total of 20 genes were identified as being involved in *Drosophila* muscle development [3]. These are: CG10293 (how), CG1429 (mef2), CG17927 (mhc), CG18251 (msp-300), CG1915 (sls), CG2096 (flw), CG2328 (eve), CG2956 (twi), CG3992 (srp), CG4376 (actn), CG4677 (lmd/gfl), CG4889 (wg), CG5596 (mlcl), CG5939 (prm), CG7107 (up), CG7438 (myo31DF), CG7445 (fln), CG7895 (tin), CG9155 (myo61F), CG9885 (dpp).

The time-dependent expression profiles of these 20 genes during the embryonic development, relative to their expression in a reference sample containing a standard mixture of cells at all developmental stages, have been experimentally characterized by DNA microarray techniques [23]; they have been deposited in NCBI's Gene Expression Omnibus [24] and are accessible through GEO Series accession number GSE4347 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE4347). The DNA microarray technique [1] proceeds by extracting the mRNA from the cell sample of interest and from the reference sample, reverse transcribing them into cDNA, labeling them by two types of fluorophores, and letting them hybridize to their complementary sequences attached to a microarray. The fluorescence intensities I_μ emitted by the fluorophores from the sample of interest are measured relative to the intensities I_μ^R emitted by the fluorophores from the reference sample; the index μ labels the mRNA molecules (or equivalently, the corresponding genes or proteins). Here we have $\mu = 1,...,20$. These intensities must be normalized to correct for different effects including the unequal quantities of RNA copies, differences in labeling or detection efficiencies between the fluorescent dyes, and systematic biases in the measured expression levels [25,26]. The gene expression levels X_μ are given as the ratio of the normalized intensities $\tilde{I}_\mu$ and $\tilde{I}_\mu^R$, under the commonly made assumption that the RNA concentrations and fluorescence intensities are proportional [27]. Time series correspond to gene expression levels of the sample taken at N different time points τ_i ($i = 1,...,N$). Here the time series contain $N = 31$ time-points and cover the 24 hours of the embryonic development, with varying sampling frequency (every 30 minutes up to time point 14 and then every hour). The time-dependent gene expression profile $X_\mu(\tau)$ is thus defined as:

$$X_\mu(\tau) = \frac{\tilde{I}_\mu(\tau)}{\tilde{I}_\mu^R}. \quad (1)$$

The droID database [28] lists the interactions between genes and gene products. For the 20 genes involved in *Drosophila* muscle development, 36 experimentally proven interactions are listed. These include 34 interactions between transcription factors and genes, and 3 genetic interactions, defined as interactions whose molecular mechanism is unknown or results from a cascade of interactions [29]. These interactions are listed in Table S1 of File S1. This table also contains four new interactions that were unknown when this work was started. They are used for validation purposes. Note that we overlooked protein-protein interactions because most of them are not directly obtained from experiment; rather, they are predicted from results on other species, and are thus less reliable.

Clustering of gene expression profiles

The expression profiles $X_\mu(\tau)$ that have a similar shape are undistinguishable for modeling purposes, and we therefore cluster them into groups. However, these profiles present a high noise level and missing values. To alleviate this drawback, we first preprocess the data. Two methods were tested. The first consists of data filtering with a mobile mean procedure: $X_\mu(\tau_i) \rightarrow X_\mu(\tau_i)/2 + X_\mu(\tau_{i-1})/4 + X_\mu(\tau_{i+1})/4$; when some values are missing in this equation, they are replaced by the neighboring values. The second procedure consists of smoothing, using the cubic splines algorithm csaps of Matlab (The MathWorks Inc., Natick, MA), with parameter value $p = 0.999$ so that the interpolated curve follows very closely the experimental points [9]. To cluster these preprocessed expression profiles, we need to

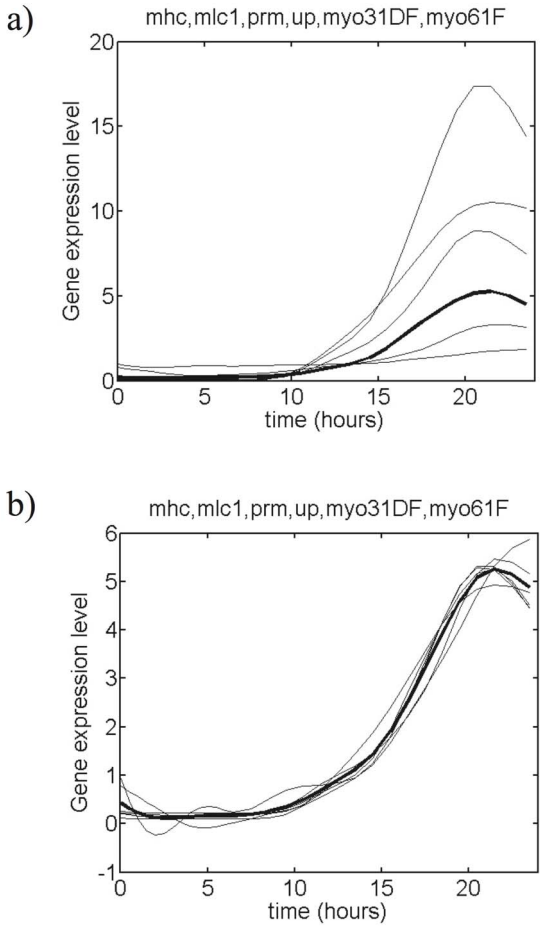

Figure 1. *Drosophila* muscle gene expression profiles belonging to a cluster. The 6 members of the cluster are listed at the top of the figure. The results for all clusters are given in Figs S1-S2 of File S1. (a) Filtered and normalized gene expression profiles contained in the cluster; the representative profile is depicted in bold. (b) Expression profiles superimposed onto the representative profile by translation and scaling; the average profile is depicted in bold.
doi:10.1371/journal.pone.0090285.g001

define a similarity measure and a clustering procedure. Several distances between expression profiles can be defined [30]. Since the expression levels $X_\mu(\tau)$ to be modeled are relative levels with respect to a gene-dependent and time-independent factor eq. (1), no difference should be made between $X_\mu(\tau)$ and $aX_\mu(\tau)$, where a is an arbitrary positive real number. Moreover, we chose not to take into account the difference between two profiles with the same shape but different average expression levels, as such profiles are merely translated with respect to each other. We thus require a symmetric, translation-invariant and scaling-invariant distance measure with zero scaling dimension: $\forall a,b\in\mathbb{R} : D(X_\mu,X_\nu+b)=D(X_\mu,X_\nu), D(X_\mu,aX_\nu)=D(X_\mu,X_\nu)$, and $D(X_\mu,X_\nu)=D(X_\nu,X_\mu)$. The distance satisfying these constraints has the form [30]:

$$D(X_\mu,X_\nu)=\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{|X_\mu(\tau_k)-\langle X_\mu\rangle|}{\zeta_\mu}-\frac{|X_\nu(\tau_k)-\langle X_\nu\rangle|}{\zeta_\nu}\right)^2}, \quad (2)$$

in terms of the mean $\langle X_\mu\rangle$ and standard deviation ζ_μ:

$$\langle X_\mu\rangle=\frac{1}{N}\sum_{k=1}^{N}X_\mu(\tau_k) \quad \text{and} \quad \zeta_\mu=\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[X_\mu(\tau_k)-\langle X_\mu\rangle\right]^2} \quad (3)$$

On the basis of this distance measure, the 20 genes were classified into groups displaying similar expression profiles, using two distinct clustering procedures, the k-means algorithm and a tree-like hierarchical clustering algorithm [31]. The latter proceeds by considering all profiles in separate classes and grouping them two by two, in such a way that the average distance between any pair of profiles in each class is minimum. The procedure is stopped when a threshold distance or a maximum number of classes is reached. Each of the C clusters so obtained, labeled by c ($c=1,...,C$), is represented by its normalized average profile, $\bar{X}_c(\tau)$. To compute this profile, we first identified the representative profile of the cluster, defined as the profile for which the distance with respect to all other members of the class is minimum. All the profiles of the cluster were then superimposed on the representative, using the translation and scaling factors that minimize the distance. The average profile corresponds to the average, at each time point, of all translated and scaled profiles in the cluster. This average profile is then normalized, by scaling and translation, so as to ensure that the new standard deviation of each profile is 1 and that the minimum expression level of all clusters is 0.

Model structures

We assumed the system to be autonomous, and considered coupled differential equations with a transcription term and a degradation term:

$$\dot{\bar{X}}_c(t)=\Theta_c(\bar{X})-\Delta_c(\bar{X})\bar{X}_c(t). \quad (4)$$

where $\bar{X}=(\bar{X}_1,...,\bar{X}_C)$ and t is the real, continuous time. The dot means the derivative with respect to t. Since the transcription term $\Theta_c(\bar{X})$ is defined to be positive, it increases the concentration $\bar{X}_c$ of cluster c, basically through the binding of transcription factors, which either activates or represses genes in this cluster. The positively defined function $\Delta_c(\bar{X})$, called degradation factor,

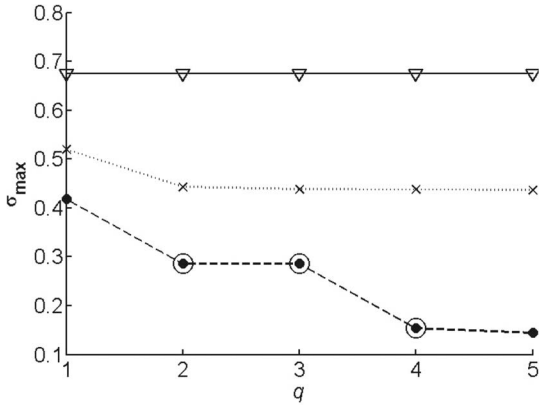

Figure 2. Value of the objective function σ_{max} as a function of the connectivity q, for different model structures. The results obtained with the model structure m_{CN}^{exp} are represented by a solid line with triangles, with m_{NC}^{exp} by a dotted line with crosses, and with m_{NN}^{exp} by a dashed line with dots. The circled points indicate the solutions selected for parameter reduction.
doi:10.1371/journal.pone.0090285.g002

Table 2. Characteristics of the full and reduced solutions using the model structure m_{NN}^{exp}.

Model	q	Solution	σ	σ_{max}	σ_{pert}	χ	NC[1]	PC[2]	AC[3]
	2	full	0.29	0.29	0.43	3.02	20	5/17(29%)	68/83 (82%)
		reduced	0.27	0.27	0.30	13.78	20	5/17(29%)	68/83 (82%)
m_{NN}^{exp}	3	full	0.28	0.29	0.43	3.01	30	7/17(41%)	60/83 (72%)
		reduced	0.28	0.29	0.43	0.95	20	4/17(24%)	67/83 (81%)
		full	0.15	0.15	0.43	3.01	40	8/17(47%)	51/83 (61%)
		reduced	**0.20**	**0.21**	**0.39**	**2.77**	**29**	**8/17(47%)**	**62/83 (75%)**
Biased m_{NN}^{exp}	3	full	0.45	0.46	0.73	1.68	30	17/17(100%)	70/83 (84%)
		reduced	**0.31**	**0.33**	**1.12**	**3.21**	**27**	**17/17(100%)**	**73/83 (88%)**

The last two lines contain the solutions biased towards the experimental network. The networks corresponding to the solutions in bold are depicted in Fig. 4b-c. [1]NC : number of connections in the estimated network; [2]PC: fraction of these connections that are among the 17 experimentally verified connections (see Table S1 in File S1); [3]AC: fraction of the non-connections that are not among the 17 experimentally verified connections (thus that are among the 10×10–17 = 83 experimental non-connections).

doi:10.1371/journal.pone.0090285.t002

describes the degradation, destabilization or activity inhibition of the gene products belonging to cluster c, or their removal from the system. We used the model structure proposed in [13] for the full *Drosophila* gene expression time series, as it showed to be flexible and to lead to good results:

$$m_{NN}^{exp}: \quad \Theta_c(\bar{\mathbf{X}}) = \frac{\lambda_c^+ + \lambda_c^- \exp\left[-\sum_{d=1}^{C} L_{cd}\bar{X}_d(t)\right]}{1+\exp\left[-\sum_{d=1}^{C} L_{cd}\bar{X}_d(t)\right]}, \qquad \Delta_c(\bar{\mathbf{X}}) = \frac{\kappa_c^+ + \kappa_c^- \exp\left[-\sum_{d=1}^{C} K_{cd}\bar{X}_d(t)\right]}{1+\exp\left[-\sum_{d=1}^{C} K_{cd}\bar{X}_d(t)\right]}, \tag{5}$$

with $\kappa_c^+,\ \kappa_c^-,\ \lambda_c^+,\ \lambda_c^- \geq 0$. For defining this structure, it was assumed that the transcription term and degradation factor are modulated by interactions between genes and/or gene products. For the transcription term, these interactions represent the binding of activating or repressing transcription factors as well as the whole cascade of protein-protein interactions occurring before the binding of the transcription factors. For the degradation term, these interactions tend to either prolong (*e.g.* through stabilizing complexes) or shorten (*e.g.* through degradation by proteases) their period of activity. The parameters κ_c^+ and κ_c^- (λ_c^+ and λ_c^-) symbolize the maximum and minimum degradation rate (transcription rate) when $\kappa_c^+ > \kappa_c^-$ ($\lambda_c^+ > \lambda_c^-$) and the converse when $\kappa_c^+ < \kappa_c^-$ ($\lambda_c^+ < \lambda_c^-$), and K_{cd} and L_{cd} give the influence (stabilizing or destabilizing according to their sign) of gene (product) d on gene (product) c.

Two other model structures were also tested, which are particular cases of the first. These are:

$$m_{CN}^{exp}: \Theta_c(\bar{\mathbf{X}}) = \rho_c,\ \Delta_c(\bar{\mathbf{X}}) = \frac{\kappa_c^+ + \kappa_c^- \exp\left[-\sum_{d=1}^{C} K_{cd}\bar{X}_d(t)\right]}{1+\exp\left[-\sum_{d=1}^{C} K_{cd}\bar{X}_d(t)\right]}, \tag{6}$$

which is obtained from (5) by posing $L_{cd} = 0$, $\lambda_c^- = 0$ and $\lambda_c^+ = 2\rho_c$, and

$$m_{NC}^{exp}: \Theta_c(\bar{\mathbf{X}}) = \frac{\lambda_c^+ + \lambda_c^- \exp\left[-\sum_{d=1}^{C} L_{cd}\bar{X}_d(t)\right]}{1+\exp\left[-\sum_{d=1}^{C} L_{cd}\bar{X}_d(t)\right]}, \quad \Delta_c(\bar{\mathbf{X}}) = \gamma_c, \tag{7}$$

obtained from (5) by posing $K_{cd} = 0$, $\kappa_c^- = 0$ and $\kappa_c^+ = 2\gamma_c$.

Parameter identification

The gene expression network was built iteratively by increasing the connectivity q which is defined as the average number of connections ending at a node (or class). In a first stage, the number of connections was considered to be identical for all nodes. The procedure starts by considering $q = 1$, and determines, for each node, the connection (defined by a series of parameters) that minimizes an objective function. It continues by incrementing q until it is large enough to get sufficiently small values of the objective function.

A two-step procedure, based on two different objective functions, was used for parameter identification so as to manage the large amount of parameters and the non-linearity of the equations. The first step consists of constructing the network by reproducing the derivatives of the gene expression levels rather than the gene expression levels themselves. The objective function $\zeta(\mathbf{J})$, where $\mathbf{J}$ denotes generically all parameters of the model, is thus the square root of the square difference of the measured and estimated expression level derivatives, summed over all time points:

$$\zeta(\mathbf{J}) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{k=1}^{N}\left[\dot{\bar{\boldsymbol{X}}}_c(\tau_k) - \dot{\hat{\bar{\boldsymbol{X}}}}_c(\tau_k, J_c)\right]^2}. \tag{8}$$

This entails considering the expression levels and their derivatives as independent variables and reducing the identification to an algebraic problem. Details of this procedure can be found in [13]. In the second step, the connections defined in the first stage for $q = 1, 2, \ldots$ are maintained and the parameters of these connections are identified so as to minimize another objective function, expressed as a function of the difference between measured and estimated profiles rather than their

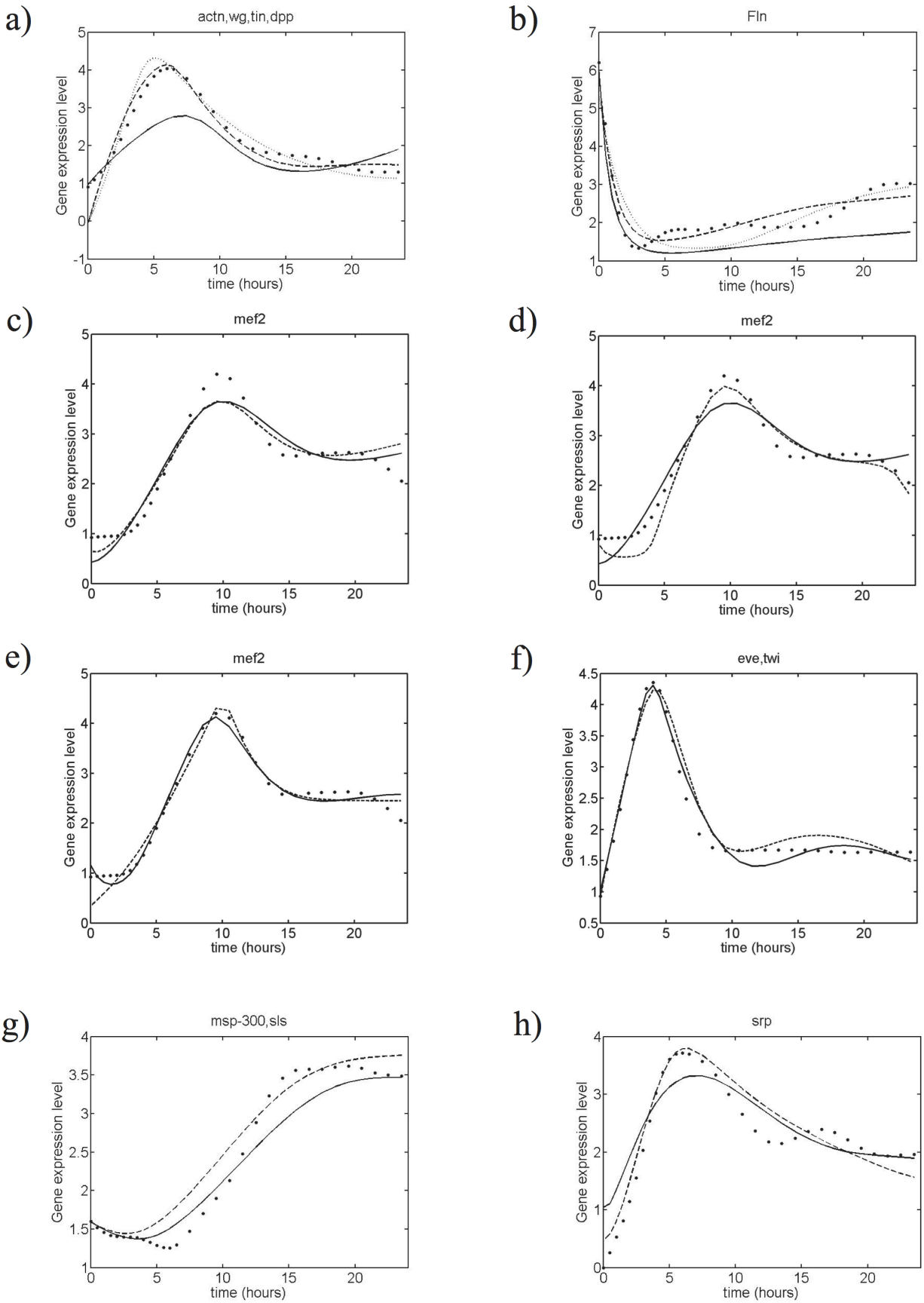
a)
actn,wg,tin,dpp
b)
Fln
c)
mef2
d)
mef2
e)
mef2
f)
eve,twi
g)
msp-300,sls
h)
srp
Gene expression level
time (hours)

Figure 3. Estimated and experimental expression profiles for a few clusters. The results for all clusters are given in Figs S3-S7 of File S1. Dots: clustered, filtered and smoothed experimental data; (a)-(b): Estimated expression profiles for clusters {*actn*, *wg*, *tin*, *dpp*} and {*fln*} using the three model structures; dashed line: m_{NN}^{exp} ; dotted line: m_{NC}^{exp} ; solid line: m_{CN}^{exp}; (c)-(f): Estimated expression profiles for cluster {*mef2*} and {*eve*, *twi*} using the model structure m_{NN}^{exp} and $q=2$ (c), $q=3$ (d) and $q=4$ (e-f); solid line: before parameter reduction; dashed line: after parameter reduction using the Ψ_v procedure; (g)-(h): Estimated expression profiles for clusters {*msp-300*, *sls*} and {*srp*} using the model structure m_{NN}^{exp}, $q=3$, and the biasing procedure towards the experimental network; solid line: before parameter reduction; dashed line: after parameter reduction using the Ψ_v procedure.
doi:10.1371/journal.pone.0090285.g003

derivatives. Two variants of this objective function are considered, $\sigma_{max}(\mathbf{J})$ and $\sigma(\mathbf{J})$. They are defined as:

$$\sigma_{max}(\mathbf{J}) = \max_c \sigma_c(\mathbf{J}) \qquad \text{and} \qquad \sigma(\mathbf{J}) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\sigma_c(\mathbf{J})^2}, \quad (9)$$

where

$$\sigma_c(\mathbf{J}) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[\bar{\boldsymbol{X}}_c(\tau_k) - \hat{\bar{\boldsymbol{X}}}_c(\tau_k,\mathbf{J})\right]^2}. \quad (10)$$

The estimate of the gene expression profiles, $\hat{\bar{\boldsymbol{X}}}_c$, is obtained by integration of the differential equations (4), using one of the model structures given by eqs (5–7), and the ode45 routine of Matlab. The parameters $\mathbf{J}$ are identified so as to minimize either $\sigma(\mathbf{J})$ or $\sigma_{max}(\mathbf{J})$, using Matlab's fmincon optimization algorithm. The initial values of the parameters are set to those obtained for the q-1 identification, with the newly added parameters set to zero.

Parameter reduction

The next step consists of eliminating unnecessary parameters among L_{cd} and K_{cd} which appear in eqs (5–7), while requiring that at least one connection per gene class be kept. We proceed by dropping one parameter at a time at each step in the iteration, according to two criteria:

1) the parameter of smallest absolute value; this procedure is referred to as Ψ_v;
2) the parameter which, when dropped, leads to the smallest increase of σ_{max}; this procedure is called Ψ_σ.

These criteria turned out to be more effective than those based on the Fisher information matrix [13]. After a parameter is eliminated the remaining parameters are optimized again using the local optimization algorithm fmincon from Matlab. The elimination procedure is then reiterated.

Evaluation of the solutions

Four criteria were used to evaluate the quality of the estimated profiles:

1) the number or remaining parameters;
2) the standard deviations σ and σ_{max} between estimated and experimental profiles, defined in eq. (9);
3) the robustness of the solution with respect to perturbations of its parameters; this is estimated by adding to each parameter in turn $\pm 1\%$ of its value, determining which perturbation leads to the largest deviation between measured and estimated expression levels, $|\bar{\boldsymbol{X}}_c(\tau_k) - \hat{\bar{\boldsymbol{X}}}_c(\tau_k)|$, for any cluster c and time point τ_k, and computing the value of the standard deviation σ obtained with this perturbed parameter, denoted σ_{pert};
4) the stability of the solution, evaluated by extrapolating the estimated profiles up to a time τ_{end} and by computing the difference between the average value of the estimated gene expression levels over the measuring period and the extrapolated level:

$$\chi = \sum_{c=1}^{C}\left|\left(\frac{1}{N}\sum_{k=1}^{N}\hat{\bar{\boldsymbol{X}}}_c(\tau_k)\right) - \hat{\bar{\boldsymbol{X}}}_c(\tau_{end})\right|. \quad (11)$$

The time τ_{end} is taken to be 3 times the measured (embryonic) time span.

Results

Clustering of the gene expression profiles

The raw data points representing the gene expression levels $X_\mu(\tau)$ of the 20 genes involved in the embryonic muscle development of *Drosophila*, were first preprocessed to fill in the missing points and to decrease the effect of measurement noise, as described in Materials and Methods. Two procedures were tested, consisting of filtering and/or smoothing. Moreover, given that some of the expression profiles have very similar shapes and are thus basically indistinguishable, we proceed to cluster them into groups. Since the profiles are defined up to a gene-dependent factor (see eq. (1)), the distance used to evaluate the similarity is a translation and scaling-invariant measure of scaling dimension zero, denoted D and defined in eq. (2). Two different classification methods were tested with this distance, *i.e.* k-means and a tree-like clustering algorithm (see Methods section).

To choose the most appropriate clustering and preprocessing method, we computed: 1) the average distance D (defined by eq. (2)) between the members of the same class; 2) the average distance between members of different classes; 3) the average distance D between the representative member of a class (defined in Methods) and the other members; and finally 4) the average distance between the representative members of different classes. To have well-defined classes, the first and third distance measures that correspond to intraclass distances must be as low as possible, while the second and fourth distance measures must be as high as possible.

The results are given in Table 1 for classifications into 10 classes and in Table S2 in File S1 for 5–15 classes. Preprocessing the data by successively filtering and smoothing decreases all the distances in general, and decreases even more the intra- than the interclass distances. We thus selected this preprocessing procedure. The lowest intraclass distances and the highest interclass distances are sometimes obtained with the tree-like clustering procedure and sometimes with k-means, depending on the number of classes. However, k-means performs more often slightly better and we thus selected it as clustering procedure. The choice of the number of

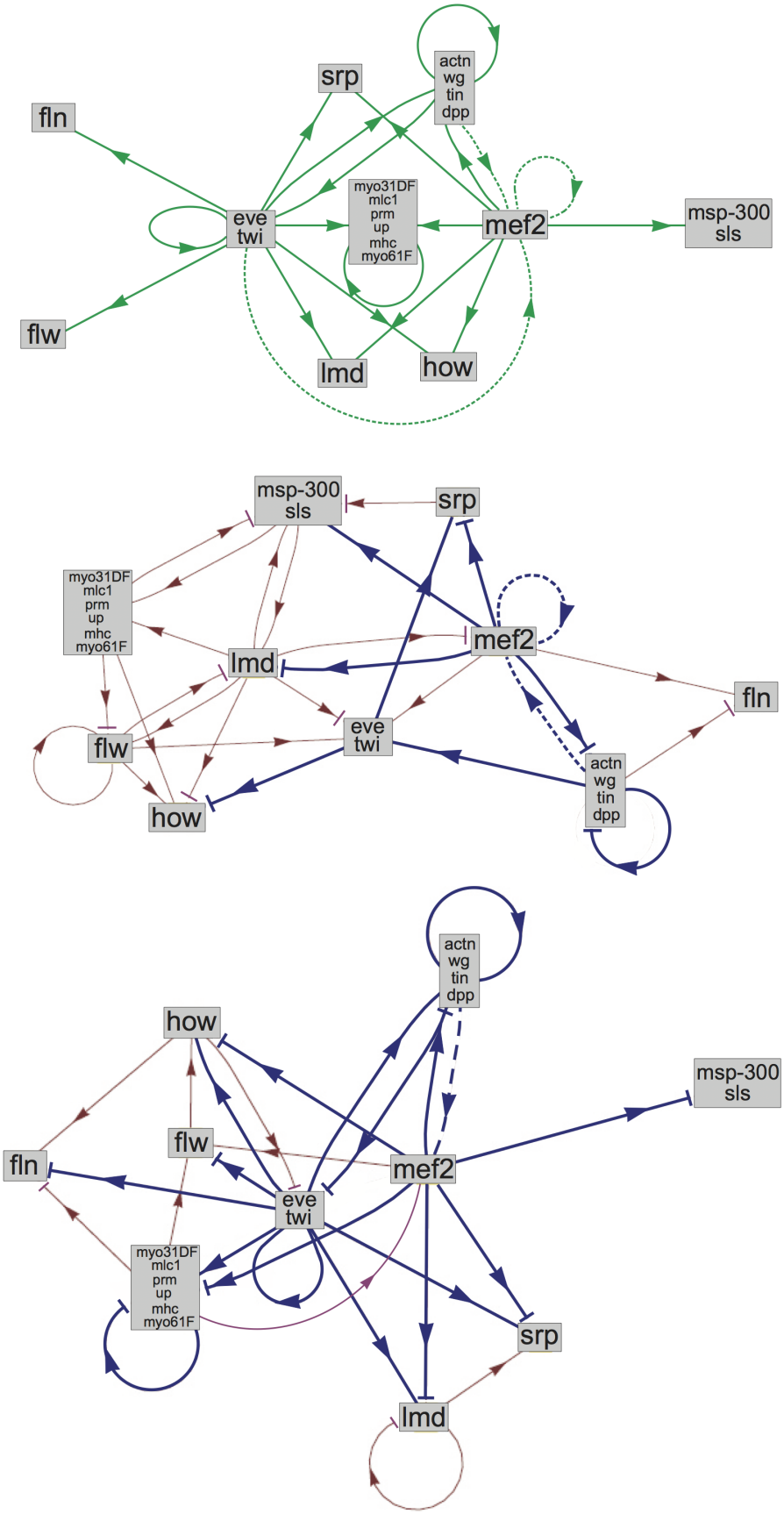
srp
actn
wg
tin
dpp
fln
myo31DF
mlc1
prm
up
mhc
myo61F
eve
twi
mef2
msp-300
sls
flw
lmd
how
msp-300
sls
srp
myo31DF
mlc1
prm
up
mhc
myo61F
mef2
lmd
fln
eve
twi
flw
actn
wg
tin
dpp
how
actn
wg
tin
dpp
how
msp-300
sls
flw
fln
mef2
eve
twi
myo31DF
mlc1
prm
up
mhc
myo61F
srp
lmd

Figure 4. Gene regulation networks. (a): Experimental connections; the dashed connections correspond to the 3 connections that were unknown when this work was performed; (b): Network obtained with the model structure m_{NN}^{exp} and $q=4$, after parameter reduction using the Ψ_v procedure; (c): Network obtained with the model structure m_{NN}^{exp}, $q=3$ and the biasing procedure towards the experimental network, after parameter reduction; (b-c): the connections in blue are the experimental connections that have been predicted; the blue dashed connections correspond to the new experimental connections that have been predicted.
doi:10.1371/journal.pone.0090285.g004

classes is somewhat arbitrary as we do not see a gap in the intra- and interclass distances when decreasing the number of classes (Table S2 in File S1).

To choose the total number of classes, we were guided by: the concern for having basically indistinguishable profiles in the same class; different profiles in different classes; and a sufficiently small number of classes to ensure that the parameters can be reliably identified from the available data. The analysis of Table S2i File S1 indicates that the average distance between members of the classes falls below 0.3 when clustering into 10 classes and more. A visual inspection of the superimposed expression levels in the classes confirmed that these 10 classes are well-defined (see Fig. S1 in File S1). We thus fixed the number of classes to 10. This classification grouped together the genes {*mhc*, *mlc1*, *prm*, *up*, *myo31DF*, *myo61F*}, {*msp-300*, *sls*}, {*actn*, *wg*, *tin*, *dpp*} and {*eve*, *twi*}; all other classes contain a single gene.

The representative and other members of these 10 classes are shown in Fig. 1 and in Fig. S1 of File S1. Each of these clusters labeled by c ($c=1,...,10$) is represented by its normalized average profile, $\bar{X}_c(\tau)$, which is defined in the Methods section and is depicted in Figs. 1, and S1 and S2 in File S1.

Dynamical modeling of gene expression profiles

The time-evolution of the 10 normalized average expression levels, $\bar{X}_c(\tau)$, of the 20 genes involved in muscle development was modeled using an autonomous model structure (eq. (4)) with three versions of the transcription term and degradation factor given by eqs (5–7).

Because of the large number of parameters and the nonlinearity of the equations, it is impossible to have a reliable direct identification of all the parameters that define all the possible connections between genes. So, assuming that the real gene expression network is sparse, we first determine the necessary connections, assuming a constant connectivity q for all nodes (see section 2.4 for details). We start from $q=1$ and increase it until the value of the objective function is sufficiently small. Moreover, we use successively two objective functions, denoted ζ and σ, defined in eqs (8–9). The first is given as a function of the difference between the derivatives of the experimental and estimated profiles, and the second as a function of the difference between the experimental and estimated profiles. The first is used to define the important connections and the second to optimize the parameters.

Moreover, two variants of the latter objective function are used: σ which is the standard deviation between estimated and experimental profiles averaged over all classes, and σ_{max} which is the largest standard deviation of all classes (eq. (9)). Using the former has the drawback that some classes may be modeled very well and others very poorly. Using the latter ensures that all classes have σ_c values lower than σ_{max}, and gives thus slightly more homogenous and satisfactory results. Hence we keep σ_{max} as the objective function.

The evolution of σ_{max} during the network construction, from connectivity $q=1$ to $q=5$, is shown in Fig. 2. Clearly, the model structure that has the largest number of parameters, m_{NN}^{exp}, in which neither the transcription term nor the degradation factor is constant, is superior to the other two structures. For this structure, the minimal connectivity required (for which the σ_{max} value is reasonably low) is equal to 2, but a connectivity of 4 yields much lower values of σ_{max}. Figs. 3a-b and Fig. S3 in File S1 illustrate the superiority of m_{NN}^{exp}: only this structure allows the correct reproduction of the data. We restrict ourselves to this model structure and connectivity in the following.

To get rid of the unnecessary connections, we proceed to parameter elimination. Two methods were tested: Ψ_v and Ψ_σ, where the eliminated parameters are those of smallest absolute value or those that lead to the smallest increase of the objective function σ_{max} (see Methods). Note that when several parameter eliminations lead to the same value of σ_{max}, the one that increases σ the least was chosen. The elimination procedure was performed until a threshold valued of σ_{max} was reached. This threshold was set at 0.3, by visual inspection, so as to ensure a fair reproduction of the experimental profiles. Moreover, in order to mimic biological reality, we selected solutions that were robust against perturbations of the parameters; in particular we required the standard deviation $\sigma_{pert}<0.5$ (see Methods). We did not put a threshold on the stability of the solution, estimated by χ (eq. (10)).

Note that these two characteristics, robustness and stability, are quite important for modeling biological systems. Indeed, all such systems have a stochastic behavior that depends, among others, on changes in the environment, the amount of biomolecules, their possible binding and function. However, these changes do not affect the main properties of the system, which continues to give similar responses to similar stimuli. Only very large or very specific perturbations can bring the system out of its correctly functioning state and lead it to another state, which can be functional or dysfunctional, depending on the perturbation. It is thus very important that the models that simulate biological systems have the same properties, and thus do not yield very different solutions for similar parameter values. The other characteristic of biological systems is its stability. Even though the available data usually cover only a part of the system's life, it is reasonable to assume that the expression levels continue to be of the same order of magnitude, never becoming unrealistically large or negative. The same property is expected to be built in the model: the solutions must take realistic values until the next perturbation or developmental stage, or the end of the organism's life.

The results of the elimination procedures Ψ_v and Ψ_σ for $q=2$–4 are given in Table 2 and Table S3 in File S1. These two procedures gave comparable results for the data reproduction, and Ψ_v gives usually better results for the stability and robustness, especially for $q=4$. We will thus in the following only detail the results obtained with Ψ_v.

The estimated profiles are shown in Figs 3c-f and Figs S4-S6 in File S1. The number of connections of the reduced solutions vary between 20 (for $q=2$–3) and 29 (for ($q=4$) and, accordingly, the reproduction of the experimental profiles is somewhat better for $q=4$ (σ and $\sigma_{max}=0.2$) than for $q=2$–3 (σ and $\sigma_{max}=0.3$). Note that 20 seems to be the minimum number of connections needed: the reduction procedure applied on the $q=2$ solution, with 20 initial connections, fails to eliminate further connections. All the solutions show a fairly good robustness with respect to the variations of the parameters, with values of σ_{pert} between 0.3 and 0.4. In contrast, the stability upon extrapolation in time differs

among the solutions. The best reduced solution is that obtained from $q = 3$ ($\chi = 0.9$), followed by the $q = 4$ solution ($\chi = 2.8$); the $q = 2$ solution is not stable at all ($\chi = 13.8$).

Comparison with the experimental network

The connections between gene clusters obtained by our dynamical modeling procedure can be compared to the experimentally determined connections, described in section 2.1 and Table S1 in File S1. For this comparison, we first have to transform the experimental network between individual genes into a network between gene clusters. This is done by considering two clusters as being connected if at least one member in each cluster is connected to at least one member in another cluster. The experimental cluster network so defined contains 17 connections; it is shown in Fig. 4a. Note that these interactions are oriented but unsigned, as the sign is not experimentally determined.

The intersection between the experimental and modeled networks is indicated in Table 2. The smallest intersection is obtained with the reduced $q = 3$ solutions (only 4 out of 17 reproduced connections), whereas the largest intersection is obtained with the full and reduced $q = 4$ solutions (8 out of 17, which amounts to 47%). The number of experimentally non-observed connections that are also absent in the modeled solutions is of course much larger (between 61% and 82%).

The number of common connections between the experimental network and the different solutions can be compared to the number of common connections that are expected at random. The most significant result is found for the $q = 4$ reduced solution: the probability of finding 8 common connections among two sets of 29 and 17 connections each, out of a total set of 100 connections, is equal to 0.048. The same value is found for the absent connections.

The network corresponding to this best solution ($q = 4$ reduced solution) is depicted in Fig. 4b, with the correctly reproduced connections highlighted. Among the 8 connections that are in agreement with experiment, four involve the gene *mef2* (myocyte enhancing factor 2). In particular, the subnetwork involving the genes and gene clusters *mef2*, *srp*, *lmd*, *how*, {*eve*, *twi*} and {*actn*, *wg*, *tin*, *dpp*} is in good agreement with experiment.

Note moreover that all existing functional connections have not yet been determined experimentally. Therefore some of the predicted connections may in fact be real ones. This is indeed the case: among the four new connections that were unknown when this work was performed (indicated in Table S1 in File S1), which correspond to three new connections between clusters (see Fig4a), two were actually predicted, as shown in Fig. 4b. These correctly predicted connections involve *mef2*, which supports the conclusion that this region of the network is well reproduced by our model. Adding these new connections increases the number of correctly predicted connections to 10, out of a total of 20 experimental connections.

Our gene regulation network connects gene clusters rather than individual genes and is thus of a different type than the networks obtained with other methods. Nevertheless, the fraction of correctly predicted connections, either between genes or gene clusters, can be compared among the different methods applied to the same ensemble of *Drososphila* muscle development genes [19,20]. These methods do not reach our 50% score.

Biasing towards the experimental network

To analyze if it is possible to find solutions that reproduce the data well but are different from those obtained in the previous section and are more consistent with the experimental data, we perform a biased modeling procedure. This procedure follows the same two steps: first the construction of the network from $q = 1$ to higher q by minimizing the cost function ζ (eq. (8)), and then the minimization of the cost function σ_{max} (eq. (9)) while keeping the same network. However, here, instead of allowing a free choice among all possible connections, the choice was biased towards the experimentally proven connections: if, for a given cluster, experimental connections do exist and have not yet been included in the network in a previous step, the choice is limited to those; otherwise the choice is free. Moreover, in the parameter reduction procedure, parameters involved in the experimental connections may be eliminated but the connection may not be dropped entirely.

The results obtained with this procedure are given in the last two lines of Table 2 and in Figs 3g-h and Fig S7 in File S1. The reproduction of the expression profiles is somewhat less accurate than with the unbiased method but remains good, with $\sigma = 0.4$ and 0.3 for the full and reduced solutions starting at $q = 3$. Note that the optimization of the solutions performs less well with some imposed connections, as σ is higher for the full than for the reduced solution. The robustness of the reduced solution is also somewhat less good than for the unbiased procedure ($\sigma_{pert} = 1.1$), whereas the stability is similar. Note that the total number of connections in the reduced solution is equal to 27, and that we had to add 10 connections in addition to the 17 experimental ones to reach a reasonable accuracy in the profile reproduction. However, the number of 27 connections is comparable to the number of connections obtained with the unbiased procedure (20–29). Note that among the three new experimental connections that were not imposed in this procedure, one appears to be correctly predicted.

Conclusion

One successful result of our work is the consistent construction of dynamical models, on the sole basis of the gene expression profiles of the genes involved in *Drosophila* muscle development obtained from DNA microarray series. The models obtained reproduce the expression profiles quite well, are robust against parameter variation and do not take unrealistically large values when extrapolated in time. However, our results present two important drawbacks. First, the solutions are not unique, and different networks are obtained with very similar data reproduction abilities. The additional requirement of robust and stable solutions filters out some of them, but the number of acceptable solutions remains high. Second, half of the experimental connections are obtained by the best of our unbiased models. To obtain all experimental connections, we had to bias the model construction towards the experimental network.

The amount of 50% of the experimental connections found by our models compares quite favorably with the results of other analyses. However, our solution is still far from perfect and it is worthwhile to question the basics of our approach. We made a number of assumptions that, although commonly made in biological modeling, could explain the limited overlap between estimated and experimental connections. These assumptions are detailed hereunder.

- We considered only the 20 genes known to be involved in muscle development of *Drosophila*. In reality, these genes are connected to other genes. We suppose here that these additional connections are not (or much less) important for the transcriptional regulation of these 20 genes. More generally, we disregarded the effect of all external factors on the regulation of these 20 genes, which is quite a bold (but common) assumption.
- As some of these 20 genes have similar experimental expression profiles, which are moreover quite noisy, we prepro-

cessed the data by filtering and smoothing them, and grouped the similar profiles together using a k-means classification procedure and a translation-invariant and scaling-invariant distance measure. Although we tested several preprocessing and classification procedures and although the selected ones appear to perform quite well, we cannot be sure that the noise is eliminated in the right way, and that the clusters are formed adequately.

• As some genes were clustered together, the transcriptional model we derive connects gene clusters instead of individual genes. The interpretation is that when a cluster is found to regulate another cluster, some of their members do so, but not necessarily all. It is indeed possible that the different members of a cluster – even though their expression profiles are similar – are not regulated by the same gene (cluster). The different – almost equivalent – solutions in terms of data reproduction and robustness that we found could well reflect the differences in connections between individual members of the clusters.

• Another possibility is that the DNA microarray data, and thus the networks predicted from them, correspond to external conditions that differ from those of the experimental inter-gene connections. It has for example been shown that networks may be different when responding to different types of stress [32].

• The model structure we selected is quite flexible and gives good results in terms of data reproduction; it could however be argued that it does not mimic the biological mechanism and that another model structure should be used.

• The experimentally determined interactions are listed as regulatory interactions. However, some of them could be involved only indirectly in regulation, and others could be side actors. Moreover, not all interactions are known today, and some of the predicted interactions – or of the non-predicted ones – will perhaps be experimentally demonstrated in the future.

It is difficult at this point to identify the reason for the limited – though substantial – overlap between experimental and estimated connections and of the large number of almost equivalent solutions. Note that the latter result could be taken as a positive result that mimics reality. Indeed, gene regulation networks have been shown experimentally to display some elasticity [33].

The results of our analysis indicate that more extensive and more specific experimental data is needed to decide between the different hypotheses. For example the existence of connections predicted by our models, depicted in Fig. 4b-c, should be verified experimentally. In particular the additional connections involving the *mef2* gene in Fig. 4b, which is at the center of a quite well predicted subnetwork, are interesting candidates to be tested.

Supporting Information

File S1 Supporting tables and figures. Table S1: The interactions between the 20 genes involved in muscle development that have been observed experimentally. Table S2: Effect of the clustering algorithm and the number of clusters on the quality of the clusters. Table S3: Characteristics of the full and reduced solutions using the model structure m_{NN}^{exp} and the reduction procedure Ψ_σ. Figure S1: The four clusters of Drosophila muscle gene expression profiles containing more than one member. Figure S2: The average profile of the ten clusters. Figure S3: Experimental and estimated gene expression profiles, with $q = 3$. Figure S4: Experimental and estimated gene expression profiles for model m_{NN}^{exp} with $q = 2$ before and after parameter reduction using the Ψ_v procedure. Figure S5: Experimental and estimated gene expression profiles for model m_{NN}^{exp} with $q = 3$ before and after parameter reduction using the Ψ_v procedure. Figure S6: Experimental and estimated gene expression profiles for model m_{NN}^{exp} with $q = 4$ before and after parameter reduction using the Ψ_v procedure. Figure S7: Experimental and estimated gene expression profiles for model with $q = 3$ before and after parameter reduction using the Ψ_v procedure, when the 17 experimentally validated connections are imposed.
(PDF)

Author Contributions

Conceived and designed the experiments: AH JA MR. Performed the experiments: AH. Analyzed the data: AH JA MR. Wrote the paper: MR.

References

1. Page GP, Zakharkin SO, Kim K, Mehta T, Chen L, et al. (2007) Microarray analysis. Meth Mol Biol 404: 409–430.
2. Dufva M (2009) Introduction to microarray technology. Methods Mol Biol 529: 1–22.
3. Bar-Joseph Z (2004) Analyzing time series gene expression data. Bioinformatics 209: 2493–503.
4. Wu X, Dewey TG (2006) From microarray to biological networks: Analysis of gene expression profiles. Methods Mol Biol 316: 35–48.
5. Androulakis IP, Yang E, Almon RR (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. Ann Rev Biomed Eng 9: 205–228.
6. Goutsias J, Lee NH (2007) Computational and experimental approaches for modeling gene regulatory networks. Curr Pharm Des 13: 1415–1436.
7. Kramer R, Xu D (2008) Projecting gene expression trajectories through inducing differential equations from microarray time series experiments. J Signal Process Syst 50: 321–329.
8. Sima C, Hua J, Jung S (2009) Inference of gene regulatory networks using time-series data: a survey. Curr Genomics 10: 416–429.
9. Haye A, Dehouck Y, Kwasigroch JM, Bogaerts P, Rooman M (2009) Modeling the temporal evolution of the drosophila gene expression from DNA microarray time series. Phys Biol 6: 016004.
10. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R (2009) Gene regulatory network inference: Data integration in dynamic models-a review. BioSystems 96: 86–103.
11. Albert J, Rooman M (2011) Dynamic modeling of gene expression in prokaryotes: application to glucose-lactose diauxie in escherichia coli. Synthetic Syst Biol 5: 33–43.
12. Hempel S, Koseska A, Nikoloski Z, Kurths J (2011) Unraveling gene regulatory networks from time-resolved gene expression data-a measures comparison study. BMC Bioinformatics 12: 292.
13. Haye A, Albert J, Rooman M (2012) Robust non-linear differential equation models of gene expression evolution across drosophila development. BMC Res Notes 5: 46.
14. Konopka T, Rooman M (2010) Gene expression model (in)validation by fourier analysis. BMC Syst Biol 4: 123.
15. Zhou T, Carlson JM, Doyle J (2002) Mutation, specialization, and hypersensitivity in highly optimized tolerance. Proc Natl Acad Sci USA 99: 2049–2054.
16. Stelling J, Sauer U, Szallasi Z, Doyle FJ 3rd, Doyle J (2004) Robustness of cellular functions. Cell 118: 675–85.
17. Kitano H (2007) Towards a theory of biological robustness. Mol Syst Biol 3: 29.
18. D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16: 707–726.
19. Zhao W, Serpedin E, Dougherty ER (2006) Inferring gene regulatory networks from time series data using the minimum description length principle. Bioinformatics 22: 2129–2135.
20. Guo F, Hanneke S, Fu W, Xing EP (2007) Recovering temporally rewiring networks: A model-based approach. In: Proceedings of the International Conference on Machine Learning. pp. 321–328.
21. Li P, Zhang C, Perkins EJ, Gong P, Deng Y (2007) Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 7: S13.
22. Müller HJ, Chiou JM, Leng X (2008) Inferring gene expression dynamics via functional regression analysis. BMC Bioinformatics 9: 60.
23. Arbeitman MN, Furlong EM, Imam F, Johnson E, Null BH, et al. (2002) Gene expression during the life cycle of drosophila melanogaster. Science 297: 2270–2275.
24. Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Research 30: 207–210.

25. Quackenbush J (2002) Microarray data normalization and transformation. Nature Genetics 32: 496–501.
26. Bolstad BM, Irizarry RA, Åstrand M (2003) Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
27. Motakis ES, Nason GP, Fryzlewicz P, Rutter GA (2006) Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. Bioinformatics 22: 2547–2553.
28. Yu J, Pacifico S, Liu G, Finley RL (2008) DroID: the Drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics 9: 461.
29. Ingold A (2002) Couplage entre les bases de données factuelles et bases de données bibliographiques: Identification dans Medline des gènes décrits dans Flybase et application à l'extraction d'informations sur les interactions génétiques ou moléculaires à partir de publications. Ph.D. thesis, Université Aix-Marseille 3.
30. Rooman M, Albert J, Dehouck Y, Haye A (2011) Detection of perturbation phases and developmental stages in organisms from DNA microarray time series data. PLoS One 6: e27948.
31. Hartigan JA (1975) Clustering algorithms. Hoboken, NJ: John Wiley & Sons Inc.
32. Hickman R, Hill C, Penfold CA, Breeze E, Bowden L, et al. (2013) A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves. The Plant Journal 75: 26–39.
33. Krishnan A, Giuliani A, Tomita M (2007) Indeterminacy of reverse engineering of gene regulatory networks: The curse of gene elasticity. PloS ONE 6: e562.