Mouli Das*, Chivukula A. Murthy and Rajat K. De

Second order optimization for the inference of gene regulatory pathways

Abstract: With the increasing availability of experimental data on gene interactions, modeling of gene regulatory pathways has gained special attention. Gradient descent algorithms have been widely used for regression and classification applications. Unfortunately, results obtained after training a model by gradient descent are often highly variable. In this paper, we present a new second order learning rule based on the Newton's method for inferring optimal gene regulatory pathways. Unlike the gradient descent method, the proposed optimization rule is independent of the learning parameter. The flow vectors are estimated based on biomass conservation. A set of constraints is formulated incorporating weighting coefficients. The method calculates the maximal expression of the target gene starting from a given initial gene through these weighting coefficients. Our algorithm has been benchmarked and validated on certain types of functions and on some gene regulatory networks, gathered from literature. The proposed method has been found to perform better than the gradient descent learning. Extensive performance comparison with the extreme pathway analysis method has underlined the effectiveness of our proposed methodology.

Keywords: flow vector; morphogenesis; node-edge incidence matrix; segment polarity gene; transcription factor.

***Corresponding author: Mouli Das,** PhD Student, Machine Intelligence Unit (MIU), Indian Statistical Institute (I.S.I), 203 B. T. Road, Kolkata – 700108, West Bengal, India, Phone: (+91) (33) 2575 2204 (Lab), (+91) (33) 2575 3112 (Lab), (+91) (33) 9331975199 (M), Fax: (+91) (33) 2578 3357, e-mail: mouli.das@gmail.com
Chivukula A. Murthy and Rajat K. De: Machine Intelligence Unit (MIU), Indian Statistical Institute (I.S.I), 203, B. T. Road, Kolkata 700 108, India

1 Introduction

Living organisms behave as complex systems that are flexible and adaptive to their surroundings. At the cellular level, organisms function through intricate networks of chemical reactions and interacting molecules. The best characterized among these biochemical networks are metabolic pathways, the biological networks that involve enzymatic reactions of chemical compounds (Stelling, 2004). Regulatory pathways are another class of pathways that represent transcriptional regulation. The concept of a gene regulatory pathway (GRP) arises from the regulatory interactions between genes and gene-products. It is basically a sequence of transcription events, from a starting gene to a target one, forming a connected path controlling the process of gene expression of the target gene. Pathways are the key to understanding how an organism reacts to perturbations in its environment (e.g., heat shock, chemical or hormone stimulus) or internal changes (e.g., disease, development, etc.).

Understanding of the gene regulation mechanisms is currently one of the most important tasks for systems biology (Schlitt and Brazma, 2007). Gene regulatory networks (GRNs) are composed of genes, proteins, metabolites and signaling components. GRNs are concerned with the control of transcription, i.e., how genes are up and down regulated in response to signals. These processes are mainly effected by special proteins with regulatory function, called transcription factors (TFs). These regulatory proteins (TFs) can act either as activators, which means they increase the rate of expression of the genes, or as repressors that decrease the rate of expression of the regulated genes. Several approaches have been developed that enable the experimental or computational identification of interactions between genes and TFs.

These interactions give rise to a huge GRN for a particular tissue. A GRN may contain several paths from a particular gene to the other. The problem is to identify an optimal pathway through which the transcription of the target gene becomes maximum. Identifying optimal GRPs from GRNs is now an area of extremely active research (Tu et al., 2006). New experimental technologies now make it possible to quickly obtain vast amounts of data on regulatory pathways in a particular organism under particular conditions (Crombach and Hogeweg, 2008).

Commonly, first order backpropagation (BP) by gradient descent (GD) and evolutionary algorithms (EAs) are used for learning the functional and structural parameters of GRNs (Lee and Tzou, 2009; Sirbu et al., 2010). Here we propose a second order approach, a variant of the first order GD computation technology, for global optimization to infer GRPs from known GRNs. In contrast with the GD optimization rule our method is independent on the learning parameter (Yeh and Cheng, 2010). This second order optimization approach is based on modified Newton's method for multiobjective optimization (Bortoletti et al., 2003). The proposed algorithm overcomes the drawbacks of the standard GD algorithm such as slow asymptotic convergence rate, bad controllability of convergence accuracy, local minimum problems and high sensitivity to learning constant (Magoulas et al., 1999). Under this framework, we determine optimal regulatory pathways through certain weighting coefficients representing TFs. A set of flow vectors is estimated based on biomass conservation constraint. A set of constraints is formulated incorporating weighting coefficients. The method calculates the maximal expression of the target gene starting from a given initial gene through these weighting coefficients. This new learning method is applied to illustrate its better performance than the GD method for various types of benchmark functions. Several GRNs such as the *Arabidopsis thaliana* regulatory network, the *Drosophila melanogaster* regulatory network, the budding yeast cell cycle network and the T helper (Th) regulatory network are also considered. These applications and the corresponding validation results suggest that our method is generally applicable to a broad variety of problems. This new modeling framework for GRPs can reveal emergent properties of the GRNs.

The paper is organized as follows. In the next Section 2 we establish our second order learning methodology as well as related concepts that we use throughout the paper. This is followed by the main results concerning derivation of optimal regulatory pathways in Section 3. We then illustrate the results for certain functions as well as for four different types of GRNs. Herein we also make a comparison of our method with the existing extreme pathway analysis (EPA) method and the GD optimization technique. In the subsequent Section 4 we provide a biological validation of the results discussed. We close with conclusions and perspectives in Section 5.

2 Method

Flux balance analysis (FBA) is a mathematical approach for studying biochemical networks, in particular metabolic networks (Benyamini et al., 2010). FBA has also been extended for modeling the qualitative behavior of GRNs reconstructions by controlling the TF level to capture many interesting biological properties as shown in Garg et al. (2007). These network reconstructions contain all of the known regulatory reactions in an organism and the genes that encode each enzyme. FBA calculates the flow of genes through this GRN, thereby making it possible to predict the growth rate of an organism or the rate of expression of a biotechnologically important gene. Our model is a new learning rule developed under the framework of FBA and is based on Newton's method (Wang and Lin, 1998). In our approach we identify optimal regulatory pathways in GRNs. Gradient-based methods are one of the most widely used error minimization methods used to train BP networks. A novel second-order learning algorithm has been developed in this article, which represents an attractive alternative to standard BP algorithms. The fast learning speed and high convergence accuracy of the proposed learning algorithm has also been discussed here (Castillo et al., 2006).

2.1 Mathematical model

Here we briefly give an overview of the structure and dynamics of GRNs. The GRNs are described by directed graphs with nodes corresponding to genes and edges to regulatory interactions (Xiong et al., 2004b). Genes with outgoing edges are the source genes. For a given source gene, we call the set of all genes with incoming edges from that source gene its target genes. A system boundary can be drawn around a GRN comprising of internal flows operating inside the network and exchange flows which allows a gene to enter or exit the system boundary. Let **g** denote the expression levels of the genes in the network and **f** denote the vector of non-linear functions. Thus, g_i denotes the expression level of gene *i* associated with node *i* in the graph. There is a flow, associated with each directed edge (*i, j*) from node *i* to node *j*, which indicates the flow of mRNA and thereby protein obtained from gene *i* transported through the edge (*i, j*). This protein now binds to gene *j* and regulates its expression level. Rate equations indicating the change of expression levels of the genes over time is (Xiong et al., 2004a)

$$d\mathbf{g}/dt = f(\mathbf{g}, \mathbf{u}), \tag{1}$$

where **u** is the set of transcriptional perturbations. For a small value of **u**, the non-linear equation (1) is transformed to the linear equation

$$d\mathbf{g}/dt = \mathbf{AX}, \tag{2}$$

where **A** is the regulatory coefficient matrix and $\mathbf{X}=[\mathbf{g}^T, \mathbf{u}^T]^T$.

GRNs are described by the $m \times n$ node-edge incidence matrix **B** (*m* is the number of genes and *n* is the total number of regulatory interactions, comprising of both internal and exchange flows). An element e_{ik} of matrix **B** is "–1" ("+1") if *k*-th edge (interaction) exits (enters) the node corresponding to gene g_i. Otherwise, $e_{ik}=0$. The matrix **B** plays an important role in the modeling of GRPs. Here we assume the rank of the node-edge incidence matrix **B** is equal to *m*. To define a GRP it is necessary to decompose the matrix **A** to

$$\mathbf{A}=\mathbf{BY}, \tag{3}$$

where **Y** is the conductance matrix which indicates the concept of regulatory flow in a GRP. As $n>m$, equation (3) may not be a unique one. So we consider $\mathbf{V}=\mathbf{YX}$, where **V** is the vector of flows comprising both internal and exchange flows. The matrices **B** and **V** define a homogeneous system that specifies the balance of flows at each gene. Therefore substituting equation (3) in equation (2) leads to the equation (Xiong et al., 2004b).

$$d\mathbf{g}/dt = \mathbf{BV}, \tag{4}$$

At steady state, the time derivative of **g** is set to zero and the above equation (4) reduces to

$$\mathbf{BV} \approx 0, \tag{5}$$

which indicates the flow balance equations for the network.

2.1.1 System definition

A simple GRN is shown in Figure 1 to illustrate the methodology. Here we consider g_P as the source gene and g_Q to be the target gene. The internal flows are denoted by *v* and the exchange flows are denoted by *b*. The GRN considered here consists of $R_1, R_2, \ldots, R_s$ interactions through *s* different paths from the starting gene g_P to the target gene g_Q. A system boundary can be drawn around all these types of reactions. A GRP determines the route of gene regulation that leads from the transcription of a particular starting gene to the transcription of another given target gene through a sequence of transcription events of the intermediate genes forming a connected path in the network. The rate of growth of the target gene g_Q on the starting gene g_P which needs to be maximized is obtained by taking algebraic sum of the weighted flows of reactions and is given by

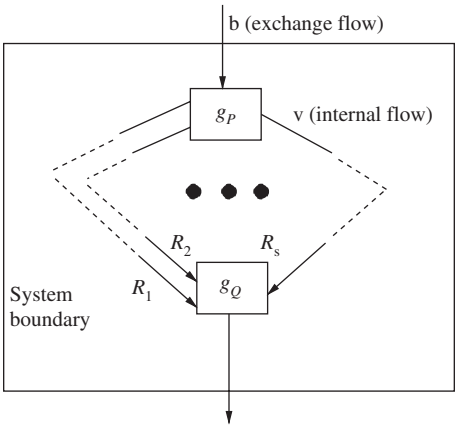

Figure 1 A simple gene regulatory network. The three dots indicates the continuation of the biochemical reactions from R_1 to R_s involving s different paths to reach the target gene. The reactions R_1, R_2 and R_s, involving the target gene, are shown in the diagram.

$$z=\sum_{k=1}^{s} c_k v_k. \tag{6}$$

Here v_k is the k-th internal flow of the reaction R_k involving only the gene g_Q. c_k denotes the weighting factor representing the concentration of other TFs.

2.1.2 Algorithm outline

A given node-edge incidence matrix consisting of rows corresponding to the number of genes and columns corresponding to the number of regulatory interactions among them is used as input to our procedure. We start with the computation on this matrix and then perform the following steps:

1. Generate the feasible flow vectors of the pathway that satisfy the inequality constraints and the quasi-steady state condition based on biomass conservation constraint,
2. Incorporate a set of weighting coefficients representing concentration of various TFs,
3. Formulate an objective function in terms of these weighting coefficients and then minimize it with respect to the weighting coefficients using the second order learning rule,
4. Determine the optimal regulatory pathway from the values of the weighting coefficients that correspond to minimum value of the objective function.

2.1.3 Generation of gene flow vectors

The node-edge incidence matrix **B** can be computed from the path diagram of the GRNs as depicted in Figures 2–5. Path diagrams (directed graphs) are able to depict genetic regulatory relationships by representing the direction and strength of causal effects in the expression levels of different genes. The path diagram consists of nodes, represented by letters, and edges, represented by lines. The nodes of the path diagram correspond to variables. The directed edges between nodes denote the direction of the regulatory relationship between the nodes (variables) connected by the edges and indicate a directed regulatory influence of one gene on another. The directed edges can represent either activation (positive control) or inhibition (negative control) (Barrett and Palsson, 2006).

The internal flows v are positive yielding $v_i \geq 0$, $\forall i$. The exchange flows b are unconstrained and operate in a bidirectional manner as $\alpha_j \leq b_j \leq \beta_j$ where α_j and β_j are either zero or negative and positive infinity, respectively,

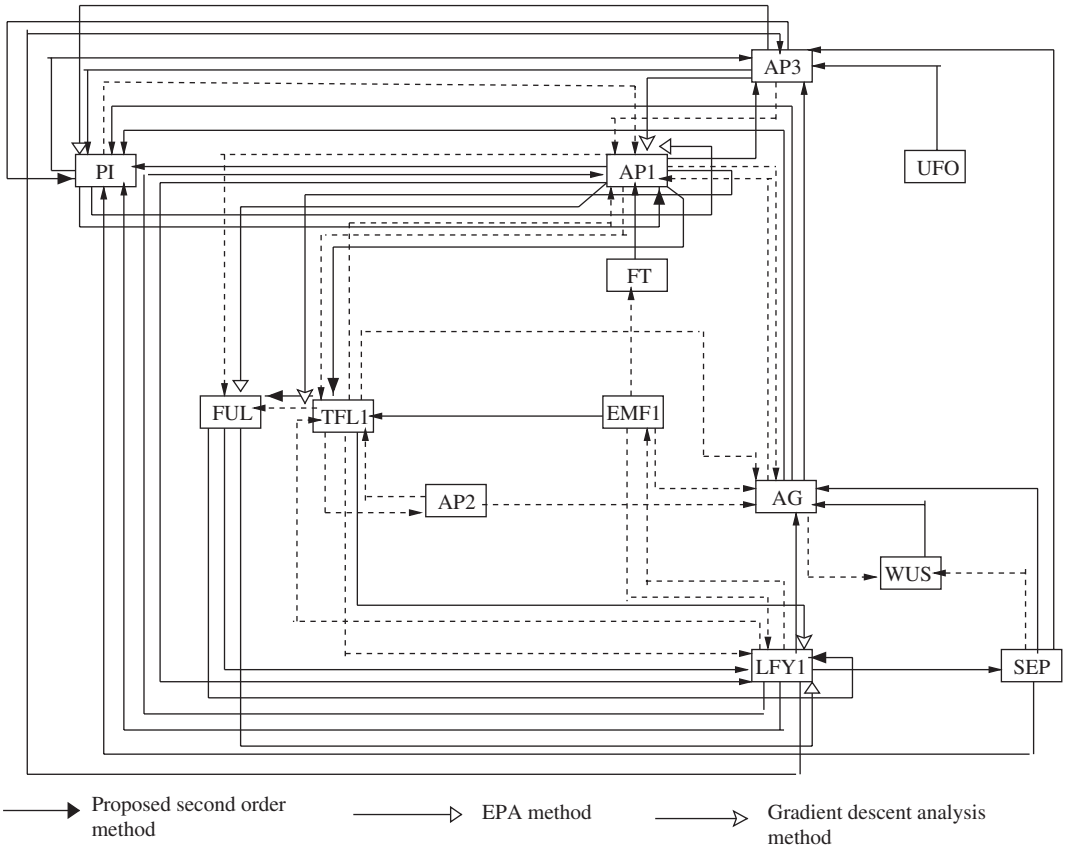

Figure 2 The *A. thaliana* GRN. There are 13 genes (AP3, PI, AP1, UFO, FT, FUL, TFL1, EMF1, AP2, AG, WUS, LFY1, SEP) and 36 interactions, of which there are 18 positive and 18 negative interactions. The starting gene is AP3 and the target gene is LFY1. Positive and negative regulatory interactions are represented by solid and dotted arrows, respectively. The bold and white arrows indicate the pathways obtained by the three methods.

based on the direction of the exchange flow. The flow vectors $\mathbf{v}$, forming the null space of $\mathbf{B}$ satisfies approximately the quasi-steady state condition equation (5) and the above mentioned inequality constraints. As $n>m$, equation (5) is under determined (Schilling et al., 2000). We generate p number of basis vectors $\mathbf{v}_b$ by using standard routines and toolboxes available in MATLAB. Then we further generate p number of random numbers a_j, $j=1, 2, \ldots, p$ and a vector

$$\mathbf{v}=\sum_{j=1}^{p} a_j \mathbf{v}_{bj}, \tag{7}$$

until certain inequality constraint on $\mathbf{v}$ is satisfied for all its components. All the TFs that are not shown in a system may not be expressed at the required level so that the corresponding target genes may not be expressed/inhibited fully. This leads to variation in the concentration of other TFs and hence another constraint can be defined as

$$\mathbf{B}.(\mathbf{C}.\mathbf{v})=0, \tag{8}$$

where $\mathbf{C}$ is an $n\times n$ diagonal matrix whose diagonal elements are the components of the vector $\mathbf{c}$, the weighting coefficient. That is, if $\mathbf{C}=[\gamma_{ij}]_{n\times n}$, then $\gamma_{ij}=\delta_{ij}c_i$, where δ_{ij} is the Kronecker delta. Thus the optimization problem of determining a GRP yielding maximum expression of the target gene g_Q starting from the initial gene g_P, reduces to a maximization problem, where z is maximized with respect to $\mathbf{c}$, subject to satisfying the constraint given in equation (8) along with the inequality constraints described in Section 2.1.3.

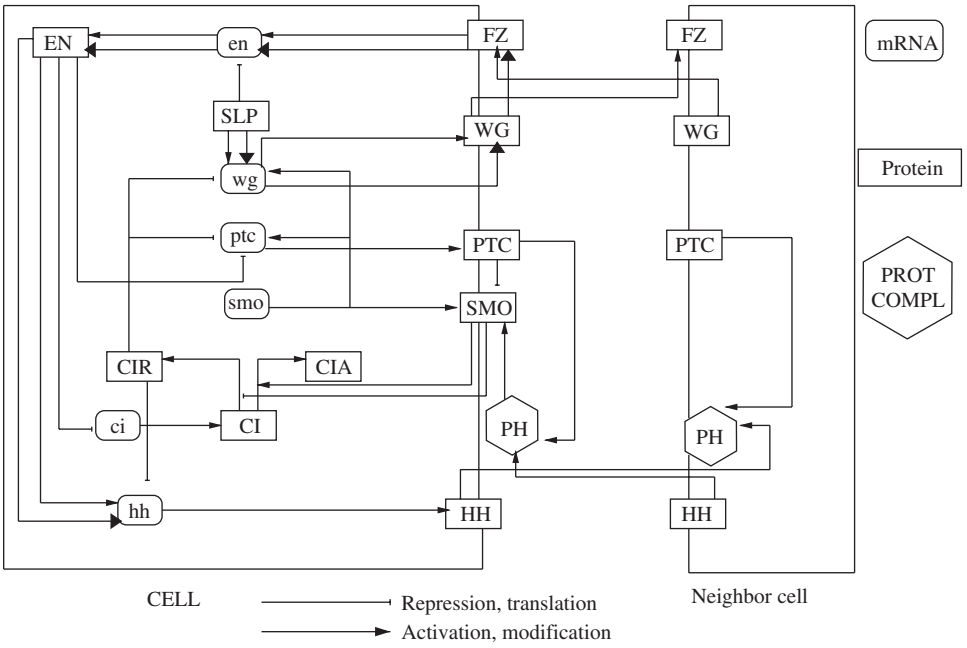

Figure 3 The network of interactions between the segment polarity genes of *Drosophila melanogaster*. There are 17 segment polarity genes (EN, en, FZ, SLP, WG, wg, ptc, PTC, smo, SMO, CIR, CIA, ci, CI, PH, hh, HH) and 29 interactions between them (21 positive and 8 negative). Here the starting gene is SLP and the target gene is HH. The shape of the nodes indicates whether the corresponding substances are mRNAs (ellipses), proteins (rectangles) or protein complexes (octagons). The edges are distinguished by their signatures, i.e., whether they are activating (→) or inhibiting (⊣). The bold black arrows indicates the optimal regulatory pathway.

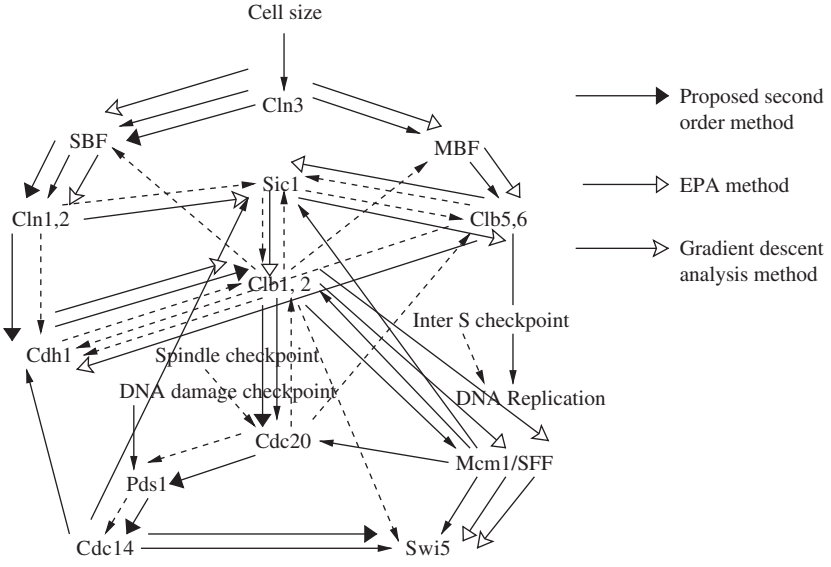

Figure 4 The budding yeast cell cycle regulatory network. There are 13 genes (Cln3, SBF, Cln1,2, Cdh1, Cdc14, Pds1, Cdc20, Clb1,2, Sic1, MBF, Clb5,6, Mcm1/SFF, Swi5) and 34 interactions (16 positive and 18 negative) in this network. The starting gene is Cln3 and the target gene as Swi5. The solid arrows indicate positive regulations and the dotted arrows indicate negative regulations (inhibition, repression or degradation). The bold and white arrows indicates the pathways obtained by the three methods.

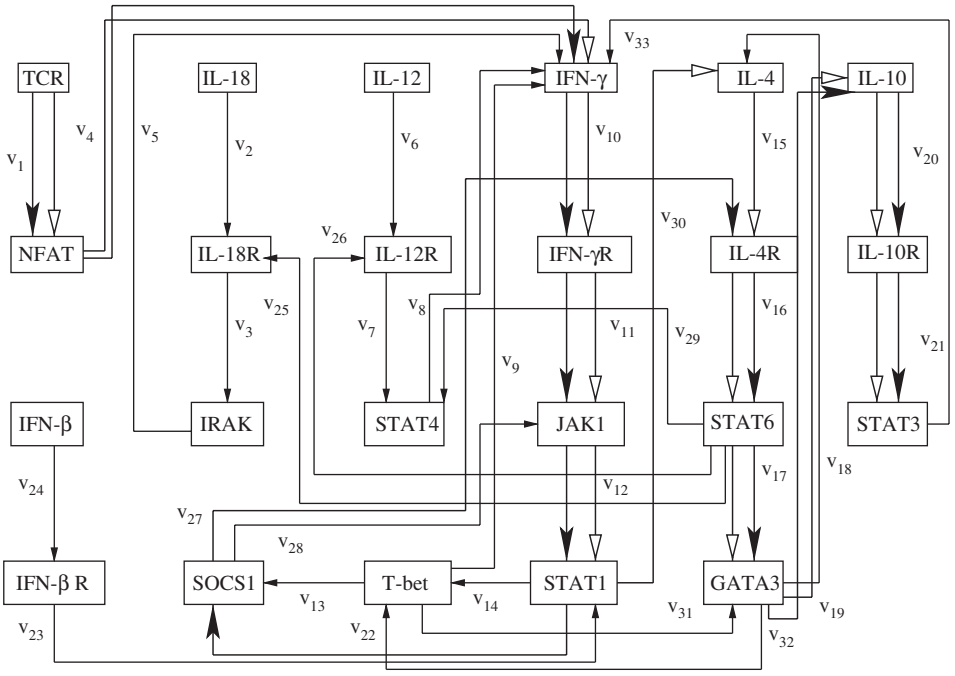

Figure 5 Th cell gene regulatory network in human. There are 23 genes namely, TCR, NFAT, IL-18, IL-12, IFN-γ, IL-4, IL-10, IL-18R, IL-12R, IFN-γR, IL-4R, IL-10R, IFN-β, IRAK, STAT4, JAK1, STAT6, STAT3, IFN-βR, SOCS1, T-bet, STAT1, GATA3 and 33 regulatory reactions among them, being either positive (i.e., activatory), or negative (i.e., inhibitory). Here the starting gene is *TCR* and the target gene is *STAT3*. The optimal regulatory pathway as obtained by the present method is shown by bold black arrows and the regulatory pathway as obtained by the EPA method and GD method is shown by white arrows.

2.1.4 Determination of weighting coefficients c_i through second order learning rule

The objective function z as mentioned in equation (6) is modified to

$$y=1/z+\Lambda^T.(\mathbf{B}.(\mathbf{C}.\mathbf{v})). \tag{9}$$

using equation (8). The new objective function, y has to be minimized with respect to the weighting factors c_i for all i. The term $\mathbf{\Lambda}=[\lambda_1, \lambda_2, \ldots, \lambda_m]^T$ is the regularizing parameter. For the sake of simplicity, we have considered here $\lambda_1=\ldots=\lambda_m=\lambda$ (say). c_i's are generated in the range [0, 1] through random numbers. The c_i's are updated by the new second order learning rule as

$$\Delta c_i=-\frac{\partial y}{\partial c_i}\Big/\left|\frac{\partial^2 y}{\partial c_i^2}\right|. \tag{10}$$

This learning rule is a modified version of the Newton Raphson's method of weight updating. The c_i's can also be modified by using the first order GD optimization technique which largely depends on the value of the learning parameter η. We have considered this issue in our earlier work in Das et al. (2010). This new learning rule is independent of the parameter η and entirely depends on the proper choice of the initial value to reach the global optima. This method based on second order optimization technique overcomes the drawbacks of the first order gradient method such as reduction of local minimum problems, faster convergence. The updated value of c_i is as

$$c_i(t+1)=c_i(t)+\Delta c_i, \qquad \forall i, \qquad t=0, 1, 2, \ldots$$

$c_i(t+1)$ is the value of c_i at iteration $(t+1)$, which is computed based on the c_i-value at iteration t. Δc_i indicates the amount of updation. The λ-value is varied heuristically from 0.1 to 1.0 in steps of 0.1. The corresponding c_i-values are observed as the y-value is minimized. The weighting factor c_i attains values between 0 and 1 as

mentioned in Section (2.1.4) corresponding to *v*-values. We take into account only those values of c_i's that are close to 1 and ignore the other c_i-values that are close to 0.

In the case of metabolic networks, the weighting factor c_i indicates the concentration level of *i*-th enzyme, which is an important factor for a metabolic reaction to proceed. For GRNs, *c*-value indicates concentration of TFs/regulators, as already mentioned in Section (2.1.1). A low *c*-value (close to zero) corresponds to less amount of TFs/regulators and hence, less amount of mRNA production as well as protein production. This will reduce the expression level of the target gene. On the other hand, the non-zero *c*-values signify the presence of required TFs/regulators, to a ceratin extent, for transcription of a gene. Here we have considered higher *c*-values for deriving an optimal regulatory pathway. The corresponding minimal value of *y* is observed. These inferences leads us to identify the optimal GRP yielding the maximal expression of the target gene g_Q starting from the initial gene g_P.

3 Results and comparison

To analyze the computational efficiency of our methodology, we first demonstrate it on a general optimization framework. We have tested our second order learning rule on various types of functions which are convex as well as concave in nature (Table 1).

Then we have used a range of GRNs of varying complexity to demonstrate the applicability of our approach. We inferred the GRNs from published experimental data. The networks considered are the network of *Arabidopsis thaliana* genes involved in flowering morphogenesis in Figure 2 (Mendoza and Alvarez-Buylla, 1998), the interaction of the segment-polarity genes in *Drosophila melanogaster* regulatory network in Figure 3 (Sanchez et al., 2008), the budding yeast cell cycle network in Figure 4 (Ay et al., 2009) and the T helper (Th) cell network in Figure 5 (Garg et al., 2007). Our methodology based on second order learning rule successfully reported the optimal regulatory pathways. A comparison with the existing EPA method (Schilling et al., 2000) and also with the pure GD algorithm has also been done (Das et al., 2010).

3.1 Theoretical analysis

The second order method produced significant and better results (Table 1) when compared to the GD method for the different functions considered. While considering the functions in serial numbers 1 and 2 (Table 1), it can be observed that the proposed method requires very few number of iterations to converge starting from any initial value. The GD method converges for very small η-value and takes large number of iterations. Similar behavior is also observed for the function in serial numbers 3–5 in Table 1. So it can be concluded that the problem in choosing the η-value is very crucial for the GD method. For all the functions in serial numbers 1–6 in Table 1, the proposed method converges faster than the GD method. The GD method diverges or oscillates for large and moderate η-value, and it converges slowly only if the η-value is very small. The function in serial number 6 in Table 1 shows oscillatory behavior by the GD method whereas the proposed method reaches the optimal value very fast. Thus it can inferred that the proposed method performs better as compared to the GD method for steep parabola like functions.

3.2 *Arabidopsis thaliana* GRN

The goal of this work is to develop a mathematical model of genetic control for determination of floral organ identity in *A. thaliana* flower morphogenesis. The topology of the network and the relative strengths of interactions among these genes were based from published genetic and molecular data (Zhang et al., 2011). A genetic control mechanism called the ABC model determining floral organ identity in floral morphogenesis has been proposed (Coen and Meyerowitz, 1991). It describes functioning of the genes determining the type

Table 1 A comparative study on the proposed method and the GD method for various types of functions.

Serial no.	Functions	Initial value	Proposed method		GD method		
			Number of iterations	Optimal value of x	η-Value	Number of iterations	Optimal value of x
1	x^4	4	40	0.00	0.5	4	∞
					0.1	5	∞
					0.01	10,000	0.0353
					0.005	10,000	0.0499
					0.001	70,000	0.0422
2	$\frac{16}{3}x^3$	10	25	0.00	0.5	5	−∞
					0.01	9	−∞
					0.005	2682	0.0046
					0.001	13,452	0.0046
		2	22	0.00	0.2	6	−∞
					0.01	1338	0.0046
					0.001	13,428	0.0046
		4	23	0.00	0.7	5	−∞
					0.5	5	−∞
					0.1	6	−∞
					0.01	1338	0.0046
		0.1	18	0.00	0.5	21	0.0044
3	$\frac{1}{9}(x-2)^9$	50	200	2.0	0.5, 0.1	2	−∞
					0.01, 0.005		
					0.001		
		10	116	2.0	0.5, 0.1, 0.01	1	−∞
					0.001, 0.005	3	−∞
		4	100	2.0	0.5, 0.1	3	−∞
					0.005	90,000	2.316
					0.001	90,500	2.397
4	$\frac{256}{3}(x-3)^3$	5	38	3.0	0.5	5	−∞
					0.01	6	−∞
					0.005	7	−∞
					0.002	52	−∞
					0.001	2151	3.0
		4	36	3.0	0.5	5	−∞
					0.01	7	−∞
					0.005	12	−∞
					0.002	1074	3.0
					0.001	2150	3.0
5	$\frac{256}{3}(x+3)^3$	2	25	−3.0	0.5	4	−∞
					0.1	5	−∞
					0.01	6	−∞
					0.0002	2000	−2.99
					0.0001	17,000	−2.99
		5	26	−3.0	0.1	5	−∞
					0.01	4	−∞
					0.001	8	−∞
					0.0001	17,000	−2.99
		10	27	−3.0	0.01	5	−∞
					0.001	7	−∞
					0.0001	17,000	−2.99
6	$10{,}000\,x^2$	10	1	0.00	0.5	10	∞
					0.1	12	∞
					0.01	16	∞
					0.001	27	−∞
					0.0001		Oscillatory

Here "optimal value of x" means the value of x at which the function attains a minimum value

of developing floral organ. The genes controlling the type of floral organs can be divided into three classes: A, B, and C. *A. thaliana* flowers consist of four concentric whorls of organs: four sepals, four petals, six stamens and two fused carpels arranged from the outermost (whorl one) to the innermost (whorl four). The genes related to flowering are grouped into a hierarchy of four sets of genes depending on their time of activation as the transition to flowering and flower morphogenesis proceeds (Mendoza and Alvarez-Buylla, 1998). Based on these considerations, we have divided the 13 network elements into four groups, A class genes (APETALA1 [AP1] and AP2) specify sepal fate in the outer (first) floral whorl, A plus B genes (AP3 and PISTILLATA [PI]) determine petal development in the second whorl, B plus C genes determine stamens in the third whorl and the C gene (AGAMOUS [AG]) alone determines carpel fate in the central (fourth) whorl (Coen and Meyerowitz, 1991).

The present second order method generates the optimal GRP as $AP3 \rightarrow PI \rightarrow AP1 \rightarrow TFL1 \rightarrow FUL \rightarrow LFY1$. The GD optimization method gives a different pathway as $AP3 \rightarrow AP1 \rightarrow TFL1 \rightarrow LFY1$. The EPA method produces another different pathway as $AP3 \rightarrow P1 \rightarrow AP1 \rightarrow FUL \rightarrow LFY1$. The three different pathways are depicted by the three different arrows as shown in Figure 2.

Table 2 provides a list of c-values and the amount (z) of the protein synthesized by the target gene for an optimal regulatory path as obtained by the present method and the GD method by varying the upper bound on flow values for the GRN in Figure 2. The z-value decreases with the decrease in upper bound of the flow value. It can be observed that the c-values changes for all the 10 instances. Thus, it can be concluded that for each of the 10 cases we have obtained the same optimal regulatory path $AP3 \rightarrow PI \rightarrow AP1 \rightarrow TFL1 \rightarrow FUL \rightarrow LFY1$ (from our proposed method), $AP3 \rightarrow AP1 \rightarrow TFL1 \rightarrow LFY1$ (by the GD method) although there are differences in the c-values. It can be inferred from Table 2 that the decrease in z-value obtained by the GD method is smaller than that obtained by our method. Consequently the decrease in y-value by the present method will be much more when compared to the GD method. Thus we can conclude that our method performs better as compared to the GD method. Similar table can also be constructed for the other examples considered in this work. This will lead to similar observations. Thus it can be concluded that the present second order method is effective in identifying the optimal GRP.

3.3 *Drosophila melanogaster* GRN

This regulatory network is extremely well-studied, and the genome sequences of many members of the Drosophila family are now available (Ma et al., 2006). We focus on a recent model of the segment polarity genes which are expressed throughout the life of the fruit fly *Drosophila melanogaster*. There are 40 genes (the

Table 2 Variation of c-values and average z with the upper bound on reaction flows for the optimal regulatory paths of the system in Figure 2.

Serial number	Upper bound on flow value	Optimal c-values		Average amount of protein (z)	
		Proposed method	GD	Proposed method	GD
1	5000	0.95, 0.94, 0.91, 0.82, 0.89	0.91, 0.90, 0.85	6379.44	4096.54
2	4000	0.93, 0.97, 0.96, 0.91, 0.76	0.90, 0.89, 0.81	5104.53	4762.71
3	3000	0.88, 0.85, 0.82, 0.86, 0.85	0.93, 0.82, 0.80	4506.71	3562.81
4	2000	0.88, 0.87, 0.96, 0.93, 0.81	0.87, 0.84, 0.81	3569.88	2373.84
5	1000	0.95, 0.92, 0.87, 0.81, 0.86	0.90, 0.87, 0.85	2678.74	1954.44
6	50	0.81, 0.84, 0.88, 0.81, 0.89	0.89, 0.87, 0.85	59.62	35.78
7	40	0.88, 0.84, 0.94, 0.92, 0.98	0.86, 0.85, 0.82	48.45	26.83
8	30	0.85, 0.93, 0.86, 0.97, 0.86	0.90, 0.89, 0.85	39.78	21.08
9	20	0.93, 0.84, 0.96, 0.94, 0.86	0.91, 0.90, 0.84	31.79	14.85
10	10	0.85, 0.82, 0.93, 0.88, 0.88	0.94, 0.92, 0.91	20.05	10.74

gap genes, the pair-rule genes and the segment polarity genes) organized hierarchically in this arthropod. Here we consider only the segment polarity genes and their associated interactions. The segment polarity genes encode for the TF engrailed (EN), the cytosolic protein cubitus interruptus (CI), the secreted proteins wingless (WG), the hedgehog (HH), the transmembrane receptor proteins patched (PTC) and the smoothened (SMO) involved in transduction of the HH signal (Ingolia, 2004). The present second order method generates the optimal GRP as *SLP→wg→WG→FZ→en→EN→hh→HH* as indicated by bold black arrows as shown in Figure 3. The GD optimization method and the EPA method generates the same optimal regulatory pathway as described above.

The pair-rule gene sloppy paired *slp* is activated before the segment polarity genes and expressed afterwards. *slp* encodes two TFs with similar functions that activate transcription and repress *en* transcription and they are co-expressed. The *wg* gene encodes a glycoprotein that is secreted from the cells and can bind to the Frizzled (FZ) receptor on neighboring cells. This leads to the transcription of engrailed *en*. EN leads to the transcription of the hedgehog gene *hh* and represses *ci* and *ptc*. The hedgehog protein (HH) is attached to the cell membrane by a linkage that is severed by the dispatched protein, freeing it to bind to the HH receptor PTC on a neighboring cell. PTC forms a complex with smoothened (SMO) in which SMO is inactivated by a post-translational conformational change. Binding of HH to PTC removes the inhibition of SMO, and activates a pathway that results in the modification of CI. The CI protein can be converted into one of two TFs, depending on the activity of SMO. When SMO is inactive, CI is cleaved to form CIR, a transcriptional repressor that represses wg, ptc and hh. When SMO is active, CI is converted to a transcriptional activator CIA that promotes the transcription of wg and ptc (Nanfack et al., 2009).

3.4 Budding yeast cell cycle GRN

The cell-cycle process is highly conserved among the eukaryotes. The cell cycle consists of four phases: G1 (cell growth and division) phase, S (DNA synthesis and replication of chromosomes) phase, G2 (a "gap" between S and M) phase and M (mitosis-chromosomes separated and the cell divided into two) phase (Ay et al., 2009). The cell enters the G1 phase after the M phase thus completing a cycle. The G1 phase is observed the most in the yeast life cycle. This GRN consists of four different types of entities: cyclins (Cln1, 2, 3 and Clb1, 2, 5, 6 which bind to the kinase Cdc28); the inhibitors, degraders and competitors of the cyclin/Cdc28 complexes (Sic1, Cdh1, Cdc20, Cdc14); TFs (SBF, MBF, Mcm1/SFF, Swi5) and checkpoints (the cell size, the DNA replication and damage and the spindle assembly) (Figure 4) (Davidich and Bornholdt, 2008). When the cell grows large under rich nutrient conditions the Cln3/Cdc28 will be "activated," which in turn activates (by phosphorylation) a pair of TF groups, SBF and MBF, which transcriptionally activate the genes of the cyclins Cln1 and 2 and Clb5 and 6, respectively. The protein Sic1 can bind to the Clb/Cdc28 complex to inhibit its function, Clb1 and 2 phosphorylates Swi5 to prevent its entry into the nucleus, whereas Cdh1 targets Clb1 and 2 for degradation.

The cell-cycle process starts when the cell starts to divide by activating Cln3 (the START). The subsequent activity of Clb5 drives the cell into the S phase. The entry into and exit from the M phase is controlled by the activation and degradation of Clb2. After the M phase, the cell comes back to the stationary G1 phase, waiting for the signal for another round of division. Thus the cell-cycle process starts with the "excitation" from the stationary G1 state by the "cell-size" signal and evolves back to the stationary G1 state through a well defined sequence of states (Bahler, 2005). So the starting gene can be considered as Cln3 and the target gene as Swi5. The present second order method generates the optimal GRP as *Cln3→SBF→Cln*1,2*→Clb*1,2 *→Cdc*20*→Pds*1*→Cdc*14*→Swi5*. The GD optimization method produces a different path as *Cln3→SBF→Cln*1,2 *→Sic*1, 2*→Clb5*, 6*→Cdh*1*→Clb*1,2*→Mcm*1/*SFF→Swi5*. The EPA generates another different pathway as *Cln*3 *→MBF→Clb*5,6*→Sic*1*→Clb*1,2*→Mcm*1/*SFF→Swi5*. The three different types of arrows in Figure 4 denote the pathways obtained by the three methods.

3.5 T helper cell GRN

With the increasing availability of experimental data on gene-gene and protein-protein interactions, modeling of GRNs has gained special attention lately. We benchmarked and validated our algorithm on the T helper model from human (Garg et al., 2007). The immune system is composed of diverse cell populations, for example antigen-presenting cells, T and B lymphocytes as well as effector cells like eosinophils, mast cells and neutrophils. B and T cells are the two major types of lymphocytes in the mammalian immune system. One type of T lymphocytes, called T helper (Th), has an important role in regulating this cellular network (Naldi et al., 2010). Th cells coordinates the immune system by secreting a series of cytokines, which activate other immune cells. Th cells can be further divided into Th1 and Th2 cells. There are five recognized lineages of Th cells, namely Th0, Th1, Th2, Th17, and Treg. Th1 and Th2 cells are thought to be mutually inhibitory and also to be involved in different diseases; Th1 cells are associated with autoimmune diseases, while Th2 cells are involved in allergies (Rodriguez et al., 2007; Pedicini et al., 2010; Hong et al., 2011).

Various kinds of molecules (secreted cytokines, receptors, signal transducers and TFs) were considered in Figure 5. An optimal GRP obtained by the present second order method is $v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11} \rightarrow v_{12} \rightarrow v_{22} \rightarrow v_{27} \rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$ as shown by bold black arrows. The EPA and GD method generates a different regulatory pathway as $v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11} \rightarrow v_{12} \rightarrow v_{30} \rightarrow v_{15} \rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$ as shown by white arrows in Figure 5.

4 Biological validation

This second order method yields more relevant results, as compared to GD formalism and EPA analysis (Das et al., 2010), from the perspective of biology. Our second order learning method identifies optimal regulatory pathways for the GRNs as discussed above and gives similar results as were reported in the literature. The biological significance of the genes reported in the optimal GRP and also present in the network are highlighted below.

We discuss the biological implications and the potential use of the genetic aspects of flowering in a highly studied plant *Arabidopsis thaliana* GRN in Figure 2. The network includes genes that codify for TFs [AG, PI, AP3, WUS, etc.; Coen and Meyerowitz (1991)], F-box proteins [UFO; Samach et al. (1999)], membrane bound signaling molecules [TFL1 and FT; Simpson et al. (1999)], cofactors involved in transcriptional regulation [EMF1; Aubert et al. (2001)] and TFs. Recently, it has been shown that EMF1 down-regulates AG (Calonje et al., 2008). The starting gene AP3 and the intermediate gene PI of the optimal GRP obtained by the proposed second order method are both able to down-regulate AP1 (another intermediate gene of the optimal GRP) (Sundstrom et al., 2006). The target gene LFY1 of the optimal GRP is involved in the transition from the non-flowering state to the flowering pathways. The meristem identity genes LFY and AP1, an intermediate gene of the optimal GRP, initiate floral development by activating floral homeotic gene expression in the floral meristem (Bowman et al., 1993). EMF1 gene is part of the proposed floral repressor, which controls the transition from vegetative to reproductive growth, and it is proposed to be under the influence of many upstream genes (Coupland, 1995). The biological significance of all the genes involved in the GRN of *A. thaliana* in Figure 2 are cited in Mendoza and Alvarez-Buylla (1998), Zhang et al. (2011), Coen and Meyerowitz (1991).

Using Arabidopsis flower development as our model system, we have begun to understand how a group of undifferentiated cells in a floral meristem develop into a complex floral structure with four types of floral organs and many different cell types. Hence, a mathematical model based on the data on genetic control of flower development in *A. thaliana* can be later applied to study flower development in angiosperms. Floral homeotic genes or floral homeotic mutants with A, B, and C functions are found in several plant species studied (Angenent et al., 1993; Pnueli et al., 1994). Understanding how domain-specific activities of floral homeotic genes are regulated in *A. thaliana* and other plant species would shed light on the evolution of key control mechanism in floral pattern formation.

The segment polarity gene network of *Drosophila melanogaster* in Figure 3 is a very robust developmental module that is adopted in a wide range of developmental programs (Sanchez et al., 2008). These segment

polarity genes are conserved among all insects, perhaps among all arthropods. The reliability of our predictions is confirmed by detailed gene specific studies from the literature. The expression pattern of the segment polarity genes [expression of wingless (wg) and engrailed (EN)] (intermediate genes of the optimal GRP in Figure 3) maintains the borders between different parasegments (the embryonic counterparts of the segments), and executes specific developmental functions such as the growth of the denticle patterns and appendage primordia (Ingolia, 2004; Ma et al., 2006). Homologs of the segment polarity genes, present in the optimal GRP of Figure 3, have been identified in vertebrates, including humans, which suggests strong evolutionary conservation of these genes (Nanfack et al., 2009).

The starting gene Cln3 of the optimal GRP in Figure 4 activates the TFs, SBF and MBF. The transcription level of several hundred START-specific genes are correlated with the level of Cln3 expression. Cdh1, an intermediate gene of the optimal GRP, plays an important role in Clb2 (another intermediate gene of the optimal GRP) degradation during telophase. Cdc14, an intermediate gene of the optimal GRP, is a phosphatase that takes part in the exit phase of mitosis while activating certain proteins such as Swi5, Sic1, Cdh1 and some unknown target proteins. Cdc20, an intermediate gene, of the optimal GRP, is needed for Clb5 degradation and partial degradation of Clb2 at anaphase. Pds1, an intermediate gene of the optimal GRP, inhibits Cdh1 activation, possibly by preventing Cdc14 release. The cyclins Cln1, 2 take part in DNA replication, duplication of the spindle pole body and also in bud formation. Clb1 encodes a B-type cyclin that activates Cdc28 to promote the transition from G2 to M phase of the cell cycle. Clb1 and Clb2 transcripts accumulate during G2 and M phase, and their transcription is repressed at the end of mitosis (Bahler, 2005; Davidich and Bornholdt, 2008; Ay et al., 2009).

The importance of all the genes in the optimal GRP of the T helper (Th) GRN in Figure 5 has been observed in Santoni et al. (2008). Th1 cells produce interferon IFN-γ (the intermediate gene in the optimal GRP), promoting cell-mediated immunity and control of intracellular pathogens. Th1 differentiation is regulated by TFs such as T-bet, Stat1 (the intermediate gene in the optimal path) and Stat4, as well as by cytokines such as IL-12, IL-23, IL-27, type I IFNs, and IFN-γ. In contrast, Th2 cells produce IL-4 (the intermediate gene in the optimal extreme path), which promotes allergic responses and is important in host defense against helminths. The TFs Stat6, GATA-3, c-Maf, NFATs (the intermediate genes in the optimal GRP), and the cytokine IL-4 promote Th2 differentiation.

5 Concluding remarks and future challenges

GRPs are the most complex, extensive control systems found in nature. To have a better understanding of these pathways it is necessary to model them using ordinary differential equations and compute their steady states (Goelzer et al., 2008). GRP inference is critically important for revealing fundamental cellular processes, investigating gene functions and understanding their relations. Herein, we have employed a second order optimization method based on Newton's method to reveal the potential optimal regulatory pathways between genes through certain weighting factors that encode the TFs (Buntine and Weigend, 1994). This second order learning rule overcomes the drawbacks of the first order BP scheme. Unlike GD this method is independent of the learning parameter. The results for different types of benchmark functions are provided in Table 1 to demonstrate the effectiveness of the proposed second order scheme, along with its superiority over the GD method. Benchmarks on different GRNs of varying complexity show that this new approach yields significant results. The results demonstrate that the new second order method can provide meaningful insights in understanding the nonlinear dynamics of the GRPs and revealing potential regulatory interactions between genes.

A variety of aspects provide possibilities for fruitful future work. The *Arabidopsis thaliana* GRN model provides a tool for exploring the relative conservation of the basic structure of flowers among angiosperms. This strongly suggests that the ABC model is widely conserved across flowering plants. As flower development proceeds, the same or other putative signals might prompt the GRN of *A. thaliana* for petals, stamens and carpels, thus completing flower development. Until now, the methodology has been used mainly on the

study of the metabolic and regulatory networks, thus applying it to diverse biological systems will help to better appreciate both its strengths and limitations.

References

Angenent, G. C., J. Franken, M. Busscher, L. Colombo and A. J. van Tunen (1993): "Petal and stamen formation in petunia is regulated by the homeotic gene fbp1," Plant J., 4, 101–112.

Aubert, D., L. Chen, Y. Moon, D. Martin, L. A. Castle, C. Yang and Z. R. Sung (2001): "Emf1, a novel protein involved in the control of shoot architecture and flowering in arabidopsis," Plant Cell, 13, 1865–1875.

Ay, F., F. Xu and T. Kahveci (2009): "Scalable steady state analysis of boolean biological regulatory networks," Plos One, 4, e7992, doi: 10.1371/journal.pone.0007992.

Bahler, J. (2005): "Cell-cycle control of gene expression in budding and fission yeast," Annu. Rev. Genet., 39, 69–94.

Barrett, C. L. and B. O. Palsson (2006): "Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach," Plos Comput. Biol., 2, e52, doi: 10.1371/journal.pcbi.0020052.

Benyamini, T., O. Folger, E. Ruppin and T. Shlomi (2010): "Flux balance analysis accounting for metabolite dilution," Genome Biol., 11, R43, doi: 10.1186/gb–2010–11–4–r43.

Bortoletti, A., C. D. Fiore, S. Fanelli and P. Zellini (2003): "A new class of quasi-newtonian methods for optimal learning in mlp-networks," IEEE T. Neural Network, 14, 263–273.

Bowman, J. L., J. Alvarez, D. Weigel, E. M. Meyerowitz and D. R. Smyth (1993): "Control of flower development in arabidopsis thaliana by apetala1 and interacting genes," Development, 119, 721–743.

Buntine, W. L. and S. Weigend (1994): "Computing second derivatives in feed-forward networks: a review," IEEE T. Neural Network, 5, 480–488.

Calonje, M., R. Sanchez, L. Chen and Z. R. Sung (2008): "Embryonic flower1 participates in polycomb group-mediated ag gene silencing in arabidopsis," Plant Cell, 20, 277–291.

Castillo, E., B. G. Berdinas, O. F. Romero and A. A. Betanzos (2006): "A very fast learning method for neural networks based on sensitivity analysis," J. Mach. Learn. Res., 7, 1159–1182.

Coen, E. and E. Meyerowitz (1991): "The war of the whorls: genetic interactions flower development," Nature, 353, 31–37.

Coupland, G. (1995): "Genetic and enviromental control of flowering time in arabidopsis," Trends Genet., 11, 393–397.

Crombach, A. and P. Hogeweg (2008): "Evolution of evolvability in gene regulatory networks," PLoS Comput. Biol., 4, e1000112, doi: 10.1371/journal.pcbi.1000112.

Das, M., S. Mukhopadhyay and R. K. De (2010): "Gradient descent optimization in gene regulatory pathways," PLoS One, 5, e12475, doi: 10.1371/journal.pone.0012475.

Davidich, M. I. and S. Bornholdt (2008): "Boolean network model predicts cell cycle sequence of fission yeast," PLoS One, 3, e1672, doi: 10.1371/journal.pone.0001672.

Garg, A., I. Xenarios, L. Mendoza and G. DeMicheli (2007): "Efficient methods for dynamic analysis of genetic networks and in silico gene perturbation experiments," Lect. Notes Bioinform., 4453, 62–76.

Goelzer, A., F. B. Brikci, I. M. Verstraete and P. e. a. Noirot (2008): "Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *bacillus subtilis,*" BMC Syst. Biol., 2, 20, doi: 10.1186/1752–0509–2–20.

Hong, T., J. Xing, L. Li and J. J. Tyson (2011): "A mathematical model for the reciprocal differentiation of t helper 17 cells and induced regulatory t cells," Plos Comput. Biol., 7, e1002122, doi: 10.1371/journal.pcbi.1002122.

Ingolia, N. T. (2004): "Topology and robustness in the Drosophila segment polarity network," PLoS Biol., 2, e123, doi: 10.1371/journal.pbio.0020123.

Lee, W. and W. Tzou (2009): "Computational methods for discovering gene networks from expression data," Brief. Bioinform., 10, 408–423.

Ma, W., L. Lai, Q. Ouyang and C. Tang (2006): "Robustness and modular design of the Drosophila segment polarity network," Mol. Syst. Biol., 2, 70, doi: 10.1038/msb4100111.

Magoulas, G. D., M. N. Vrahatis and G. S. Androulakis (1999): "Improving the convergence of the backpropagation algorithm using learning rate adaptation methods," Neural Comput., 11, 1769–1796.

Mendoza, L. and E. R. Alvarez-Buylla (1998): "Dynamics of the genetic regulatory network for arabidopsis thalina flower morphogenesis," J. Theor. Biol., 193, 307–319.

Naldi, A., J. Carneiro, C. Chaouiya and D. Thieffry (2010): "Diversity and plasticity of Th cell types predicted from regulatory network modelling," Plos Comput. Biol., 6, e1000912, doi: 10.1371/journal.pcbi.1000912.

Nanfack, Y. F., M. Postma and J. A. Kaandorp (2009): "Inferring drosophila gap gene regulatory network: a parameter sensitivity and perturbation analysis," BMC Syst. Biol., 3, 94, doi: 10.1186/1752–0509–3–94.

Pedicini, M., F. Barrenas, T. Clancy and F. e. a. Castiglione (2010): "Combining network modeling and gene expression microarray analysis to explore the dynamics of Th1 and Th2 cell regulation," Plos Comput. Biol., 6, e1001032, doi: 10.1371/journal.pcbi.1001032.

Pnueli, L., D. Hareven, S. D. Rounsley, M. F. Yanofsky and E. Lifshitz (1994): "Isolation of the tomato agamous gene tag1 and analysis of its homeotic role in transgenic plants," Plant Cell, 6, 163–173.

Rodriguez, J. S., L. Simeoni, J. A. Lindquist and R. e. a. Hemenway (2007): "A logical model provides insights into T cell receptor signaling," Plos Comput. Biol., 3, 163, doi: 10.1371/journal.pcbi.0030163.

Samach, A., J. E. Klenz, S. E. Kohalmi, E. Risseeuw, G.W. Haughn and W. L. Crosby (1999): "The unusual floral organs gene of arabidopsis thaliana is an f-box protein required for normal patterning and growth in the floral meristem," Plant J., 20, 433–445.

Sanchez, L., C. Chaouiya and D. Thieffry (2008): "Segmenting the fly embryo: logical analysis of the role of the segment polarity cross-regulatory module," Int. J. Dev. Biol., 52, 1059–1075.

Santoni, D., M. Pedicini and F. Castiglione (2008): "Implementation of a regulatory gene network to simulate the th1/2 differentiation in an agent-based model of hypersensitivity reactions," Bioinformatics, 24, 1374–1380.

Schilling, C. H., D. Letscher and B. O. Palsson (2000): "Theory for the systemic defnition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective," J. Theor. Biol., 203, 229–248.

Schlitt, T. and A. Brazma (2007): "Current approaches to gene regulatory network modelling," BMC Bioinformatics, 8 (Suppl 6), S9, doi: 10.1186/1471–2105–8–S6–S9.

Simpson, G., A. Gendall and C. Dean (1999): "When to switch to flowering," Annu. Rev. Cell Dev. Biol., 99, 519–550.

Sirbu, A., H. J. Ruskin and M. Crane (2010): "Comparison of evolutionary algorithms in gene regulatory network model inference," BMC Bioinformatics, 11, 59, doi: 10.1186/1471–2105–11–59.

Stelling, J. (2004): "Mathematical models in microbial systems biology," Curr. Opin. Microbiol., 7, 513–518.

Sundstrom, J. F., N. Nakayama, K. Glimelius and V. F. Irish (2006): "Direct regulation of the floral homeotic apetala1 gene by apetala3 and pistillata in arabidopsis," Plant J., 46, 593–600.

Tu, Z., L. Wang, M. N. Arbeitman, T. Chen and F. Sun (2006): "An integrative approach for causal gene identification and gene regulatory pathway inference," Bioinformatics, 22, 489–496.

Wang, Y. J. and C. T. Lin (1998): "A second-order learning algorithm for multilayer networks based on block hessian matrix," Neural Networks, 11, 1607–1622.

Xiong, M., J. Zhao and H. Xiong (2004a): "Network-based regulatory pathways analysis," Bioinformatics, 20, 2056–2066.

Xiong, M. M., J. Li and X. Z. Fang (2004b): "Identification of genetic networks," Genetics, 166, 1037–1052.

Yeh, I. C. and W. L. Cheng (2010): "First and second order sensitivity analysis of mlp," Neurocomputing, 73, 2225–2233.

Zhang, X., A. J. Cal and J. O. Borevitz (2011): "Genetic architecture of regulatory variation in arabidopsis thaliana," Genome Res., 21, 725–733.