A Second-Order Learning Algorithm for Computing Optimal Regulatory Pathways

Mouli Das[1], C.A. Murthy[1], Subhasis Mukhopadhyay[2], and Rajat K. De[1]

[1] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
{mouli_r,murthy,rajat}@isical.ac.in
[2] Department of Bio-Physics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata 700 009, India
sm.bmbg@gmail.com

Abstract. Gene regulatory pathways play an important role in the functional understanding and interpretation of gene function. Many different approaches have been developed to model and simulate gene regulatory networks. In this paper we present the results of an iterative new second-order learning algorithm based on the multilayer perceptron (MLP) for generating optimal gene regulatory pathways by using ordinary differential equations. The algorithm based on Newton's method is independent on the learning parameter and overcomes the drawbacks of the standard backpropagation (BP) algorithm. The methodology generates flow vectors which indicate the flow of mRNA and thereby the protein produced from one gene to another gene. A set of weighting coefficients representing concentration of various transcription factors is incorporated. The gene regulatory pathways are obtained through optimization of an objective function with respect to these weighting coefficients. Two gene regulatory networks are used to demonstrate the efficiency of the proposed learning algorithm. A comparative study with the existing extreme pathway analysis (EPA) also forms a part of this study. Results reported in the paper were corroborated by the same reported in the literature.

Keywords: FBA, extreme pathway, backpropagation, MLP, transcription factor.

1 Introduction

The increasing availability of genomic, transcriptomic and related data allows detailed analysis of properties of complex biochemical reaction networks composed of gene networks, protein networks, metabolic networks and signaling networks. Flux balance analysis (FBA) [6] has been useful for large scale analysis of metabolic networks, and methods have been developed to extend this approach for transcriptional regulation. High-throughput technologies allow studying aspects of gene regulatory networks (GRNs) on a genome-wide scale and we will discuss recent advances as well as limitations and future challenges for gene network modelling. GRNs have an important role in every process of life, including cell differentiation, metabolism, the cell cycle and signal transduction. GRNs are

M.K. Kundu et al. (Eds.): PerMIn 2012, LNCS 7143, pp. 227–234, 2012.

concerned with the control of transcription, i.e. how genes are up and down regulated in response to signals [3]. GRNs consist of interactions between proteins, known as transcription factors, and genes, which in turn encode other proteins. By understanding the dynamics of GRNs we can shed light on the mechanisms of diseases that occur when the cellular processes are dysregulated. Inference of gene regulatory pathways is a key goal in the quest for understanding fundamental cellular processes and revealing underlying relations among genes [14].

Optimization by gradient descent is widely used by various machine learning algorithms such as back-propagation (BP) of the error in Multi-Layer Perceptrons (MLPs) and Radial Basis Functions [5]. However, several drawbacks of the BP learning method have been observed; its convergence speed is usually too low, its convergence accuracy is hard to control, it is easily stuck in bad local minima and the choice of proper learning constant largely depends on trial and error [11]. One common approach is to upgrade the normal BP, which is a first-order learning algorithm, to a second-order one [1]. Since the second order method is an optimization algorithm with quadratic convergence speed, it can be used to improve the learning speed and accuracy of the normal BP [13]. We describe an extension of the back propagation algorithm which uses a simple approximation to the second derivative terms. Also, the proposed method is independent of the learning constant in contrary to the difficulty in the choice of a proper learning constant for normal BP.

The proposed method generates the possible flow vectors in the pathway by taking convex combination of the basis vectors spanning the null space of the given node-edge incidence matrix. These flow vectors satisfy the quasi-steady state condition along with other inequality constraints. A set of constraints involving the weighting coefficients representing concentration of various transcription factors is formulated. An objective function, in terms of these weighting coefficients, is formed, and minimized through the new learning technique. The weighting coefficients corresponding to a minimum value of the objective function represent an optimal regulatory pathway yielding the maximal expression of the target gene starting from the initial gene. These optimal pathways determine the gene regulatory routes leading from the transcription of a given gene to the transcription of another gene, and represent the structural and functional properties of the network as a whole. Two benchmark regulatory networks are given to illustrate this approach. The results are biologically validated and compared with the standard extreme pathway analysis (EPA) [12].

2 Second Order Optimization Algorithm

The second order method is derived from Newton's method [2] whose principle is discussed here. The Taylor expansion of a function $E(w)$ of a single variable w in the vicinity of a minimum w^* is given by

$$E(w) = E(w^*) + 1/2(w - w^*)^2(d^2E/dw^2)_{w=w^*} + O(w^3) \quad (1)$$

The gradient of the cost function is zero at the minimum. Differentiating the above equation (1) with respect to w gives an approximation of the gradient of the cost function in the neighborhood of a minimum,

$$dE/dw = (w - w^*)(d^2E/dw^2)_{w=w^*} \tag{2}$$

Therefore, if variable w is in the neighborhood of w^*, the minimum could be reached in a single iteration if the second derivative of the cost function at the minimum were known. w would simply be updated by an amount

$$\Delta w = -\frac{(dE/dw)}{(d^2E/dw^2)_{w=w^*}} \tag{3}$$

Thus by contrast to simple gradient descent, the direction of motion, in parameter space, is not the direction of the gradient, but a linear transformation of the gradient. Our proposed methodology is an iterative technique that is based on the above formula (equation (3)) with certain modifications in the double derivative term.

3 Proposed Second Order Optimization Methodology

Here we develop a new learning rule based on Newton's method [9] for identification of an optimal regulatory pathway in gene regulatory networks starting from a given gene to a target gene through which the expression level of the target gene becomes maximum. The gene regulatory networks are described by directed graphs [14] with nodes corresponding to genes and edges to regulatory interactions. 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. One important concept that we will use below is a representation of a graph by a so-called node-edge incidence matrix $\mathbf{B}$, where the element b_{ij} in a row i and column j equals 1 (i.e., $b_{ij} = 1$), if node i is connected to node j, otherwise $b_{ij} = 0$. Let g_i be 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.

A system boundary is drawn around a gene regulatory network which consists of both internal flows, constrained to be positive and exchange flows, constrained to be either positive, negative or bidirectional depending on the direction. There are n flows/regulatory interactions and m genes in the network. Let n_I be the number of internal flows and n_E be that of exchange flows, and then $n = n_I + n_E$. The i-th internal flow is denoted by v_i and the j-th exchange flow is denoted by b_j. So there are $v_1, \ldots, v_{n_I}$ internal flows and $v_{n_I+1}, \ldots, v_n$ exchange flows where $v_{n_I+l} = b_l$. The target gene can be reached through k biochemical reactions $R_1, R_2, \ldots R_k$ from the starting gene. The algebraic sum of the weighted flows of reactions $R_1, R_2, \ldots R_k$ to reach the target gene is given by

$$z = \sum_{i=1}^{k} c_i v_i \tag{4}$$

which needs to be maximized for yielding maximum expression level of the target gene. The term c_i denotes the weighting factor, representing concentration of other transcription factors to get the corresponding flow v_i.

The gene flow vectors $\mathbf{v}$ satisfy approximately the quasi-steady state condition

$$\mathbf{Bv} \approx \mathbf{0} \tag{5}$$

where $\mathbf{B}$ is the $m \times n$ node-edge incidence matrix which can be computed from a given gene regulatory network. As $n > m$, equation (5) is under determined. We generate p number of basis vectors $\mathbf{v}_b$ that span the null space of $\mathbf{B}$. p random numbers a_j, $j = 1, 2, \ldots, p$ are further generated. Finally a vector $\mathbf{v}$ is generated as a linear combination of the basis vectors using a_p i.e., $\mathbf{v} = \sum_{j=1}^{p} a_j \mathbf{v}_{bj}$. The flow vectors $\mathbf{v}$ satisfy the following inequality constraints [12]. All the internal fluxes are positive yielding: $v_i \geq 0, \forall i$. The constraints on the exchange fluxes depending on their direction can be expressed as $\alpha_j \leq b_j \leq \beta_j$ where $\alpha_j \epsilon \{-\infty, 0\}$ and $\beta_j \epsilon \{0, \infty\}$.

All the transcription factors 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 transcription factors and hence another constraint can be defined as

$$\mathbf{B}(\mathbf{Cv}) = \mathbf{0} \tag{6}$$

where $\mathbf{C}$ is an $n \times n$ diagonal matrix, whose diagonal elements are the components of the vector $\mathbf{c}$. That is, if $\mathbf{C} = [\gamma_{ij}]_{n \times n}$, then $\gamma_{ij} = \delta_{ij} c_i$, where δ_{ij} is the Kronecker delta.

The objective function y can be formulated by using equations (4) and (6)

$$y = 1/z + \mathbf{\Lambda}^T(\mathbf{B}(\mathbf{Cv})) \tag{7}$$

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. Initially, a set of random values in $[0, 1]$ corresponding to c_i's are generated. Then c_i's are determined iteratively through a new learning technique based on second order derivatives using equation (3) [5],

$$\Delta c_i = -\frac{\partial y}{\partial c_i} / |\frac{\partial^2 y}{\partial c_i^2}| \tag{8}$$

Thus by contrast to simple gradient descent, this second order gradient method is independent of the learning parameter. This being a modified version of the Newton's method of weight updating, uses the second-order derivative in addition to the gradient to determine the next updating direction and step size. The modulus of the second order derivative in the denominator of equation (8) indicates the amount of updation necessary to reach the optima and prevents it from converging in the wrong direction. Thus the modified value of c_i is

$c_i(t+1) = c_i(t) + \Delta c_i,\ \forall i,\ t = 0, 1, 2, \ldots\ c_i(t+1)$ is the value of c_i at iteration $(t+1)$, which is computed based on the c_i-value at the iteration t.

Regularization parameter λ is chosen empirically from 0.1 to 1.0 in steps of 0.1. The c_i-values for which y attains a minimum value at a particular λ value is observed. c_i attains values between 0 to 1 as mentioned previously corresponding to some values of v_i and is negligible for other values of v_i. We take into account the values of c_i's that are close to 1, corresponding to the minimum value of y. This enables us to identify the optimal regulatory pathway yielding the maximal expression of the target gene starting from the initial gene.

4 Experimental Results and Comparison

T Helper Cell Network

The vertebrate immune system in Fig. 1 is made of diverse cell populations; some of them are antigen presenting cells, natural killer cells, B and T lymphocytes. There is a subpopulation of T lymphocytes, the T-helper, or Th, cells that have received much attention from the modeling point of view. Th cells can be divided into precursor Th0 cells and effector Th1 and Th2 cells, depending on the pattern of secreted molecules. Th1 and Th2 cell types play a central role in cellular immunity and humoral responses, respectively. The regulatory network presented constitutes the most extensive attempt to model the regulatory network controlling the differentiation of Th lymphocytes to date, and it has been implemented both as a discrete and a continuous dynamical system. Here the starting gene is TCR and the target gene is $STAT3$ [8]. There are 33 reactions and 23 genes in the network. The average amount of protein synthesis z for this network is $z = c_{21}v_{21} - c_{33}v_{33}$. The optimal pathway obtained by the new learning rule is $v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11} \rightarrow v_{12} \rightarrow v_{22} \rightarrow v_{27} \rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$. The extreme regulatory pathway obtained by EPA is different from that obtained by the present method and is as follows $v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11} \rightarrow v_{12} \rightarrow v_{30} \rightarrow v_{15} \rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$. The biological significance of the sequence of steps that leads to the optimal path can be found in [4].

The set of all pathways from the starting gene TCR to the target gene $STAT3$ along with c-values and the average amount of the protein synthesized (z) by the target gene are shown in Table 1. The pathway corresponding to serial number 1 yields the highest average z and the corresponding c-values for this pathway is large compared to the c-values of other pathways. The results show that the second order optimization method is able to correctly identify the optimal regulatory pathway.

Transcriptional regulatory network of *E. coli*

The transcriptional network of *E. coli* is the most complete experimentally characterized network of a single cell [7]. The regulatory network in Fig. 2 is important in oxidative stress response of plant cells and is also an important component in the acid resistance system of *E. coli*. In Fig. 2A, the starting gene is *crp* and

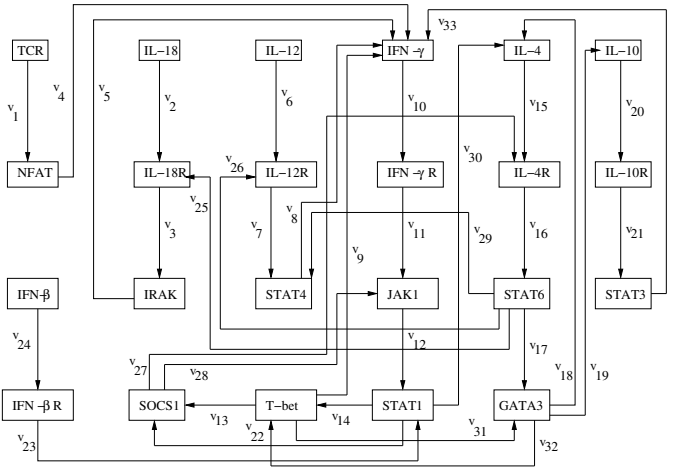

Fig. 1. Th Cell Gene Regulatory Network

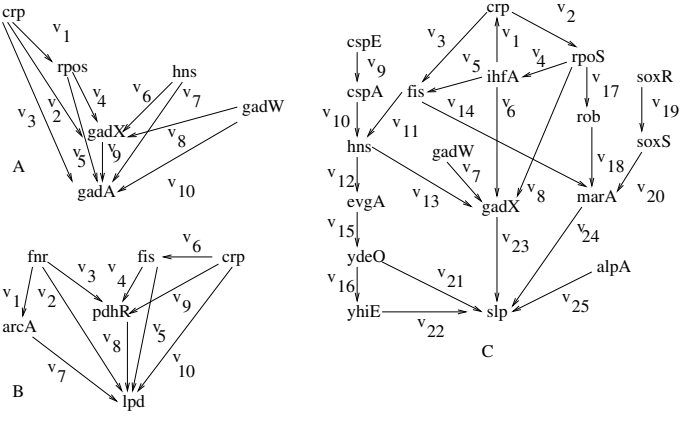

Fig. 2. The three complex regulatory circuits of the extended transcriptional regulatory network of *E. coli*

the target gene is *gadA*. The starting gene in part B, is fnr and the target gene is *lpd*, and in part C the starting gene is *crp* and the target gene is *slp*. The target gene gadA in part A codes for glutamate decarboxylase, an important metabolic enzyme in the gammaaminobutyric acid (GABA) shunt.

The expression of z for the network in Fig. 2A is $z = c_3v_3 + c_5v_5 + c_7v_7 + c_9v_9 + c_{10}v_{10}$. The corresponding expressions for part B is $z = c_2v_2 + c_5v_5 + c_7v_7 + c_8v_8 + c_{10}v_{10}$ and for part C is $z = c_{21}v_{21} + c_{22}v_{22} + c_{23}v_{23} + c_{24}v_{24} + c_{25}v_{25}$. The optimal pathway obtained for Fig. 2A is $p_1 : v_1 \rightarrow v_4 \rightarrow v_9$. Similarly, the optimal pathways are $p_1 : v_3 \rightarrow v_8$ for part B and $p_1 : v_2 \rightarrow v_{17} \rightarrow v_{18} \rightarrow v_{24}$

Table 1. c-values and z-values for the Th regulatory network in Fig. 1

Sl. No.	Some possible paths	Optimal c-values	Average quantity (z) of protein synthesis
1	$v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11}$ $\rightarrow v_{12} \rightarrow v_{22} \rightarrow v_{27}$ $\rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19}$ $\rightarrow v_{20} \rightarrow v_{21}$	$c_1 = 0.97, c_4 = 0.92, c_{10} = 0.89$ $c_{11} = 0.87, c_{12} = 0.81, c_{22} = 0.95$ $c_{27} = 0.96, c_{16} = 0.93, c_{17} = 0.92$ $c_{19} = 0.86, c_{20} = 0.85, c_{21} = 0.84$	53.89
2	$v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11}$ $\rightarrow v_{12} \rightarrow v_{30} \rightarrow v_{15}$ $\rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19}$ $\rightarrow v_{20} \rightarrow v_{21}$	$c_1 = 0.97, c_4 = 0.92, c_{10} = 0.89$ $c_{11} = 0.87, c_{12} = 0.81, c_{30} = 0.52$ $c_{15} = 0.45, c_{16} = 0.93, c_{17} = 0.92$ $c_{19} = 0.86, c_{20} = 0.85, c_{21} = 0.84$	21.45
3	$v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11}$ $\rightarrow v_{12} \rightarrow v_{14} \rightarrow v_{13}$ $\rightarrow v_{27} \rightarrow v_{16} \rightarrow v_{17}$ $\rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$	$c_1 = 0.97, c_4 = 0.92, c_{10} = 0.89$ $c_{11} = 0.87, c_{12} = 0.81, c_{14} = 0.25$ $c_{13} = 0.16, c_{27} = 0.96, c_{16} = 0.93$ $c_{17} = 0.92, c_{19} = 0.86, c_{20} = 0.85$ $c_{21} = 0.84$	18.72
4	$v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11}$ $\rightarrow v_{12} \rightarrow v_{14} \rightarrow v_{31}$ $\rightarrow v_{19} \rightarrow v_{20} \rightarrow v_{21}$	$c_1 = 0.97, c_4 = 0.92, c_{10} = 0.89$ $c_{11} = 0.87, c_{12} = 0.81, c_{14} = 0.25$ $c_{31} = 0.05, c_{19} = 0.86, c_{20} = 0.85$ $c_{21} = 0.84$	10.92
5	$v_1 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{11}$ $\rightarrow v_{12} \rightarrow v_{14} \rightarrow v_{31}$ $\rightarrow v_{32} \rightarrow v_{13} \rightarrow v_{27}$ $\rightarrow v_{16} \rightarrow v_{17} \rightarrow v_{19}$ $\rightarrow v_{20} \rightarrow v_{21}$	$c_1 = 0.97, c_4 = 0.92, c_{10} = 0.89$ $c_{11} = 0.87, c_{12} = 0.81, c_{14} = 0.25$ $c_{31} = 0.05, c_{32} = 0.11, c_{13} = 0.16$ $c_{27} = 0.96, c_{16} = 0.93, c_{17} = 0.92$ $c_{19} = 0.86, c_{20} = 0.85, c_{21} = 0.84$	15.56

for part C. The extreme regulatory pathways derived by EPA method for Fig. 2A and B are the same as derived from our proposed second order algorithm. We have obtained a different extreme regulatory pathway by EPA method $v_3 \rightarrow v_{11} \rightarrow v_{12} \rightarrow v_{15} \rightarrow v_{16} \rightarrow v_{22}$ for the network in Fig. 2C.

This application is particularly challenging as the model bacterium *E.coli* has direct experimental supports. The pyruvate dehydrogenase (PDHR) (the intermediate gene in Fig. 2B) complex of *E.coli* is a master regulator of the genes involved in the main pathway [10]. PdhR is an important regulator for the steady state maintenance of the central metabolism for energy production. The target gene *lpd* in Fig. 2B functions as glycine cleavage system L protein. The target gene *slp* in Fig. 2C is regulated by 17 regulators. These regulators participate in cellular responses to various environmental conditions, such as oxidative stress (soxRS), acid stress (gadW, gadX, evgA, ydeO and yhiE), cold shock (cspE, cspA) and multiple antibiotic resistance (marA). This underlies the importance of this gene in stress response.

5 Conclusions

This paper proposes a novel second-order learning algorithm for exploring gene regulatory networks in which the underlying optimal regulatory pathways from a starting gene to a target gene can be determined in terms of concentration of various transcription factors regulating the genes in the network. The second order system can be shown to have much better convergence properties than the first order gradient descent system. The entire method is based on well known flux balancing approach [6]. The methodology presented here was tested on two

biologically significant networks, the T helper cell network and the regulatory network of *E. coli*. The results demonstrate the effectiveness of the methodology in retrieving biologically valid regulatory relations and providing meaningful insights for better understanding the dynamics of gene regulatory networks. We can expect that as more and more gene regulatory networks are reconstructed, the second-order analysis will gradually emerge as an important paradigm for studies of complex biological systems.

References

1. Castillo, E., Berdinas, B.G., Romero, O.F., Betanzos, A.A.: A very fast learning method for neural networks based on sensitivity analysis. Journal of Machine Learning Research 7, 1159–1182 (2006)
2. Dreyfus, G.: Neural Networks: Methodology and Applications. Springer, Heidelberg, Germany (2005)
3. Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105 (2003)
4. Garg, A., Di Cara, A., Xenarios, I., Mendoza, L., De Micheli, G.: Synchronous versus asynchronous modeling of gene regulatory networks. Bioinformatics 24, 1917–1925 (2008)
5. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan College Publishing Co. Inc., New York (1994)
6. Lee, J.M., Gianchandani, E.P., Papin, J.A.: Flux balance analysis in the era of metabolomics. Briefings in Bioinformatics 7, 1–11 (2006)
7. Ma, H., Kumar, B., Ditges, U., Gunzer, F., Buer, J., Zeng, A.P.: An extended transcriptional regulatory network of escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 32, 6643–6649 (2004)
8. Mendoza, L., Xenarios, I.: A method for the generation of standardized qualitative dynamical systems of regulatory networks. Theoretical Biology and Medical Modelling 3, 1–18 (2006)
9. Mizutani, E., Dreyfus, S.E.: Second-order stagewise backpropagation for hessian-matrix analyses and investigation of negative curvature. Neural Networks 21, 193–203 (2008)
10. Ogasawara, H., Ishida, Y., Yamada, K., Yamamoto, K., Ishihama, A.: Pdhr (pyruvate dehydrogenase complex regulator) controls the respiratory electron transport system in escherichia coli. Journal of Bacteriology 189, 5534–5541 (2007)
11. Parlos, A.G., Femandez, B., Atiya, A.F., Muthusami, J., Tsai, W.K.: An accelerated learning algorithm for multilayer perceptron networks. IEEE Transactions on Neural Networks 5, 493–497 (1994)
12. Schilling, C.H., Letscher, D., Palsson, B.O.: 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 (2000)
13. Wang, Y.J., Lin, C.T.: A second-order learning algorithm for multilayer networks based on block hessian matrix. Neural Networks 11, 1607–1622 (1998)
14. Xiong, M., Zhao, J., Xiong, H.: Network-based regulatory pathways analysis. Bioinformatics 20, 2056–2066 (2004)