Identification of Gene Regulatory Pathways: A Regularization Method

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

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

Abstract. Network based pathways are emerging as an important paradigm for analysis of biological systems. In the present article, we introduce a new method for identifying a set of extreme regulatory pathways by using structural equations as a tool for modeling genetic networks. The method, first of all, generates data on reaction flows in a pathway. A set of constraints is formulated incorporating weighting coefficients. The effectiveness of the present method is demonstrated on two genetic networks existing in the literature. A comparative study with the existing extreme pathway analysis also forms a part of this investigation.

Keywords: flux balance analysis, gene regulatory networks, apoptosis, incidence matrix, yeast cell cycle.

1 Introduction

The abundance of genomic data currently available has led to the construction of genome-scale models of metabolism [1]. Recently several mathematical and computational approaches for defining the functions of networks have emerged. These properties analyze the systems properties of networks and move beyond the traditional pathway definitions present in biochemistry networks. A genetic network being a dynamic system provides important information on how a biological network changes from one state to another. But they need extensive quantitative information which is difficult to obtain [2]. The development of dynamic models of genetic networks is severely hampered due to the lack of experimental procedures to measure the dynamic quantities.

Due to the recent advances in genomics, the reconstruction of the networks of microorganisms has become feasible by using biochemical knowledge and information from genetic databases. However the analysis of such large scale systems remains a major challenge in computational biology. To study complex biological networks that are assumed to operate in the steady state it is necessary to develop a mathematical framework [3]. The stationary state condition allows for detecting routes in the system which are coupled with the stoichiometric coefficients. Constraint based approaches have become a major tool to analyze

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 425–432, 2007.

the network of microorganisms [4]. Pathway analysis is becoming increasingly important for assessing inherent network properties of biochemical reaction networks [5]. Of the two most promising concepts for pathway analysis, one relies on elementary flux modes [6] and the other on extreme pathways popularly known as flux balance analysis. Flux balance analysis [7] is based on the fundamental law of mass conservation and the application of optimization principles to determine the optimal distribution of resources within a network.

Here we develop a method for identification of extreme regulatory pathways in genetic networks. The method, first of all, generates the possible flow vectors in the pathway. We consider only those flow vectors which, by taking convex combination of the basis vectors spanning the null space of the given node-edge incidence matrix, satisfy the quasi-steady state condition. Then a set of weighting coefficients is incorporated. A set of constraints incorporating these weighting coefficients is formulated. An objective function, in terms of these weighting coefficients, is formed, and then minimized under regularization method. The weighting coefficients corresponding to a minimum value of the objective function represent the extreme regulatory pathway. The effectiveness of the present method is demonstrated on two genetic systems designed in [8]. The method is compared with the existing extreme pathway analysis [9].

2 Genetic Network Model

Most of the structural and regulatory analyses consider the networks as unweighted (directed or undirected) graphs, with the genes as nodes and the interactions among them as edges [10]. The limitations of these methods is that the strengths of the interactions, the actual magnitude of flow through individual genes, the concentration of intermediates, and the allosteric interactions known to be crucial to the regulation of intracellular biochemistry are not considered. Gene regulatory networks [11] is based on a map of allosteric interactions, and are composed of individual elements that interact with each other in a complex fashion to regulate and control the production of proteins necessary for cell function. There are two important aspects of every genetic network that have to be modeled and analyzed. The first is the topology (connectivity structure) and the second is the set of interactions between the elements, i.e. determining the dynamical behavior of the system. A genetic network can be represented as a directed graph where the nodes represent genes and the directed edge represents the regulatory relationship between two connected genes. Let g_i be the expected 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 measures the amount of expression of gene i transported through the edge (i, j). The regulatory coefficient measures the regulatory strength between two connected genes. A system boundary can be drawn around a network which consists of both internal and exchange flows. There are n_I number of internal flows and n_E number of exchange flows. The k-th internal flow is denoted by V_k and the l-th exchange flow is denoted by b_l. The internal flows are constrained to be positive and the exchange

flows are either positive, negative or either positive or negative depending on the flow of the gene across the network.

2.1 Linear Structural Equation Model

The structure and dynamics of biochemical reaction networks, at the lowest level of detail, we distinguish the stoichiometric structure of a biochemical reaction network. It is a description of all biochemical conversions that take place in the network (e.g., of catalysis, transport, and binding). It represents the topology of mass flow through the network. It does not incorporate inhibitory and activatory effects of allosteric effectors. Linear structural equations can be used for constructing a first order approximation model of a genetic network using steady state gene expression measurements [3]. Let $\mathbf{g}$ denote the expression levels of the genes in the network and $\mathbf{f}$ denote the vector of non-linear functions. Rate equations indicating the expression levels of the genes in the network are given in a simplified form as [12].

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

where $\mathbf{u}$ is the set of transcriptional perturbations. When the system reaches a steady state which is equivalent to setting the time derivative of $\mathbf{g}$ to zero, the system can be approximated by a linear set of equations

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

where $X = [\mathbf{g}^T, \mathbf{u}^T]^T$. The above equation can be mathematically formulated

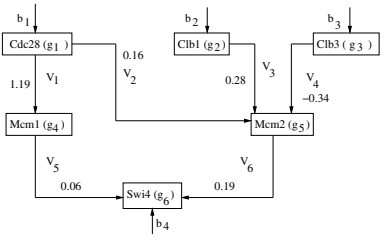

Fig. 1. Path diagram for a genetic network reconstructed from yeast cell cycle data

by a node-edge incidence matrix, B where the column in the matrix associated with edge (i, j) contains a '-1' in row i, a '+1' in row j and zeros elsewhere. The rank of the matrix B is equal to the number of genes in the network. So we decompose the matrix A into $A = BY$. As the number of columns in B is quite large than the number of rows the above decomposition may not be unique. Let $V = YX$. V is the vector of flows consisting of both internal and exchange flows. Thus at steady state, we get

$$BV = 0 \tag{3}$$

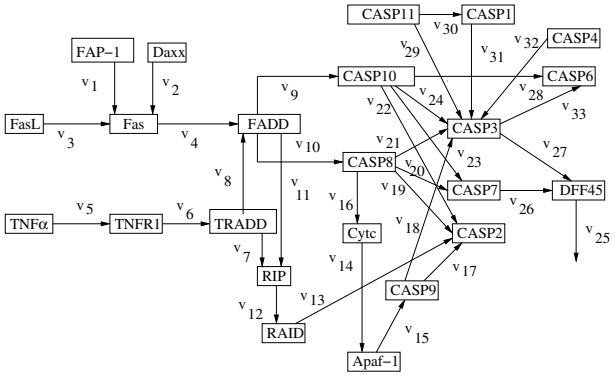

Fig. 2. Path diagram for apoptotic genetic network

which indicates the flow balance equations for the network. We model genetic networks by introducing two path diagrams (Fig. 1) and (Fig. 2) [8] which is a directed graph representing a system of structural equations. The path diagram consists of nodes represented by letters and edges represented by lines. The directed edges between the nodes denote the direction of the regulatory relationship between the nodes connected by the edges and this indicates a directed regulatory influence of one gene on another. The directed edges can represent either activation (positive control) or inhibition (negative control). The genetic network in the Fig. 1 reconstructed from the yeast cell cycle data consists of 6 genes from which the node-edge incidence matrix B can be constructed. Here for convenience of presentation we arrange the matrix B so that the first series of columns represent the internal flows and the remaining columns represent the exchange flows. All internal flows are positive yielding, $v_i \geq 0, i\epsilon n_I$. Like the stoichiometric matrix S in metabolic networks, the node-edge incidence matrix B plays a similar role in genetic networks. Here the b_{ij} element of the node-edge incidence matrix is the coefficient of the i-th gene in the regulatory process j. Here we further give an explanation for interpreting extreme directions as extreme pathways and develop a regularization approach for generating these pathways for genetic networks. Any cycle or a path having a starting point with entering exchange flow and an ending point with exiting exchange flow is an extreme direction, and, is referred to as an extreme regulatory pathway. The genes $Cdc28$ and $Clb1$ have an entering exchange flow. So $b1$ and $b2$ are entering roots. As the regulatory coefficient of the gene $Clb3$ on the gene $Mcm2$ is negative, the actual flow of the gene is from gene $Mcm2$ to the gene $Clb3$ which implies exchange flow $b3$ is negative and hence it is an exiting root. To balance the flow at the gene $Swi4$, the exchange flow $b4$ is negative and hence $b4$ is an exiting root.

3 Proposed Method

We consider the genetic network in (Fig. 1) with starting genes as $Cdc28$ and $Clb1$ and the target gene as $Swi4$ [12]. The target gene can be reached through

s different paths. That is, there are s flows/paths $V_1, V_2, \ldots, V_s$ in the network involving $Swi4$. We take the algebraic sum of the weighted flows of reactions $V_1, V_2, \ldots, V_s$ to reach the target $Swi4$ and it is given by

$$z = \sum_{k=1}^{s} c_k v_k \tag{4}$$

Let us also consider that there are n flows comprising of both internal and exchange flows and m genes in the network. Here v_k is the gene flow involving genes $Swi4$ and $Clb3$. The term c_k denotes the weighting factor corresponding to the flow V_k. Here we have considered the genetic network where there is no feedback loop. The role of c_i in extreme pathway analysis is different from our method. In the earlier case, $\mathbf{c}$ is a *unit vector*, along a particular flow of the gene, whereas in the present method, $\mathbf{c}$ indicates the connection of other transcription factors (not shown in the diagram). The same procedure is applied for the genetic network in (Fig. 2) [8] where the starting genes are $FasL$ and $TNF\alpha$ and the target gene is $DFF45$.

3.1 Generation of Gene Flow Vectors

We require the values of the gene flow vectors $\mathbf{v} = [v_1, v_2, \ldots, v_n]^T$. We propose a method for generating flow vectors that approximately satisfies the quasi-steady state condition. That is, we generate those $\mathbf{v}$ which satisfies

$$\mathbf{B}.\mathbf{v} \approx \mathbf{0} \tag{5}$$

where $\mathbf{B}$ is the $m \times n$ node-edge incidence matrix that describes the relationship between genes and their regulatory interactions. $\mathbf{B}$ is computed from the diagram. As $m > n$, equation (5) is under determined. So we proceed in the following ways:
Step I: Generate basis vectors $\mathbf{v}_b$ that form the null space of the node-edge incidence matrix $\mathbf{B}$. Let the number of such basis vectors be l. (This is done by standard functions available in MATLAB).
Step II: Generate l number of random numbers a_p, $p = 1, 2, \ldots, l$. Then generate a vector $\mathbf{v}$ as a linear combination of the basis vectors using a_p.

3.2 Formulation of a New Constraint

As the genes are not expressed at the required level there comes further restrictions on the system, and we define a new constraint as

$$\mathbf{B}.(\mathbf{C}.\mathbf{v}) = \mathbf{0} \tag{6}$$

$\mathbf{C}$ 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. Thus the problem of determining the extreme regulatory pathways starting from the genes $Cdc28$ and $Clb1$ to the target genes $Swi4$ and $Clb3$ boils down to an optimization problem, where z has to be optimized with respect to $\mathbf{c}$, subject to the inequality constraints and the new constraint.

3.3 Estimation of Weighting Coefficients c_i

Combining equations (4) and (6), we can reformulate the objective function as

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

that needs 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). Initially, a set of random values in $[0, 1]$ corresponding to c_i's are generated. Then c_i's are modified iteratively using gradient descent technique, where the amount of modification for c_i in each iteration is defined as

$$\Delta c_i = -\eta \frac{\partial y}{\partial c_i} \tag{8}$$

The term η is a small positive quantity indicating the rate of modification. 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. We now analyze the results in Section 4.

4 Results

Following the method described in Section 3.1 we have generated a set of flow vectors. Then the objective function y (Equation(7)) is minimized, where the expression for z is defined as $z = c_5 v_5 + c_6 v_6 + c_{10} v_{10}$ for the genetic network in the (Fig. 1) and for the genetic network in the (Fig. 2) the expression for z is defined as $z = c_{26} v_{26} + c_{27} v_{27} - c_{25} v_{25}$. We vary the value of λ from 0.1 to 1.0. Initially λ should be kept small. For each value of λ, we minimize y, and consider that set of c_i-values corresponding to λ as the final solution, for which y becomes minimum. The genetic network in the (Fig. 1) reconstructed from the yeast cell cycle data consists of 6 genes where we have obtained the 5 extreme regulatory pathways as $p_1 : g_1 \rightarrow g_4 \rightarrow g_6, p_2 : g_1 \rightarrow g_5 \rightarrow g_6, p_3 : g_2 \rightarrow g_5 \rightarrow g_6, p_4 : g_1 \rightarrow g_5 \rightarrow g_3, p_5 : g_2 \rightarrow g_5 \rightarrow g_3$. The genetic network in the (Fig. 2) reconstructed from the yeast cell cycle data consists of 23 genes where we have obtained the 2 extreme regulatory pathways as $p_1 : v_3 \rightarrow v_4 \rightarrow v_{10} \rightarrow v_{20} \rightarrow v_{26}, p_2 : v_5 \rightarrow v_6 \rightarrow v_8 \rightarrow v_{10} \rightarrow v_{16} \rightarrow v_{14} \rightarrow v_{15} \rightarrow g_{21} \rightarrow v_{18} \rightarrow v_{27}$.

These are the two major experimentally confirmed pathways (extrinsic and intrinsic apoptosis pathways) [13]. The pathway p_1 involves response to binding of death ligands to their receptor FasL which triggers apoptosis via activating FADD and CASP8. Activation of complex of ligand receptor, FADD and CASP8 leads to the formation of a death inducing signaling complex (DISC), which in turn activates downstream effecters CASP7 and DFF45, resulting in DNA fragmentation. The seond apoptotic initiator pathway p_2 is induced by the formation of cytochrome c (Cytc) released from mitochondria with the adaptor Apaf-1, which in turn activates CASP9 and CASP3. The gene CASP3 will again trigger DNA fragmentation factor DFF45 and lead to DNA fragmentation.

A structure is the most essential feature of the networks. It provides information to assess the function of the gene. The extreme regulatory pathways helps to identify key components of the network structure and evaluate the relative importance of the gene in the network. These regulatory flows play an important role in apoptosis. The flows from the gene Fas to the gene FADD, and from the gene TRADD to the gene FADD involve responses to DISC [13]. Binding of death ligands to their receptor FADD activates the genes CASP8 and CASP10 and initiates a major extrinsic pathway. The genes CASP3, FADD, CASP8 and CASP10 are essential in apoptosis. A large number of extreme pathways are lost in apoptosis network due to the deletion of the above mentioned genes. The gene CASP3 is a major effecter gene and carries out the majority of substrate proteolysis during apoptosis [13]. FADD, CASP8 and CASP10 participate in DISC formation and play important roles in apoptosis initiation.

These pathways determines the gene regulatory route leading from the transcription of a given gene to the transcription of another gene. Genes communicate (interact) via the proteins they encode and protein production (transcription and translation) is controlled by a series of biochemical reactions, which are in turn influenced by many factors, both internal and external to the cell. In the case of metabolic networks, we can directly find a link between the concentration of the starting metabolite with that of the target metabolite, but it is not the same in case of genetic networks. Here the amount of mRNA produced by transcription and hence the amount of protein synthesized by translation by a starting gene will affect the protein synthesis of the target gene but they have no direct link between them. The gene flows at the steady state are combinations of the gene flows of the set of extreme regulatory pathways. We also use the algorithms developed in [9] for the genetic networks in (Figs. 1, 2) for generating extreme regulatory pathways. The extreme pathways generated by these two methods are the same for the genetic network in (Fig. 1). The extreme regulatory pathways generated by [9] and our regularization method is the same for the genetic network in (Fig. 2).

5 Conclusions and Discussions

The extreme regulatory pathways represent the regulatory capabilities i.e. the structural and functional properties of the genetic network [14]. These pathways determine the route starting from a particular gene to a given target gene. It is the set of interactions occurring between a group of genes which depend on each other's individual functions in order to make the aggregate function of the network available to the cell. The decrease in network functionality due to deletion of the genes and identifying the most important genes in a network can be done as these pathways indicate the flexibility and robustness of the networks. Selecting a model for a genetic network can have a great impact on how well the model follows the underlying dynamics of the actual genetic network. The balance between the available data, estimation techniques and the model complexity determines the usefulness of a given model.

Here we have given a brief overview of the methods and modeling descriptions available in computational systems biology. With the ability to reconstruct genetic networks on a large scale, the need to develop network-based pathway definitions and pathway analysis procedures has grown. We need to bring such pathway definitions into biological reality and use them productively to enhance our understanding of the systemic functions of real reconstructed networks.

References

1. Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., Palsson, B.O.: Metabolic pathways in the post-genome era. Trends in Biochemical Sciences 28, 250–258 (2003)
2. Gardner, T.S., Bernardo, D., Lorenz, D., Collins, J.J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105 (2003)
3. Datta, S.: Exploring relationships: a partial least square approach. Gene Exp. 9, 257–264 (2001)
4. Urbanczik, R., Wagner, C.: An improved algorithm for stoichiometric network analysis: theory and applications. Bioinformatics 21(7), 1203–1210 (2005)
5. Schilling, C.H., Edwards, J.S., Letscher, D.: Combining pathway analysis with flux balance analysis fot the comprehensive study of metabolic systems. Biotechnology and Bioengineering 71(4), 286–306 (2000)
6. Klamt, S., J., G., Kamp, A.V.: Algorithmic approaches for computing elementary modes in large biochemical reaction networks. IEEE Proc.- Syst. Biol. 152(4), 249–255 (2005)
7. Klamt, S., Stelling, J.: Two approaches for metabolic pathway analysis? Trends in Biotechnology 212(2), 64–69 (2003)
8. Xiong, M., Zhao, J., Xiong, H.: Network-based regulatory pathways analysis. Bioinformatics 20, 2056–2066 (2004)
9. 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)
10. Kriete, A., Elis, R. (eds.): Computational systems Biology. Elsevier, San Diego, California, USA (2006)
11. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of escherichia coli. Nat. Genet. 31, 64–68 (2002)
12. Xiong, M.M., Li, J., Fang, X.Z.: Identification of genetic networks. Genetics 166, 1037–1052 (2004)
13. Shivapurkar, N., Reddy, J., Chaudhary, P.M., Gazdar, A.F.: Apoptosis and lung cancer: A review. J. Cell. Biochem. 88, 885–898 (2003)
14. Covert, M.W., Schilling, C.H., Palsson, B.O.: Regulation of gene expression in flux balance models of metabolism. Journal of theoretical biology 223, 73–88 (2001)