The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC, pp. 35–44

An Optimization Model for Gene Regulatory Network Reconstruction with Known Biological Information*

Jinshan Li[1,2,†] Xiang-Sun Zhang[2]

[1] College of Science, Beijing Forestry University, Beijing 100083, China
[2] Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing 100080, China

Abstract Grounded on linear ordinary differential formulation, we propose an optimization model for inferring gene regulatory networks, which integrates with not only the sparsity principle of gene networks but also some extra priori knowledge between two genes that can be found in existing publications or biological web sites. The model is applied to an artificial data-set and a real gene expression profile data related to breast cancer metastasis, and the computational outcome shows that this model can find solution with biological plausibility and reliability in a sense.

Keywords Optimization model; gene regulatory networks; microarray technique; known interaction between genes

1 Introduction

It is believed that knowledge of mRNA levels under different conditions can help people understanding how the expression levels of each gene depend on an external stimuli and on the expression levels of other genes. With high throughput experimental methods, such as DNA microarrays, mRNA expression levels of a group of genes can be measured simultaneously[10]. While the amount of available gene expression data has been increasing rapidly, the required mathematical techniques to analyze such data is still in development and inferring a gene regulatory network from gene expression data has been proved to be difficult.

Therefore, the model of construction of gene regulatory networks has become one of important topics in bioinformatics. Logically the large number of regulatory components requires a large experimental data to infer the network structure. Recently, DNA microarrays has become one of the main tools in this research areas[16]. Microarray technique enables people to monitor the activities of thousands of genes in parallel but only at a few of time instants. To reverse-engineering the regulatory networks from these microarray-measured data sets, one has to hunt for a valid computational model that can create a candidate network that yields a similar time series

*This work is partially supported by the National Natural Science Foundation of China (Grant No.10631070) and the Ministry of Science and Technology of China (Grant No.2006CB503905).

†Email:lijinshan@gmail.com

data. In other words, the model is able to reflect the true regulatory networks, i.e. the dependencies of the biological components.

There are several approaches applied to address this problem. Although some of them based only on the distance between the real data and the simulated data from the mathematical model can excavate some biological knowledge in a sense, some important and known information of the systems has been neglected, e.g. the sparsity of the gene network, that is, one gene only depends on the average on a small proportion of the component in the system, and the knowledge of interaction, that is, some interactions between genes have already been known from existing publications or biological web sites.

In this paper we propose an optimization model for reverse engineering of time series data, and apply it to a simulated data and a real time-course gene expression data related to breast cancer metastasis (see the table in appendix in [8]). The model is a constrained mathematical programming based on the ordinary differential equations which have been widely used to analyze genetic regulatory systems. Firstly, a special solution in terms of least L_2 norm is obtained by singular valued decomposition (SVD) technique based on the ordinary differential equations. Then a mathematical programming is set to improve the special solution. That is, the math program is to be optimized to gain a regulatory network, which fits the data and realizes the sparse connection of the system, and the known interactions between genes is reflected by the maximization of objective functions.

The rest of this paper is organized as follows. Section 2 presents an overview over related works and a list of associated publications. In section 3, we detail the method proposed in this paper. And its application to a simulated data and a real biological microarray data will be shown in section 4. Conclusion and an outlook on future research will be given in the last section.

2 Related Work

Understanding the mechanisms of gene regulatory system is very interesting and enables researchers derive the underlying networks. In this section, a brief description of related works is given.

The first computational models for inferring gene network are boolean or random boolean networks [5, 6, 15, 2]. In contrast to discrete models such as boolean networks, continuous models in publication allow for the expression of gene regulation to be continuous. An example for this kind of approach is the differential equation model given in [17, 19, 11]. Another popular model for inferring gene networks is the Bayesian networks or dynamic Bayesian networks[13, 12, 9]. The computational biology literature abounds in various modelling approaches, all of which have particular goals along with their strengths and weaknesses[4].

Particularly, linear differential equations are attractive because of their lower number of parameters which imply that we are less likely to over-fit data and sufficiency of modeling complex interactions between genes. Although gene regulations are often nonlinear in nature, almost all of the existing approaches for GRN inference use linear or additive models due to unclear structures of biological systems and scarcity of data [14, 3, 8].

The idea of modeling gene expression data by differential equations has been explored by a considerable numbers of authors [17, 11, 19]. Differential equations are used to model gene interactions under the assumption that the transcription rate over time of each gene expression level is a function of the expression level of some (usually a few) other genes. Such modeling assumption is based on the reaction kinetics at the biochemical level.

An optimization model based on linear ordinary differential equations, defining a reverse engineering to tackle this problem, is considered in [8]. More precisely, the optimization model is introduced to realize the sparsity of gene regulatory system, which is helpful for inferring GRNs.

3 Method

A simple linear model has proved to be useful in a number of cases [14] even if it is clear that nonlinearity is an unavoidable issue since it reflects also the nature of biochemical interactions. From the viewpoint of dynamical systems, linear equations can at least capture the main features of the network or the function. Therefore, as in [18, 8], we will assume the system to behave linearly. We consider the linear system described by the following differential equations:

$$\dot{x_i}(t) = -\lambda_i x_i(t) + \sum_{j=1}^{N} W_{ij} x_j(t) + b_i(t) + \varepsilon_i(t) \tag{1}$$

for $i = 1, 2, \cdots, N$, where the state variables x_i's are the concentration of mRNA of gene i, the λ_i's are the self-degration rates, the b_i's are the external stimuli, or environment conditions, which are set to zero when there is no external input, and the ε_i's represent noise. W_{ij} describe the type and strength of the effect of the j-th gene on the i-th gene with a positive, zero or negative sign indicating activating, naught and repressing influence respectively. However, we sometimes do not have the information of the external stimuli, that is to say, b_i is nonexisting, so (1) changed to be

$$\dot{\mathbf{x}} = A\mathbf{x} + \varepsilon \tag{2}$$

in a usual compact form, where the matrix A is an $N \times N$ matrix which incorporates both self-degradation rates and the strength of the gene-to-gene interaction.

Microarray experiments often result in discrete time series of measured values. We assume that the number of measured time points to be m, $t_1, t_2, \cdots, t_m$, and (2) can be described in a discrete form without the error part in our model,

$$\Delta X = AX, \tag{3}$$

where the element in the $N \times (m-1)$ matrix ΔX is $\triangle x(t_i) = (x(t_i) - x(t_{i-1}))/(t_i - t_{i-1}), i = 2, 3, \cdots, m$, and the element in $N \times (m-1)$ matrix X incorporates all the expressive values of every gene from time t_1 to t_{m-1}.

It is well known that the data sets created by microarray technology contain the the express levels of a large number of genes at relatively much less time points, which is the so-called dimensional problem, and leads in the solution of the above matrix equation (3) not uniquely determined. We can get a particular solution by the singular valued decomposition (SVD) technique, $X = VSU'$, where V and U are orthogonal matrix with order $N \times N$ and $(m-1) \times (m-1)$ respectively, and S is a diagonal matrix. Then

$$A_0 = \triangle XUS^{-1}V', \tag{4}$$

which is a least-L_2-norm solution. But the solution does not have the sparse property [8].

As discussed in [8], a more reliable solution of gene regulatory networks should locate near to the margin of the hyper-polygon which is the part of the solution hyperplane in the quadrant where the least-L_2-norm solution lies.

We have known that $A_0 + \alpha Y, (\forall Y \in X^{\perp} = \{Y | YX = 0\}, \forall \alpha \in \mathbb{R})$ are also the solutions of (3), in which the real solution is contained. In the following discussion, $X^{\perp}$ also represents a base of the null space. For the aim of enhancing veracity of the network, some priori regulatory information among genes are imported in our model, e.g., the known interaction between gene i_0 and j_0. Let $G(j_0)$ be the set of all the known genes who regulate gene j_0 . Since the known regulatory relations in $G(j_0)$ should be easy to determined in various biological experiments from the viewpoint of probability if they have higher strength than others, therefore we should give enough emphasis on them, and propose the following optimization model:

$$\begin{aligned} &\max_{Z \in \mathbb{R}^{N-m+1}} \sum_{i_0 \in G(j_0)} |a_{i_0 j_0} + X^{\perp}(i_0) Z| \\ s.t. \quad & sign(a_{ij_0})(a_{ij_0} + X^{\perp}(i)Z) \geq 0 \quad i = 1, 2, \cdots, N, \end{aligned} \tag{5}$$

where $A_0^T(j_0) = (a_{1j_0}, \cdots, a_{Nj_0})^T$ denotes the j_0th column of A_0^T, $X^{\perp}(i)$ is the ith vector of the null space base $X^{\perp}$, and $Z \in \mathbb{R}^{N-m+1}$, $j_0 = 1, 2, \cdots, N$, where N is the number of genes. In order to avoid the objective functions going to infinite, the class of constraints is presented, which keeps the sign of $a_{ij_0} + X^{\perp}(i)Z$ same as that of a_{ij_0}, i.e., the more reliable network will be searched in the hyper-polygon mentioned above. In the following, we assume there is only one element in $G(j_0)$, that is to say, the number of the known genes regulating gene j_0 is one. The known regulatory information provides a direction for finding a more reliable network in the process of the optimization than that in [8].

4 Application

In this section, we apply our model to an artificial data set with dimensional problem and also a real gene expression data related to breast cancer metastasis with 27 genes and 6 time points in the appendix in [1, 8]. The results of computational experiment indicates that our models can find more reliable solution which possesses biological plausibility.

4.1 Application to Artificial Data

The example is a small artificial network with five genes governed by

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \\ \dot{x}_5(t) \end{pmatrix} = \begin{pmatrix} 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & -1 & 2 \\ 2 & 0 & 0 & 0 & -1 \\ -1 & 2 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \varepsilon_3(t) \\ \varepsilon_4(t) \\ \varepsilon_5(t) \end{pmatrix} \tag{6}$$

where $x_i(t)$ is the expression level of the gene-i, $\dot{x}_i(t)$ is the transcription rate over time of $x_i(t)$, and $\varepsilon_i(t)$ is noise for $i = 1,2,3,4,5$. Clearly, every gene is regulated by two genes, one for repressing it and the other for enhancing it.

To test our model, we randomly choose the initial condition of the system and take 4 time points of $\mathbf{x}$ as a measured time-course dataset with time span 0.1 and without noise, see table 6. Clearly, the number of variables (genes) is larger than that of samples (time points), which indicates that the dataset has dimensional problem. Now we apply (5) to the artificial dataset with some known relationship such as $3 \in G(1)), 4 \in G(2), 5 \in G(3), 1 \in G(4))$, and $2 \in G(5)$ with high strength to infer the gene regulatory network . Figure 1 shows the inferred network. The red solid lines present the known relations between genes and the blue dashed lines are the inferred gene relations. The corresponding gene regulatory network is found correctly by the model, except for the last row in equation (6), but the least square solution $(-1.3434, 1.6160-0.2280, -0.1863, -0.2691)$ also reflects their regulatory relation fundamentally. In the model, for example, the maximum of the objective function for $3 \in G(1))$ is 2 and the corresponding strength $(0,-1,2,0,0)$ with other genes are correctly inferred. This artificial dataset shows that our model can find the real solution of the networks with added known interaction which provides a direction for solving the optimization problem (5).

4.2 Application to Data of Breast Cancer Metastasis

In this section the optimization model is applied to a real gene expression data related to breast cancer metastasis, containing 27 gene and 6 time points, see the table in Appendix in [8, 1]. The content of the dataset contains gene expression data of surgical samples, including both breast cancer primary tissue and metastasis tissue, collected from 30 patients in different clinical staging. The oligonucleotide microarray technique was used to identify the gene expression profiling and screen the differential expression genes in breast cancer samples with a special emphasis on

gene	$t_0 = 0$	$t_0 = 0.1$	$t_0 = 0.2$	$t_0 = 0.3$
1	-1.0000	-1.3000	-1.6850	-2.1365
2	2.0000	2.1500	2.3250	2.4925
3	-0.5000	-0.8500	-1.0950	-1.2120
4	0.5000	0.4500	0.2900	-0.0030
5	-1.5000	-1.0000	-0.4400	0.1935

Table 1: Artificial data created by simulated gene networks (6).

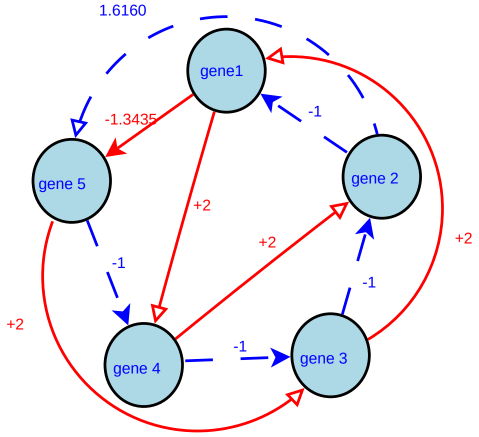

Figure 1: The optimization model is applied to infer the simulated gene network. The red solid lines present the known relations between genes and the blue dashed lines are the inferred relations. The artificial gene regulatory network is found correctly by the model, except for the last equation in equation (6), but the least L_2 solution can reflects their basical regulatory relations.

metastasis factors. 27 differential expression genes were identified, 14 of which are up-regulation genes whose Ratio is large for 3, and the rest are down-regulation gene whose Ratio is small for 0.33[1].

We now employ added priori information between two genes, besides the sparsity of biological networks, to derive the gene regulatory networks. For example, self-regulation of gene 14 and 19 will be taken as priori knowledge to infer gene networks by using equality (5). We maximize the objective function for the known information and here the feasible solution has more biological plausibility, see table 2. The left of the table lists the regulation to gene 19, and the right is to gene 14. The second row lists the multiples β who times maximum, and its corresponding column is the solution of optimization problem (5) with the condition $\beta\times$ maximum for $19 \in G(19)$ or $\beta\times$ maximum for $14 \in G(14)$. And their corresponding sub-networks which regulating gene 19 and 14 respectively are showed in the column 2 and 4 of table 2. It is worthy of attention that the two sub-networks do not contain gene 16 which regulates gene 19 and 14 as proved in [1]. But if we decrease the maximums a little respectively, i.e. 0.97× maximum for $19 \in G(19)$ or 0.98× maximum for $14 \in G(14)$, we can find that there will be not only the regulating strength of gene 16 to gene 19 and 14, which is correct and appear in [1], but also the strength of gene 22 and 23 to gene 14, which have not been found in [1]. It is showed in column 3 and 6, which might be caused by the error in biological experiments.

Gene-16 (geneID: NM-002091, name: GRP) regulating many genes is a very important one. GRP is in the information path of nerve, and a check point in cell cycle. GRP is also a transcriptional control factor on which DNA depends. GRP and its receptor usually express in cancers such as breast cancer, and play an important role in proliferation and differentiation of cell[1, 7]. The expression of GRP in the cancer organization is obviously lower than that in normal organization, and GRP influences and regulates proliferation and differentiation of cell of cancer tumor[7]. Therefor, if the optimization model find the regulation of gene GRP to the objective genes, the solution found by the model is thought as a reliable one, see figure 2. $14 \in G(14), 19 \in G(19), 25 \in G(2), 25 \in G(26)$ marked with solid line arrows are the known regulation in the model. The longer dashed lines mark the inferred gene regulation in various colors to their corresponding genes respectively, which are correct and appear in [1]. The yellow nodes denote the genes that do not appear in [1], and their regulatory relations are labeled in shorter dashed arrows.

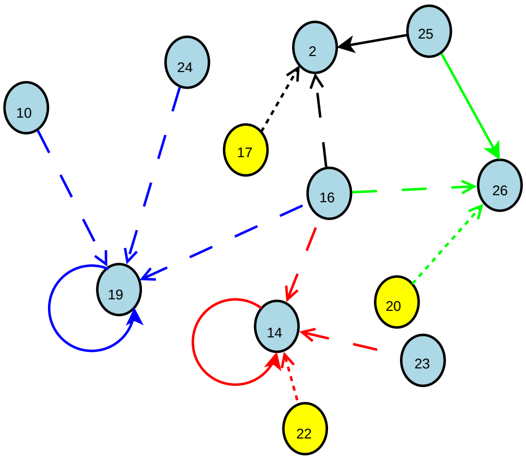

Figure 2: $14 \in G(14), 19 \in G(19), 25 \in G(2), 25 \in G(26)$ marked solid line arrows in different colors are the known regulations in the optimization model. The longer dashed arrows in various colors denote the inferred gene regulation to the different genes respectively, which appear in [1]. The yellow nodes denote the genes that do not appear in [1], and their regulatory relations are labeled in shorter dashed arrows.

	$19 \in G(19)$	$19 \in G(19)$	$14 \in G(14)$	$14 \in G(14)$
gene	$\beta = 1$	$\beta = 0.97$	$\beta = 1$	$\beta = 0.98$
1	0	0	0	0
2	0	0	0	0
3	-0.1424	-0.1646	0	0
4	0	0	0.1757	0.1267
5	0	0	0	0
6	0	0	0	0
7	0	0	0.4279	0
8	0	0	0	0
9	0	0	0	0
10	0.8005	0.8677	0	0
11	-0.2612	-0.3688	0.0760	0.4852
12	0	0	0	0
13	0	0	0	0
14	0	0	-1.6023	-1.5702
15	0	0	0	0
16	0	1.4653	4.0555	4.6255
17	0	0	0	0
18	0	0	0	0
19	-3.3628	-3.2619	0	0
20	0	0.0132	0	0
21	0	0	0	0
22	0	0	0	-2.9778
23	0	0	0	2.7693
24	2.2436	1.6618	0	0
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0

Table 2: The left of the table lists the regulation to gene 19, and the right is about the regulation to gene 14. The second row lists the multiples β who times maximum for $19 \in G(19)$ and $14 \in G(14)$ and their corresponding columns are their solution respectively, where the maximums for $19 \in G(19)$ and $14 \in G(14)$ are 3.362803, 1.6022547 respectively.

5 Outlook and Future Work

Grounded on linear ordinary differential equation, we proposed an optimization model for inferring gene regulatory networks. The model integrates with not only the sparsity property of biological networks but also some extra priori knowledge between genes. The model is applied to a simulated data and a real microarray data related to breast cancer metastasis, and the computational outcome shows that this model can find the solution with biological plausibility. The known relationship between genes provides a direction for the optimization, so the solution obtained by the model is reliable than the solution in [8] in a proper sense.

Due to the ambiguity in the data, it is difficult for our proposed models to find the true solution as concluded above. As one future enhancement of the proposed methods, we plan to incorporate some additional methods to identify the correct network. In future work more priori information will be imported into the inference process of real microarray data like partially known pathways or information about co-regulated genes, which can be found in literature or in public databases. This would enable us to search for models consistent with current biological knowledge, but would also allow for alternative solutions where biological information is missing or faulty. Furthermore, non-linear interaction will be considered in our models for enhancing precise of gene regulatory networks to overcome the insufficiency of the currently proposed models.

References

[1] C. Ni, B. Sun, Y. Feng, D. Zhang, X. Li, L. Dong and L. Zhang: Primary Research of Linear Differential Model for the Genetic Regulatory Networks of Gene Related to Breast Cancer Metastasis. Progress in Biochemistry and Biophysics, 2005, 33(12), pp. 1165–1172. (in Chinese)

[2] E. R. Dougherty and I. Shmulevich: Mappings Between Probabilistic Boolean Networks, Signal Processing, 2003, 83(4), pp. 799–809.

[3] M. Gustafsson, M. Hornquist and A. Lombardi: Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(3), pp. 254–261.

[4] H. D. Jong: Modeling and simulation of genetic regulatory systems: a literature review. Journal of computational biology, 2002, 9(1), pp. 67–103.

[5] I. Shmulevich, E.R. Dougherty, and W. Zhang: From Boolean to probabilistic Boolean networks as models of genetic regulatory networks, Proceedings of the IEEE, 2002, 90(11), pp. 1778–1792.

[6] I. Shmulevich and S. A. Kauffman: Activities and Sensitivities in Boolean Network Models. Physical Review Letters, 2004, 93(4), pp. 048701(1-4).

[7] J. A. Jensen, R. E. Carroll and R. V. Benya: The case for gastrin-releasing peptide acting as a morphogen when it and its receptor are aberrantly expressed in cancer. Peptides, 2001, 22(4): 689–699.

[8] JS Li and X.-S. Zhang: An Optimiazation Model for Achieving Sparsity of Gene Regulatory Networks. Operations research and its applications 2006, 6, 368–379. In Lecture Notes in Operations Research edited by X.-S. Zhang, D.-G. Liu, L.-Y. Wu, World publishing cooperation, Beijing.

[9] S.Y. Kim, S. Imoto and S. Miyano: Inferring gene networks from time series microarray data using dynamic Bayesian networks. 2003, 4(3), Bioinformatics, pp. 228–235.

[10] L. Pournara and L. Wernisch: Reconstruction of gene netwoeks using Bayesian learning and manipulation experiments. Bioinformatics, 2004, 20(17), pp. 2934–2942.

[11] M. de Hoon, S. Imoto, S. Miyano: Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data Using Differential Equations. Discovery Science, 5th International Conference, 2002, 2534, pp. 267–274.

[12] M. Zou, D. Suzanne and Conzen: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics, 2005, 21(1), pp. 71–79.

[13] N. Friedman, M. Linial, I. Nachman, and D. Peér: Using Bayesian networks to analyze expression data. Journal of Computational Biology, 2000, 7(3), pp. 601–620.

[14] P. D'haeseleer, X. Wen, S. Fuhrman and R. Somogyi: Linear modeling of mrna levels during cns development and injury. Pacific Symposium on Biocomputing, 1999, 4, pp. 41–52.

[15] R. F. Hashimoto, S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner and E. R. Dougherty: Growing genetic regulatory networks from seed genes. Bioinformatics, 2004, 20(8), pp. 1241–1247.

[16] M. Schena, D. Shalon, R. W. Davis, and P. Q. Brown: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, 270, pp. 467–470.

[17] T. Chen, H.L. He, and G.M. Church: Modeling gene expression with differential equations. Pacific Symposium on Biocomputing, 1999, 4, pp. 29–40.

[18] T. S. Gardner, D. di Bernardo, D. Lorentz, and J.J. Collins: Reverse engineering gene networks and identifying compound mode of action via expression profiling. Science, 2003, 301, pp. 102–105.

[19] Y. Wang, T. Joshi, X. S. Zhang, D. Xu and L. Chen: Inferring Gene Regulatory Networks From Multiple Microarray Datasets. Bioinformatics, 2006, 22(19), pp. 2413–2420.