The Tenth International Symposium on Operations Research and Its Applications (ISORA 2011)
Dunhuang, China, August 28–31, 2011
Copyright © 2011 ORSC & APORC, pp. 22–27

On Two Issues of Molecular Network Models in Systems Biology – A Review

Katsuhisa Horimoto[1]

1 Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064

Abstract Two points are definitely different between network models in systems biology and those in engineering fields. One is that most network models (structures) in biology are not static but change depending on external conditions, and the other is that the models including unmeasured variables frequently emerged so as to measure the variables under the conditions with as less perturbation to the organism as possible. To consider the two points specific to systems biology, we have designed two methods by statistical and symbolic-computation approaches, respectively. Here, we will describe the methods with their merits and pitfalls.

Keywords Biological Network Model; Network Screening; Regulatory Network; Parameter Optimization; Differential Elimination

1 Introduction

Huge information on most molecules in a cell, which can be measured by recently developed technologies such as microarray and deep sequencer, provides a chance to build and investigate the behaviors of many molecules in systems biology. In addition, recent advances of the computer performance and the improvement of computational algorithms facilitate the study of a large scale of molecular models in biology. For example, signal-response relationships in various phosphorylation pathways are well investigated [1]. In each pathway, one kind of signal from outside of cell is received by one kind of receptors on the membrane, and its signal is transduced by the phosphorylation chains of a set of defined molecules into molecules in the inner cell structure, named nucleus. In this process, more than 100 molecules are generally involved in the signal transduction. Indeed, various features of signal transduction are well known, and many lists and figures of signal transduction pathways are seen in various databases. However, it is well known that precise signal transduction pathways still remain to be solved and the finding of new molecular relationships responsible for various conditions is still one of the important issues in experimental biology.

With these situations in mind, two points are definitely different between the models in biology, especially molecular biology, and those in the fields of engineering. One is a conception issue: the molecular relationships in many biological phenomena highly depend on the external conditions. While the relationships may be kept in the basic metabolic pathways such as EM pathway responsible for energy production, molecular relationships in most pathways are flexible, in response to the external conditions. This flexibility means that the model cannot be defined by a static structure. The other point is a technical issue: the model including unmeasured variables is inevitable in some cases. Since the molecular relationships are highly flexible, most experiments are designed to be as less perturbation to the cell as possible. This naturally means that all variables in the model are rarely measured, and the model behavior should be investigated by the measured data of a limited set of variables. Unfortunately, standard numerical parameter optimization frequently assumes that the data of all variables are obtained, and subsequently produces unstable estimation of parameter values.

In this paper, we would like to describe our tiny trials to overwhelm the above two issues: one is the network screening for catching the active networks (models) in particular conditions [2-4], and the other is the symbolic-numeric optimization procedure for improving the accuracy of parameter estimation [5,6].

2 Network Screening

2.1 Concept

The standard approach of network structure analysis in molecular biology is to infer the network structure from the measured data by using mathematical models, such as Boolean network and graphical model. This approach has possibility of finding new relationships between transcriptional factors and their regulated genes, guaranteed by mathematical soundness, such as (partial) dependence between two variables. As mentioned earlier, however, a static network cannot correspond well to the dynamical changes of molecular network structures. In addition, verification of inferred network structure for further analysis such as simulation needs huge experiments, due to new relationships, which cannot be found in the previous knowledge, in the inferred structure.

Our approach is a reverse approach to the standard approach. First, we gather a set of known network structures that are guaranteed by previous experiments. Then, we rationally select some networks among them, by using the measured data under the particular conditions. In the selection, we utilize a statistical procedure for estimating the consistency of graph structure with measured data [7]. Since the simple application is frequently not suitable for selecting the networks from the actual data, we have slightly modified its procedure for the practical applications. Although this approach cannot find any new relationships between molecules, all of the relationships in the selected networks are guaranteed by previous experiments.

2.2 Procedure

Network screening was performed as described previously [2-4]. This analysis is based on the procedure for estimating the consistency of a network structure (directed acyclic graph) with the measured data for the constituent variables in the graph [7]. The joint density function for a given network (reference network) was recursively factorized into conditional density functions, according to the parent-child relationship in the graph. The conditional functions were quantified into log-likelihoods, using linear regression for the measured data, with the assumption that the data followed a normal distribution (Gaussian network). The probability of the log-likelihood for the network structure (graph consistency probability; GCP) was then estimated from the distribution of log-likelihoods for 2,000 networks, generated under the condition that the networks shared the same numbers of nodes and edges as those of the given network.

The GCP was estimated for the ensemble of reference networks, to extract the candidate activated networks in the particular conditions, in a process termed 'network screening'. The reference networks were composed of the sub-networks that were constructed using the previous knowledge on the relationship between the transcriptional factors and their regulated genes, which is frequently compiled in databases such as TRANSFAC [8] or the combination of the data measured by the ChIP-on-Chip experiments and the following classification of the data based on the functional gene sets such as MSigDB [9].

Network-based analysis based on high throughput data is a challenging issue, which is expected to help us understand complex disease, and further elucidate the essential mechanisms of living organisms which would escape conventional single gene-based analysis. Instead of picking up differently expressed genes from high-throughput data, we use known functional networks to screen datasets and evaluate significantly activated networks. Then the network shows a whole picture of activated TF regulated functional gene sets under certain conditions, which cannot be achieved by single gene based method, and is much easier to bring the biological insights to us.

3 Symbolic-Numeric Optimization Procedure

3.1 Concept

In the studies of model dynamics, in general, a model to describe the relationship between constituent variables is first constructed with reference to the empirical knowledge, and then the model is mathematically expressed by differential equations, on the basis of the variable relations in the elementary process, such as molecule reactions. Finally, the parameters in the model are estimated by various parameter optimization techniques, from the time-series data measured for the constituent variables. While the computational time for parameter estimation has been greatly reduced, by the improvements in optimizing algorithms and the advent of high performance computers, the accurate numerical estimation of parameter values for a given model remains a limiting step. Indeed, the range of parameter values estimated by various optimization techniques is frequently broad, due to the conditions for parameter estimation, such as the initial values. In particular, we cannot always obtain the data measured for all of the constituent molecules in systems biology, due to limitations of experimental conditions measurement techniques. In this case, one of the issues we should resolve is the fact that the parameters are estimated from the data for only some of the constituent molecules. Unfortunately, it is more difficult to estimate the parameters in such a network model including unmeasured molecules.

Recently, we proposed a novel method for optimizing the parameters [5,6], by using differential elimination. Differential elimination was used in previously in the context of a system identification based on physical laws [10,11]. In our method, we use part of a technique from a previous study [12], in which differential elimination is introduced into the parameter optimization in a model including unmeasured variables. Instead of using differential elimination for estimating the initial values for the following parameter optimization, as done in the previous study [12], the equations derived by differential elimination are directly introduced as the constraints into the objective function for the parameter optimization.

3.2 Procedure

The key point is the introduction of new constraints obtained by differential elimination into the objective function, to improve the parameter accuracy. Following an explanation of differential elimination, the method of introducing the constraints [5,6] is briefly described.

Differential algebra aims at studying differential equations from a purely algebraic point of view [13,14]. Differential elimination theory is a sub theory of differential algebra [15], based on Rosenfeld-Gröbner [16]. The differential elimination rewrites the inputted system of differential equations to another equivalent system according to ranking (order of terms).

We assume a model, which is described by the system of parametric ordinary differential equations. Then, the differential elimination produces the equations equivalent to the original system. The values of rewritten system can be calculated, if we have time-series data of one of variables, and they would be zero, if all parameters were exactly estimated. Thus, rewritten system can be regarded as a kind of error function that expresses the difference between the measured and estimated data.

In general, the typical objective function for evaluating the reproducibility of an experimentally measured time-series for a parameter set is the total relative error. The parameter set is then estimated when the total relative error falls below a given threshold. However, the immense searching space of parameter values frequently hinders correct parameter estimation. Furthermore, all of the time series data for a parameter set are not always measured, especially in systems biology. To overcome this problem, we introduce the constraint between the estimate obtained by differential elimination (DE constraints), into the objective function, and a linear combination of the typical objective function of the total relative error and the DE constraints is defined as a new objective function.

One of the features of the DE constraint is that it includes the derivatives of the original system for the model. Since the derivatives generally contain the curve form information of the measured time-series data, such as slope, extremal point and inflection point, the new objective function estimates the difference of not only the values but also the comprehensive forms between the measured and estimated data, while the standard objective function estimates only the value difference. Note that the DE constraint is rationally reduced from the original system of differential equations for a given model, in a mathematical sense. Thus, our approach is expected to become a general approach in parameter optimization for improving the parameter accuracy.

4 Concluding Remarks

We introduced two methods for network structure and dynamics analyses, which we considered specific situations in modeling and analysis of systems biology. One method was designed to consider the model flexibility depending on external conditions by a statistical approach, and the other was designed to consider the model including unmeasured variables for less external perturbation to the organisms by a symbolic computation approach. Two methods still have some pitfalls, but partially overwhelm some difficulties of the analyses by previous standard methods. Further improvements of the two methods may be useful for uncovering new molecular mechanisms underlying complex biological phenomena.

Acknowledgements

We would like to express our gratitude to the collaborators, Drs. Saito, S., Wang, Y., Zhou, H., Chen, L., Nakatsui, M., Ürgüplü, A., Sedoglavic, A., Lemaire, F., Boulier, F., Tanimura, N., and Okamoto, M., for their technical supports, critical comments, and discussion, during the courses of the two works. The works were partially supported by “Development of Analysis Technology for Gene Functions with Cell Arrays” and “Development of Analysis Technology for Induced Pluripotent Stem (iPS) Cell”, from New Energy and Industrial Technology Development Organization (NEDO), Japan, and by Grant-in-Aids for Scientific Research on Priority Areas "Systems Genomics" (grant 20016028) and for Scientific Research (A) (grant 19201039), from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Birgit Schoeberl, Claudia Eichler-Jonsson, Ernst Dieter Gilles and Gertraud Müller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nature Biotechnology, 2002, 20: 370-375.

[2] Shigeru Saito, Sachiyo Aburatani and Katsuhisa Horimoto. Network evaluation from the consistency of the graph structure with the measured data. BMC Sys. Biol., 2008, 2, 84.

[3] Huarong Zhou, Shigeru Saito, Guanying Piao, Zhi-Ping Liu, Jiguang Wang, Katsuhisa Horimoto and Luonan Chen. Network Screening of Goto-Kakizaki Rat Liver Microarray Data during Diabetic Progression. BMC Sys. Biol.,2011, 5(Suppl 1), S16.

[4] Shigeru Saito, Yasuko Onuma, Yuzuru Ito, Hiroaki Tateno, Masashi Toyoda, Akutsu Hidenori, Koichiro Nishino, Emi Chikazawa, Yoshihiro Fukawatase, Yoshitaka Miyagawa, Hajime Okita, Nobutaka Kiyokawa, Yohichi Shimma, Akihiro Umezawa, Jun Hirabayashi, Katsuhisa Horimoto and Makoto Asashima. Possible linkages between the inner and outer cellular states of human induced pluripotent stem cells. BMC Sys.

Biol., 2011, 5(Suppl 1), S17.

[5] Masahiko Nakatsui, Katsuhisa Horimoto, Masahiro Okamoto, Yasuhito Tokumoto and Jun Miyake. Parameter Optimization by Using Differential Elimination: a General Approach for Introducing Constraints into Objective Function. BMC Sys. Biol., 2010, 4(Suppl 2), 59.

[6] Masahiko Nakatsui, Katsuhisa Horimoto, François Lemaire, Asli Ürgüplü, Alexandre Sedoglavic and François Boulier. Brute force meets Bruno force in parameter optimization: Introduction of novel constraints for parameter accuracy improvement by symbolic computation. IET Systems Biology, in press.

[7] Judea Pearl. Causality. Cambridge University Press. NY, 2000.

[8] Wingender, E. TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief. Bioinformatics, 2008, 9: 326-332.

[9] Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA, 2005, 102: 15545-15550.

[10] Lilianne, D.-V., Ghislaine, J.-B. and Celine, N. System identifiability (symbolic computation) and parameter estimation (numerical computation). Numerical Algorithms. 2003, 34: 282-292.

[11] Ljung, L. and Glad, T. On global identifiability for arbitrary model parametrizations. Automatica. 1994, 30: 265-276.

[12] Boulier, F. Differential algebra and system modelling in cellular biology. Proceedings of Algebraic Biology. 2008, 22-39.

[13] Ritt, J. F. Differential Algebra. Dover Publications Inc. NY, 1950.

[14] Kolchin, E. E. Differential Algebra and Algebraic Groups. Academic Press. NY, 1973.

[15] Boulier, F. Differential Elimination and Biological Modelling. Johann Radon Institute for Computational and Applied Mathematics (RICAM) Book Series Vol. 2. 2007, 111-139.

[16] Boulier, F., Lazard, D., Ollivier, F. and Petitot, M. Representation for the radical of a finitely generated differential ideal. Proceedings of ISSAC 1995, 1995, 158-166.