Efficient Inferring Method for Extracting Reliable Interactions from Time Series of Gene Expression Profile

Masahiko Nakatsui
nakatsui@brs.kyushu-u.ac.jp

Takanori Ueda
takanori.ueda@cytopathfinder.com

Maki Yukihiro
maki@fukuoka-int-u.ac.jp

Isao Ono
isao@dis.titech.ac.jp

Masahiro Okamoto
okahon@brs.kyushu-u.ac.jp

1 Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, Fukuoka 812-8581 Japan
2 Research and Development Division, CytoPathfinder Inc, AIST Tsukuba Central 6, Tsukuba, Ibaraki 305-8566, Japan
3 Department of Digital Media, Fukuoka International University, Fukuoka 818-0193, Japan
4 Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502, Japan

Keywords: inverse problem, S-system formalism, real-coded genetic algorithm, gene expression network

1 Introduction

Recent advances of powerful new technology such as DNA microarrays provide a mass of gene expression data on a genomic scale. One of the most important projects in post-genome-era is the system identification of gene expression network by using these observed data. We previously proposed an efficient numerical optimization technique [6] by using time-course data of system components. This is based on real-coded genetic algorithm (RCGAs [4,5]) to estimate the interrelated coefficients among system components of a dynamic network model called S-system that is a type of power-law formalism and is suitable for description of organizationally complex systems such as gene expression networks and metabolic pathways. Furthermore, in order to find the skeletal structure (small-size system) of the S-system formalism that can explain the experimentally observed responses, some of the interrelated coefficients, absolute values of which are less than a given threshold value are to be removed during optimization procedures [6]. The system identification by S-system is not suitable for large scale genetic network unless developing efficient numerical optimization method because the number of estimated parameters increases with $O(n^2)$. In order to overcome this problem, we introduced the Step-by-step strategy [1]. This strategy can be summarized as follows; 1) Focused on one temporal profile of gene expression ($i$), and it is supposed that other temporal profiles are treated as known and fixed data. 2) Estimate the interrelated parameter values for gene ($i$). 3) Repeat above procedures ($n-1$) times. It is inverse problem to infer the internal structure of gene expression from experimentally observed time-course data. So there are many network candidates of gene expression which can realize the same experimentally observed facts, however, the structures of these network candidates are different each other. Therefore, we should propose the analytical method for extracting useful information from many network candidates. We previously proposed the analyzing procedure for extracting common core binomial genetic interactions from many kinds of network candidates [2], and confirmed that the parameter sensitivity for common core binomial genetic interactions is significantly greater than that for other unique interactions [3]. The interactions with high sensitivity much contribute to realize experimentally obtained time-course data of gene expression network. In this study, we shall describe about the combination of Step-by-step strategy with analyzing method for extracting common core binomial genetic interactions. The inferring accuracy and efficiency of our proposed method were examined in the case for applying to large-scale network.

2 Method and Materials
In S-system model, the sign of interrelated coefficient shows the kinds of interactions such as activation, inhibition, or no relation. The common core binomial genetic interactions are defined by the corresponding binomial interactions with sign of which are same among all inferred network candidates.

We prepared artificial model network of gene expression containing 30 genes shown in Fig. 1. Subsequently we calculated time-course data sets which were given as experimentally observed facts. We prepared 31 kinds of time-course data sets for wild type and that for one gene disrupted strain. First, we inferred 30 network candidates by using Step-by-step strategy. Subsequently we extracted common core binomial genetic interactions for each step from 30 kinds of inferred network candidates. In the next step, combining and summing up all common core binomial genetic interactions for each step, we obtained the collection of interactions containing whole genes in model network. We tried to infer network candidates under 6 kinds of error allowance on RCGAs, 3%, 5%, 7%, 10%, 20%, and 30%.

3 Result and Discussion

By using 30 genes network model, we applied our proposed analyzing procedure for extracting common core binomial genetic interactions from all inferred network candidates. First we inferred network candidates and extracted common core binomial genetic interactions based on all 31 time-series. In the case of error allowance on RCGAs is 3% and 5%, we could extract all interactions in model network as common core binomial genetic interactions without including false positive interactions. The inferring efficiency decreases with error allowance on RCGAs, however, the inferring accuracy is extremely high (there are no false positive interactions) in all parameter optimizing conditions. Subsequently, in order to study the accuracy and efficiency of our proposed method with using a little number of experimentally observed facts, we tried to extract common core binomial genetic interactions with changing the number of given time-series for 3 to 25. We could extract correct interactions almost perfectly when we used more than 15 sets of time-series that is half number of genes in model network, however, the inferring accuracy in the case with a little number of experimentally observed facts sometimes becomes low. We should revise the efficient and reliable analyzing algorithms and the experimental design for inferring gene expression networks.

References