Efficient inferring method of genetic interactions based on time-series of gene expression profile

Masahiko Nakatsui
Laboratory for Bioinformatics,
Graduate School of Systems Life Sciences
Kyushu University
Fukuoka, 812-8581, Japan
nakatsui@brs.kyushu-u.ac.jp

Takanori Ueda
Research and Development Division
CytoPathfinder Inc.
Kiba KI Bldg. 4F 6-4-2, Koto-ku,
Tokyo, 135-0042, Japan
takanori.ueda@cytopathfinder.com

Yukihiro Maki
Department of Digital Media
Fukuoka International University
Dazaifu, Fukuoka, 818-0193, Japan
maki@fukuoka-int-u.ac.jp

Isao Ono
Department of Computational Intelligence
and Systems Sciences,
Interdisciplinary Graduate School of
Science and Engineering,
Tokyo Institute of Technology,
Yokohama 226-8502, Japan
isao@dis.titech.ac.jp

Masahiro Okamoto
Laboratory for Bioinformatics,
Graduate School of Systems Life Sciences
Kyushu University
Fukuoka, 812-8581, Japan
okahon@brs.kyushu-u.ac.jp

Abstract

Recent advances in technologies such as DNA microarrays have provided a mass of gene expression data on the genomic scale. One of the most important projects in post-genome-era is the systemic identification of gene expression networks. However, inferring internal gene expression structure from experimentally observed time-series data is an inverse problem. We have therefore developed a system for inferring network candidates based on experimental observations. Moreover, we have proposed an analytical method for extracting common core binomial genetic interactions from among various network candidates. Common core binomial genetic interactions are reliable interactions and are important in understanding the dynamic behavior of gene expression network. Here, we discuss an efficient method for inferring genetic interactions that combines a Step-by-step strategy [1] with an analytical method for extracting common core binomial genetic interactions.

keywords: inverse problem, S-system formalism, gene expression network, system identification.

1 Introduction

The expression profiles of hundreds of thousands of genes can be measured simultaneously on a genomic scale using recent technologies such as DNA microarrays and DNA chips. These data depend on environmental conditions and are typically obtained as snapshots, but can be generated as dense time-series that indicate dynamic behavior. Experimentally observed time-course data contain enormous amounts of information regarding the regulation of genetic networks *in vivo*. However, as this information is entirely implicit, it requires adequate analytical and computational methods for retrieval and interpretation. This aspect of genetic networks based on the experimentally observed time-course data is generally referred to as an “inverse problem” and can be defined as function optimization of parameters involved in a suitable model-representation of the genetic network. In other words, system parameter values must be estimated in a model that can realize the given experimentally observed time-course data.

The key points in solving such inverse problems are setting up canonical representations of mathematical modeling for genetic networks and exploring and exploiting parameter values within vast search space. We initially proposed a novel inferring method for genetic networks by combining a dynamic network model called S-system [2] with a computational technique for parameter estimation based on simple genetic algorithms [3][4]. S-system is suitable for conceptual modeling and describing organizationally complex systems involving looping or cyclic interactions between system components, such as metabolic pathways and gene expression networks. The value of interrelated coefficients in the above formalism is directly or indirectly related to the regulation mechanism in the modeled network, and the inferred network structure resulting from the estimation of parameters is one of the better candidates for genetic network structure. Genetic networks described by S-system formalism are suitable for systematic analysis because the dynamic behavior of the network can be obtained by numerical simulation. S-system formalism, however, has a major disadvantage in that this formalism includes a large number of parameters that must be estimated; the number of estimated parameters is $2n(n+1)$ (where n is the number of system components).

Simple genetic algorithm (SGA) is a well-known heuristic optimizer of such large numbers of parameters. However, SGA has two intrinsic problems; early convergence in the first stage of the search, and evolutionary stagnation in the last stage of the search. Real-coded genetic algorithms (RCGAs) have recently attracted attention as alternative numerical optimizing methods to SGA. One of the crossover operators for RCGAs, known as *unimodal normal distribution crossover* (UNDX), has shown good performance in optimizing various functions, including multi-modal functions, and benchmark functions with epistasis among the parameters [5]. Furthermore, Sato *et al*. proposed a new generation-alternation model, known as *minimal generation gap* (MGG), to avoid early convergence in the first stage and to suppress evolutionary stagnation in the last stage [6].

Using S-system modeling and RCGAs, with a combination of UNDX and MGG, we proposed efficient procedures for inferring genetic interactions based on experimentally observed time-course data sets of system components (mRNA) [7][8][9]. We were able to obtain numerous network candidates for gene expression based on experimental observations; however, the structure of these candidate networks differs from one another. Therefore, we proposed an efficient analytical method for extracting useful and reliable information from various network candidates. Here, we describe an analysis method for extracting common core binomial genetic interactions, and the combination of this method with the efficient network inferring engine called Step-by-step strategy [1][10].

2 Method for System Identification

2. 1 S-system formalism

S-system is suitable for dealing with gene expression network structures. It can sufficiently represent the structure of organizationally complex systems to capture the essence of experimentally observed response:

$$\frac{dX_i}{dt}=\alpha_i\prod_{j=1}^{n}X_j^{g_{ij}}-\beta_i\prod_{j=1}^{n}X_j^{h_{ij}}\quad(i=1,2,\ldots,n)\qquad(1)$$

where, n is the number of system components (genes) in the network, X_i is the gene expression quantity, α_i and β_i are apparent rate constants, and g_{ij} and h_{ij} are interrelated coefficients between X_is. The first term on the right hand side of Eq. (1) corresponds to the synthetic process of X_i, and the second term expresses the degradation process of X_i. The value of g_{ij} and h_{ij} determine the structure of interactions between X_i and X_j; $g_{ij}(h_{ij})$ express the interactive effects of X_j to the synthetic process (degradation process) of X_i. If $g_{ij}(h_{ij})$ is positive, gene X_j induces a synthetic process (degradation process) in gene X_i. On the other hand, if $g_{ij}(h_{ij})$ is negative, gene X_j suppresses the synthetic process (degradation process) of gene X_i. If the value of $g_{ij}(h_{ij})$ is zero, there are no effects of gene X_j on the synthetic process (degradation process) of gene X_i. The gene expression network can be inferred by estimating α_i, β_i, g_{ij} and h_{ij} in the S-system formula. The S-system formalism has, however, a major disadvantage in that this formula includes a large number of estimated parameters, $2n(n+1)$. To overcome this disadvantage, we used RCGAs to estimate these parameters.

2. 2 Real-coded Genetic Algorithms

Because the S-system is a formalism of ordinary nonlinear differential equations, the system can easily be solved numerically by using numerical calculation programs customized specifically for this structure [11]. However, when an adequate time-course of relevant state variables is given, the set of parameter values α_i, β_i, g_{ij}, and h_{ij}, in many cases, will not be uniquely determined, as it is highly possible that other sets of parameter values will also show a similar time-course. Therefore, even if one set of parameter values that explain the observed time-course is obtained, this set is still one of the best candidates to explain the observed time-courses. Our strategy is to explore and exploit these candidates within the immense searching space of parameter values. In this problem, each set of parameter values to be estimated is evaluated using the following procedure: Suppose that $X_{d,i,t}^{\mathrm{cal}}$ is the numerically calculated time-course ant time t of state variable X_i in the d-th. data set and that $X_{d,i,t}^{\exp}$ represents the experimentally observed time-course at time t of X_i in the d-th. data set. The sum of the square values of the relative error between $X_{d,i,t}^{\mathrm{cal}}$ and $X_{d,i,t}^{\exp}$ gives the total relative error E;

$$E=\sum_{d=1}^{D}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{X_{d,i,t}^{\exp}-X_{d,i,t}^{\mathrm{cal}}}{X_{d,i,t}^{\exp}}\right)^2$$

where, D is the total number of data sets experimentally observed under different experimental conditions, such as gene disruption, N is the total number of experimentally observed state variables, and T is the total number of sampling points over time in one experimental condition.

The computational task is to determine a set of parameter values that minimizes the objective function E. We have developed an efficient computational technique based on RCGAs as a nonlinear numerical optimization method that is much less likely to be stranded in local minima. This technique is based on the combination of UNDX [5], with the alternation of the MGG model [6]. To find the skeletal structure (small-size system) of the S-system formalism that explain the experimentally observed response, some parameters (g_{ij} and h_{ij}), the absolute values of which are less than a given threshold value, are removed (reset to zero) during optimization procedure [9]. We also restricted the number of interactions to each synthesis or degradation process of each gene. We previously developed a genetic inferring system implementing RCGAs in a grid computing network [12].

2. 3 Step-by-step Strategy

System identification using S-system formalism is not suitable for large-scale genetic networks without an efficient numerical optimization method, as the number of estimated parameters increases with $O(n^2)$. To overcome this problem, we introduced the Step-by-step strategy. This strategy can be summarized as follows;

1. Focuses on one temporal profile of gene expression (i), and other temporal profiles are treated as known and fixed data.
2. Estimates the interrelated parameter values for gene (i).
3. Repeats the above procedure (n-1) times.

2. 4 Extracting Common Core Binomial Genetic Interactions

In the S-system model, interrelated coefficients show interactions such as activation, inhibition or no relationship. The common core binomial genetic interactions are defined by the corresponding binomial interactions, the sign of which are the same among all network candidates of inferred under the same parameter optimizing conditions. Contradicted interactions among inferred network candidates are extracted as having no relationship in common core binomial genetic interactions. The common core binomial genetic interactions represent relationships between two genes; however these interactions do not have information on strength. Therefore, in the collection of common core binomial genetic interactions obtained by the proposed method, we are not able to obtain dynamic behavior by numerical simulation. We previously confirmed that the sensitivities for common core binomial genetic interactions are significantly larger than those for other unique interactions [13]. As interactions having high sensitivity contributes to accurately identifying gene expression networks among experimental time-course data, these interactions appear to be rigid and essential to organizationally complex systems.

2. 5 Evaluation of Extracted Common Core Binomial Genetic Interactions

We defined the correctness ratio (CR) and reproduction ratio (RR) for evaluating inferred gene expression network candidates and extracted common core binomial genetic interactions. The CR is defined as follows:

$$\mathrm{CR}=\frac{\mathrm{TP_{all}}}{\mathrm{TP_{all}}+\mathrm{FP_{all}}}$$

$$\mathrm{TP_{all}}=\sum_{i=1}^{n}\mathrm{TP}_i$$

$$\mathrm{FP_{all}}=\sum_{i=1}^{n}\mathrm{FP}_i$$

where, TP_i is the number of true-positive interactions in i-th network candidate, FP_i is the number of false-positive interactions in i-th network candidate, and n is the number of inferred network candidates. The value of CR shows the inferring accuracy of gene expression network candidates or extracted common core binomial genetic interactions under investigation. We also defined RR, which indicates the inferring efficiency of network candidates or common core binomial genetic interactions as follows;

$$\mathrm{RR}=\frac{\mathrm{TP_{all}}}{\mathrm{TP_{all}}+\mathrm{FN_{all}}}$$

$$\mathrm{TP_{all}}=\sum_{i=1}^{n}\mathrm{TP}_i$$

$$\mathrm{FN_{all}}=\sum_{i=1}^{n}\mathrm{FN}_i$$

where, FN_i is the number of false-negative interactions in i-th network candidate. Both CR and RR values are between 0.0 and 1.0, and the best values of CR and RR are 1.0.

3 Case Study

We prepared an artificial gene expression network model containing 30 genes, as shown in Fig. 1. Subsequently, we calculated time-course data sets, which were considered to be experimental observations. We prepared 31 time-course data sets for wild-type and one gene-disrupted strains under the following conditions; the number of sampling points in each time-course data set is 70, the initial value for all genes is 0.25. We applied the Step-by-step strategy to infer genetic interactions based on the 31 time-course data sets. The optimizing conditions were follows:

- ✓ Crossover Operator: UNDX
- ✓ Generation-alternation model: MGG
- ✓ Error allowance on RCGAs: 3%, 5%, 7%, 10%, 20%, and 30%
- ✓ Threshold for obtaining skeletal structure of gene expression network: 0.05
- ✓ Restricted number of interactions to synthesis or degradation process of each gene: 3

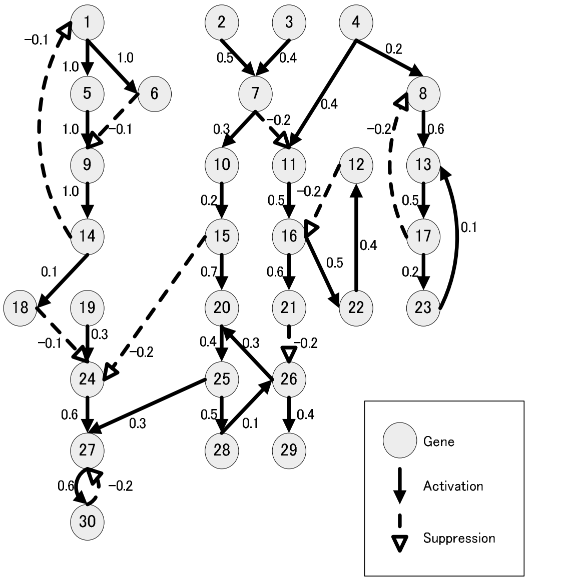
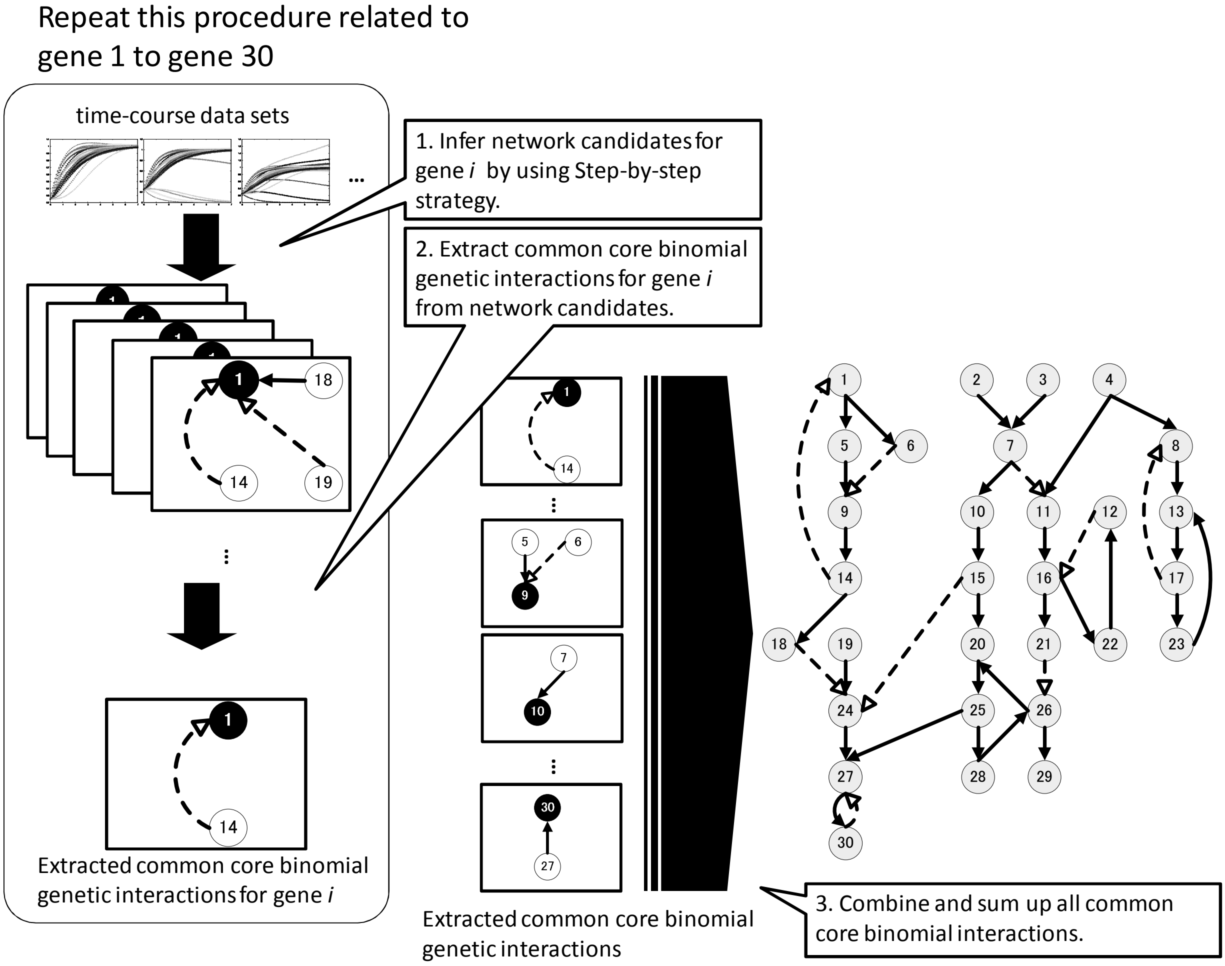


Fig. 1 Network model containing 30 genes. We set h_{ii} which represents the interrelated coefficient for self-degradation, at 1.0 and set h_{ij} at 0. The values accompanying arrows show the value of g_{ij}. The number of estimated interactions is 38.

We initially inferred 30 network candidates (30 trials) in each step of the Step-by-step strategy. As shown in procedure 1 and 2 in Fig. 2, we extracted common core binomial genetic interactions for each step from the 30 inferred network candidates. In the next step, combining and determining the totals for all common core binomial genetic interactions in each step, we obtained a collection of interactions that included all genes in the model network (procedure 3 in


Repeat this procedure related to gene 1 to gene 30

1. Infer network candidates for gene *i* by using Step-by-step strategy.

2. Extract common core binomial genetic interactions for gene *i* from network candidates.

Extracted common core binomial genetic interactions for gene *i*

Extracted common core binomial genetic interactions

3. Combine and sum up all common core binomial interactions.

Fig. 2 Combination of Step-by-step strategy and analysis procedure for extracting common core binomial genetic interactions.

Table 1 Results for 30-gene network inference.

Error allowance on RCGAs	3%	5%	7%	10%	20%	30%
Number of inferred interactions	38	53	53	56	54	48
Number of true-positive interactions	38	38	35	34	25	20
Number of false-positive interactions	0	0(15)*	0(18)*	0(22)*	0(29)*	0(28)*
Number of false-negative interactions	0	0	3	4	13	18
Correctness Ratio	1.0	1.0	1.0	1.0	1.0	1.0
Reproduction Ratio	1.0	1.0	0.92	0.89	0.66	0.53

*The number in the parentheses shows the number of interactions which represents the suppression of self-synthesis (negative value of g_{ii})

Fig. 2). As the number of trials for each step in the Step-by-step strategy was 30, the total inferring frequency in 30 steps is 900.

4 Results

The results of inference and extraction of common core binomial genetic interactions are shown in Table 1. With 3% and 5% error allowance on RCGAs, as the numbers of true-positive interactions are the same as that of inferred interactions, we were able to extract all interactions in the model network shown in Fig. 1 as common core binomial genetic interactions without including false-positive interactions. Under these optimizing conditions, both the CR and RR are 1.0. When the error allowance on RCGAs was 5% or more, we obtained some false-positive interactions, which are shown in parenthesis in Table 1. All of these false-positive interactions represent suppression of the own synthesis process ($g_{ii} < 0$). These interactions play the same role as activation of self-degradation ($h_{ii} > 0$), which was set to all genes in the model network (see Fig. 1 : $h_{ii} = 1$, $h_{ij} = 0$ ($i \neq 0$)). Therefore, we ignored these interactions in subsequent analysis. The RR, which indicates inferring efficiency, decreased with error allowance on RCGAs; however, the CR, which indicates inferring accuracy, was extremely high (there were no false-positive interactions) under all parameter optimizing conditions.

4.1 Changing the number of given time-series

Subsequently, in order to study the accuracy and efficiency of our proposed network inferring method with experimentally observed time-course data sets, we attempted to infer gene expression network candidates by changing the number of time-series from 3 to 25. We then extracted common core binomial genetic interactions from the network candidates inferred from the same time-course data sets. We applied the Step-by-step strategy to infer network candidates with a less than 10% error allowance on RCGAs. The number of trials for each step in Step-by-step strategy was 30. We made 5 attempts to infer network candidates while changing the combination of randomly selected time-series in the one gene disrupted strain. The RR and CR of the extracted common core binomial genetic interactions for each step of the Step-by-step strategy are shown in Fig. 4. In Fig. 4, the RR increased with the number of given time-course data sets. We were able to extract correct interactions almost perfectly as common core binomial genetic interactions when we used more than 15 time-series, which is half the number of genes in the target network.

5 Conclusion

We proposed a reliable analyzing procedure for extracting common core binomial genetic interactions from all inferred network candidates of gene expression. Using an artificial network

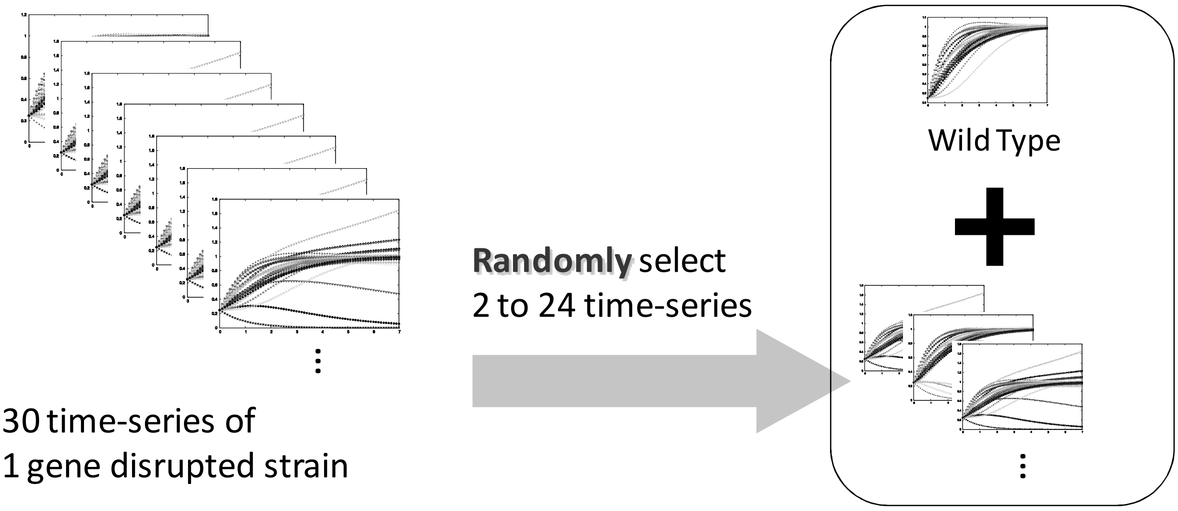


Fig. 3 Changing the number of time-series.

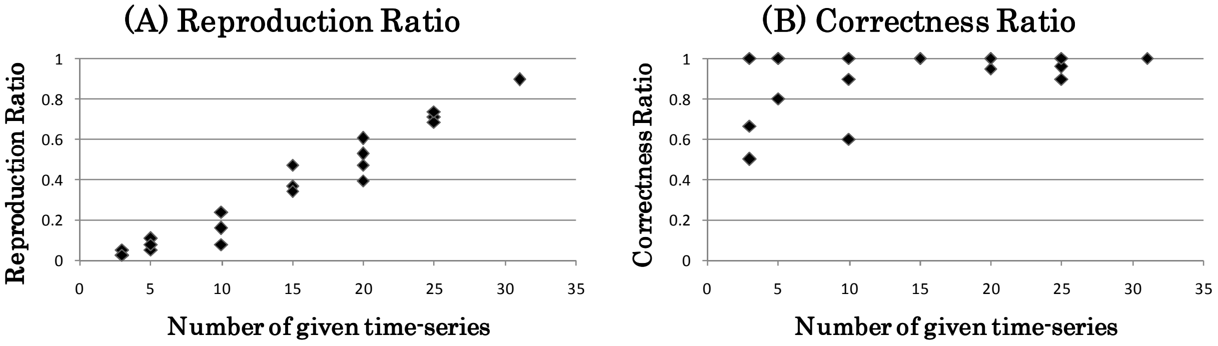


Fig. 4 Reproduction ratio and correctness ratio.

model including 30 genes, we examined the accuracy and efficiency of our proposed method, which combines the network inferred engine known as Step-by-step Strategy with analysis by extracting common core binomial genetic interactions. Efficiency decreases with error allowance on RCGAs; however, the accuracy is extremely high under all parameter optimizing conditions. To examine the inferring accuracy and efficiency of our proposed method with small numbers of experimental data, we attempted to extract common core binomial genetic interactions with a changing number of time-series. We were able to extract correct interactions almost perfectly when we used more than 15 time-series, including 14 time-series from the one gene disrupted strain; however, the inferring accuracy in cases with small numbers of experimental observations can sometimes be very low.

We obtained reliable, rigid and essential interactions by extracting common core binomial genetic interactions; however, we cannot simulate or systematically analyze the dynamic behavior of genetic networks because common core binomial genetic interactions contain only information on interactions between genes. In the future, we plan to discuss method to reproduce networks that can be simulated numerically by using common core binomial genetic interactions. Furthermore, we will revise our efficient and reliable analysis algorithms, as well as the experimental design.

References

[1] Maki, Y., Takahashi, Y., Arikawa, Y., Watanabe, S., Aoshima, K., Eguchi, Y., Ueda, T., Aburatani, S., Kuhara, S., Okamoto, M., "AN INTEGRATED COMPREHENSIVE WORKBENCH FOR INFERRING GENETIC NETWORKS: *VOYAGENE*," *Journal of Bioinformatics and Computational Biology*, Vol. 2 No. 3 533-550. 2004.

[2] Savageau, M., A., Biochemical Systems Analysis: A study of function and design in molecular biology, Addison-Wesley, Reading. 1976.

[3] Maki, Y., Tominaga, D., Okamoto, M., Wataneb, S., and Eguchi, Y., "Development of a system for the inference of large scale genetic networks," *Pacific Symposium on Biocomputing 2001* (*PSB2001*), 446-458. 2001.

[4] Tominaga, D., Koga, N., Okamoto, M., "Efficient Numerical Optimization Algorithm Based on Genetic Algorithm for Inverse Problem," *Proc. Genetic and Evolutionary Computation Conference*, 251-258. 2000.

[5] Ono, I., and Kobayashi, S., "A real-coded genetic algorithm for function optimization using unimodal distribution crossover," *Proc 7th ICGA*, 249-253. 1997.

[6] Sato, H., Ono, I., and Kobayashi, S., "A new generation alternation model of genetic algorithm and its assessment," *J. of Japanese Society for Artificial Intelligence*, 15 (2) 743-744. 1997.

[7] Nakatsui, M., Ueda, T., Okamoto, M., "Integrated System for Inference of Gene Expression Network," *Genome Informatics*," 14 282-283. 2003.

[8] Ueda, T., Koga, N., and Okamoto, M., "Efficient numerical optimization technique based on real-coded genetic algorithm," *Genome Informatics* 12 451-453. 2001.

[9] Ueda, T., Ono, I., and Okamoto, M., "Development of system identification technique based on real-coded genetic algorithm," *Genome Informatics* 13 386-387. 2002.

[10] Maki, Y., Ueda, T., Okamoto, M., Uematsu, N., Inamura, K., Uchida, K., Takahashi, Y., Eguchi, Y., "Inference of Genetic Network Using the Expression Profile Time Course Data of Mouse P19 Cells," *Genome Informatics*, 13 382-383. 2002.

[11] Irvine, D, H., Savageau, M. A., "Efficient solution of nonlinear ordinary differential equation expressed in S-system canonical form," *SIAM Journal on Numerical Analysis*, 27(3) 704-735. 1990.

[12] Imade H., Mizuguchi, N., Ono, I., Ono, N., Okamoto, M., "Gridifying: An Evolutionary Algorithm for Inference of Genetic Networks Using the Improved GOGA Framework and its Performance Evaluation on OBI Grid," *Lecture Note in Bioinformatics*, 3370 171-186. 2005.

[13] Nakatsui, M., Ueda, T., Ono, I., Okamoto, M., "Control Aspect of Common Interactions Extracted from Inferred Network Candidates of Gene Expression," *Genome Informatics*, P008. 2004.