JOURNAL OF BIOSCIENCE AND BIOENGINEERING
Vol. 101, No. 5, 377–384. 2006
DOI: 10.1263/jbb.101.377

REVIEW

Application of Bioinformatics for DNA Microarray Data to Bioscience, Bioengineering and Medical Fields

Taizo Hanai,[1]* Hiroyuki Hamada,[1] and Masahiro Okamoto[1]

Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan[1]

Received 7 November 2005/Accepted 23 December 2005

In the 1990s, DNA microarray or DNA chip as a novel biological experimental technology was developed, which enables the comprehensive measurement of the expression levels of hundreds of genes, simultaneously. Using this technique, a comprehensive understanding of the cell can be achieved. However, because even simple life forms, such as microorganisms, have more than a thousand kinds of genes, the data from a DNA chip cannot be analyzed without statistical and informational technology. Bioinformatics is the interdisciplinary research field integrating molecular biology with informatics, and it is expected to have a huge impact on the bioscientific, bioengineering and medical fields. There are many techniques in bioinformatics for the analysis of DNA microarray data; however, these are mainly divided into fold-change analysis, clustering, classification, genetic network analysis, and simulation. In this review, these techniques are briefly explained by using some examples.

[**Key words:** bioinformatics, DNA microarray, clustering, pattern classification, genetic network]

Proteins have many important functions in a cell. If the comprehensive measurements of the concentrations of all kinds the proteins in a cell were performed, the condition of the cell could be predicted. However, at present, it is difficult to measure all kinds of protein concentrations, simultaneously. During the translation process of a protein from a gene, mRNA is first transcribed and the protein corresponding to this mRNA is subsequently synthesized. Therefore, if the expression levels for all mRNAs were measured, the condition of a cell could be known. For example, the comprehensive measurement in the expression levels for all kinds of mRNAs during cell proliferation, mRNAs corresponding to DNA and protein synthesis for cell division were mainly observed. In the 1990s, DNA microarray or DNA chip technology was developed, which enables the simultaneous measurement of the expression levels of over one hundred genes (1–3). Measuring the time course of the expression profiles for all kinds of mRNA using this technique, would allow temporal changes in the condition of the cell to be evaluated. When an external stimulus is introduced to a cell, we can observe which and how many mRNAs are predominantly transcribed as a trigger of control in the cell against this stimulus, and which and how many mRNAs are transcribed subsequently. Using such data, it is also possible to estimate the control system of transcription against this stimulation. However, because even simple life forms, such as microorganisms, have more than a thousand kinds of gene, it is impossible to analyze the data from a DNA chip without using statistical and informational technology. Bioinformatics or computational biology, is the interdisciplinary research field integrating biology with informatics, and is expected to a huge impact on the bioscientific, bioengineering and medical fields. There are many techniques in bioinformatics for DNA microarray data; however, these are mainly divided into fold-change analysis, clustering, classification, genetic network analysis, and simulation (4). As shown in Fig. 1, we have applied these bioinformatics techniques for DNA microarray data to the bioscience, bioengineering, and medical fields for 6 years. In this paper, we briefly explain these techniques and show the results of our research as examples.

GROUPING GENES BY CLUSTERING ANALYSIS OF DNA MICROARRAY DATA

One of the most basic and useful analyses for application to the bioscience, bioengineering and medical fields using DNA microarray data is clustering analysis, because the results may facilitate the understanding of such as the functions of uncharacterized genes. Therefore, various clustering methods, such as hierarchical clustering (5), *k*-means clustering (6), self-organized maps (SOMs) (7) and the other methods (8), have been examined and used to elucidate fundamental and/or characteristic expression patterns. The classification of cell lines, particularly human cancers

* Corresponding author. e-mail: taizo@brs.kyusyu-u.ac.jp
phone: +81-(0)92-642-2899 fax: +81-(0)92-642-3030

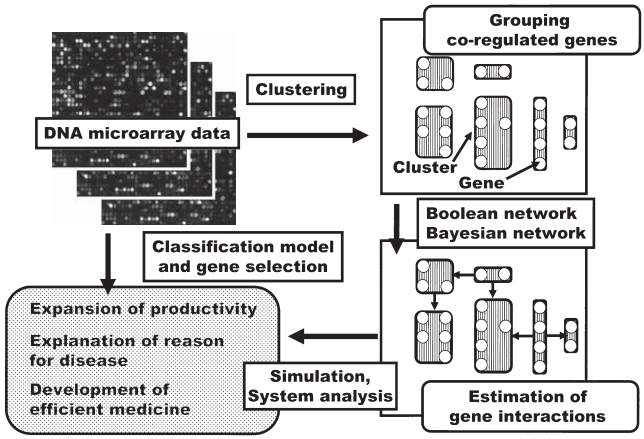

FIG. 1. Strategies for applied bioinformatics for DNA microarray data.

(9), as well as the analysis of temporally expressed genes of *Saccharomyces cerevisiae* (5) were examined using these clustering methods. However, one problem is that the microarray data include a large amount of noise due to experimental error and this significantly affects the results of clustering analysis. However, because fuzzy logic is considered to display robustness to noise, we have applied fuzzy *k*-means clustering (10), which combines fuzzy logic with *k*-means clustering, to the DNA microarray data.

The fuzzy *k*-means clustering (Fig. 2) is performed using the following equation:

$$J(K, m)=\sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(x_i,c_k) \quad (1)$$

where K and N are the number of clusters and genes in the data sets, respectively, m is a parameter which relates to the fuzziness of resulting clusters, u_{ki} is the degree of membership of gene x_i in cluster k, and $d^2(x_i, c_k)$ is the distance from gene x_i to centroid c_k. The parameters in this equation are cluster centroid vector c_k and membership vectors u_{ki}. The values of these unknown parameters can be estimated by the Lagrange method. The calculated u_{ki} shows the ratio of belonging to cluster k and centroid c_k shows the representative gene expression profile of a cluster k.

In this study, we used the expression data published by Chu *et al*. (11). The growth of *Saccharomyces cerevisiae* was synchronized by transferring cell to a sporulation medium (SPM) at $t=0$ to maximize the synchrony of sporulation. RNA was harvested at time $t=0$, 0.5, 2, 5, 7, 9 and 11.5 h after transfer to the SPM. Each gene's mRNA expression level just before transfer to the SPM was used as a control. The expression profiles of approximately 6100 genes are included in this data. Using these profiles, we followed the same method as that of Chu *et al*. (11) to extract the genes that showed a significant increase in mRNA levels during sporulation. We finally selected 45 genes, the functions of which have been biologically characterized by Kupiec *et al*. (12). In this study, the parameter m of the fuzzy *k*-means clustering equation was set to 1.55 and the number of clusters (K) was set to 6 on the basis of biological knowledge (11).

The result of the fuzzy *k*-means clustering is shown in Fig. 3. This figure shows the representative time course data for each cluster and these values correspond to those at the centroid in each cluster. Early I, Early II, Early-middle, Middle, Middle-late and Metabolic, which were characterized by Chu (11), were used as index genes. As a result, Early I, Early II, Early-middle, Middle, Middle-late and Metabolic genes were found to belong mainly to clusters 1, 2, 4, 5, 5, and 3, respectively. From this result, it was shown that fuzzy *k*-means clustering can be used to group genes with the same biological characterization.

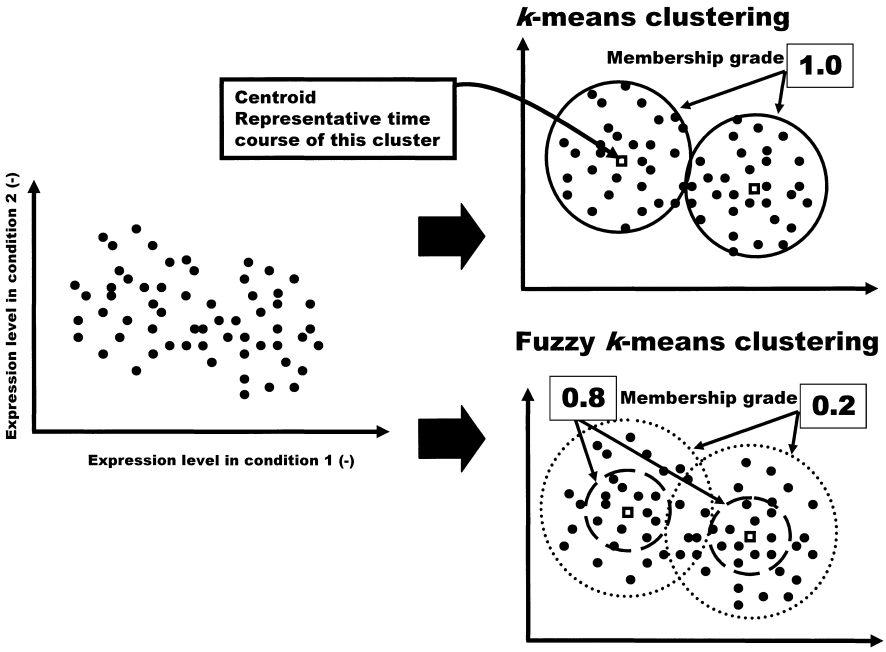

FIG. 2. *k*-means and fuzzy *k*-means clustering.

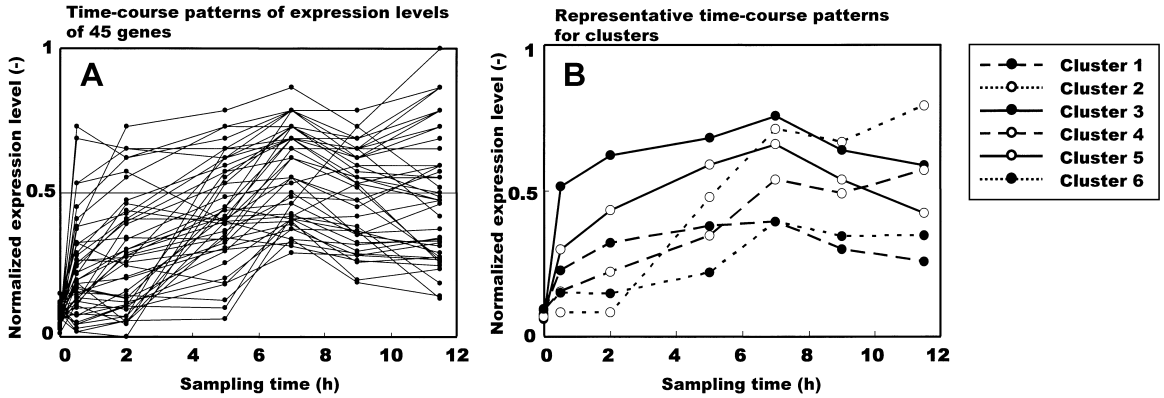

FIG. 3. Time-course patterns of expression levels of 45 genes (A) and representative time-course patterns for clusters (B).

Generally, microarray experiments include several types of noise, and therefore, quantitative microarray data vary from experiment to experiment. Because this noise might significantly affect the result, it is important to investigate the tolerance of the clustering analysis of microarray data to noise. For this purpose, we intentionally added noise to the experimental expression value, and tested, using the ordinal *k*-means clustering and fuzzy *k*-means clustering, whether the same clustering result could be obtained repeatedly (13). The noise was randomly determined according to the normal distribution, setting the maximum variance of the noise to be half (50%) and the same (100%) as the expression value. We prepared 10 data sets including noise at every maximum variance and noise tolerance was evaluated repeatedly. Table 1 shows the results of clustering robustness using 10 sets of noised data in two clustering methods. In the case of *k*-means clustering with 100% maximum noise level, 393 genes in total among 450 (45 genes × 10 sets) genes were clustered into the same clusters as those using unnoised data; 87.3% of genes had same clustering result even when random noise was added. We defined robustness ratio as the ratio of genes the clustering results of which were the same with or without the addition of random noise. In the case of fuzzy *k*-means clustering with 100% maximum noise level, according to the threshold of the membership grade, robustness ratios increased from 87.8% to 100.0%. It is obvious that fuzzy *k*-means clustering exhibits higher robustness than the ordinary *k*-means clustering.

From these results, it is considered that fuzzy *k*-means clustering is a more suitable clustering method than the ordinary *k*-means clustering and is a powerful tool for gene clustering from DNA microarray data.

TABLE 1. Comparison of robustness ratio for *k*-means and fuzzy *k*-means clustering for microarray data with artificial noise

Method	Threshold for membership grade (–)	Maximum noise level	
		50%	100%
k-means	–	0.942	0.873
fuzzy *k*-means	0.5	0.987	0.987
	0.6	0.995	0.993
	0.7	0.993	1.000
	0.8	1.000	1.000

PROGNOSTIC PREDICTION BY CLASSIFICATION ANALYSIS OF DNA MICROARRAY DATA

Despite recent progress in clinical studies and biological technology for cancer, prognostic predictions for patients remain difficult and inaccurate. If prognostic prediction computer software for cancer patients using DNA microarray data could be developed with high accuracy, it would considerably benefit the quality of life (QOL) of patients. Such software is one representative application of bioinformatics for DNA microarray data to the medical field. As shown in Fig. 4, many research groups have attempted to develop classifying software on the basis of artificial neural networks, such as the fuzzy neural network and multiple regression analysis (14–16). By using published microarray data, we have also predicted the survival of patients using a support vector machine (SVM) (17), which is one of the powerful supervised machines learning methods for classification problems (Fig. 5). Furthermore, the estimations of missing microarray measurement values and gene selection were investigated, and we have constructed models on the basis of a linear SVM.

Gene selection will lead to an improvement in classification accuracy by eliminating genes unnecessary for the classification model. We selected the important genes for the classification by the parameter increasing method (PIM) (15) using SVM after the selection of 100 genes with the higher rank by signal-to-noise ratio (18). The procedure for PIM can be summarized as follows: (i) set an empty subset

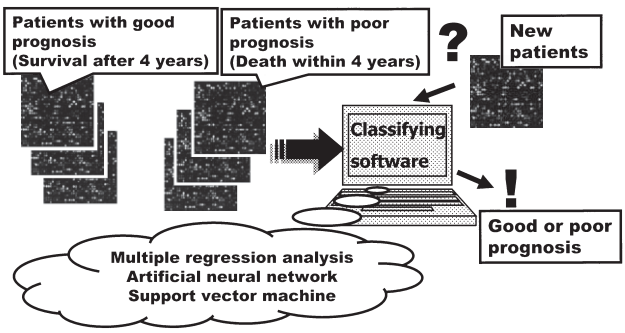

FIG. 4. Classification analysis for prognostic prediction from microarray data.

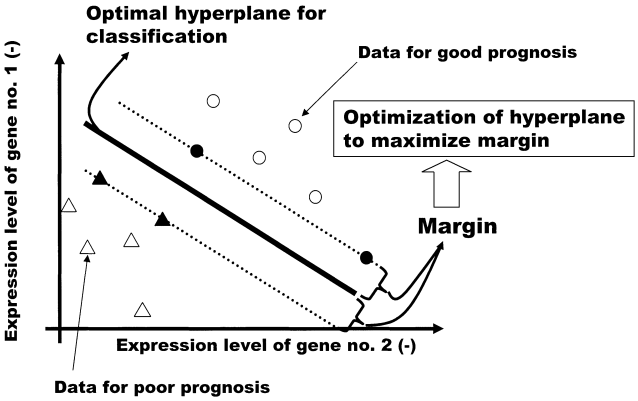

FIG. 5. Concept of classification for prognostic prediction using support vector machine.

of genes as the input variables, (ii) add one gene to the subset to minimize the sum-squared error of SVM, (iii) repeat procedure ii while the sum-squared error continues to decrease, (iv) determine the final member of genes (the optimal member of genes for the classification model) in the subset.

In this study, we used the gene expression data of a diffuse large B-cell lymphoma (DLBCL) which was published by Alizadeh *et al*. (19). This data set consists of 40 patient samples including survival information on patients, and each sample has an expression profile of 4026 genes. By setting the threshold value of overall survival at 4 years, the prognostic data can be categorized into two groups, survival or death.

Among the 4026 genes, 1980 genes could not be subjected to classification because of missing values in microarray measurements. In conventional analysis, all values of such genes with missing microarray measurement values were deleted and disregarded. However, because there is a possibility that these genes with missing values are necessary for classification, we estimated missing values by the *k*-nearest neighbor (*k*-NN) method (20), followed by SVM analysis. The *k*-NN method selects the genes having the minimum Euclid distance to the target gene with the missing value and then substitutes the corresponding value of the selected gene for that of the target gene.

The leave-one-out cross validation (LOO) (18) was carried out for the fair test predictions for 40 patients. As shown in Table 2, the predictive accuracy was improved by applying the gene selection and the estimation of the missing value. The SVM model consisting of five selected genes shows the highest predictive accuracy (95%). These five genes might be biologically important genes for prognosis. The five selected genes for prognosis are JNK3, E2F-3, fvt-1, with the remaining two genes being labeled unknown. It was reported that JNK3 is related to apoptosis (21), E2F-3 to the cell cycle (22), and fvt-1 to lymphomas (23).

From these results, it is expected that SVM in conjunction with *k*-NN and gene selection will be a powerful classification method for prognostic predictions from DNA microarray data.

GENETIC NETWORK ANALYSIS FROM DNA MICROARRAY DATA

Recent advances in the technology of bioinformatics have enabled gene expression to be comprehensive, whereas several approaches have been proposed to infer the corresponding genetic networks (24–28). A systematic approach for modeling genetic networks was employed by inferring the global state of a genetic network and by highlighting the function and structure of each component of the network in terms of its function within the whole network. The clarification of transcription regulation networks has been tackled with a top-to-bottom method. The method is based on collected, ordered information of genetic expression profiles and transcription factors, which arises massively and simultaneously from DNA microarrays. The inference problem of genetic networks using experimentally observed data is generally referred to as an inverse problem and can be defined as the function optimization of the values of parameters involved in a suitable model-representation of a genetic network. We previously proposed an inferring engine in which any of the following four completely different network models work independently. The first and second models are given by a static Boolean network based on a Threshold-Test model (29, 30) and a multi-level digraph model (31), respectively, which can treat a large quantity of expression data. The third model is a Bayesian model (32) that statistically investigates the characteristics of dependence and conditional independence within the gene expression data set. The fourth model is a dynamic network model such as the S-system (30–32), which can infer a genetic network including groups of interdependent genes. The S-system belongs to a type of power-law formalism because it is based on a particular type of ordinary differential equation in which the component processes are characterized by power-law functions, namely,

$$\frac{d}{dt}X_i = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} \qquad (2)$$

where n is the total number of state variables or reactants (X_i), i, and j ($1 \le i, j \le n$) are the suffixes of state variables. The non-negative parameters α_i and β_i are apparent rate constants, and the real-valued exponents g_{ij} and h_{ij} show the interrelated effectivity of X_j to X_i. The first term represents all influences that increase X_i, whereas the second term represents all influences that decrease X_i. From a mathematical point of view, the S-system is the representation form in

TABLE 2. Comparison of predictive accuracy of SVM between treatments for missing value estimation and gene selection

Method	Initial number of genes	Selected number of genes	Predictive accuracy by LOO (%)
Without *k*-NN	2046	2046	62.5 (25/40)
Without *k*-NN, with gene selection	2046	11	90.0 (36/40)
With *k*-NN, with gene selection	4026	5	95.0 (38/40)

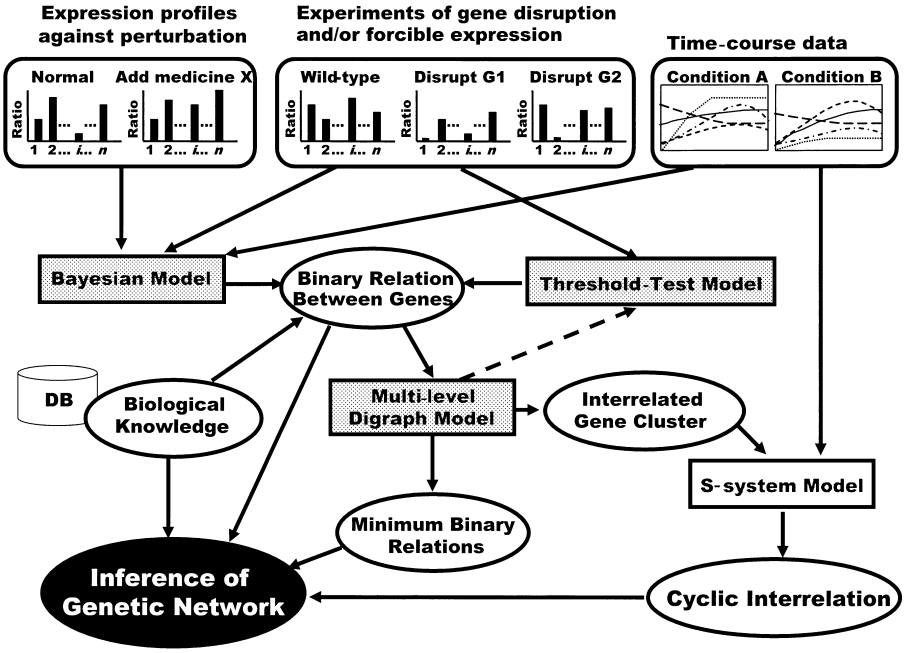

FIG. 6. Procedure of genetic network analysis using microarray data.

Cartesian space with the assumption of linear approximation in logarithmic space. The g_{ij} and h_{ij} exponents uniquely describe the interrelated coefficient of X_j to the synthetic process of X_i and to the degradation process of X_i, respectively. For instance, a certain g_{ij} with a positive value signifies that X_j induces the production of X_i, and one with a negative value signifies that X_j suppresses the production of X_i, one with a value of zero indicates that X_j does not affect the production of X_i. The S-system formalism includes a large number of parameters that must be estimated (α_i, β_i, g_{ij} and h_{ij}); the number of estimated parameters in the S-system formalism is $2n(n+1)$, where n is the number of state variables (X_i). We have to estimate a set of system parameters in the model which can realize experimentally observed time-course data. For the optimization of the estimation of large numbers of parameters, the real-coded genetic algorithm (RCGA) (33, 34) is introduced as a nonlinear numerical optimization method which is much less likely to be stranded in local minima.

The strategy for the inference of genetic interactions is summarized in Fig. 6. Given data of hold-change in the intensity of gene disruption or forcible expression under stationary state, the threshold-test model and Bayesian model are applied to infer binary relationships between target genes. In the case of expression profiles before and after adding an appropriate drug, or expression profiles between diseased and normal samples, the Bayesian model is again applicable. Additionally, the multi-level digraph model infers consistent minimal binary relationships starting from many binary relationships derived from the threshold-test or Bayesian models. If the dynamic changes (time-course) in intensity can be obtained, the S-system model is applied to infer cyclic interactions within a target gene cluster.

MATHEMATICAL MODEL AND SYSTEM ANALYSIS OF CELL CYCLE

Cell cycle is necessary for maintaining the homeostasis of a living body. The eukaryotic cell cycle is divided into four phases: the gap 1 (G1), synthesis (S), gap 2 (G2), and mitosis (M) phases. This cycle is orchestrated by the expression of the cell cycle genes, which form a complex and highly integrated network (35). The abnormal control of the cell cycle disrupts normal cell proliferation, which frequently results in the development of cancer. Thus, it is very important to understand the control mechanism of cell cycle for the comprehensive understanding of the development of cancer. The disruption in the control mechanism for the G1-to-S transition triggers an acquisition of infinite proliferation ability which is one of the characteristic features of cancer. In order to understand the complex and highly integrated G1-to-S transition, some mathematical models related to the G1-to-S transition have already been proposed by some researchers. One of the most detailed mathematical models was proposed by Aguda and Tang (36). In this study, we propose a new expanded mathematical model of the G1-to-S transition, and perform a system analysis for the new model. In the system analysis, by changing the values of all kinetic rate constants involved in this mathematical model, we analyzed how the stability of a mathematical model changes (stability analysis), and how the dynamic behavior of each chemical species is affected (sensitivity analysis). This mathematical model was constructed as an open system which assumed the constant fluxes of biological species from the outside. As a result, we estimated the dominant factors of the control mechanism of the G1-to-S transition.

In the reaction scheme for the G1-to-S transition (Fig. 7), the central reaction is RB protein (retinoblastoma: RB) phosphorylation. In the G1 phase, RB binds E2F (E2 promoter-

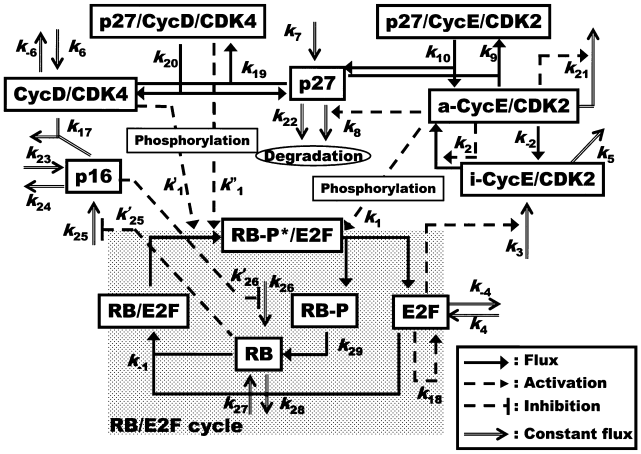

FIG. 7. Reaction scheme of G1-to-S transition in mammalian cell cycle.

binding factor) which is a transcription factor responsible for the induction of the S phase. By binding with E2F, RB inhibits the transcription by E2F. Next, RB is phosphorylated to RB-P* by kinases, and dissociates from E2F. E2F dissociated from RB induces the transcription of some genes coding for proteins that induce the S phase, and the cell cycle progresses to the S phase. Two cyclin complexes participating in this RB phosphorylation reaction are cyclin D/CDK4 (a complex of cyclin D and cyclin-dependent kinase 4), and cyclin E/CDK2 (a complex of cyclin E and cyclin-dependent kinase 2) (37).

We derived 12 simultaneous differential equations from our model. Based on the standard rate constants reported in Aguda's model, we numerically calculated these equations using the fourth-order Runge–Kutta method. By calculating the level of chemical species in the steady state, we calculated the values of elements in the Jacobian matrix analytically. Then, we calculated the eigenvalues of the Jacobian matrix by employing the Householder and QR methods. Finally, we evaluated the sensitivity of the rate constants, and analyzed the stability of the model using eigenvalues and the levels of chemical species in the steady state. This analysis was carried out by changing each rate constant with every 10% of a standard rate constant from 0% to 200%.

We checked the stability of the model to perturbation, and the level of each chemical species in the new steady state. As a result, a rate constant k_9 for the reaction where p27 binds with a-CycE/CDK2 was the most sensitive and important parameter for the stability of the model and the level of each chemical species in the steady state. Focusing on k_9, the signs of the real part of 11 of 12 eigenvalues were negative, one sign was positive from 40% to 200% of the standard rate constant. When one of all signs of the real part of eigenvalues is positive for the closed system without assuming fluxes of the biochemical species from the outside, the level of chemical species will diverge. However, our model was constructed as an open system, and the levels did not diverge. When k_9 was reduced to 30% of the value of the standard rate constant, all signs became negative, and the model was strictly stable.

When k_9 was set to 30% of the standard rate constant, the levels of some species in the steady state became over 140% compared with the case of the standard rate constant. The level of a-CycE/CDK2 in the steady state became approximately 150-fold higher than that using the standard rate constants. The level of p27 in the steady state became approximately 1/750, and the level of RB-P*/E2F in the steady state became approximately 1/250. These analysis results can be interpreted as follows: If a certain functional disorder makes k_9 decrease, then the p27 degradation pathway, which is dependent on the level of a-CycE/CDK2, is activated because of the increasing level of a-CycE/CDK2, which is followed by a marked decrease in the level of p27. With the decrease in the level of p27, the increase in a-CycE/CDK2 activates the RB/E2F cycle through the activation of RB-P*/E2F phosphorylation. Finally, with the activation of the RB/E2F cycle, the levels of E2F and a-CycE/CDK2 increase. It has been speculated that an activated RB/E2F cycle results in a rapid G1-to-S transition quickly, and the control mechanism of the G1-to-S transition is disrupted. Based on the analysis of the stability, it was revealed that the model becomes strictly stable by setting k_9 under 30% of the standard value. Thus, it was assumed that deviation from this stable state is very difficult. This phenomenon qualitatively agreed well with the acquisition of the infinite proliferation ability of cancer cell. It is obvious that a chronic decrease in k_9 disrupts the control mechanism of the G1-to-S transition and that k_9 is one of the dominant factors in the control mechanism of the G1-to-S transition. The report that the p27 level of patients with gastrointestinal stromal tumor is low compared with that of normal subjects (38) qualitatively supports the results of our system analysis.

Mathematical model and system analysis is a powerful tool for understanding biological phenomena and application to bioscience, bioengineering and medical fields.

PERSPECTIVES

In this review, the application of bioinformatics using DNA microarray data to the bioscience, bioengineering and medical fields as briefly introduced showing the results of our studies as the examples. Current researchers have access to and can use not only DNA microarray data but also proteome and metabolome data. Thus, we are currently in a data rich environment, and it is difficult to understand these data comprehensively using only human cogitation. It is well known that the bioinformatics techniques described in this review are powerful tools for the analysis of such data in a variety of fields. The next goal in the area of bioinformatics research is the development of a novel method for integrating different types of data and for understanding the cross-control mechanism between different layers, such as between the translation of protein and the transcription of mRNA, or between the genome sequence and transcription, or between the concentrations of metabolites and proteins, and so on. The application of bioinformatics in the treatment of large quantities of these –omics data is becoming more indispensable in the bioscience, bioengineering and medical fields.

ACKNOWLEDGMENTS

The authors are grateful to Ms. Chinatsu Arima, Ms. Chihoko Yoshimura, Mr. Yoshihiko Tashima, Dr. Kazumi Hakamada, Mr. Masahiko Nakatsui, Dr. Yukihiro Maki and Dr. Takanori Ueda for their substantial contributions to these studies. Our research in this review was supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Informatics Science" (nos. 13208008 and 12208008) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and the Project for Development of a Technological Infrastructure for Industrial Bioprocesses on R&D of New Industrial Science and Technology Frontiers by the Ministry of Economy, Trade and Industry (METI), and entrusted by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

1. **Schena, M., Shalon, D., Davis, R. W., and Brown, P. O.:** Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, **270**, 467–470 (1995).
2. **Yamakawa, S., Ando, K., Chisaka, A., Yoshida, K., Shinmyo, A., and Kohchi, T.:** Systematic transient assays of promoter activities for leaf-specific genes identified by gene-expression profiling cDNA microarrays in *Arabidopsis thaliana*. J. Biosci. Bioeng., **98**, 140–143 (2004).
3. **Mera, N., Aoyagi, H., Nakasono, S., Iwasaki, K., Saiki, H., and Tanaka, H.:** Analysis of gene expression in yeast protoplasts using DNA microarrays and their application for efficient production of inverse and α-glucosidase. J. Biosci. Bioeng., **97**, 169–183 (2004).
4. **Knudsen, S.:** Guide to analysis of DNA microarray data, p. 63–100. Wiley, Hoboken (2004).
5. **Eisen, M. B., Spellman, P. T., Brown, P. O., and Bostein, D.:** Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, **95**, 14863–14868 (1998).
6. **Somogyi, R.:** Making sense of gene-expression data. Pharmainformatics, 17–24 (1999).
7. **Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R.:** Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, **96**, 2907–2912 (1999).
8. **Huang, J., Shimizu, H., and Shioya, S.:** Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. J. Biosci. Bioeng., **96**, 421–428 (2003).
9. **Perou, C. M., Jeffrey, S. S., Rijn, M. V. D., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D.:** Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA, **96**, 9212–9217 (1999).
10. **Gasch, A. and Eisen, M.:** Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering. Genome Biol., **3**, 1–22 (2002).
11. **Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I.:** The transcriptional program of sporulation in budding yeast. Science, **282**, 699–705 (1998).
12. **Kupiec, M., Byers, B., Esposito, R. E., and Mitchell, A. P.:** The molecular and cellular biology of the yeast *Saccharomyces*, p. 889–1036. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1997).
13. **Tomida, S., Hanai, T., Honda, H., and Kobayashi, T.:** Analysis of expression profile using fuzzy adaptive resonance theory. Bioinformatics, **18**, 1073–1083 (2002).
14. **Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S.:** Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., **7**, 673–679 (2001).
15. **Hanai, T., Yatabe, Y., Nakayama, Y., Takahashi, T., Honda, H., Mitsudomi, T., and Kobayashi, T.:** Prognostic models in patients with no-small-cell lung cancer using artificial neural networks in comparison with logistic regression. Cancer Sci., **94**, 473–477 (2003).
16. **Ando, T., Suguro, M., Hanai, T., Kobayashi, T., Honda, H., and Seto, M.:** Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn. J. Cancer Res., **93**, 1207–1212 (2002).
17. **Vapnik, V. N.:** Statistical learning theory, p. 375–441. Wiley, Hoboken (1998).
18. **Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., and Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S.:** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, **286**, 531–537 (1999).
19. **Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., and other 20 authors:** Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, **403**, 503–511 (2000).
20. **Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B.:** Missing value estimation methods for DNA microarrays. Bioinformatics, **17**, 520–525 (2001).
21. **Yang, D. D., Kuan, C. Y., Whitmarsh, A. J., Rincon, M., Zheng, T. S., Davis, R. J., Rakic, P., and Flavell, R. A.:** Absence of excitotoxicity-induced apoptosis in the hippocampus of mice lacking the Jnk3 gene. Nature, **389**, 865–870 (1997).
22. **Dirks, P. B., Rutka, J. T., Hubbard, S. L., Mondal, S., and Hamel, P. A.:** The E2F-family proteins induce distinct cell cycle regulatory factors in p16-arrested, U343 astrocytoma cells. Oncogene, **17**, 867–876 (1998).
23. **Rimokh, R., Gadoux, M., Bertheas, M. F., Berger, F., Garoscio, M., Deleage, G., Germain, D., and Magaud, J. P.:** FVT-1, a novel human transcription unit affected by variant translocation t(2;18)(p11;q21) of follicular lymphoma. Blood, **81**, 136–142 (1993).
24. **Wessels, L., Someren, E., and Reinders, M.:** A comparison of genetic network models. Pac. Symp. Biocomput., **6**, 508–519 (2001).
25. **Wagner, A.:** Estimating coarse gene network structure from large-scale gene perturbation data. Genome Res., **12**, 309–315 (2002).
26. **Friedman, N., Linial, M., Nachman, I., and Pe'er, D.:** Using Bayesian network to analyze expression data. J. Comput. Biol., **7**, 601–620 (2000).
27. **Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S.:** Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. Pac. Symp. Biocomput., **8**, 17–28 (2003).
28. **Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M.:** Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics, **19**, 643–650 (2003).
29. **Hakamada, K., Hanai, T., Honda, H., and Kobayashi, T.:** A preprocessing method for inferring genetic interaction from gene expression data using Boolean algorithm. J. Biosci. Bioeng., **98**, 457–463 (2004).
30. **Arikawa, Y., Takahashi, Y., Watanabe, S., Maki, Y., Okamoto, M., and Eguchi, Y.:** Inference of a gene network from the experimentally observed expression data by using

AIGNET. Genome Inform., **12**, 274–275 (2001).
31. **Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., and Eguchi, Y.:** Development of a system for the inference of large scale genetic networks. Pac. Symp. Biocomput., **6**, 446–458 (2001).
32. **Maki, Y., Takahashi, Y., Arikawa, Y., Watanabe, S., Aoshima, K., Eguchi, Y., Ueda, T., Aburatani, S., Kuhara, S., and Okamoto, M.:** An integrated comprehensive workbench for inferring genetic networks: VoyaGene. J. Bioinform. Comput. Biol., **2**, 533–550 (2004).
33. **Ueda, T., Koga, N., Ono, I., and Okamoto, M.:** Efficient numerical optimization technique based on real-coded genetic algorithm for inverse problem, p. 290-293. *In* Sugisaka, M. and Tanaka, H. (ed.), Proceedings of the 7th International Symposium on Artificial Life and Robotics (AROB 7th. '02). Shubundo Insatsu, Oita (2002).
34. **Imade, H., Mizuguchi, N., Ono, I., Ono, N., and Okamoto, M.:** "Gridifying" an evolutionary algorithm for inference of genetic networks using the improved GOGA framework and its performance evaluation on OBI grid, p. 171–186. *In* Konagaya, A. and Satou, K. (ed.), Lecture notes in bioinformatics, vol. 3370. Springer, Heidelberg, Germany (2004).
35. **Kohn, K. W.:** Molecular interaction map of the mammalian cell cycle control and DNA repair systems. Mol. Biol. Cell, **10**, 2703–2784 (1999).
36. **Aguda, B. D. and Tang, T.:** The kinetic origins of the restriction point in the mammalian cell cycle. Cell Prolif., **32**, 321–335 (1999).
37. **Ekholm, S. V. and Steven, I. R.:** Regulation of G1 cyclin-dependent kinases in the mammalian cell cycle. Curr. Opin. Cell Biol., **12**, 676–684 (2000).
38. **Gelen, M. T., Elpek, G. O., Aksoy, N. H., Ogus, M., Suleymanlar, I., and Isitan, F.:** p27 expression and proliferation in gastrointestinal stromal tumors. Turk. J. Gastroenterol., **14**, 132–137 (2003).