

Inferring Genetic Networks from DNA Microarray Data by Multiple Regression Analysis

Mamoru Kato¹

mkato@ims.u-tokyo.ac.jp

Tatsuhiko Tsunoda²

tatsu@ims.u-tokyo.ac.jp

Toshihisa Takagi³

takagi@ims.u-tokyo.ac.jp

¹ Department of physics, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

² SNP Research Center, RIKEN, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

³ Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Abstract

Inferring gene regulatory networks by differential equations from the time series data of a DNA microarray is one of the most challenging tasks in the post-genomic era. However, there have been no studies actually inferring gene regulatory networks by differential equations from genome-level data. The reason for this is that the number of parameters in the equations exceeds the number of measured time points. We here succeeded in executing the inference, not by directly determining parameters but by applying multiple regression analysis to our equations. We derived our differential equations and steady state equations from the rate equations of transcriptional reactions in an organism. Verification with a number of genes related to respiration indicated the validity and effectiveness of our method. Moreover, the steady state equations were more appropriate than the differential equations for the microarray data used.

Keywords: gene networks, DNA microarray, differential equations, multiple regression analysis

1 Introduction

Along with the recent advancements in genome science, information on gene sequences has been exhaustively clarified. In the post-genomic era, interest has arisen regarding the elucidation of interactions between genes. The DNA microarray is capable of profiling the expression levels of many genes simultaneously, and is a promising technology for the elucidation of gene interactions. A number of studies have disclosed gene interactions from the time series data of the DNA microarray [1]–[5]. For example, DeRisi *et al.* [1] obtained 6153 mRNA expression levels of *Saccharomyces cerevisiae* at 7 time points and discovered that genes sharing a consensus regulatory sequence show similar expression patterns. Cho *et al.* [2] observed 6601 mRNA expression levels of *Saccharomyces cerevisiae* at 17 time points and investigated the upstream regulatory sequences of cell cycle-regulated genes. However, these studies did not disclose how genes regulate each other, *i.e.*, the *gene regulatory networks*.

To clarify the gene regulatory networks, some studies have attempted to apply differential equations to microarray data. D'haeseleer *et al.* [6] applied their linear differential equations to the data of 65 mRNA expression levels at 28 time points, which was an artificial combination of three different data sets of the rat, and discussed GAD/GABA interactions. Chen *et al.* [7] proposed linear differential equations on the concentrations of mRNA and protein. Akutsu *et al.* [8] applied the non-linear differential equations known as the S-system [9] to artificial data and inferred artificial regulation networks.

However, there has been no research into which differential equations are applied to actual data at the genome level. The reason for this is that the number of parameters in the equations thus far

suggested well exceeds the number of time points required for deciding them. For instance, when one simply applies the differential equations described in [6]–[9] to the data of about 6000 genes, the number of time points required is at least 6000, while the number of actually measured time points is about 7-17.

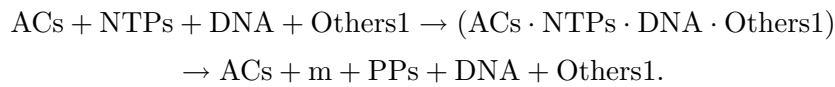
Rather than directly determining the parameters, we applied multiple regression analysis to our equations and thereby inferred gene regulatory networks from actual microarray data [1] at 7 time points. We derived our differential equations and steady state equations from the rate equations of transcriptional reactions in an organism, whereas other authors [6]–[9] have given their equations without consideration of transcriptional reactions in an organism. Verification with a number of genes related to respiration indicated the validity and effectiveness of our method. Moreover, we discovered that the steady state equations were more appropriate than the differential equations for the microarray data.

2 Materials and Methods

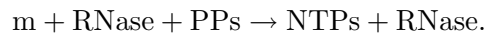
2.1 Modeling of the genetic control system

We model the genetic control system as a chemical reaction system as follows.

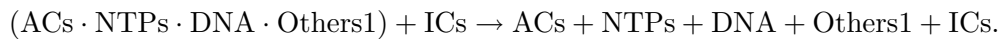
1. An mRNA is generated by the chemical reaction among NTPs as substrates, the DNA as a template, and activator complexes (or monomers) specifically bound on the enhancer or the promoter:



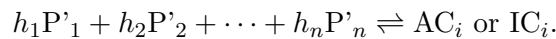
2. An mRNA is decomposed by reacting with an RNase:



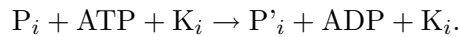
3. Inhibitor complexes (or monomers) specifically bind on the silencer or the promoter and inhibit the transcription, which is originally activated with the activators:



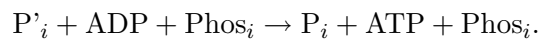
4. An activator or inhibitor complex is generated by the association of activated proteins:



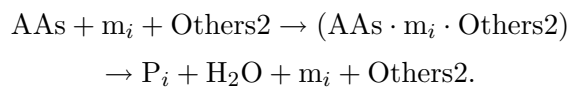
5. A protein is activated by the phosphorylation of a kinase, *etc.*:



6. The activated protein becomes inactive by the dephosphorylation of a phosphatase, *etc.*:



7. A protein is generated by the chemical reaction between amino acids as substrates and an mRNA as a template:



8. The generated protein disappears by being decomposed with a protease, or by changing chemically to other materials:

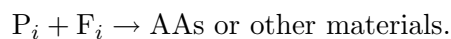


Table 1: Abbreviations in the reaction formulas 1-8.

| | |
|---|--|
| AC _{<i>i</i>} : The activator complex <i>i</i> , which is specific to the gene. In special cases, it is the activator monomer <i>i</i> . | F _{<i>i</i>} : The factor that decomposes or chemically changes P _{<i>i</i>} in specific. |
| IC _{<i>i</i>} : The inhibitor complex <i>i</i> , which is specific to the gene. In special cases, it is the inhibitor monomer <i>i</i> . | Others1: Non-specific factors necessary for transcription; in other words, a general transcription factor, an RNA polymerase, and Mg ²⁺ ions. |
| ACs: AC ₁ , AC ₂ , ... | Others2: Non-specific factors necessary for translation; in other words, ribosomes, tRNAs, an initiation factor, an elongation factor, and a termination factor. |
| ICs: IC ₁ , IC ₂ , ... | PP: Pyrophosphoric acid. |
| m _{<i>i</i>} : The mRNA <i>i</i> . | AA: Amino acid. |
| P' _{<i>i</i>} : The activated protein <i>i</i> . | h ₁ , h ₂ , ..., h _{<i>n</i>} : Stoichiometric coefficients. |
| P _{<i>i</i>} : The simple protein <i>i</i> . | |
| K _{<i>i</i>} : The kinase, <i>etc.</i> specific to P _{<i>i</i>} . | |
| Phos _{<i>i</i>} : The phosphatase, <i>etc.</i> specific to P' _{<i>i</i>} . | |

2.2 Deriving Model Equations by reaction kinetics

If a material, A, reacts with another material, B, the changes of the material concentrations follow rate equations of the materials. Conversely, if the rate equations are satisfied with the changes of the concentrations of A and some kind of material existing in a reactor with some materials at an arbitrary time, it is likely that the material will be B. We will infer transcriptional regulatory factors from this idea.

We derive rate equations from reaction formulas. We assume that a reaction order of a material concentration in a rate equation is equal to the stoichiometric coefficient in a reaction formula. From reaction formulas 1 and 2, we have

$$\frac{d[m]}{dt} = k_{m,1}[(ACs \cdot NTPs \cdot DNA \cdot Others1)] - k_{m,2}[m][RNase][PPs], \quad (1)$$

and from reaction formulas 1 and 3, we have

$$\begin{aligned} \frac{d[(ACs \cdot NTPs \cdot DNA \cdot Others1)]}{dt} &= k_{sc,1}[ACs][NTPs][DNA][Others1] \\ &\quad - k_{sc,2}[(ACs \cdot NTPs \cdot DNA \cdot Others1)] \\ &\quad - k_{sc,3}[(ACs \cdot NTPs \cdot DNA \cdot Others1)][ICs], \end{aligned} \quad (2)$$

where k is a rate constant. We consider the case where only the mRNA concentration is measured. If all material concentrations in these two equations are expressed by the mRNA concentration or constants, it will be decided whether all the concentrations satisfy the two rate equations. We express all the concentrations by the mRNA concentration or constants as follows.

Non-specific materials, NTP, ATP, ADP, PP, AA, H₂O, RNase, Others1, and Others2 are used for various reactions or various gene reactions. The materials are always needed in abundance in a cell. Therefore, the change in the amount of each material would be negligible compared with the total amount. We assume the concentration of each material to be a constant. Because the amount of DNA per cell is constant, the concentration of DNA is also assumed to be a constant.

Next, we express all the concentrations of the materials, except mRNA and the materials above, only by the mRNA concentration as follows. From reaction formula 4, we have

$$\frac{d[AC_i]}{dt} = k_{AC,1} \prod_j [P'_j]^{h_j} - k_{AC,2}[AC_i] \quad , \quad \frac{d[IC_i]}{dt} = k_{IC,1} \prod_j [P'_j]^{h_j} - k_{IC,2}[IC_i]. \quad (3)$$

From reaction formulas 5 and 6, we have

$$\frac{d[P'_i]}{dt} = k_{P',1}[P_i][K_i] - k_{P',2}[P'_i][Phos_i]. \quad (4)$$

From reaction formulas 7 and 8, we have

$$\frac{d[P_i]}{dt} = k_{P,1}[(AAs \cdot m_i \cdot Others2)] - k_{P,2}[P_i][F_i]. \quad (5)$$

And from reaction formula 7, we have

$$\frac{d[(AAs \cdot m_i \cdot Others2)]}{dt} = k_{la,1}[m_i] - k_{la,2}[(AAs \cdot m_i \cdot Others2)]. \quad (6)$$

By integrating the rate equation of each material, the concentration of each material at an arbitrary time will be expressed only by the mRNA concentration if the concentration of each material at the initial time is known.

In the case where no material concentration at the initial time is known, it is necessary to think about the approximation which does not at least contradict the rate equations. We assume that each material is in steady state (at least in local time), and apply a steady state approximation ($d[*]/dt \simeq 0$) to the rate equations 2-6 as follows:

$$\begin{aligned} [(ACs \cdot NTPs \cdot DNA \cdot Others1)] &\simeq \frac{k_{sc,1}[ACs]}{k_{sc,2} + k_{sc,3}[ICs]} \quad (\geq 0) \\ &\simeq K_{sc,1}[ACs] - K_{sc,2}[ACs][ICs] \quad (K_{sc,1} \geq K_{sc,2}[ICs]), \end{aligned} \quad (7)$$

where $[ICs]$ is assumed to be small, and the first-order Maclaurin's approximation is applied.

$$[ACs] = \prod_i [AC_i] \simeq K_{AC} \prod_i [P'_i]^{h_i}, \quad [ICs] = \prod_i [IC_i] \simeq K_{IC} \prod_j [P'_j]^{h_j} \quad (8)$$

$$[P'_i] \simeq K_{P'_i} [P_i][K_i]/[Phos_i] \quad (9)$$

$$[P_i] \simeq K_{P_i} [(AAs \cdot m_i \cdot Others2)]/[F_i] \quad (10)$$

$$[(AAs \cdot m_i \cdot Others2)] \simeq K_{la} [m_i] \quad (11)$$

Here, the concentration of each material at an arbitrary time will be expressed only by the mRNA concentration. However, the concentrations of the chemical species, K_i , $Phos_i$, and F_i , for which the rate equations are not set up, cannot be expressed by the mRNA concentration. In this research, it is assumed that the concentrations of the chemical species are constants (we will discuss the chemical species in Section 5.3).

When Equations 1 and 7-11 are rearranged, we have

$$\frac{d[m]}{dt} = a[ACs] - b[ACs][ICs] - c[m] \quad (a \geq b[ICs]) \quad (12)$$

$$[ACs] \equiv \prod_i [m_i]^{h_i}, \quad [ICs] \equiv \prod_j [m_j]^{h_j}, \quad (13)$$

where a , b , and c are parameters derived from rate constants. This equation is a non-linear differential equation of mRNA concentrations. If the steady state approximation is also applied to Equation 12,

$$[m] = a[ACs] - b[ACs][ICs]. \quad (14)$$

When we assume no effects of inhibitors, the $b[ACs][ICs]$ term is excluded from Equations 12 and 14. We show the rate equations that express only the effects of activators on mRNA in Table 2.

Table 2: Model Equations for activators. $[ACs] = \prod_i [m_i]^{h_i}$. $h_i \in \text{Integer}$.

| | |
|------------------|---------------------------|
| Model Equation 1 | $d[m]/dt = a[ACs] - c[m]$ |
| Model Equation 2 | $[m] = a[ACs]$ |

2.3 Materials

We applied Model Equations 1 and 2 to the DNA microarray data [1] made available to the public by Patrick Brown laboratory. The data contains fluorescence intensity values that correspond to the mRNA expression levels of 6153 genes at 7 time points taken at two-hour intervals. The values are the direct fluorescence intensity and the background noise intensity of both a test sample and a reference sample. The data was measured under the condition that the metabolism of *Saccharomyces cerevisiae* cultured at 30°C switched from fermentation to respiration upon the depletion of glucose.

We will now describe the relation between the fluorescence intensity and the mRNA concentration used in our Model Equations. T (g) is substituted for the mass of the total RNA of a test sample at each time point. The amount of the mRNA of gene i is $r_{i,t}T/M_i$ (mol), where $r_{i,t}$ is the ratio of the amount of the total RNA to that of the mRNA i at a time point t , and M_i is the molecular weight of the mRNA i .

After the reverse transcription of the RNA, a cDNA solution was prepared. The cDNA solution was hybridized with DNA probes on a slide glass. The fluorescence intensity emitted from the probed cDNA of gene i is expressed as:

$$I'_{i,t} = i_i(e_{H,i,t}V_{i,t})(e_{R,i}/V)(r_{i,t}T/M_i) \text{ (arbitrary units)}, \quad (15)$$

where $e_{R,i}$ is the efficiency of incorporating Cy3 and Cy5 during reverse transcription, V (l) is the volume of the cDNA solution, $e_{H,i,t}$ is the efficiency of the hybridization, $V_{i,t}$ is the volume of the printed spot, which contains DNA probes, and i_i is the fluorescence intensity emitted from the probed cDNA per mole. I' is equal to the remainder when the fluorescence intensity of the background noise is subtracted from the direct fluorescence intensity. In the same way, the fluorescence intensity of the reference sample is expressed as:

$$I'_{i,0} = i_i(e_{H,i,t}V_{i,t})(e_{R,i}/V)(r_{i,0}T/M_i) \text{ (arbitrary units)}. \quad (16)$$

The concentration of mRNA i is expressed as: $[m_i]_t = r_{i,t}n/v$ (mol/l), where n (mol) is the amount of total RNA per cell and v (l) is the volume of a cell. When $r_{i,t}$, $e_{H,i,t}$, and $V_{i,t}$ are eliminated using Equations 15 and 16, we have

$$[m_i]_t = (r_{i,0}n/v)I'_{i,t}/I'_{i,0} \text{ (mol/l)}.$$

Because the coefficient $r_{i,0}n/v$ is independent of time, it is included in the parameters of the Model Equations. We used the fluorescence intensity ratio $I'_{i,t}/I'_{i,0}$ instead of the mRNA concentration in the Model Equations.

2.4 Methods

As long as the factors in our model greatly influence the generation of mRNA, and the approximations are valid for the data, we can infer transcriptional regulatory factors by our Model Equation. That is, if our Model Equation is satisfied with the concentration of an mRNA and that of a regulatory factor candidate at an arbitrary time, the candidate is likely to be the true regulatory factor of the mRNA (see Section 2.2). One of the indices to measure how well measured values satisfy a model equation is a multiple correlation coefficient. We calculated multiple correlation coefficients on the measured concentration of an mRNA and that of every candidate. (A multiple correlation coefficient

does not change according to the linear transformation of measured values. Including the coefficient $r_{i,0}n/v$ in the parameters has no influence on a multiple correlation coefficient.) And we inferred that, if a candidate had a large multiple correlation coefficient, the candidate would be the true regulatory factor.

We will now describe the way to choose a regulatory factor candidate. A regulatory factor candidate is selected by determining the value of each h_i in a set of the stoichiometric coefficients, (h_1, h_2, \dots, h_N) (cf. Equation 13). When the value is determined, it means that we select a regulatory factor candidate of a $(h_1 + h_2 + \dots + h_N)$ -mer, which is divided into a h_1 -mer of gene 1, a h_2 -mer of gene 2, ..., a h_N -mer of gene N . The order of the computation time is $O(N^M)$ when we search all of the regulatory factor candidates of a M -mer in the system of N genes.

We will now explain the procedure for calculating a multiple correlation coefficient. Though only the procedure in Equation 12 is described here, the procedures in the other Model Equations are similar. First, using the measured concentrations of an mRNA, a selected activator candidate, and a selected inhibitor candidate, we calculate the regression coefficients: \hat{a} , \hat{b} , and \hat{c} . The regression coefficients are obtained by solving the normal equation, which is derived from the following equations:

$$\frac{\partial T}{\partial a} = 0, \quad \frac{\partial T}{\partial b} = 0, \quad \frac{\partial T}{\partial c} = 0,$$

where T is the residual sum of squares:

$$T = \sum_t e_t = \sum_t \left\{ \frac{d[m]_t}{dt} - (a[\text{ACs}]_t - b[\text{ACs}]_t[\text{ICs}]_t - c[m]_t) \right\}^2.$$

Because the data is given at discrete time points, the differential equation is approximated by difference approximation.

$$\frac{d[m]_t}{dt} \simeq \frac{([m]_{t+\Delta t} - [m]_t)}{\Delta t}$$

Because the rate constants are not negative, \hat{a} , \hat{b} , and \hat{c} are also not negative. The candidate which has negative \hat{a} , \hat{b} , or \hat{c} is excluded. Moreover, because the concentration of $(\text{ACs} \cdot \text{NTPs} \cdot \text{DNA} \cdot \text{Others}1)$ is not negative, the condition, $a \geq b[\text{ICs}]$, must be satisfied. If this inequality is not satisfied, the candidate is excluded. Using the regression coefficients, a multiple correlation coefficient is calculated. A multiple correlation coefficient is a (simple) correlation coefficient between the measured value, $d[m]/dt$, and the predicted value, \hat{y} , of a multiple regression equation. A multiple regression equation is expressed as:

$$\hat{y} = \hat{a}[\text{ACs}] - \hat{b}[\text{ACs}][\text{ICs}] - \hat{c}[m].$$

3 Results

3.1 Computational time

We will first report on the computational time when we selected all activator candidates of an M -mer in the system of N genes and calculated all the multiple correlation coefficients in Model Equation 1. Because the algorithm for Model Equation 2 is similar to that for Model Equation 1, the computational time in Model Equation 2 is also similar. The computational time was defined as the mean of the total time measured for ten cis-regulated genes, which were chosen at random. The machine we used was the Sun Ultra2, which had two 200-MHz Ultra SPARC CPUs and a memory of 512MB.

When $N \sim 6000$ and $M \leq 2$, the computational time was 507.3 (± 17.6) seconds per cis-regulated gene. This time corresponds to about 8.5 minutes. When $M \leq 3$, so much time was required that we were unable to obtain a computational time. We estimated the computational time from the order, $O(N^M)$, described in the Methods and measured time at N values of 100, 150, and 200. The measured time was 18.2, 61.4, and 144.0 seconds, respectively. Therefore, the computational time we estimated

Table 3: Activator candidates inferred by Model Equation 1.

| cis | trans1 | trans2 | trans3 | trans4 | ... |
|---------|------------------|------------------|------------------|------------------|-----|
| YNR001C | YBL108W/ YPR184W | YML090W/ YMR031C | YGL217C/ YLR299W | YDR380W/ YLR299W | ... |
| YNR002C | YLR039c/ YLR248W | YDL137W/ YNL040W | YBR111C/ YOR263C | YLR257W/ YML127W | ... |
| YNR003C | YLR272C/ YDL138w | YLR301W/ YMR019W | YLL015W/ YPL207W | YGR001C/ YLR437C | ... |
| YNR004W | YOL100W/ YPR194C | YNL297C/ YPR194C | YBL022C/ YDL019c | YDR135C/ YDR511W | ... |
| YNR005C | YAL017W/ YHR187W | YHR018C/ YHR162W | YML090W/ YML127W | YHR136C/ YMR186W | ... |
| ... | ... | ... | ... | ... | ... |

Table 4: Activator candidates inferred by Model Equation 2.

| cis | trans1 | trans2 | trans3 | trans4 | ... |
|---------|------------------|------------------|------------------|------------------|-----|
| YNR001C | YCR004C/ YKL103C | YKL142W/ YOR275C | YNL052W/ YPR070W | YIL029C/ YKL103C | ... |
| YNR002C | YIL062C/ YMR156C | YBR256C/ YGR129W | YMR148W/ YPR049C | YMR149W/ YPL223C | ... |
| YNR003C | YLR229C/ YPL127C | YIL133C/ YLR250W | YHR078W/ YNL210W | YDL121c/ YOR172W | ... |
| YNR004W | YGR033C/ YOR026W | YGR066C/ YLR103C | YGR247W/ YPL141C | YGR258C/ YOR174W | ... |
| YNR005C | YML095C/ YPL064C | YFR034C/ YKL052C | YBR154C/ YLR370C | YBL002W/ YHL034C | ... |
| ... | ... | ... | ... | ... | ... |

was about 45 days. If we search all candidates of a trimer or more, it is impossible to complete the computation in a short time. If we search all candidates of a monomer or a dimer, we can complete the computation.

3.2 Inferred results

The inferred activators are shown in Tables 3 and 4, and correspond to the results by Model Equations 1 and 2 (listed in Table 2), respectively. The two equations express only the effects of activators. We searched all the monomer and dimer candidates. The left-hand column contains the cis-regulated genes, and the other columns contain the trans-regulating genes of the activator candidates, which are arranged according to the order of multiple correlation coefficients.

4 Verification

4.1 *CIT1* and *CYC1*

We next verified whether our Model Equations were satisfied by the measured concentration of a true activator and that of the mRNA controlled by the activator. The degree of satisfaction was measured by a multiple correlation coefficient. Because there is no well-known dimeric activator that has been experimentally confirmed in the metabolic change from fermentation to respiration, we used the concentration of HAP 2/3/4 trimeric activator and the mRNA concentrations of *CIT1* and *CYC1*. It has been shown experimentally that *CIT1* and *CYC1* contain the binding site of the HAP 2/3/4 trimeric activator, and are activated with the activator when the metabolism of *Saccharomyces cerevisiae* shifts from fermentation to respiration [10]–[12]. Because no inhibitors of *CIT1* and *CYC1* are known to participate in the metabolic change, we used Model Equations 1 and 2 (Table 2), which do not express the effects of inhibitors. In Model Equation 1, multiple correlation coefficients were not calculated because neither condition $\hat{a} \geq 0$ nor $\hat{c} \geq 0$ of the regression coefficient was satisfied for

Table 5: Multiple correlation coefficients of HAP 2/3/4 complex for *CIT1* and *CYC1*.

| Gene | Model Equation 1 | Model Equation 2 |
|-------------|------------------|------------------|
| <i>CIT1</i> | - | 0.928 |
| <i>CYC1</i> | - | 0.906 |

both *CIT1* and *CYC1* (Table 5). In Model Equation 2, multiple correlation coefficients were 0.928 for *CIT1* and 0.906 for *CYC1*. These results suggest that the measured concentration would satisfy Model Equation 2, which is a steady state equation, better than Model Equation 1, which is a differential equation.

The multiple correlation coefficient of the HAP 2/3/4 complex was large for both *CIT1* and *CYC1* in Model Equation 2 (Table 5). However, not only the complex had a large value of a multiple correlation coefficient on *CIT1* and *CYC1*. In Model Equation 2, we chose 6000 activator candidates of a monomer, a dimer, and a trimer at random, and calculated the multiple correlation coefficients on the concentrations of the candidates and those of *CIT1* and *CYC1*. In the case of *CIT1*, 19.5% of the total candidates had multiple correlation coefficients of more than 0.9. In the case of *CYC1*, 2.0% of the total candidates had multiple correlation coefficients of more than 0.9. The ratios will show effectiveness for the inference when the threshold of the multiple correlation coefficient is considered to be 0.9. Because the ratio of *CYC1* was small, our method would be effective if we inferred *CYC1*. However, there were a number of false-positive candidates among the 2.0% of the total. Though our method would be expected to exclude the negative candidates, which were 98.0% of the total, it would not readily find the true activator out of the false-positive candidates.

4.2 Genes involved in respiration

Table 6 shows the multiple correlation coefficients of the genes involved in respiration. The genes are known to have the binding site of HAP 2/3/4 [1], [13]. The multiple correlation coefficients were calculated from the concentrations of the respiration genes and HAP 2/3/4 in Model Equations 1 and 2. Clearly, none of the genes have large multiple correlation coefficient values in Model Equation 1 (and most of the genes do not have multiple correlation coefficients because neither condition $\hat{a} \geq 0$ nor $\hat{c} \geq 0$ of the regression coefficient was satisfied). However, most of the genes have large multiple correlation coefficient values in Model Equation 2 (and for all of the genes, the condition $\hat{a} \geq 0$ was satisfied). These results indicate the inappropriateness of our differential equation method and the appropriateness of our steady-state equation method. *ATP1*, *ACO1*, etc., which have small multiple correlation coefficient values in Model Equation 2, might not be regulated by, or might not be regulated only by, the HAP 2/3/4 complex. Otherwise, the data on the genes might contain noticeable errors.

5 Discussion

5.1 The differential equation and the steady state equation

Why was the differential equation not appropriate? When a differential equation is applied to actual data at discrete time points, the differential equation must be approximated by difference approximation. However, the approximation is not reasonable if the intervals of discrete time are long. In that case, the differential equation will lead to erroneous results. Because the time width of the data used was long, two hours, the differential equation would not have been able to derive the correct result.

Why was the steady state equation appropriate? First of all, we consider the genetic control system of yeast to be an open chemical reaction system. An open chemical reaction system reaches steady state if no disturbance adds to the system from the outside. However, if a disturbance adds to the

Table 6: Multiple correlation coefficients of HAP 2/3/4 complex for genes involved in respiration. M. E. 1 means Model Equation 1, and M. E. 2 means Model Equation 2.

| Gene | M. E. 1 | M. E. 2 | Gene | M. E. 1 | M. E. 2 |
|--------------|---------|---------|--------------|---------|---------|
| <i>KGD2</i> | - | 0.990 | <i>QCR7</i> | - | 0.898 |
| <i>KGD1</i> | 0.835 | 0.959 | <i>COX5A</i> | - | 0.895 |
| <i>COR1</i> | 0.567 | 0.955 | <i>RIP1</i> | - | 0.885 |
| <i>ACH1</i> | - | 0.954 | <i>COX9</i> | - | 0.867 |
| <i>ACE1</i> | - | 0.944 | <i>CYT1</i> | - | 0.855 |
| <i>COX13</i> | - | 0.919 | <i>COX12</i> | - | 0.823 |
| <i>COX6</i> | - | 0.916 | <i>ACO1</i> | - | 0.817 |
| <i>COX4</i> | - | 0.911 | <i>ATP1</i> | 0.755 | 0.733 |

system, the steady state is broken and each material concentration changes. After enough time elapses without a disturbance, the system reaches steady state again. The material concentrations at that time are generally different from those at the first steady state. In this way, the mRNA concentration of the yeast would change over time. Next, we consider the case in which the non-steady state reaches the steady state quickly, compared with the intervals at which we sample data. Then, the state we macroscopically observe would be the steady state. In other words, the data we sample at long intervals would be in steady state. In conclusion, if data is sampled at long intervals, a steady state equation would be more appropriate than a differential equation.

5.2 The number of parameters in equations and that of measured time points

The previous studies [6]–[8] have not been able to infer genetic networks at the genome level because the number of parameters in equations exceeds the number of measured time points. Using the present method, however, we were able to infer them. The reason for this is that we did not directly decide the parameters in our equation from time series data. Rather, we gave the values of the integer parameters (h_1, h_2, \dots, h_N) beforehand and left the real number parameters (a, c in Model Equation 1, a in Model Equation 2) unknown, and measured how well the integer values and time series data satisfied our model equation. We compared the multiple correlation coefficients on various sets of the integer values and searched for the sets that satisfied our model equation well (Section 2.4). In this method, the number of measured time points that is needed for the inference depends on the number of the real number parameters. The number is more than two in Model Equation 1, and more than one in Model Equation 2.

5.3 Chemical species interacting with transcriptional regulatory factors

In this work, we treated the concentrations of chemical species interacting with transcriptional regulatory factors (K_i , Phos_i , and F_i) as constants. However, the concentrations should be originally treated as variables. The concentrations will be expressed by the activated protein concentrations because the chemical species are proteins that are activated with certain factors. Nevertheless, it is difficult to express the activated protein concentrations only by the mRNA concentrations. The reason is that the mRNA concentrations also accompany the activated protein concentrations when the activated protein concentrations expressing the chemical species are eliminated (Equations 9, 10, and 11).

5.4 Future work

This study indicated that our method is able to infer genetic networks from DNA microarray data. In future work, it will be necessary to verify the ability of this method for a number of gene regulatory systems that are experimentally understood. We treated the concentrations of the chemical species interacting with transcriptional regulatory factors as constants. In future studies, we will express the concentrations only by the mRNA concentration, and develop a method of inferring not only the transcriptional regulatory factors but also the chemical species interacting with the factors.

Acknowledgments

We thank *Tatsuya Akutsu* for his useful input.

References

- [1] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.
- [2] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz L., Conway A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W., A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell*, 2:65–73, 1998.
- [3] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- [4] 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.
- [5] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M., Systematic determination of genetic network architecture, *Nature Genetics*, 22:281–285, 1999.
- [6] D’haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R., Linear modeling of mRNA expression levels during CNS development and injury, *Proc. Pacific Symp. Biocomputing*, 4:41–52, 1999.
- [7] Chen, T., He, H.L., and Church, G.M., Modeling gene expression with differential equations, *Proc. Pacific Symp. Biocomputing*, 4:29–40, 1999.
- [8] Akutsu, T., Miyano, S., and Kuhara, S., Algorithms for inferring qualitative models of biological networks, *Proc. Pacific Symp. Biocomputing*, 5:293–304, 2000.
- [9] Irvine, D.H. and Savageau, M.A., Efficient solution of nonlinear ordinary differential equations expressed in S-system canonical form, *SIAM J. Numer. Anal.*, 27(3):704–735, 1990.
- [10] Forsburg, S.L. and Guarente, L., Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer, *Genes Dev.*, 3:1166–1178, 1989.
- [11] Olesen, J.T. and Guarente, L., The HAP2 subunit of yeast CCAAT transcriptional activator contains adjacent domains for subunit association and DNA recognition: model for the HAP2/3/4 complex, *Genes Dev.*, 4:1714–1729, 1990.
- [12] Rosenkrantz, M., Kell, C.S., Pennell, E.A., and Devenish, L.J., The HAP2,3,4 transcriptional activator is required for derepression of the yeast citrate synthase gene, *CIT1*, *Mol. Microbiol.*, 13(1):119–131, 1994.

- [13] Fondrat, C. and Kalogeropoulos, A., Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome 3, *Comput. Appl. Biosci.*, 12(5):363–374, 1996.