ICROS-SICE International Joint Conference 2009
August 18-21, 2009, Fukuoka International Congress Center, Japan

Estimation of Protein Networks for Cell Cycle in Yeast Based on Least–Squares Method

Noriko Takahashi[1] , Takehito Azuma[2] and Shuichi Adachi[1]

[1]Department of Applied Physics and Physico-Informatics, Keio University, Yokohama, Japan
(E-mail: takahashi_y8@arx.appi.keio.ac.jp)
[2]Department of Electrical and Electronic Engineering, Utsunomiya University, Utsunomiya, Japan
(E-mail: tazuma@cc.utsunomiya-u.ac.jp)

Abstract: In this paper, a new approach to estimation problems of protein networks is proposed, based on an idea of systems biology. Generally, it is difficult to estimate complicated networks by molecular biology. However, it will be possible to solve the difficulty by using the proposed approach. This approach is based on system identification using least–squares method for state–space models. Moreover, the proposed approach is applied to an estimation problem of protein networks for cell cycle in yeast. Nine proteins are selected from 48 proteins concerned with cell cycle in yeast, then 9–dimensional protein networks are estimated.

Keywords: System identification, Systems biology, Cell cycle, Protein networks, Least–squares method.

1. INTRODUCTION

Molecular biology is based on an idea to understand basic components of life such as genes, m–RNAs and proteins. By developments in this field, various researches on some genes that cause diseases have been reported. These researches might lead to establishments of efficient cure methods or to developments of medicines, so those are certainly important. By molecular biology, however, it is difficult to understand dynamic behaviors of life phenomena. Then, systems biology, which is based on an idea of regarding life as a system, was proposed [1][2] and various researches have been carried out [3].

One of the important themes in systems biology is cell cycle, which is the basic part of activities of cells and is known to be closely concerned with mechanisms of aging, cancer, and so on. In the researches for this cell cycle, many experiments about cell cycle in yeast have been performed. The reason is that in spite of being a unicellular organism, yeast's mechanisms of basic life phenomena are similar to that of multicellular organisms including humans. And as for yeast, there are budding yeast and fission one, whose cell cycle mechanisms are different each other. Especially for the budding yeast, various experimental results have been reported and some protein networks representing chemical reactions have been shown [4][5]. However, not all the networks in budding yeast have been discovered, so there can be still unknown proteins and networks. It is the protein networks that control cell cycle to go on regularly. So, it is desired to estimate these protein networks. For an estimation of protein networks, there was the experimental method in molecular biology. By using this method, however, it is difficult to estimate complicated networks because being subject to the constraints of facilities where experiments are preformed.

To estimate complicated networks, a new approach is proposed in which system identification based on least–squares method for state–space models is used. In this approach, considering that wave forms of protein concentrations can be obtained by experiments, the protein networks are estimated by identifying the linear system described by state–space models. Considering that each element of the identified matrix means strength of protein combinations, the protein networks can be estimated. Moreover, by using this approach, 9–dimensional protein networks for cell cycle in budding yeast are estimated, then the resulting networks show the efficacy of the proposed approach.

2. PROTEINS CONCERNED WITH THE CELL CYCLE IN BUDDING YEAST

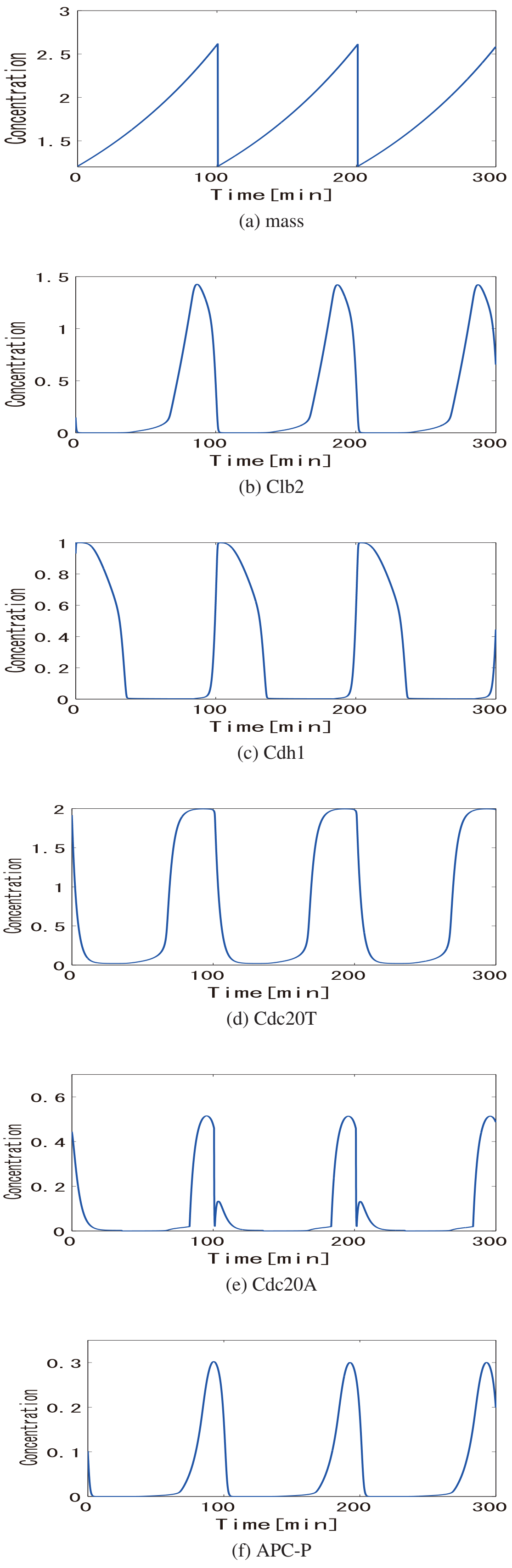
In this paper, for an estimation problem of protein networks in budding yeast, it is assumed that wave forms of the protein concentrations are obtained. Before using the experimental data, wave forms of the protein concentrations can be obtained from a set of non–linear differential equations by Chen, et al [5]. According to the research by Chen, 48 proteins are mainly concerned with the cell cycle in budding yeast, and their relations are represented by 48 non–linear differential equations and some reset rules. For example, differential equations about 9 proteins whose networks will be estimated in the next section are given as follows:

$$\frac{\mathrm{d}[\mathrm{mass}]}{\mathrm{d}t} = k_g[\mathrm{mass}] \tag{1}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Clb2}]}{\mathrm{d}t} =& (k'_{s,b2} + k''_{s,b2}[\mathrm{Mcm1}])[\mathrm{mass}] + (k_{d3,c1}[\mathrm{C2P}] \\ &+k_{di,b2}[\mathrm{C2}]) + (k_{d3,f6}[\mathrm{F2P}] + k_{di,f2}[\mathrm{F2}]) \\ &-(V_{d,b2} + k_{as,b2}[\mathrm{Sic1}] + k_{as,f2}[\mathrm{Cdc6}])[\mathrm{Clb2}]\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Cdh1}]}{\mathrm{d}t} =& k_{s,cdh} - k_{d,cdh}[\mathrm{Cdh1}] \\ &+\frac{V_{a,cdh}([\mathrm{Cdh1}]_{\mathrm{T}} - [\mathrm{Cdh1}])}{J_{a,cdh}([\mathrm{Cdh1}]_{\mathrm{T}} - [\mathrm{Cdh1}])} - \frac{V_{i,cdh}[\mathrm{Cdh1}]}{J_{i,cdh} + [\mathrm{Cdh1}]}\end{aligned} \tag{3}$$

PR0002/09/0000-0196 ¥400 © 2009 SICE



(a) mass

(b) Clb2

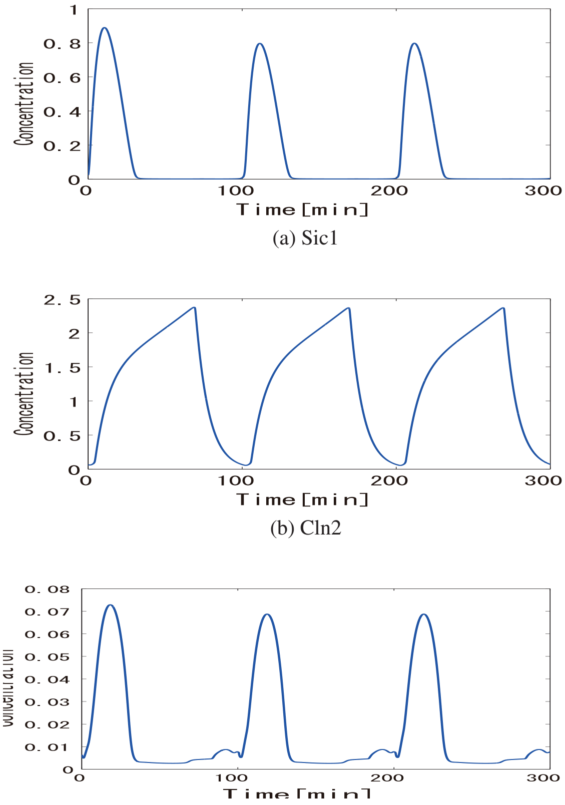
(c) Cdh1

(d) Cdc20T

(e) Cdc20A

(f) APC-P

Fig. 1 Time histories of concentration of proteins (1)



(a) Sic1

(b) Cln2

(c) Sic1P

Fig. 2 Time histories of concentration of proteins (2)

$$\frac{\mathrm{d}[\mathrm{Cdc20}]_{\mathrm{T}}}{\mathrm{d}t} = k'_{s,20} + k''_{s,20}[\mathrm{Mcm1}] - k_{d,20}[\mathrm{Cdc20}]_{\mathrm{T}} \tag{4}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Cdc20}]_{\mathrm{A}}}{\mathrm{d}t} &= (k'_{a,20} + k''_{a,20}[\mathrm{APC\text{-}P}])([\mathrm{Cdc20}]_{\mathrm{T}} - [\mathrm{Cdc20}]_{\mathrm{A}}) \\ &\quad -(k_{mad2} + k_{d,20})[\mathrm{Cdc20}]_{\mathrm{A}}\end{aligned} \tag{5}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{APC\text{-}P}]}{\mathrm{d}t} &= \frac{k_{a,apc}[\mathrm{Clb2}](1-[\mathrm{APC\text{-}P}])}{J_{a,apc}+1-[\mathrm{APC\text{-}P}]} \\ &\quad -\frac{k_{i,apc}[\mathrm{APC\text{-}P}]}{J_{i,apc}+[\mathrm{APC\text{-}P}]}\end{aligned} \tag{6}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Sic1}]}{\mathrm{d}t} &= (k'_{s,c1} + k''_{s,c1}[\mathrm{Swi5}]) + (V_{d,b2} + k_{di,b2})[\mathrm{C2}] \\ &\quad +(V_{d,b5} + k_{di,b5})[\mathrm{C5}] + k_{pp,c1}[\mathrm{Cdc14}][\mathrm{Sic1P}] \\ &\quad -(k_{as,b2}[\mathrm{Clb2}] + k_{as,b5}[\mathrm{Clb5}] + V_{kp,c1})[\mathrm{Sic1}] \\ &\quad +V_{kp,c1})[\mathrm{Sic1}]\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Cln2}]}{\mathrm{d}t} &= (k'_{s,n2} + k''_{s,n2}\cdot[\mathrm{SBF}])\cdot[\mathrm{mass}] \\ &\quad -k_{d,n2}\cdot[\mathrm{Cln2}]\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{Sic1P}]}{\mathrm{d}t} &= V_{kp,c1}[\mathrm{Sic1}] - (k_{pp,c1}[\mathrm{Cdc14}] + k_{d3,c1})[\mathrm{Sic1P}] \\ &\quad +V_{d,b2}[\mathrm{C2P}] + V_{d,b5}[\mathrm{C5P}]\end{aligned} \tag{9}$$

If this cell cycle model is constructed on a computer, wave forms of 48 protein concentrations can be obtained by running the program. For example, 9 waves of protein concentrations described from Eqs.(1) – (9) are shown in Figs.1 and 2. They show that these waves are periodic signals. By using these obtained concentrations of proteins, protein networks will be estimated in the next section.

3. ESTIMATION OF PROTEIN NETWORKS FOR CELL CYCLE

In general, the purpose of system identification is to derive a mathematical model of the object by using input and output data. However, because cell cycle in budding yeast is an autonomous system, there is no input data but protein concentrations can be obtained. Therefore, it is necessary to identify the object by using only output data. Then, in this section, it is assumed that state variables are known, then system identification based on least–squares method for state–space models is used. Moreover, this method is applied to an estimation problem of protein networks.

3.1 An estimation problem based on least–squares method for state–space models

It is assumed that the system is represented by

$$\begin{aligned} \boldsymbol{Y}(k) &= \boldsymbol{\Theta\Phi}(k) \qquad (10)\\ \boldsymbol{Y}(k) &= \boldsymbol{x}(k+1) \qquad (11)\\ \boldsymbol{\Phi}(k) &= \boldsymbol{x}(k), \qquad (12) \end{aligned}$$

where $\boldsymbol{\Theta}$ is the unknown parameter to be estimated. Let the criterion be

$$J = \frac{1}{N}\sum_{k=1}^{N}[\boldsymbol{Y}(k) - \boldsymbol{\Theta\Phi}(k)]^2. \qquad (13)$$

Then $\boldsymbol{\Theta}$ that minimizes J is given by

$$\hat{\boldsymbol{\Theta}} = \left(\sum_{k=1}^{N}\boldsymbol{Y}(k)\boldsymbol{\Phi}^T(k)\right)\left(\sum_{k=1}^{N}\boldsymbol{\Phi}(k)\boldsymbol{\Phi}^T(k)\right)^{-1}. \qquad (14)$$

Then, this method is applied to an estimation problem of the protein networks. If state variable $\boldsymbol{x}(k)$ denotes the obtained protein concentrations, $\boldsymbol{x}(k)$ is known. So, when rewriting $\boldsymbol{\Theta}$ in Eq.(10) as

$$\boldsymbol{\Theta} = \boldsymbol{A}_{cell}, \qquad (15)$$

then the protein networks describing the cell cycle model can be represented as the linear regression equation, that is

$$\boldsymbol{Y}(k) = \boldsymbol{A}_{cell}\boldsymbol{\Phi}(k), \qquad (16)$$

where

$$\begin{aligned} \boldsymbol{Y}(k) &= \boldsymbol{x}(k+1) \qquad (17)\\ \boldsymbol{\Phi}(k) &= \boldsymbol{x}(k). \qquad (18) \end{aligned}$$

In Eq.(16), $\boldsymbol{A}_{cell}$ is the matrix represented by

$$\boldsymbol{A}_{cell} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix},$$

where p is the number of proteins whose networks will be estimated.

Then Eq.(16) can be represented as

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_p(k+1) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_p(k) \end{bmatrix},$$

where each element of $\boldsymbol{A}_{cell}$ is the strength between one protein and the another. For example, the element a_{12} in $\boldsymbol{A}_{cell}$ means the effect that the protein x_2 at the time k gives to the protein x_1 at the time $k+1$. Hence, for an estimation problem of protein networks, the unknown parameter $\boldsymbol{\Theta}$ will be the matrix $\boldsymbol{A}_{cell}$ that indicates the strength of protein combinations. The larger certain element of the matrix is, the stronger the combination is. So, it can be considerd that the combination exists if the element is large enough.

3.2 An estimation result of 9–dimensional protein networks

By using the proposed approach, it is able to estimate the protein networks of 48 proteins. However, it is a little difficult to estimate 48–dimensional networks from the beginning. So at first, from 48 proteins concerned with the cell cycle, 9 proteins which are working throughout the cell cycle are selected, and then 9–dimensional networks are estimated. Because the state variable $\boldsymbol{x}(k)$ is represented by concentrations of selected proteins, $\boldsymbol{x}(k)$ is given by

$$\begin{aligned} \boldsymbol{x}(k) = \big[\; & [mass] \quad [Clb2] \quad [Cdh1] \\ & [Cdc20]_T \quad [Cdc20]_A \quad [APC-P] \\ & [Sic1] \quad [Cln2] \quad [Sic1P] \; \big]^T. \end{aligned}$$

From Figs.1 and 2, each protein concentration behaves differently, so they are normalized before estimating the networks. Number of data samples was 20000 where sampling period was 0.6sec, so the simulation time was 200min. Then from Eqs.(14), (15) and (16), $\boldsymbol{A}_{cell}$ was estimated as follows:

$$\boldsymbol{A}_{cell} = \left[\begin{array}{rrr} 9.95\text{E}{-1} & 1.63\text{E}{-3} & -1.36\text{E}{-3} \\ -1.10\text{E}{-3} & 1.00 & -4.37\text{E}{-4} \\ 2.37\text{E}{-3} & -3.85\text{E}{-3} & 1.00 \\ 1.13\text{E}{-3} & -2.49\text{E}{-4} & -1.32\text{E}{-3} \\ 2.55\text{E}{-3} & 4.43\text{E}{-3} & 7.70\text{E}{-4} \\ -5.74\text{E}{-4} & 4.45\text{E}{-3} & -2.60\text{E}{-4} \\ -7.54\text{E}{-4} & 8.25\text{E}{-4} & 5.68\text{E}{-4} \\ 5.13\text{E}{-4} & 3.01\text{E}{-5} & 7.20\text{E}{-4} \\ -9.28\text{E}{-5} & -8.27\text{E}{-4} & -9.15\text{E}{-4} \end{array} \right.$$

$$\begin{array}{rrr} 1.07\text{E}{-3} & 1.96\text{E}{-4} & 2.04\text{E}{-3} \\ -3.47\text{E}{-5} & 5.94\text{E}{-4} & -2.91\text{E}{-3} \\ 8.64\text{E}{-4} & -8.09\text{E}{-4} & 1.68\text{E}{-3} \\ 1.00 & -6.30\text{E}{-5} & -8.23\text{E}{-4} \\ -2.95\text{E}{-3} & 1.00 & -3.30\text{E}{-3} \\ -1.21\text{E}{-3} & 7.25\text{E}{-4} & 9.97\text{E}{-1} \\ 3.43\text{E}{-4} & 9.99\text{E}{-4} & -1.33\text{E}{-3} \\ -1.23\text{E}{-3} & -3.73\text{E}{-4} & 1.14\text{E}{-3} \\ 5.46\text{E}{-4} & -5.97\text{E}{-5} & 5.16\text{E}{-4} \end{array}$$

$$\left.\begin{matrix} 1.26\mathrm{E}{-3} & 2.45\mathrm{E}{-3} & -2.72\mathrm{E}{-4} \\ 3.06\mathrm{E}{-4} & 5.27\mathrm{E}{-4} & 3.28\mathrm{E}{-5} \\ 2.37\mathrm{E}{-4} & -1.19\mathrm{E}{-3} & 1.58\mathrm{E}{-4} \\ 4.41\mathrm{E}{-4} & -2.49\mathrm{E}{-4} & 9.67\mathrm{E}{-4} \\ -6.77\mathrm{E}{-4} & -1.20\mathrm{E}{-3} & 5.25\mathrm{E}{-5} \\ 2.76\mathrm{E}{-4} & 2.46\mathrm{E}{-4} & -1.04\mathrm{E}{-5} \\ 1.00 & 3.00\mathrm{E}{-4} & -1.67\mathrm{E}{-3} \\ 3.17\mathrm{E}{-4} & 1.00 & -6.57\mathrm{E}{-4} \\ 2.58\mathrm{E}{-3} & 1.02\mathrm{E}{-5} & 9.99\mathrm{E}{-1} \end{matrix}\right] .(19)$$

From this result, the fact that all the diagonal elements of $\boldsymbol{A}_{cell}$ are about 1 shows that each protein has self-feedback.

Next, examing the each row of $\boldsymbol{A}_{cell}$, and decide that the combination exists if the value is greater than 1/1000 of the maximum of the row. On the other hand, if the value is less than 1/1000 of the maximum of the row, the combination does not exist and the value is rewritten as 0. The criterion of 1/1000 is not derived theoretically, but is chosen so that already known networks can be estimated. Then the matrix $\boldsymbol{A}_{cell}$ is rewritten as

$$\boldsymbol{A}'_{cell} = \left[\begin{matrix} 9.95\mathrm{E}{-1} & 1.63\mathrm{E}{-3} & -1.36\mathrm{E}{-3} \\ -1.10\mathrm{E}{-3} & 1.00 & 0 \\ 2.37\mathrm{E}{-3} & -3.85\mathrm{E}{-3} & 1.00 \\ 1.13\mathrm{E}{-3} & 0 & -1.32\mathrm{E}{-3} \\ 2.55\mathrm{E}{-3} & 4.43\mathrm{E}{-3} & 0 \\ 0 & 4.45\mathrm{E}{-3} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}\right.$$

$$\begin{matrix} 1.07\mathrm{E}{-3} & 0 & 2.04\mathrm{E}{-3} \\ 0 & 0 & -2.91\mathrm{E}{-3} \\ 0 & 0 & 1.68\mathrm{E}{-3} \\ 1.00 & 0 & 0 \\ -2.95\mathrm{E}{-3} & 1.00 & -3.30\mathrm{E}{-3} \\ -1.21\mathrm{E}{-3} & 0 & 9.97\mathrm{E}{-1} \\ 0 & 9.99\mathrm{E}{-4} & -1.33\mathrm{E}{-3} \\ -1.23\mathrm{E}{-3} & 0 & 1.14\mathrm{E}{-3} \\ 0 & 0 & 0 \end{matrix}$$

$$\left.\begin{matrix} 1.26\mathrm{E}{-3} & 2.45\mathrm{E}{-3} & 0 \\ 0 & 0 & 0 \\ 0 & -1.19\mathrm{E}{-3} & 0 \\ 0 & 0 & 9.67\mathrm{E}{-4} \\ 0 & -1.20\mathrm{E}{-3} & 0 \\ 0 & 0 & 0 \\ 1.00 & 0 & -1.67\mathrm{E}{-3} \\ 0 & 1.00 & 0 \\ 2.58\mathrm{E}{-3} & 0 & 9.99\mathrm{E}{-1} \end{matrix}\right] . (20)$$

When estimating 9–dimensional protein networks from this obtained $\boldsymbol{A}'_{cell}$, the estimation result of the network connection is shown in Fig.3. It shows that not only known networks but also new ones were estimated, so the efficacy of the proposed approach for an estimation of protein networks is demonstrated.

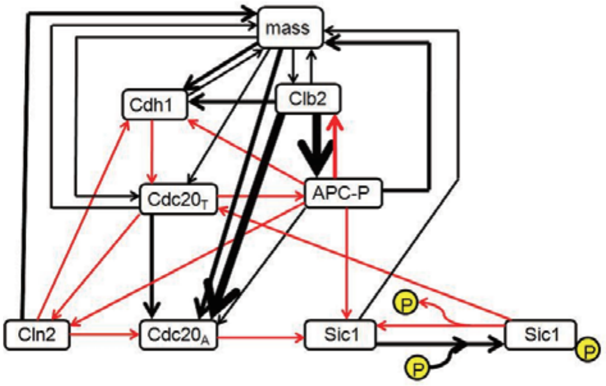


Fig. 3 An estimated 9–dimensional protein networks. Known ones (black) and new ones (red). Bolder lines indicate stronger protein combinations.

4. CONCLUSION

In this paper, for estimation problems of the protein networks, a new approach has been proposed in which system identification based on least–squares method for state–space models is employed. Moreover, the proposed approach was applied to an estimation problem of the protein networks for cell cycle in yeast. In this approach, the linear system described by state–space models was identified, and then each element of the identified matrix meant strength of protein combinations. By using this method, 9–dimensional protein networks can be estimated. Then the resulting networks show that not only known networks but also new ones were estimated, so the efficacy of the proposed approach for an estimation of protein networks is illustrated.

REFERENCES

[1] H. Kitano : Systems Biology : A brief overview, Science,Vol. 295, No. 5560, pp. 1662-1664 (2002)

[2] H. Kitano : Computational Systems Biology, Nature, Vol. 420, No. 6912, pp. 206-210 (2002)

[3] H. Kitano and T. Azuma : Systems Biology and control, system/control/information, Vol. 48, No. 3, pp. 104-111 (2004)

[4] J. Tyson : Modeling the Cell Division Cycle, Cdc2 and Cyclin Interactions, PNAS Vol. 88, pp. 7328-7332 (1991)

[5] K. C. Chen, et al : Integrative analysis of cell cycle control in budding yeast, Molecular Biology of the Cell, Vol.15, pp.3841–3862 (2004)