Application of Structural Equation Modelling for Oil Accumulation System Control in Oleaginous yeast

S Aburatani¹, Y Shida², W Ogasawara³, H Yamazaki³ and H Takaku³

¹. Computational Bio Big Data Open Innovation Lab. (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Main Bldg., 2-3-26, Aomi, Koto-ku, Tokyo, 135-0064, Japan
². Department of Bioengineering, Nagaoka University of Technology, 1603-1, Kamitomioka Nagaoka, Niigata, 940-2188, Japan
³. Department of Applied Life Sciences, Niigata University of Pharmacy and Applied Life Sciences, 265-1 Higashijima, Akiha-ku, Niigata city, Niigata, 956-8603, Japan

E-mail: s.aburatani@aist.go.jp and htakaku@nupals.ac.jp

Abstract. Recently, we developed a new statistical method for revealing the regulatory systems in living cells. Our method is based on Structural Equation Modelling combined with factor analysis. In generally, Structural Equation Modelling is utilized to detect the model adaptability with the measured data, such as large-scale questionnaire data. In this study, we improved our developed iteration algorithm and gene selection procedure to infer the causalities between variables as a regulatory network from limited numerical data. Our improved gene selection method is based on cross correlation to summarize the time preceding information from gene expression profiles, which were systematically measured at 8 time points. Cross correlation is usually utilized as a measure of similarity between two waves by a time-lag application, and we defined the values of lags ranging from −2 to +4. By this improved method, we selected 14 genes as regulatory factors for the specific system in oleaginous yeast. In the inferred model, only 6 genes among the selected 14 genes were considered to affect the volume of oil accumulation in a closed and specific system. Our method will be useful to artificially control cell systems in the bioproduction and biotechnology fields.

1. Introduction

In the bioproduction and biotechnology fields, expanding the ability of microorganisms is a fascinating theme, and numerous empirical approaches have been developed. Recently, systematic and synthetic approaches have been applied to this important theme for controlling the cell systems in microorganisms. Among the several types of microorganisms, oleaginous yeast is a suitable model. It has a specific system for oil accumulation, and its genomic information has been identified [1]. During oil accumulation, specific biosynthesis pathways are activated to generate triacylglyceride (TAG) inside the cell. In an oil accumulation system, the conversion of compounds is executed through sequential chemical reactions, which are branched into multiple pathways, and finally turn into TAG [2]. Although the reactions for generating TAG have been extensively studied, the regulation mechanism of this specific system has remained unexplained, and we are far from complete control of the system.

To deepen our knowledge and to control the oil accumulation system, the transcriptional regulatory systems in oleaginous yeast must be determined. In this regard, a gene regulatory network is useful to obtain better insight of the regulatory mechanisms in living cells. In our former investigations, we
developed a possibly useful technique for inferring causal relationships between un-observed regulatory factors and observed genes, based on structural equation modelling (SEM) [3]. The noteworthy features of SEM are the inclusion of latent variables into the constructed model and the ability to infer the network, including cyclic structures. We applied our approach to reveal the causalities within the well-studied serial transcriptional regulation systems in *S. cerevisiae*, *C. elegans*, and *D. melanogaster* [4,5,6]. The expressive features of SEM allowed the inferred model to include transcription factor proteins and other regulatory factors as latent variables, and genes as observed variables.

In this study, we developed a new method to select the suitable genes for the SEM calculations, and modified our SEM approach to infer the effective factors for controlling specific cell systems in living cells. We applied our improved approach to the expression of 14 genes and one set of phenotypic data.

2. Data Processing and Gene Selection

2.1. Experimental Data

For gene selection, we utilized the expression profiles of 7,785 genes, the cell concentration data (x10^8 cells/ml), and the ability of oil production data (mg/10^8 cells) measured in 4 different oleaginous yeast strains, which are known to have high oil accumulation ability. All data were measured at 8 time points from 0 to 240 hours (0, 24, 48, 72, 96, 144, 192, and 240), covering when the oil accumulation occurs [7]. First, we defined the ability of the oil accumulation system in each condition, as follows:

\[
OIL_{\text{time}}^{st} = \text{cell} \times \text{oil production}
\]  

Here, *cell*, and *oil production* are the cell concentration data and the oil production data, which were measured as the phenotypic data at every time point in each strain, and *OIL* ^{st} _{time} was calculated as the total amount of accumulated oil. Usually, the growth of living cells is inversely proportion to the oil production activity. To control the oil accumulation system, we defined the total amount of accumulated oil as an objective variable.

2.2. Data processing

To construct a suitable model, we should select genes that are arranged as exploratory variables in a network model from the expression profiles measured at 8 time points. From a biological viewpoint, it is considered that some time-gap will occur between the defined *OIL* ^{st} _{time} and the state of gene expression. To detect the time-gap between two variables, we developed a new gene selection method based on cross correlation, which is generally used in the signal processing field. In a time-series analysis, the cross correlation between two time-series describes the normalized cross covariance function. Let \( X_t = \{ X_1, \ldots, X_N \} \), \( Y_t = \{ Y_1, \ldots, Y_N \} \) represent two time-series data including \( N \) time points. The causalities between \( X_t \) and \( Y_t \) corresponding with the time discrepancy were detected by the cross correlation, given by

\[
\rho_{XY}^{\tau}(\tau) = \left\{ \frac{1}{(N-\tau)} \sum_{i=\tau+1}^{N} (X_i^{\tau} - \mu_x^{\tau}) (Y_i^{\tau} - \mu_y^{\tau}) \right\} \left\{ \frac{1}{(N-\tau)} \sum_{i=\tau+1}^{N} (X_i^{\tau} - \mu_x^{\tau}) \right\}^{1/2} \left\{ \frac{1}{(N-\tau)} \sum_{i=\tau+1}^{N} (Y_i^{\tau} - \mu_y^{\tau}) \right\}^{1/2}
\]

where \( X_i \) and \( Y_i \) are the gene expression, defined *OIL* ^{st} _{time} with \( N \) time points in each strain, and \( \tau \) as the time-lag between variables \( X \) and \( Y \). We calculated the cross correlations between all genes and the total amount of accumulated oil with \( \tau = -2, \ldots, +4 \), and the seven cross correlation values of the genes were calculated in each strain.
2.3. Gene Selection Procedure

In this study, we inferred the control system of oil accumulation, and thus the time-lag differences between strains could be merged to select the genes that are related to the oil accumulation system. The seven cross correlations were calculated for all genes in each strain with the lags -2, -1, 0, +1, +2, +3, and +4, and the seven absolute values of the cross correlations were compared. The highest absolute value of the cross correlation was selected as the fundamental relationship between the gene and the total amount of accumulated oil, and the selected cross correlation value was arranged as an element of a relative strength matrix. The value $\tau$ with the highest absolute value was considered as the causal information between the gene and the total amount of accumulated oil, and the selected lag value $\tau$ was arranged as a matrix element in a causality matrix. In a relative strength matrix and a causality matrix, the information of the causality strength and the causal direction between the gene and the oil accumulation system in the four strains were merged as a matrix, respectively. The schema of our developed procedure is displayed in Figure 1.

We subsequently selected the genes by a combination of these two matrices data. To select the genes with a strong relationship to the oil accumulation system, we applied a logical operation to the relative strength matrix. The absolute values of each element $\rho_{i,oil}(\tau)$ were tested for each gene. If the $\forall \rho_{i,oil}(\tau) \geq 0.9$, the gene was detected as having a strong relationship with the oil accumulation system. Among 7,785 genes, only 212 genes were selected. Those selected 212 genes include the effects of the oil accumulation system activity. To extract the causal genes, we calculated the summation of the raw elements $\tau_{i,oil}$ in the causality matrix. If $\forall \tau_{i,oil} > 0 \cap \sum(\tau_{i,oil}) \geq 9$, we selected the gene as a causal factor of the oil accumulation system. Finally, we selected 14 genes.

![Figure 1. The Schema of Gene Selection. Cross correlation values were calculated from measured data in each strain. The highest absolute values of cross correlation in each strain were merged as a relative strength matrix and a causality matrix.](image)

3. Modelling

3.1. Construction of Initial Model
In the first assumption, the genes were the effect variables and the total amount of oil, $OIL_{total}$, was defined as an objective variable. We arranged the selected 14 genes as the parent nodes for the total amount of oil accumulation as a child node in an initial model. The parent nodes were assumed to be independent, and the causality between the parent nodes was not identified in the initial model. In the matrix form, the initial model was expressed by

$$
\begin{bmatrix}
g \\
p
\end{bmatrix} = \begin{bmatrix}
\mathbf{O} & \mathbf{O} \\
\Lambda & \mathbf{O}
\end{bmatrix}\begin{bmatrix}
g \\
p
\end{bmatrix} + \begin{bmatrix}
g \\
\epsilon
\end{bmatrix}
$$

(3)

Here, $g$ is a vector of 14 genes, and $p$ is the data of the total amount of accumulated oil. Since the genes were assumed to be independent, the weights of the relationships between genes were defined as $O$ matrices. The matrix $\Lambda$ is a 14x1 vector representing the effectiveness of genes to the total amount of oil. In the initial model, all genes were defined as exogenous variables, which are not regulated by the other variables in the model. Thus, the data of the gene values were expressed as error terms of $g$. The errors that affect the total oil accumulation are denoted by $\epsilon$.

3.2. Structural Equation model without Latent Variables

After the construction of an initial model, we applied the SEM calculation to infer the network model that fit the measured data. Generally, SEM includes two types of variables within a model: observed and latent. In this study, the 14 genes and the total amount of oil have observed data, and all variables in the network model were defined as observed variables. None were defined as latent variables. The common regulator of several genes or the total amount of oil was dispensable for the inferred network, and the effect from the other control system in oleaginous yeast was calculated as error terms. The regulatory network model was simplified as follows:

$$
\mathbf{v} = \Lambda \mathbf{v} + \delta
$$

(4)

where $\mathbf{v}$ is a vector of 15 observed variables, including the 14 genes and the total amount of oil, and $\Lambda$ is a 15x15 matrix representing the strength of the regulatory relationships between the observed variables. Errors that affect all variables are denoted by $\delta$.

In the SEM calculation, the covariance analysis is defined as $\Sigma \approx \Sigma(\theta)$, where $\Sigma$ is the covariance matrix calculated from the observed data, and $\Sigma(\theta)$ is the matrix-valued function of the parameter $\theta$ in the model. Let $\Phi$ denote the covariance matrix of the error terms $\delta$, and $\mathbf{I}$ and $\Lambda$ denote the 15 x 15 matrices of the identity matrix and the inferred parameter matrix, respectively. The covariance matrix of model $\Sigma(\theta)$ was given by

$$
\Sigma(\theta) = (\mathbf{I} - \Lambda)^{-1} \Phi (\mathbf{I} - \Lambda)^{-1}^t
$$

(5)

Each element of the covariance matrix model $\Sigma(\theta)$ was expressed as a function of the parameters that appear in the model. The unknown parameters were estimated, in order to minimize the difference between the model covariance matrix $\Sigma(\theta)$ and the calculated covariance matrix $\Sigma$. The parameter estimation was performed to minimize the difference between $\Sigma$ and $\Sigma(\theta)$, and the maximum likelihood method was utilized as a fitting function to estimate the model parameters:

$$
F_{ML} (\Sigma, \Sigma(\theta)) = \log |\Sigma(\theta)| - \log |\Sigma| + tr(\Sigma(\theta)^{-1} \Sigma) - q
$$

(6)

Here, $|\Sigma|$ is the determinant of matrix $\Sigma$, $tr(\Sigma)$ is the trace of matrix $\Sigma$, and $q$ is the number of observed variables. The SEM software package SPSS AMOS 25 (IBM, USA) was used.

3.3. Model Evaluation

The quality of model adaptation was evaluated by the following criteria: CMIN, GFI, CFI, RMSEA, and AIC. The value of CMIN was calculated by the Chi-square statistic divided by the degrees of freedom [8]. The goodness-of-fit index (GFI) measured the relative discrepancy between the measured
data and the constructed model [9]. We also utilized the comparative fit index (CFI) and RMSEA as fitting scores [10,11]. These indices indicate whether the model is suitable to the measured data independent of a huge sample number. Furthermore, we evaluated the model fitting by AIC [12]. The AIC value of the constructed model was compared with those of the independent model and the saturation model.

These indices have threshold values as criteria to decide whether the model is suitable with the measured data. The value of CMIN higher than 0.05 was considered as a well fitted model, and the values of GFI and CFI greater than 0.90 were also evaluated as well fitted models. In the RMSEA criteria, a value of about 0.05 or less would represent a close fit of the model in relation to the degrees of freedom, and a value of about 0.08 or less would indicate a reasonable error, even though values higher than 0.10 were considered to indicate that the constructed model is far from the actual data.

3.4 Model Optimization

In our former investigations, we developed a model optimization algorithm for the SEM calculation [3,4]. In this study, we improved our iteration algorithm to escape from a local optimal solution. In our optimization algorithm, the non-significant edge was detected by the probability calculated from the inverse matrix of the Fisher information matrix of parameters, and the possible relationships between the variables were estimated by the modification index (MI) scores, which measure how much the chi-square statistic is expected to decrease if a particular parameter setting is constrained. In our former algorithm, we applied this iteration algorithm to the observed and latent variables at first, and then the error terms were executed.

In this study, we applied this iteration algorithm to the error terms when the MI scores became smaller than 10.0. Since the error terms can influence the other control systems in a living cell, we can obtain a general solution with optimized error terms. Furthermore, we improved the iteration algorithm combined with the genetic algorithm (GA). To avoid the local minima, a point mutation was generated when the largest MI score became less than 5.0. The improved iteration algorithm is displayed in Figure 2.
This improved iteration algorithm was stopped when added random edges were detected as non-significant edges. To avoid the local minima, the significance level was set as $p<0.05$ at the early stage of this iteration process, and the significance level became restrictive at the latter stage.

4. Results

By our developed gene selection method, we detected 14 genes, and the SEM calculation was applied to those selected 14 genes and the total oil accumulation as one phenotypic variable calculated by equation (1). We estimated 27 causal relationships between 15 variables. The optimal model inferred by our improved iteration algorithm is displayed in Figure 3a. Among the 14 selected genes, 8 genes were not related to the amount of oil accumulation in the network model, even though the measured data indicated those 8 genes had a strong correlation with the total amount of oil. In the optimal model, 7 genes among those 8 genes were apparent from the oil accumulation system, but the super parent node SPC2 was common between those 7 genes and the oil accumulation system. The one remaining gene YDL114W was included within the oil accumulation system, but it was not a control factor of oil accumulation.

The subgraph, which was related with oil accumulation, of the inferred model is displayed in Figure 3b. From this figure, we can consider that the oil accumulation system in oleaginous yeast had two control systems. One is a simple positive control system, which started from K-72537 through K_61530. We can expect to activate the oil accumulation system by the overexpression of these genes. The other system was more complicated, and started from SPC2. This system had four control branches for oil accumulation, and three of them were negative controls.
Figure 3. The inferred network models. The rounded squares indicate genes, and the squares indicate the total amount of OIL. The solid lines indicate positive regulations, and the dashed lines indicate negative regulations. The bold line represents the strong regulation with the absolute value of the edge weight higher than 0.5, and that of the thin line is less than 0.5. The relationships between error terms are not displayed here.

5. Conclusions
Here we have shown the inferred network model of an oil accumulation system in oleaginous yeast by SEM without latent variables. In this model, some genes with unknown function were identified as control factors of the oil accumulation system. This model does not represent the molecular mechanism occurring in a cell, but rather serves as a guide of an empirical approach, identifying some genes that control the total amount of oil accumulation. In our analysis, the effect of the other cellular systems in oleaginous yeast can be included as error terms. We will empirically verify the accuracy of this inferred model in the near future.

Acknowledgements
The experimental data were supported by funding from the Science and Technology Research Promotion Program for Agriculture, Forestry, Fisheries and Food Industry.

References
[1] Riley R, Haridas S, Wolfe K H, Lopes M R, Hittinger C T, Goker M, Salamov A A, Wisecaver J H, Long T M, Calvey C H, Aerts A L, Barry K W, Choi C, Clum A, Coughlan AY, Deshpande S, Douglass A P, Hanson S J, Klenk H P, LaButti K M, Lapidus A, Lindquist E A, Lipzen A M, Meier-Kolthoff J P, Ohm R A, Otiillar R P, Pangilinan J L, Peng Y, Rokas A, Rosa C A, Scheuner C, Sibirny A A, Slot J C, Stielow J B, Sun H, Kurtzman C P, Blackwell M, Grigoriev I V and Jeffries T W 2016 Proc. Natl. Acad. Sci. U.S.A. 113(35) 9882
[2] Sutanto S, Zullaikah S, Tran-Nguyen P L, Ismadji S and Ju Y H 2018 Fuel Process. Technol., 177, 39
[3] Aburatani S 2011 Gene Regul. Syst. Biol. 5 S7569
[4] Aburatani S 2012 Bioinformation 8(14) 652
[5] Aburatani S and Toh H 2012 Basic Methods in Protein Purification and Analysis 14 (Hong Kong: iConcept Press Ltd) Chapter 5
[6] Aburatani S and Toh H 2014 J. Phys. Conf. Ser. 490 (1) 012145
[7] Yamazaki H, Kobayashi S, Ebina S, Abe S, Ara S, Shida Y, Ogasawara W, Yaoi K, Araki H and Takaku H Appl. Microbiol. Biotechnol. 2019 Accepted
[8] Wheaton B, Muthen B, Alwin D F and Summers G F 1977 Sociol. Methodol. 1977, ed. Heise D R (San Francisco: Jossey-Bass) pp.84-136
[9] Joreskog K G and Sorbom D 1984 LISREL-V1 user’s guide (3rd ed.) (Scientific software)
[10] Browne M W and Cudeck R 1993 Testing structural equation models, ed. Boolen K A and Long J S (Newbury Park, CA: Sage) pp.136-162
[11] Bentler P M 1990 Psychological Bulletin 107 238
[12] Akaike H 1987 Psychometrika 52 317