Remark 1. If *comp* is a positive dimensional, then we can always perturb the set of polynomials in *comp* in order to obtain a zero-dimensional variety. Although here we cannot discuss the stability and convergency issues related to such perturbations, it is an important research issue on its own light (see [7] for an example).

In this paper, we have calculated the other consistency measure (in short, *CM2*) as the smallest $g(\vec{k})$ under the following constraint:

$$k_1 \geq 0, k_2 \geq 0, \ldots, k_m \geq 0. \tag{2.6}$$

The difference between Constraints: (2.4) and (2.6) is that one takes account of the zero value of the kinetic constants $\vec{k}$, corresponding to the non-existence of edges in the network. This account yields a finer model selection where all of the subnetworks of the presupposed network are also considered. We can calculate the smallest value of $g(\vec{k})$ under Constraint (2.6), using the following recursive definition:

Let $MinimumValue(q(\vec{l}))$ denote the *minimum value* of function q with variables: $\vec{l} = \{l_1, l_2, \ldots, l_m\}$ by the following procedure:

1. If the cardinality of $\vec{l}$, namely m, is zero, then the *minimum value* is infinity.
2. Otherwise, let v_0 denote the minimum value of q under Constraint (2.4) via 'NSolve.' Furthermore, let v_i $(i = 1, 2, \ldots, m)$ denote the value calculated by $MinimumValue(q(\vec{l}_i))$, where $\vec{l}_i$ is the vector: $\{l_1, l_2, \ldots, l_{i-1}, 0, l_{i+1}, \ldots, l_m\}$.
3. The *minimum value* is the smallest value among $v_0, v_1, \ldots, v_m$.

Model Selection. Using the consistency measure defined in §2.3, we performed a model selection. We, first, calculated the consistency measures among all of the combinations of the presupposed models with the sampling data. Next, we arranged the combinations of the models with the data in ascending order by the consistency measure. Last, we estimated the most consistent model having the first element (the smallest value).

3 Results and Discussion

3.1 Preparations: Transformation into Laplace Domain

Model Formula. We analyzed the models for a relationship between specific leaf area, leaf nitrogen, and leaf gas exchange in botany [9]. In the original paper, six models for the kinetics of four biomolecules are listed, and the consistency of the models with the observed data, which are composed of various properties of the molecules, rather than time series data, are tested by the d-sep test. In this paper, four of the six original models (models A, B, C, and D) and one model (model E) modified from the original one are considered, to show how cyclic relationships can be handled. The models considered in this paper are shown in Fig. 3. Each model expressed the relationship between four biomolecules, *SLA*, *N*, *A*, and *G*. According to the definition in §2.2, each relationship

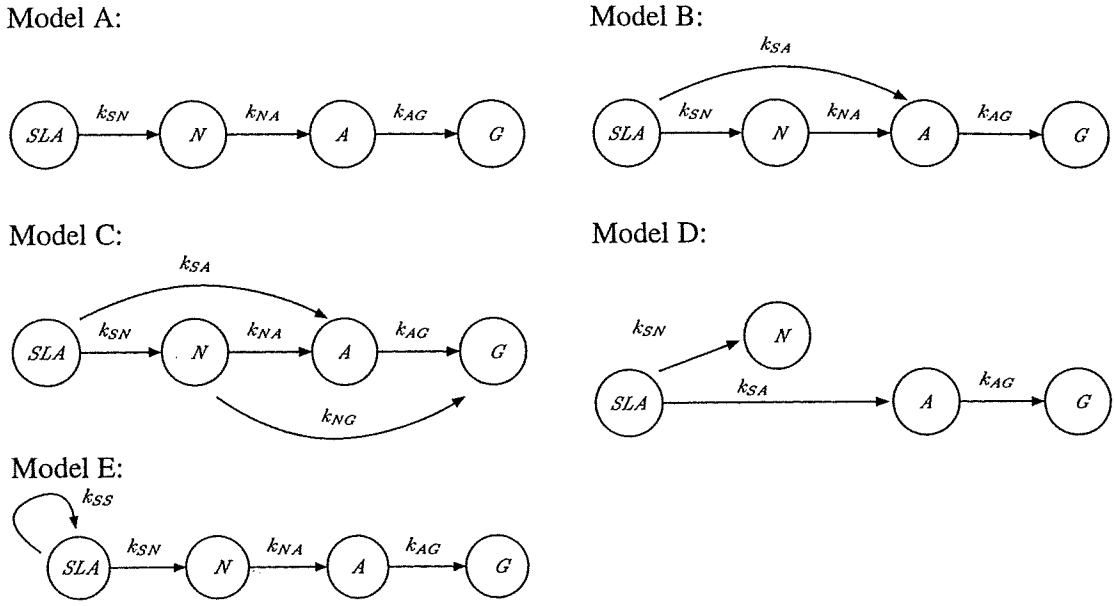

Fig. 3. Models analyzed in the present study. In the above models, the causal relationships between molecules are denoted by arrows. The molecules corresponding to the variables, denoted within the circles, are SLA, N, A, and G, and the kinetic parameters, denoted over the arrows, are k_{SN}, k_{NA}, k_{AG}, k_{SA}, k_{NG}, and k_{SS}.

between the variables is assumed to be linear, and then the differential equations for the five models can be formulated as follows:

Model A:

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, SLA(t) = -k_{SN}\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, N(t) = k_{SN}\, SLA(t) - k_{NA}\, N(t), \\ \mathrm{d}/\mathrm{d}t\, A(t) = k_{NA}\, N(t) - k_{AG}\, A(t), \\ \mathrm{d}/\mathrm{d}t\, G(t) = k_{AG}\, A(t). \end{cases} \tag{3.1}$$

Model B:

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, SLA(t) = -(k_{SN} + k_{SA})\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, N(t) = k_{SN}\, SLA(t) - k_{NA}\, N(t), \\ \mathrm{d}/\mathrm{d}t\, A(t) = k_{SA}\, SLA(t) + k_{NA}\, N(t) - k_{AG}\, A(t), \\ \mathrm{d}/\mathrm{d}t\, G(t) = k_{AG}\, A(t). \end{cases} \tag{3.2}$$

Model C:

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, SLA(t) = -(k_{SN} + k_{SA})\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, N(t) = k_{SN}\, SLA(t) - (k_{NA} + k_{NG})\, N(t), \\ \mathrm{d}/\mathrm{d}t\, A(t) = k_{SA}\, SLA(t) + k_{NA}\, N(t) - k_{AG}\, A(t), \\ \mathrm{d}/\mathrm{d}t\, G(t) = k_{AG}\, A(t) + k_{NG}\, N(t). \end{cases} \tag{3.3}$$

Model D:

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, SLA(t) = -(k_{SN} + k_{SA})\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, N(t) = k_{SN}\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, A(t) = k_{SA}\, SLA(t) - k_{AG}\, A(t), \\ \mathrm{d}/\mathrm{d}t\, G(t) = k_{AG}\, A(t). \end{cases} \tag{3.4}$$

Model E:

$$\begin{cases} \mathrm{d}/\mathrm{d}t\, SLA(t) = (k_{SS} - k_{SN})\, SLA(t), \\ \mathrm{d}/\mathrm{d}t\, N(t) = k_{SN}\, SLA(t) - k_{NA}\, N(t), \\ \mathrm{d}/\mathrm{d}t\, A(t) = k_{NA}\, N(t) - k_{AG}\, A(t), \\ \mathrm{d}/\mathrm{d}t\, G(t) = k_{AG}\, A(t). \end{cases} \tag{3.5}$$

In the above equations, k_{SN}, k_{NA}, k_{AG}, k_{SA}, k_{NG}, and k_{SS} are the kinetic parameters between the molecules. Notice that the relationships between the molecules in the actual kinetics cannot be expressed by the above equations. In the actual case, some relationships are non-linear, such as the well-known Michaelis–Menten kinetics in enzyme reactions. In the present study, we have adopted the relationships between molecules as typical ones, but do not consider the details of the kinetics between molecules.

According to the definitions in §2.2, we transform the above systems of differential equations of (3.1)–(3.5) into the system of algebraic equations over the Laplace domain, and solve the equations for the five models. For instance, the solution to the system of differential equations for Model A is expressed over the Laplace domain, as follows:

$$\begin{cases} L[SLA(t)](s) = \dfrac{SLA(0)}{s + k_{SN}}, \\ L[N(t)](s) = \dfrac{N(0)\,s + N(0)\,k_{SN} + k_{SN}\,SLA(0)}{s^2 + (k_{SN} + k_{NA})\,s + k_{NA}\,k_{SN}}, \\ L[A(t)](s) = (A(0)\,s^2 + (k_{NA}\,N(0) + A(0)\,k_{SN} + A(0)\,k_{NA})\,s + k_{NA}\,N(0)\,k_{SN} \\ \quad + k_{NA}\,k_{SN}\,SLA(0) + A(0)\,k_{NA}\,k_{SN})/(s^3 + (k_{AG} + k_{SN} + k_{NA})\,s^2 + (k_{AG}\,k_{SN} \\ \quad + k_{AG}\,k_{NA} + k_{NA}\,k_{SN})\,s + k_{AG}\,k_{NA}\,k_{SN}), \\ L[G(t)](s) = (G(0)\,s^3 + (G(0)\,k_{AG} + G(0)\,k_{NA} + k_{AG}\,A(0) + G(0)\,k_{SN})\,s^2 \\ \quad + (G(0)\,k_{AG}\,k_{NA} + k_{AG}\,k_{NA}\,N(0) + G(0)\,k_{AG}\,k_{SN} + k_{AG}\,A(0)\,k_{NA} \\ \quad + k_{AG}\,A(0)\,k_{SN} + G(0)\,k_{NA}\,k_{SN})s + k_{AG}\,k_{NA}\,N(0)\,k_{SN} + k_{AG}\,k_{NA}\,k_{SN}\,SLA(0) \\ \quad + G(0)\,k_{AG}\,k_{NA}\,k_{SN} + k_{AG}\,A(0)\,k_{NA}\,k_{SN})/(s^4 + (k_{AG} + k_{SN} + k_{NA})\,s^3 \\ \quad + (k_{AG}\,k_{SN} + k_{AG}\,k_{NA} + k_{NA}\,k_{SN})\,s^2 + s\,k_{AG}\,k_{NA}\,k_{SN}). \end{cases} \tag{3.6}$$

In the above equations, the initial values for each molecule are denoted by $SLA(0)$, $N(0)$, $A(0)$, and $G(0)$.

Sampling Data Fitting. To estimate the consistency of the above equations derived from the models with the data, we should presuppose the equations for the sampling data. For this purpose, first, a series of exponentials with parameters are set. For instance, the equations for fitting to the data in Model A are expressed as follows:

$$\begin{cases} SLA_O(t) = \beta_{SLA,1} \exp(-\alpha_{SLA,1} t), \\ N_O(t) = \beta_{N,1} \exp(-\alpha_{N,1} t) + \beta_{N,2} \exp(-\alpha_{N,2} t), \\ A_O(t) = \beta_{A,1} \exp(-\alpha_{A,1} t) + \beta_{A,2} \exp(-\alpha_{A,2} t) + \beta_{A,3} \exp(-\alpha_{A,3} t), \\ G_O(t) = \beta_{G,1} \exp(-\alpha_{G,1} t) + \beta_{G,2} \exp(-\alpha_{G,2} t) + \beta_{G,3} \exp(-\alpha_{G,3} t) + \beta_{G,4}. \end{cases} \tag{3.7}$$

Then, the corresponding algebraic equations are obtained by the Laplace transformation. The corresponding algebraic equations in Model A are as follows:

$$\begin{cases} L[SLA_O(t)](s) = \dfrac{\beta_{SLA,1}}{s + \alpha_{SLA,1}}, \\ L[N_O(t)](s) = \dfrac{\beta_{N,1}}{s + \alpha_{N,1}} + \dfrac{\beta_{N,2}}{s + \alpha_{N,2}}, \\ L[A_O(t)](s) = \dfrac{\beta_{A,1}}{s + \alpha_{A,1}} + \dfrac{\beta_{A,2}}{s + \alpha_{A,2}} + \dfrac{\beta_{A,3}}{s + \alpha_{A,3}}, \\ L[G_O(t)](s) = \dfrac{\beta_{G,1}}{s + \alpha_{G,1}} + \dfrac{\beta_{G,2}}{s + \alpha_{G,2}} + \dfrac{\beta_{G,3}}{s + \alpha_{G,3}} + \dfrac{\beta_{G,4}}{s}. \end{cases} \tag{3.8}$$

Notice that the parameters in the above equations are estimated by numerically fitting them to the data.

3.2 Estimation of Model Consistency

Data Generation for Simulation. In the present study, we have no actual data for the molecules in the models, and thus we need to generate the time series of data for the constituent molecules for the simulation study, before the model consistency estimation. Notice that, if the data for the constituent molecules in the models are actually observed, then this process is not necessary. First, the system of differential equations of (3.1)–(3.5) is solved over the time domain. For instance, the solution of the Model A is expressed as follows:

$$
\begin{cases}
SLA(t) = SLA(0)\ \exp(-k_{SN}\,t), \\
N(t) = (N(0) - \dfrac{SLA(0)\,k_{SN}}{k_{NA} - k_{SN}})\ \exp(-k_{NA}\,t) + \dfrac{SLA(0)\,k_{SN}}{k_{NA} - k_{SN}}\ \exp(-k_{SN}\,t), \\
A(t) = \dfrac{k_{NA}\,(k_{SN}\,SLA(0) - k_{NA}\,N(0) + N(0)\,k_{SN})}{(k_{NA} - k_{SN})\,(k_{NA} - k_{AG})}\ \exp(-k_{NA}\,t) \\
\qquad + \dfrac{k_{NA}\,k_{SN}\,SLA(0)}{(k_{AG} - k_{SN})\,(k_{NA} - k_{SN})}\ \exp(-k_{SN}\,t) \\
\qquad + (A(0) + \dfrac{k_{NA}\,(-k_{SN}\,SLA(0) + k_{AG}\,N(0) - N(0)\,k_{SN})}{(k_{AG} - k_{SN})\,(k_{NA} - k_{AG})})\ \exp(-k_{AG}\,t), \\
G(t) = \dfrac{k_{AG}\,(-k_{SN}\,SLA(0) + k_{NA}\,N(0) - N(0)\,k_{SN})}{(k_{NA} - k_{AG})\,(k_{NA} - k_{SN})}\ \exp(-k_{NA}\,t) \\
\qquad + \dfrac{k_{AG}\,(-k_{NA} + k_{AG})\,SLA(0)\,k_{NA}}{(k_{NA} - k_{AG})\,(k_{AG} - k_{SN})\,(k_{NA} - k_{SN})}\ \exp(-k_{SN}\,t) \\
\qquad + (-A(0) + \dfrac{(k_{SN}\,SLA(0) - k_{AG}\,N(0) + N(0)\,k_{SN})\,k_{NA}}{(k_{NA} - k_{AG})\,(k_{AG} - k_{SN})})\ \exp(-k_{AG}\,t) \\
\qquad + SLA(0) + N(0) + A(0) + G(0).
\end{cases}
\tag{3.9}
$$

In the above equations, we have no information about the actual values of the kinetic parameters and their initial values. Thus, we set them as follows: $k_{SN} = 1$, $k_{NA} = 0.1$, $k_{AG} = 0.5$, $k_{NG} = 0.2$, $k_{SA} = 0.4$, and $k_{SS} = 0.7$ for the kinetic parameters, and $SLA(0) = 10$, $N(0) = 7$, $A(0) = 3$, and $G(0) = 1$ for the initial values. By using the above values, the differential equations of (3.9) are simulated from $t = 0$ to 100 with intervals of 1. Then, we obtain the time series of data for each molecule at 101 sample points. We then numerically estimate the parameters by fitting the equations of (3.7) over the time domain to the above-generated data by the Maple 10 Global Optimization tool (©MapleSoft). In Fig. 4, the sampling data at 101 points and the corresponding equations (fitted curve) are plotted in Model A, together with the given and estimated parameters. Notice that, besides the estimation, all of the parameters in (3.7) can be exactly obtained from the given values for the kinetic parameters and the initial values in (3.9). In the present case, it is natural that the estimated values of the parameters are quite consistent with the given values of the parameters for generating the data.

Consistency Measure. As the first step for the model consistency estimation, we construct a set of polynomials, *comp*, from the algebraic equations of (3.6) for the

models and those of (3.8) for the sampling data. The following equations are *comp* for Model A:

$$
\begin{aligned}
comp = \{ & k_{SN} - \alpha_{SLA,1}, \\
& k_{NA}\, k_{SN} - \alpha_{N,1}\, \alpha_{N,2}, \\
& k_{SN} + k_{NA} - \alpha_{N,2} - \alpha_{N,1}, \\
& N(0)\, k_{SN} + k_{SN}\, SLA(0) - \beta_{N,1}\, \alpha_{N,2} - \beta_{N,2}\, \alpha_{N,1}, \\
& k_{AG} + k_{SN} + k_{NA} - \alpha_{A,1} - \alpha_{A,3} - \alpha_{A,2}, \\
& k_{AG}\, k_{SN} + k_{AG}\, k_{NA} + k_{NA}\, k_{SN} - \alpha_{A,1}\, \alpha_{A,3} - \alpha_{A,1}\, \alpha_{A,2} - \alpha_{A,2}\, \alpha_{A,3}, \\
& k_{NA}\, N(0) + A(0)\, k_{SN} + A(0)\, k_{NA} - \beta_{A,1}\, \alpha_{A,3} - \beta_{A,1}\, \alpha_{A,2} - \beta_{A,2}\, \alpha_{A,3} - \beta_{A,2}\, \alpha_{A,1} \\
& \quad -\beta_{A,3}\, \alpha_{A,2} - \beta_{A,3}\, \alpha_{A,1}, \\
& k_{NA}\, N(0)\, k_{SN} + k_{NA}\, k_{SN}\, SLA(0) + A(0)\, k_{NA}\, k_{SN} - \beta_{A,1}\, \alpha_{A,2}\, \alpha_{A,3} - \beta_{A,2}\, \alpha_{A,1}\, \alpha_{A,3} \\
& \quad -\beta_{A,3}\, \alpha_{A,1}\, \alpha_{A,2}, \\
& k_{AG} + k_{SN} + k_{NA} - \alpha_{G,2} - \alpha_{G,1} - \alpha_{G,3}, \\
& k_{AG}\, k_{NA}\, k_{SN} - \alpha_{A,1}\, \alpha_{A,2}\, \alpha_{A,3}, k_{AG}\, k_{NA}\, k_{SN} - \alpha_{G,1}\, \alpha_{G,2}\, \alpha_{G,3}, \\
& k_{AG}\, k_{SN} + k_{AG}\, k_{NA} + k_{NA}\, k_{SN} - \alpha_{G,1}\, \alpha_{G,2} - \alpha_{G,2}\, \alpha_{G,3} - \alpha_{G,1}\, \alpha_{G,3}, \\
& k_{AG}\, k_{NA}\, N(0)\, k_{SN} + k_{AG}\, k_{NA}\, k_{SN}\, SLA(0) + G(0)\, k_{AG}\, k_{NA}\, k_{SN} + k_{AG}\, A(0)\, k_{NA}\, k_{SN} \\
& \quad -\beta_{G,4}\, \alpha_{G,1}\, \alpha_{G,2}\, \alpha_{G,3}, \\
& G(0)\, k_{AG}\, k_{NA} + k_{AG}\, k_{NA}\, N(0) + G(0)\, k_{AG}\, k_{SN} + k_{AG}\, A(0)\, k_{NA} + k_{AG}\, A(0)\, k_{SN} \\
& \quad +G(0)\, k_{NA}\, k_{SN} - \beta_{G,3}\, \alpha_{G,1}\, \alpha_{G,2} - \beta_{G,1}\, \alpha_{G,2}\, \alpha_{G,3} - \beta_{G,2}\, \alpha_{G,1}\, \alpha_{G,3} \\
& \quad -\beta_{G,4}\, \alpha_{G,1}\, \alpha_{G,3} - \beta_{G,4}\, \alpha_{G,2}\, \alpha_{G,3} - \beta_{G,4}\, \alpha_{G,1}\, \alpha_{G,2}, \\
& G(0)\, k_{AG} + G(0)\, k_{NA} + k_{AG}\, A(0) + G(0)\, k_{SN} - \beta_{G,2}\, \alpha_{G,1} - \beta_{G,2}\, \alpha_{G,3} - \beta_{G,3}\, \alpha_{G,1} \\
& \quad -\beta_{G,3}\, \alpha_{G,2} - \beta_{G,1}\, \alpha_{G,3} - \beta_{G,4}\, \alpha_{G,1} - \beta_{G,4}\, \alpha_{G,2} - \beta_{G,4}\, \alpha_{G,3} - \beta_{G,1}\, \alpha_{G,2} \}.
\end{aligned}
$$

In the *comp*, the parameters and the initial values can be expressed as numerical values by the sample data fitting. Thus, only the set of kinetic parameters in the model remains as the unknown parameters in the *comp*. In the following section, we will estimate the kinetic parameters under the constraints in equations (2.4) and (2.6), and will select the model by considering the smallest value of $g(\vec{k})$, the sum-square value of the elements in *comp*.

Model Selection. The model selections by estimating the consistency of the models with the simulated data under the two constraints of equations (2.4) and (2.6) are shown in Table 1. In the first column, the query models, from which the simulated data are generated, are listed, and the models with consistencies that are estimated for the query model are listed in the second column. In the following column, the smallest values of the consistency measure are sorted in ascending order, and the corresponding kinetic measures are listed. As easily seen in this table, the present method has successfully identified the query models. Indeed, all of the models and four of the five models under the two constraints of (2.4) and (2.6) are correctly selected in Table 1, respectively. In addition to the successful selection, the characteristic features for the model selection are observed in the selections by the two constraints. The details of the features are as follows.

As for the selection under the constraint of (2.4), all of the models are clearly selected. By each query model, the corresponding models show the smallest consistency measure ($CM1$) in the constraint of (2.4). For example, when the query model is Model A, the corresponding value for the model consistency for Model A is 1.34×10^{-11}, which

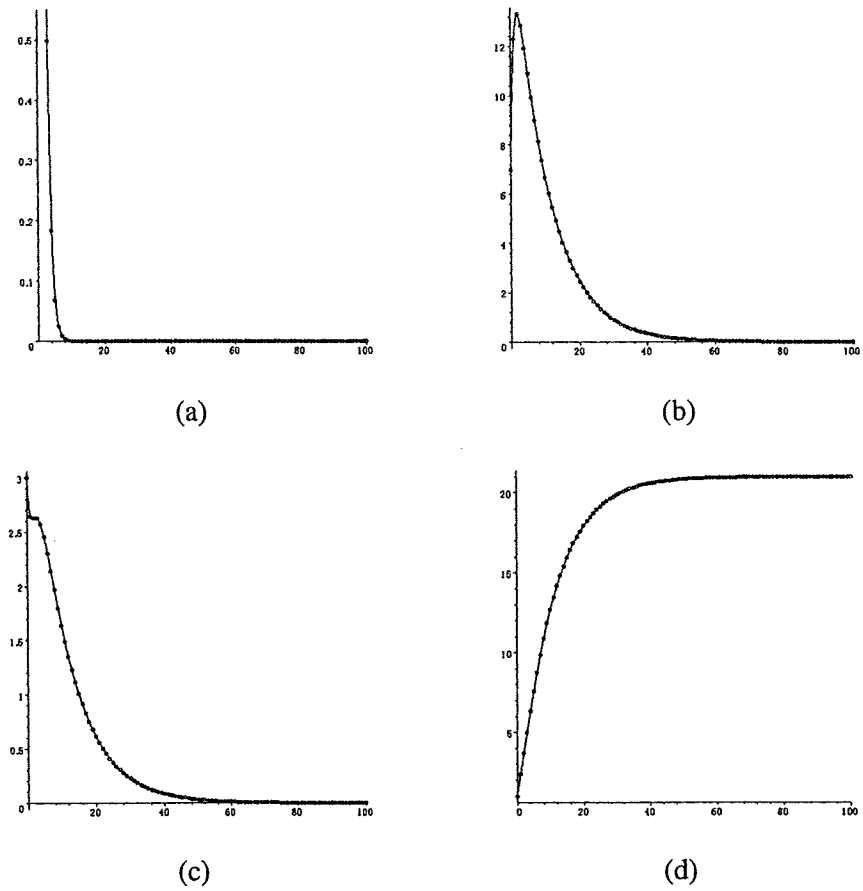

Fig. 4. Sample data for numerical fitting (circles), together with fitted curves (solid lines). The data were generated by numerical calculation from the differential equations (3.9), and the curves were fitted by commercial software (see details in the text). The given and estimated parameters are as follows: $\alpha_{SLA,1}$, 1 (given) and 1.00 (estimated); $\beta_{SLA,1}$, 10 and 10.0; $\alpha_{N,1}$, 1/10 and 0.100; $\alpha_{N,2}$, 1 and 1.00; $\beta_{N,1}$, 163/9 and 18.1; $\beta_{N,2}$, 100/9 and 11.1; $\alpha_{A,1}$, 1/10 and 0.100; $\alpha_{A,2}$, 1/2 and 0.500; $\alpha_{A,3}$, 1 and 1.00; $\beta_{A,1}$, 163/36 and 4.53; $\beta_{A,2}$, −15/4 and −3.75; $\beta_{A,3}$, 20/9 and 2.22; $\alpha_{G,1}$, 1/10 and 0.100; $\alpha_{G,2}$, 1/2 and 0.500; $\alpha_{G,3}$, 1 and 1.00; $\beta_{G,1}$, −815/36 and −22.6; $\beta_{G,2}$, 15/4 and 3.75; $\beta_{G,3}$, −10/9 and −1.11; $\beta_{G,4}$, 21 and 21.0. Each figure corresponds to the four variables (molecules) in the model: (a) SLA, (b) N, (c) A, (d) G.

is the smallest among the values of the five models. The magnitude is slightly smaller than 1.36×10^{-11} for Model E. Interestingly, the parameter value for k_{SS} in Model E is estimated to be nearly zero, 1.40×10^{-6}, and when k_{SS} is zero, Model E is identical to Model A. In the remaining models, the parameters cannot be estimated under the constraint of (2.4). In the other query models, the model corresponding to the query model shows the smallest values for the model consistency, and the remaining models show relatively large values or no values, due to the constraint of (2.4). In particular, Model E, in which a cyclic relationship is included, is successfully selected from the other models, especially Model A, which differs from Model E, only in the cyclic part. Furthermore, in all cases, the values of the kinetic parameters are estimated to be equal to the values that are set for the data generation. Thus, the model selection by using the constraint of (2.4) has completely succeeded in all of the models.

Four of the five models are successfully selected under the constraint of (2.6). In the model selection for Model A, Model C is selected. However, Models C, A, E, and B show small values for the consistency measure ($CM2$). Furthermore, three models, Models C, D, and B, become the same form as Model A, by considering the values of the kinetic parameters. Indeed, k_{NG} and k_{SA} in Model C and k_{SA} in Model B are estimated to be exactly zero values, and k_{SS} in Model E is estimated to be a very small value, 1.40×10^{-6}. A similar situation is also found when the query model is Model B. In this case, while the model showing the smallest value is Model B, a similar value is also found in Model C. However, the value of k_{NG} is estimated to be exactly zero, and this indicates that Model C, with the estimated values for kinetic parameters, is the same form as Model B. Thus, the constraint of (2.6) effectively excludes the false relationship between the molecules by estimating the values of the kinetic parameters. As for the model selection for Model E, the small value appears only in the query model, and the relatively large values appear in the other models. In the models with the large values, the $CM2$ values in Models A, B, and C are relatively smaller than the $CM2$ value in Model D. Interestingly, the former models share common chain relationships between SLA, N, A, and G with Model E, as seen in Fig 3, while the latter model is a distinctive form from Model E. Even in the inconsistent models, $CM2$ may reflect the similarity of the model form between the query and the estimated models. At any rate, the model selection under the constraint of (2.6) also has succeeded in all of the models.

In summary, the present model selection algorithm shows high performance under the constraints of both (2.4) and (2.6). The constraint of (2.4) focuses on only the selection of a model consistent with the data by a simple algorithm, and the constraint of (2.6) focuses on finer model selection, with the exclusion of false relationships, by a slightly and complicated algorithm. Thus, the algorithm with the constraint of (2.4) is useful to select a model consistent with the data among many candidate models, and that with the constraint of (2.6) is effective to select a model among the candidate models including similar forms.

3.3 Discussion

We have proposed a method for selecting a model that is the most consistent with the data in the present study. In small but distinctive networks, our algorithm has successfully selected the query model, from which the sampling data are generated. The present study partly exploits the previous studies of Cobelli et al. [5, 6] about the relationship between observational parameters and model parameters over the Laplace domain. In these studies, they dealt with the case of differential equations adjoined by a higher dimensional ideal to survey whether the model parameters themselves can be determined uniquely or non-uniquely. In our work, the combination of the transformation of equations over the Laplace domain with the numerical fitting to the observed data enables us to estimate the model's consistency with the data as well as with the values of the kinetic parameters. Although the robustness for data including noise should be further tested, our algorithm is expected to be feasible for actual biological issues regarding the selection of a kinetics model.

The scalability of the present algorithm also remains to be tested. Actually, the present model selection algorithm required several hours for one model. In addition,

Table 1. Model selections. The five models in Fig. 3 were examined for the model selection and the determination of kinetic parameters with the simulated data by the two constraints (see the details in the text). The 'query' and 'estimated' indicate the model from which the simulated data are generated, and the model the consistency of which is estimated by the corresponding query model, respectively.

Query	$CM1$, under the constraint (2.4)								$CM2$, under the constraint (2.6)							
	estimated	smallest	k_{SN}	k_{NA}	k_{AG}	k_{NG}	k_{SA}	k_{SS}	estimated	smallest	k_{SN}	k_{NA}	k_{AG}	k_{NG}	k_{SA}	k_{SS}
A	A	1.34×10^{-11}	1.00	0.100	0.500	-	-	-	C	7.66×10^{-12}	1.00	0.100	0.500	0*	0*	-
	E	1.36×10^{-11}	1.00	0.100	0.500	-	-	1.40×10^{-6}	A	1.34×10^{-11}	1.00	0.100	0.500	-	-	-
	D	×	×	-	×	-	×	-	E	1.36×10^{-11}	1.00	0.100	0.500	-	-	1.40×10^{-6}
	B	×	×	×	×	-	×	-	B	1.35×10^{-10}	1.00	0.100	0.500	-	0*	-
	C	×	×	×	×	×	×	-	D	1.68	0*	-	0.435	-	0.0439	-
B	B	4.20×10^{-11}	1.00	0.100	0.500	-	0.400	-	B	4.20×10^{-11}	1.00	0.100	0.500	-	0.400	-
	A	20.1	1.19	0.167	0.637	-	-	-	C	6.44×10^{-11}	1.00	0.100	0.500	0*	0.400	-
	D	×	×	-	×	-	×	-	E	20.1	1.19	0.167	0.637	-	-	0*
	E	×	×	×	×	-	-	×	A	20.1	1.19	0.167	0.637	-	-	-
	C	×	×	×	×	×	×	-	D	1050	0*	-	0.0351	-	1.97	-
C	C	2.78×10^{-9}	1.00	0.100	0.500	0.200	0.400	-	C	2.78×10^{-9}	1.00	0.100	0.500	0.200	0.400	-
	B	0.558	0.994	0.160	0.913	-	0.408	-	B	0.558	0.994	0.160	0.913	-	0.408	-
	A	28.0	1.19	0.213	1.17	-	-	-	D	23.9	0*	-	1.22	-	0.418	-
	D	×	×	-	×	-	×	-	E	28.0	1.19	0.213	1.17	-	-	0*
	E	×	×	×	×	-	-	×	A	28.0	1.19	0.213	1.17	-	-	-
D	D	1.83×10^{-14}	1.00	-	0.500	-	0.400	-	D	1.83×10^{-14}	1.00	-	0.500	-	0.400	-
	A	576	1.02	3.98	0.623	-	-	-	E	358.	1.13	3.63	0.285	-	-	0*
	E	×	×	×	×	-	-	×	B	399.	1.10	3.74	0.395	-	0*	-
	B	×	×	×	×	-	×	-	C	434.	1.18	3.43	0.454	0.528	0*	-
	C	×	×	×	×	×	×	-	A	576.	1.02	3.98	0.623	-	-	-
E	E	9.26×10^{-11}	1.00	0.100	0.500	-	-	0.700	E	9.26×10^{-11}	1.00	0.100	0.500	-	-	0.700
	A	1.46	0.702	0.0564	0.367	-	-	-	C	1.46	0.702	0.0564	0.367	0*	0*	-
	D	×	×	-	×	-	×	-	A	1.46	0.702	0.0564	0.367	-	-	-
	B	×	×	×	×	-	×	-	B	1.46	0.702	0.0564	0.367	-	0*	-
	C	×	×	×	×	×	×	-	D	2.57	0*	-	0.258	-	0.0284	-

0*: *exact* zero value.

−: no corresponding parameters.

×: no real positive solutions.

the limit of the nodes and edges in the tested network approximately ranged within 10 edges between 10 nodes. However, the present algorithm over the Laplace domain may overcome the issue of scalability. In a local network within a large-scale network, the relationships of the molecules in the local network with those outside of it are regarded as inputs from the outside, and the variables corresponding to the inputs may easily be eliminated, if the relationships are treated over the Laplace domain. Indeed, we have successfully eliminated the *unnecessary* variables to estimate the parameter values in complex compartmental models for Parkinson's disease by PET measurements [13]. If the *unnecessary* variables in the local network can be eliminated, then the present algorithm can be applied to estimate the model's consistency. Thus, the iteration of the elimination and the consistency estimation may be applicable for the consistency estimation, even in a large-scale network model. Further examinations of the present algorithm for a large-scale network and for noisy data will appear in the near future.

4 Conclusion

In the present model selection, an algebraic manipulation of the differential equations over the Laplace domain, formulated based on the assumption of linear relationships between the variables, is combined with the numerical fitting of the sampling data. The performance of our approach is illustrated with simulated data, in the distinctive forms of models, one of which includes a cyclic relationship hitherto unavailable in previous methods. Although some further examinations of the present method are necessary, especially of the analyzed data and its robustness with noise, the extension of our approach to a large-scale network is promising.

Acknowledgments

H. Y. and K. H. were partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (grant 18016008), by a Grant-in-Aid for Scientific Research (grant 19201039) and by a Grant-in-Aid for Young Scientists (B) (grant 19790881) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). This study was supported in part by the New Energy and Industrial Technology Development Organization (NEDO) of Japan, by Core Research for Evolutional Science and Technology (CREST), by Program for Improvement of Research Environment for Young Researchers from Special Coordination Funds for Promoting Science and Technology (SCF) commissioned by Japan Science and Technology Agency (JST,MEXT).

References

[1] Audoly, S., D'Angiò, L., Saccomani, M.P., Cobelli, C.: Global identifiability of linear compartmental models — A computer algebra algorithm. IEEE Trans. Biomed. Eng. 45, 36–47 (1998)

[2] Bisits, A.M., Smith, R., Mesiano, S., Yeo, G., Kwek, K., MacIntyre, D., Chan, E.C.: Inflammatory aetiology of human myometrial activation tested using directed graphs. PLoS Comput. Biol. 1, 132–136 (2005)

[3] Buchberger, B.: An Algorithmic Criterion for the Solvability of a System of Algebraic Equations. In: Buchberger, B., Winkler, F. (eds.) Gröbner Bases and Applications. London Mathematical Society Lecture Notes Series, vol. 251, pp. 535–545. Cambridge University Press, Cambridge (1998)

[4] Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G., Lowry, S.F.: Inflammation and Host Response to Injury Large Scale Collab. Res. Program: A network-based analysis of systemic inflammation in humans. Nature 437, 1032–1037 (2005)

[5] Cobelli, C., Foster, D., Toffolo, G.: Tracer Kinetics in Biomedical Research: From Data to Model. Kluwer Academic/Plenum Publishers (2000)

[6] Cobelli, C., Toffolo, G.: Theoretical aspects and practical strategies for the identification of unidentifiable compartmental systems. ch. 8, pp. 85–91. Pergamon Press, Oxford (1987)

[7] Hanzon, B., Jibetean, D.: Global minimization of a multivariate polynomial using matrix methods. Journal of Global Optimization 27, 1–23 (2003)

[8] Joreskog, K.G.: A general method for analysis of covariance structures. Biometrika 57, 239–251 (1970)

[9] Meziane, D., Shipley, B.: Direct and Indirect Relationships Between Specific Leaf Area, Leaf Nitrogen and Leaf Gas Exchange. Effects of Irradiance and Nutrient Supply. Annals of Botany 88, 915–927 (2001)

[10] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Francisco (1988)

[11] Shipley, B.: A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling 7, 206–218 (2000)

[12] Wright, S.: The method of path coefficients. Ann. Math. Statist. 5, 161–215 (1934)

[13] Yoshida, H., Nakagawa, K., Anai, H., Horimoto, K.: Exact parameter determination for Parkinson's disease diagnosis with PET using an algebraic approach. In: Anai, H., Horimoto, K., Kutsia, T. (eds.) Algebraic Biology 2007. LNCS, vol. 4545, pp. 110–124. Springer, Heidelberg (2007)

The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 288–301

Integer Programming-based Approach to Allocation of Reporter Genes for Cell Array Analysis

Morihiro Hayashida[1,*] Fuyan Sun[2]
Sachiyo Aburatani[2] Katsuhisa Horimoto[2]
Tatsuya Akutsu[1]

[1] Bioinformatics Center, Institute for Chemical Research,
Kyoto University, Gokasho, Uji, 611-0011, Japan
[2] Computational Biology Research Center, National Institute of Advanced
Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Abstract Observing behaviors of protein pathways and genetic networks under various environments in living cells is essential for unraveling disease and developing drugs. For that purpose, the biological experimental technique using transfected cell microarrays (cell arrays) has been developed. In order to apply cell arrays to identification of the subnetworks that are significantly activated or inactivated by external signals or environmental changes, it is useful to allocate several or several tens of reporter genes. In this paper, we consider the problem of selecting the most effective set of reporter genes.

We propose two graph theoretic formulations of the reporter gene allocation problem, and show that both problems are hard to approximate. We propose integer programming-based methods for solving practical instances of these problems optimally. We apply them to apoptosis pathway maps, and discuss biological significance of the result. We also apply them to artificial scale-free networks. The result shows that optimal solutions can be obtained within several seconds even for networks with 10,000 nodes.

Keywords integer programming; reporter gene; cell array; signaling network; set cover; NP-hard.

1 Introduction

Identification of novel target genes for the treatment of diseases is an important topic in drug design and systems biology. Because of its importance, various approaches have been proposed. Among these, *transfected cell microarrays* (*cell arrays* for short) are regarded as a potentially powerful approach [1, 2, 3, 4]. Cell arrays are complementary technique to DNA microarrays. The most important difference is that each spot in a DNA microarray corresponds to a gene, whereas each spot in a cell array corresponds to a cluster of several tens or hundreds of *living cells*. This property enables us to observe times series data of gene expression in living

*Corresponding Author. morihiro@kuicr.kyoto-u.ac.jp

cells. Furthermore, upon the addition of cells and a lipid transfection reagent, slides printed with cDNA become living microarrays, in which some specific gene is over-expressed. On the other hands, it is also possible to knock out some specific gene by using siRNA [1, 3]. Therefore, we may be able to observe effects of gene over-expression or gene knockdown by using cell arrays. We may also be able to observe effects of external signals on gene expressions in living cells. In order to observe the effects using cell arrays, we may need *reporter genes*, which are designed to measure the expression level of gene or the corresponding product through the magnitude of fluorescence. Over the past decade, a battery of powerful tools that encompass forward and reverse genetic approaches have been developed to dissect the molecular and cellular processes that regulate disease. In particular, the advent of genetically-encoded fluorescent proteins, together with advances in imaging technology, make it possible to study these biological processes in many dimensions [5]. Importantly, these technologies allow direct visual access to complex events as they happen in their native environment, which provides greater insights into human diseases than ever before [6, 7]. However, the cost (both in labor and money) of introduction of reporter genes to a cell is very high. Thus, we cannot use a lot of reporter genes. Instead, we should allocate several or several tens of reporter genes which are the most efficient for identifying the pathways that are significantly activated or inactivated by means of external signals or environmental changes.

There exist related studies. Several studies have been done for developing hypothesis generation techniques that use model checking and formal verification in order to qualitatively reason about signaling networks [8, 9, 10]. These techniques may be useful for computational analysis of effects of external signals and/or environmental changes. However, these techniques require statements about the property of individual reactions in networks, details of which are often unavailable. Ruths et al. recently proposed a framework for computational hypothesis testing in which signaling networks are represented as bipartite directed graphs [11]. In their framework, each network contains two types of nodes: nodes corresponding to molecules and nodes corresponding to reactions. They considered two problems: the constrained downstream problem and the minimum knockdown problem. The latter one is closely related to our problem and is to find a minimal set of nodes removal of which disconnects two given sets of compounds. They defined the minimum knockdown problem as a graph theoretic problem. They proved that the problem is NP-hard and proposed an iterative and randomized heuristic algorithm.

In this paper, we consider graph theoretic formulations of the reporter gene allocation problem. Since there is no consensus mathematical model of genetic networks or signaling pathways, we do not assume any specific models such as Boolean networks and Bayesian networks. Instead, we treat each network as a directed graph, where each edge can have a weight. Then, we formulate the reporter gene allocation problem as problems of selecting a set of nodes that covers as many nodes as possible, or selecting a minimal set of nodes that covers all the nodes in a network, where we say that node v is covered by node u if there exists a directed path from u to v within a

specified length. We prove that these problems are NP-hard. Furthermore, we prove that these problems are hard to approximate. We also show that some connection between these problems and the set cover problem (along with its variant). In order to solve realistic instances, we formulate these problems as integer programs (IPs) and apply a famous IP solver (CPLEX) to solving instances of these IPs. This approach is reasonable because a close relationship between integer programming and the set cover is known [12]. It should be noted that our approach is significantly different from that in [11]: (i) problems and network representations are different from each other, (ii) optimality of the solution is not guaranteed in [11], whereas optimality is guaranteed in our approach.

We perform computational experiments using both artificially generated networks and a real biological network. Though our IP formulations are simple, the results are quite surprising: the proposed method can find optimal solutions within several seconds even for networks with 10,000 nodes. Furthermore, the set of allocated reporters for a real network is reasonable from a biological viewpoint. These suggest that the proposed approach is practically useful for finding an optimal set of reporter genes.

2 Allocation Problems

In this section, we define two optimal allocation problems, P1 and P2. Biological networks such as gene regulatory networks and signaling pathways can be considered as a directed graph $G = (V,E)$ with a set of nodes $V = \{v_1,...,v_n\}$ and a set of directed edges from v_i to v_j, $(v_i,v_j) \in E$. In gene regulatory networks, a node means a gene, and in signaling pathways, a node means a protein. It should be noted that a reporter gene can be used both for measuring gene expression and for measuring abundance of proteins.

We define that a node v is a *neighboring upstream node* of a node v_r if there is a directed path within the length of a constant L from v to v_r in G. In this case, we also say that v is *covered* by v_r. For a set of nodes R, we say hat v is covered by R if v is covered by some node in R. This definition can be justified as follows: if some node v covered by v_r is affected by external signals and/or environmental changes, it is highly expected (for small L) that v_r is also be affected. That is, we may infer that a subnetwork around v_r is affected by external signal or environmental change if v_r is affected, and we want to cover as many parts of the network as possible.

We assume in this paper that L does not depend on the reporter node and each edge has unit length. This assumption is reasonable because it is difficult to determine L for each gene or protein and the length of each edge. However, the proposed methods can be modified for a general case in which L depends on the reporter node and each edge has distinct length (or weight). Figure 1 shows an example of covered nodes by using a reporter when $L = 2$.

Problem P1 maximizes the number of covered nodes by using K reporters, and is defined as follows.

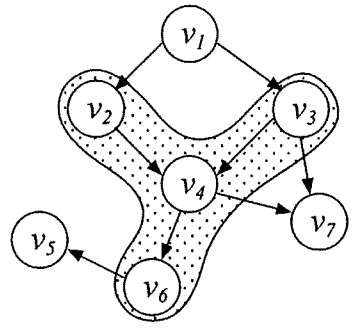

Figure 1: Example of nodes covered by a reporter node when $L = 2$ in a directed graph $G = (V,E)$ with $V = \{v_1,...,v_7\}$. In this case, v_2, v_3, v_4 and v_6 are covered by v_6.

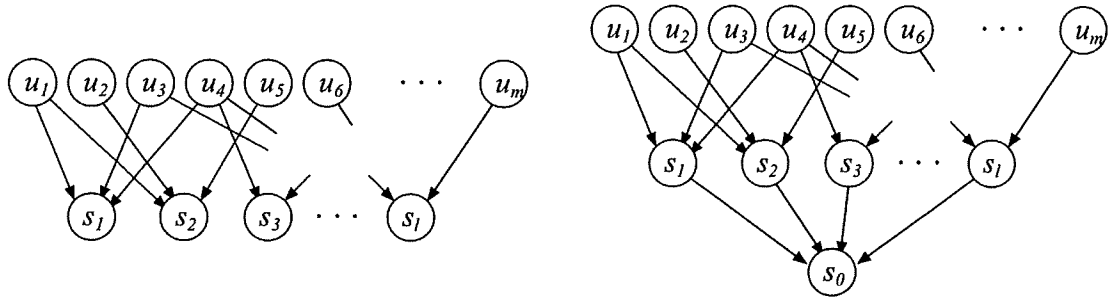

Figure 2: Left: Transformation of an instance $I = \langle U = \{u_1,...,u_m\}, S = \{s_1,...,s_l\}, k\rangle$ of the maximum coverage problem to Problem P1. Right: Transformation of $I = \langle U, S\rangle$ of the set cover problem to Problem P2.

Difinition 1[Problem P1] Given a directed graph $G = (V,E)$ and two integers L and K ($\leq |V|$), find a set $R \subseteq V$ of cardinality at most K maximizing the number of nodes covered by R.

It should be noted that R corresponds to a set of reporters. For sufficiently large K, we can cover all nodes of V using the solution of Problem P1. In some cases, we may want to cover all the nodes by using a minimum number of reporter nodes. Thus, we also consider the following problem.

Difinition 2[Problem P2] Given a directed graph $G = (V,E)$ and an integer L, find a minimum cardinality set $R \subseteq V$ such that all nodes of V are covered by R.

3 Theoretical Results

We show that Problem P1 is MAX SNP-hard, which means that no PTAS exists unless P=NP. It should be noted that MAX SNP-hardness also implies NP-hardness. For terminology on approximation algorithms, refer to [12].

Theorem 1. *Problem P1 is MAX SNP-hard.*

Proof. We show an L-reduction from the maximum coverage problem [12, 13], which is known to be MAX SNP-hard [14], to Problem P1. The maximum coverage

problem is defined as follows: Given a family of sets S over U, and an integer k, find $C \subseteq S$ of cardinality at most k which maximizes the number of covered elements in U. From an instance $I = \langle U = \{u_1, ..., u_m\}, S = \{s_1, ..., s_l\}, k(\leq l) \rangle$ of the maximum coverage problem, we construct an instance $I' = \langle G = (V, E), L, K \rangle$ of P1 in the following way (See Figure 2):

$$V = \{u_1, ..., u_m, s_1, ..., s_l\},$$
$$E = \bigcup_{j=1}^{l} \bigcup_{u_i \in s_j} \{(u_i, s_j)\},$$
$$L = 1,\ K = k.$$

It should be noted that $|V| = m + l, |E| = \sum_{j=1}^{l} |s_j|$. Thus, I' can be constructed in polynomial time.

Let $OPT(I)$ and $OPT(I')$ be optimal solutions of I and I', respectively. Then, $OPT(I') = OPT(I) + k$ holds. Without loss of generality, we can assume that $OPT(I) \geq k$. Therefore, $OPT(I') \leq 2OPT(I)$.

Given any solution $R \subseteq V$ of I' with cost (i.e., the number of covered nodes) c', we produce a solution C of I in polynomial time by letting $C = R - U$, where $R - U = \{r | r \in R \text{ and } r \notin U\}$. Then, $|C| \leq |R| \leq k$. Let c be the cost (i.e., the number of covered elements) of C. Since $c' \leq c + k$ holds,

$$OPT(I') - c' = OPT(I) + k - c' \geq OPT(I) - c.$$

Therefore, the above reduction is an L-reduction and thus Problem P1 is MAX SNP-hard. □

For Problem P2, we can show a much stronger hardness result as follows.

Theorem 2. *There is no polynomial time algorithm for Problem P2 with approximation ratio less than* $\frac{1-\delta}{4} \log n$ *for any constant* $0 < \delta < 1$ *unless* $NP \subseteq DTIME(n^{polylog(n)})$.

Proof. We prove the theorem by contradiction. Suppose that there is a polynomial time algorithm for Problem P2 with approximation ratio less than $\frac{1-\delta}{4} \log n$ for any constant $0 < \delta < 1$.

The set cover problem is defined as follows: Given a family of sets S over U, find a minimum cardinality set $C \subseteq S$ such that all elements of U are covered by $\bigcup_{s_i \in C} s_i$. From an instance $I = \langle U = \{u_1, ..., u_m\}, S = \{s_1, ..., s_l\} \rangle$ of the set cover problem, we construct an instance $I' = \langle G = (V, E), L \rangle$ of P2 in the following way (See Figure 2):

$$V = \{u_1, ..., u_m, s_1, ..., s_l, s_0\},$$
$$E = \bigcup_{j=1}^{l} \left(\{(s_j, s_0)\} \cup \bigcup_{u_i \in s_j} \{(u_i, s_j)\} \right),$$
$$L = 1,$$

where s_0 is a node not in S.

Let $OPT(I)$ and $OPT(I')$ be optimal solutions of I and I', respectively. Then, $OPT(I') = OPT(I) + 1$ holds.

Given any solution $R \subseteq V$ of I' with cost c' (i.e., the number of selected nodes), we produce a solution C of I in polynomial time by letting $C = (R - U - \{s_0\}) \cup \{s_j |$ for $u_i \in R - S - \{s_0\}, u_i \in \exists s_j\}$. Let c be the cost (i.e., the number of selected elements) of C. Since $c = |C| \leq |R| = c'$ holds,

$$\frac{c}{OPT(I)} = \frac{c}{OPT(I') - 1} \leq \frac{c'}{OPT(I') - 1}.$$

For any constant $0 < \delta < 1$,

$$\frac{c'}{OPT(I') - 1} \leq \frac{1}{1 - \delta} \frac{c'}{OPT(I')} < \frac{1}{4} \log n$$

holds for sufficient large $n = m + l + 1$. Therefore,

$$\frac{c}{OPT(I)} < \frac{1}{4} \log n.$$

This contradicts to the fact that there is no polynomial time algorithm for the set cover problem with approximation ratio less than $\frac{1}{4} \log n$ unless $NP \subseteq DTIME(n^{polylog(n)})$. Thus, the theorem is proved. □

We can also show positive results on approximation ratios using a well-known greedy algorithm for the set cover [12]. For that purpose, we let $U = V$ and $S = \{s_v | s_v$ is the set of nodes covered by $v \in V\}$, and simply apply the greedy algorithm. Then, the following propositions are directly obtained from the results on the greedy algorithm [12, 13, 14].

Proposition 3. *P1 can be approximated within a factor of* $e/(e-1)$.

Proposition 4. *P2 can be approximated within a factor of* $O(\log n)$.

4 Integer Programming Formulation

In this section, we propose methods to solve Problem P1 and P2 using integer programming. In the previous section, we showed that both Problem P1 and P2 are very hard to find optimal or approximate solutions. However, efficient algorithms such as branch-and-bound methods have been developed for *integer programming*, which is also NP-hard. Therefore, we formulate Problem P1 and P2 as integer programs, and call IP1 and IP2 respectively. In the next section, we show that IP1 and IP2 are solved in practical time through computational experiments.

Problem P1 is formulated as follows.

(IP1) Maximize $\sum_{i=1}^{n} y_i,$

$$
\begin{aligned}
&\text{Subject to} \\
&y_i \leq \sum_{j \in S_i^L} x_j \quad \text{for } i = 1, .., n, \\
&\sum_{i=1}^{n} x_i \leq K, \\
&x_i = \{0, 1\}, \\
&y_i = \{0, 1\},
\end{aligned}
$$

where S_i^L is the set of nodes covered by v_i. Thus, for $j \in S_i^L$, the length of a directed path from the node v_i to v_j is less than or equal to L. $x_i = 1$ if v_i is selected as a reporter, otherwise $x_i = 0$. $y_i = 1$ if v_i is covered by some reporter, otherwise $y_i = 0$. IP1 maximizes the number of covered nodes using at most K reporter nodes.

Similarly, Problem P2 is formulated as follows.

$$
\begin{aligned}
&\text{(IP2)} \quad \text{Minimize} \quad \sum_{i=1}^{n} x_i, \\
&\text{Subject to} \\
&\sum_{j \in S_i^L} x_j \geq 1 \quad \text{for } i = 1, .., n, \\
&x_i = \{0, 1\}.
\end{aligned}
$$

IP2 minimizes the number of reporters such that all nodes are covered. If the parameter K of IP1 is greater than or equal to the optimal solution of IP2, the optimal solution of IP1 is always n.

5 Computational Experiments

We applied the proposed methods to two kinds of data, apoptosis pathway maps as a real network and artificial scale-free networks for validating the practicality of our methods in large networks.

All of these computational experiments were done on a PC with a Xeon 5160 3GHz CPU and 8GB RAM running under the Linux (version 2.6.19) operating system. We used ILOG CPLEX (version 10.1)[15] for solving IP1 and IP2, and measured execution time of the optimization function CPXmipopt() for mixed integer programming problems in CPLEX. We must calculate S_i^L for all i in order to give integer programming problems to the function. However, the preparation takes at most $O(n^2)$ time.

5.1 Apoptosis Pathway Maps

We used apoptosis pathway maps in a HeLa cell (See Figure 3). The maps are composed of major signal pathways of apoptosis, which are initiated by TRAIL (tumour necrosis factor apoptosis inducing ligand) ligation [16]. The maps were constructed by a commercial software, MetaCore (GeneGo Corp.) [17], in which

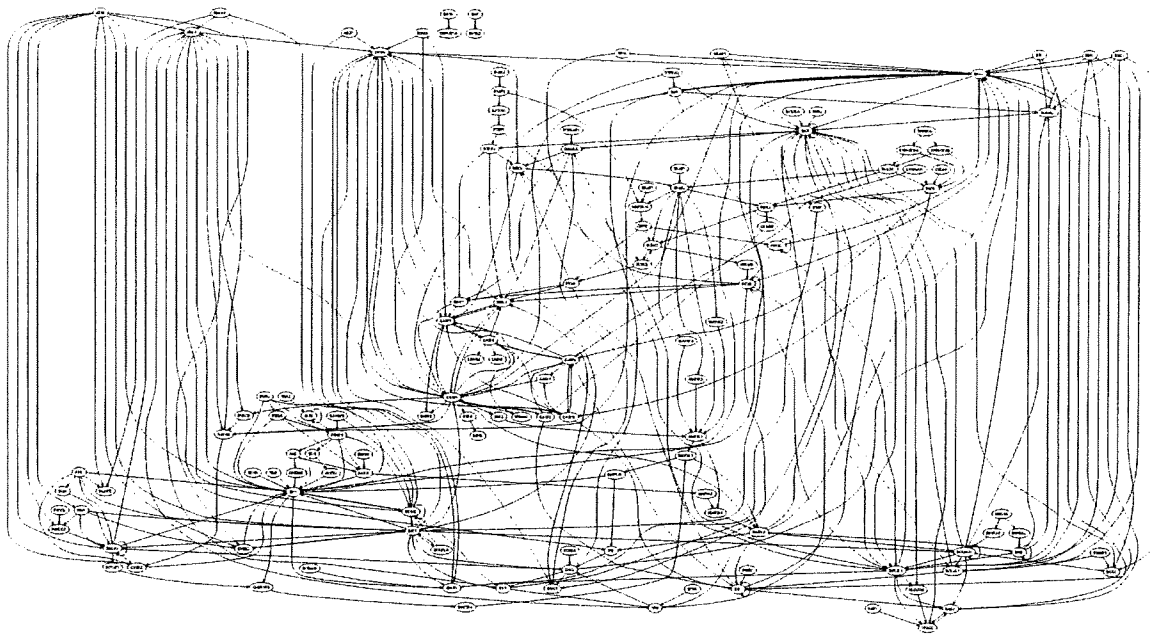

Figure 3: Apoptosis pathway maps in a HeLa cell, which contain 132 proteins and 337 binomial relations.

findings presented in peer-reviewed scientific publications were systematically encoded into an ontology by content and modelling experts, and a molecular network of direct physical, transcriptional and enzymatic interactions was computed from this knowledge base. The maps thus constructed contain 132 proteins and 337 binomial relations.

Table 1 shows the results on the optimal solution of IP1 and IP2 for each $L(= 1,...,6,132)$ and $K(= 1,...,6)$. The solution of IP2 for each L gives the required number of nodes to cover all nodes of V. For example, 42 reporters are required for $L = 1$, and 9 reporters for $L = 6$.

In the case that L is equal to the number of nodes $n = 132$, a node v_i is always covered by another v_j if there is a directed path from v_i to v_j. Since 121 proteins among 132 proteins are covered by protein BAK1 in the case of both $L = 6$ and $L = 132$, we can see that the distance between almost all pairs of proteins in this network is at most 12. Thus, it is considered that the network also has a small-world property [18]. It should be noted that most nodes (126 nodes) are covered by 6 reporters in the case of $L = 6$. It is also observed that 104 nodes are covered by 6 reporters even in the case of $L = 2$. For $L = 1,...,3$, TP53, BCL2 and BAX were selected as the most significant reporters respectively. These proteins are considered as hubs of the network because they have large indegrees and outdegrees. On the other hand, BAK1 is not considered as a hub, but is as an accumulation node of the network, and is selected as a reporter. Moreover, it seems that some of the selected proteins have significant biological meanings as follows. p53, a tumour suppressor gene that responds to DNA-damage, is influential on TRAIL-induced apoptosis by up-regulating TRAIL receptor [19]. Bcl-2 superfamily regulates cell death that is

Table 1: The optimal solution of IP1 and IP2 for each L and K in apoptosis pathway maps, where the numbers of covered nodes and the numbers of the selected reporters are shown for IP1 and IP2, respectively.

L	IP1 for each K						IP2	Reporter in $K=1$ (indegree/outdegree)
	1	2	3	4	5	6		
1	20	36	47	56	62	68	42	TP53 (19/5)
2	60	76	85	92	98	104	22	BCL2 (17/4)
3	88	103	110	116	118	120	15	BAX (16/6)
4	109	116	120	122	124	126	12	BAX (16/6)
5	118	121	123	125	127	128	10	BAK1 (6/1)
6	121	123	125	127	128	129	9	BAK1 (6/1)
132	121	123	125	127	128	129	9	BAK1 (6/1)

amplified via the mitochondrial pathway [20]. BAX may be related with possible amplification of apoptosis via the intrinsic pathway in response to JNK. The caspase-9 may be essential for border-cell migration in the Drosophila ovary [21], and the regulation of cell migration may also point to a roll in the cleavage of several adhesion- and cell motility- related proteins during mammalian apoptosis [22].

Table 2 shows the selected proteins as reporters for each L and K. The protein selected as a reporter for smaller K was not always selected for larger K. For example, for $L=2$, BCL2 was selected as a reporter in the case of $K=1$, but was not in the cases of $K=2,...,4$. If we use a simple greedy algorithm for solving P1, we may not be able to find CASP9 and BAX for $K=2$, or CASP9, BAX and IKBKG for $K=3$ since the greedy algorithm often tends to add a new node to the solution for $K-1$. On the other hand, our integer programming-based methods can always find optimal solutions if any. For each case, the elapsed time of optimizing IP1 or IP2 was at most 0.023 seconds. These results suggest that our methods are practical.

5.1.1 Effects of Specific Nodes

It is also important to observe the effects of signals on specific proteins or genes using cell arrays. In this section, we used CASP8, which is a protease located at the upstream of the caspase cascade that is a main pathway of the apoptosis initiated by TRAIL [23], as a specific protein among the apoptosis pathway maps. Then, we extracted the downstream proteins within the distance 2 from CASP8 (See Figure 4). We excluded CASP8 from this downstream subnetwork not to select it as a reporter. Thus, we obtained the subnetwork with 23 proteins and 58 binomial relations excluding CASP8.

Table 3 shows selected proteins as reporters for each L and K as Table 2. In both the whole network and the subnetwork, the same proteins such as BCL2, BAK1 and CASP9 were selected as reporters. It is reasonable because they have similar connections in both networks. For $L=4,...,n(=23)$, five proteins without outward

Table 2: Selected proteins as reporters for each L and K in apoptosis pathway maps.

L	K	IP1	Reporters
1	1	20	TP53
1	2	36	TP53, BCL2
1	3	47	TP53, BCL2, BAX
1	4	56	TP53, BCL2, BAX, CASP9
1	5	62	TP53, BCL2, BAX, CASP9, FADD
1	6	68	TP53, BCL2, BAX, CASP9, FADD, MAP3K1
1	7	73	TP53, BCL2, BAX, CASP9, FADD, MAP3K1, BIRC4
2	1	60	BCL2
2	2	76	CASP9, BAX
2	3	85	CASP9, BAX, IKBKG
2	4	92	CASP9, BAX, IKBKG, MAP2K7
2	5	98	CASP9, IKBKG, MAP2K7, BCL2, VDAC2
2	6	104	CASP9, IKBKG, MAP2K7, BCL2, VDAC2, TP53
3	1	88	BAX
3	2	103	BAX, IKBKG
3	3	110	IKBKG, BCL2, VDAC2
3	4	116	IKBKG, BCL2, BAK1, MAP2K7
3	5	118	IKBKG, BAK1, MAP2K7, CASP9, TP53
4	1	109	BAX
4	2	116	BCL2, BAK1
4	3	120	BAX, VDAC2, IKBKG
4	4	122	BAX, VDAC2, IKBKG, FASLG
5	1	118	BAK1
5	2	121	BAK1, BCL2
5	3	123	BCL2, VDAC2, TNFRSF1A
5	4	125	BCL2, VDAC2, TNFRSF1A, DFFB
6	1	121	BAK1
6	2	123	BAK1, FASLG
6	3	125	BAK1, FASLG, TNFRSF1A
132	1	121	BAK1
132	2	123	BAK1, TNFRSF1A
132	3	125	BAK1, TNFRSF1A, FASLG