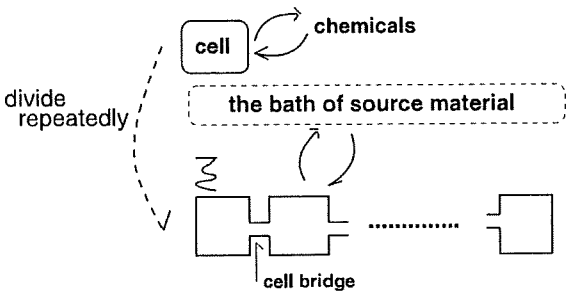


FIGURE 1. Schematic representation of our previous model. The cells are surrounded by a bath of source material with a constant concentration. After a division, the cells are connected to one another by forming cell bridges. Thus, the cells are connected to one another as a one-dimensional chain.

computation using QE are given in Section 4, which describes algebraic relations between the proliferation and transition rates. In Section 4.2 and 4.3, these relations are used to show that the cell-type order conservation rule plays a key role in high cell-type diversity.

2. Overview of previous model

In this section, we give a brief overview of our previous work [22], which is the basis of the construction and analysis of the model in this paper.

In a multi-cellular organism, a single cell – an egg – correctly develops into a prospectively determined pattern. This morphogenesis is robust against environmental perturbations, and the same pattern is always generated from eggs of a particular species. In other words, recursive production is repeated. At the same time, the developmental process in a multi-cellular organism produces a variety of cell types. The compatibility of these two points is surprising, because 'recursive production' is the reproduction of the same pattern of an individual cell, while 'cell-type diversity' is the existence of various patterns, namely, various cell types, within an individual. The question we addressed in our previous work was the selection of initial cells to allow for compatibility between recursive production and cell-type diversity.

We present our previously developed model of a multi-cellular organism in Figure 1. Within each cell, catalytic and auto-catalytic chemical reactions maintain the cell itself and synthesize some chemicals for the cell membrane. Our numerical results indicated that, by starting with an initial object consisting of both the chaotic cell type with diverse chemicals and the regular-dynamics cell type with less chemical diversity, the recursive production of a multi-cellular organism with cell-type diversity has been realized. As illustrated in Figure 2, starting with the two cells corresponding to I_1 and I_n, the regeneration pattern corresponding

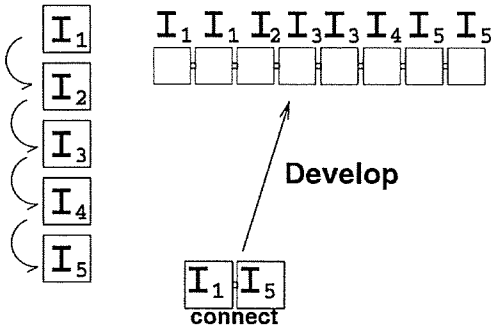


FIGURE 2. Regeneration of cell-type sequence, which was observed in [22]. The cell differentiates from I_1 to I_5 sequentially. Starting with $I_1 I_5$, patterns without non-contiguous numbers, such as $I_1 I_1 I_2 I_3 I_3 I_4 I_5 I_5$, are eventually produced. Thus, no contiguity will disappear during the development process.

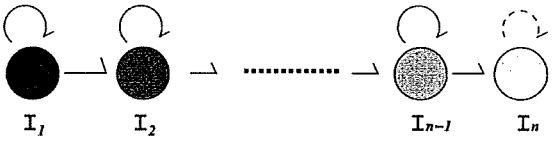


FIGURE 3. Schematic representation of our model. Cell differentiation proceeds as follows: $I_1 \to I_2 \to \cdots \to I_n$.

to $I_1 I_2 \ldots I_n$ is eventually produced. Here, such regeneration phenomena can be described as the following rewriting rule, named a *cell-type order conservation rule*:

$$I_i I_j \to I_i I_{i+1} \cdots I_{j-1} I_j\,, \quad I_j I_i \to I_j I_{j-1} \cdots I_{i+1} I_i \quad (j > i+1)\,. \tag{2.1}$$

This rewriting rule appears as interaction terms in the IL-system in the next section.

3. Present model

In this section, we present a simple model of a multi-cellular organism in which the cell lineage can be represented as a line, that is, only sequential differentiation occurs. Our model is schematically illustrated in Figure 3. We assume that cell differentiation starts with an initial type, I_1, and then the cell differentiates into several intermediate types $I_2 \to I_3 \to \cdots \to I_{n-1}$ before differentiating into the final type, I_n. The proliferation and transition rates of cell type i $(1 \le i \le n)$ are

defined as follows:

$$I_i \rightarrow \begin{cases} I_iI_i & p_{i,i} \\ I_{i+1} & p_{i,i+1} \\ I_i & 1-p_{i,i}-p_{i,i+1} \end{cases} \quad (1 \leq i < n),$$

$$I_n \rightarrow \begin{cases} I_nI_n & p_{n,n} \\ I_n & 1-p_{n,n} \end{cases} \tag{3.1}$$

with $0 \leq p_{i,i} < 1$ $(1 \leq i \leq n)$, $0 < p_{i,i+1} < 1$ $(1 \leq i < n)$, $p_{i,i}+p_{i,i+1} < 1$ $(1 \leq i < n)$. After the rewriting rules above are once applied, the following rewriting rules are once applied: $I_iI_j \rightarrow I_iI_{i+1}\cdots I_{j-1}I_j$, $I_jI_i \rightarrow I_jI_{j-1}\cdots I_{i+1}I_i$ $(j > i+1)$, which describe interactions between cells. We repeat this manipulation. These rules are termed cell-type order conservation rule as mentioned in Section 2.

4. Results and discussion

4.1. Analysis of the growth matrix in the stochastic l-system

Now, we calculate the *growth matrix* M of the two contiguous cell types I_iI_i, I_iI_{i+1}, $I_{i+1}I_i$ $(1 \leq i < n)$, which enables us to estimate the expected composition of $I_\ell I_k$ $(k = \ell-1, \ell, \ell+1)$ at step m. It should be noted that the other two contiguous cell types (e.g., I_iI_{i+3}) never appear at any step according to the cell-type order conservation rule. Although we could use the growth matrix concerning more than two cell types, the simple growth matrix with the two contiguous cell types suffices in this work.

If we start with I_1I_1, then the composition at step m can be calculated by the following formula:

$$(1,0,0,\ldots)M^m. \tag{4.1}$$

Here, we have studied the case with $n = 3$, showing the existence of three cell types. Let A, B and C denote I_1, I_2 and I_3, respectively, in the following. The growth matrix M is then:

1st to 4th column:

$$M_3 = \begin{pmatrix} 2p_{1,1}+(1-p_{1,2})^2 & (1-p_{1,2})p_{1,2} & (1-p_{1,2})p_{1,2} & p_{1,2}^2 \\ p_{1,1} & 1-p_{1,2} & 0 & p_{1,2}+p_{2,2}-p_{1,2}p_{2,3} \\ p_{1,1} & 0 & 1-p_{1,2} & p_{1,2}+p_{2,2}-p_{1,2}p_{2,3} \\ 0 & 0 & 0 & 2p_{2,2}+(1-p_{2,3})^2 \\ 0 & 0 & 0 & p_{2,2} \\ 0 & 0 & 0 & p_{2,2} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

5th to 7th column:

$$\begin{pmatrix} 0 & 0 & 0 \\ p_{2,3} & 0 & 0 \\ 0 & p_{2,3} & 0 \\ (1-p_{2,3})p_{2,3} & (1-p_{2,3})p_{2,3} & p_{2,3}^2 \\ 1-p_{2,3} & 0 & p_{2,3}+p_{3,3} \\ 0 & 1-p_{2,3} & p_{2,3}+p_{3,3} \\ 0 & 0 & 1+2p_{3,3} \end{pmatrix}$$

with its eigenvalues:

$$1-p_{1,2},\, 1+2p_{1,1}-p_{1,2},\, (1-p_{1,2})^2,\, 1-p_{2,3},$$
$$1+2p_{2,2}-p_{2,3},\, (1-p_{2,3})^2 \quad \text{and} \quad 1+2p_{3,3}\,. \tag{4.2}$$

Let S denote the diagonal matrix: $Diag(1-p_{1,2}, 1+2p_{1,1}-p_{1,2}, (1-p_{1,2})^2, 1-p_{2,3}, 1+2p_{2,2}-p_{2,3}, (1-p_{2,3})^2, 1+2p_{3,3})$. The features of the growth matrix M are the followings:

If the eigenvalues differ from one another, then there exists a regular matrix P such that $M = PSP^{-1}$. In this case, $(1,0,0,0,0,0,0)P$ is:

$$(0, 2, e_3, 0, e_5, e_6, e_7)\,,$$

where e_3, e_5, e_6 and e_7 are nonzero values. These facts lead the composition (4.1): $(1,0,0,0,0,0,0)PS^mP^{-1}$ to

$$\big(0, 2(1+2p_{1,1}-p_{1,2})^m, e_3(1-p_{1,2})^{2m}, 0, e_5(1+2p_{2,2}-p_{2,3})^m\,,$$
$$e_6(1-p_{2,3})^{2m}, e_7(1+2p_{3,3})^m\big)P^{-1}\,.$$

As m approaches infinity, because only $1+2p_{1,1}-p_{1,2}, 1+2p_{2,2}-p_{2,3}$ and $1+2p_{3,3}$ can be greater than 1, the composition above can be described as follows:

$$\big(0, 2(1+2p_{1,1}-p_{1,2})^m, 0, 0, e_5(1+2p_{2,2}-p_{2,3})^m, 0, e_7(1+2p_{3,3})^m\big)P^{-1}\,.$$

The fifth and seventh rows of P^{-1}, corresponding to the fifth and seventh columns: $e_5(1+2p_{2,2}-p_{2,3})^m$ and $e_7(1+2p_{3,3})^m$ in the first row vector, have zero elements at the AA, AB, BA and AA, AB, BA, BB, BC, CB columns, respectively. Therefore, only the second column: $2(1+2p_{1,1}-p_{1,2})^m$ can give rise to AA, AB, and BA as m approaches infinity. This indicates that one of the necessary conditions for AA, AB, BA, BB, BC, CB and CC to be well mingled as m approaches infinity is:

$$1+2p_{1,1}-p_{1,2} > 1 \;\wedge\; 1+2p_{1,1}-p_{1,2} > 1+2p_{2,2}-p_{2,3}$$
$$\wedge\; 1+2p_{1,1}-p_{1,2} > 1+2p_{3,3}\,. \tag{4.3}$$

In addition, for real biological systems, the following constraints are assumed:

$$\gamma N(AA) = N(BB) = N(CC) \;\wedge\; \gamma N(AB) = N(BC)\,. \tag{4.4}$$

Under the condition (4.4), let m approach infinity, and the following equations are derived:

$$N(AB) = N(BA) = \frac{\gamma(p_{1,2}-p_{2,3})(1-p_{1,2}-p_{2,3})}{\gamma(p_{1,2}-p_{2,3})+p_{2,3}}\,,$$
$$N(BC) = N(CB) = \gamma N(AB)\,,$$
$$N(BB) = N(CC) = \gamma\,,$$

$$p_{1,1} = p_{1,2}(1-p_{1,2})\big(p_{2,3}+\gamma(p_{1,2}-p_{2,3})\big)/\big(2\gamma(p_{1,2}-p_{2,3})(1-p_{1,2}-p_{2,3})\big)\,,$$

$$p_{2,2} = \Big(p_{1,2}p_{2,3}\big(-(1-p_{1,2})p_{1,2}^2+p_{2,3}-p_{1,2}p_{2,3}^2\big)+\big(p_{1,2}^4(3-5p_{2,3})$$

$$
\begin{aligned}
&- (2 - p_{2,3})(1 - p_{2,3})p_{2,3}^3 - p_{1,2}^5(1 - 2p_{2,3}) + p_{1,2}p_{2,3}^2\big(- 1 + 2(2 \\
&- p_{2,3})(1 - p_{2,3})p_{2,3}\big) + p_{1,2}^2p_{2,3}(5 - 9p_{2,3} + 6p_{2,3}^2) - p_{1,2}^3(2 + p_{2,3} \\
&- 7p_{2,3}^2 + 4p_{2,3}^3)\Big)\gamma + (p_{1,2} - p_{2,3})^2(1 - p_{2,3})p_{2,3}(2 - p_{1,2} \\
&- p_{2,3})\gamma^2\Big)\Big/\Big(2(p_{1,2} - p_{2,3})(-1 + p_{1,2} + p_{2,3})\gamma\big((-1 + p_{1,2})p_{1,2} \\
&- p_{2,3}^2 + (p_{1,2} - p_{2,3})(-2 + p_{1,2} + p_{2,3})\gamma\big)\Big), \\
p_{3,3} = &\Big(p_{2,3}\Big((1 - p_{1,2})p_{1,2}p_{2,3} - (p_{1,2} - p_{2,3})(p_{1,2}^2 + (1 - p_{2,3})p_{2,3}^2 \\
&- p_{1,2}(1 + p_{2,3} + p_{2,3}^2))\gamma - (p_{1,2} - p_{2,3})^2(2 - p_{1,2} \\
&- p_{2,3})(1 - 2p_{1,2} + p_{2,3})\gamma^2\Big)\Big)\Big/\Big(2(p_{1,2} - p_{2,3})(-1 + p_{1,2} \\
&+ p_{2,3})\gamma\Big((-2 + p_{1,2})p_{1,2}\gamma - p_{2,3}\big(1 - (2 - p_{2,3})\gamma\big)\Big)\Big), \qquad (4.5)
\end{aligned}
$$

where $N(XY)$ denotes the number of sequences XY as m approaches infinity and γ denotes that the ratio of the initial cells to the developed cells is $1/\gamma$. Notice that $N(AB) = N(BA)$ and $N(BC) = N(CB)$ always hold true because of the construction of the rewriting rules (3.1). In the equations above (4.5), $N(AA)$ is normalized, i.e., $N(AA) = 1$. Thus, $N(XY)$, $(X, Y \in \{A, B, C\})$, $p_{1,1}$, $p_{2,2}$ and $p_{3,3}$ can explicitly be represented as functions of $p_{1,2}$ and $p_{2,3}$. Notice that as $N(AB) = (N(BA))$ approaches 1, the cell-type diversity approaches its maximum.

4.2. Inference of the proliferation and transition rates by QE

Now, let us infer relations between the proliferation and transition rates for which the cell-type diversity is high under the constraints: (4.3), (4.4) and (4.5). For this purpose, it is sufficient to calculate the relations that maximize $N(AB)$ under the constraints of (4.3), (4.5) and (4.4) because of $N(AA) = 1$, $N(AB) = N(BA)$, $N(BC) = N(CB)$ and the constraint (4.4).

It may be worth noting that it seems difficult to calculate relations between the rates under such complicated constraints by numerical methods. Indeed, in our previous analysis by brute-force numerical simulations [16], we estimated a set of rates that realize high cell-type diversity by searching a large number of points, but could not obtain definite relations between the rates. In [10], Janssen and Lindenmayer described some plant as an IL-system and investigated the development of highly branched inflorescences using various developmental parameter sets by numerical analysis. They succeeded in producing a proper model of the plant, but did not obtain definite relations between the parameters. Although the rate values provide a snapshot for the system behaviour, the relation between rates will provide more profound insights into the mechanism of the system. Therefore, in this paper, we have utilized the QE approach to obtain algebraic relations between rates.

Firstly, we determine the maximum values of $N(AB)$ by solving the following QE problem:

$$^{\exists}p_{1,2}{}^{\exists}p_{2,3}\left(\psi(p_{1,2},p_{2,3},\gamma)\wedge N(AB)\geq\epsilon\right), \tag{4.6}$$

where $\psi(p_{1,2},p_{2,3},\gamma)$ is a formula derived by combining all equations and inequalities appearing in (4.3), (4.4) and (4.5), conjunctively. For a fixed value of γ, the QE procedure (4.6) produces the following inequalities: $\epsilon\leq(\sqrt{17}+1)/8\sim 0.64039$, $(\sqrt{881}-9)/40\sim 0.517041$ and $(\sqrt{89801}-99)/400\sim 0.50167$ when the γ values are 1, 10 and 100, respectively. Thus, we have determined the maximum values. To summarize, we have obtained the composition describing the highest cell-type diversity:

$$(AA,AB,BA,BB,BC,CB,CC)=\left(1,f(\gamma),f(\gamma),\gamma,\gamma f(\gamma),\gamma f(\gamma),\gamma\right), \tag{4.7}$$

with $f(1)=(\sqrt{17}+1)/8$, $f(10)=(\sqrt{881}-9)/40$ and $f(100)=(\sqrt{89801}-99)/400$. By QE method, we have also successfully derived the algebraic equation between the maximum value: h $(0<f<1)$ and γ $(\gamma>0)$ as follows:

$$\begin{cases} h^2\gamma^2-\gamma^2-2h^3\gamma+3h^2\gamma-h\gamma+2h^3=0 & (8\gamma^3-11\gamma^2+3\gamma-1\leq 0)\,, \\ 2h^2\gamma+h\gamma-\gamma-h=0 & (8\gamma^3-11\gamma^2+3\gamma-1>0)\,. \end{cases} \tag{4.8}$$

The relations between rates which maximize the diversity can be derived using formula (4.5). The rate-relations for $\gamma=1$, 10 and 100 are illustrated in Figure 4. These values were chosen because in our previous simulation [22], the constraint (4.4) over $N(XY)$, $(X,Y\in\{A,B,C\})$ was observed, and partly because there are few initial-type cells (A in this work) corresponding to stem cells in real biological systems [8]. Remember that $1/\gamma$ was defined as the ratio of the initial cells to the developed cells. Interestingly, three modes for the highest diversity of cell types emerge. The three modes correspond to three curves separated by discontinuous points of the derivative. The three modes when γ is 100 are explicitly expressed as follows:

- Mode I:

$$\begin{aligned} p_{2,3}=&\text{ the minimum real root of the equation in } x,\\ &19900p_{1,2}^2-4900p_{1,2}^3+20000p_{1,2}^4+(-39901p_{1,2}+69801p_{1,2}^2\\ &-10000p_{1,2}^3)x+(20000+10200p_{1,2}-30100p_{1,2}^2)x^2+(-30100\\ &+10000p_{1,2})x^3+10100x^4=0\quad(0<p_{1,2}<p_0)\,, \end{aligned}$$

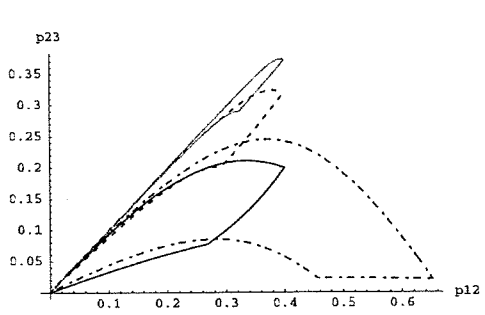
where p_0 is the minimum real root of the equation in x,

$$39999-320794x+883988x^2-966392x^3+363200x^4=0\,,$$

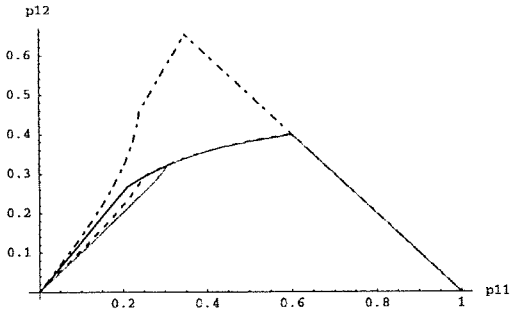
and is approximately 0.321746.

- Mode II:

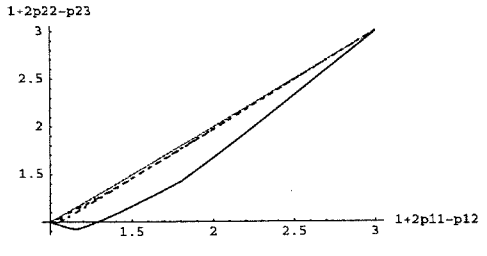
$$p_{2,3}=\left(1+198p_{1,2}-\sqrt{1+396p_{1,2}-796p_{1,2}^2}\right)/200\quad(p_0\leq p_{1,2}<2/5)\,.$$



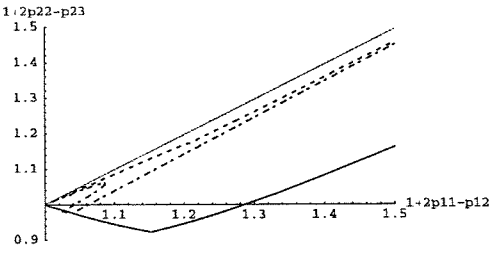
(a) Relation between $p_{1,2}$ and $p_{2,3}$



(b) $p_{1,1}$ and $p_{1,2}$



(c) $1 + 2p_{1,1} - p_{1,2}$ and $1 + 2p_{2,2} - p_{2,3}$



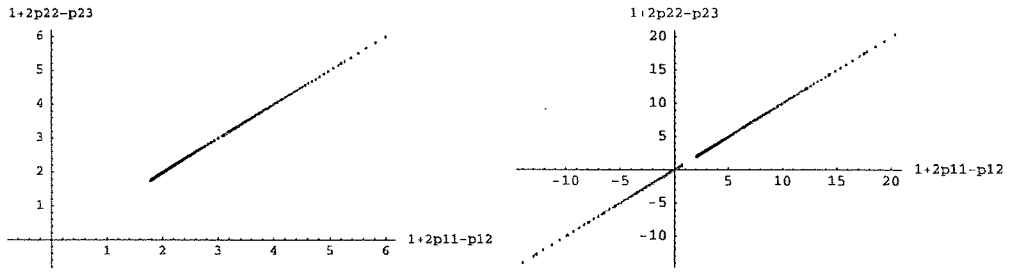
(d) Magnified graph of (c)

FIGURE 4. Relations between the proliferation $p_{i,i}$ and transition rates $p_{i,i+1}$ when the maximum values above are satisfied. The black line, dashed line, gray line denote the relations when γ is 1, 10 and 100, respectively *with* the cell-type order conservation rule. The dot-dash line shows the case with $\gamma = 10$ *without* the cell-type order conservation rule. (b) Modes I, II and III correspond to the three curves (or lines) into which the points where the curve is not smooth separate the whole region. Mode I includes the origin. (d) is the graph of (c) magnified around $(1, 1)$. Note in (c) and (d), the line $1 + 2p_{1,1} - p_{1,2} = 1 + 2p_{2,2} - p_{2,3}$ is much the same as the gray curve.

- Mode III:

$$p_{2,3} = \left(200 - 99p_{1,2} - \sqrt{40000 - 199600p_{1,2} + 249801p_{1,2}^2}\right)/400 \qquad (0 < p_{1,2} \leq 2/5)\,.$$

Modes I, II and III show the existence of three stages, in which the cell-type diversity is highest. We have observed the existence of the three stages when $\gamma \geq 1$. Notice that Modes I, II and III, described by algebraic functions of $p_{1,2}$, have been inferred using the QE method.



(a) 0.01 (b) 0.05

FIGURE 5. Rate-relations between the points that are lowered from the highest cell-type diversity curve by 0.01 (a) and 0.05 (b). The black points were calculated with the conservation rule, the gray points without the rule.

We have focused on the case of $\gamma \geq 10$ because in our previous simulation [22], the constraint (4.4) over $N(XY), (X, Y \in \{A, B, C\})$ was observed, and there are few initial-type cells (A in this work) corresponding to stem cells in real biological systems [8]. Remember that $1/\gamma$ is defined as the ratio of the initial cells to the developed cells in Section 4.1.

We have also calculated the relation between the proliferation and transition rates when $N(AB)$ is the maximum *without* the cell-type order conservation rule (2.1), to evaluate the effect of the conservation rule. Figures 4 (c) and (d) show that, with the cell-type order conservation rule, the $(1+2p_{1,1}-p_{1,2}, 1+2p_{2,2}-p_{2,3})$ curve (dashed and gray) is close to the line $1 + 2p_{1,1} - p_{1,2} = 1 + 2p_{2,2} - p_{2,3}$; by contrast, the curve for $\gamma = 10$ without the conservation rule (the dot-dashed line) is separate from $1 + 2p_{1,1} - p_{1,2} = 1 + 2p_{2,2} - p_{2,3}$. Such a tendency is observed as long as $\gamma \geq 10$.

4.3. Relation between cell-type diversity and the order conservation rule

We also evaluated the robustness of high cell-type diversity when γ is 10 with and without the cell-type order conservation rule. This evaluation was performed by deriving the relation between the proliferation and transition rates obtained when the points are lowered by 0.01 and 0.05 from the highest cell-type diversity curve inferred exactly in Section 4.2. As illustrated in Figures 5(a) and (b), the set of points (gray) without the conservation rule are more separate from the original set than the set (black) with the rule.

This indicates that, without the cell-type order conservation rule, the relation between the proliferation and transition rates wherein high cell-type diversity is realized is less robust. Taking account of the results in Section 4.2 and these results, we can safely state that the cell-type order conservation rule plays a key role in high cell-type diversity.

5. Conclusion

One of the remarkable features in this study is that algebraic relations have been inferred over the IL-system with the aid of quantifier elimination. Indeed, the inferred relations between cell-type diversity and cell-type order conservation have revealed that cell-type diversity appears robustly if and only if the cell-type order conservation rule exists.

Although our model assumes only three cell types, our approach of combining IL-systems and algebraic computation will shed some light on the important role of the cell-type order conservation rule for multi-cellular organisms and in inference problems over IL-systems.

Acknowledgements

We wish to express our gratitude to Professor Christopher W. Brown for helpful calculations and useful suggestions on QEPCAD-B and to Professor Kunihiko Kaneko for valuable discussions. H. Yoshida and K. Horimoto were partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (grant 18016008) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. This study was supported in part 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] H. Anai, K. Horimoto. Symbolic-numeric estimation of parameters in biochemical models by quantifier elimination, *J. Bioinfo. Comput. Biol.*, Vol. 4 (2006), 1097–1107.

[2] D. L. Balkwill, S. A. Nierzwicki-Bauer, and S. E. Stevens, Jr. Modes of cell division and branch formation in the morphogenesis of the cyanobacterium, *Mastigocladus laminosus*, *J. Gen. Microbiol.*, Vol. 130 (1984), 2079–2088.

[3] B. F. Caviness, J. R. Johnson. *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Springer-Verlag, Vienna, 1998.

[4] P. Doucet. The syntactic inference problem for D0L-sequences, In *Lecture in Notes in Computer Science*, Vol. 15 of *L-Systems*, Springer-Verlag, 1974, 146–161.

[5] P. Eichhorst, F. Ruskey. On unary stochastic Lindenmayer systems, *Inf. Control*, Vol. 48 (1981), 1–10.

[6] P. Eichhorst, W. J. Savitch. Growth functions of stochastic Lindenmayer systems, *Inf. Control*, Vol. 45 (1980), 217–228.

[7] H. Feliciangeli, G. T. Herman. Algorithms for producing grammars from sample derivations: a common problem of formal language theory and developmental biology, *J. Comput. Syst. Sci.*, Vol. 7 (1973), 97–118.

[8] S. F. Gilbert. *Developmental Biology*, Sinauer Associates, 7th edition, 2003.

[9] G. T. Herman, A. D. Walker. The syntactic inference problem applied to biological systems, In B. Meltzer and D. Michie eds., *Machine Intelligence*, Vol. 7, Edinburgh University Press, 1972, chapter 18, 314–356.

[10] J. M. Janssen, A. Lindenmayer. Models for the control of branch positions and flowering sequences of capitula in *Mycelis Muralis* (*L.*) *Dumont* (*Compsitae*), *New Phytol.*, Vol. 105 (1987), 191–220.

[11] H. Jürgensen, A. Lindenmayer. Inference algorithms for developmental systems with cell lineages, *Bull. Math. Biol.*, Vol. 49 (1987), 93–123.

[12] G. Kókai, Z. Tóth, and R. Ványi. Modelling blood vessels of the eye with parametric L-systems using evolutionary algorithms, in W. Horn, et al. (eds), *Lecture Notes in Artificial Intelligence*, Vol. 1620 of *Artificial Intelligence in Medicine*, Springer-Verlag, Berlin Heidelberg, 1999, 433–442.

[13] A. Lindenmayer. Mathematical models for cellular interactions in development. I. Filaments with one-sided inputs, *J. Theor. Biol.*, Vol. 18 (1968), 280–299.

[14] A. Lindenmayer. Mathematical models for cellular interactions in development. II. Simple and branching filaments with two-sided inputs, *J. Theor. Biol.*, Vol. 18 (1968), 300–315.

[15] S. Orii, H. Anai, and K. Horimoto. Symbolic-numeric estimation of parameters in biochemical models by quantifier elimination, in *Bioinfo*, 2005, Int. Joint Conf. of InCoB, AASBi, and KSBI.

[16] private communications with Prof. Kunihiko Kaneko * Graduate School of Arts and Sciences Department of Basic Science, The University of Tokyo.

[17] R. Ványi, G. Kókai, Z. Tóth, and T. Pető. Grammatical Retina description with enhanced methods, In R. Poli, W. Banzhaf, W. B. Langdon, J. F. Miller, P. Nordin, and T. C. Fogarty, eds., *Genetic Programming, Proceedings of EuroGP*, Vol. 1802 of *LNCS*, Springer-Verlag, 2000.

[18] T. Yokomori. Stochastic characterization of E0L languages, *Inf. Control*, Vol. 45 (1980), 26–33.

[19] T. Yokomori. Inductive inference of 0L languages, In G. Rozenberg and A. Salomaa, eds., *Lindenmayer Systems: Impacts on Theoretical Computer Science, Computer Graphics, and Developmental Biology*, Springer-Verlag, 1992, chapter 2, 115–132.

[20] H. Yoshida, H. Anai, and K. Horimoto. Derivation of rigorous conditions for high cell-type diversity by algebraic approach, *Biosystems*, Vol. 90 (2007), 486–495, (doi:10.1016/j.biosystems.2006.11.008).

[21] H. Yoshida, H. Anai, S. Orii, and K. Horimoto. Inquiry into conditions for cell-type diversity of multicellular organisms by quantifier elimination, in *Algebraic Biology*, Vol. 1, 2005.

[22] H. Yoshida, C. Furusawa, and K. Kaneko. Selection of initial conditions for recursive production of multicellular organisms, *J. Theor. Biol.*, Vol. 233 (2005), 501–514, (doi:10.1016/j.jtbi.2004.10.026).

[23] H. Yoshida, T. Yokomori, and A. Suyama. A simple classification of the volvocine algae by formal languages, *Bull. Math. Biol.*, Vol. 67 (2005), 1339–1354, (doi:10.1016/j.bulm.2005.03.001).

Hiroshi Yoshida
Faculty of Mathematics
Organization for the Promotion of Advanced Research
Kyushu University
Hakozaki 6–10–1
Higashi-ku
Fukuoka 812–8581
Japan
e-mail: phiroshi@math.kyushu-u.ac.jp

Katsuhisa Horimoto
Computational Biology Research Centre (CBRC)
National Institute of Advanced Industrial Science and Technology (AIST)
Aomi 2–42
Koto-ku
Tokyo 135–0064
Japan
e-mail: k.horimoto@aist.go.jp

Hirokazu Anai
IT Core Laboratories
Fujitsu Laboratories Ltd./CREST, JST.
Kamikodanaka 4–1–1
Nakahara-ku
Kawasaki 211–8588
Japan
e-mail: anai@jp.fujitsu.com

Received: January 15, 2007.
Revised: November 3, 2007.
Accepted: November 12, 2007.

Integer programming-based approach to allocation of reporter genes for cell array analysis

Morihiro Hayashida*

Bioinformatics Center,
Institute for Chemical Research,
Kyoto University, Gokasho, Uji, 611-0011, Japan
E-mail: morihiro@kuicr.kyoto-u.ac.jp
*Corresponding author

Fuyan Sun, Sachiyo Aburatani, and Katsuhisa Horimoto

Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology,
2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan
E-mail: sun-fuyan@aist.go.jp
E-mail: s.aburatani@aist.go.jp
E-mail: k.horimoto@aist.go.jp

Tatsuya Akutsu

Bioinformatics Center,
Institute for Chemical Research,
Kyoto University, Gokasho, Uji, 611-0011, Japan
E-mail: takutsu@kuicr.kyoto-u.ac.jp

Abstract: In this paper, we consider the problem of selecting the most effective set of reporter genes for analysis of biological networks using cell microarrays. 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 the biological significance of the result. We also apply them to artificial networks, the result of which shows that optimal solutions can be obtained within several seconds for networks with 10,000 nodes.

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

Reference to this paper should be made as follows: Hayashida, M., Sun, F., Aburatani, S., Horimoto, K. and Akutsu, T. (2008) 'Integer programming-based approach to allocation of reporter genes for cell array analysis', *Int. J. Bioinformatics Research and Applications*, Vol. 4, No. 4, pp.385–399.

Biographical notes: Morihiro Hayashida is an Assistant Professor in Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Japan. He received his MSc Degree in Information Science from University of Tokyo, Japan, in 2002 and his PhD Degree in Informatics from Kyoto University, Japan, in 2005. His research interests include issues related to protein function prediction and bioinformatics.

Fuyan Sun is a research staff of the Biological Network Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan. She received her PhD Degree in the Nagasaki University Graduate School of Biomedical Sciences, Japan, in 2003. From 2003 to 2005, she was a research staff at the Center for Developmental Biology, RIKEN. Her research focuses on biological and medical informatics.

Sachiyo Aburatani is a research scientist of the Biological Network Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan. She received her PhD Degree in Agricultural Science from Kyushu University, Japan, in 2003. From 2003 to 2006, she was an Assistant Professor at the Institute of Medical Science, University of Tokyo. Her research interests are in the areas of gene regulatory networks with the use of DNA microarrays and bioinformatics.

Katsuhisa Horimoto is a leader of the Biological Network Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan. He received his PhD Degree in Biophysics in 1991, from Science University of Tokyo, Japan. From 1991 to 1997, he worked at Science University of Tokyo as a research associate. He was an Associate Professor at Saga Medical School, Japan, from 1997 to 2001, and a Professor at the Laboratory of Biostatistics, University of Tokyo, from 2002 to 2006. His interests include the development of computational methods to elucidate the properties of biological networks.

Tatsuya Akutsu is a Professor in Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. He received his MEng Degree in Aeronautics in 1996 and a Dr Eng Degree in Information Engineering in 1989, both from University of Tokyo, Japan. From 1989 to 1994, he was with Mechanical Engineering Laboratory, Japan. He was an Associate Professor in Gunma University from 1994 to 1996, and in Human Genome Center, University of Tokyo from 1996 to 2001 respectively. He joined Kyoto University in October 2001. His research interests include bioinformatics and discrete algorithms.

1 Introduction

One of the important topics in drug design and bioinformatics is identification of novel target genes for the treatment of diseases. For that purpose, various approaches have been proposed. Among these, *transfected cell microarrays* (*cell arrays* for short) are regarded as a potentially powerful approach (Bailey et al., 2002; Kato et al., 2004; Yoshikawa et al., 2004; Ziauddin and Sabatini, 2001). 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 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 overexpressed. On the other hands, it is also possible to knock out some specific gene by using siRNA (Bailey et al., 2002; Yoshikawa et al., 2004). Therefore, we may be able to observe effects of gene overexpression 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 some additional technology. 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 (Hadjantonakis et al., 2003). 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 (Stearman et al., 2007; Golzio et al., 2007). *Repoter genes* are genes encoding these fluorescent proteins, by which we can observe the expression level of gene or the corresponding product through the magnitude of fluorescence. Combining repoter genes with the cell array technology, we may be able to visually observe effects of gene overexpression, gene knockdown or external signals on gene expressions in living cells. However, the cost (both in labour 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 (Chabrier-Rivier et al., 2004; Eker et al., 2002; Tran et al., 2005). 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 (Ruths et al., 2006). 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 randomised 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 well-known 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 (Vazirani, 2001). It should be noted that our approach is significantly different from that in Ruths et al. (2006):

- problems and network representations are different from each other
- optimality of the solution is not guaranteed in Ruths et al. (2006), 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 $V = \{v_1, \ldots, v_n\}$ 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 *neighbouring 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 that 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 $L = 2$.

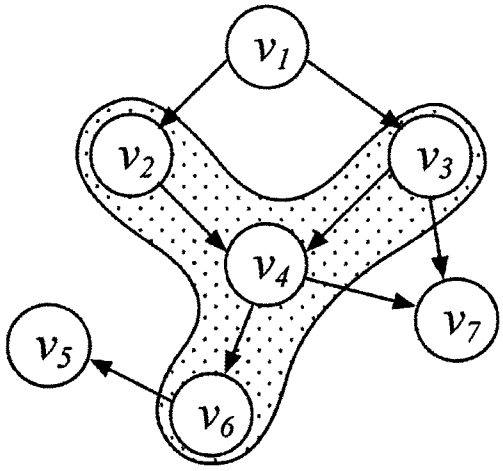
Problem P1 maximises the number of covered nodes by using K reporters, and is defined as follows.

Definition 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 maximising 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.

Definition 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.

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



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 Vazirani (2001).

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

Proof: We show an L-reduction from the maximum coverage problem (Vazirani, 2001; Hochbaum, 1982), which is known to be MAX SNP-hard (Akutsu and Bao, 1996), 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 maximises the number of covered elements in U. From an instance $I = \langle U = \{u_1, \ldots, u_m\}, S = \{s_1, \ldots, s_l\}, k(\leq l)\rangle$ 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):

$$V = \{u_1, \ldots, u_m, s_1, \ldots, s_l\},$$

$$E = \bigcup_{j=1}^{l} \bigcup_{u_i \in s_j} \{(u_i, s_j)\},$$

$$L = 1, \quad K = k.$$

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 the costs of 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, $OPT(I') \leq 2OPT(I)$.

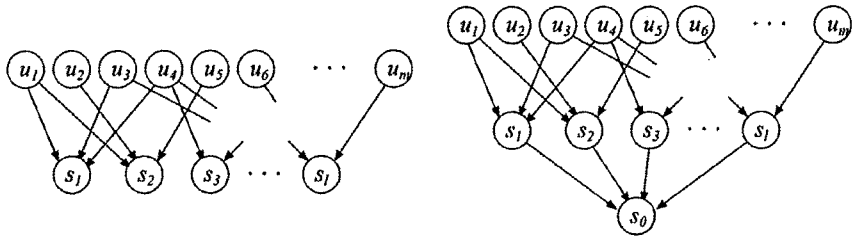
Given any solution $R \subseteq V$ of I' with cost (i.e., the number of covered nodes) c', we produce a solution C of I in polynomial time by letting $C = R - U$, where $R - U = \{r | r \in R \text{ and } r \notin U\}$. Then, $|C| \leq |R| \leq k$. Let c be the cost (i.e., the number of covered elements)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

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



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 some 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 $I = \langle U = \{u_1, \ldots, u_m\}, S = \{s_1, \ldots, s_l\}\rangle$ 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):

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

where s_0 is a node not in S.

Let $OPT(I)$ and $OPT(I')$ be the costs of 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 $C = (R - U - \{s_0\}) \cup \{s_j | \text{ for } u_i \in R - S - \{s_0\}, u_i \in \exists s_j\}$. 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 from the assumption 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. □

It is to be noted that the reduction in the proof of Theorem 2 also provides a proof of NP-hardness of Problem P2.

Though we have shown negative results on approximation of problems P1 and P2, we can also show positive results on approximation ratios using a well-known greedy algorithm for the set cover (Vazirani, 2001; Hochbaum, 1982; Akutsu and Bao, 1996).

Proposition 1: *P1 can be approximated within a factor of* $e/(e-1)$ *in polynomial time, where* e *is the base of the natural logarithm.*

Proof: We reduce P1 to the maximum coverage problem. From an instance $I = \langle G = (V, E), L, K\rangle$ of P1. we construct an instance I' of the maximum coverage problem by letting $U = V$, $S = \{s_v | s_v$ is the set of nodes covered by $v \in V\}$, and $k = K$. It is clear that this reduction can be done in linear time.

Then, by identifying a node v with a set s_v, we can see the following.

- $OPT(I) = OPT(I')$ holds
- From a solution R of I' with cost c, we can obtain a solution C of I with cost c.

Since the maximum coverage problem can be approximated within a factor of $e/(e-1)$ using the simple greedy algorithm for the set cover problem (Vazirani, 2001; Hochbaum, 1982; Akutsu and Bao, 1996), P1 can also be approximated within a factor of $e/(e-1)$. □

Proposition 2: *P2 can be approximated within a factor of* $O(\log n)$ *in polynomial time.*

Proof: We reduce P2 to the set cover problem as in the proof of Proposition 1, where k is not relevant in this case. Then, it is straight-forward to see that P2 is approximated within a factor of $O(\log n)$ since the set cover problem can be approximated within a factor of $O(\log n)$ using the simple greedy algorithm (Vazirani, 2001; Hochbaum, 1982). □

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 IPs, 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.

$$
\begin{aligned}
\text{(IP1)}\quad & \text{Maximise} && \sum_{i=1}^{n} y_i, \\
& \text{Subject to} \\
& && y_i \leq \sum_{j \in S_i^L} x_j \quad \text{for } i = 1, .., n, \\
& && \sum_{i=1}^{n} x_i \leq K, \\
& && x_i = \{0, 1\}, \\
& && y_i = \{0, 1\},
\end{aligned}
$$

where S_i^L is the set of nodes covering 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 maximises the number of covered nodes using at most K reporter nodes.

Similarly, Problem P2 is formulated as follows.

$$
\text{(IP2)}\quad \text{Minimise} \quad \sum_{i=1}^{n} x_i,
$$

Subject to

$$\sum_{j \in S_i^L} x_j \geq 1 \quad \text{for } i = 1, .., n,$$
$$x_i = \{0, 1\}.$$

IP2 minimises 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, http://www.ilog.com/products/cplex/) for solving IP1 and IP2, and measured execution time of the optimisation 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 (Kimberley and Screaton, 2004). The maps were constructed by a commercial software, MetaCore (GeneGo Corp., http://www.genego.com/metacore.php), in which 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 $L(= 1, \ldots, 6, 132)$ and $K(= 1, \ldots, 6)$. 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 (Watts and Strogatz, 1998). 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 $L = 1, \ldots, 3$, 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