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.

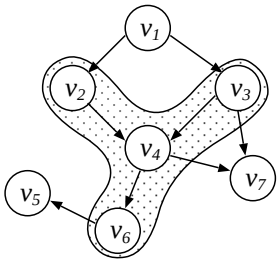

Figure 1: Example of nodes covered by a reporter node when $L = 2$ in a directed graph $G = (V,E)$ with $V = \{v_1, ..., v_7\}$. In this case, v_2, v_3, v_4 and v_6 are covered by v_6.

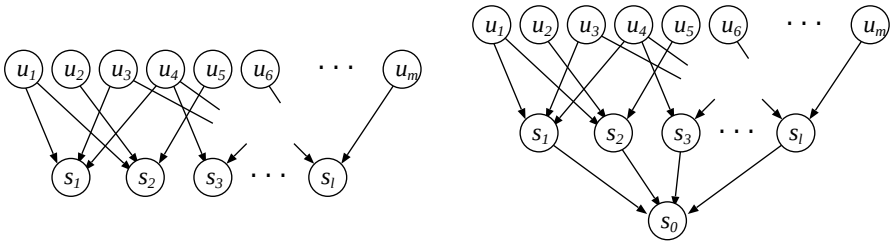

Figure 2: Left: Transformation of an instance $I = \langle U = \{u_1, ..., u_m\}, S = \{s_1, ..., s_l\}, k\rangle$ of the maximum coverage problem to Problem P1. Right: Transformation of $I = \langle U, S\rangle$ of the set cover problem to Problem P2.

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 $\sum_{i=1}^{n} y_i,$

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

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.

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

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

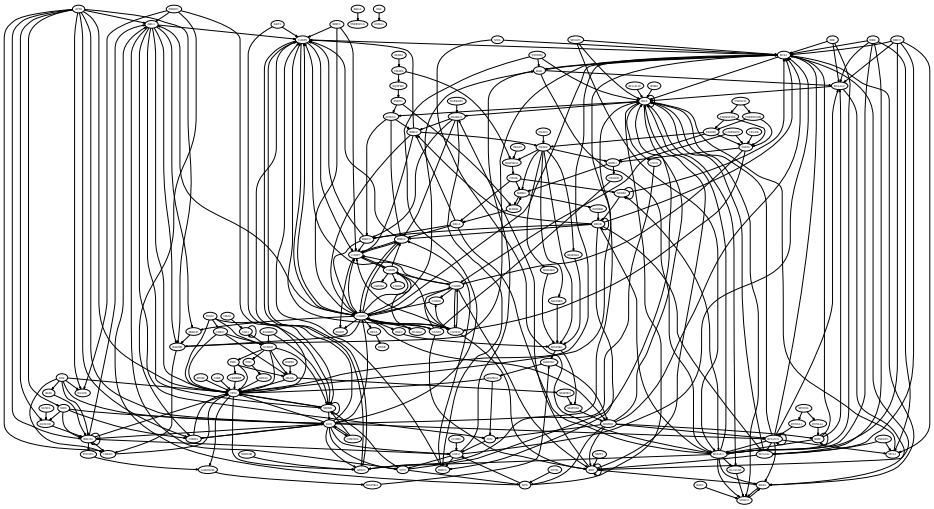

Figure 3: Apoptosis pathway maps in a HeLa cell, which contain 132 proteins and 337 binomial relations.

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 $K=1$
	1	2	3	4	5	6		(indegree/outdegree)
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

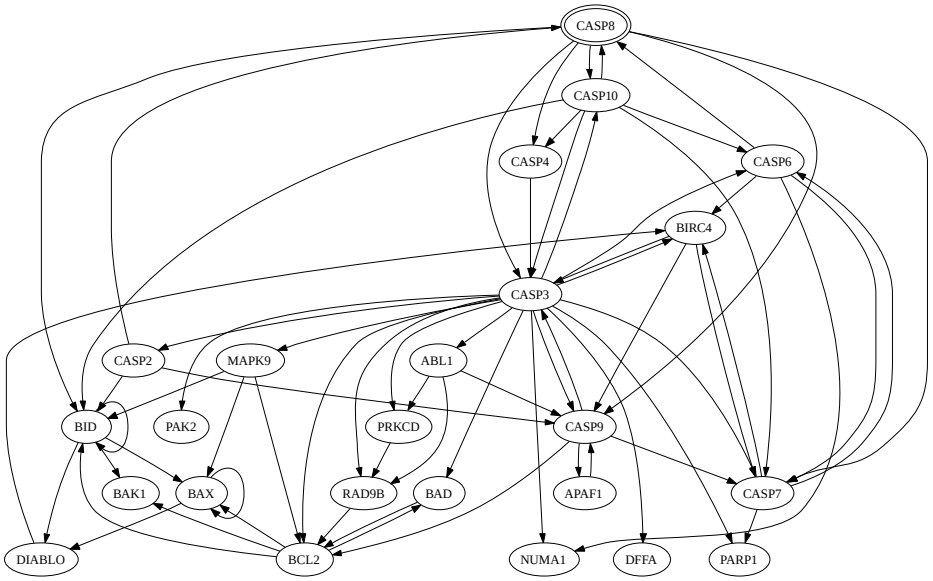

Figure 4: Downstream proteins of CASP8 within the distance 2 in apoptosis pathway maps. CASP8 is highlighted with the double circles. We excluded CASP8 from this subnetwork not to select it as a reporter.

edges were selected as the optimal reporter nodes in IP2.

5.2 Artificial Scale-free Networks

It is known that many real biological networks have the scale-free property [24]. In particular, it is observed that gene regulatory networks have the power-law outdegree distribution and the Poisson indegree distribution [25]. Thus, we generated scale-free networks with a power-law outdegree distribution ($\propto k^{-2.5}$) and Poisson indegree distribution as follows. We first choose the outdegree for each node from a power-law distribution. That is, the outdegree d_i of node v_i is drawn from a power-law distribution. Then, we choose d_i output nodes randomly with uniform probability from n nodes. Thus, the indegree distribution should follow a Poisson distribution.

Table 4 shows the average CPU time over 100 networks for each case. For large $n(= 1000, 5000, 10000)$, the elapsed time was sufficiently short (even in the case of $L = 3$ and $K = 5$). This result suggests that the proposed methods are scalable to realistic size instances. The elapsed time of IP2 was shorter than that of IP1 for almost all cases. It is reasonable because IP1 has twice as many integer variables as IP2, and the number of constraints in IP1 is larger than that in IP2.

6 Concluding Remarks

We defined two problems P1 and P2 to allocate reporter genes that are effective for observing behaviors of various biological networks. We showed hardness results

Table 3: Selected proteins as reporters for each L and K in the downstream proteins of CASP8.

L	K	IP1	Reporters
1	1	6	BCL2
1	2	10	BID, CASP7
1	3	13	BCL2, BID, BIRC4
1	10 (IP2)	23	CASP9, RAD9B, BCL2, BAK1, DIABLO, CASP3, DFFA, NUMA1, PAK2, PARP1
2	1	13	BCL2
2	2	18	BCL2, BIRC4
2	3	19	BCL2, DIABLO, NUMA1
2	7 (IP2)	23	BCL2, BAK1, DIABLO, DFFA, NUMA1, PAK2, PARP1
3	1	16	BAD
3	6 (IP2)	23	CASP9, BAK1, DFFA, NUMA1, PAK2, PARP1
4	5 (IP2)	23	BAK1, DFFA, NUMA1, PAK2, PARP1
23	1	19	BAK1
23	5 (IP2)	23	BAK1, DFFA, NUMA1, PAK2, PARP1

on approximation of these problems. On the other hand, by means of reduction to the set cover problem, we showed that P1 and P2 can be approximated within a factor of $e/(e-1)$ and $O(\log n)$, respectively.

We proposed integer programming-based methods IP1 and IP2 for solving practical instances of P1 and P2, respectively. We applied them to apoptosis pathway maps, and found that such proteins as TP53, BCL2 and BAX selected by our methods often correspond to hubs in the network. These proteins are also considered to play important biological roles. Furthermore, we applied our methods to artificial scale-free networks with up to 10,000 nodes, and we showed that our methods can compute optimal solutions for these networks in practical time.

Table 4: Elapsed time (sec.) of solving IP1 and IP2 for each n, L and K.

n	L	K	IP1	IP2
1000	1	1	0.0147972	0.00932519
1000	3	5	0.904964	0.0526494
5000	1	1	0.102972	0.0485728
5000	3	5	2.90922	0.841976
10000	1	1	0.276991	0.101553
10000	3	5	5.62986	4.01971

Though we considered directed and unweighted networks in this paper, IP1 and IP2 can be modified for undirected and/or weighted networks. Furthermore, we can add various kinds of constraints to IP1 and IP2 because these are based on integer programming. Such a flexibility would be useful for modifying the proposed methods according to requirements from experimental biologists.

Acknowledgments

We would like to thank Prof. Yuichi Sugiyama in University of Tokyo for valuable suggestions. This work is partially supported by the Cell Array Project from NEDO, Japan and by a Grant-in-Aid "Systems Genomics" from MEXT, Japan.

References

[1] S. N. Bailey, R. Z. Wu and D. M. Sabatini. Applications of transfected cell microarrays in high-throughput drug discovery. *Drug Discovery Today*, 7, S113–S118, 2002.

[2] K. Kato, K. Umezawa, M. Miyake, J. Miyake and T. Nagamune. Transfection microarrays of nonadherent cells on an oleyl poly (ethylene glycol) ether-modified glass slide. *Biotechniques*, 37, 444–452, 2004.

[3] T. Yoshikawa, E. Uchimura, M. Kishi, D. P. Funeriu, M. Miyake and J. Miyake. Transfection microarray of human mesenchymal stem cells and on-chip siRNA gene knockdown. *Journal of Controlled Release*, 96, 227–232, 2004.

[4] J. Ziauddin and D. M. Sabatini. Microarray of cells expressing defined cDNAs. *Nature*, 411, 107–110, 2001.

[5] A. K. Hadjantonakis, M. E. Dickinson, S. E. Fraser and V. E. Papaioannou. Technicolour transgenics: imaging tools for functional genomics in the mouse. *Nature Reviews Genetics*, 4, 613–625, 2003.

[6] R. S. Stearman, M. C. Grady, P. Nana-Sinkam, M. Varella-Garcia and M. W. Geraci. Genetic and epigenetic regulation of the human prostacyclin synthase promoter in lung cancer cell lines. *Molecular Cancer Research*, 5, 295–308, 2007.

[7] M. Golzio, L. Mazzolini, A. Ledoux, A. Paganin, M. Izard, L. Hellaudais, A. Bieth, M. J. Pillaire, C. Cazaux, J. S. Hoffmann, B. Couderc and J. Teissié. In vivo gene silencing in solid tumors by targeted electrically mediated siRNA delivery. *Gene Therapy*, 14, 752–759, 2007.

[8] N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages and V. Schächter. Modeling and querying biomolecular interaction networks. *Theoretical Computer Science*, 325, 25–44, 2004.

[9] S. Eker, M. Knapp, K. Laderoute, P. Lincoln and C. L. Talcott. Pathway Logic: executable models of biological networks. *Electric Notes in Theoretical Computer Science*, 71, 144–161, 2002.

[10] N. Tran, C. Baral, V. J. Nagaraj and L. Joshi. Knowledge-based framework for hypothesis formation in biochemical networks. *Bioinformatics*, 21, ii213–ii219, 2005.

[11] D. A. Ruths, L. Nakhleh, M. S. Iyengar, S. A. G. Reddy and P. T. Ram. Hypothesis generation in signaling networks. *Journal of Computational Biology*, 9, 1546–1557, 2006.

[12] V. V. Vazirani. *Approximation Algorithms*. Springer, 2001.

[13] D. S. Hochbaum. Approximation algorithms for the set covering and vertex cover problems. *SIAM Journal on Computing*, 11, 555–556, 1982.

[14] T. Akutsu and F. Bao. Approximating minimum keys and optimal substructure screens. *Lecture Notes in Computer Science 1090 (Proc. COCOON 96)*, 290–299, 1996.

[15] http://www.ilog.com/products/cplex/

[16] F. C. Kimberley and G. R. Screaton. Following a TRAIL: Update on a ligand and its five receptors. *Cell Research*, 14, 359–372, 2004.

[17] http://www.genego.com/metacore.php

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 440–442, 1998.

[19] G. S. Wu, T. F. Burns and E. R. McDonald III, W. Jiang, R. Meng, I. D. Krantz, G. Kao, D.-D. Gan, J.-Y. Zhou, R. Muschel, S. R. Hamilton, N. B. Spinner, S. Markowitz, G. Wu and W. S. El-Deiry. KILLER/DR5 is a DNA damage-inducible p53-regulated death receptor gene. *Nature Genetics*, 17, 141–143, 1997.

[20] M. R. Sprick and H. Walczak. The interplay between the Bcl-2 family and death receptor-mediated apoptosis. *Biochim Biophys Acta*, 1644, 125–132, 2004.

[21] E. R. Geisbrecht and D. J. Montell. A role for Drosophila IAP1-mediated caspase inhibition in Rac-dependent cell migration. *Cell*, 118, 111–125, 2004.

[22] U. Fischer, R. U. Janicke and K. Schulze-Osthoff. Many cuts to ruin: a comprehensive update of caspase substrates. *Cell Death Differentiation*, 10, 76–100, 2003.

[23] M. Lamkanfi, N. Festjens, W. Declercq, T. Vanden Berghe and P. Vandenabeele. Caspases in cell survival, proliferation and differentiation. *Cell Death and Differentiation.*, 14, 44–55, 2007.

[24] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 509–512, 1999.

[25] N. Guelzim, S. Bottani, P. Bourgine and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31, 60–63, 2002.