Lecture Notes in Operations Research 7

Series Editors:
Ding-Zhu Du and Xiang-Sun Zhang

OPTIMIZATION AND SYSTEMS BIOLOGY

The First International Symposium, OSB'07
Beijing, China, August 8–10, 2007
Proceedings

Edited by
Xiang-Sun Zhang
Luonan Chen
Ling-Yun Wu
Yong Wang

WORLD PUBLISHING CORPORATION

The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 7–9, 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.

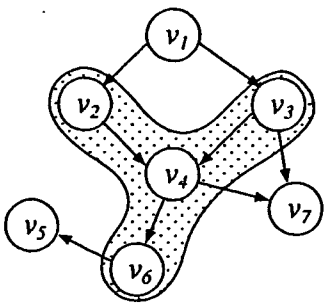

Figure 1: Example of nodes covered by a reporter node when $L = 2$ in a directed graph $G = (V,E)$ with $V = \{v_1,...,v_7\}$. In this case, v_2, v_3, v_4 and v_6 are covered by v_6.

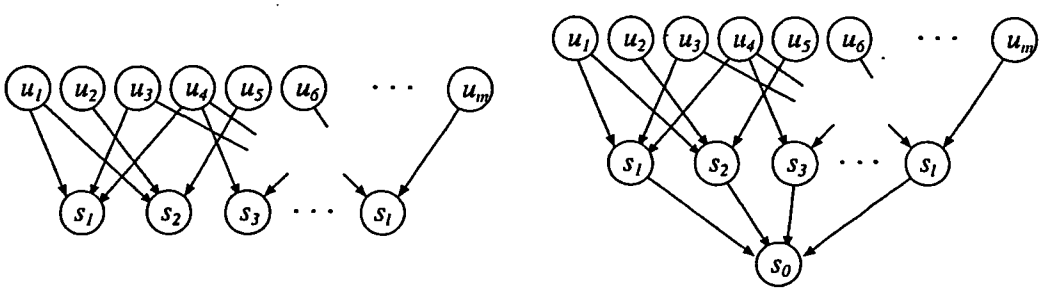

Figure 2: Left: Transformation of an instance $I = \langle U = \{u_1,...,u_m\}, S = \{s_1,...,s_l\}, k\rangle$ of the maximum coverage problem to Problem P1. Right: Transformation of $I = \langle U, S\rangle$ of the set cover problem to Problem P2.

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

Subject to

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

$$\sum_{i=1}^{n} x_i \leq K,$$

$$x_i = \{0,1\},$$

$$y_i = \{0,1\},$$

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.

(IP2) Minimize $\sum_{i=1}^{n} x_i,$

Subject to

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

$$x_i = \{0,1\}.$$

IP2 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

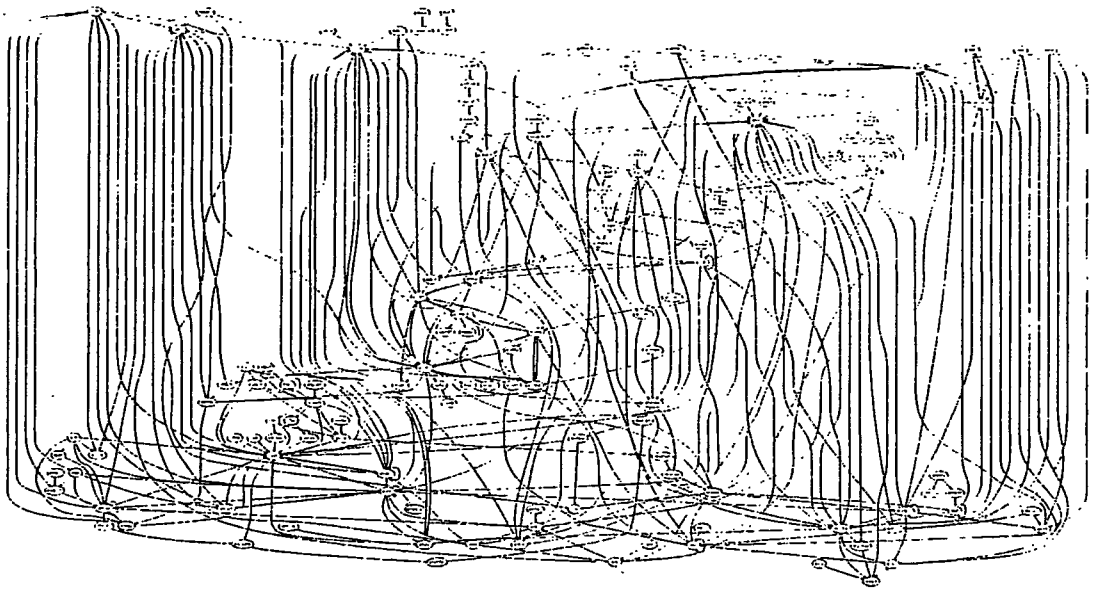

Figure 3: Apoptosis pathway maps in a HeLa cell, which contain 132 proteins and 337 binomial relations.

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

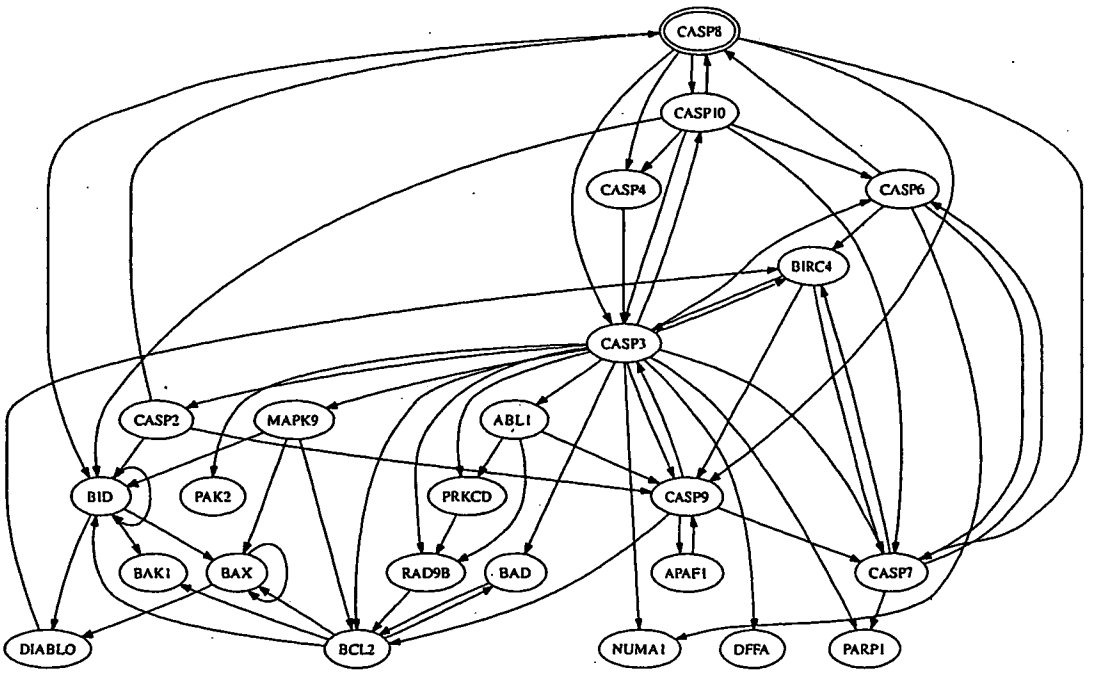

Figure 4: Downstream proteins of CASP8 within the distance 2 in apoptosis pathway maps. CASP8 is highlighted with the double circles. We excluded CASP8 from this subnetwork not to select it as a reporter.

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.

Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 366 (2008) 186–192
BBRC
www.elsevier.com/locate/ybbrc

Comparative proteomic and transcriptomic profiling of the human hepatocellular carcinoma

Hirotaka Minagawa [a], Masao Honda [b,*], Kenji Miyazaki [c], Yo Tabuse [c,*], Reiji Teramoto [c], Taro Yamashita [b], Ryuhei Nishino [b], Hajime Takatori [b], Teruyuki Ueda [b], Ken'ichi Kamijo [a], Shuichi Kaneko [b]

[a] *Nano Electronics Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan*
[b] *Department of Gastroenterology, Kanazawa University Graduate School of Medical Science, Kanazawa, 13-1 Takara-machi, Kanazawa 920-8641, Japan*
[c] *Bio-IT Center, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan*

Received 14 November 2007
Available online 4 December 2007

Abstract

Proteome analysis of human hepatocellular carcinoma (HCC) was done using two-dimensional difference gel electrophoresis. To gain an understanding of the molecular events accompanying HCC development, we compared the protein expression profiles of HCC and non-HCC tissue from 14 patients to the mRNA expression profiles of the same samples made from a cDNA microarray. A total of 125 proteins were identified, and the expression profiles of 93 proteins (149 spots) were compared to the mRNA expression profiles. The overall protein expression ratios correlated well with the mRNA ratios between HCC and non-HCC (Pearson's correlation coefficient: $r = 0.73$). Particularly, the HCC/non-HCC expression ratios of proteins involved in metabolic processes showed significant correlation to those of mRNA ($r = 0.9$). A considerable number of proteins were expressed as multiple spots. Among them, several proteins showed spot-to-spot differences in expression level and their expression ratios between HCC and non-HCC poorly correlated to mRNA ratios. Such multi-spotted proteins might arise as a consequence of post-translational modifications.

Keywords: Hepatocellular carcinoma; Proteome; Two-dimensional difference gel electrophoresis; Transcriptome; cDNA microarray

Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide, and a leading cause of death in Africa and Asia [1]. Although several major risks related to HCC, such as hepatitis B and/or hepatitis C virus infection, aflatoxin B1 exposure, and alcohol consumption, and genetic defects, have been revealed [2], the molecular mechanisms leading to the initiation and progression of HCC are not well known. To find the molecular basis of hepatocarcinogenesis, comprehensive gene expression analyses have been done using many systems such as hepatoma cell lines and tissue samples [3,4]. Previously, we have carried out a comprehensive mRNA expression analysis using the serial analysis of gene expression (SAGE) [5] and cDNA microarray-based comparative genomic hybridization [6] to acquire the outline of gene expression profile of HCC. Although these genomic approaches have yielded global gene expression profiles in HCC and identified a number of candidate genes as biomarkers useful for cancer staging, prediction of prognosis, and treatment selection [7], the molecular events accompanying HCC development are not yet understood. In general, proteins rather than transcripts are the major effectors of cellular and tissue function [8] and it is accepted that protein expression do not always correlate with mRNA expression [9,10]. Thus, protein expression analysis, which could complement the available mRNA data, is also important to understand the molecular mechanisms of HCC.

* Corresponding authors. Fax: +81 76 234 4250 (M. Honda), +81 29 856 6136 (Y.Tabuse).
E-mail addresses: mhonda@medf.m.kanazawa-u.ac.jp (M. Honda), y-tabuse@cd.jp.nec.com (Y. Tabuse).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.11.101

The technique of two-dimensional difference gel electrophoresis (2D-DIGE), developed by Unlu et al. [11] is one of major advances in quantitative proteomics. Several groups have recently utilized 2D-DIGE to examine protein expression changes in HCC samples [12,13], whereas reports on the analysis combining both transcriptomic and proteomic approach are rare.

In the present study, we compared quantitatively protein expression profiles of HCC to non-HCC (non-cancerous liver) samples derived from 14 patients by 2D-DIGE. We also compared the protein expression profiles of the same HCC and non-HCC samples to the mRNA profiles which have been obtained using a cDNA microarray. The expression ratios of 93 proteins showed significant correlations with the mRNA ratios between HCC and non-HCC. Proteins involved in metabolic processes showed more prominent correlation. Our study describes an outline of gene and protein expression profiles in HCC, thus providing us a basis for better understanding of the disease.

Materials and methods

Patients. A total of 14 HCC patients who had surgical resection done in the Kanazawa University Hospital were enrolled. The clinicopathological characteristics of them are shown in Table 1. The HCC samples and adjacent non-tumor liver samples were snap frozen in liquid nitrogen, and used for cDNA microarray and 2D-DIGE analysis. All HCC and non-tumor samples were histologically diagnosed and quantitative detection of hepatitis C virus RNA by Amplicore analysis (Roche Diagnostic Systems) showed positive. The grading and staging of chronic hepatitis associated with non-tumor lesion were histologically assessed according to the method described by Desmet et al. [14] and histological typing of HCC was assessed according to Ishak et al. [15]. All strategies used for gene expression and protein expression analysis were approved by the Ethical Committee of Kanazawa University Hospital.

Preparation of cDNA microarray slides. In addition to in-house cDNA microarray slides consisting of 1080 cDNA clones as previously described [6,16–18], we made new cDNA microarray slides for detailed analysis of the signaling pathway of metabolism and enzyme function in liver disease [19]. Besides cDNA microarray analysis, a total of 256,550 tags were obtained from hepatic SAGE libraries (derived from normal liver, CH-C, CH-C related HCC, CH-B, and CH-B related HCC), including 52,149 unique tags. Among these, 16,916 tags expressing more than two hits were selected to avoid the effect of sequencing errors in the libraries. From these candidate genes, 9614 non-redundant clones were obtained from Incyte Genomics (Incyte Corporation), Clontech (Nippon Becton Dickinson), and Invitrogen (Invitrogen). Each clone was sequence validated and PCR amplified by Dragon Genomics (Takara Bio), and the cDNA microarray slides (Liver chip 10k) were constructed using SPBIO 2000 (Hitachi Software) as described previously [6,16–18].

Table 1
Characteristics of patients involved in this study

Patient No.	Age	Sex[a]	Histology of non-tumor lesion[b]	Tumor histology	Viral status
1	64	M	F4A1	Moderate	HCV
2	65	M	F4A1	Well	HCV
3	48	M	F3A1	Moderate	HCV
4	69	F	F4A2	Moderate	HCV
5	66	F	F4A2	Well	HCV
6	45	M	F4A1	Well	HCV
7	75	F	F4A1	Well	HCV
8	46	M	F4A2	Moderate	HCV
9	66	M	F2A2	Well	HCV
10	75	M	F3A1	Moderate	HCV
11	67	F	F4A2	Well	HCV
12	64	M	F4A1	Moderate	HCV
13	68	M	F4A0	Well	HCV
14	74	M	F1A0	Moderate	HCV

[a] M, male; F, female.
[b] F, fibrosis; A, activity.

RNA isolation and antisense RNA amplification. Total RNA was isolated from liver biopsy samples using an RNA extraction kit (Stratagene). Aliquots of total RNA (5 μg) were subjected to amplification with antisense RNA (aRNA) using a Message AmpTM aRNA kit (Ambion) as recommended by the manufacturer. About 25 μg of aRNA was amplified from 5 μg total RNA, assuming that 500-fold amplification of mRNA was obtained. The quality and degradation of the isolated RNA were estimated after electrophoresis using an Agilent 2001 bioanalyzer. In addition, 10 μg of aRNA was used for further labeling procedures.

Hybridization on cDNA microarray slides and image analysis. As a reference for each microarray analysis, aRNA samples prepared from the normal liver tissue from one of the patients were used. Test RNA samples fluorescently labeled with cyanine (Cy) 5 and reference RNA labeled with Cy3 were used for microarray hybridization as described previously [6,16–18]. Quantitative assessment of the signals on the slides was done by scanning on a ScanArray 5000 (General Scanning) followed by image analysis using GenePix Pro 4.1 (Axon Instruments) as described previously [6,16–18].

Protein expression analysis using 2D-DIGE. Protein samples were homogenized with lysis buffer (7 M urea, 2 M thiourea, 4% w/v CHAPS, 0.8 μM aprotinin, 15 μM pepstatin, 0.1 mM PMSF, 0.5 mM EDTA, 30 mM Tris–HCl, pH 8.5) and centrifuged at 13,000 rpm for 20 min at 4 °C. The supernatants were used as protein samples. The protein concentrations were determined with a protein assay reagent (Bio-Rad). The non-HCC and HCC samples (50 μg each) labeled with either Cy3 or Cy5 according to the manufacture's manual were combined and separated on 2-DE gels together with the Cy2-labeled internal standard (IS), which was prepared by mixing equal amounts of all samples. Analytical 2-DE was performed as described previously [20] using Immobiline DryStrip (pH 3–10, 24 cm, GE Healthcare) in the first dimension and 12.5% SDS-polyacrylamide gels (24 × 20 cm) in the second dimension. Samples were run in triplicate to obtain statistically reasonable results. After scanning with a Typhoon 9410 scanner (GE Healthcare), gels were silver stained for protein identification. For protein identification, 400 μg of the IS sample was also separately run on a 2-DE gel and stained with SYPRO Ruby (Invitrogen). All analytical and preparative gel images were processed using ImageQuant (GE Healthcare) and the protein level analysis was done with the DeCyder software (GE Healthcare). To detect phosphoproteins, 400 μg of HCC and non-HCC samples were separately run on 2-DE gels and stained with ProQ Diamond (Invitrogen). After acquiring images, gels were counterstained with SYPRO Ruby to visualize total proteins as described above.

Protein identification. The excised protein spots were in-gel digested with porcine trypsin (Promega). For LC-ESI-IT MS/MS analysis using LCQ Deca XP (Thermo Electron), the digested and dried peptides were dissolved in 10 μl of 0.1% formic acid in 2% acetonitrile (ACN). The dissolved samples were loaded onto C18 silica gel capillary columns (Magic C18, 50 × 0.2 mm), and the elution from the column was directly connected through a sprayer to an ESI-IT MS. Mobile phase A was 2% ACN containing 0.1% formic acid, and mobile phase B was 90% ACN containing 0.1% formic acid. A linear gradient from 5% to 65% of concentration B was applied to elute peptides. The ESI-IT MS was operated in positive ion mode over the range of 350–2000 (m/z) and the database search was carried out against the IPI Human using MASCOT (Matrixscience). The following search parameters were used: the cutting enzyme, trypsin; one missed cleavage allowed, mass tolerance window, ±1 Da, the MS/MS tolerance window, ±0.8 Da; carbamidomethyl cystein and oxidized methionine as fixed and variable modifications, respectively.

Detection of phosphorylated peptide. Possible phosphorylation sites were investigated by MALDI–TOF–MS using monoammonium phosphate (MAP) added matrix mainly according to Nabetani et al. [21]. An additive of MAP was mixed with α-CHCA matrix solution (5 mg/mL, 0.1% TFA, 50% ACN aqueous) to 40 mM in final concentration. Tryptsin digests of the spots positively stained with ProQ were dissolved into 4 μL of 0.1% TFA, 50% ACN aqueous solution and 1 μL of the peptides solution was spotted on the MALDI target plate. After drying up, 1 μL of the MAP matrix was dropped on the dried peptide mixture. Voyager DE-STR (ABI) was used to obtain mass spectra both in negative and positive ion mode. MS peaks that had relatively stronger intensities in negative ion mode than in positive ion mode were selected as candidates for acidically modified peptides.

Results and discussion

We identified 195 spots representing 125 proteins (Suppl. Table 1) and obtained the corresponding mRNA expression data for a total of 93 proteins (149 spots) (Suppl. Table 2). These 93 proteins were classified according to their biological processes and subcellular localizations into categories described by the Gene Ontology Consortium (http://www.geneontology.org/index.shtml) and about a half of them were related to metabolic processes (Fig. 1A). It is a general agreement that proteins with extremely high or low p*I* as well as hydrophobic proteins are difficult to be detected by 2-DE. Being consistent with this notion, our analysis detected many cytoplasmic proteins (Fig. 1B). Therefore, the protein expression data presented here were biased in favor of cytoplasmic and soluble proteins. The protein expression abundance between non-HCC and HCC was calculated using the normalized spot volume, which was the ratio of spot volume relative to IS (Cy3:Cy2 or Cy5:Cy2) and we used the Student's paired *t*-test ($p < 0.05$) to select the protein spots which were expressed differentially between non-HCC and HCC, using 2-DE gel images run in triplicate. The spot volume of a multi-spotted protein was indicated as a total volume by integrating the intensities of multiple spots as was done by Gygi et al. [10]. Comparison of protein expression profiles revealed that several proteins were expressed differentially between HCC and non-HCC. Proteins whose abundances increased >2-fold or decreased <1/2 in HCC are listed in Table 2. While glutamine synthetase, vimentin, annexin A2 and aldo-keto reductase were up-regulated, carbonic anhydrase 2, argininosuccinate synthetase 1, carbonic anhydrase 1, fructose-1,6-bisphosphatase 1, and betaine-homocysteine methyltransferase were down-regulated in HCC. Up- or down-regulation of most of these proteins in HCC has been reported previously [22–27]. Up-regulation of vimentin and annexin A2, and reduced expression of carbonic anhydrase 1 and 2 was suspected to be associated with cellular motility and metastasis [23,24,26].

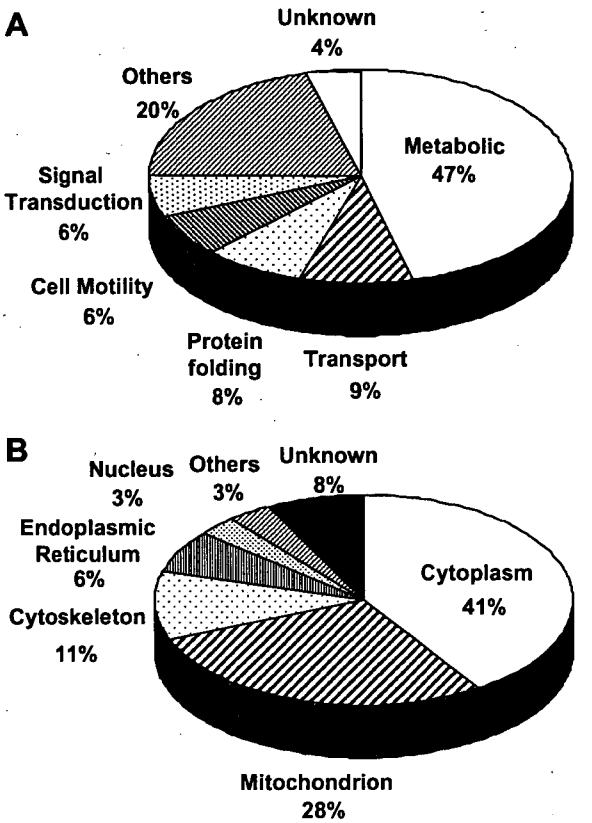

Fig. 1. Classification of identified proteins according to their cellular function (A) and subcellular localization (B).

The mRNA expression abundance was calculated from cDNA microarray data. Hierarchical clustering of

Table 2
Proteins expressed differentially between HCC and non-HCC

Spot ID	Protein name	Refseq ID	Theoretical		Fold change (HCC/non-HCC)		References
			p*I*	MW (kDa)	Protein[a]	mRNA	
1353, 1354	Glutamine synthase	NP_002056.2	6.43	42.7	2.06	3.08	[22]
1039, 1046	Vimentin	NP_003371	5.09	53.6	2.30	1.51	[23]
1716	Annexin A2	NP_001002857.1	7.57	38.8	2.57	1.82	[24]
1685, 1699	Aldo-keto reductase 1B10	NP_064695	7.12	36.2	4.29	4.73	[25]
1977	Carbonic anhydrase 2	NP_000058	6.87	29.3	0.39	0.62	[26]
1307, 1312, 1331	Argininosuccinate synthetase 1	NP_000041.2	8.08	46.8	0.41	0.30	[27]
1941	Carbonic anhydrase 1	NP_001729	6.59	28.9	0.47	1.25	[26]
1582	Fructose-1,6-bisphosphatase 1	NP_000498	6.54	37.2	0.48	0.36	
1256	Betaine-homocysteine methyltransferase	NP_001704	6.41	45.4	0.48	0.40	

[a] Integrated spot volume was used to calculate the fold change of multi-spotted proteins.

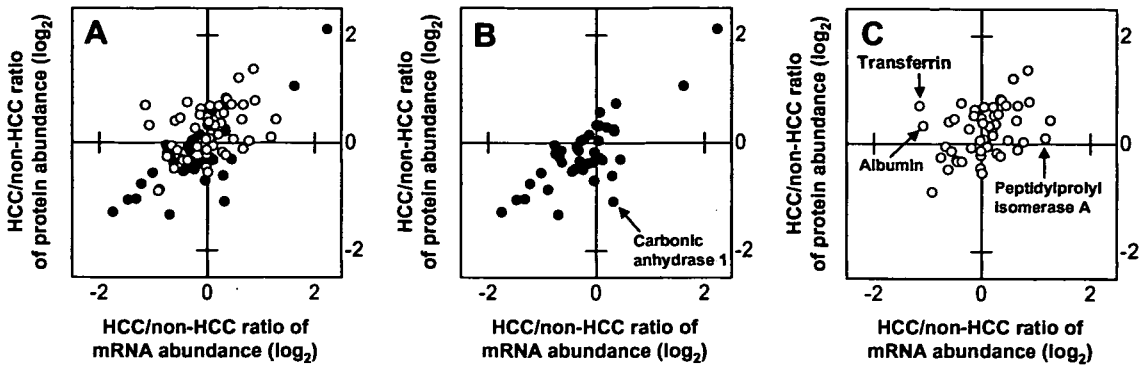

Fig. 2. Comparative analysis of protein and mRNA expression profiles between HCC and non-HCC. (A) The HCC/non-HCC ratios of averaged protein expression levels for 93 proteins were plotted against those of mRNA. Proteins related to metabolic pathways were indicated in closed circles and were shown again in (B). Proteins related to the other biochemical pathways were indicated in open circles and shown in (C). Proteins listed in Table 3 were indicated in (B) and (C). All graphs were depicted in $\log_2$ scale.

Table 3
Proteins whose expression changes between HCC and non-HCC show poor correlation to mRNA expression changes

Spot ID	Protein name	Refseq ID	Theoretical		Spot[a] Av. Ratio	Spot *p* value	Protein ratio	Micro array Av. ratio	Micro array *p* value
			*p*I	MW (kDa)					
564	Transferrin	NP_001054	6.8	79.3	2.23	0.035	1.61	0.45	3.3E−06
565					1.87	0.079			
566					2.28	0.13			
605					0.73	0.098			
1489	Albumin	NP_000468	5.9	71.3	—	0.63	1.25	0.47	2.3E−03
1941	Carbonic anhydrase 1	NP_001729	6.6	28.9	—	3.5E−03	0.47	1.25	0.39
2290	Peptidylprolyl isomerase A	NP_066953	7.7	18.1	—	5.0E−01	1.07	2.29	1.1E−01

[a] Since transferrin was detected in multiple spots, averaged ratio and spot *p* value of each spot is shown.

Table 4
Multi-spotted proteins showing spot-to-spot differences in expression level between non-HCC and HCC

Spot ID	Spot Av. ratio	Spot *p* value	Protein name	Refseq ID	Theoretical		Protein[a] ratio
					*p*I	MW (kDa)	
436	1.92	5.3E−04	Tumor rejection antigen (gp96)	NP_003290	4.8	92.7	1.2
537	0.79	0.16					
564	2.23	0.035	Transferrin	NP_001054	6.8	79.3	1.61
565	1.87	0.079					
566	2.28	0.13					
605	0.73	0.098					
1257	1.02	0.92	Fumarate hydratase	NP_000134	8.8	54.8	0.8
1261	0.6	1.3E−03					

[a] HCC/non-HCC protein ratios were calculated using integrated spot abundances.

gene expression was done with BRB-ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools.htm). The filtered data were log-transferred, normalized, centered, and applied to the average linkage clustering with centered correlation. BRB-ArrayTools contains a class comparison tool based on univariate *F* tests to find genes differentially expressed between predefined clinical groups. The permutation distribution of the *F* statistic, based on 2000 random permutations, was also used to confirm statistical significance. A *p* value of less than 0.05 for differences in HCC/non-HCC gene expression ratio was considered significant.

The average HCC/non-HCC expression ratios of the 93 proteins were plotted against the mRNA ratios in Fig. 2, where a positive value indicates increased expression in HCC and a negative ratio indicates reduced expression. The overall expression ratio of HCC/non-HCC indicated noticeable correlation between protein and mRNA

(Fig. 2A), and the Pearson's correlation coefficient for this data set (93 proteins/genes) was 0.73. Next, we divided 93 proteins into those related to metabolism and others biological processes. The HCC/non-HCC ratios of protein expression for metabolism-related proteins showed substantial correlation with those of mRNA (Fig. 2B, $r = 0.9$), whereas those of other proteins were poorly correlated (Fig. 2C, $r = 0.36$). Extreme care must be taken in a direct comparison of proteomic data with transcriptome because of multiple layers of discrepancies caused by the distinct sensitivities of cDNA array hybridization and 2-DE, the inability of a cDNA array to distinguish mRNA isoforms and post-translational modifications of proteins. Nevertheless, our results suggest that the expression of considerable portion of proteins with metabolic function listed here is regulated at transcriptional level. On the other hand, post-transcriptional and/or post-translational processes seem to be involved in the regulation of expression level for proteins with other cellular functions as a whole. Four proteins (albumin, transferrin, peptidylproryl isomerase A, and carbonic anhydrase 1) showed apparent poor correlation in protein and mRNA expression profiles (Table 3 and Fig. 2). Transcriptional control might have little effect on the expression changes of these proteins between HCC and non-HCC.

A number of proteins were expressed as multiple spots on 2-DE gels and most multi-spotted proteins showed little spot-to-spot variations in the averaged HCC/non-HCC ratio. Although we do not know how these multiple spots were generated, many of them might be due to the conformational equilibrium of proteins under electrophoresis rather than to any post-translational modifications [28]. On the other hand, the HCC/non-HCC expression ratios of several multi-spotted proteins varied from spot to spot, and three proteins (transferrin, fumarate hydratase, and tumor rejection antigen gp96) were categorized as these multi-spotted proteins (Table 4).

For example, gp96 was detected in two spots (spot #436 and 537) with distinct molecular mass and p*I* and they showed different HCC/non-HCC expression ratio (Fig. 3A and B and Table 4). The expression of these two isoforms was observed to change in the opposite direction between non-HCC and HCC: #436 was up-regulated in HCC (HCC/non-HCC ratio: 1.96) while #537 was down-regulated (HCC/non-HCC ratio: 0.79) (Table 4 and Fig. 3C and D). Gp96 is a glycoprotein present in endoplasmic reticulum and is supposed to function as a molec-

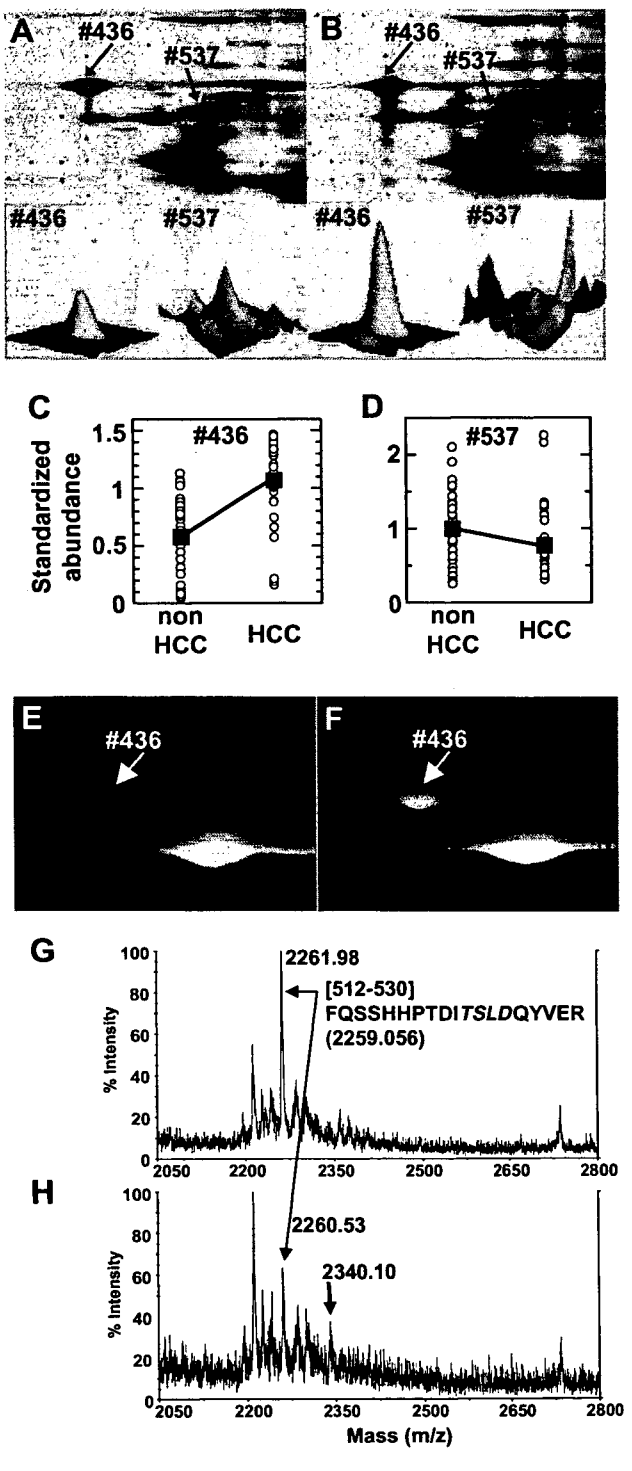

Fig. 3. Comparison of expression profiles of two gp96 spots between HCC and non-HCC. The expression profile and phosphorylation of tumor rejection antigen gp96 in HCC and non-HCC was investigated. Magnified gel images and 3D views of two gp96 spots in non-HCC (A) and HCC (B) were shown. Differences in expression level of two gp96 spots, #436 (C) and #537 (D), between non-HCC and HCC were shown. The open circle indicates the standardized abundance of the individual spot in each sample. The closed square represents the averaged abundance of each gp96 spot. Magnified gel images of non-HCC (E) and HCC (F) stained with ProQ. The #436 spot was positively stained with ProQ, while unambiguous staining of the #537 spot was not observed. Tryptic peptides prepared from the spot #436 were analyzed by MALDI–TOF mass-spectrometry in the positive ion mode (G) and the negative ion mode (H). A peak of 2261.98 detected in positive ion mode corresponds to the amino acid sequence from 512 to 530. In addition to the original peak (*m/z*: 2260.53), a peak mass shifted by +80 Da was detected in the negative ion mode. A predicted phosphorylation consensus motif for protein kinase CK2 is indicated in italics (G).