Table 1
Characteristics of the samples used for RT- PCR and SAGE

	NL $N=7$	CLD $N=20$	HCC $N=26^a$	ICC $N=16^a$
Mean age, years (SE)	64.6 (4.3)	62.7 (2.1)	65.0 (1.9)	67.9 (2.7)
Sex				
Female	3	5	6	5
Male	4	15	20	11
Hepatitis virus infection				
HBsAg-positive	0	8	9	
HCV-Ab-positive	0	12	16	
Mean ALT (SE) (IU/L)	14.2 (2.0)	61.3 (13.9)	52.6 (11.0)	47.0 (13.8)
Mean ALP (SE) (IU/L)	234.4 (36.2)	289.0 (33.1)	321.3 (31.5)	482.0 (87.1)
Mean GGT (SE) (IU/L)	93.3 (72.0)	88.0 (14.4)	82.9 (14.1)	185.5 (61.1)
Mean AFP (SE) (ng/mL)	NA	81.8 (29.7)	80.9 (24.6)	NA
Mean CEA (SE) (ng/mL)	NA	NA	5.3 (0.8)	20.9 (14.0)
Mean CA19-9 (SE) (U/L)	NA	NA	36.7 (3.7)	1171.5 (767.1)

NL, normal liver; CLD, chronic liver disease; HCC, hepatocellular carcinoma; ICC, intrahepatic cholangiocarcinoma; ALT, alanine aminotransferase; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase; AFP, alpha-foetoprotein; CEA, carcinoembryonic antigen; CA19-9, carbohydrate antigen 19-9; NA, not available.

[a] We used one ICC sample and three HCC samples for SAGE.

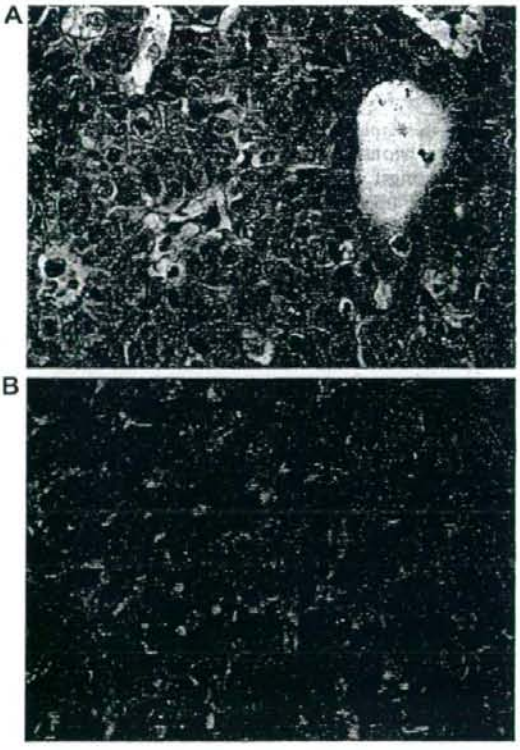


Fig. 1. Histopathological findings of the samples used for the serial analysis of gene expression. (A) Moderately differentiated intrahepatic cholangiocellular carcinoma. (B) Well-differentiated hepatocellular carcinoma. [This figure appears in colour on the web.]

ogy, Santa Cruz, CA). The blots were then washed and exposed to rabbit anti-goat IgG (1:1000 dilution) for 30 min and visualised using the ECL™ kit (GE Healthcare UK Ltd., Little Chalfont, UK).

2.5. Immunohistochemical analysis

For immunohistochemical analysis, we used 16 ICC samples and 16 HCC samples. Deparaffinised sections were treated with primary antibodies against CLDN4 (diluted 1:24, rabbit polyclonal antibodies; Cat. No. sc-17664; Santa Cruz Biotechnology) for 40 min at room temperature after appropriate antigen-retrieval treatment. Primary antibody/antigen binding was detected using a standard streptavidin–biotin–peroxidase technique (LSAB2 Kit Universal/HRP Rabbit/Mouse; DAKO, Glostrup, Denmark) and visualised using a DAB+ (3,3′-diaminobenzidine tetrahydrochloride) Liquid System (DAKO).

2.6. Discrimination analysis

We performed discrimination analysis based on the mRNA expression data using Dr. SPSS II software (SPSS Inc., Chicago, IL) to create an ICC classifier. Fifty-three of the seventy-four samples mentioned above were selected randomly: seven ICC, four NL, 20 CLD, 17 HCC and five extrahepatic adenocarcinoma samples. These 53 samples, which were subjected to RT-PCR analysis, were considered to be the training group. To verify the validity of the analysis, we performed RT-PCR using another 21 samples: nine ICC, three NL and nine HCC samples.

2.7. Statistical analyses

One-way ANOVA and Tukey–Kramer analyses were used for the statistical analysis of RT-PCR data.

3. Results

3.1. SAGE profiles of ICC

We obtained a total of 93,874 SAGE tags: 34,079 from the ICC library and 59,795 from the HCC library. Of these, 30,859 tags were unique: 14,168 and 16,689

Table 2
Clinicopathological features and discrimination analysis of the ICC samples used for RT-PCR

No.	Age (yr)	Gender	Virus	Tumour location	Size (cm)	Stage[a]	Gross appearance	Diff.	Serological marker		Z-Score	Predicted result[d]
									CEA[b] (ng/mL)	CA19-9[c] (IU/mL)		
ICC samples of the training group												
1	58	M	HCV	P	3.0 × 2.5	3	MF	Mod	5.20	23.00	4.02	ICC
2	73	M	HCV	A/P	6.5 × 5.5	4B	MF	Poor	NA	NA	7.09	ICC
3	72	M	No	A/P	7.0 × 6.5	4A	MF	Mod	7.50	NA	1.56	ICC
4	76	M	HBV	L/Med	5.3 × 4.7	4B	MF+PDI	Mod	46.00	8,600.00	2.16	ICC
5	68	F	HCV	A/P	9.0 × 6.0	4B	MF+PDI	Poor	NA	NA	0.56	ICC
6	65	M	HCV	Mu	4.0 × 3.0	4A	MF	Poor	0.00	87.00	1.31	ICC
7	53	M	HCV	Med	4.5 × 3.0	4A	PDI+MF	Poor	5.20	1,161.00	6.31	ICC
ICC samples of test group												
8	35	F	No	A/P	6.8 × 4.5	4A	MF + PDI	Poor	455.43	168.80	1.34	Not ICC
9	81	F	No	L	2.6 × 2.0	3	MF	Mod	164.58	5.10	1.63	ICC
10	73	M	No	P	5.0 × 4.0	3	MF	Mod	191,886.86	1.10	206.40	ICC
11	71	F	No	A/P	10.0 × 9.0	3	MF	Mod	3335.47	NA	13.28	ICC
12	72	F	No	M	5.0 × 4.5	3	PDI	Mod	28,614.09	NA	43.17	ICC
13	70	M	No	M	4.0 × 3.0	3	MF	Well	4533.44	4.10	16.17	ICC
14	61	M	HCV	P	1.8 × 1.5	3	MF	Mod	573.55	3.80	2.05	ICC
15	80	M	HCV	A/P	3.5 × 2.7	3	MF	Mod	2914.25	0.00	9.16	ICC
16	62	M	No	L/Med	5.2 × 3.7	3	MF	Mod	11,024.51	3.40	28.45	ICC

ICC, intrahepatic cholangiocarcinoma; Diff, differentiation; M, male; F, female; A, anterior segment; P, posterior segment; Med, medial segment; L, lateral segment; Mu, multiple segment; MF, mass-forming type; PDI, periductal infiltrating type; well, well differentiated; mod, moderately differentiated; poor, poorly differentiated.

[a] International Hepato–Pancreato–Biliary Association (IHBPA) classification [20].

[b] The normal range of CEA is <5 ng/mL.

[c] The normal range of CA19-9 is <37 IU/mL.

[d] The predicted results were determined using leave-one-out cross-validation.

tags from the ICC and HCC library, respectively. Scatterplots using all tags indicated that the ICC expression profile was different from that of HCC ($R = 0.4629$) and NL ($R = 0.3438$; Supplementary Fig. 2), whereas the HCC expression profiles were correlated with those of NL ($R = 0.8632$; data not shown). We also observed a strong correlation between the expression profile of the HCC used here and that of another HCC ($R = 0.8027$; data not shown) [17]. These findings suggest that the expression profile of ICC differed from those of HCC and NL.

To eliminate as much sequence error as possible, only tags with two or more hits were selected from the ICC library, resulting in 3707 tags. When we searched for these tags in the NCBI SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE/), we found that 1348 (36.4%) corresponded to single known genes, 341 (9.2%) corresponded to single expressed sequence tags (EST) and 1740 (46.9%) had multiple matches. The remaining 278 tags (7.5%) did not correspond to any known gene.

3.2. Gene signatures of ICC

We compared the transcript abundance in the ICC and HCC libraries without excluding tags with one hit and found that the levels of expression of 1898 genes were enhanced over fivefold in ICC. As expected, several genes, previously reported to be upregulated in ICC, were amongst the 20 genes that showed the greatest degree of upregulation (Table 3), including *KRT7*, *KRT19* and *S100 calcium-binding protein A6* (*S100A6*) [22]. We also identified several genes not previously well characterised as associated with ICC, such as *BGN* and *IGFBP5*. Some of the ICC-specific genes were not detected in the NL library, suggesting that these gene signatures may be specific to cholangiocytes.

To determine the molecular functions of the 1898 genes that were overexpressed in ICC than in HCC, we analysed all the SAGE tags representing these genes in the MetaCore™ from GeneGo Inc. (Supplementary Table 2). The biological processes of these genes included the regulation of cell adhesion molecules, translation initiation and the regulation of wingless-type MMTV integration site (Wnt) signalling ($P < 0.01$; statistical significance calculated using the basic equation in Supplementary Fig. 1).

We also found that 702 genes were underexpressed over fivefold in ICC relative to HCC without excluding tags with one hit; the 20 genes that showed the greatest degree of downregulation in ICC are listed in Supplementary Table 3. Amongst these were genes encoding apolipoprotein, fibronectin and haptoglobin, which were also underexpressed in ICC relative to NL. The biological processes associated with these genes (Supplementary Table 4) included immune response and the

Table 3
Twenty genes overexpressed in ICC compared to HCC

SAGE Tag	RefSeq ID	Gene name	Tag count			Fold (ICC/HCC)
			ICC	HCC	NL	
GACGCCGAAC	NM_133467	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4	111	1	9	122.00
CCTGGTCCCA	NM_005556	Keratin 7	103	0	0	113.00
AATAGAAATT	NM_000582	Secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T lymphocyte activation 1)	83	1	0	92.00
CCCCCTGGAT	NM_014624	S100A6 S100 calcium-binding-protein A6 (calcyclin)	66	0	0	73.00
CTTCCAGCTA	NM_004039	Annexin A2	55	1	3	61.00
GCAATCCTGT	NM_006998	Secretagogin, EF hand calcium-binding protein	52	0	0	58.00
GACATCAAGT	NM_002276	Keratin 19	50	0	0	55.00
GCAAAGAAAA	NM_006769	LIM domain only 4	44	0	3	49.00
CAGGCCCCAC	NM_005620	S100 calcium-binding protein A11 (calgizzarin)	41	1	0	46.00
GGAGACTTCC	NM_001153	Annexin A4	36	0	3	39.00
TGGCCCCACC	NM_002654	Pyruvate kinase, muscle	36	1	3	39.00
GCATTTGACA	NM_001046	Solute carrier family 12 (sodium/potassium/chloride transporters), member 2	33	0	0	36.00
AACTTGGCCA	NM_002423	Matrix metalloproteinase 7 (matrilysin, uterine)	30	1	0	33.00
AGGTCCTAGC	NM_000852	Glutathione-*S*-transferase pi	30	0	6	33.00
ATCTTTCTGG	NM_003406	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	30	1	0	33.00
TGATGTCTGG	NM_020182	Transmembrane, prostate androgen induced RNA	30	0	0	33.00
GACCGCAGGA	NM_001845	Collagen, type IV, alpha 1	27	1	0	30.00
GCCTGTCCCT	NM_001711	Biglycan	27	1	6	30.00
GGAACAAACA	NM_013230	CD24 molecule	27	0	0	30.00
TATGAATGCT		No reliable match	27	0	0	30.00

SAGE, serial analysis of gene expression; ICC, intrahepatic cholangiocarcinoma; HCC, hepatocellular carcinoma; NL, normal liver.

regulation of inflammation via complement activation ($P < 0.01$). These results suggest that the gene expression pattern of ICC is different from that of HCC.

Amongst the genes that were upregulated more than 20 times in the ICC library compared with the HCC library, we selected nine that were not previously well characterised to be overexpressed in ICC (Table 4) for further analysis. We defined these genes as the candidate markers of ICC and compared the SAGE libraries originating from extrahepatic adenocarcinoma (Supplementary Table 5). By comparing the mean transcript abundance in ICC with those in the other libraries, excluding tags with one hit, we found that eight genes (*CITED4*, *BGN*, *IGFBP5*, *CLDN4*, *PFKP*, *TM4SF1*, *CAPN1* and *CLDN10*) showed a threefold greater expression in ICC than in at least three other organs. These genes appear to be useful not only for the differentiation of ICC and HCC, but also for the differentiation of ICC and metastatic liver cancer.

3.3. RT-PCR and protein expression analyses

The above findings are based on the library analysis of one ICC sample; thus, their reliability must be validated with multiple samples. Moreover, the application of SAGE results to routine diagnostic examination is indispensable. We performed RT-PCR using RNA samples isolated from seven ICC, four NL, 20 CLD, 17 HCC and five extrahepatic adenocarcinomas (breast, colon, stomach, ovary and lung) to validate the expression data for the eight genes mentioned above. Of these eight genes, three (*BGN*, *IGFBP5* and *CLDN4*) were more highly expressed in ICC than in NL, CLD, HCC or extrahepatic adenocarcinomas ($P < 0.05$; Fig. 2; Supplementary Table 1), suggesting that these three genes are the specific gene signatures for ICC.

Of *BGN*, *IGFBP5* and *CLDN4*, the expression of CLDN4 was confirmed by immunoblotting (Fig. 3A) and immunohistochemical analyses (Fig. 3B and C). Immunoblotting revealed strong CLDN4 expression in six of seven patients with ICC and KMBC. In contrast, slight or no expression was detected in HCC and NL (Fig. 3A). Immunohistochemical analysis showed intense membrane staining in ICC, whereas no CLDN4 expression was observed in HCC (Fig. 3B and C).

3.4. Discrimination analysis

Based on the expression patterns of *BGN*, *IGFBP5* and *CLDN4* in the 53 samples of the training group, we performed discrimination analysis to obtain coefficients for these three genes, which were then used to

Table 4
Nine genes not previously well characterised and showing >20-fold overexpression from the ICC library compared to the HCC library

RefSeq ID	Gene name	Tag count		Ratio of RT-PCR					P value[a]
		ICC	HCC	ICC	NL	CLD	HCC	Other	
NM_133467	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4	122	1	27.56	1.37	1.81	1.13	0.13	0.0305
NM_000852	Glutathione-*S*-transferase pi	33	0	10.61	2.33	2.54	1.8	1.4	–
NM_001711	Biglycan	30	1	5.4	1.06	1.19	0.68	0.35	<0.0001
NM_000599	Insulin-like growth factor-binding protein 5	24	0	11.13	1.86	0.68	0.83	0.21	<0.0001
NM_001305	Claudin-4	24	1	344.98	1.71	7.92	8.56	0.44	0.0036
NM_002627	Phosphofructokinase, platelet	24	0	160.68	0.5	0.13	0.08	6.16	0.0089
NM_014220	Transmembrane 4 superfamily member 1	24	0	4,687.43	25.12	37.64	498.02	0.44	0.1882
NM_005186	Calpain 1, (μ/l) large subunit	21	1	4.32	1.33	0.92	0.85	5.76	<0.0001
NM_006984	Claudin-10	21	0	6.25	0.56	2.69	0.11	21.26	0.0638

ICC, intrahepatic cholangiocarcinoma; NL, normal liver; CLD, chronic liver disease; HCC, hepatocellular carcinoma; Other, extrahepatic adenocarcinoma; ICC, intrahepatic cholangiocarcinoma; HCC, hepatocellular carcinoma.
[a] One-way ANOVA.

define the *Z*-score as positive in ICC (Fig. 4A). The prediction performance of the discrimination analysis showed an efficient ROC curve (AUC = 0.987) (Fig. 4B). To verify the validity of this classifier, we calculated the *Z*-score for the 21 samples in the test group and found that the score was positive for all the ICC samples (Fig. 4C). Thus, we could distinguish ICC from the other cancers using this classifier.

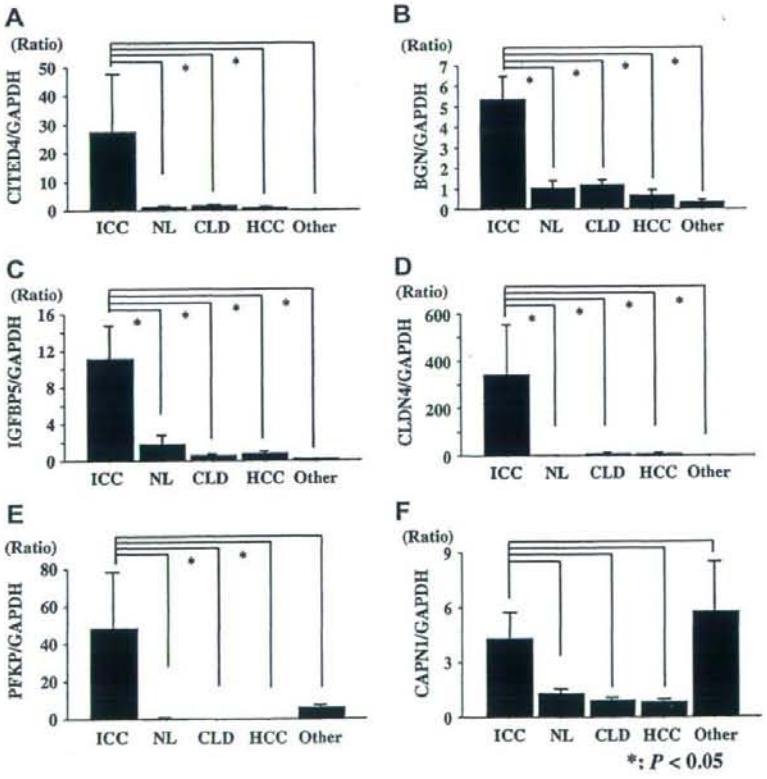


Fig. 2. Results of RT-PCR for intrahepatic cholangiocarcinoma (ICC), normal liver (NL), chronic liver disease (CLD), hepatocellular carcinoma (HCC) and other adenocarcinomas. GAPDH, glyceraldehyde-3-phosphate dehydrogenase; CITED4, Cbp/p300-interacting transactivator with Glu/Asp-rich carboxy-terminal domain 4; BGN, biglycan; IGFBP5, insulin-like growth factor-binding protein 5; CLDN4, claudin-4; PFKP, platelet-type phosphofructokinase; CAPN1, calpain 1 (μ/l) large subunit.

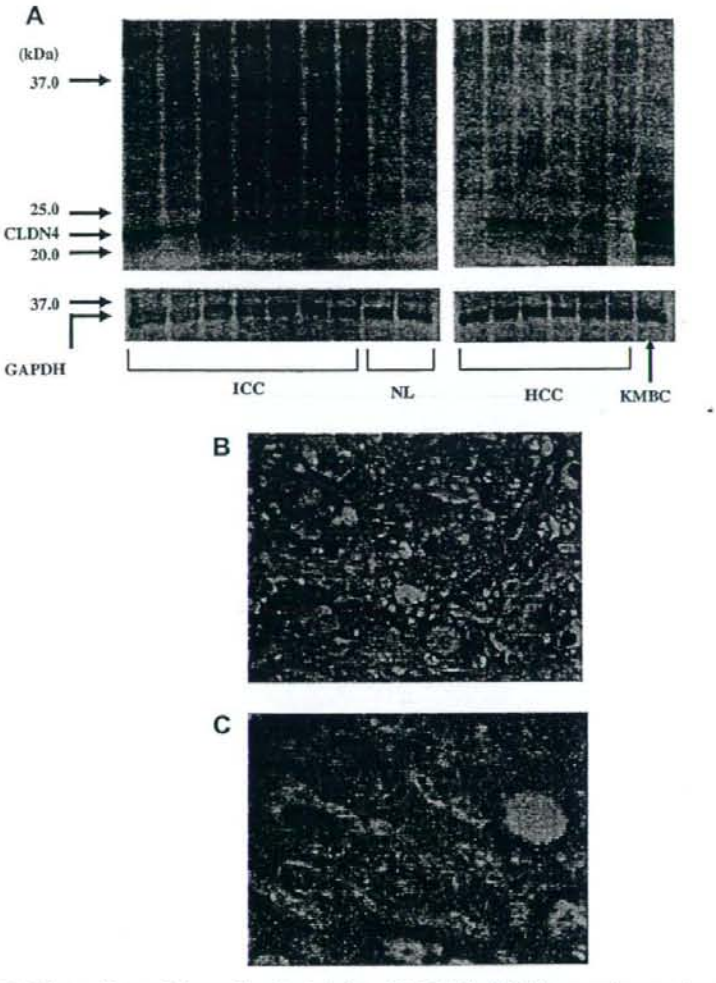


Fig. 3. Representative CLDN4 Western blots and immunohistochemical data. (A) CLDN4 (25 kDa) expression was strong in the intrahepatic cholangiocarcinoma (ICC) samples. The molecular mass marker (kDa) is indicated. (B) Intense membrane staining in intrahepatic cholangiocarcinoma. (C) Hepatocellular carcinoma showing no CLDN4 expression. GAPDH, glyceraldehyde-3-phosphate dehydrogenase; CLDN4, claudin-4; NL, normal liver; HCC, hepatocellular carcinoma; KMBC, a human extrahepatic bile duct carcinoma cell line [21]. [This figure appears in colour on the web.]

4. Discussion

Using gene expression profiling, we identified a distinct gene signature for ICC. We chose the SAGE technique, which is used to comprehensively identify and quantify the expression of known and unknown genes to analyse the gene expression in ICC. In SAGE methods, gene expressions of different organs or tissues are comparable because the tag sequence directly reflects RNA copy number, and the global standardised methods of SAGE enable the analysis and comparison of the data obtained from different laboratories through SAGE maps (http://www.ncbi.nlm.nih.gov/SAGE/). Another advantage of SAGE is that it allows the identification of genes not previously reported, including expressed sequence tags (ESTs). In fact, we detected some ESTs that were overexpressed in ICC compared to HCC (data not shown) and these novel sequences may represent new tumour markers for ICC.

When we compared the ICC and HCC libraries, *KRT7* and *KRT19* were amongst the genes showing overexpression in ICC. Both are expressed in bile duct epithelium, indicating that ICC arises from bile duct epithelial cells. In addition, we observed highly abundant

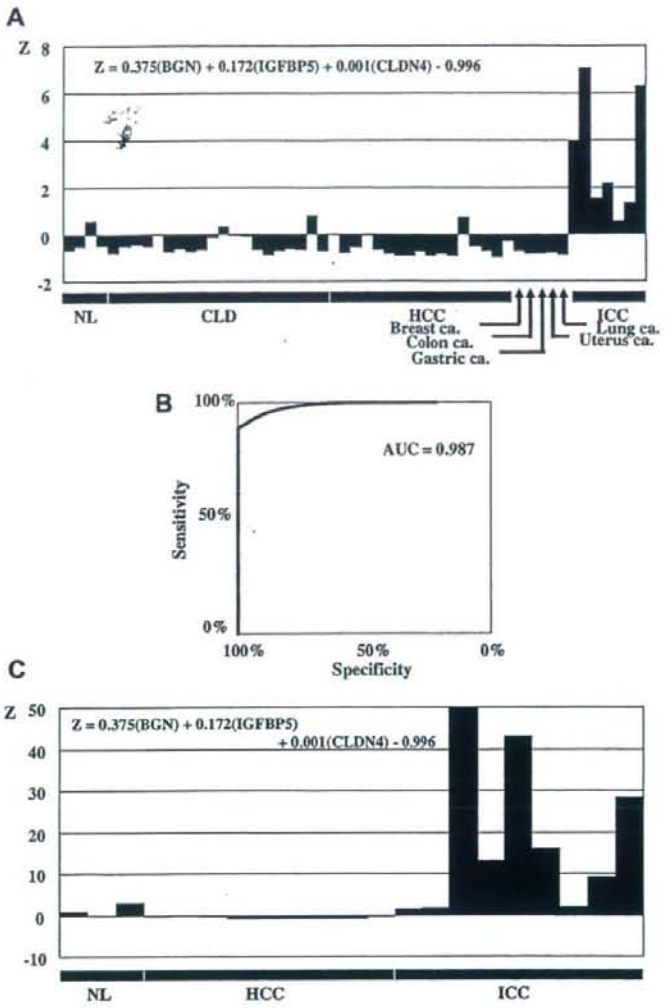


Fig. 4. Discrimination analysis of gene expression data. From the coefficient obtained for each gene, the *Z*-score was calculated as $Z = 0.375$ (BGN) + 0.172 (IGFBP5) + 0.001 (CLDN4) − 0.996. This score was positive for the intrahepatic cholangiocarcinoma (ICC) sample but negative for the other samples. (A) Results from the samples of training group. (B) Prediction performance of *Z*-score using the ROC curve. (C) Results from samples of the test group. BGN, biglycan; IGFBP5, insulin-like growth factor-binding protein 5; CLDN4, claudin-4; NL, normal liver; CLD, chronic liver disease; HCC, hepatocellular carcinoma; AUC, area under the curve.

expression in ICC of the genes encoding S100A6 and matrix metalloproteinase 7 [22–24], both previously reported in ICC. Therefore, our findings confirm the expression patterns reported previously for ICC.

We also analysed the molecular functions of the genes showing an over fivefold higher expression in ICC than in HCC and found that many of them were associated with the regulation of cell adhesion molecules, translation initiation and the regulation of Wnt signalling. Conversely, amongst the genes with an over fivefold overexpression in HCC than in ICC, the majority were associated with immunity and inflammation. These findings indicate that the characteristic gene expression patterns in ICC differ from those in HCC.

In addition to tissue-specific genes, cancer-associated genes, such as those encoding annexin A2 [25–27], S100 calcium-binding protein A11 [28], glutathione-*S*-transferase pi [29], transmembrane, prostate androgen-induced RNA [30] and CD24 antigen [31] (Table 3), were also overexpressed in ICC compared to HCC.

Thus, genes that were differentially expressed between ICC and HCC do not merely reflect differences in the cell types or tissue types from which the tumours arise, but differences in the ICC and HCC tumours themselves.

Analyses using SAGE libraries showed that nine genes, which were not previously well characterised as expressed in ICC, were overexpressed in ICC than in HCC. By comparing the mean transcript abundance in ICC with those in other libraries, including gastric, colon, prostate and breast cancer, we found that eight genes were overexpressed in ICC than in the other organs: *CITED4*, *BGN*, *IGFBP5*, *CLDN4*, *PFKP*, *TM4SF1*, *CAPN1* and *CLDN10*. As these findings were based on one ICC sample, we confirmed their validity using RT-PCR and found that the expression of *BGN*, *IGFBP5* and *CLDN4* was higher in ICC than in HCC, NL, CLD or adenocarcinomas originating from other organs.

BGN encodes a matrix proteoglycan involved in the metabolism of connective tissue by binding to collagen and TGF-beta. It is expressed in the lung, spleen and liver of mice [32]. Immunohistochemically, the BGN protein has been shown to be expressed in Disse's space, cholangioles and the vascular wall, but not in normal liver [33].

IGFBP5 encodes a protein that binds to IGF-I and -II, but its precise function is not yet known. This protein has been reported to promote osteoblast mitosis [34] and to cause the involution of mammary gland cells by inducing apoptosis via IGF-dependent and -independent pathways [35]. However, its function in liver and bile duct cells is unknown.

CLDN4 encodes an essential membrane protein of the claudin family. This protein is a component of tight junctions, a membrane receptor for *Clostridium perfringens* enterotoxin, and is thought to be involved in organogenesis. In addition, CLDN4 is highly expressed in the small intestine, but only weakly expressed in mouse liver [36]. Whilst our study was under way, Lodi et al. [14] reported that the level of CLDN4 expression was higher in biliary tract cancers than in HCC.

Other candidate genes more suitable for the differentiation of ICC from HCC and metastatic liver cancers may exist, but our three genes – BGN, IGFBP5 and CLDN4 – were not previously reported to be overexpressed in ICC, although CLDN4 was cited whilst our study was under way.

Moreover, the prediction performance of ICC discrimination analysis using these three genes showed an efficient ROC curve (AUC = 0.987) and the equation allowed us to completely distinguish ICC from HCC, CLD and NL, although further evaluation is needed using, for example, metastatic liver cancers. The ROC curve was also generated for the test group and the AUC (0.963) was as high as that generated using the training set (data not shown). The discrimination of ICC using the known markers such as CEA and CA19-9 was less effective than this model. The AUC for CEA and CA19-9 was 0.676 and 0.722, respectively (data not shown).

However, the inclusion of other ICC markers will enhance the sensitivity and specificity; for example, both CEA and CA19-9 were enhanced in 50% and 62% of patients with ICC, respectively (Table 2). In addition, we found no correlation between the level of CEA or CA19-9 and the expression of the three genes, suggesting that they are independent markers of ICC. Thus, a combined evaluation of these markers could enhance sensitivity and specificity.

Despite the limited numbers of patients with ICC and SAGE libraries examined, this is the first report of ICC-specific, novel candidate tumour-marker genes. Further validation using a larger cohort is needed to confirm the expression of these genes in more patients with ICC and to examine the functional relevance of these genes in the development and prognosis of ICC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhep.2008.03.025.

References

[1] Shaib Y, El-Serag HB. The epidemiology of cholangiocarcinoma. Semin Liver Dis 2004;24:115–125.
[2] Patel T. Increasing incidence and mortality of primary intrahepatic cholangiocarcinoma in the United States. Hepatology 2001;33:1353–1357.
[3] Taylor-Robinson SD, Toledano MB, Arora S, Keegan TJ, Hargreaves S, Beck A, et al. Increase in mortality rates from intrahepatic cholangiocarcinoma in England and Wales 1968–1998. Gut 2001;48:816–820.
[4] Khan SA, Taylor-Robinson SD, Toledano MB, Beck A, Elliott P, Thomas HC. Changing international trends in mortality rates for liver, biliary and pancreatic tumours. J Hepatol 2002;37:806–813.
[5] Davila JA, El-Serag HB. Cholangiocarcinoma: the "other" liver cancer on the rise. Am J Gastroenterol 2002;97:3199–3200.
[6] Marrero JA, Lok AS. Newer markers of hepatocellular carcinoma. Gastroenterology 2004;127:S113–S119.
[7] Johnson DE, Herndier BG, Medeiros LJ, Warnke RA, Rouse RV. The diagnostic utility of the keratin profiles of hepatocellular carcinoma and cholangiocarcinoma. Am J Surg Pathol 1988;12:187–197.
[8] Lai YS, Thung SN, Gerber MA, Chen ML, Schaffner F. Expression of cytokeratins in normal and diseased livers and in primary liver carcinomas. Arch Pathol Lab Med 1989;113:134–138.
[9] Maeda T, Adachi E, Kajiyama K, Sugimachi K, Tsuneyoshi M. Combined hepatocellular and cholangiocarcinoma: proposed criteria according to cytokeratin expression and analyses of clinicopathologic features. Hum Pathol 1995;26:956–964.
[10] Altmannsberger M, Weber K, Holscher A, Schauer A, Osborn M. Antibodies to intermediate filaments as diagnostic tools: human

gastrointestinal carcinomas express prekeratin. Lab Invest 1982;46:520–526.
[11] Saintigny P, Coulon S, Kambouchner M, Ricci S, Martinot E, Danel C, et al. Real-time RT-PCR detection of CK19, CK7 and MUC1 mRNA for diagnosis of lymph node micrometastases in non small cell lung carcinoma. Int J Cancer 2005;115:777–782.
[12] Khan SA, Davidson BR, Goldin R, Pereira SP, Rosenberg WM, Taylor-Robinson SD, et al. Guidelines for the diagnosis and treatment of cholangiocarcinoma: consensus document. Gut 2002;51:VI1–VI9.
[13] Gores GJ. Cholangiocarcinoma: current concepts and insights. Hepatology 2003;37:961–969.
[14] Lódi C, Szabó E, Holczbauer A, Batmunkh E, Szijártó A, Kupcsulik P, et al. Claudin-4 differentiates biliary tract cancers from hepatocellular carcinomas. Modern Pathol 2006;19:460–469.
[15] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
[16] Yamashita T, Hashimoto S, Kaneko S, Nagai S, Toyoda N, Suzuki T, et al. Comprehensive gene expression profile of a normal human liver. Biochem Biophys Res Commun 2000;269:110–116.
[17] Yamashita T, Kaneko S, Hashimoto S, Sato T, Nagai S, Toyoda N, et al. Serial analysis of gene expression in chronic hepatitis C and hepatocellular carcinoma. Biochem Biophys Res Commun 2001;282:647–654.
[18] Kawai HF, Kaneko S, Honda M, Shirota Y, Kobayashi K. Alpha-fetoprotein-producing hepatoma cell lines share common expression profiles of genes in various categories demonstrated by cDNA microarray analysis. Hepatology 2001;33:676–691.
[19] Shirota Y, Kaneko S, Honda M, Kawai HF, Kobayashi K. Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. Hepatology 2001;33:832–840.
[20] Makuuchi M, Belghiti J, Belli G, Fan ST, Lau JW, Ringe B, et al. IHPBA concordant classification of primary liver cancer: working group report. J Hepatobiliary Pancreat Surg 2003;10:26–30.
[21] Yano H, Maruiwa M, Iemura A, Mizogichi A, Kojiro M. Establishment and characterization of a new human extrahepatic bile duct carcinoma cell line (KMBC). Cancer 1992;69:1664–1673.
[22] Kim J, Kim J, Yoon S, Joo J, Lee Y, Lee K, et al. S100A6 protein as a marker for differential diagnosis of cholangiocarcinoma from hepatocellular carcinoma. Hepatol Res 2002;23:274–286.
[23] Lichtinghagen R, Helmbrecht T, Árndt B, Böker KH. Expression pattern of matrix metalloproteinases in human liver. Eur J Clin Chem Clin Biochem 1995;33:65–71.
[24] Miwa S, Miyagawa S, Soeda J, Kawasaki S. Matrix metalloproteinase-7 expression and biologic aggressiveness of cholangiocellular carcinoma. Cancer 2002;94:428–434.
[25] Kumble KD, Hirota M, Pour PM, Vishwanatha JK. Enhanced levels of annexins in pancreatic carcinoma cells of Syrian hamsters and their intrapancreatic allografts. Cancer Res 1992;52:163–167.
[26] Emoto K, Sawada H, Yamada Y, Fujimoto H, Takahama Y, Ueno M, et al. Annexin II overexpression is correlated with poor prognosis in human gastric carcinoma. Anticancer Res 2001;21:1339–1345.
[27] Emoto K, Yamada Y, Sawada H, Fujimoto H, Ueno M, Takayama T, et al. Annexin II overexpression correlates with stromal tenascin-C overexpression: a prognostic marker in colorectal carcinoma. Cancer 2001;92:1419–1426.
[28] Ohuchida K, Mizumoto K, Ohhashi S, Yamaguchi H, Konomi H, Nagai E, et al. S100A11, a putative tumour suppressor gene, is overexpressed in pancreatic carcinogenesis. Clin Cancer Res 2006;12:5417–5422.
[29] Cai L, Mu LN, Lu H, Lu QY, You NC, Yu SZ, et al. Dietary selenium intake and genetic polymorphisms of the GSTP1 and p53 genes on the risk of esophageal squamous cell carcinoma. Cancer Epidemiol Bio Prev 2006;15:294–300.
[30] Xu LL, Shi Y, Petrovics G, Sun C, Makarem M, Zhang W, et al. PMEPA1, an androgen-regulated NEDD4-binding protein, exhibits cell growth inhibitory function and decreased expression during prostate cancer progression. Cancer Res 2003;63:4299–4304.
[31]] Karahan N, Güney M, Oral B, Kapucuoglu N, Mungan T. CD24 expression is a poor prognostic marker in endometrial carcinoma. Eur J Gynaecol Oncol 2006;27:500–504.
[32] Wegrowski Y, Pillarisetti J, Danielson KG, Suzuki S, Iozzo RV. The murine biglycan: complete cDNA cloning, genomic organization, promoter function, and expression. Genomics 1995;30:8–17.
[33] Hogemann B, Edel G, Schwarz K, Krech R, Kresse H. Expression of biglycan, decorin and proteoglycan-100/CSF-1 in normal and fibrotic human liver. Pathol Res Pract 1997;193:747–751.
[34] Schneider MR, Wolf E, Hoeflich A, Lahm H. IGF-binding protein-5: flexible player in the IGF system and effector on its own. J Endocrinol 2002;172:423–440.
[35] Butt AJ, Dickson KA, McDougall F, Baxter RC. Insulin-like growth factor-binding protein-5 inhibits the growth of human breast cancer cells in vitro and in vivo. J Biol Chem 2003;278:29676–29685.
[36] Katahira J, Sugiyama H, Inoue N, Horiguchi Y, Matsuda M, Sugimoto N. *Clostridium perfringens* enterotoxin utilizes two structurally related membrane proteins as functional receptors in vivo. J Biol Chem 1997;272:26652–26658.

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 $L = 2$.

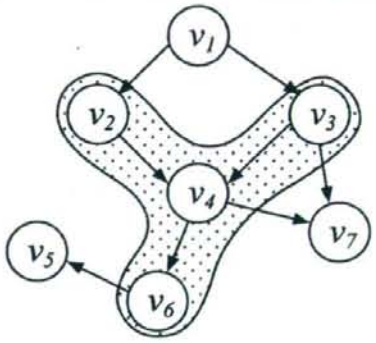
Problem P1 maximises the number of covered nodes by using K reporters, and is defined as follows.

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, $OPT(I') \leq 2OPT(I)$.

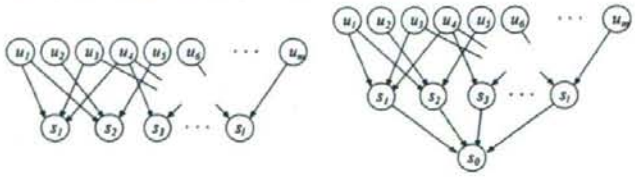
Given any solution $R \subseteq V$ of I' with cost (i.e., the number of covered nodes) c', we produce a solution C of I in polynomial time by letting $C = R - U$, where $R - U = \{r | r \in R \text{ and } r \notin U\}$. Then, $|C| \leq |R| \leq k$. Let c be the cost (i.e., the number of covered elements)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

$$\text{(IP1)} \quad \text{Maximise} \quad \sum_{i=1}^{n} y_i,$$

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 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

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 (Wu et al., 1997). Bcl-2 superfamily regulates cell death that is amplified via the mitochondrial pathway (Sprick and Walczak, 2004). BAX may be related with possible amplification of apoptosis via the intrinsic pathway in response to JNK. The caspase-9 (CASP9) may be essential for border-cell migration in the Drosophila ovary (Geisbrecht and Montell, 2004), 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 (Fischer et al., 2003).

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

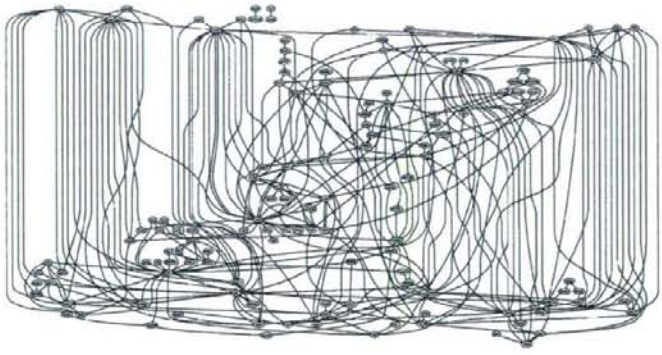


Table 1 The optimal solution of IP1 and IP2 for each L and K in apoptosis pathway maps, where the numbers of covered nodes and the numbers of the selected reporters are shown for IP1 and IP2, respectively

	IP1 for each K							
L	*1*	*2*	*3*	*4*	*5*	*6*	*IP2*	*Reporter in K = 1 (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)

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 $K = 2, \ldots, 4$. 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 $K - 1$. Actually, for $L = 1$ the greedy algorithm selected 44 reporters to cover all nodes of V although only 42 reporters are required as we see from the result of IP2. On the other hand, our integer programming-based methods can always find optimal solutions if any. For each case, the elapsed time of optimising IP1 or IP2 was at most 0.023 seconds. These results suggest that our methods are practical.

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

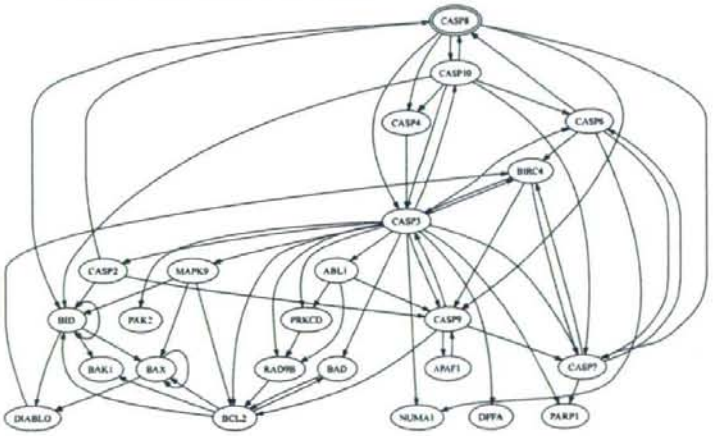
5.1.1 Effects of specific nodes

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 (Lamkan et al., 2007), 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 $L = 4, \ldots, n(= 23)$, five proteins without outward edges were selected as the optimal reporter nodes in IP2.

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



5.2 *Artificial scale-free networks*

It is known that many real biological networks have the scale-free property (Barabási and Albert, 1999). The degree distribution, $P(k)$, of a scale-free network follows a power-law relationship ($P(k) \propto k^{-\gamma}$). In the network, most nodes have one connection, and a few nodes have many connections. In particular, it is observed that gene regulatory networks have the power-law outdegree distribution and the Poisson indegree distribution (Guelzim et al., 2002). Thus, we generated scale-free networks with power-law outdegree distributions 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. Since it is known that the exponent $\gamma = 2 \sim 3$ for many real biological networks, we performed experiments for $\gamma = 2.0, 2.2, 2.5$, and 3.0. For large $n(= 1000, 5000, 10000)$, the elapsed