Available online at www.sciencedirect.com
ScienceDirect
Toxicology Letters 167 (2006) 106–113
Toxicology Letters
www.elsevier.com/locate/toxlet

Altered gene expression of transcriptional regulatory factors in tumor marker-positive cells during chemically induced hepatocarcinogenesis

Shigehiro Osada[a,b,*,1], Ayako Naganawa[a,1], Masashi Misonou[a], Soken Tsuchiya[c,d], Shigero Tamba[c], Yasushi Okuno[d,e], Jun-ichi Nishikawa[a], Kimihiko Satoh[f], Masayoshi Imagawa[b], Gozoh Tsujimoto[e], Yukihiko Sugimoto[c], Tsutomu Nishihara[a]

[a] *Laboratory of Environmental Biochemistry, Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamada-Oka, Suita, Osaka 565-0871, Japan*
[b] *Department of Molecular Biology, Graduate School of Pharmaceutical Sciences, Nagoya City University, 3-1 Tanabe-dori, Mizuho-ku, Nagoya, Aichi 467-8603, Japan*
[c] *Department of Physiological Chemistry, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*
[d] *Department of Pharmacoinformatics, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*
[e] *Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*
[f] *Department of Organic Function, Hirosaki University, School of Health Science, Hon-Cho 66-1, Hirosaki 036-8564, Japan*

Received 10 August 2006; received in revised form 29 August 2006; accepted 29 August 2006
Available online 25 September 2006

Abstract

Glutathione-*S*-transferase placental form (GST-P) is markedly and specifically inducible in rat chemical hepatocarcinogenesis and is a reliable marker protein for pre-neoplasia. To gain insights into the molecular mechanisms at the early stage of hepatocarcinogenesis and hepatotoxicity, we investigated the gene expression profile by DNA microarray analysis. We prepared RNA from GST-P-positive foci in three individual rats and compared with normal liver sections from three individual rats, and labeled RNA was individually hybridized onto Affymetrix GeneChip Rat Expression Array 230A. DNA microarray analysis showed distinctly different profiles of dysregulated gene expression and supported the previous finding that some enzymes involved in metabolism and detoxification are overexpressed and suppressed. Here we discovered that several DNA-binding transcription factors and cofactors, including sterol-regulatory-element binding protein 1 (SREBP1) and Wilms' tumour 1 (WT1)-interacting protein, and their target genes were dysregulated in GST-P-positive foci. Moreover, genes involved in chromatin components, histone modification enzymes, and centrosome duplication were highly expressed. These genes were not previously known to be up-regulated during chemically induced hepatocarcinogenesis. DNA microarray analysis using RNA prepared from tumor marker-positive foci and control tissues provided a candidate gene link to the early stage of carcinogenesis and hepatotoxicity.

Keywords: Chemical hepatocarcinogenesis; Tumor marker; Gene expression; Transcription factor; Chromatin; Histone

* Corresponding author at: Department of Molecular Biology, Graduate School of Pharmaceutical Sciences, Nagoya City University, 3-1 Tanabe-dori, Mizuho-ku, Nagoya, Aichi 467-8603, Japan. Tel.: +81 52 836 3456; fax: +81 52 836 3456.
E-mail address: osada@phar.nagoya-cu.ac.jp (S. Osada).
[1] These authors contributed equally to this work.

0378-4274/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.toxlet.2006.08.014

1. Introduction

Rat glutathione-*S*-transferase placental form (GST-P) is a phase II detoxification enzyme and its expression is completely repressed in normal liver. GST-P is also a well-known tumor marker that is specifically induced during chemical hepatocarcinogenesis in rats (Sato, 1989; Satoh et al., 1985). GST-P expressed single cells are detected in the liver after treatment of diethylnitrosamine (DEN) and might be precursors of preneoplastic foci and nodules (Satoh et al., 1989; Satoh et al., 2005). Hepatocyte nodules in six models of liver carcinogenesis were analyzed and the amount of GST-P was elevated in all types of nodules (Eriksson et al., 1983). Measurement of GST-P-positive foci is rapid detection of carcinogenic agents in the medium-term rat liver bioassay (Ito test), which is considered to be a reliable tool for prediction of promoting or reducing activity of chemicals on hepatocarcinogenesis. Over 300 chemicals have already observed and this test was recommended as an alternative to long-term carcinogenicity testing at the International Conference on Harmonization (ICH) (Ito et al., 2003). GST-P-positive foci are induced by not only DEN but many chemicals. Gamma-glutamyltranspeptidase is also expressed in GST-P-positive foci derived from precursor cells and GST-P-positive foci are important for detoxification for carcinogen (Satoh et al., 2005). Carcinogenic activity of nitroso compounds, including DEN, is well studied, but the mechanisms of hepatotoxicity of these compounds are poorly understood. Analysis of GST-P-positive foci is valuable for the understanding of the molecular mechanisms of hepatocarcinogenesis, detoxification and hepatotoxicity.

Transgenic rats using the regulatory element of the GST-P gene revealed that a gene involved in liver cell transformation is not physically linked with the GST-P gene, but the expression is regulated by common transcription factors (Morimura et al., 1993). Further, we identified the enhancer element responsible for tumor-specific expression of the GST-P gene (Sakai and Muramatsu, 2005; Suzuki et al., 1995). This indicates that analysis of the expression profile in GST-P-positive foci leads to the identification of the responsible gene for liver cancer and understanding the mechanism of hepatocarcinogenesis.

Carcinogenesis is a genetic and epigenetic disease arising from multiple molecular changes and these events lead to changes in gene expressions. Recently, it was reported that specific differences in the gene expression profile revealed by cDNA microarray analysis of GST-P-positive foci and the surrounding tissue, and metabolic enzymes were found as up- and down-regulated genes (Suzuki et al., 2004). Direct comparisons of gene expressions between normal liver and chemically induced preneoplastic foci provide more useful information related to the molecular mechanisms of carcinogenesis. Although genes involved in transcriptional regulation are one of the most important factors in carcinogenesis, their expression levels are generally lower than those of metabolic enzymes and are hard to evaluate by DNA microarray analysis.

In this study, we conducted microarray analysis of mRNA from GST-P-positive foci in three individual rats and compared with normal liver sections from three individual rats. Labeled RNA was individually hybridized onto GeneChip Rat Expression Array 230A, and several differentially expressed genes were found to be involved in transcriptional regulation, which were not previously known to be regulated during chemically induced hepatocarcinogenesis.

2. Materials and methods

2.1. Chemical hepatocarcinogenesis of rats

Carcinogenic experiments were done according to the Solt–Farber protocol (Solt and Farber, 1976). Experiments were initiated by intraperitoneal injection of DEN (200 mg/kg) (Wako Pure Chemical Industries, Ltd., Osaka, Japan) into 5-week-old Sprague–Dawley rats. After the animals had been fed basal diets for 2 weeks, they were changed to basal diets containing 0.02% 2-acetylaminofluorene (Nacalai Tesque, Kyoto, Japan). Three weeks after DEN injection, partial hepatectomy was performed and livers were extirpated 8 weeks after DEN injection. Control rats were injected with saline and fed basal diets. All animal care and handling procedures were approved by the animal care and use committee of Osaka University.

2.2. Preparation of RNA from rat liver

To map the exact location of GST-P-positive foci, one of the serial frozen sections (10 μm) from the liver was treated with rabbit anti-GST-P antibody and immunohistochemical staining was performed with the DAKO ENVISION System (DAKO Co., Tokyo, Japan). RNA was prepared from the area corresponding to GST-P-positive foci in hyperplastic nodules induced in three individual rats or sections from three control rats by RNeasy Mini Kit (QIAGEN, Hilden, Germany).

2.3. Oligonucleotide microarray and data analysis

Target RNA amplification and labeling with biotinylated nucleotides were carried out using MEGAscript T7 Kit (Ambion, Austin, TX) and Enzo BioArray High Yield RNA Transcript Labeling Kit (Enzo Diagnostics, Farmingdale, NY) as specified by the manufacturer. The quality and

size distribution of the targets were determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Labeled and fragmented RNA of individual rats was hybridized onto GeneChip Rat Expression Array 230A (Affymetrix, Santa Clara, CA) using standard methods. We calculated the background correction and normalization of the array data using Robust Multi-Array (RMA) method in the R package. Statistics of differential expression between genes was estimated using the linear modeling features of the limma library of the R. Limma computes p-values of moderated t-statistics by emprical Bayes shrinkage of the standard error toward a common value.

3. Results and discussion

3.1. Detection of GST-P-positive foci

To examine the expression profile of GST-P-positive foci during hepatocarcinogenesis, hyperplastic nodule-induced rats were prepared according to the Solt–Farber procedure (Solt and Farber, 1976). Eight weeks after DEN treatment, the livers, which had a large number of foci and nodules, were excised and immunohistochemical experiments indicated that an approximately 70–80% region contained GST-P-positive foci (data not shown).

3.2. DNA microarray analysis of gene expression

We prepared biotinylated target RNA from GST-P-positive foci and normal liver sections in three hyperplastic nodule-induced rats and three control rats, respectively. Each target was individually hybridized with the Rat 230A Array containing the primary probe sets against well-annotated full-length genes. The scatter plot of the gene expression pattern between three independent control rats showed excellent reproducibility of results with an average correlation coefficient ± S.D. (0.93 ± 0.035). In the case of GST-P-positive foci, good reproducibility was also obtained (average of correlation coefficient ± S.D., 0.95 ± 0.0068). On the other hand, the average of the correlation coefficient ± S.D. derived from comparisons of control versus GST-P-positive foci was 0.74 ± 0.026. These results indicate that expression profiles in the same groups were indistinguishable, but dysregulation in many genes was observed during hepatocarcinogenesis.

3.3. Expression profile of enzymes involved in metabolism and detoxification

Genes were examined in which the expression was enhanced or reduced in GST-P-positive foci compared with control liver. Significantly changed transcripts were selected by moderated t-statistics. There were 15,923 probes on the chip, and 375 and 199 genes were significantly up- and down-regulated, respectively, with log ratio values outside of 1 to -1 ($p<0.05$). Of these, the twenty most up- and down-regulated genes are shown in Tables 1 and 2 together with p-values for statistical significance. Significant up-regulation of the GST-P gene (*Gstp1/Gstp2*) expression was observed in GST-P-positive foci (Table 1). It is known that enzymes involved in metabolism and detoxification are induced or repressed during chemical hepatocarcinogenesis (Sato, 1989; Suzuki et al., 2004). Overexpression of metabolic enzymes, which were reported to demonstrate increased expression in hyperplastic nodules, including aldehyde dehydrogenase, aflatoxin B1 aldehyde reductase, NAD(P)H dehydrogenase and glutathione peroxidase 2, and the suppression of carbonic anhydrase 3, were detected by DNA microarray analysis (Tables 1 and 2). Semi-quantitative reverse transcriptase-coupled PCR experiments were performed on several selected genes and it was confirmed that the expression patterns were similar to those observed with microarray (data not shown). These results indicate that our study would be suitable for discovering new genes to provide new information on hepatocarcinogenesis, detoxification, and hepatotoxicity.

3.4. Expression profile of transcripts involved in transcription

Probes on the chip were divided into various categories based on Gene Ontology (Ashburner et al., 2000). Observation and analysis of the expression profile for genes involved in transcription, one of the categories, provides valuable information to understand the mechanism of carcinogenesis. Transcripts categorized as transcription with significantly changed expression with log ratio values outside 1 to -1 are listed in Tables 3 and 4, and most have not previously been found to be differentially expressed during chemically induced hepatocarcinogenesis. For example, Pawr was overexpressed in GST-P-positive foci. Pawr also termed par-4, which interacts with Wilms' tumor 1 (WT1) and modulates functions of WT1 (Johnstone et al., 1996). WT1 is a sequence-specific DNA-binding protein and functions as both a tumor suppressor and an oncogenic factor (Loeb and Sukumar, 2002). The WT1 gene exerts an oncogenic function rather than a tumor-suppressor gene function in solid tumors as well as leukemias (Sugiyama, 2001). In prostate cancer cell line, ectopic expression PAWR repressed Bcl-2 expression through WT1 (Cheema et al., 2003). However, Loeb revealed that

Table 1
A list of the twenty genes most highly induced in GST-P-positive foci

Gene symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Akr1b8	Aldo-keto reductase family 1, member B8	6.65	8.40E−07	NM_173136
Yc2	Glutathione-*S*-transferase Yc2 subunit	5.38	1.32E−06	NM_001009920
Gstp1/Gstp2	Glutathione-*S*-transferase, pi 1/2	5.14	4.31E−06	NM_012577 NM_138974
Aldh1a1	Aldehyde dehydrogenase family 1, member A1	4.42	9.18E−05	NM_022407
Akr7a3	Aflatoxin B1 aldehyde reductase	4.17	4.17E−05	NM_013215
Aldh3a1	Aldehyde dehydrogenase family 3, member A1	3.89	1.07E−04	NM_031972
Nqo1	NAD(P)H dehydrogenase, quinone 1	3.72	9.79E−05	NM_017000
Serpinb 1 a_predicted	Serine (or cysteine) proteinase inhibitor, clade B, member 1a (predicted)	3.72	3.22E−02	NM_001031642
LOC294067	Similar to ww domain binding protein 5	3.71	1.92E−05	XM_215278
RDG:621458	Neurofilament, light polypeptide	3.64	8.68E−04	NM_031783
RGD1310542_predicted	Similar to RIKEN cDNA 4930457P18 (predicted)	3.59	3.13E−04	NM_001014154
–	Brain expressed X-linked 1	3.56	6.18E−04	NM_001037365
Rnf30_predicted	RING finger protein 30 (predicted)	3.54	2.56E−04	NM_001013217
Anxa2	Annexin A2	3.53	6.18E−04	NM_019905
Ca2	Carbonic anhydrase 2	3.51	4.17E−05	NM_019291
Ddit4l	DNA-damage-inducible transcript 4-like	3.45	5.30E−06	NM_080399
Gpx2	Glutathione peroxidase 2	3.41	8.23E−05	NM_183403
RGD:1303152	Ectodermal-neural cortex 1	3.31	1.36E−04	NM_001003401
Dscr1l1	Down syndrome critical region gene 1 -like 1	3.27	3.17E−05	NM_175578
LOC500040	Similar to testis-derived transcript	3.23	2.40E−04	XM_575396

Log ratio indicates a logarithm of the fold-change vs. the expression level of the control rats. Statistics of differential expression between genes was estimated using the linear modeling features of the limma library of the R. Limma computes *p*-values of moderated *t*-statistics by emprical Bayes shrinkage of the standard error toward a common value.

Table 2
A list of the twenty genes most highly repressed in GST-P-positive foci

Gene symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Pgcl4	Alpha-2u globulin PGCL4	−4.75	1.88E−02	NM_147215
Pgcl3/5/1/4	Alpha-2u globulin PGCL3/5/1/4	−4.71	1.79E−02	NM_147212 NM_147213 NM_147214 NM_147215
Pgcl4	Alpha-2u globulin PGCL4	−4.48	8.86E−03	NM_147215
Ca3	Carbonic anhydrase 3	−3.72	2.20E−02	NM_019292
Apoa4	Apolipoprotein A-IV	−3.63	1.01E−04	NM_012737
Cyp3a13	Cytochrome P450, family 3, subfamily a, polypeptide 13	−3.53	6.18E−04	NP_671739.1
Cyp2c	Cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase)	−3.50	8.20E−03	NM_019184
Ca3	Carbonic anhydrase 3	−3.45	6.62E−03	NM_019292
LOC368066	Similar to thioether S-methyltransferase	−3.25	2.21E−03	XM_347233
Fasn	Fatty acid synthase	−3.04	2.66E−03	NM_017332
Cyp2a2	Cytochrome P450, subfamily 2A, polypeptide 1	−3.03	2.28E−02	NM_012693
Sult1a2	Sulfotransferase family 1 A, member 2	−2.83	2.52E−03	NM_031732
Ust5r	integral membrane transport protein UST5r	−2.75	2.21E−02	NM_134380
Apoa2	Apolipoprotein A-II	−2.74	1.83E−03	NM_013112
Slc27a5	bile acid CoA ligase	−2.69	2.10E−03	NM_024143
–	Ab2-060	−2.68	1.36E−04	AI411138
Avpr1a	Arginine vasopressin receptor 1A	−2.53	7.61E−03	NM_053019
–	Malic enzyme 3, NADP(+)-dependent, mitochondrial (predicted)	−2.51	6.18E−04	AA964869
Thrsp	Thyroid hormone responsive protein	−2.46	2.54E−04	NM_012703
Slc21a10	Solute carrier family 21, member 10	−2.41	3.05E−02	NM_031650

Log ratio and *p*-value are described in Table 1.

Table 3
A list of genes involved in transcription induced in GST-P-positive foci

Gene Symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Rnf30_predicted	Ring finger protein 30 (predicted)	3.54	2.56E−04	NM_001013217
Copeb	Core promoter element binding protein	1.92	2.21E−02	NM_031642
Baspl	Brain acidic membrane protein	1.69	1.50E−02	NM_022300
Copeb	Core promoter element binding protein	1.66	1.70E−03	NM_031642
Htatip2_predicted	HIV-1 Tat interactive protein 2 (predicted)	1.61	6.54E−04	XM_214927
L3mbtl2_predicted	l(3)mbt-like 2 (Drosophila) (predicted)	1.59	1.29E−03	NM_001033695
Ppp2ca	Protein phosphatase 2a, catalytic subunit, alpha isoform	1.58	3.77E−03	NM_017039
Ppp2ca	Protein phosphatase 2a, catalytic subunit, alpha isoform	1.56	2.53E−03	AI009467
Maged1	Melanoma antigen, family D, 1	1.41	9.41E−03	NM_053409
Als2cr3	Amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 3 homolog (human)	1.41	4.83E−02	NM_133560
Hmgb2	High mobility group box 2	1.35	4.16E−02	XM_573272
Npm1	Nucleophosmin 1	1.34	4.18E−02	NM_012992
Pdlim1	PDZ and LIM domain 1	1.27	1.22E−02	NM_017365
Mdm2_predicted	Transformed mouse 3T3 cell double minute 2 (predicted)	1.24	2.64E−02	XM_235169
Sox4_predicted	SRY-box containing gene 4 (predicted)	1.19	2.36E−02	XM_344594
Npm1	Nucleophosmin 1	1.18	3.92E−03	NM_012992
Carm1_predicted	Coactivator-associated arginine methyltrnsferase1 (predicted)	1.14	2.07E−02	NM_001030041
Tgif_predicted	TG interacting factor predicted	1.14	3.74E−03	NM_001015020
Ivns 1 abp_predicted	Influenza virus NS1A binding protein (predicted)	1.09	1.09E−02	XM_213898
Pawr	PRKC, apoptosis, WT1, regulator	1.07	4.28E−03	NM_033485
RGD1304726_predicted	Similar to RIKEN cDNA 6330509G02 (predicted)	1.06	2.49E−02	NM_001024993
Ets2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	1.03	1.24E−02	XM_239510
Rbbp7	Retinoblastoma binding protein 7	1.02	4.15E−03	NM_031816

Log ratio and *p*-value are described in Table 1.

WT1 transcriptionally up-regulates anti-apoptotic genes such as Bcl-2 in rhavdoid cell line (Loeb, 2006; Mayo et al., 1999). In GST-P-positive foci, we found mRNA overexpression of Bcl-2 (log ratio, 0.776; $p = 0.0330$) by microarray. The different regulation mechanism of Bcl-2 expression is caused by cell lineage and isoform-specific differences in WT1 function (Loeb, 2006; Mayo et al., 1999). Further characterization of pawr would lead the

Table 4
A list of genes involved in transcription repressed in GST-P-positive foci

Gene symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Thrsp	Thyroid hormone responsive protein	−2.46	2.54E−04	NM_012703
Thrsp	Thyroid hormone responsive protein	−1.82	2.14E−02	NM_012703
Thrsp	Thyroid hormone responsive protein	−1.68	6.14E−03	NM_012703
Srebf1	Sterol regulatory element binding factor 1	−1.57	4.24E−03	XM_213329
Atf5	Activating transcription factor 5	−1.39	8.82E−03	NM_172336
Sec 14l2	SEC14-like 2 (S. cerevisiae)	−1.36	5.06E−03	NM_053801
	Protocadherin 1 (cadherin-like 1) (predicted)	−1.27	8.40E−03	XM_225997
Gls2	Liver mitochondrial glutaminase	−1.22	1.40E−02	NM_138904
Per2	Period homolog 2	−1.14	7.45E−03	NM_031678
Idb4	Inhibitor of DNA binding 4	−1.12	9.70E−03	NM_175582
Clp1	Cardiac lineage protein 1	−1.11	3.12E−02	NM_001025136
Tgfb1i4	Transforming growth factor beta 1 induced transcript 4	−1.10	5.86E−03	L25785
Rxra	Retinoid X receptor alpha	−1.02	1.24E−03	NM_012805
Hes6_predicted	Hairy and enhancer of split 6 (Drosophila) (predicted)	−1.01	5.41E−03	NM_001013179

Log ratio and *p*-value are described in Table 1.

understanding of oncogenic or tumor suppressor gene function of WT1 during hepatocarcinogenesis.

On the other hand, several sequence-specific DNA-binding transcription factors were repressed during hepatocarcinogenesis (e.g. Sterol-regulatory-element binding factor 1 (Srebf1)/Sterol-regulatory-element binding protein 1 (Srebp1) and retinoid X receptor alpha (RXRalpha)) (Table 4). SREBPs have been established as lipid synthetic transcription factors for cholesterol and fatty acid synthesis (Eberle et al., 2004). The expression of fatty acid synthase and apolipoprotein A-II are mainly regulated by SREBP1, and these genes were suppressed in GST-P-positive foci (Table 2). Further, SREBP1 is required for the induction of thyroid hormone-responsive protein (THRSP) in hepatocytes (Martel et al., 2006). Brown et al. (1997) reported that exposure of Thrsp antisense oligonucleotide inhibited the expression of mRNAs encoding lipogenic (fatty acid synthase, ATP citrate lyase and malic enzyme) and glycolytic (pyruvate kinase) enzymes. The log ratios of these genes were −3.04 ($p = 0.00266$), −1.51 ($p = 0.00105$), −0.508 ($p = 0.0221$) and −2.05 ($p = 0.000531$), respectively. These observations suggest that the aberrant decrease of lipogenic and glycolytic enzymes may be caused by the suppression of SREBP1. This raises the possibility that hepatotoxicity induced by nitroso compounds would be cased by the down regulation of SREBP1.

One of nuclear receptors, retinoid X receptor alpha (RXRalpha) was also decreased in GST-P-positive foci. RXRalpha dimerizes with constitutive androstane receptor (CAR), pregnane X receptor (PXR) and peroxisome proliferator-activated receptor (PPARalpha). Hepatocyte RXRalpha-deficient mice revealed that hepatocyte RXRalpha is required for induction of metabolic enzymes by the ligands of CAR, PXR, and PPARalpha, and is essential for xenobiotic metabolism in vivo (Cai et al., 2002). Hepatotoxicity may be caused by decrease of RXRalpha expression in GST-P-positive foci.

3.5. Expression of transcripts coding chromatin modification enzymes

Sequence-specific transcription factors require cofactors for transcription from the chromatin context and chromatin components affect gene expression (Sterner and Berger, 2000). The expression of cofactors and chromatin components during hepatocarcinogenesis has not been studied well. Recent studies demonstrated that cofactors possess histone modification activities, which are required for the change of chromatin conformation and the regulation of gene function. Generally, histone acetylation promotes transcription, although histone methylation both positively and negatively regulates gene expression dependent on the position of the lysine residue on histone. An epigenetic program including histone and DNA modifications is important for the maintenance of inheritable information and the disturbance of epigenetic balances may lead to alterations in gene expression, resulting in cellular transformation and malignant growth (Lund and van Lohuizen, 2004). Microarray analysis revealed that coactivator-associated arginine methyltransferase (Carm1) and Rbbp7, also termed retinoblastoma suppressor-associated protein 46 (RbAp46), were induced in GST-P-positive foci. CARM1 catalyzes the methylation of histone H3 at Arg17 and can also function as a coactivator for transcription factor NF-E2-related factor 2 (Nrf2), which regulates the induction of Phase II detoxifying enzymes, including GST-P, through its transactivation domain (Lin et al., 2006; Miao et al., 2006). Although increased Nrf2 was detected in hyperplasic nodules, the extremely high level of GST-P expression during hepatocarcinogenesis was difficult to explain by the slight induction of Nrf2 alone (Ikeda et al., 2004). Here we found the overexpression of Carm1 in GST-P-positive foci. Increase of both Nrf2 and Carm1 expression and the cooperative regulation of gene expression may lead to the induction of GST-P expression during hepatocarcinogenesis.

We also found the induction of RbAp46, which contributed to the regulation of gene expression as a subunit of histone acetyltransferase, histone deacetylase and chromatin remodeling complexes NURD (Zhang et al., 1999). Li et al., 2003 reported that the expression of RbAp46 suppressed colony formation in soft agar, and inhibited tumor formation in nude mice. They also showed that high levels of RbAp46 expression promoted apoptotic cell death, resulting in the inhibition of tumorigenicity of neoplastigenic breast epithelial cells. These results suggest that overexpressed RbAp46 in GST-P-positive cells may function as a suppressor of tumorigenicity in the early stage of hepatocarcinogenesis.

3.6. Expression of transcripts coding chromatin components and related factors

High mobility group box 2 (Hmgb2), a member of HMGB family proteins, was up-regulated in GST-P-positive cells. HMGB proteins are abundant nonhistone nuclear proteins that have been found in association with chromatin. HMGB family proteins contain two DNA-binding HMG-box domains and bind to DNA without sequence specificity, but play important architectural roles in the assembly of nucleoprotein complexes in a variety of biological processes including the initiation

of transcription and DNA repair (Thomas, 2001). Further, HMGB2, while showing no coactivator activity on its own, can promote transcription activity together with histone acetyltransferase. HMGB2 acts mainly at the level of elongation and is a coactivator for transcription from chromatin templates (Guermah et al., 2006). Hmgb2 is frequently overexpressed in malignant gastrointestinal stromal tumors and ovarian cancer (Koon et al., 2004; Ouellet et al., 2006). Overexpression of Hmgb2 may be common feature of carcinogenesis. HMGB2 binds with high affinity to DNA modified with the cancer chemotherapeutic drug cisplatin and enhancement of cisplatin sensitivity in Hmgb2 transfected human lung cancer cells (Arioka et al., 1999; Farid et al., 1996). Cisplatin-induced hepatotoxicity may be promoted by overexpressed Hmgb2.

Nucleophosmin was identified as a positively regulated gene in GST-P-positive cells. Nucleophosmin is a key regulator for centrosome duplication, the maintenance of genomic integrity, and ribosome assembly. At the steady state, nucleophosmin localizes mainly in the nucleolus, whereas aberrant cytoplasmic localization of nucleophosmin is observed in acute myeloid leukemias (Mariano et al., 2006). Observation of localization of nucleophosmin would be important for the understanding of roles of overexpressed nucleophosmin in GST-P-positive foci. Recent studies suggest that nucleophosmin may be a Ran-Crm1 substrate that controls centrosome duplication and utilizes a conserved Crm1-dependent nuclear export sequence in its amino terminus to enable shuttling between the nucleolus/nucleus and cytoplasm (Wang et al., 2005; Yu et al., 2006). Further, purification of nucleophosmin binding protein revealed that nucleophosmin directly interacted with ribosomal protein L5. This interaction mediated the colocalization of nucleophosmin with both maturing nuclear 60S ribosomal subunits and newly exported and assembled 80S ribosomes (Yu et al., 2006). Interestingly, Crm1 (log ratio, 0.908; $p = 0.0427$) and ribosomal protein L5 (log ratio, 0.830; $p = 0.00429$) were also up-regulated in GST-P-positive foci. Overexpression of these genes may disturb multiple processes involved in nucleophosmin, which accelerate oncogenesis.

DNA microarray analysis in this study uncovered several genes, which expression was induced or repressed during hepatocarcinogenesis, and some of these genes possess anti-oncogenic as well as oncogenic activities and may be involved in regulation of GST-P expression. Our study provided a candidate gene link to the early stage of carcinogenesis and hepatotoxicity. To elucidate the mechanisms of the early stage of hepatocarcinogenesis mediated by these genes, further characterization of aberrantly expressed genes in GST-P-positive cells is necessary. We proceed to observe the effect of overexpression of these genes up-regulated during hepatocarcinogenesis, especially epigenetics regulatory factors, on transformation and the induction of GST-P expression.

Acknowledgements

This research was supported in part by grants from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Japan Society for the Promotion of Science (JSPS), a Grant-in-Aid for Scientific Research from the Ministry of Health and Labor of Japan, Long-range Research Initiative (LRI) by Japan Chemical Industry Association (JCIA) and Sankyo Foundation of Life Science.

References

Arioka, H., Nishio, K., Ishida, T., Fukumoto, H., Fukuoka, K., Nomoto, T., Kurokawa, H., Yokote, H., Abe, S., Saijo, N., 1999. Enhancement of cisplatin sensitivity in high mobility group 2 cDNA-transfected human lung cancer cells. Jpn. J. Cancer Res. 90, 108–115.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Brown, S.B., Maloney, M., Kinlaw, W.B., 1997. "Spot 14" protein functions at the pretranslational level in the regulation of hepatic metabolism by thyroid hormone and glucose. J. Biol. Chem. 272, 2163–2166.

Cai, Y., Konishi, T., Han, G., Campwala, K.H., French, S.W., Wan, Y.J., 2002. The role of hepatocyte RXR alpha in xenobiotic-sensing nuclear receptor-mediated pathways. Eur. J. Pharm. Sci. 15, 89–96.

Cheema, S.K., Mishra, S.K., Rangnekar, V.M., Tari, A.M., Kumar, R., Lopez-Berestein, G., 2003. Par-4 Transcriptionally Regulates Bcl-2 through a WT1-binding Site on the bcl-2 Promoter. J. Biol. Chem. 278, 19995–20005.

Eberle, D., Hegarty, B., Bossard, P., Ferre, P., Foufelle, F., 2004. SREBP transcription factors: master regulators of lipid homeostasis. Biochimie 86, 839–848.

Eriksson, L.C., Sharma, R.N., Roomi, M.W., Ho, R.K., Farber, E., Murray, R.K., 1983. A characteristic electrophoretic pattern of cytosolic polypeptides from hepatocyte nodules generated during liver carcinogenesis in several models. Biochem. Biophys. Res. Commun. 117, 740–745.

Farid, R.S., Bianchi, M.E., Falciola, L., Engelsberg, B.N., Billings, P.C., 1996. Differential binding of HMG1, HMG2, and a single HMG box to cisplatin-damaged DNA. Toxicol. Appl. Pharmacol. 141, 532–539.

Guermah, M., Palhan, V.B., Tackett, A.J., Chait, B.T., Roeder, R.G., 2006. Synergistic functions of SII and p300 in productive activator-dependent transcription of chromatin templates. Cell 125, 275–286.

Ikeda, H., Nishi, S., Sakai, M., 2004. Transcription factor Nrf2/MafK regulates rat placental glutathione-*S*-transferase gene during hepatocarcinogenesis. Biochem. J. 380, 515–521.

Ito, N., Tamano, S., Shirai, T., 2003. A medium-term rat liver bioassay for rapid in vivo detection of carcinogenic potential of chemicals. Cancer Sci. 94, 3–8.

Johnstone, R.W., See, R.H., Sells, S.F., Wang, J., Muthukkumar, S., Englert, C., Haber, D.A., Licht, J.D., Sugrue, S.P., Roberts, T., Rangnekar, V.M., Shi, Y., 1996. A novel repressor, par-4, modulates transcription and growth suppression functions of the Wilms' tumor suppressor WT1. Mol. Cell. Biol. 16, 6945–6956.

Koon, N., Schneider-Stock, R., Sarlomo-Rikala, M., Lasota, J., Smolkin, M., Petroni, G., Zaika, A., Boltze, C., Meyer, F., Andersson, L., Knuutila, S., Miettinen, M., El-Rifai, W., 2004. Molecular targets for tumour progression in gastrointestinal stromal tumours. Gut 53, 235–240.

Li, G.C., Guan, L.S., Wang, Z.Y., 2003. Overexpression of RbAp46 facilitates stress-induced apoptosis and suppresses tumorigenicity of neoplastigenic breast epithelial cells. Int. J. Cancer 105, 762–768.

Lin, W., Shen, G., Yuan, X., Jain, M.R., Yu, S., Zhang, A., Chen, J.D., Kong, A.N., 2006. Regulation of Nrf2 transactivation domain activity by p160 RAC3/SRC3 and other nuclear co-regulators. J. Biochem. Mol. Biol. 39, 304–310.

Loeb, D.M., 2006. WT1 Influences apoptosis through transcriptional regulation of Bcl-2 family members. Cell Cycle 5, 1249–1253.

Loeb, D.M., Sukumar, S., 2002. The role of WT1 in oncogenesis: tumor suppressor or oncogene? Int. J. Hematol. 76, 117–126.

Lund, A.H., van Lohuizen, M., 2004. Epigenetics and cancer. Genes Dev. 18, 2315–2335.

Mariano, A.R., Colombo, E., Luzi, L., Martinelli, P., Volorio, S., Bernard, L., Meani, N., Bergomas, R., Alcalay, M., Pelicci, P.G., 2006. Cytoplasmic localization of NPM in myeloid leukemias is dictated by gain-of-function mutations that create a functional nuclear export signal. Oncogene 25, 4376–4380.

Martel, P.M., Bingham, C.M., McGraw, C.J., Baker, C.L., Morganelli, P.M., Meng, M.L., Armstrong, J.M., Moncur, J.T., Kinlaw, W.B., 2006. S14 protein in breast cancer cells: direct evidence of regulation by SREBP-1c, superinduction with progestin, and effects on cell growth. Exp. Cell Res. 312, 278–288.

Mayo, M.W., Wang, C.Y., Drouin, S.S., Madrid, L.V., Marshall, A.F., Reed, J.C., Weissman, B.E., Baldwin, A.S., 1999. WT1 modulates apoptosis by transcriptionally upregulating the bcl-2 proto-oncogene. EMBO J. 18, 3990–4003.

Miao, F., Li, S., Chavez, V., Lanting, L., Natarajan, R., 2006. Coactivator-associated arginine methyltransferase-1 enhances nuclear factor-kappaB-mediated gene transcription through methylation of histone H3 at arginine 17. Mol. Endocrinol. 20, 1562–1573.

Morimura, S., Suzuki, T., Hochi, S., Yuki, A., Nomura, K., Kitagawa, T., Nagatsu, I., Imagawa, M., Muramatsu, M., 1993. Transactivation of glutathione transferase P gene during chemical hepatocarcinogenesis of the rat. Proc. Natl. Acad. Sci. U.S.A. 90, 2065–2068.

Ouellet, V., Page, C.L., Guyot, M.C., Lussier, C., Tonin, P.N., Provencher, D.M., Mes-Masson, A.M., 2006. SET complex in serous epithelial ovarian cancer. Int. J. Cancer 119, 2119–2126.

Sakai, M., Muramatsu, M., 2005. Regulation of GST-P gene expression during hepatocarcinogenesis. Methods Enzymol. 401, 42–61.

Sato, K., 1989. Glutathione transferase as markers of preneoplasia and neoplasia. Adv. Cancer Res. 52, 205–255.

Satoh, K., Hatayama, I., Tateoka, N., Tamai, K., Shimizu, T., Tatematsu, M., Ito, N., Sato, K., 1989. Transient induction of single GST-P positive hepatocytes by DEN. Carcinogenesis 10, 2107–2111.

Satoh, K., Kitahara, A., Soma, Y., Inaba, Y., Hatayama, I., Sato, K., 1985. Purification, induction, and distribution of placental glutathione transferase: a new marker enzyme for preneoplastic cells in the rat chemical hepatocarcinogenesis. Proc. Natl. Acad. Sci. U.S.A. 82, 3964–3968.

Satoh, K., Takahashi, G., Miura, T., Hayakari, M., Hatayama, I., 2005. Enzymatic detection of precursor cell populations of preneoplastic foci positive for gamma-glutamyltranspeptidase in rat liver. Int. J. Cancer 115, 711–716.

Solt, D., Farber, E., 1976. New principle for the analysis of chemical carcinogenesis. Nature 263, 701–703.

Sterner, D.E., Berger, S.L., 2000. Acetylation of histones and transcription-related factors. Microbiol. Mol. Biol. Rev. 64, 435–459.

Sugiyama, H., 2001. Wilms' tumor gene WT1: its oncogenic function and clinical application. Int. J. Hematol. 73, 177–187.

Suzuki, S., Asamoto, M., Tsujimura, K., Shirai, T., 2004. Specific differences in gene expression profile revealed by cDNA microarray analysis of glutathione-*S*-transferase placental form (GST-P) immunohistochemically positive rat liver foci and surrounding tissue. Carcinogenesis 25, 439–443.

Suzuki, T., Imagawa, M., Hirabayashi, M., Yuki, A., Hisatake, K., Nomura, K., Kitagawa, T., Muramatsu, M., 1995. Identification of an enhancer responsible for tumor marker gene expression by means of transgenic rats. Cancer Res. 55, 2651–2655.

Thomas, J.O., 2001. HMG1 and 2: architectural DNA-binding proteins. Biochem. Soc. Trans. 29, 395–401.

Wang, W., Budhu, A., Forgues, M., Wang, X.W., 2005. Temporal and spatial control of nucleophosmin by the Ran-Crm1 complex in centrosome duplication. Nat. Cell Biol. 7, 823–830.

Yu, Y., Maggi Jr., L.B., Brady, S.N., Apicelli, A.J., Dai, M.S., Lu, H., Weber, J.D., 2006. Nucleophosmin is essential for ribosomal protein L5 nuclear export. Mol. Cell. Biol. 26, 3798–3809.

Zhang, Y., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Bird, A., Reinberg, D., 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev. 13, 1924–1935.

J Pharmacol Sci 101, 267 – 270 (2006)

Journal of Pharmacological Sciences

Current Perspective

MicroRNA: Biogenetic and Functional Mechanisms and Involvements in Cell Differentiation and Cancer

Soken Tsuchiya[1], Yasushi Okuno[1], and Gozoh Tsujimoto[1,*]

[1]*Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

Received June 2, 2006

Abstract. MicroRNAs (miRNAs) are endogenous small noncoding RNAs (20 – 23 nucleotides) that negatively regulate the gene expressions at the posttranscriptional level by base pairing to the 3' untranslated region of target messenger RNAs. Hundreds of miRNAs have been identified in humans and evolutionarily conserved from plants to animals. It is revealed that miRNAs regulate various physiological and pathological pathways such as cell differentiation, cell proliferation, and tumoriogenesis. By the computational analysis, it is predicted that 30% of protein-encoding genes are regulated by miRNAs. In this review, we discuss recent remarkable advances in the miRNA biogenetic and functional mechanisms and the involvements of miRNAs in cell differentiation, especially in hematopoietic lineages, and cancer. These evidences offer the possibility that miRNAs would be potentially useful for drug discovery.

Keywords: microRNA, RNA cleavage, translational repression, target mRNA, base pairing

Introduction

MicroRNAs (miRNAs) are endogenous short non-coding RNA molecules (20 – 23 nucleotides) that regulate cell differentiation, cell proliferation, and apoptosis through post-transcriptional suppression of gene expression by binding to the complementary sequence in the 3' untranslated region (3'UTR) of target messenger RNAs (mRNAs) (1). Hundreds of miRNAs have been identified in humans and they are evolutionarily conserved (1, 2). In addition, the presence of up to 1000 miRNAs is estimated by computational analysis (3). Strikingly, 30% of protein-encoding genes in humans are predicted to be regulated by miRNAs (4). Recently, it has been revealed that altered expression of specific miRNA genes contributes to the initiation and progression of diseases such as cancer (5 – 10). This review focuses on the biogenetic and functional mechanisms and the involvements in cell differentiation and cancer in mammalian miRNAs and the utility of miRNAs in drug discovery.

*Corresponding author. gtsuji@pharm.kyoto-u.ac.jp
Published online in J-STAGE
doi: 10.1254/jphs.CPJ06013X

Invited article

Mechanisms of biogenesis and function

Most miRNA genes are located in the introns of host genes or outside genes. Unlike *Drosophila*, most of the human miRNA genes individually exist, although some human miRNAs are found in polycistronic clusters (5, 8, 9).

The miRNAs are synthesized through multiple steps (Fig. 1). Initially, the miRNAs are transcribed as long RNA precursors (pri-miRNAs) (11). As pri-miRNAs usually contain the cap structure and the poly(A) tail, it is suggested that the transcription of miRNAs is carried out by RNA polymerase II (12). The pri-miRNAs are processed into the precursors of approximately 70 nucleotides (pre-miRNAs) with a stem-loop structure and a two nucleotide 3' overhang by the RNase III enzyme Drosha and the double-stranded-RNA-binding protein DGCR8/Pasha (13, 14), and pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin-5 in a Ran guanosine triphosphate-dependent manner (15). Pre-miRNAs exported in the cytoplasm are processed by another RNase III enzyme, Dicer, and only one strand (guide strand) as a mature

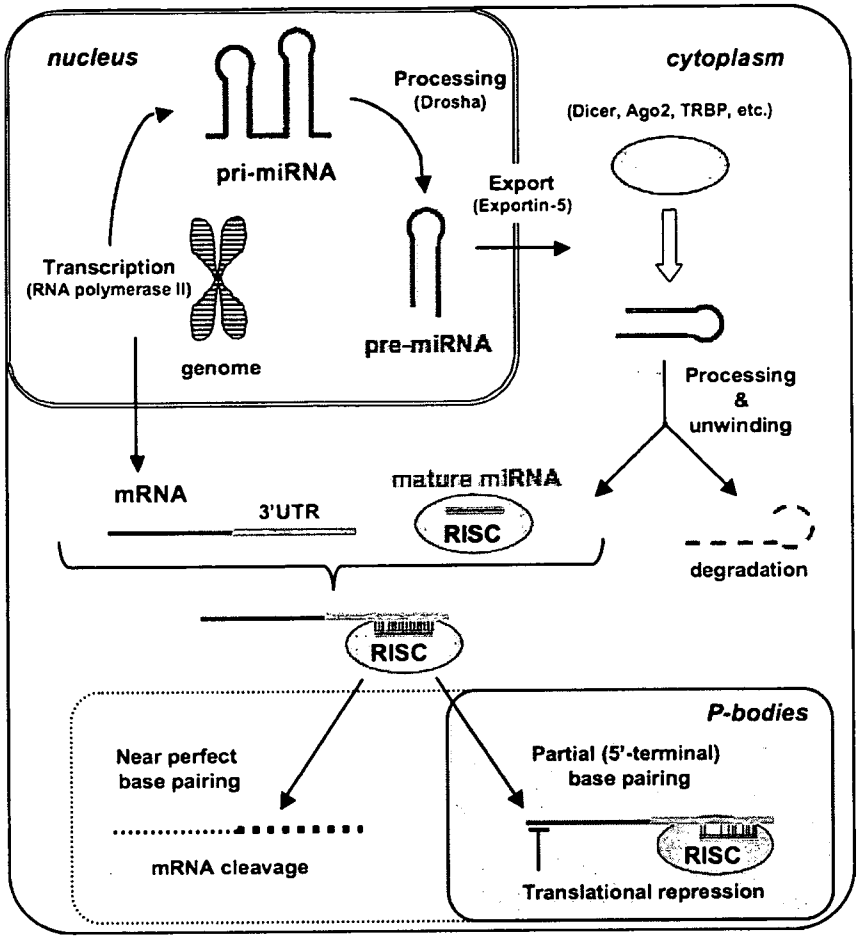

Fig. 1. Diagram of the miRNA biogenetic and functional mechanisms. Whether the target mRNA cleavage by RISC occurs in the cytoplasm or P-bodies remains unknown.

miRNA is incorporated into a RNA-induced silencing complex (RISC) that mediates either target RNA cleavage or translational inhibition, while the another strand (passenger strand) is excluded. Which strand is incorporated in RISC is determined by the stability of the base pairs at the 5' end of the duplex (16, 17). The incorporated guide strand guides the RISC to the complementary sequence in the 3'UTR of target mRNA. When the guide strand shares perfect or near perfect base pairing with the 3'UTR of target mRNA, the target mRNA is degraded by Argonaute2 (Ago2), a component of RISC (18). On the contrary, when the guide strand shares partial base pairing, translation is target-specifically repressed without the target mRNA degradation (19). Recent studies have revealed that RISC is at least composed of Dicer, Ago2, and the double-strand RNA binding protein TRBP, and RISC efficiently processes pre-miRNAs to mature miRNAs (20). Furthermore, RISC more efficiently cleaves target mRNAs by using the pre-miRNAs than the duplex miRNAs that do not have the stem-loop. These results suggest that miRNA processing by Dicer, assembly of the mature miRNA into RISC, and target RNA cleavage by Ago2 are coupled. Compared to the RNA cleavage mechanism by Ago2, the translational repression mechanism by miRNAs had been poorly understood. Recently, it was revealed that the target mRNAs binding to RISC through partial base pairing are accumulated in the cytoplasmic foci referred to as processing bodies (P-bodies) (21, 22). P-bodies, in which the mRNAs are stored or degraded by the decapping enzymes and exonucleases, do not contain the translational machinery (23). Furthermore, the disruption of P-bodies by the silencing of GW182, a key protein in P-body, inhibits translational silencing in not only partial base pairing but also perfect base pairing (24), although the localization of target mRNA with perfect base pairing is not detected in P-bodies (21). These results suggest that, at least in part, translational repression appears to be caused by the recruitment of target mRNAs to P-bodies. However, whether localiza-

tion of the RISC-target mRNA complex in P-bodies is a cause or a result of the translational repression and whether the target mRNA cleavage by RISC occurs in the cytoplasm or P-bodies remain controversial issues.

Cell differentiation

Increasing evidence indicates that miRNAs have distinct expression patterns among tissues and cells in different differentiation stage (25). It is reported that overexpression of miR-124, which is preferentially expressed in brain, shifted the gene expression profile of HeLa cells towards that of the brain. Similarly, overexpression of miR-1 shifted the expression profile towards that of the muscle in that miR-1 is preferentially expressed (25). These results indicate that miRNAs play important roles in cell differentiation and characterization.

Recently, it was revealed that miRNAs also played critical roles in the differentiation of mammalian hematopoietic lineage. For example, miR-181 is preferentially expressed in the thymus and B-lymphoid cells of mouse bone marrow and promotes B cell differentiation by overexpression in hemapoietic stem/progenitor cells (26). Conversely, overexpression of the miR-181a, one member of the miR-181 family, was reported to repress megakaryoblast differentiation in humans (27). By the induction of megakaryoblast differentiation, the expression of endogenous miR-181a is downregulated through the acetylcholinesterase, protein kinase (PK) C, and PKA cascade. The expression of miR-130a is also downregulated by the induction of megakaryoblast differentiation (28). miR-130a targets the transcriptional factor MAFB that is a transcriptional activator of GPIIB, an important protein for platelet physiology. Furthermore, miR-223 is upregulated by the retinoic acid-induced replacement of NFI-A with CCAAT/Enhancer binding protein (C/EBP) α, and promotes human granulopoiesis (29). As miR-223 repressed NFI-A translation, the upregulation of miR-223 by C/EBPα and granulopoiesis further accelerated through positive feedback by miR-223.

Cancer

It has been revealed that the change of miRNA expressions contributes to the initiation and progression of cancer. More than 50% of miRNAs are located in cancer-associated genomic regions or in fragile sites (5). The expression of miR-15a and miR-16, which locate as a cistronic cluster at 13q14, is deleted or decreased in most cases (approx. 68%) of B cell chronic lymphocytic leukemia (B-CLL) (6). Both these miRNAs negatively regulate the expression of B cell lymphoma 2 (Bcl2), that is reported to be expressed in many types of cancer including leukemias, and inhibit cell death (7). Overexpression of miR-15 and miR-16 in the MEG-01 cell line actually induces the apoptosis. Inversely, one cluster of miRNAs, miR-17 – 92 polycistron, was found to increase in the cancers such as B-CLL (8). The expression of six miRNAs in this cluster is upregulated by c-myc, whose expression and/or function are one of the most common abnormalities in human cancers, and miR-17-5p and miR-20a included in this cluster negatively regulate the expression of transcriptional factor E2F1 (9). Furthermore, mice reconstituted with hemotopoietic stem cells overexpressing miR-17 – 19b exhibit accelerated c-myc-induced lymphomagenesis (8). Furthermore, it was revealed that miRNA expression profiles enable researchers to successfully classify poorly characterized human tumors that can not be accurately classified by mRNA expression profiles (10). These results show the possibility that miRNAs have clinical benefits as not only therapeutic targets but also a tool for cancer diagnosis.

Drug discovery

miRNAs are expected to be potential targets of therapeutic strategies applied to drug discovery for a number of reasons. Firstly, in addition to the initiation and progression of tumor, miRNAs play critical roles in various biological pathways such as differentiation of adipocyte and insulin secretion and diseases such as diabetes and hepatitis. Therefore, the possibility that various human diseases are caused by abnormalities in miRNAs is indicated. Actually, miR-15 and miR-16 have been deleted or decreased in most cases of B-CLL and are identified as tumor suppressor genes (6, 7). Secondly, miRNA expression profiles are correlated with clinical severity of cancer malignancy, and because of this, miRNAs are expected to be powerful tools for cancer diagnosis (10). Thirdly, miRNAs are applicable in gene therapy. The expression of miRNAs can be introduced in vivo by using viral vectors and chemical modifications. Finally, antisense oligonucleotides are potent inhibitors of miRNA, and they can be applied to gene therapy. Actually, it was reported that introduction of 2'-*O*-methoxyethyl phosphorothioate antisense oligonucleotide of miR-122, which is abundant in the liver and regulates cholesterol and fatty-acid metabolism, decreases plasma cholesterol levels and improves liver steatosis in mice with diet-induced obesity (30). These findings indicate that miRNAs and the antisense oligonucleotides are potential targets for drug discovery.

Perspective

It has been established that miRNAs play critical roles in cell differentiation, proliferation, and apoptosis, and the abnormalities of specific miRNA expression contribute to the initiation and progression of tumor. However, identification of target mRNAs negatively regulated by miRNAs remain largely to be explored. Although up to hundreds of target genes toward a single miRNA were predicted by bioinformatics approaches (4), there is no comprehensive assay to biologically validate the prediction algorithm. Therefore, establishment of a method to comprehensively and rapidly identify target mRNAs for the miRNA is necessary for understanding biological and functional mechanisms of miRNA.

References

1 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–297.
2 Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 2000;408:86–89.
3 Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. Phylogenetic shadowing and computational identification of human microRNA genes. Cell. 2005;120:21–24.
4 Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.
5 Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A. 2004;101:2999–3004.
6 Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002;99:15524–15529.
7 Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A. 2005;102:13944–13949. Erratum in: Proc Natl Acad Sci U S A. 2006;103:2464–2565.
8 He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, et al. A microRNA polycistron as a potential human oncogene. Nature. 2005;435:828–833.
9 O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. c-Myc-regulated microRNAs modulate E2F1 expression. Nature. 2005;435:839–843.
10 Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435:834–838.
11 Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 2002;21:4663–4670.
12 Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23:4051–4060.
13 Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003;425:415–419.
14 Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, et al. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432:235–240.
15 Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 2003;17:3011–3016.
16 Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. Cell. 2003;115:209–216. Erratum in: Cell. 2003;115:505.
17 Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell. 2003;115:199–208.
18 Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell. 2004;15:185–197.
19 Hutvágner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science. 2002;297:2056–2060.
20 Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell. 2005;123:631–640.
21 Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, et al. Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science. 2005;309:1573–1576.
22 Liu J, Valencia-Sanchez MA, Hannon GJ, Parker R. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol. 2005;7:719–723.
23 Andrei MA, Ingelfinger D, Heintzmann R, Achsel T, Rivera-Pomar R, Lu hrmann R. A role for eIF4E and eIF4E-transporter in targeting mRNPs to mammalian processing bodies. RNA. 2005;11:717–727.
24 Liu J, Rivas FV, Wohlschlegel J, Yates JR 3rd, Parker R, Hannon GJ. A role for the P-body component GW182 in microRNA function. Nat Cell Biol. 2005;7:1261–1266.
25 Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005;433:769–773.
26 Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. Science. 2004;303:83–86.
27 Guimaraes-Sternberg C, Meerson A, Shaked I, Soreq H. MicroRNA modulation of megakaryoblast fate involves cholinergic signaling. Leuk Res. 2006;30:583–595.
28 Garzon R, Pichiorri F, Palumbo T, Iuliano R, Cimmino A, Aqeilan R, et al. MicroRNA fingerprints during human megakaryocytopoiesis. Proc Natl Acad Sci U S A. 2006;103:5078–5083.
29 Fazi F, Rosa A, Fatica A, Gelmetti V, De Marchis ML, Nervi C, et al. A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. Cell. 2005;123:819–831.
30 Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, et al. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. Cell Metab. 2006;3:87–98.

Nucleic Acids Research, 2006, Vol. 34, Database issue D673–D677
doi:10.1093/nar/gkj028

GLIDA: GPCR-ligand database for chemical genomic drug discovery

Yasushi Okuno*, Jiyoon Yang, Kei Taneishi, Hiroaki Yabuuchi and Gozoh Tsujimoto

Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Received August 15, 2005; Revised and Accepted September 22, 2005

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GPCR-LIgand DAtabase (GLIDA) is a novel public GPCR-related chemical genomic database that is primarily focused on the correlation of information between GPCRs and their ligands. It provides correlation data between GPCRs and their ligands, along with chemical information on the ligands, as well as access information to the various web databases regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of GPCR-related drug discovery to easily retrieve such information from either biological or chemical starting points. GLIDA includes structure similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their interactions and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. GLIDA is publicly available at http://gdds.pharm.kyoto-u.ac.jp:8081/glida. We hope that it will prove very useful for chemical genomic research and GPCR-related drug discovery.

INTRODUCTION

The superfamily of G-protein coupled receptors (GPCRs) forms the largest class of cell surface receptors. These molecules regulate various cellular functions responsible for physiological responses (1). GPCRs represent one of the most important families of drug targets in pharmaceutical development (2). A large majority of human-derived GPCRs still remain 'orphans' with no identified natural ligands or functions, and thus a key goal of GPCR research related to drug design is to identify new ligands for such orphan GPCRs.

With the unprecedented accumulation of the genomic information, databases and bioinformatics have become essential tools to guide GPCR research. The GPCRDB (http://www.gpcr.org/7tm/) (2) and IUPHAR (http://iuphar-db.org/iuphar-rd/index.html) (3) receptor databases are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in collecting and applying chemical information in the post-genome era. This new trend is called 'chemical genomics', in which biological information and chemical information are integrated on the genome scale (4,5). PubChem (http://pubchem.ncbi.nlm.nih.gov/) (6), KEGG/LIGAND (http://www.genome.jp/kegg/ligand.html) (7) and ChEBI (http://www.ebi.ac.uk/chebi/) (8) have been developed as databases related to chemical genomics. KEGG/LIGAND and ChEBI contain primarily biochemical information on reported enzymatic reactions. Recently, NIH (the National Institutes of Health) opened PubChem, a public database providing information on the chemical structures of small molecules. However, one cannot retrieve direct information relating these chemical structures to gene or protein entries. Although chemical genomic approaches have thrown new light on relationships between receptor sequences and compounds that interact with particular receptors, the GPCR-ligand information is not well represented in these large-scale databases for chemical genomics.

There are still very few publicly available databases or tools for GPCR-specialized drug discovery from the viewpoint of chemical genomics. Herein, we have developed a novel relational database, GLIDA (GPCR-LIgand DAtabase) (9).

*To whom correspondence should be addressed. Tel: +81 75 753 9264; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp

Table 1. The current numbers of GLIDA ligands and GPCRs and their respective links

Information item	Number of entries
GPCR entries	3738
Links to Entrez Gene	3073
Links to GPCRDB	3738
Links to UniProt	3738
Links to IUPHAR	389
Links to KEGG	595
Ligand entries	649
Cas registry number	320
Lolecular structure	364
Links to PubChem	242
Links to ChEBI	28
Links to KEGG	109
GPCR-ligand pair entries	1989
GPCR entries	281
Ligand entries	632

GLIDA contains biological information on GPCRs and chemical information on their ligand compounds. Furthermore, it provides various analytical data on GPCR-ligand correlations by incorporating bioinformatics and chemoinformatics methods, and thus it should prove very useful for chemical genomic research in GPCR-related drug discovery.

DATA CONTENTS

GLIDA contains three types of primary data: biological information on GPCRs, chemical information on their ligands and information on binding of specific GPCR-ligand pairs. The GPCR entries were acquired from the deposits of human, mouse and rat entries in the GPCRDB because these three species include sufficient information regarding ligands, and rats and mice are representative model animals for drug discovery. The ligand information was manually collected and curated using various public web sites and commercial DBs, such as the IUPHAR Receptor Database, PubMed, PubChem and MDL ISIS/Base 2.5. Table 1 indicates the size and scope of the GLIDA database.

GPCR and ligand data

The database lists general information on GPCR and ligand data, respectively. The general information table of GPCR contains gene names, family names, protein sequences and links to other biological databases, such as GPCRDB, UniProt, IUPHAR, Entrez Gene and KEGG. The ligand result page provides a general information table containing names, molecular structures, CAS registry numbers, formulas, molecular weights, MOLfiles and links to the other chemical databases KEGG, PubChem and ChEBI.

Information on binding of GPCR-ligand pairs

The correlation information relating GPCRs to particular ligands, a key issue for GPCR-related drug discovery, is stored in a relational database. GLIDA allows users to retrieve GPCR-ligand binding information dynamically and continuously. When users retrieve a GPCR (or ligand) entry, its result page displays all entries showing the corresponding ligands (or GPCR entries) with their binding activity types, as well as references. The references are hyperlinked with the corresponding PubMed literature or the IUPHAR pages that were used to collect the information regarding GPCR-ligand binding. The activity types include agonist, inverse agonist, antagonist and so on. An agonist will bind to and activate the corresponding GPCRs, whereas an antagonist will bind to and block the activity of the corresponding GPCRs. An inverse agonist binds to GPCRs and reduces the fraction of them that are in an active conformation, and a partial agonist is an agonist that in a given tissue, under specified conditions, cannot elicit as large an effect as another agonist acting through the same GPCRs in the same tissue can.

WEB INTERFACE AND APPLICATION

GLIDA was constructed on the LAMP (Linux, Apache, MySQL and PHP) platform. GLIDA is available at http://gdds.pharm.kyoto-u.ac.jp:8081/glida. The web interface of GLIDA includes a GPCR search page (Figure 1a) and a ligand search page (Figure 1b). Each page consists of a classification table and a keyword search box. The user can search a GPCR (or ligand) manually from the guide-tree of the classification table, or automatically by using the keyword search function of MySQL. Every GPCR (or ligand) has its own result page (Figure 1c or d) containing a general information table for a GPCR (or ligand), a table of its correlated ligands (or GPCRs) and a button to carry out a similarity search and correlation analysis. Clicking the button starts the calculation, and an analytical report page (Figure 1e) then appears with a list of the top 25 entries that are most similar to the GPCR (or ligand) and a correlation map of the 25 GPCRs (or ligands) and their corresponding binding pairs. A search starting from ligand retrieval proceeds in the same way.

Hierarchical classification

The GPCR classification table on the search page was adapted from the phylogenetic tree of the GPCRDB information system (http://www.gpcr.org/7tm/phylo/phylo.html). As for the ligand classification table, GLIDA offers an original one (Figure 1b) that is based on a cluster analysis of the ligand structures as follows. We converted the structural images of the ligands into computational MDL Mol files using ISIS/Draw software. Next, we calculated distance metrics among all of the ligands using the frequency profiles of the atoms and the bonds of the KEGG atom types (10), and carried out complete-linkage clustering. We manually defined subclusters based on their common structural skeletons. Both the GPCR and ligand classification tables display the entries of the corresponding GPCRs or ligands at the end of the tree, and these are hyperlinked with their respective result pages.

Similarity search and GPCR-LIGAND correlation maps

GLIDA has a structure similarity search function on its result pages. Alignment scores of protein sequences generated by the BLAST algorithm provide similarity measures for GPCRs. Ligand similarity is defined by the dissimilarity (distance) of frequency profile patterns generated from the constitutive atoms and bonds of the chemical structure, using the KEGG atom types (10,11). From this similarity search, the 25 most

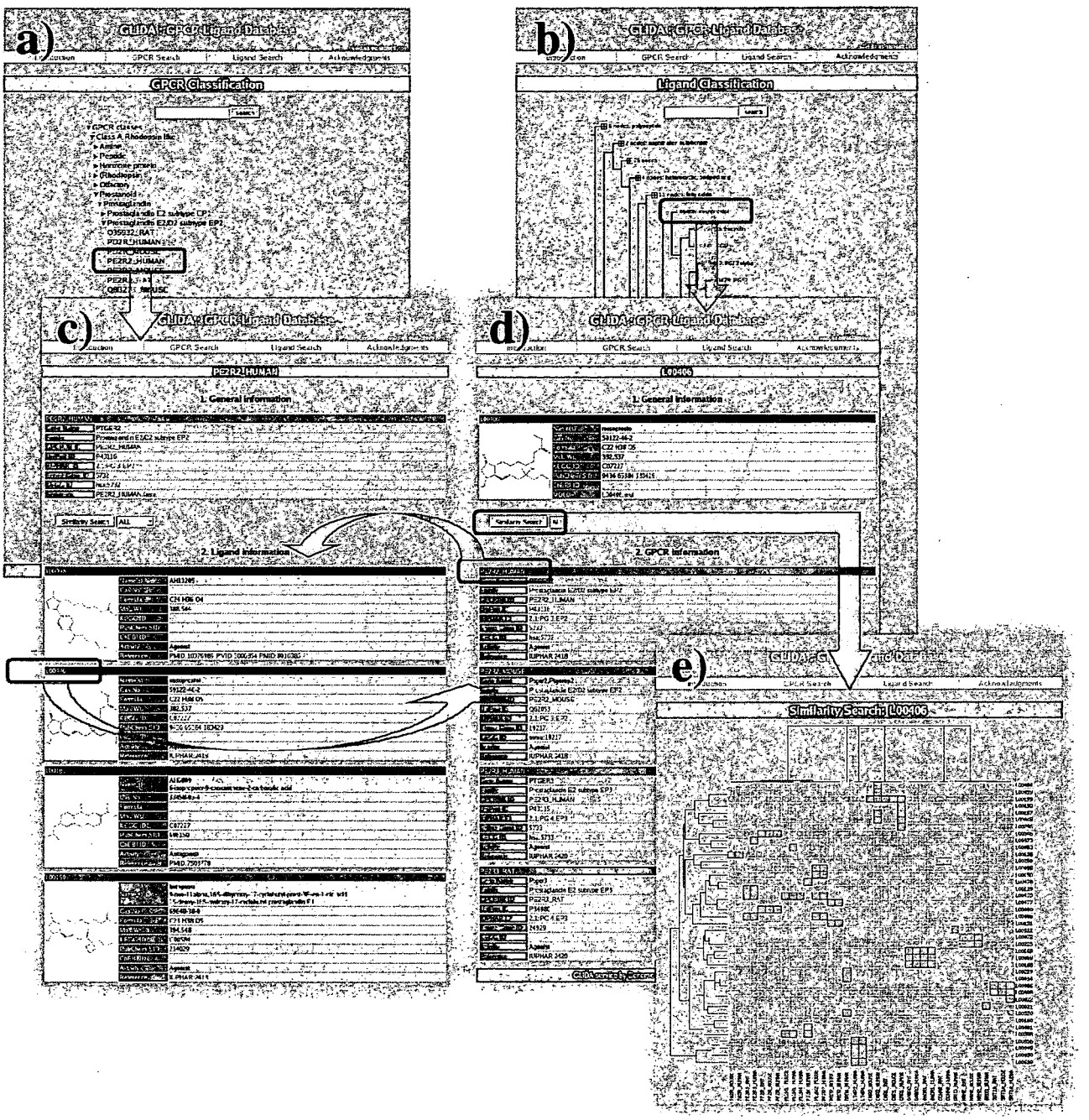

Figure 1. A screenshot of GLIDA showing its linked relations among search pages (**a**, **b**), result pages (**c**, **d**) and an analytical report page (**e**).

similar GPCRs (or 40 ligands) are retrieved and listed with their similarity scores on an analytical report page.

As the similarity search calculation is proceeding, GLIDA illustrates the correlation map (Figure 2e) showing the homologous GPCRs (or ligands) and their ligands (or GPCRs) that are retrieved. This map shows spots that match the GPCRs and their ligands in a two-dimensional matrix. The ordering along the x-axis and the y-axis are calculated respectively by two-way clustering of the GPCRs and the ligands based on their similarities. In particular, the ordering along the x- and y-axis allows users to evaluate information regarding similarities and correlations between GPCRs and ligands simultaneously. By analyzing the correlation patterns between GPCRs and ligands that are illustrated by these maps, we can gain detailed knowledge about their interactions and utilize this information to infer possible candidates for development

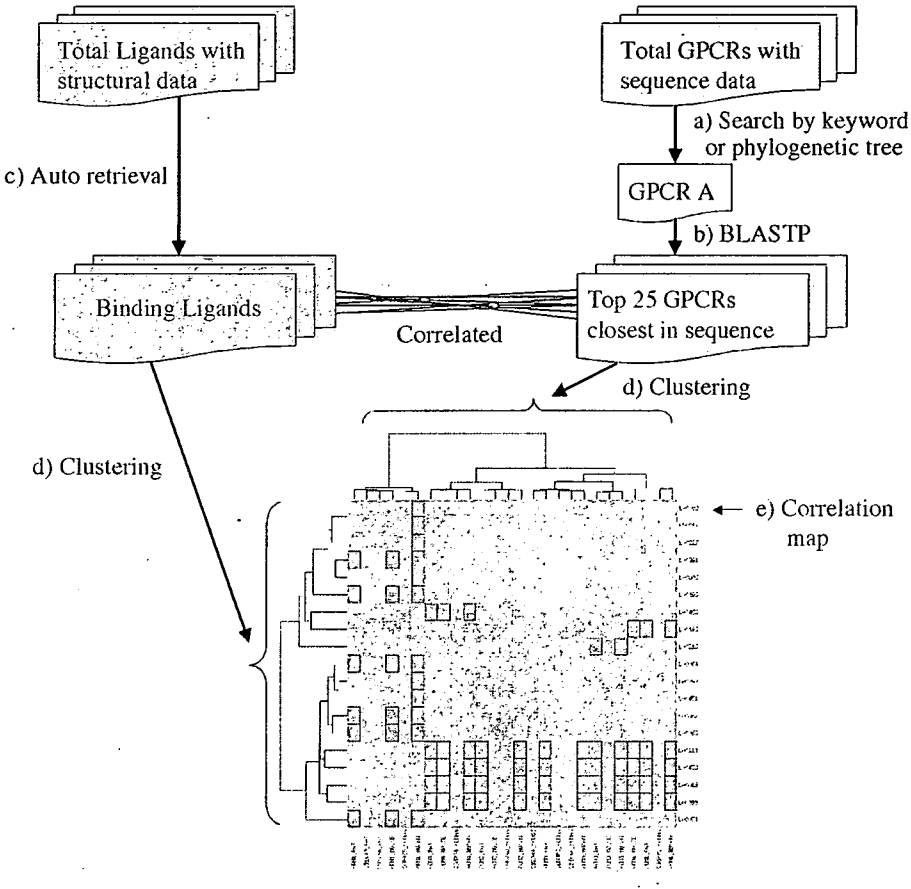

Figure 2. A schematic example of the search and analysis process showing GPCR-ligand correlations produced from a GPCR query using GLIDA. (a) If GPCR A is selected using a keyword search or a guide-tree search on the GPCR search page, its retrieved data will be displayed in its result page, (b) By clicking an analysis button on the result page, a list of the top 25 GPCRs that are most similar in sequence, including GPCR A, are obtained by the BLASTP calculation. (c) The server retrieves a list of corresponding ligands, which are respectively correlated with the 25 GPCRs. (d) Finally, a map is displayed to help visualize the matching spots linking GPCRs with particular ligands. The *x*-axis and *y*-axis respectively indicate the clustering results for GPCRs and ligands, calculated using sequence alignment scores among the GPCRs and structural profile distances among the ligands.

of GPCR-specific drugs. Figure 2 shows an example of the GPCR-ligand search and analysis process starting from a GPCR query using GLIDA.

DISCUSSION AND FUTURE DIRECTIONS

GLIDA provides a unique database for GPCR-related chemical genomic research and drug discovery. GLIDA is distinct from other public chemical genomic databases because it contains original, GPCR-specific chemical entries, although the total scale of its contents is not yet large (Table 1). GLIDA provides several advantages over other databases, in that a search can be started either from a GPCR or from a ligand. Thus, searches may be carried out in a dynamic and user-friendly way. GLIDA's coverage of chemical and biological information simultaneously also provides an advantage to users by saving them the time and labor required to search multiple databases. The ligand search page is another distinct characteristic of GLIDA in that it displays the structural distribution of ligands, and thereby facilitates research on GPCR-related drugs by incorporating structural aspects of the ligand compounds. The analytical report pages resulting from the calculated structural similarities of GPCRs and ligands can give the user deep insights into the GPCR-ligand relationships. The lists of neighboring ligands (or GPCRs) and the correlation maps are useful visualizing tools for analyzing correlations among their structural features and their GPCR-ligand binding properties. Because the GLIDA algorithms can be applied to proteins other than the GPCR family, it may also be considered as a promising database for chemical genomics research.

GLIDA will be updated continuously. In particular, we are planning to computationally extract GPCR-ligand information from the literature and from patents using a text-mining tool, and to increase the number of ligand entries immediately. Further information on ligands from various computable chemical descriptors is currently being incorporated, and GLIDA will be combined with a system for predicting novel ligands of orphan GPCRs in the future. Furthermore, we also plan to carry out XML publication of GLIDA.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, from the Ministry of Health, Labor and Welfare of Japan' and from the 21st Century COE program 'Knowledge information infrastructure for Genome Science'. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Health, Labor and Welfare of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. George,S.R., O'Dowd,B.F. and Lee,S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nature Rev. Drug Discov.*, **1**, 808–820.
2. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
3. Fredholm,B.B., Fleming,W.W., Vanhoutte,P.M. and Godfraind,T. (2002) The role of pharmacology in drug discovery. *Nature Rev. Drug Discov.*, **1**, 237–248.
4. Lipinski,C. and Hopkins,A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.
5. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
6. Zerhouni,E. (2003) Medicine: the NIH roadmap. *Science*, **302**, 63–72.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic. Acids Res.*, **30**, 402–404.
8. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
9. Yang,J., Okuno,Y. and Tsujimoto,G. (2004) GLIDA: GPCR and Ligand Database. *Genome Informatics*, **15**, P057.
10. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
11. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii245–ii251
doi:10.1093/bioinformatics/bti1141

Text Mining

A probabilistic model for mining implicit 'chemical compound–gene' relations from literature

Shanfeng Zhu[1], Yasushi Okuno[2], Gozoh Tsujimoto[2] and Hiroshi Mamitsuka[1,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan and
[2]Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

Motivation: The importance of chemical compounds has been emphasized more in molecular biology, and 'chemical genomics' has attracted a great deal of attention in recent years. Thus an important issue in current molecular biology is to identify biological-related chemical compounds (more specifically, drugs) and genes. Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to find the association of these entities. Our focus is to mine implicit 'chemical compound and gene' relations from the co-occurrence in the literature.

Results: We propose a probabilistic model, called the mixture aspect model (MAM), and an algorithm for estimating its parameters to efficiently handle different types of co-occurrence datasets at once. We examined the performance of our approach not only by a cross-validation using the data generated from the MEDLINE records but also by a test using an independent human-curated dataset of the relationships between chemical compounds and genes in the ChEBI database. We performed experimentation on three different types of co-occurrence datasets (i.e. compound–gene, gene–gene and compound–compound co-occurrences) in both cases. Experimental results have shown that MAM trained by all datasets outperformed any simple model trained by other combinations of datasets with the difference being statistically significant in all cases. In particular, we found that incorporating compound–compound co-occurrences is the most effective in improving the predictive performance. We finally computed the likelihoods of all unknown compound–gene (more specifically, drug–gene) pairs using our approach and selected the top 20 pairs according to the likelihoods. We validated them from biological, medical and pharmaceutical viewpoints.

Contact: mami@kuicr.kyoto-u.ac.jp

1 INTRODUCTION

Traditional molecular biology tells us that genetic information is transferred from DNA to protein and ultimately shows up as protein functions. The final goal of molecular biology in this 'central dogma' is to identify and understand biological activities regulated by proteins so that they can be managed. The most important protein function is to catalyze biochemical reactions for the synthesis of one chemical compound from another. Thus the first step of the above goal may be compared with detecting one or more chemical compounds for which each protein can catalyze.

Recently the importance of chemical compounds has been emphasized more in molecular biology, and a new research field, called 'chemical genomics', has attracted a great deal of attention. In fact, one of the five items to be taken up by the National Institute of Health (NIH) roadmap initiative is a chemoinformatics project for building small molecular libraries. This chemoinformatics project will develop a new compound database of chemical structures and their biological activities, with the idea of promoting pharmaceutical research, such as discovering new drugs. This database, called PubChem, will house compound information on the screening and probe data newly obtained by the Molecular Libraries Screening Centers Network (MLSCN) as well as those from the current scientific literature. A related fact is that databases of chemical compounds and their biochemical reactions have also been developed in recent years. For example, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa *et al.*, 2004), which is a database of metabolic pathways generated by gathering biochemical reactions, has drastically grown both in the size of stored reactions and the number of citations in the biological and medical sciences. European Bioinformatics Institute (EBI) also developed a freely available database of small molecular entities, ChEBI (Brooksbank *et al.*, 2005), which stands for 'Chemical Entities of Biological Interest'.

Thus an important issue in current molecular biology is to identify biological-related chemical compounds and genes, which is a fundamental step of chemical genomics research. Mining biomedical and biological literature databases, such as Medline (Wheeler *et al.*, 2005), for identifying these kinds of biological-related entities has been actively tackled in the last few years (Blasckel *et al.*, 2002; Yandell *et al.*, 2002). Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to identify the association of these entities (Stapley *et al.*, 2000; Jenssen *et al.*, 2001; Chang *et al.*, 2004). This technique is based on the following hypothesis: if a biological entity appears with another biological entity in the same document, these two entities should be biologically related with high probability. This hypothesis was already experimentally testified by many researchers (Jenssen *et al.*, 2001; Chang *et al.*, 2004). We will describe the details of this and related issues in Section 2.

Thus we focus on the co-occurrence information in the literature to discover implicit 'chemical compound–gene' relations, being those which are not in existing compound–gene co-occurrences in the literature but could be discovered from the co-occurrence data. All possible combinations of compounds and genes are very large in number, but obviously known co-occurrences of them are very limited, even though the literature is very abundant in size. Thus we attempt to use not only the co-occurrence of a chemical compound

*To whom correspondence should be addressed.

and a gene but also various other types of co-occurrences data, such as gene–gene and compound–compound co-occurrence data.

We propose a probabilistic model, which we call a mixture aspect model (MAM), coupled with an efficient algorithm for estimating its parameters. MAM is an extension of a probabilistic model, called the aspect model (AM) developed in natural language processing (Hofmann, 2001), with one significant difference. MAM can incorporate different types of co-occurrence data efficiently. More formally, the probabilistic structure of MAM is a weighted mixture of (normalized) AMs, and each component (i.e. AM) handles one type of co-occurrence data. For example, we can have three different components corresponding to three different co-occurrences of compound–gene, gene–gene and compound–compound. These three datasets might be handled by AM by regarding all three as only one type of co-occurrence data. However MAM has roughly two significant advantages, compared with this way of using AM. First, it has a weight for each component, so that the users can control the weight for each co-occurrence dataset. Obviously, AM cannot do this. The second advantage is both time and space efficiency. When we have T datasets all having co-occurrences of N events, MAM considers $T \cdot N^2$ combinations at most, whereas AM must consider a maximum of $T^2 \cdot N^2$. In practice, N (i.e. the number of compounds or genes) reaches at least a few thousands, so that even if T is a relatively small number, this difference would be pronounced.

Our algorithm for estimating the probability parameters of MAM is based on the EM (Expectation–Maximization) algorithm (Dempster *et al.*, 1977) that locally maximizes the likelihoods of given data. Once the probability parameters of MAM are estimated, MAM can predict the likelihood for any pair of events, such as a pair of a chemical compound and a gene. MAM can find new biological-related compound–gene pairs that have not yet been found in current biology and medical literatures.

In our experiments, we generated three types of co-occurrence datasets: gene–gene, compound–compound and compound–gene from the Medline records (Wheeler *et al.*, 2005). We evaluated our method by not only these datasets, but also an independent (human-curated) dataset of chemical compound and gene relationships in the ChEBI database. We first checked the performance of MAM to predict the co-occurrences of compounds and genes by using cross-validation, starting with compound–gene pairs and then adding compound–compound pairs, followed by gene–gene pairs. Experimental results have shown that adding gene–gene (or compound–compound) pairs improved the performance of using compound–gene pairs only, with the difference being statistically significant. In particular, we found that adding compound–compound pairs is the most effective in improving the performance of predicting compound–gene pairs. We then performed the experiment on predicting the biological-related compounds and genes in the ChEBI database, and found that the performance improvement was obtained in almost the same way. These results indicate that combining all these datasets is effective in our problem setting, and that MAM and its learning algorithm are extremely useful for obtaining the results. Finally, we computed the likelihood of each of all unknown compound–gene (more precisely, drug–gene) pairs and selected the top 20 of them according to the likelihoods. We thus showed a list of them that have the highest likelihoods given by MAM trained by all given datasets and examined the validity of these pairs from biological, medical and pharmaceutical viewpoints.

2 RELATED WORK

Mining the Medline text for biomedical knowledge discovery has become a very active field in bioinformatics recently. One of the important applications is to discover the relationship among genes, proteins, disease phenotype and chemical compounds. Co-occurrence in Medline is a simple, effective and popular technique to identify biological relationships among different entities. This technique is based on the hypothesis that entities appearing in the same Medline record are more likely to be biologically related. This hypothesis has been verified by many researchers. Jenssen *et al.* (2001) presented a gene-to-gene co-occurrence network called PubGene using over 10 million Medline records. They randomly selected 500 pairs of genes that co-occurred once and 500 pairs of genes that co-occurred more than five times in the Medline, then manually analyzed the biological relationship of these pairs by expertise. They found that the accuracy of biological relationship identification is ~60% for the first group, and 72% for the second. In further analysis, they found that almost all errors were owing to the failures in gene name recognition. Chang *et al.* (2004) also identified related genes and drugs based on their co-occurrence in the titles and abstracts of publications in Medline. They manually examined the biological relationship of 100 gene–drug pairs. They found that out of the 100 pairs (50 of them with largest number of co-occurrence, and another 50 of them randomly selected), 70 shared some biological relationships. From these studies, we can see that co-occurrence methods can successfully find biological relationships, and most of the failures are because of the difficulty of biological entity name identification in extracting Medline texts. We emphasize that in our experiment we generated our co-occurrence data not directly from Medline texts, but from human curated datasets (for further details, see Section 4), consequently avoiding errors that may occur in gene name or chemical compound name identification.

Some studies have combined co-occurrence methods with natural language processing techniques, such as shallow parsing, full parsing and constructing templates (Yandell *et al.*, 2002; Blasckel *et al.*, 2002). Their goal is to extract and clarify the detailed relationships among biological entities, such as protein–protein interaction (Blasckel *et al.*, 1999), protein–drug interaction (Rindflesch *et al.*, 2000) and gene–mutation pairs (Rebholz-Schuhmann *et al.*, 2004).

Some researchers have, however, attempted to find implicit relationships between biological entities of not having direct co-occurrences in the literature (Wiren *et al.*, 2004). For example, Perez-Iratxeta *et al.* (2002) used the fuzzy set theory to analyze the relationships between the co-occurrence of MeSH terms in different categories and the co-occurrence of a MeSH term and a GO (Gene Ontology) term in Medline records and scored the implicit associations between symptoms of diseases and GO terms by fuzzy relations.

In contrast to these existing approaches, our focus is placed on implicit 'compound–gene' relations in the literature, and our approach is based on statistical learning using a probabilistic model that is an extension of the so-called AM (Hofmann, 2001). This AM has already proved effective in a lot of applications for analyzing co-occurrence data, such as informational retrieval, computational linguistics and collaborative filtering (Hofmann, 2004; Si *et al.*, 2003). We emphasize that our statistical learning based approach develops a noise-robust probabilistic model and a systematic and

efficient algorithm for estimating the parameters of our model from different types of multiple co-occurrence data.

3 METHODS

3.1 Notations

We define the notations that are used throughout this paper. We denote a variable by a capitalized letter, e.g. U, and its value as the same letter in lower case, e.g. u. To explain a particular model for the co-occurrence of a gene and a compound, we define the following symbols in particular. Let G be an observable random variable taking on values $g_1, \ldots, g_S$, each of which corresponds to a gene. Similarly, let C be an observable random variable taking on $c_1, \ldots, c_T$, each of which corresponds to a chemical compound. Let Z be a discrete-valued latent variable taking on values $z_1, \ldots, z_H$, each of which corresponds to a latent cluster, where H is the number of clusters. Let θ be a set of parameters for the model to be optimized in the learning process, and let π be a mixture parameter (i.e. weight) of a component of our model that the users can specify. Let D be a set of all examples.

3.2 Mixture aspect model (MAM)

We begin by describing the AM for two-mode and co-occurrence data (Hofmann, 2001). With latent clusters z_h $(h = 1, \ldots, H)$, AM gives the log-likelihood for a co-occurrence of (u, v) in the following form:

$$\log p(u, v; \theta) = \log \sum_h p(u|z_h;\theta)p(v|z_h;\theta)p(z_h;\theta).$$

So the log-likelihood for D by this model is given as follows:

$$\log p(D;\theta) = \sum_{i,j} N_{i,j} \log p(u_i, v_j;\theta),$$

where $N_{i,j}$ is the number of co-occurrences of (u_i, v_j).

The purpose of this paper is to handle multiple different types of co-occurrence data with overlapping variable. More concretely, we can assume that we have two datasets, in which one has two random variables U and V, and the other has V and W. For these two datasets, we now define a new probabilistic model that is a mixture of two AMs, which we call two-component mixture aspect model (2MAM). The log-likelihood for D with two datasets for this model is given as follows:

$$\log p(D;\theta) = \pi_{UV} \sum_{i,j} \frac{N_{i,j}}{N_{UV}} \log \sum_h p(u_i|z_h;\theta)p(v_j|z_h;\theta)p(z_h;\theta) + \pi_{VW} \sum_{j,k} \frac{M_{j,k}}{N_{VW}} \log \sum_h p(v_j|z_h;\theta)p(w_k|z_h;\theta)p(z_h;\theta),$$

where $\pi_{UV} + \pi_{VW} = 1$ for U and V, $N_{i,j}$ and $M_{j,k}$ are the number of co-occurrences of (u_i, v_j) and (v_j, w_k), respectively, $N_{UV} = \sum_{i,j} N_{i,j}$ for U and V, and $N_{VW} = \sum_{j,k} M_{j,k}$ for V and W.

We note that both the first and second terms in this equation use the same probability parameter $p(v|z;\theta)$. Therefore, the parameter must be controlled by both datasets. We can easily see that this mixture model for two datasets can be extended to a mixture model for an arbitrary number of datasets. We note that if each of these datasets has a random variable that appears in more than one dataset, this model is different from AM. This is particularly true when estimating its parameters, each of which corresponds to a variable appearing more than once. These parameters must be trained (controlled) by more than one dataset. The detailed algorithm for estimating this type of parameters is described for a particular case of the co-occurrence of a chemical compound and a gene in Section 3.4.

3.3 Mixture aspect model for predicting co-occurrences of compound–gene

When there is only one type of co-occurrence data (i.e. compound–gene pairs), this dataset can be handled by AM. If another dataset like gene–gene pairs is added to this dataset, these two datasets can be handled by 2MAM. For example, if we have two types of co-occurrence data, such as compound–gene and gene–gene pairs, the log-likelihood for all the data D by 2MAM is written as follows:

$$\log p(D;\theta) = \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} \log \sum_h p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta) + \pi_{GG} \sum_{j,j'} \frac{M_{j,j'}}{N_{GG}} \log \sum_h p(g_j|z_h;\theta)p(g_{j'}|z_h;\theta)p(z_h;\theta),$$

where $N_{CG} = \sum_{i,j} N_{i,j}$ and $N_{i,j}$ is the number of co-occurrences of (c_i, g_j), and $N_{GG} = \sum_{j,j'} M_{j,j'}$ and $M_{j,j'}$ is the number of co-occurrences of $(g_j, g_{j'})$.

In this paper, we consider three types of co-occurrence data: compound–gene, gene–gene and compound–compound pairs. We also present a probabilistic model for this data, which we call three-component mixture aspect model (3MAM). The log-likelihood for all data D can be given by 3MAM as follows:

$$\log p(D;\theta) = \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} \log \sum_h p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta) + \pi_{GG} \sum_{j,j'} \frac{M_{j,j'}}{N_{GG}} \log \sum_h p(g_j|z_h;\theta)p(g_{j'}|z_h;\theta)p(z_h;\theta) + \pi_{CC} \sum_{i,i'} \frac{L_{i,i'}}{N_{CC}} \log \sum_h p(c_i|z_h;\theta)p(c_{i'}|z_h;\theta)p(z_h;\theta).$$

In the above equation, $\pi_{CG} + \pi_{GG} + \pi_{CC} = 1$, $N_{CC} = \sum_{i,i'} L_{i,i'}$ and $L_{i,i'}$ is the number of $(c_i, c_{i'})$ pairs.

In this paper, even though we used three types of data, it is evident that 3MAM can incorporate another type of co-occurrence data if it can improve the predictive performance of 3MAM.

3.4 Estimating probability parameters

Given training data D and the number of clusters H, a popular criterion for estimating the probabilities of a probabilistic model is the maximum likelihood (ML). Parameters are estimated to maximize the log-likelihood of data D:

$$\theta^{ML} = \arg\max_\theta \log p(D;\theta).$$

The most popular approach for obtaining an ML estimator of a probabilistic model is a time-efficient general scheme called the EM (Expectation–Maximization) algorithm (Dempster *et al.*, 1977) that provides a local maximum. In general, the EM algorithm starts with a random set of initial parameter values and iterates both the expectation step (E-step) and the maximization step (M-step) alternately until a certain convergence criterion is satisfied.

3.4.1 AM We begin to explain the EM algorithm for AM for only one type of co-occurrence data, i.e. compound–gene pairs. The log-likelihood for D is given in Section 3.2, and the E- and M-steps can be given as follows:

E-step:

$$p(z_h|c_i, g_j;\theta) = \frac{p(c_i|z_h;\theta)p(g_j|z_h;\theta)p(z_h;\theta)}{\sum_{h'} p(c_i|z_{h'};\theta)p(g_j|z_{h'};\theta)p(z_{h'};\theta)}.$$

M-step:

$$\theta_{c_i|z_h} \propto \sum_j N_{i,j}\, p(z_h|c_i, g_j;\theta_{\text{old}}),$$

$$\theta_{g_j|z_h} \propto \sum_i N_{i,j}\, p(z_h|c_i, g_j;\theta_{\text{old}}),$$

$$\theta_{z_h} \propto \sum_{i,j} N_{i,j}\, p(z_h|c_i, g_j;\theta_{\text{old}}).$$