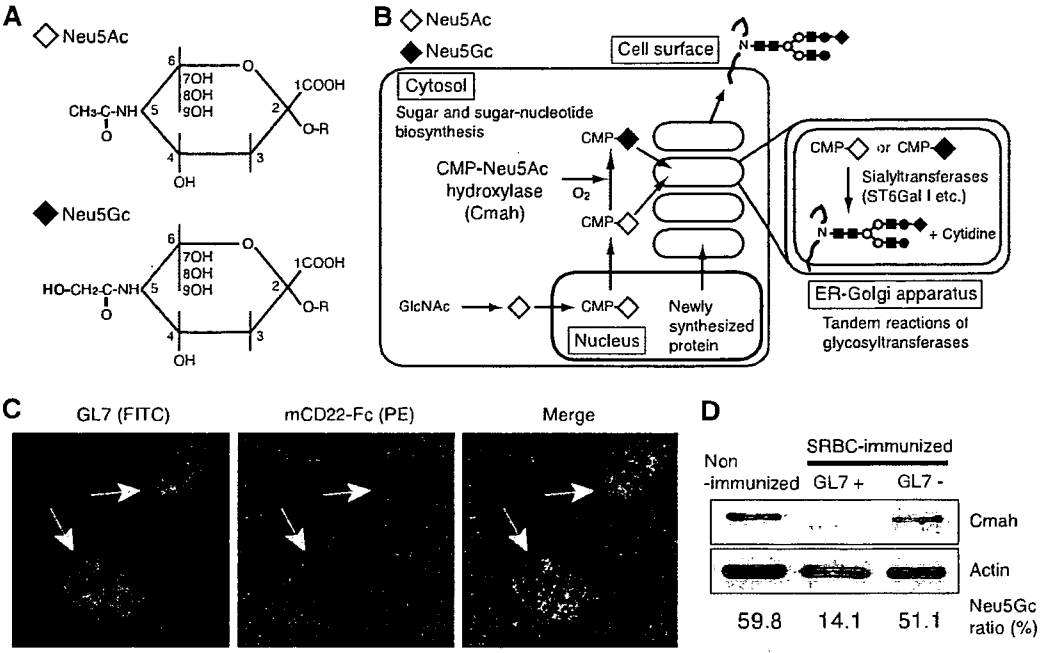

FIG. 3. Change in Sia species in germinal centers. (A) Structural differences between two major molecular species of Sia. The metabolic precursor Neu5Ac and its modified form Neu5Gc differ only by an oxygen atom at the C-5 position. The conversion of CMP-Neu5Ac to CMP-Neu5Gc is catalyzed by the enzyme Cmah. (B) Biosynthesis of sialylated glycoproteins destined for the cell surface. Cytosolic metabolism of Sia is responsible for the abundance of the molecular species of Sia on the cell surface, as a given ratio of cytosolic CMP-Sia is imported into the Golgi apparatus and then used by the sialyltransferases for the biosynthesis of glycoproteins en route to the plasma membrane. (C) Loss of CD22 ligand in germinal centers. Spleen sections of SRBC-immunized mice (10 days after immunization) were costained with FITC-conjugated GL7 and mCD22-Fc precomplexed with R-PE-conjugated anti-human IgG. The mCD22-Fc is a chimeric probe that binds to the CD22 ligand. Arrows indicate germinal centers. (D) Downregulation of Cmah expression in germinal center B cells. GL7-positive germinal center cells and GL7-negative cells were prepared from a B-cell-enriched fraction derived from the spleen of a mouse 12 days after immunization with SRBC. Ultracentrifugation supernatant fractions (cytosolic fractions) of untreated mouse B cells (nonimmunized; control), GL7-positive B cells (GL7+), and GL7-negative B cells (GL7−) were subjected to immunoblotting with anti-mouse Cmah antibody and antiactin antibody (to demonstrate equal loading of samples). The Neu5Gc/(Neu5Ac+Neu5Gc) ratio of the ultracentrifugation pellets (membrane fractions) of each cell type was measured by HPLC.

LPS-stimulated B blasts (Fig. 4C), further confirming the responsibility of *Cmah* for the repression of the appearance of the GL7 epitope. After 48 h of stimulation with LPS, *Gapdh* expression increased by about 30% (Fig. 4A). This may be attributable to the blastic transformation of LPS-stimulated proliferating B cells (B blasts), which produce much more cytosolic space and subsequent metabolism than resting B cells. GL7 staining of LPS-stimulated B cells showed heterogeneity in the degree of staining. Thus, cells used to prepare RNA for this real-time PCR experiment were a mixture of $GL7^{high}$ and $GL7^{low}$ cells. When these finding are taken into consideration, the reduction of *Cmah* expression in $GL7^{high}$ germinal center B cells could be more drastic. The expression of *Cd22*, an α2,6-linked Neu5Gc binding protein, on B cells was reduced to around 40% after 48 h, even though its cell surface expression was still comparable to that of unstimulated cells in flow cytometry (Fig. 4A and B).

Targeted disruption of the Cmah gene in mice. To further examine the in vivo function of Neu5Gc-bearing glycans, we targeted the *Cmah* gene in mice by inserting the neomycin resistance gene cassette into the second coding exon (Fig. 5A and B). Biochemical analysis of mouse tissues made it clear that gene inactivation was achieved, as homozygous null mice lacked enzyme expression in the liver ultracentrifugation supernatant, as shown by immunoblotting using antiserum against the N terminus of Cmah (Fig. 5C). We also did not detect a signal with a different molecular mass from the *Cmah*-disrupted allele. We further analyzed the effect of the enzyme deficiency on the level of its product by HPLC. *Cmah*-null tissues lacked detectable production of Neu5Gc throughout the normal adult mouse body (Fig. 5D). We concluded that the *Cmah* gene is indispensable for most of the cellular biosynthesis of Neu5Gc, as previously suggested in humans (6, 22). The development of the null mice appeared to be grossly normal; however, the numbers of null and heterozygote mutant offspring derived from F_1 crosses were subtly reduced from wild-type littermates in the rate expected from Mendelian rules (wild-type:heterozygote:null, 508:881:449), even though the mice were bred in a specific-pathogen-free mouse facility.

Normal B-cell maturation in *Cmah*-deficient mice. We found that Neu5Gc expression was severely repressed during B-cell activation in germinal centers, and thus we examined the development of the immune system in *Cmah*-null mice. In null mice, the values from blood counts and blood chemistry analyses were normal in every category examined (white blood cell, red blood cell, blood hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular

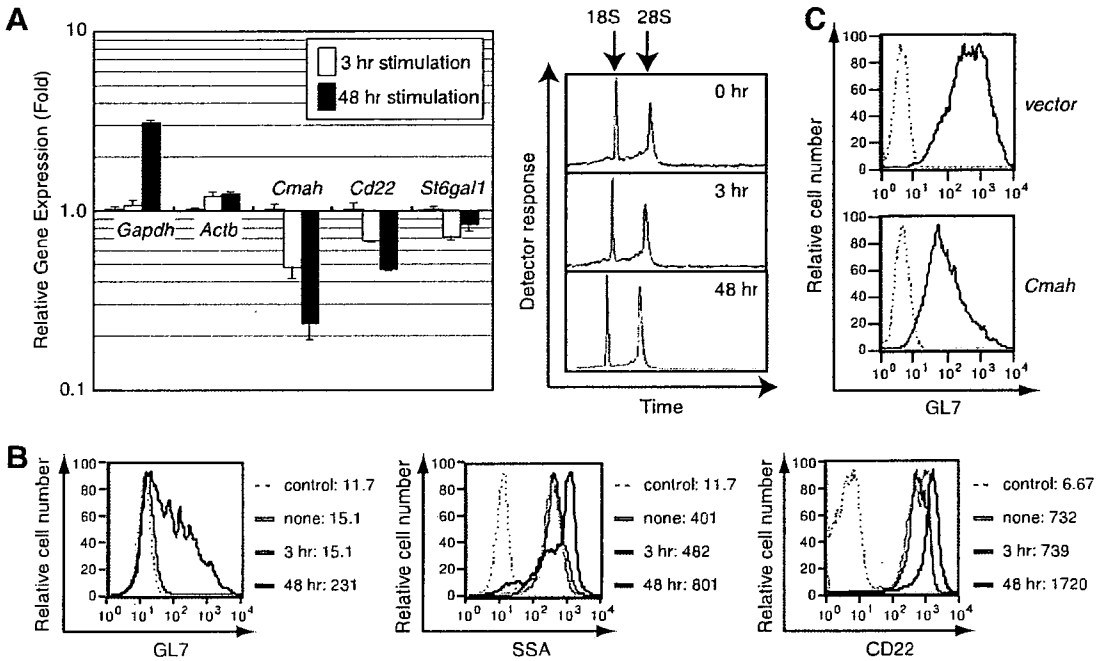

FIG. 4. Downregulation of *Cmah* mRNA in primary cultured B cell blasts, causing GL7 epitope expression. (A and B) Cmah repression caused by in vitro B-cell activation. Splenic B cells were stimulated with 30 μg/ml LPS for the indicated times. Reverse-transcribed cDNAs prepared from total RNA of these cells were subjected to real-time PCR analysis. The right box shows capillary electrophoresis analysis results indicating the lack of RNA degradation in the RNA used for cDNA synthesis. The expression levels of the mRNA of *Gapdh*, *Actb* (beta actin), *Cmah*, *Cd22*, and *St6gal1* are shown as the relative change compared with the mRNA expression in untreated B cells (A). The same set of cells that was used to prepare total RNA was stained with FITC-conjugated GL7, SSA, and anti-CD22 (B). The MFI of each stain is indicated at the right of each panel. (C) Reduced expression of the GL7 epitope by ectopic *Cmah* expression. *Cmah* was ectopically expressed in LPS-stimulated splenic B blasts using retrovirus. Retrovirus-infected cells were sorted and stained with FITC-conjugated GL7.

hemoglobin concentrate, and platelet). The development of immune cells in *Cmah*-null mice appeared to be grossly normal for T-cell and B-cell maturation, as indicated by routine flow cytometric analysis profiles. The indicators analyzed included the ratio of B1 to B2 cells, the ratio of marginal zone to follicular B cells, and the expression level of surface IgM, major histocompatibility complex class II (MHC-II), and CD22 (Fig. 5E; also see Table S2 in the supplemental material). We also examined the staining profile of activation markers for B cells. The only probe with a significant change in the null B cells was GL7 (Fig. 5F), which recognizes α2,6-linked Neu5Ac on LacNAc (Fig. 2C). Serum Ig measurements using the sandwich ELISA method revealed a significant ($P = 0.074$) increase in the serum IgG1 level of the *Cmah*-null population (Table 2).

Hyperreactive B cells in *Cmah*-deficient mice. We examined the mouse phenotype after immunization. When mice were immunized with the T-dependent antigen DNP-KLH or the T-independent (II) antigen DNP-Ficoll, the response to the T-independent antigen (serum titer against the hapten, DNP conjugated to BSA, by ELISA) was enhanced in null mice compared with controls, most prominently for IgM but also significantly for IgG3 (Fig. 6A). In contrast, the T-dependent response of the null group to DNP-KLH with potent complete Freund's adjuvant was not significantly different from that of the control group (Fig. 6B). Thus, the Neu5Gc deficiency in B cells resulted in a hyperresponsive phenotype to the T-independent antigen, indicating the importance of Neu5Gc-mediated negative regulation of B-cell activation. To further study the regulatory mechanism of the B-cell response by Neu5Gc-bearing glycans, mature splenic B cells were isolated and used in an in vitro proliferation assay with various stimuli. In this assay, compared with the cells from littermate controls, *Cmah*-null B cells proliferated robustly in response to the $F(ab')_2$ fragment against BCR (anti-μ chain), regardless of interleukin-4 (IL-4) addition (Fig. 6C). The FBS routinely used to support the cell culture contains around 5% Neu5Gc and represents a possible supply for *Cmah*-null cells. Therefore, we also examined the difference in proliferation using serum from chickens and humans, which contain only Neu5Ac as a Sia source (as determined by HPLC analysis [data not shown]). Under such conditions, *Cmah*-null B cells also showed augmented proliferation compared with control cells, although the degree of overall proliferation was much stronger in medium with FBS, perhaps because of differences in the growth factor(s) contained in each type of serum (data not shown). When anti-CD40 was used as the stimulus in a model mimicking T-dependent stimulation, B cells with both genotypes proliferated equally (data not shown); thus, Neu5Gc glycan-mediated regulation appeared to be stimulation dependent, and the effect seemed to be more related to T-independent activation. When T-cell proliferation was assessed using anti-CD3 as the stimulant, both *Cmah*-null and control splenic T cells proliferated to the same extent (see Fig. S2A in the supplemental material). No obvious bias toward either Th1 or Th2 was found in the cytokine production pattern of anti-CD3-stimulated *Cmah*-null T cells; however, a significant reduction of gamma

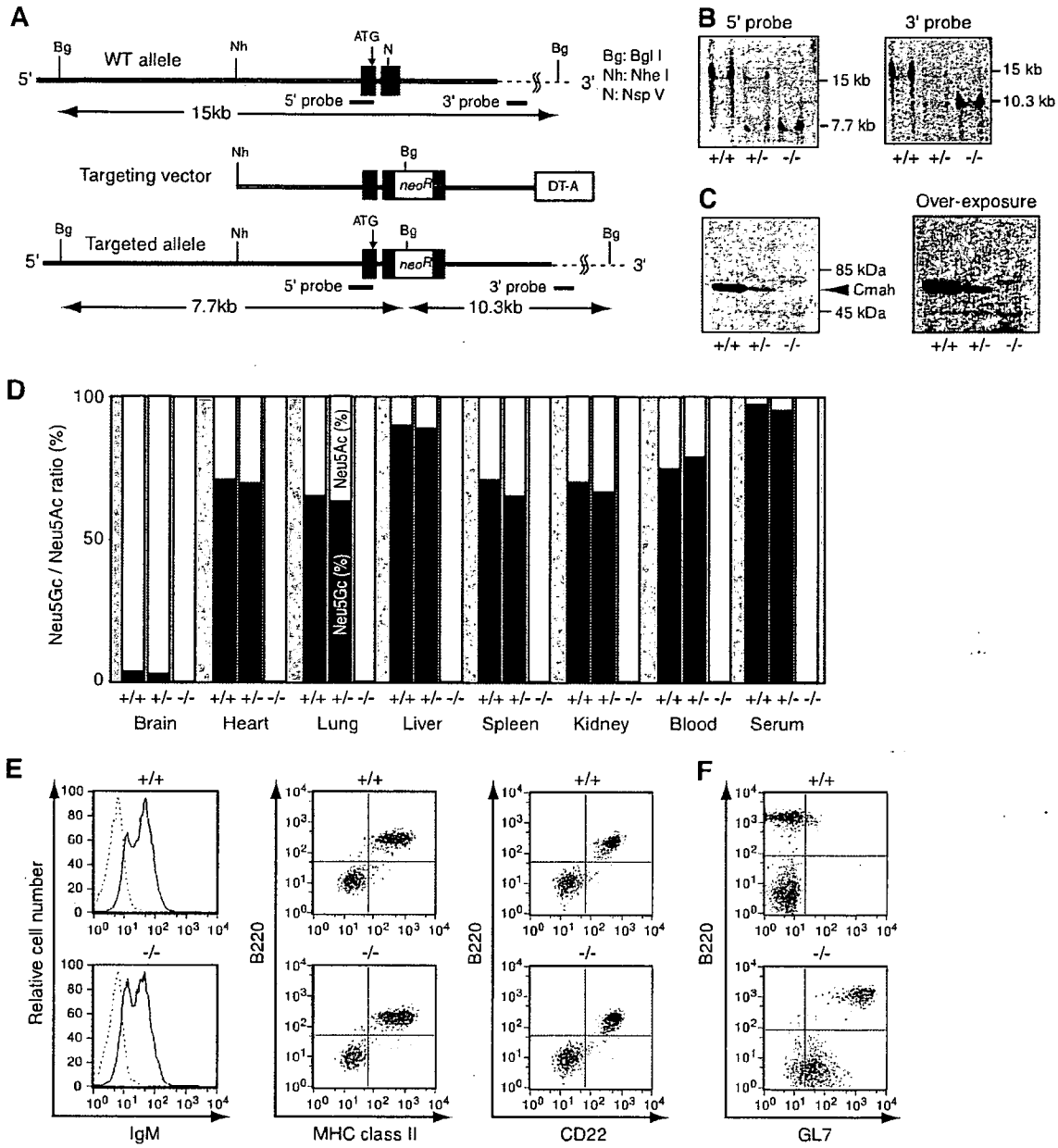

FIG. 5. Generation and biochemical analyses of *Cmah* knockout mice. (A) Allele for targeted *Cmah*. A targeting vector was created by inserting the *PGK-neoR* cassette into the NspV site of the second coding exon (exon 5) of the *Cmah* gene. (B) Genotype of homologous recombination of selected ES cell lines. The genotypes of G418-selected cell lines were determined by Southern blotting analysis of genomic DNA digested with BglI, using both radiolabeled 5′ internal and 3′ external probes. The genetic status of the *Cmah* allele is indicated as follows: +/+, wild type; +/−, heterozygote; and −/−, null (B to F). (C) Loss of Cmah enzyme demonstrated by immunoblotting analysis of liver cytosolic fractions. Ultracentrifugation supernatant fractions of livers were assessed for the expression of Cmah using anti-mouse Cmah immunoblotting. Staining of a ~67-kDa band (arrowhead) in wild-type and heterozygote livers represents the signal of Cmah, which is not detectable in *Cmah*-null liver samples. (D) Loss of Neu5Gc production throughout the body in mutated mice. Acid-hydrolyzed Sia from the indicated tissues was derivatized using DMB, and the ratios of Neu5Ac and Neu5Gc to total Sia were measured by reverse-phase HPLC. Solid columns represent the percentage of Neu5Gc in various tissues, and open columns represent the percentage of Neu5Ac. The detection limit for Neu5Gc in this assay was around 0.1%. (E) Flow cytometry profile of *Cmah*-null mice splenocytes. The expression of IgM, MHC-II (I-A and I-E), and CD22 on splenocytes from wild-type and *Cmah*-null mice was detected by flow cytometry. In anti-MHC-II and anti-CD22 staining, splenocytes were costained with anti-B220, a marker for B cells. (F) Strong expression of the GL7 epitope on *Cmah*-null mice B cells. Splenocytes from wild-type and *Cmah*-null mice were costained with anti-B220 and GL7 and subjected to flow cytometry.

interferon and IL-4 secretion was found in these cells (see Fig. S2B in the supplemental material). Based on these findings, we conclude that B cells from *Cmah*-deficient mice acquire hyperresponsiveness to stimuli, and thus the null animals show hyperresponsiveness (hyperproduction of antibodies) to the T-independent antigen.

Retrovirus-mediated rescue of hyperproliferative B-cell response in null mice. The LPS stimulation-dependent prolifer-

TABLE 2. Serum Ig isotype levels of nonimmunized *Cmah*-null mice

Isotype	Serum Ig level (μg/ml)[a]		
	Wild type	Heterozygote	*Cmah* null
IgM	169.6 ± 24.7	205.3 ± 38.9	190.0 ± 33.3
IgG1[b]	115.5 ± 14.9	151.3 ± 19.5	197.4 ± 41.2
IgG3	20.4 ± 2.3	23.9 ± 3.1	19.6 ± 3.8
IgA	242.7 ± 9.7	280.0 ± 29.2	260.2 ± 10.6

[a] Serum Ig levels were measured in nonimmunized mice at 7 to 13 weeks of age (at least 20 per genotype). Values are expressed as the means ± standard errors of the means.

[b] The serum IgG1 level was slightly increased in *Cmah*-null mice (Student's *t* test; $P = 0.074$ for wild type versus *Cmah* null).

ative response is also related to the T-independent response. In *Cmah*-null B cells, LPS stimulation caused enhanced proliferation (Fig. 7A). Given that LPS induces a considerable percentage of cells to progress through the cell cycle, retroviral infection-mediated gene rescue is possible. To determine whether the B-cell hyperreactivity was caused by the *Cmah* mutation, we expressed Cmah ectopically in LPS-stimulated proliferating *Cmah*-null B cells and found that the introduction of *Cmah* did result in repression of the hyperproliferation of *Cmah*-null B cells (Fig. 7B). This rescued hyperproliferative phenotype produced by ectopic Cmah expression in *Cmah*-null B cells indicates that the phenotypes in *Cmah*-null mice are caused by the loss of Cmah expression and probably not by effects on the expression of other genes owing to the insertion of the neomycin-resistance cassette during ES cell-based mutagenesis. This conclusion is also supported by the consistent phenotype resulting from the *Cmah*-disrupted allele in an extensively backcrossed C57BL/6J background. Moreover, our RT-PCR results confirmed equal expression levels of *Lrrc16* and *6330500D04Rik*, the genes located adjacent to the *Cmah* gene in the genome, in splenocytes of wild-type and *Cmah*-null mice (data not shown). To infect control and *Cmah*-encoding retrovirus, we used the same *Cmah*-null B-cell fractions. Since attenuated proliferation was found in *Cmah*-infected B-cell blasts, the augmented proliferation found in the *Cmah*-null B cells compared to the wild type (Fig. 6C) was not due to any subtle population difference in the B-cell fraction. Thus, we conclude that Cmah expression determines the proliferation of B cells when activated and that the difference in the in vivo response to the T-independent antigen is caused by differential expression of Neu5Gc in B cells.

Normal germinal center formation in the Cmah-deficient spleen. As shown in Fig. 5F, *Cmah*-null B cells strongly express the GL7 epitope, and GL7 has been used to detect the germinal center reaction in mice (5, 17, 41, 55). GL7-negative mature B cells turn GL7 positive during germinal center reactions upon T-dependent immunization. Germinal center B cells further develop to CD79b-positive memory B cells, which are no longer stained by GL7 (52). Therefore, it was of interest to assess whether these *Cmah*-null mice could undergo normal germinal center formation. PNA binds to glycan moieties with a terminal β-galactose residue at the core-1 branch of O-linked glycans, and it has been used as a marker for germinal center B cells (8). We compared the staining profiles of the two germinal center probes using spleen sections of wild-type and *Cmah*-null mice, either with or without SRBC immunization. In the wild-type spleen without immunization, PNA showed some staining in the marginal zone area, whereas GL7 did not (Fig. 8A). As expected from flow cytometric staining, GL7 widely stained the B-cell zone of the *Cmah*-null spleen even without immunization (Fig. 8A). When wild-type mice were immunized with SRBC, in addition to the marginal zone staining, intense PNA-positive germinal center follicles were observed. When PNA and GL7 staining results were compared on merged images, PNA appeared to stain a larger number of cells in the germinal center than did GL7, which stained a limited number of cells in the area, most probably centrocytes (Fig. 8B). In SRBC-immunized *Cmah*-null spleen, the staining pattern of GL7 was not different from that of the nonimmunized spleen section. These results confirmed that the appearance of GL7 epitope via the conversion of Neu5Gc to Neu5Ac is an activation-dependent event in the wild-type spleen, whereas *Cmah*-null mice lose Neu5Gc throughout; thus, *Cmah*-null spleen was stained by GL7 regardless of the immunization. In contrast, with GL7 staining, the *Cmah*-null spleen formed PNA-positive follicles that resembled the germinal centers of wild-type sections (Fig. 8B). These results suggest that *Cmah*-null mice could develop germinal centers upon SRBC immunization, which is consistent with the normal T-dependent antigen response found in *Cmah*-null mice.

Change in ligand expression for Siglecs in Cmah-null mice. The cell surface change in Sia species (Neu5Gc to Neu5Ac) by *Cmah* disruption could potentially cause a global change in sialylated glycan recognition throughout the body, as Neu5Gc is the predominant form of Sia in the mouse body, except in the neural system (Fig. 5D). In the immune system, various members of the Siglec family of Sia-binding lectins are expressed in a variety of immune cells. The counter-receptors for sialylated glycans affected by the C-5 position oxygen atom include sialoadhesin (Siglec-1, or CD169), which requires α2,3-linked Neu5Ac on galactose as a ligand (10), and CD22 (Siglec-2), which has a strong preference for Neu5Gc over Neu5Ac in the α2,6 linkage to LacNAc in mice (3, 26, 44, 50). To explore the change in ligand expression for Siglecs in *Cmah*-null mice, we prepared Siglec-Fc fusion probes that were free from intramolecular sialylation. In null B cells, the expression of the CD22 ligand was reduced roughly 20-fold compared with that in wild-type cells (Fig. 9A). We also histochemically examined the expression of the CD22 ligand on spleen sections from *Cmah*-null mice. Regardless of immunization, the mCD22-Fc probe failed to detect any staining in the sections of *Cmah*-null spleen, as in the germinal centers of immunized wild-type mice (Fig. 9B). Therefore, *Cmah* disruption caused the reduction of the optimal ligand for CD22. At the same time, ligand expression for sialoadhesin was greatly increased in *Cmah*-null mice (Fig. 9A). Sialoadhesin is expressed on macrophages, whereas CD22 is expressed on B cells. Ligand(s) for Siglec-G, another Siglec molecule presumably expressed on B cells, was not detected on B cells (data not shown); thus, the Siglec-related effects in *Cmah*-null B cells could be a loss of CD22 ligand.

Normal tyrosine phosphorylation upon BCR cross-linking in Cmah-null B cells. In addition to its biochemical activity as a lectin, CD22 also contains immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in its cytoplasmic tail (4, 48). These ITIMs are phosphorylated as part of the phosphorylation

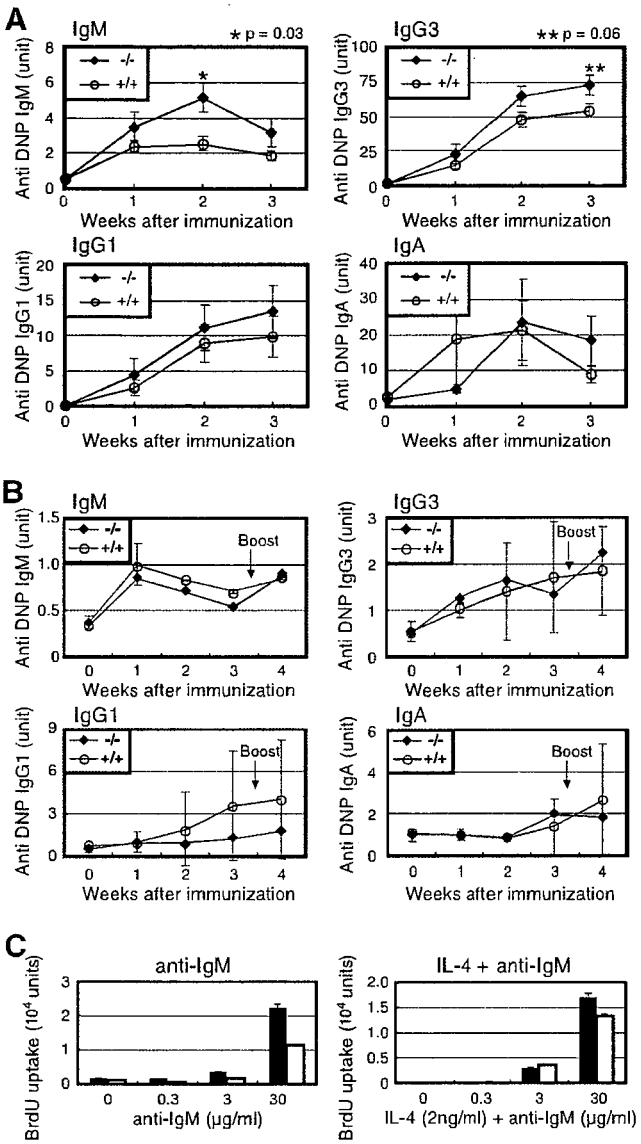

FIG. 6. Hyperresponsive phenotypes of *Cmah*-null mice. (A) T-independent hyperresponse of *Cmah*-null mice. DNP-Ficoll was used to immunize 8-week-old mice. Serum was collected each week and analyzed for reactivity with DNP-conjugated BSA coated on ELISA plates. The titer of hapten-reacting mouse Igs from each animal was determined by isotype-specific ELISA. The measured optical density at 405 nm was normalized to anti-DNP units by comparison with the value from standard pooled serum against DNP on the same plate. The results are presented as the mean responses of 10 animals for each genotype measured in two sets of experiments. The bars represent standard errors of the means. Open circles indicate the responses of wild-type mice, and filled diamonds indicate the responses of *Cmah*-null mice for each isotype. Genotypes are indicated as follows: +/+, wild-type; −/−, *Cmah*-null (A and B). (B) Normal T-dependent immune response of *Cmah*-null mice. DNP-KLH in complete Freund's adjuvant was used to immunize 8-week-old mice. The titers of hapten-reacting mouse Igs from each animal were determined by isotype-specific ELISA as above. Arrows indicate the time of secondary immunization with DNP-KLH. Open circles indicate the responses of wild-type mice, and filled diamonds indicate the responses of *Cmah*-null mice for each isotype. (C) In vitro hyperproliferation response of *Cmah*-null B cells. Splenic B cells from wild-type (open columns) and *Cmah*-null (filled columns) mice were assessed for proliferation using the $F(ab')_2$ fragment of anti-mouse IgM (μ chain) or anti-IgM plus

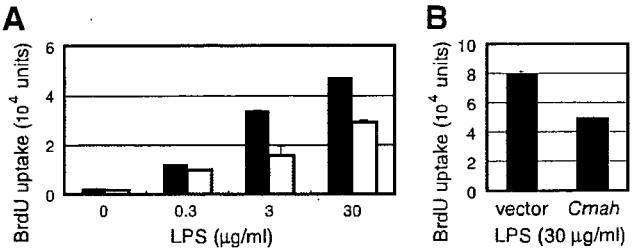

FIG. 7. Rescue of augmented proliferation of *Cmah*-null B cells by Cmah expression. (A) In vitro hyperproliferation response of *Cmah*-null B cells to LPS. Splenic B cells from wild-type (open columns) and *Cmah*-null mice (filled columns) were assessed for proliferation using LPS from *S. enterica* serovar Enteritidis as the stimulating reagent. Proliferation assays were performed as described in the legend of Fig. 6C. Data are shown as the means of triplicate cultures, and the bars represent standard errors of the means. (B) Reduction of B-cell proliferation by retrovirus-mediated Cmah expression. Cmah was ectopically expressed by mouse stem cell virus in *Cmah*-null splenic LPS B blasts. After being cultured for 2.5 days in the presence of 30 μg/ml LPS, the virus-infected B cells were subjected to a proliferation assay. As a control, cells were infected with an empty vector. Data are shown as the means of triplicate cultures, and the bars represent standard errors of the means.

cascade after BCR cross-linking. CD22 recruits SHP-1 tyrosine phosphatase to negatively regulate BCR signaling (11, 39). Given that CD22 is believed to be a regulator of BCR signaling and B-cell apoptosis (7, 13, 34, 58, 63) and that the level of BCR in *Cmah*-null mice was not different from that of the wild-type control (Fig. 5E), we analyzed the immediate-early CD22 phosphorylation status of mature B cells upon activation by BCR ligation. The overall tyrosine phosphorylation profile of B cells was not different for the two types of mice when the $F(ab')_2$ fragment of the anti-IgM (anti-μ chain) was used as a stimulant (Fig. 9C), although this may not be an optimal stimulant for CD22 phosphorylation (21). We further confirmed the tyrosine phosphorylation of CD22, possibly by Lyn kinase at the ITIM motif, upon BCR ligation. Consistently, the phosphorylation profile of CD22 assessed after immunoprecipitation by immunoblotting with an anti-phosphotyrosine antibody was almost identical in *Cmah*-null B cells and controls (Fig. 9D). In contrast, *Cmah*-null B cells showed augmented proliferation when a combination of tetradecanoyl phorbol acetate and ionomycin was used as a stimulant to directly activate classical protein kinase C(s). Thus, a downstream event of protein kinase C activation probably affects the hyperproliferative phenotype of *Cmah*-null B cells (Fig. 9E).

DISCUSSION

Change in Sia species in the germinal center. In the present study, we showed that activated B cells undergo a dramatic

2 ng/ml IL-4 as stimulating reagents. After stimulation for 24 h, BrdU was added. Following incubation overnight, incorporated BrdU was detected by ELISA. Data are shown as the means of triplicate cultures, and the bars represent standard errors of the means. The results shown here were obtained in one of the experiments using 10% FBS-containing medium.

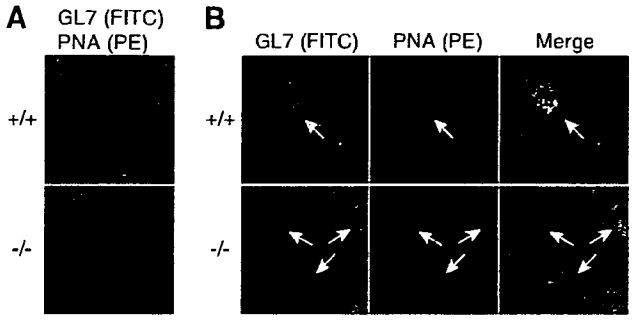

FIG. 8. Changes in staining of germinal center markers in normal or SRBC-immunized *Cmah*-null mice. (A) Histochemical analyses of spleen sections without immunization. Spleen sections from wild-type and *Cmah*-null mice were costained with FITC-conjugated GL7 and biotin-conjugated PNA visualized by R-PE-conjugated streptavidin. (B) Histochemical analyses of spleen sections after T-dependent immunization. Wild-type and *Cmah*-null mice were immunized with SRBC, and the spleens were removed 8 days after immunization. The frozen spleen sections were costained with FITC-conjugated GL7 and biotin-conjugated PNA followed by R-PE-conjugated streptavidin. Arrows indicate germinal centers. Genotypes are indicated as follows: +/+, wild type; −/−, *Cmah* null.

alteration of surface-sialylated glycans and that this alteration of Sia species from Neu5Gc to Neu5Ac can be probed with GL7. This is the first report regarding the epitope identification of GL7, which is routinely used to stain germinal center B cells in mice. We demonstrated that the GL7 epitope is the Neu5Acα2-6LacNAc-containing N-glycan, which is prominently expressed in activated B cells upon the repression of Cmah. Gain of GL7 epitope expression coincided with the loss of optimal ligand expression of CD22 in germinal center B cells, presumably centrocytes. Considering the rather strong degree of GL7 positivity in germinal center B cells in comparison with in vitro stimulated B-cell blasts, the degree of Cmah reduction might have been severe in these cells. In general, it is thought that Neu5Gc is easy to accumulate but difficult to turn over in cells. This is attributable to the one-way direction of the metabolic pathway; Neu5Gc is biosynthesized by Cmah from Neu5Ac (24, 36, 54), whereas no conversion activity was found to biosynthesize Neu5Ac from Neu5Gc. Therefore, the reduction of Neu5Gc found in the GL7-enriched germinal center cells is remarkable. Such rapid clearance of Neu5Gc could be attributable to several characteristics of germinal center cells. Most importantly, as shown in Fig. 3D, these cells repressed Cmah, the enzyme responsible for the de novo biosynthesis of Neu5Gc. Moreover, because lymphocytes are small cells with limited cytosolic space, the cytosolic pool of Sia in these cells is likely limited and easily turned over. In addition, centrocytes undergo extremely fast cell cycles (66), which probably leads to rapid passive dilution of the cytosolic pool in these cells. At the same time, new protein synthesis should be a primary event that happens in germinal center B cells, as shown by cDNA microarray analysis (51). The transcriptional repression of Cmah, together with these features of germinal center cells, could contribute to the efficient conversion of the major Sia species from Neu5Gc to Neu5Ac.

Negative regulation of B-cell activation by Cmah and its product, Neu5Gc. To clarify the biological role of Neu5Gc in vivo, we disrupted the *Cmah* gene in mice and examined their B-cell activation phenotypes. *Cmah*-null mice showed a hyperreactive B-cell phenotype to T-independent stimulation. In contrast, the T-dependent immunization response was similar to that in wild-type mice. This is consistent with the findings that Cmah expression is severely repressed in the germinal centers of wild-type spleen upon T-dependent immunization and that *Cmah*-null mice could develop follicles stained with PNA, another marker for germinal centers. Forced expression of Cmah caused repression of the proliferative response of *Cmah*-null B cells, indicating that Neu5Gc-containing sialoglycan functions to suppress B-cell reactivity though the mechanism is still unknown. This suppression via Neu5Gc-containing sialoglycan appears to be canceled by Cmah repression in germinal center B cells that are "activation committed" or "activation competent." The hyperreactive B-cell phenotypes observed in *Cmah*-null B cells could mirror differences in cellular reactivity between germinal center and nongerminal center B cells, as indicated by differential cell surface expression of the GL7 epitope (5).

Possible change in sialoglycan-receptor interaction in *Cmah*-null mice. As *Cmah* disruption results in a single oxygen atom change in these mice, it is expected that this mutation leaves both the Sia amount and Sia linkage intact in terms of sialoglycans, which could change the stability or turnover of the proteins modified with Sia (14). Although only limited information is available, sialyltransferases that biosynthesize sialylated glycans in the Golgi apparatus do not show strong preferences for CMP-Neu5Ac or CMP-Neu5Gc as substrates (59). When we probed linkage-specific protein sialylation by using α2,6-linked Sia-binding plant lectins such as *Sambucus nigra* agglutinin, we did not observe a change (data not shown). Thus, the molecular event affected in *Cmah*-null mice is likely to be lectin recognition of a single oxygen atom on sialoglycans expressed on the cell surface, although a single responsible lectin may not explain the phenotype. One of the candidate lectins as the receptor of sialoglycans is the Siglec family (9, 12, 62), though a yet-to-be-characterized Sia-binding molecule could be affected.

When ligand expression for Siglecs was detected using Siglec-Fc probes, *Cmah*-null mice lost optimal ligand expression for CD22 (Siglec-2). The ligand function of CD22 in a mouse model has been addressed in two different ways. One study was done using *St6gal1*-knockout mice (20), and another study analyzed gene-targeted mice expressing mutant CD22 molecules that do not interact with ligands (43). The phenotypes found in *Cmah*-null mice are considerably different from these two previous studies; therefore, *Cmah*-null phenotypes might be caused by the combination of loss/gain of a Sia-mediated interaction. Additional studies using a combination of various knockout strains related to sialoglycan recognition are required to address such possibilities.

Apart from the phenotypic contribution of CD22 to the assays in the present study, CD22 ligand expression is not static but is, instead, a regulated event during in vivo B-cell activation. We showed that mCD22-Fc probe staining was downregulated in germinal centers. Moreover, it was reported that in vitro activated human B cells unmask CD22 from a *cis*-ligand (45). Thus, the regulation of CD22 ligand expression

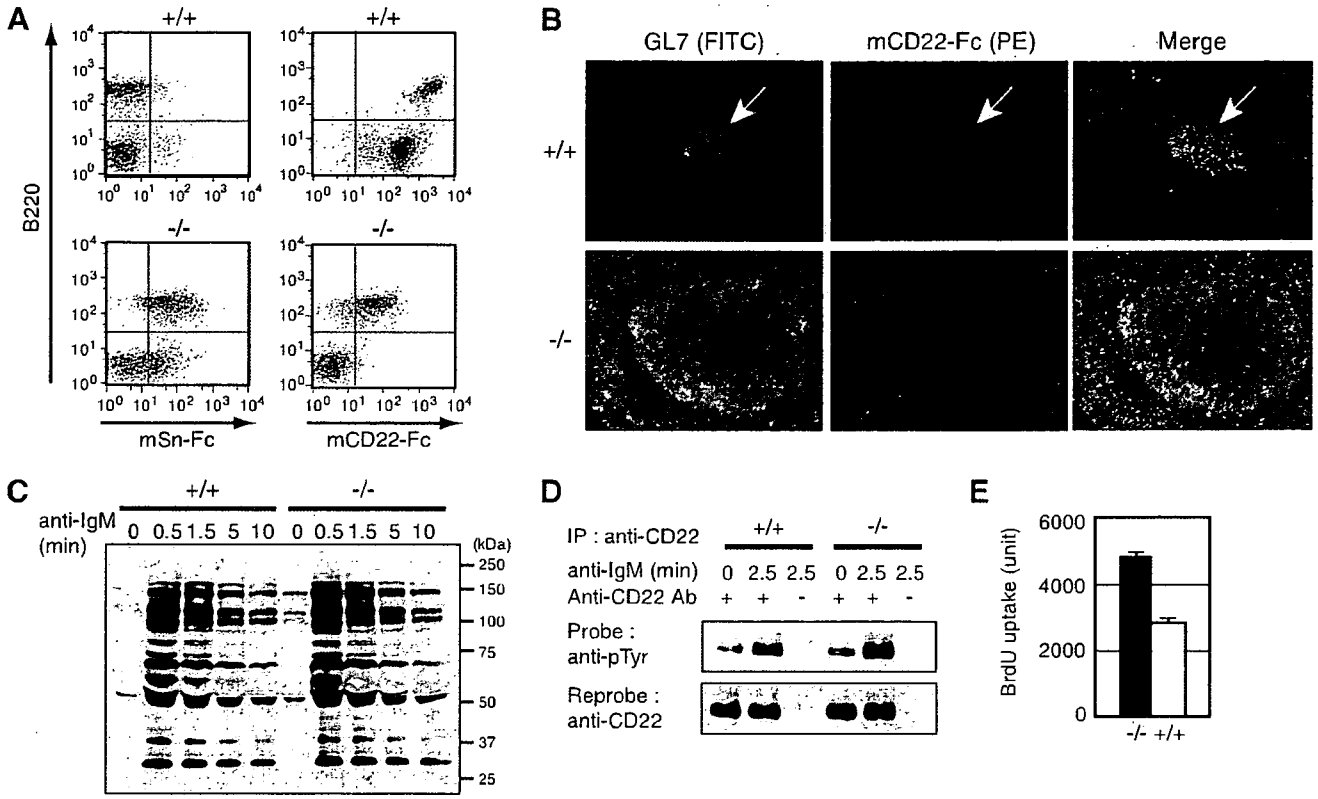

FIG. 9. Loss of optimal CD22 ligand and normal immediate response upon BCR cross-linking in *Cmah*-null mice. (A) Loss of optimal ligand for CD22 in *Cmah*-null mice. The expression of surface ligands for sialoadhesin and CD22 was detected by flow cytometry. Splenocytes from wild-type and *Cmah*-null mice were costained with FITC-conjugated anti-B220 and mSn/mCD22-Fc precomplexed with R-PE-conjugated anti-human IgG. Wild-type B cells were strongly stained with mCD22-Fc. In contrast, the level of mCD22-Fc staining showed a marked decrease in *Cmah*-null mice. The weak signal found on *Cmah*-null splenocytes was detected only with the chimeric probe mCD22-Fc prepared from Lec2 cell culture medium and not with the probe prepared from COS7 cells, possibly because of the autosialylation. (B) Histochemical analyses of CD22 ligand expression in spleen sections. Spleen sections from wild-type and *Cmah*-null mice 8 days after SRBC immunization were costained with FITC-conjugated GL7 and mCD22-Fc precomplexed with R-PE-conjugated anti-human IgG. Arrows indicate germinal centers. (C) Overall tyrosine phosphorylation upon anti-IgM stimulation. Splenic B cells from wild-type and *Cmah*-null mice were stimulated with the $F(ab')_2$ fragment of anti-mouse IgM (μ chain) for the indicated times. Whole-cell lysates were subjected to immunoblotting with antiphosphotyrosine antibody (PT-66). (D) Phosphorylation of CD22. Splenic B cells were stimulated with the $F(ab')_2$ fragment of anti-mouse IgM (μ chain) for the indicated times. The cell lysates were subjected to immunoprecipitation with anti-CD22 antibody (Cy34.1). The precipitated proteins were immunoblotted with antiphosphotyrosine (pTyr) antibody (PT-66) and then reprobed with anti-CD22 polyclonal antibody. (E) In vitro hyperproliferation response of *Cmah*-null B cells to calcium signaling. Splenic B cells were assessed for proliferation using tetradecanoyl phorbol acetate (10 ng/ml) plus ionomycin (5 μg/ml) as stimulating reagents. The proliferation assay was performed as described in the legend of Fig. 6C. The open column represents the mean proliferation of wild-type B cells, and the filled column represents the mean proliferation of *Cmah*-null B cells. The bars represent the standard errors of the mean for triplicate cultures. +/+, wild type; −/−, *Cmah* null; IP, immunoprecipitation.

could be an important event to modulate B-cell activation in vivo.

Loss of Neu5Gc in relation to human deficiency for the *CMAH* gene. *Homo sapiens* is the sole mammalian species that lacks Neu5Gc expression throughout the body; indeed, Neu5Gc is antigenic to humans (31). This is a striking difference between humans and chimpanzees, which express Neu5Gc as the major species of Sia throughout their bodies. Recently, it was shown that, unlike gene expression in the extant great apes, the *CMAH* gene is inactivated in humans (6, 22). Here, we demonstrated that Cmah is the sole enzyme responsible for the production of Neu5Gc in cells since our mouse model reproduced the human-like deficiency in Neu5Gc biosynthesis. This result confirmed that a genetic mutation in the human lineage caused the lack of Neu5Gc in humans.

Sia is commonly used in the host recognition system of microbes, and human-specific microbes are reported to recognize epitope(s) containing Neu5Ac on human cells. The mouse described here is thus the first mammalian line that could be used as an animal model system to assess Sia-targeted human infectious diseases (15).

ACKNOWLEDGMENTS

We thank Motomi Osato and Yoshiaki Itoh for the blood chemistry and blood counting experiments. We also thank Ajit Varki, Takeshi Tsubata, and Reiji Kannagi for helpful discussions during the preparation of the manuscript.

This work was supported by CREST, Japanese Science and Technology; a grant-in-aid program from the Ministry of Education, Culture, Sports, Science, and Technology of Japan; and RIKEN.

REFERENCES

1. **Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt.** 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403**:503–511.
2. **Baum, L. G., K. Derbin, N. L. Perillo, T. Wu, M. Pang, and C. Uittenbogaart.** 1996. Characterization of terminal sialic acid linkages on human thymocytes. Correlation between lectin-binding phenotype and sialyltransferase expression. J. Biol. Chem. **271**:10793–10799.
3. **Blixt, O., B. E. Collins, I. M. Van Den Nieuwenhof, P. R. Crocker, and J. C. Paulson.** 2003. Sialoside specificity of the Siglec family assessed using novel multivalent probes: identification of potent inhibitors of myelin associated glycoproteins. J. Biol. Chem. **278**:31007–31019.
4. **Buhl, A. M., and J. C. Cambier.** 1997. Co-receptor and accessory regulation of B-cell antigen receptor signal transduction. Immunol. Rev. **160**:127–138.
5. **Cervenak, L., A. Magyar, R. Boja, and G. Laszlo.** 2001. Differential expression of GL7 activation antigen on bone marrow B cell subpopulations and peripheral B cells. Immunol. Lett. **78**:89–96.
6. **Chou, H. H., H. Takematsu, S. Diaz, J. Iber, E. Nickerson, K. L. Wright, E. A. Muchmore, D. L. Nelson, S. T. Warren, and A. Varki.** 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc. Natl. Acad. Sci. USA **95**:11751–11756.
7. **Clark, E. A.** 1993. CD22, a B cell-specific receptor, mediates adhesion and signal transduction. J. Immunol. **150**:4715–4718.
8. **Coico, R. F., B. S. Bhogal, and G. J. Thorbecke.** 1983. Relationship of germinal centers in lymphoid tissue to immunologic memory. VI. Transfer of B cell memory with lymph node cells fractionated according to their receptors for peanut agglutinin. J. Immunol. **131**:2254–2257.
9. **Crocker, P. R.** 2002. Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling. Curr. Opin. Struct. Biol. **12**:609–615.
10. **Crocker, P. R., S. Kelm, C. Dubois, B. Martin, A. S. McWilliam, D. M. Shotton, J. C. Paulson, and S. Gordon.** 1991. Purification and properties of sialoadhesin, a sialic acid-binding receptor of murine tissue macrophages. EMBO J. **10**:1661–1669.
11. **Crocker, P. R., and A. Varki.** 2001. Siglecs in the immune system. Immunology **103**:137–145.
12. **Crocker, P. R., and A. Varki.** 2001. Siglecs, sialic acids and innate immunity. Trends Immunol. **22**:337–342.
13. **Cyster, J. G., and C. C. Goodnow.** 1997. Tuning antigen receptor signaling by CD22: Integrating cues from antigens and the microenvironment. Immunity **6**:509–517.
14. **Ellies, L. G., D. Ditto, G. G. Levy, M. Wahrenbrock, D. Ginsburg, A. Varki, D. T. Le, and J. D. Marth.** 2002. Sialyltransferase ST3Gal-IV operates as a dominant modifier of hemostasis by concealing asialoglycoprotein receptor ligands. Proc. Natl. Acad. Sci. USA **99**:10042–10047.
15. **Gagneux, P., and A. Varki.** 1999. Evolutionary considerations in relating oligosaccharide diversity to biological function. Glycobiology **9**:747–755.
16. **Han, H., D. J. Bearss, L. W. Browne, R. Calaluce, R. B. Nagle, and D. D. Von Hoff.** 2002. Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray. Cancer Res. **62**:2890–2896.
17. **Han, S., S. R. Dillon, B. Zheng, M. Shimoda, M. S. Schlissel, and G. Kelsoe.** 1997. V(D)J recombinase activity in a subset of germinal center B lymphocytes. Science **278**:301–305.
18. **Han, S., B. Zheng, D. G. Schatz, E. Spanopoulou, and G. Kelsoe.** 1996. Neoteny in lymphocytes: Rag1 and Rag2 expression in germinal center B cells. Science **274**:2094–2097.
19. **Hathcock, K. S., C. E. Pucillo, G. Laszlo, L. Lai, and R. J. Hodes.** 1995. Analysis of thymic subpopulations expressing the activation antigen GL7: expression, genetics, and function. J. Immunol. **155**:4575–4581.
20. **Hennet, T., D. Chui, J. C. Paulson, and J. D. Marth.** 1998. Immune regulation by the ST6Gal sialyltransferase. Proc. Natl. Acad. Sci. USA **95**:4504–4509.
21. **Hokazono, Y., T. Adachi, M. Wabl, N. Tada, T. Amagasa, and T. Tsubata.** 2003. Inhibitory coreceptors activated by antigens but not by anti-Ig heavy chain antibodies install requirement of costimulation through CD40 for survival and proliferation of B cells. J. Immunol. **171**:1835–1843.
22. **Irie, A., S. Koyama, Y. Kozutsumi, T. Kawasaki, and A. Suzuki.** 1998. The molecular basis for the absence of *N*-glycolylneuraminic acid in humans. J. Biol. Chem. **273**:15866–15871.
23. **Itohara, S., P. Mombaerts, J. Lafaille, J. Iacomini, A. Nelson, A. R. Clarke, M. L. Hooper, A. Farr, and S. Tonegawa.** 1993. T cell receptor delta gene mutant mice: independent generation of alpha beta T cells and programmed rearrangements of gamma delta TCR genes. Cell **72**:337–348.
24. **Kawano, T., S. Koyama, H. Takematsu, Y. Kozutsumi, H. Kawasaki, S. Kawashima, T. Kawasaki, and A. Suzuki.** 1995. Molecular cloning of cytidine monophospho-*N*-acetylneuraminic acid hydroxylase. Regulation of species- and tissue-specific expression of *N*-glycolylneuraminic acid. J. Biol. Chem. **270**:16458–16463.
25. **Kawano, T., Y. Kozutsumi, T. Kawasaki, and A. Suzuki.** 1994. Biosynthesis of *N*-glycolylneuraminic acid-containing glycoconjugates. Purification and characterization of the key enzyme of the cytidine monophospho-*N*-acetylneuraminic acid hydroxylation system. J. Biol. Chem. **269**:9024–9029.
26. **Kelm, S., A. Pelz, R. Schauer, M. T. Filbin, S. Tang, M. E. de Bellard, R. L. Schnaar, J. A. Mahoney, A. Hartnell, P. Bradfield, et al.** 1994. Sialoadhesin, myelin-associated glycoprotein and CD22 define a new family of sialic acid-dependent adhesion molecules of the immunoglobulin superfamily. Curr. Biol **4**:965–972.
27. **Koyama, S., T. Yamaji, H. Takematsu, T. Kawano, Y. Kozutsumi, A. Suzuki, and T. Kawasaki.** 1996. A naturally occurring 46-amino-acid deletion of cytidine monophospho-*N*-acetylneuraminic acid hydroxylase leads to a change in the intracellular distribution of the protein. Glycoconj. J. **13**:353–358.
28. **Laszlo, G., K. S. Hathcock, H. B. Dickler, and R. J. Hodes.** 1993. Characterization of a novel cell-surface molecule expressed on subpopulations of activated T and B cells. J. Immunol. **150**:5252–5262.
29. **Lossos, I. S., A. A. Alizadeh, M. B. Eisen, W. C. Chan, P. O. Brown, D. Botstein, L. M. Staudt, and R. Levy.** 2000. Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. Proc. Natl. Acad. Sci. USA **97**:10209–10213.
30. **MacLennan, I. C.** 1994. Germinal centers. Annu. Rev. Immunol. **12**:117–139.
31. **Martin, M. J., A. Muotri, F. Gage, and A. Varki.** 2005. Human embryonic stem cells express an immunogenic nonhuman sialic acid. Nat. Med. **11**:228–232.
32. **McHeyzer-Williams, L. J., M. Cool, and M. G. McHeyzer-Williams.** 2000. Antigen-specific B cell memory: expression and replenishment of a novel B220$^-$ memory B cell compartment. J. Exp. Med. **191**:1149–1166.
33. **McHeyzer-Williams, L. J., D. J. Driver, and M. G. McHeyzer-Williams.** 2001. Germinal center reaction. Curr. Opin. Hematol. **8**:52–59.
34. **Mills, D. M., J. C. Stolpa, and J. C. Cambier.** 2004. Cognate B cell signaling via MHC class II: differential regulation of B cell antigen receptor and MHC class II/Ig-$\alpha\beta$ signaling by CD22. J. Immunol. **172**:195–201.
35. **Morita, S., T. Kojima, and T. Kitamura.** 2000. Plat-E: an efficient and stable system for transient packaging of retroviruses. Gene Ther. **7**:1063–1066.
36. **Muchmore, E. A., M. Milewski, A. Varki, and S. Diaz.** 1989. Biosynthesis of *N*-glycolylneuraminic acid. The primary site of hydroxylation of *N*-acetylneuraminic acid is the cytosolic sugar nucleotide pool. J. Biol. Chem. **264**:20216–20223.
37. **Muramatsu, M., V. S. Sankaranand, S. Anant, M. Sugai, K. Kinoshita, N. O. Davidson, and T. Honjo.** 1999. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. J. Biol. Chem. **274**:18470–18476.
38. **Murasawa, M., S. Okada, S. Obata, M. Hatano, H. Moriya, and T. Tokuhisa.** 2002. GL7 defines the cycling stage of pre-B cells in murine bone marrow. Eur. J. Immunol. **32**:291–298.
39. **Nitschke, L.** 2005. The role of CD22 and other inhibitory co-receptors in B-cell activation. Curr. Opin. Immunol. **17**:290–297.
40. **Novoradovskaya, N., M. L. Whitfield, L. S. Basehore, A. Novoradovsky, R. Pesich, J. Usary, M. Karaca, W. K. Wong, O. Aprelikova, M. Fero, C. M. Perou, D. Botstein, and J. Braman.** 2004. Universal reference RNA as a standard for microarray experiments. BMC Genomics **5**:20.
41. **Pasare, C., and R. Medzhitov.** 2005. Control of B-cell responses by Toll-like receptors. Nature **438**:364–368.
42. **Paulson, J. C., J. Weinstein, and A. Schauer.** 1989. Tissue-specific expression of sialyltransferases. J. Biol. Chem. **264**:10931–10934.
43. **Poe, J. C., Y. Fujimoto, M. Hasegawa, K. M. Haas, A. S. Miller, I. G. Sanford, C. B. Bock, M. Fujimoto, and T. F. Tedder.** 2004. CD22 regulates B lymphocyte function in vivo through both ligand-dependent and ligand-independent mechanisms. Nat. Immunol. **5**:1078–1087.
44. **Powell, L. D., and A. Varki.** 1994. The oligosaccharide binding specificities of CD22 beta, a sialic acid-specific lectin of B cells. J. Biol. Chem. **269**:10628–10636.
45. **Razi, N., and A. Varki.** 1998. Masking and unmasking of the sialic acid-binding lectin activity of CD22 (Siglec-2) on B lymphocytes. Proc. Natl. Acad. Sci. USA **95**:7469–7474.
46. **Reichert, R. A., W. M. Gallatin, I. L. Weissman, and E. C. Butcher.** 1983. Germinal center B cells lack homing receptors necessary for normal lymphocyte recirculation. J. Exp. Med. **157**:813–827.
47. **Schauer, R.** 1982. Sialic acids: chemistry, metabolism and function. Cell biology monographs, vol. 10. Springer-Verlag, New York, NY.
48. **Schulte, R. J., M. A. Campbell, W. H. Fischer, and B. M. Sefton.** 1992. Tyrosine phosphorylation of CD22 during B cell activation. Science **258**:1001–1004.
49. **Schwarzkopf, M., K. P. Knobeloch, E. Rohde, S. Hinderlich, N. Wiechens, L. Lucka, I. Horak, W. Reutter, and R. Horstkorte.** 2002. Sialylation is essential for early development in mice. Proc. Natl. Acad. Sci. USA **99**:5267–5270.
50. **Sgroi, D., A. Varki, S. Braesch-Andersen, and I. Stamenkovic.** 1993. CD22,

a B cell-specific immunoglobulin superfamily member, is a sialic acid-binding lectin. J. Biol. Chem. 268:7011–7018.
51. **Shaffer, A. L., A. Rosenwald, E. M. Hurt, J. M. Giltnane, L. T. Lam, O. K. Pickeral, and L. M. Staudt.** 2001. Signatures of the immune response. Immunity 15:375–385.
52. **Shapiro-Shelef, M., K. I. Lin, L. J. McHeyzer-Williams, J. Liao, M. G. McHeyzer-Williams, and K. Calame.** 2003. Blimp-1 is required for the formation of immunoglobulin secreting plasma cells and pre-plasma memory B cells. Immunity 19:607–620.
53. **Shaw, L., and R. Schauer.** 1988. The biosynthesis of *N*-glycolylneuraminic acid occurs by hydroxylation of the CMP-glycoside of *N*-acetylneuraminic acid. Biol. Chem. Hoppe-Seyler 369:477–486.
54. **Shaw, L., and R. Schauer.** 1989. Detection of CMP-*N*-acetylneuraminic acid hydroxylase activity in fractionated mouse liver. Biochem. J. 263: 355–363.
55. **Shih, T. A., E. Meffre, M. Roederer, and M. C. Nussenzweig.** 2002. Role of BCR affinity in T cell dependent antibody responses in vivo. Nat. Immunol. 3:570–575.
56. **Sjoberg, E. R., L. D. Powell, A. Klein, and A. Varki.** 1994. Natural ligands of the B cell adhesion molecule CD22 beta can be masked by 9-O-acetylation of sialic acids. J. Cell Biol. 126:549–562.
57. **Takematsu, H., S. Diaz, A. Stoddart, Y. Zhang, and A. Varki.** 1999. Lysosomal and cytosolic sialic acid 9-*O*-acetylesterase activities can be encoded by one gene via differential usage of a signal peptide-encoding exon at the N terminus. J. Biol. Chem. 274:25623–25631.
58. **Tedder, T. F., J. Tuscano, S. Sato, and J. H. Kehrl.** 1997. CD22, a B lymphocyte-specific adhesion molecule that regulates antigen receptor signaling. Annu. Rev. Immunol. 15:481–504.
59. **Tsuji, S.** 1996. Molecular cloning and functional analysis of sialyltransferases. J. Biochem. (Tokyo) 120:1–13.
60. **Varki, A.** 1992. Diversity in the sialic acids Glycobiology 2:25–40. (Erratum, 2:168.)
61. **Varki, A.** 1997. Sialic acids as ligands in recognition phenomena. FASEB J. 11:248–255.
62. **Varki, A., and T. Angata.** 2006 Siglecs—the major subfamily of I-type lectins. Glycobiology 16:1R–27R.
63. **Wakabayashi, C., T. Adachi, J. Wienands, and T. Tsubata.** 2002. A distinct signaling pathway used by the IgG-containing B cell antigen receptor. Science 298:2392–2395.
64. **Wuensch, S. A., R. Y. Huang, J. Ewing, X. Liang, and J. T. Lau.** 2000. Murine B cell differentiation is accompanied by programmed expression of multiple novel β-galactoside α2, 6-sialyltransferase mRNA forms. Glycobiology 10: 67–75.
65. **Yamaji, T., T. Teranishi, M. S. Alphey, P. R. Crocker, and Y. Hasbimoto.** 2002. A small region of the natural killer cell receptor, Siglec-7, is responsible for its preferred binding to α2,8-disialyl and branched α2,6-sialyl residues. A comparison with Siglec-9. J. Biol. Chem. 277:6324–6332.
66. **Zhang, J., I. C. MacLennan, Y. J. Liu, and P. J. Lane.** 1988. Is rapid proliferation in B centroblasts linked to somatic mutation in memory B cell clones? Immunol. Lett. 18:297–299.

Application of a New Probabilistic Model for Mining Implicit Associated Cancer Genes from OMIM and Medline

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

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University
[2]Graduate School of Pharmaceutical Sciences, Kyoto University

Abstract: An important issue in current medical science research is to find the genes that are strongly related to an inherited disease. A particular focus is placed on cancer-gene relations, since some types of cancers are inherited. As bio-medical databases have grown speedily in recent years, an informatics approach to predict such relations from currently available databases should be developed. Our objective is to find implicit associated cancer-genes from biomedical databases including the literature database. Co-occurrence of biological entities has been shown to be a popular and efficient technique in biomedical text mining. We have applied a new probabilistic model, called mixture aspect model (MAM) [48], to combine different types of co-occurrences of genes and cancer derived from Medline and OMIM (Online Mendelian Inheritance in Man). We trained the probability parameters of MAM using a learning method based on an EM (Expectation and Maximization) algorithm. We examined the performance of MAM by predicting associated cancer gene pairs. Through cross-validation, prediction accuracy was shown to be improved by adding gene-gene co-occurrences from Medline to cancer-gene co-occurrences in OMIM. Further experiments showed that MAM found new cancer-gene relations which are unknown in the literature. Supplementary information can be found at http://www.bic.kyotou.ac.jp/pathway/zhusf/CancerInformatics/Supplemental2006.html

Key Words: Cancer genetics, Cancer gene discovery, Machine learning, Text mining, Probabilistic model.

Introduction

Cancer is attributed to complex interactions of multiple factors, such as inheritance, gene mutation and environment. It is characterized by genetic alteration such as DNA amplification, deletion, translocation and point mutation, as well as distortion in gene expression [25]. Most known cancer-causing genes, oncogenes and tumor suppressor genes, have the crucial function of regulating cell proliferation, differentiation and death for cancer genesis and progression. New cancer therapy could target the proteins encoded by these genes to kill cancer cells or inhibit the propagation of them. Some other genes are highly expressed in cancer cells than normal cells, which could be utilized for early detection of oncogenesis [16]. Thus, the discovery of the cancer associated genes is extremely helpful for the understanding of tumor pathogenesis, and potential diagnosis and treatment of the cancer.

Linkage studies were first successfully used to find some cancer-susceptibility genes with high penetrance, such as *BRCA1* and *BRCA2* in breast cancer [6]. It examines the genotypes and phenotypes of parents and offspring in cancer families to locate the susceptibility genes, which will be further assessed and screened for validation. However, it lacks the power to detect multiple susceptibility alleles with moderate risks. Genetic association studies [7] alleviate this problem by comparing the genotype distribution between diseased individuals and non-diseased individuals for finding allelic variants that predispose to cancer. Because of the existence of linkage disequilibrium, genotype variants within a region can be captured by a subset of single-nucleotide polymorphisms (SNPs) [40]. Then the association candidate gene or genomic region with cancer could be examined by a tagging-SNP approach. With the increasing accumulation of SNPs data in genomic databases, such as the HapMap project [41], selecting a set of tagging SNPs that covers all common genetic variants in whole genome becomes possible [37].

To increase the success rate, the candidate genes could be selected for carrying out association studies. For example, with the complete sequencing of whole human genome, given a known cancer associated gene, we

*Both authors equally contributed to this work.

Correspondence: Shanfeng Zhu, Kyoto University, Gokasho, Uji, 611-0011, Japan.
Email: zhusf@kuicr.kyoto-u.ac.jp, Phone: +81-774-383038, Fax: +81-774-383037.

can find some possible homologous susceptibility-genes that have similar sequences by using sequence alignment programs, such as BLAST [1] and FASTA [35], or similar structures in the encoded protein. Furthermore, due to the rapid development of bioinformatics, more and more high throughput genomic data such as genomics, transcriptomics, proteomics and metabonomics data, as well as novel algorithms for effectively and efficiently integrating and analyzing these data, could be utilized to improve the selection of candidate genes. The genetic alteration in cancer cells could be identified by molecular cytogenetic techniques and comparative genomic hybridization (CGH) approaches [23, 11]. Subsequent gene expression pattern changes could be captured (or dissected) by analyzing the microarray gene expression profile, and digital expression pattern data such as expression sequence tags (ESTs) [4] and serial analysis of gene expression (SAGE) [42]. Proteomic and metabolic data can also provide valuable biological insights on cancer gene discovery.

By contrast, in this work, we attempt to mine implicit associated cancer genes that do not appear in the literature by applying a new probabilistic model, mixture aspect model (MAM) [48] on cancer gene co-occurrence data in OMIM and Medline. Online Mendelian Inheritance in Man (OMIM), a comprehensive human curated knowledgebase of human genes and genetic disorders, was first created by Victor McKusick at Johns Hopkins University, and now updated by him and other scientists [29, 17]. Until December 2005, it consists of more than 16,000 records, which can be divided into several categories based on genes, phenotypes or both. There are around 2,200 entries including both disease phenotype description and associated genes. Bajdik et al [2] wrote a software tool CGMIM to extract these entries to identify genetically-associated cancers and candidate genes by mapping those diseases into 21 type of cancers. Using this software, we can obtain two types of co-occurrence datasets: cancer gene and cancer-cancer co-occurrence datasets. MAM was proposed by us to mine implicit "chemical compound-gene" relations by integrating three types of co-occurrence datasets in the literature, ie gene-gene, compound-compound, and compound-gene. MAM was extended from a classical probabilistic model, aspect model (AM), which has been successfully applied in natural language processing, information retrieval, and collaborative filtering in E-commerce [19, 20]. The advantage of MAM, comparing with AM, is that MAM can handle different type of co-occurrence data, keeping the same time and space efficiency as those of AM. Thus, we can say AM is a special case, handling only one co-occurrence dataset, of MAM. We emphasize that this extension of AM to MAM is significant in the situation where we can use a lot of different types of co-occurrence datasets.

In addition to applying MAM on existing cancer-gene and cancer-cancer co-occurrence datasets from OMIM, we further incorporated gene-gene co-occurrences from a different data source, Medline [45], which can capture biological relationships among co-occurred genes. We first examined the performance of our model by cross-validation and found that combining all three types of co-occurrence datasets achieves the best result. This result indicates that MAM would be especially effective to predict an unknown gene, which is implicitly associated with some cancer, with a high accuracy. We then trained our model using all obtained co-occurrence datasets and predicted the likelihoods of unknown cancer-gene pairs, which are expected to be strongly related. We finally focused on unknown genes which are specific to each type of cancer and ranked them for each cancer, according to the likelihoods predicted by our trained model. The top 20 of these genes for each cancer are given as an online supplement material for cancer biologists' reference, and we analyzed some of these genes from biological, medical and genetic viewpoints.

Related Work

Genetic alteration of chromosomal aberrations and rearrangement, especially structural chromosome aberrations, could be discerned by using cytogenetic and molecular genetics techniques, such as G banding, fluorescence in situ hybridization (FISH) and spectral karyotyping (SKY) [38]. In contrast to above techniques, Comparative Genomic Hybridization (CGH) [23, 11] can scan entire genome in a single step to identify segmental DNA copy number changes by taking advantage of the complete sequencing of human genome project. Although FISH, SKY and CGH techniques have already been widely used and made significant impacts on cancer research, they could only achieve limited resolution of 5-20Mb in genomic DNA alteration identification. By incorporating latest microarray techniques, array-based CGH such as bacterial artificial chromosome (BAC) array CGH, cDNA array CGH and oligonucleotide array CGH, can achieve much higher resolution for discerning genomic DNA alteration [32, 33, 28]. Another high resolution technique digital karyotyping is based on

enumerating the sequence tags to quantitatively measure DNA copy number change [44].

After the identification of amplified or deleted chromosomal regions, bioinformatics approaches can facilitate the discovery of cancer associated genes by analyzing the high-throughput biological data. Many studies have been carried out to analyze microarray gene expression data to find cancer related genes, which assumes that the expression level of one gene could be reflected by the abundance of corresponding mRNA. The most popular technique is to find differential expressed genes with high fold change between normal and tumor cells. For example, novel gastric cancer-related genes, specifically, such as potential marker CDC20 and MT2A, were discovered using a cDNA microarray [24]. Unlike microarray technology, digital expression profiling using expressed sequence tags (ESTs) or serial analysis of gene expression (SAGE) can be also used to identify cancer associated genes [4, 42]. In digital expression profiling, we assume that the expression level of one gene is proportional to the relative frequency of corresponding sequence tag in cDNA library data. Recently, Shen and his colleagues identified breast cancer related genes by analyzing differential gene expression between healthy and tumor breast tissue in EST and SAGE high throughput data [39]. After combining multiple analyses, they found six interesting genes related to breast cancer, with four down-regulated genes, ANXA1, CAV1, KRT5 and NMP7 and two up-regulated genes, ERBB2 and G1P3.

Although many studies analyzed high-throughput biological data to identify cancer associated genes, there are very few work that made use of literature mining. Mining biomedical text is attracting a great deal of interest because it can acquire accumulated biological and medical information and facilitate further knowledge discovery [47]. Some researchers already discovered disease gene candidates by text mining. For example, Freudenberg et al clustered diseases according to their phenotypic similarity and characterized genes with related GO function terms [13]. Potential disease genes from the human genome are then scored by their functional similarity to known disease genes in the same cluster of query disease. Perez-Iratxeta et al [30] used the fuzzy set theory to analyze the relationships between co-occurred MeSH terms in different categories , as well as the co-occurrence of a MeSH term and a GO (Gene Ontology) term in Medline records. Furthermore, they scored the implicit associations between symptoms of diseases and GO terms by fuzzy relations. In this work, we focus on mining the relationship of genetically-associated cancers and candidate genes, which can be obtained from the OMIM text database.

Most of text mining studies made use of co-occurrence techniques to discover possible biological relationships among different entities. This technique is based on the following hypothesis: if biological entity A co-occurs with biological entity B in the same biomedical document (eg a Medline record), A and B should be biologically related with high probability. This hypothesis was experimentally testified by many researchers [22, 8]. Here we also employ this method to obtain cancer-gene and cancer-cancer pairs by using a public available software CGMIM, which mines the description section of OMIM record. Since OMIM is a human curated database, the accuracy of our dataset is high. Furthermore, we incorporate gene-gene co-occurrence pairs from Medline. Although these gene-gene pairs are derived from a different source other than OMIM, we assume that co-occurred gene pairs in Medline should have much higher probability of associating with the same cancer than randomly generated gene pairs, which may help improve the prediction of cancer associated genes. This assumption is verified in our experiment (See the Data section for details).

Method

Notations

We use the same set of notations throughout the paper. A variable is denoted by a capitalized letter, and its value by corresponding lowercase letter. To explore the co-occurrence of a cancer and a gene in literature, let G be an observable random variable taking values $g_1, g_2, \ldots, g_S$, each of which stands for a specific gene, and let C be an observable random variable taking value $c_1, c_2, \ldots, c_T$, each of which stands for a specific type of cancer. Similarly, 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 (ie weight) of a component of our model that the users can specify. Let D be a set of all examples.

Mixture Aspect Model for Predicting Cancer-Gene Co-occurrences

Aspect model (AM) was proposed by Hofmann for tackling problems in natural language processing

[19, 20]. With latent clusters z_h ($h = 1, ..., H$), AM gives the log-likelihood for a co-occurrence of (c_i, g_j) in the following form:

$$\log p(c_i, g_j) = \log \sum_h p(c_i \mid z_h) p(g_j \mid z_h) p(z_h).$$

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

$$\log p(D) = \sum_{i,j} N_{i,j} \log p(c_i, g_j),$$

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

The objective of this work is to integrate different types of co-occurrence datasets, to identify cancer-associated genes with high accuracy. We used Mixture of Aspect Model (MAM), which was extended from AM by us in our previous work, to efficiently integrate different types of co-occurrence datasets. MAM has a general framework, and in this paper, we explain MAM briefly. Interested readers should refer to our previous paper [48], where the details of MAM are described. We denote the model built from k types of co-occurrence datasets as kMAM. For example, two types of co-occurrence datasets can be integrated by 2MAM. In this work, we have three types of co-occurrence datasets: cancer-gene from OMIM, cancer-cancer from OMIM, and gene-gene from Medline. Thus, we finally used 3MAM.

Here we focus on 3MAM which integrates all the three types of co-occurrence datasets. The models for other kinds of combinations among co-occurrence datasets could be derived similarly.

The log-likelihood for all data D can be given by 3MAM as follows:

$$\log p(D) = \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} \log \sum_h p(c_i \mid z_h) p(g_j \mid z_h) p(z_h)$$

$$+ \pi_{GG} \sum_{j,j'} \frac{M_{j,j'}}{N_{GG}} \log \sum_h p(g_j \mid z_h) p(g_{j'} \mid z_h) p(z_h)$$

$$+ \pi_{CC} \sum_{i,i'} \frac{L_{i,i'}}{N_{CC}} \log \sum_h p(c_i \mid z_h) p(c_{i'} \mid z_h) p(z_h).$$

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

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 [10] 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.

Aspect Model

We begin to explain the EM algorithm for AM for only one type of co-occurrence dataset, ie cancer 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:

$$p(z_h \mid c_i, g_j) = \frac{p(c_i \mid z_h) p(g_j \mid z_h) p(z_h)}{\Sigma_{h'} p(c_i \mid z_{h'}) p(g_j \mid z_{h'}) p(z_{h'})}.$$

M-step:

$$\hat{p}(c_i \mid z_h) \propto \sum_j N_{i,j} \cdot p(z_h \mid c_i, g_j)$$

$$\hat{p}(g_j \mid z_h) \propto \sum_{i,j} N_{i,j} \cdot p(z_h \mid c_i, g_j)$$

$$\hat{p}(z_h) \propto \sum_{i,j} N_{i,j} \cdot p(z_h \mid c_i, g_j)$$

Mixture Aspect Model

Now we show the EM algorithm for 3MAM which can use all the three types of co-occurrence datasets: cancer-gene, gene-gene and cancer-cancer pairs. To maximize the log-likelihood described in Section 3.2, the E- and M-steps for 3MAM can be given as follows:

E-step:

$$p(z_h \mid c_i, g_j) = \frac{p(c_i \mid z_h) p(g_j \mid z_h) p(z_h)}{\Sigma_{h'} p(c_i \mid z_{h'}) p(g_j \mid z_{h'}) p(z_{h'})}$$

$$p(z_h \mid g_j, g_{j'}) = \frac{p(g_j \mid z_h) p(g_{j'} \mid z_h) p(z_h)}{\Sigma_{h'} p(g_j \mid z_{h'}) p(g_{j'} \mid z_{h'}) p(z_{h'})}$$

$$p(z_h \mid c_i, c_{i'}) = \frac{p(c_i \mid z_h) p(c_{i'} \mid z_h) p(z_h)}{\Sigma_{h'} p(c_i \mid z_{h'}) p(c_{i'} \mid z_{h'}) p(z_{h'})}$$

M-step:

$$\hat{p}(g_j \mid z_h) \propto \pi_{cg} \sum_i \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{CC} \sum_{i'} \frac{L_{i,i'}}{N_{CC}} p(z_h \mid c_i, c_{i'})$$

$$\hat{p}(g_j \mid z_h) \propto \pi_{CG} \sum_i \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{GG} \sum_{j'} \frac{M_{j,j'}}{N_{GG}} p(z_h \mid g_j, g_{j'})$$

$$\hat{p}(z_c) \propto \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{GG} \sum_{j',j''} \frac{M_{j',j''}}{N_{GG}} p(z_h \mid g_{j'}, g_{j''}) + \pi_{CC} \sum_{i',i''} \frac{L_{i',i''}}{N_{CC}} p(z_h \mid c_{i'}, c_{i''})$$

Parameter Settings in Our Experiments

We set the number of latent clusters, H, at 128 and used a uniform distribution for the weights (ie π) of both 2MAM and 3MAM in all cases. We iterated the EM algorithm until the improvement of the observed log-likelihoods between two successive iterations is less than 0.001.

Data

Cancer-Gene and Cancer-Cancer Co-occurrences

OMIM (Online Mendelian in Man) is a human-curated database, containing the comprehensive and authoritative information on human genes and genetic disorders. Our focus is placed on genes which are related with cancers, and we used a software tool CGMIM, which extracts the description section of OMIM records to obtain cancers and associated genes. The CGMIM builds a synonym list from International Classification of Disease for Oncology (ICD-O) [14]. The list maps genetic disorders into 21 different types of cancers, which are defined by the National Cancer Institute of Canada. They are bladder, brain, breast, cervix, colorectal, esophagus, kidney, larynx, leukemia, lung, lymphoma, melanoma, myeloma, oral, ovary, pancreas, prostate, stomach, testis, thyroid and body-of-uterus. We obtained the two types of co-occurrence datasets from the OMIM database downloaded in Oct 2005. Our datasets are altogether 2,017 genes associated to cancers, 3,743 cancer-gene pairs and 206 cancer-cancer pairs.

Gene-Gene Co-occurrences

Since gene-gene co-occurrences are not available in OMIM, we obtained this kind of co-occurrences from the Medline database. We used Locuslink [34], ie a human curated database, to avoid errors that may occur in identifying gene names in Medline. The Locuslink has a list of links, each of which connects a Locus ID with a PubMed ID, meaning that we can see whether a gene (specified by a Locus ID) is in an abstract (specified by a PubMed ID) or not.

We used a file available at the following ftp site, and the file we used was generated at Dec 2004:

ftp://ftp.ncbi.nih.gov/refseq/LocusLink

From this list, we selected Medline records containing one or more human genes, focusing on "human" genes only. We then generated gene-gene co-occurrences from the selected Medline records. That is, if two genes are in a same Medline record, we can say that these two genes co-occur.

We found some Medline records have a large number of genes. For example, a record with PubMed ID 12477932 contains more than 9,000 human genes by showing all genes in a microarray experiment. Thus, we removed the record, each of which has more than 10 genes. We note that this is a normal procedure in dealing with Medline records. For example, Wilkinson et al also put this kind of restriction to filtering Medline records for finding communities of related genes [46].

Our focus is on cancer associated genes, and a gene-gene co-occurrence pair was removed unless both genes of the pair are in the 2,017 genes of our cancer-gene co-occurrence dataset. Finally we obtained 3,118 gene-gene pairs from Medline. Table 1 shows a summary of the data information.

Table 1: The size of co-occurrence datasets.

Item	Size
gene type	2,017
gene-gene	3,118
cancer type	21
cancer-cancer	206
cancer-gene	3,743

Preliminary Verification on Gene-Gene Co-occurrence Dataset

Focusing on genes in cancer-gene co-occurrence pairs from OMIM, we attempted to confirm that two genes in each gene-gene pair from Medline are associated to a same cancer with high probability. When both two genes in a gene-gene pair are associated with at least one same cancer, we call such a gene-gene pair a *positive pair*, and we computed the ratio of positive pairs to all gene-gene pairs, which we call the *positive ratio*.

We found that among total 3,118 gene-gene co-occurrence pairs, 1,804 (57.86%) are positive pairs. We then reduced the size of gene-gene pairs by the number of co-occurrences and checked the positive ratio. Table 2 summarizes the obtained results.

As shown in the table, with increasing the co-occurrence number of gene-gene pairs, the positive ratio increased. For example, when the number of co-occurrences is set at more than one, 490 (64.64%) out of 758 gene-gene pairs are positive pairs. Furthermore, as a baseline, we checked the positive ratio of randomly generated pairs. That is, we randomly generated 3,118 gene-gene pairs 1,000 times using our 2,017 cancer associated genes and checked the average positive ratio for them. The average positive ratio was only 26.65%, with minimum 24.05%, maximum 29.76% and standard deviation 0.0083, which is far less than those obtained by our gene-gene co-occurrence dataset. These results clearly indicate that the motivation of adding gene-gene co-occurrence data in Medline to the cancer-gene and cancer-cancer data from OMIM would be reasonable.

Experimental Results

Predictive Performance of Mixture Aspect Model

Evaluation Procedure

We evaluated the performance of MAM by cross-validation on predicting associated cancer-gene pairs. We examined four types of MAM (including AM). That is, we first built AM using only the cancer-gene co-occurrence dataset. We then tested two different 2MAM by adding cancer-cancer or gene-gene pairs to the cancer-gene pairs, which correspond to 2MAM (CG+CC) or 2MAM (CG+GG), respectively. Finally 3MAM was examined by using all these three types of co-occurrence datasets.

To examine the effect of the training data size on the performance of our models, we checked three different data-size ratios of training to test datasets, 3:1, 1:1 and 1:3, in our cross-validation experiment. For example, in the 1:1 case, we randomly divided the original cancer-gene dataset into two subsets of roughly equal size, and then alternately selected one subset as a test set and the other as a training set. We carried out 50 rounds of the cross-validation to reduce the possible biases caused by random partitioning. In each round, to compare the performance of different models, we kept the testing dataset unchanged while adding another type of co-occurrence dataset. In this way, we made predictions on the same test dataset. We note that AM cannot compute the likelihood for a cancer gene pair in the test dataset unless a gene of this pair appears in the training data. So we removed all the pairs which are not in the training data but in the test dataset. We then used all remaining pairs as positive test examples. Please note that this experimental setting is advantageous to AM and not to MAM. Negative examples, which were used for evaluation only, were randomly generated to be included in neither the training dataset nor the positive test dataset. The size of negative test dataset was set as the same as that of positive test dataset.

Evaluation Measures

1) Area Under the ROC Curve (AUC)

The performance of each probabilistic model is evaluated by the ability to discriminate positive examples from negative examples in test data of our cross-validation. We used AUC (Area Under the ROC curve) to evaluate the discriminative performance of a model. The AUC is computed from an ROC (Receiver Operator Characteristic) curve. The ROC curve is drawn by plotting "sensitivity" against "false positive rate", using the ranked cancer-gene pairs. The sensitivity

Table 2: The ratio of positive pairs in gene-gene co-occurrence dataset.

# co-occurrences	- (random)	> = 1	>1	>2	>3	>4	>5	>6
Dataset size	3,118	3,118	758	379	276	152	122	99
Positive ratio (%)	26.65	57.86	64.64	68.34	69.91	70.2	72.13	76.77

(or true positive rate) is the proportion of the number of correctly predicted positive examples to the total number of positive examples. The false positive rate is the proportion of the number of false positive examples to the total number of negative examples. More concretely, once we estimated the parameters of a probabilistic model from training data, we computed the likelihood of each cancer-gene pair in test data and ranked them according to their likelihoods. We then set a cut-off value to separate positive examples from negatives and computed the sensitivity and the false positive rate by changing the cut-off value from the highest likelihood to the lowest. We finally plotted all obtained values of the sensitivity and the false positive rate to draw an ROC curve.

The AUC, a popular metric for measuring the performance of different models [5, 18], can be computed as the area under this ROC curve. We can see that the larger the AUC, the better the performance of the model. We further used the paired sample two-tailed t-test to statistically evaluate the performance difference between 3MAM and another model. Since we run crossvalidation 50 times, we have at least 100 values in each of the three different ratios, and so if the t-value is greater than 3.50 (2.36) then the difference is more than 99.9% (98%) statistically significant.

2) Log-likelihood Distribution on Positive Test

All these four probabilistic models are trained in an unsupervised manner and the maximum likelihood setting, meaning that they are trained to provide the maximum likelihoods to given training data. In addition, conveniently enough, they have the same (common) set[1] of parameters, ie $p(c_i|z_k)$, $p(g_j|z_k)$ and $p(z_k)$. Thus, we can compare the four models each other by the distribution of the likelihoods for positive test examples, given by each of the models. If a model provides positive examples with higher likelihoods than those of another, we can say that this model is better than the other.

Results

1) AUC

Table 3 shows the AUC for each of the four models at different data settings and the t-value (in parenthesis) between the AUC of 3MAM and that of another model.

Table 3: AUCs and t-values (in parenthesis) obtained by 50 rounds of cross-validation on cancer-gene pairs.

Model	Ratio of training to test data		
	3:1	1:1	1:3
3MAM (CG+CC+GG)	**76.1**	**74.6**	**73.2**
2MAM (CG+CC)	75.8 **(2.56)**	74.2 **(2.44)**	71.8 **(12.9)**
2MAM (CG+GG)	73.9 **(17.2)**	71.4 **(22.5)**	68.3 **(38.0)**
AM (CG)	74.1 **(14.7)**	70.5 **(26.3)**	64.9 **(55.1)**

This table clearly shows that 3MAM outperformed the other three models, and the second best model is 2MAM (CG+CC). We can easily see that, compared with AM, the 3MAM improved around 2 to 9% in the discriminative accuracy. Furthermore, the t-values showed that 3MAM outperformed all other models by a statistically significant factor in all cases. These results indicate that incorporating cancer-cancer and gene-gene pairs from diverse sources improved the predictive performance obtained by cancer-gene pairs only.

In addition, we note the following two points on these results: First, interestingly, 2MAM (CG+GG) outperformed AM in 1:1 and especially 1:3 cases, but not 3:1 case. This is probably because gene-gene co-occurrence data comes from the different source, Medline, which can supplement original data, when it is scarce, and can achieve better performance. Second, since we have only 21 type of cancers and 2,017 genes, some putative negative test examples must be positive. This means that the performance of our model may be under-estimated.

2) Log-likelihood Distribution on Positive Test

When the probability parameter has a uniform distribution, a randomly generated cancer-gene pair has the following log-likelihood:

$$\log\left(\frac{1}{21}\times\frac{1}{2,017}\right)=-4.63$$

In our unsupervised setting, the log-likelihood of a positive example should be larger than the above value. In other words, when positive (test) examples are given, a better trained probabilistic model would provide a larger number of examples whose log-likelihoods are larger than the above value.

[1]We note that trained models have different parameter values because the training algorithms are different.

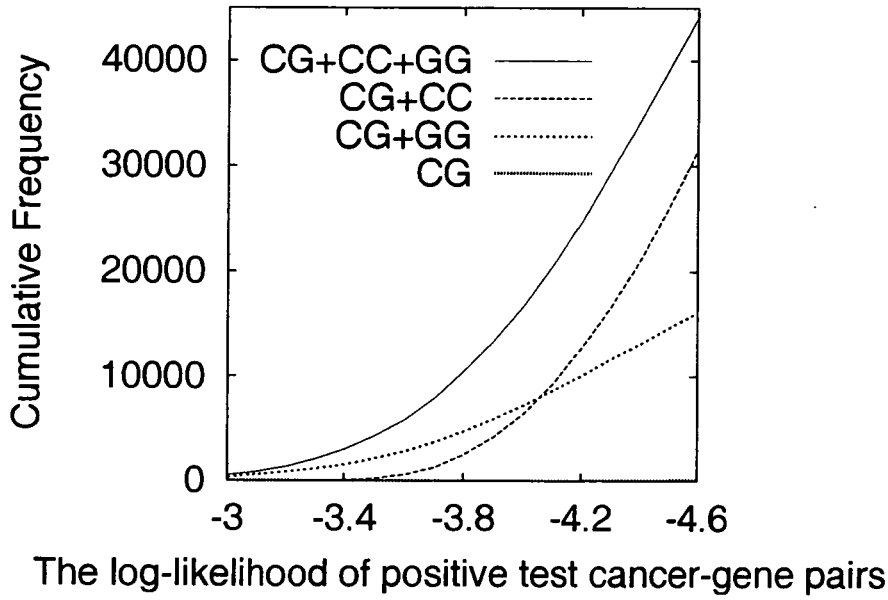

Figure 1: Cumulative number of positive examples with higher log-likelihoods.

Thus, given a cut-off value, we checked the number of positive test examples having log-likelihoods larger than the given cut-off value. Figure 1 shows the counted cumulative number of positive test pairs with higher likelihoods against a given cut-off value. This figure is drawn from the average over the 50 rounds of our cross-validation at the 3:1 ratio of training to test data. We found that 3MAM is clearly the best among the four models, always keeping the largest number of examples whose likelihoods higher than a given cut-off value. These results also confirmed the performance advantage of 3MAM over other models and showed adding cancer-cancer and cancer-gene datasets is effective. Another empirical finding in this analysis is that 2MAM (CG+GG) outperformed 2MAM (CG+CC) in the range of larger than – 4, while 2MAM (CG+CC) outperformed 2MAM (CG+GG) in the range between – 4.6 and – 4.

Mining and Analyzing Unknown Cancer Associated Genes

Mining New Cancer-Gene Co-occurrences

We trained 3MAM using all three types of co-occurrence data and tried to find new associated cancer gene pairs which are unknown in the current literature. The procedure is as follows: We first trained 3MAM using all the three types of co-occurrence data and then computed the log-likelihoods of all cancer-gene paris that are not in the current cancer-gene co-occurrence data. We repeated this procedure 100 times and ranked the new pairs according to the average log-likelihoods over 100 times. Table 4 shows the list of top 20 pairs with their log-likelihoods, and a more

Table 4: 20 Cancer-gene pairs with highest log-likelihoods that are not in our training dataset.

Cancer Type	Gene Name	Log-likelihood
OVARY	TP53	– 3.078
COLORECTAL	BCL2	– 3.085
STOMACH	TP53	– 3.113
LEUKEMIA	CDKN1A	– 3.176
LYMPHOMA	BAX	– 3.191
PANCREAS	TP53	– 3.199
BREAST	NFKB1	– 3.222
THYROID	TP53	– 3.234
LYMPHOMA	TNF	– 3.235
LUNG	BCL2	– 3.244
BREAST	BCL2	– 3.266
KIDNEY	TP53	– 3.269
BREAST	TNF	– 3.293
LEUKEMIA	TNF	– 3.300
COLORECTAL	TNF	– 3.312
LYMPHOMA NF	NFKB1	– 3.316
LUNG	TNF	– 3.323
COLORECTAL	CASP8	– 3.330
LEUKEMIA	NFKB1	– 3.336
BRAIN	BCL2	– 3.340

detailed list of top 1,000 pairs is given in Table 1 of the on-line supplementary information. The first, second, third and fourh columns of the on-line information show cancer names, HUGO IDs [43], genes and log-likelihoods, respectively.

As shown in Table 4, the top 20 list has some famous oncogenes such as TP53, BCL2 and TNF. This result implies that our prediction worked well, because these popular genes must be related with a lot of different types of cancers. So we can expect that these relations must exist, even if the cancer-gene co-occurrences in Table 4 are not in OMIM. In other words, we may say that these relations are easily expected. Thus in the next section, we focused on genes which are specific to some cancer but unknown and tried to analyze how the found genes are related with the corresponding cancer.

Mining New Genes Specific to Cancer

We computed the following score for all cancer-gene pairs by using the probability parameters of 3MAM, which was trained by using all three types of training data.

$$R(g_j, c_i) = \frac{p(g_j \mid c_i)}{\Sigma_i p(g_j \mid c_i)}$$

where

$$p(g_j \mid c_i) = \frac{\Sigma_h p(c_i \mid z_h) p(g_j \mid z_h) p(z_h)}{\Sigma_{j',k} p(c_i \mid z_k) p(g_{j'} \mid z_k) p(z_k)}.$$

The $p(g_j \mid c_i)$ is the conditional probability that given a cancer type c_i, g_j is related with the c_i. Thus the score $R(g_j, c_i)$ is the ratio that a gene g_j is related with c_i, comparing to all the other cancer types. That is, it is the probability over cancer types and shows to what extent gene g_j is specific to cancer c_i. Once we computed the score for each pair, we sorted the values for each cancer and selected the top 20 genes which are not in the cancer-gene pairs in the training data. Table 2 of the on-line supplementary information shows the list of top 20 genes of each cancer. The first, second, third and fourh columns of this file show cancer names, HUGO IDs, genes and parameter values, respectively.

These pairs are unknown pairs in OMIM and Medline, but our method suggested that each of them has a strong relationship between a cancer and a gene. In fact, we can see a biological relationship for each pair from the literature. Below we briefly describe the biological, medical and genetic relationships on each pair of the list, for only the top gene of seven cancers out of all 21 cancers, owing to the space limitations.

Brain:

The top is MMP17. According to Puente et al [36], they revealed that MMP17 is expressed mainly in the brain, leukocytes, colon, ovary and testis, using northern blot analysis of polyadenylated RNAs isolated from a variety of human tissues. This implies MMP17 can be related with brain cancer.

Breast:

The top is ZAP70, a member of the Syk tyrosine kinase family. Recently, Gatalica and Bing [15] pointed out that the loss of Syk tyrosine kinase expression characterises a subset of breast carcinomas. This implies a relationship between ZAP70 and breast cancer.

Colorectal:

The top is CYP1A1. Hou et al [21] recently reported the relationship between the CYP1A1 polymorphism and the risk for colorectal adenoma. Their summary is that the joint carriage of CYP1A1 and NQO1 polymorphisms, particularly in smokers, was related to colorectal adenoma risk, with a propensity for formation of multiple lesions. This would be an evidence for the relationship between CYP1A1 and colorectal cancer. The second is MAD2. The expression profile of MAD2 in colorectal cancer was investigated by Li et al [26]. Their result shows that the defect of spindle checkpoint gene MAD2 is involved mainly in colorectal carcinogenesis. So this clearly indicates the relationship between MAD2 and colorectal cancer.

Lymphoma:

The top is LMO1. In the recent study of leukemogenesis, Lin et al [27] found that almost 60% of transgenic mice that overexpressed both OLIG2 and LMO1 developed pre-T LBL with large thymic tumor masses. This reveals the association between LMO1 and lymphoma cancer.

Pancreas:

The top is NR5A2. NR5A2, a member of a nuclear receptor subfamily, is a liver recepter homolog1 (LRH-1). Fayard et al [12] showed that LRH-1 is abundantly expressed in pancreas. Furthermore, their in situ hybridization and gene expression studies demonstrated that both LRH and carboxyl ester lipase (CEL) are co-expressed and confined to the exocrine pancreas.

Prostate:
The top is KLK10, ie kallikrein 10. Bharaj et al [3] showed the association between single nucleotide polymorphisms in the human KLK10 and prostate cancer. Petraki et al [31] studied the localization of human KLK10 in benigh and malignant prostatic tissues and the correlation between the expression of KLK10 and prostate cancer (PC) prognosis. They pointed out that kallikreins may function as tumor suppressors or are down-regulated during cancer progression. These results imply the relationship between KLK10 and prostate cancer.

Testis:
GAGEB1 is the top. Chen et al [9] isolated GAGEB1 by differential display PCR. They found that GAGEB1 expression was restricted to testes and placenta on human multiple tissue Northern blots. This shows some relationship GAGEB1 and testis cancer.

Concluding Remarks

We have applied a new probabilistic model MAM, which was proposed by us in our research on mining implicit chemical compound-gene relationship, to the problem of finding new cancer associated genes from OMIM and Medline. MAM can integrate different types of co-occurrence datasets effectively, and we found that MAM performed very well even when co-occurrence datasets are gathered from heterogeneous sources.

In this work, we used a uniform distribution for the component weights (π) of our mixture model to allow users additional control. Interesting future work would adjust the weights to achieve the maximum predictive performance. On the other hand, the gene-gene co-occurrence data can come from a different source other than Medline. Since microarray expression data can reveal the biological relationship of genes, it would be very interesting to integrate gene-gene co-occurrence data from microarray expressions.

Acknowledgement

This work is supported in part by Bioinformatics Education Program "Education and Research Organization for Genome Information Science" and Kyoto University 21st Century COE Program "Knowledge Information Infrastructure for Genome Science" with support from MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan.

References

[1] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990), Basic local alignment search tool, *J Mol Biol*, 215(3): 403–410.

[2] Bajdik CD, Kuo B, Rusaw S, Jones S and Brooks-Wilson A. (2005), CGMIM: Automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and Candidate genes, *BMC Bioinformatics*, 6:78–84.

[3] Bharaj BB, Luo LY, Jung K, Stephan C, Diamandis EP (2002) Identification of single nucleotide polymorphisms in the human kallikrein 10 (KLK10) gene and their association with prostate, breast, testicular, and ovarian cancers. *Prostate*, 51(1):35–41.

[4] Boguski MS, Lowe TM, Tolstoshev CM. (1993) dbEST–database for "expressed sequence tags" *Nat Genet.* 4(4):332–3.

[5] Bradley A. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30:1145–1159.

[6] Brancolini V and Devoto M. (1996) Genetic linkage studies for the identification of cancer-related genes. *Ann Ist Super Sanita.* 32(1):173–180.

[7] Cardon LR and Bell JI. (2001) Association study designs for complex diseases. *Nat Rev Genet.* 2(2):91–99.

[8] Chang JT and Altman RB. (2004) Extracting and characterizing gene-drug relationships from the literature, *Pharmacogenetics*, 14:577–586.

[9] Chen ME, Lin SH, Chung LW, Sikes RA. (1998) Isolation and characterization of PAGE-1 and GAGE-7. New genes expressed in the LNCaP prostate cancer progression model that share homology with melanoma-associated antigens. *J. Biol. Chem.*, 273(28):17618–17625.

[10] Dempster A, Laird N and Rubin D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1-38.

[11] Forozan F, Karhu R, Kononen J, Kallioniemi A and Kallioniemi OP.(1997) Genome screening by comparative genomic hybridization. *Trends Genet.* 13(10):405–409.

[12] Fayard E, Schoonjans K, Annicotte JS and Auwerx J. (2003) Liver receptor homolog 1 controls the expression of carboxyl ester lipase. *J. Biol. Chem.* 278(37):35725–35731.

[13] Freudenberg J and Propping P. (2002), A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18, Suppl. 2:S110–S115.

[14] Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM and Whelan S. *International Classification of Diseases for Oncology* Third edition. World Health Organization; 2000.

[15] Gatalica Z and Bing Z., Syk tyrosine kinase expression during multistep mammary carcinogenesis. *Croat Med J.*, 46(3):372–376.

[16] Guo QM, DNA microarray and cancer. (2003) *Curr Opin Oncol*, 15:36–43.

[17] Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. (2005) Online Meddelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33:D514–D517.

[18] Hand DJ and Till RJ. (2001) A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186.

[19] Hofmann T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

[20] Hofmann T. (2004) Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22: 89–115.

[21] Hou L, Chatterjee N, Huang WY, Baccarelli A, Yadavalli S, Yeager M, Bresalier RS, Chanock SJ, Caporaso NE, Ji BT,Weissfeld JL and Hayes RB. (2005) CYP1A1 Val462 and NQO1 Ser187 polymorphisms, cigarette use, and risk for colorectal adenoma. Carcinogenesis, 26(6):1122–1128.

[22] Jenssen T, Laegreid A, Komorowski J and Hovig E. (2001), A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28:21–28.

[23] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258(5083):818–821.

[24] Kim JM, Sohn HY, Yoon SY, Oh JH, Yang JO, Kim JH, Song KS, Rho SM, Yoo HS, Kim YS, Kim JG and Kim NS. (2005) Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. *Clinical Cancer Research* 11:473–482.

[25] Kinzler KW and Vogelstein B, (2002) *The genetic basis of human cancer* edn 2, Toronto, McGraw-Hill.

[26] Li GQ, Li H and Zhang HF (2003) Mad2 and p53 expression profiles in colorectal cancer and its clinical significance. *World J Gastroenterol.*, 9(9):1972–1975.

[27] Lin YW, Deveney R, Barbara M, Iscove NN, Nimer SD, Slape C and Aplan PD (2005) OLIG2 (BHLHB1), a bHLH transcription factor, contributes to leukemogenesis in concert with LMO1. *Cancer Research,* 65(16):7151–7158.

[28] Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L and Wigler M. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* 13(10):2291–2305.

[29] McKusick VA (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn. Johns Hopkins University Press, Baltimore, MD.

[30] Perez-Iratxeta C, Bork P and Andrade MA (2002), Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 31:316–319.

[31] Petraki CD, Gregorakis AK, Papanastasiou PA, Karavana VN, Luo LY and Diamandis EP. (2003) Immunohistochemical localization of human kallikreins 6, 10 and 13 in benign and malignant prostatic tissues. *Prostate Cancer Prostatic Dis.* 6(3):223–227.

[32] Pinkel D, Segraves R, Sudar D, et al. (1998) High resolution analysis of DNA copy-number variation using comparative genomic hybridization to microarray. *Nat. Genet,* 20:207–211.

[33] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays.*Nat Genet.* 23(1):41–46.

[34] Pruitt K and Maglott D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, 29:137–140.

[35] Pearson WR and Lipman DJ.(1988) Improved tools for biological sequence comparison. *PNAS,* 85(8):2444–2448.

[36] Puente XS, Pendas AM, Llano E, Velasco G and Lopez-Otin C. (1996) Molecular cloning of a novel membrane-type matrix metalloproteinase from a human breast carcinoma. *Cancer Research,* 56(5): 944–949.

[37] Qiu P, Wang L, Kostich M, Ding W, Simon JS and Greene JR.(2004) Genome wide in silico SNP-tumor association analysis. *BMC Cancer.* 4:4.

[38] Roylance R,(2002) Methods of molecular analysis: assessing losses and gains in tumors. *Mol Pathol* 55:25–28

[39] Shen D, He J and Chang HR. In silico identification of breast cancer genes by combined multiple hight throughput analyses. *Int J Mol Med,* 15(2):205–212.

[40] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1):308–311.

[41] Thorisson GA, Smith AV, Krishnan L and Stein LD. (2005) The International HapMap Project web site. *Genome Research.* 15: 1592–1593.

[42] Velculescu VE, Zhang L, Vogelstein B and Kinzler KW. (1995) Serial analysis of gene expression. *Science,* 270:484–487.

[43] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW and Povey S. (2002) Guidelines for human gene nomenclature. *Genomics.* 79(4):464–470.

[44] Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW and Velculescu VE. (2002). Digital karyotyping. *PNAS.* 99(25):16156–16161.

[45] Wheeler D, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S and Helmberg W et al (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33: D39–D45.

[46] Wilkinson DM and Huberman BA. (2004), A method for finding communities of related genes. *PNAS:* 101, 5241–5248.

[47] Yandell MD and Majoros WH. (2002) Genomics and natural language processing. *Nat. Rev. Genet.*, 3: 601–610.

[48] Zhu S, Okuno Y, Tsujimoto G, and Mamitsuka H. (2005), A probabilistic model for mining implicit "Chemical compound-gene" relations from literature. *Proc. of ECCB2005 (Bioinformatics 21 Supplement 2):* ii245–ii251.