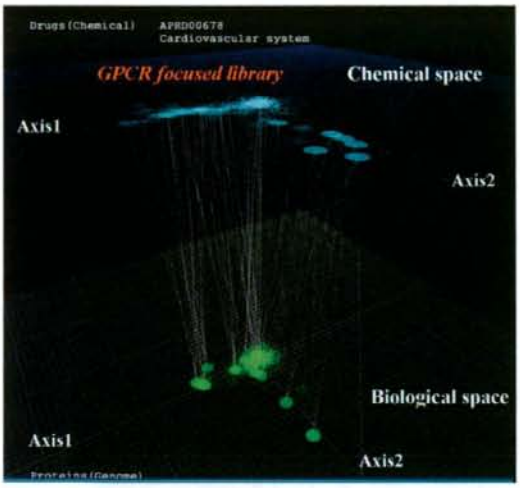

図５．DrugBank データを用いた実際のケミカル空間とバイオ空間

７、おわりに

医薬品の開発プロセスにおいて、現在用いられている In silico 技術には、我々が研究開発を行う情報科学的アプローチの他に、立体構造モデルを用いたドッキングシミュレーションのような計算化学的アプローチが有名である。情報科学的アプローチと計算化学的アプローチには、それぞれ一長一短があるが、欠点を互いに補完し合い In silico 創薬の確度向上を図ることが今後の課題であろう。特に、ケミカルゲノミクスが盛んな今日、日々増加し続ける莫大なデータを処理することは必須であり、情報科学的アプローチである創薬インフォマティクスのさらなる研究開発が必要である。

謝辞

本研究の一部は、文部科学省、厚生労働省の支援によって行われている。また、検証実験等の共同研究を行って頂いている京都大学薬学研究科ゲノム創薬科学分野の辻本豪三教授に深く感謝申し上げる。

参考文献

1) Nature, 432 (7019) (Insight), 823-865
2) Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., Tsujimoto, G., Nucleic Acids Research, 34, D673-677
3) http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
4) http://redpoll.pharmacy.ualberta.ca/drugbank/

◆略　歴◆　奥野　恭史　(Yasushi OKUNO)：1995 年京大薬・修士課程修了、京大薬・博士後期課程進学、1996 年京大薬・博士後期課程中途退学、京都大学化学研究所教務職員、2000 年京大薬博士号取得、2001 年京都大学化学研究所博士研究員、2002 年京都大学化学研究所特任助手、2003 年京都大学大学院薬学研究科特任助手、2005 年産業技術総合研究所外来研究員（併任）、2006 年京都大学大学院薬学研究科助教授

MOLECULAR AND CELLULAR BIOLOGY, Apr. 2007, p. 3008–3022 Vol. 27, No. 8
0270-7306/07/$08.00+0 doi:10.1128/MCB.02047-06

Germinal Center Marker GL7 Probes Activation-Dependent Repression of *N*-Glycolylneuraminic Acid, a Sialic Acid Species Involved in the Negative Modulation of B-Cell Activation▽†

Yuko Naito,[1,7] Hiromu Takematsu,[1,7] Susumu Koyama,[2] Shizu Miyake,[2] Harumi Yamamoto,[5] Reiko Fujinawa,[5] Manabu Sugai,[4] Yasushi Okuno,[3] Gozoh Tsujimoto,[3] Toshiyuki Yamaji,[5] Yasuhiro Hashimoto,[5,7] Shigeyoshi Itohara,[6] Toshisuke Kawasaki,[2]‡ Akemi Suzuki,[5] and Yasunori Kozutsumi[1,5,7]*

Laboratory of Membrane Biochemistry and Biophysics, Graduate School of Biostudies,[1] Department of Biological Chemistry,[2] and Department of Genomic Drug Discovery, Graduate School of Pharmaceutical Sciences,[3] and Center for Genomic Medicine, Graduate School of Medicine,[4] Kyoto University, Sakyo, Kyoto 606-8501, Japan; Supra-Biomolecular System Research Group, RIKEN Frontier Research System,[5] and Laboratory for Behavioral Genetics, RIKEN Brain Science Institute,[6] RIKEN, Wako, Saitama 351-0198, Japan; and CREST, Japan Science and Technology, Kawaguchi, Saitama, Japan[7]

Received 2 November 2006/Returned for modification 9 January 2007/Accepted 30 January 2007

Sialic acid (Sia) is a family of acidic nine-carbon sugars that occupies the nonreducing terminus of glycan chains. Diversity of Sia is achieved by variation in the linkage to the underlying sugar and modification of the Sia molecule. Here we identified Sia-dependent epitope specificity for GL7, a rat monoclonal antibody, to probe germinal centers upon T cell-dependent immunity. GL7 recognizes sialylated glycan(s), the α2,6-linked *N*-acetylneuraminic acid (Neu5Ac) on a lactosamine glycan chain(s), in both Sia modification- and Sia linkage-dependent manners. In mouse germinal center B cells, the expression of the GL7 epitope was upregulated due to the in situ repression of CMP-Neu5Ac hydroxylase (Cmah), the enzyme responsible for Sia modification of Neu5Ac to Neu5Gc. Such Cmah repression caused activation-dependent dynamic reduction of CD22 ligand expression without losing α2,6-linked sialylation in germinal centers. The in vivo function of *Cmah* was analyzed using gene-disrupted mice. Phenotypic analyses showed that Neu5Gc glycan functions as a negative regulator for B-cell activation in assays of T-cell-independent immunization response and splenic B-cell proliferation. Thus, Neu5Gc is required for optimal negative regulation, and the reaction is specifically suppressed in activated B cells, i.e., germinal center B cells.

The germinal center is a special microenvironment which occurs in secondary lymphoid organs, mainly in response to T-cell-dependent antigen immunization. Mature B cells entering the germinal center edit their immunoglobulin gene through somatic hypermutation and class-switching recombination, differentiating into memory cells and plasma cells (30, 33). The activated B cells during the germinal center reaction in mice can be probed with peanut (*Arachis hypogaea*) lectin, peanut agglutinin (PNA) (8, 37, 46), or a rat monoclonal antibody (MAb), GL7 (5). GL7 was originally reported as a marker for polyclonally activated T and B cells (28) in mice. GL7 stains a subpopulation of T cells (19) and a subpopulation of the large pre-B-cell stage during differentiation in the bone marrow (38). Activated B cells express the GL7 epitope, but mature B cells do not; thus, GL7 serves as a marker for germinal centers in the immunized spleen (18, 41, 52) or lymph nodes, and $GL7^{high}$ B cells have been shown to have higher functional activity for producing antibodies and presenting antigens (5). Despite growing knowledge about the use of this antibody as a marker for lymphocytes in various conditions, the molecular epitope of GL7 is poorly defined to date. In the original article characterizing GL7, Laszlo et al. (28) showed that GL7 could immunoprecipitate a 35-kDa cell surface protein from metabolically labeled activated B cells. However, no other studies have been published on this subject.

In the present study, we found that GL7 recognizes a glycan moiety containing terminal sialic acid (Sia) in both linkage- and modification-dependent manners. Sia is a family of acidic nine-carbon sugars that often occupies the nonreducing terminus of mammalian glycan chains (47), and Sia is essential for early development of mice (49). The localization of Sia-bearing glycan chains on the cell surface makes sialylated molecules seem to be likely targets for various cellular and molecular recognition molecules, such as the mammalian lectins that are abundant in the immune system (61). A family of enzymes, sialyltransferases, is responsible for the formation of the Sia linkage to the underlying glycan chains. To determine the

* Corresponding author. Mailing address: Laboratory of Membrane Biochemistry and Biophysics, Graduate School of Biostudies, Kyoto University, Yoshida-shimoadachi, Sakyo-ku, Kyoto 606-8501, Japan. Phone: 81 75 753 7684. Fax: 81 75 753 7686. E-mail: yasu@pharm.kyoto-u.ac.jp.

‡ Present address: Research Center for Glycobiotechnology, Ritsumeikan University, Kyoto, Japan.

† Supplemental material for this article may be found at http://mcb.asm.org/.

▽ Published ahead of print on 12 February 2007.

linkage specificity of GL7 recognition, we used the gene expression profiles of sialylation-related genes obtained by DNA microarray analysis to screen for a responsible sialyltransferase gene for the biosynthesis of the GL7 determinant.

Apart from the linkage variations, Sia also occurs in various molecular species as a result of modifications at its C-4, C-5, C-7, C-8, and C-9 positions; these modifications are spatially and temporary regulated (60). We also found that the determinant recognition by GL7 is specific to a Sia modification at the C-5 position. In mice, Sia occurs in two main forms with respect to the moiety at the C-5 position: *N*-acetylneuraminic acid (Neu5Ac), which is a precursor form of the diverse Sia family, and its major modified form, *N*-glycolylneuraminic acid (Neu5Gc). The structural difference between Neu5Ac and Neu5Gc is a single oxygen atom in the C-5 position. The modification reaction that produces Neu5Gc is catalyzed at the sugar-nucleotide level in the cytosol by the enzyme CMP-Neu5Ac hydroxylase (Cmah) (24, 53). Cmah determines the cell surface expression ratio of these two Sia species, as the cytosolic Cmah reaction occurs prior to the sialyltransferase reaction, which takes place in the Golgi apparatus during the biosynthesis of glycoconjugates. We found that GL7 recognizes only Neu5Ac-bearing glycans and that the reduction of *Cmah* expression plays a major role in the formation of the GL7 epitope in activated B cells in the germinal center, which was in sharp contrast to the dominant expression of Neu5Gc in mouse lymphocytes.

To examine the in vivo function of Neu5Gc-bearing glycans, we disrupted the *Cmah* gene in mice. *Cmah* disruption is expected to modify the Sia-mediated Sia species-specific recognition event without affecting overall sialylation, which can affect the behavior of the protein in various ways. We primarily focused on the phenotypic consequences of *Cmah* disruption in B cells since Cmah is regulated in B cells, especially in response to activation. *Cmah*-null mice exhibited hyperresponsive B cell phenotypes in assays measuring B-cell functions, i.e., antibody production and proliferation.

MATERIALS AND METHODS

Materials. Most of the materials used were obtained from Wako Chemical (Osaka, Japan) or Nacalai Tesque (Kyoto, Japan). The human immunoglobulin G1 (IgG1)-Fc fusion construct was provided by Paul Crocker and Ajit Varki. The Lec2 cells were provided by Pamela Stanley. The Plat-E cells were provided by Toshio Kitamura. Human B-cell lines were obtained from the Japanese Collection of Research Bioresources.

Antibodies and lectins. The antibodies used were as follows: donkey $F(ab')_2$ against mouse IgM (Jackson ImmunoResearch, West Grove, PA); R-phycoerythrin (R-PE)-conjugated anti-mouse B220 (RA3-6B2); R-PE-conjugated goat $F(ab')_2$ anti-human IgG-Fc; R-PE-conjugated streptavidin (CALTAG Laboratories, Burlingame, CA); fluorescein isothiocyanate (FITC)-conjugated and purified GL7; FITC-conjugated anti-mouse B220 (RA3-6B2); R-PE-conjugated anti-mouse I-A/I-E (M5/114.15.2); biotin-conjugated anti-CD22 (Cy34.1) (BD Pharmingen, San Diego, CA); horseradish peroxidase (HRP)-conjugated goat anti-rat IgM; alkaline phosphatase-conjugated isotype-specific goat anti-mouse IgA, IgG1, IgG3, and IgM; unlabeled isotype-specific goat anti-mouse IgA and IgG3; R-PE-conjugated anti-mouse IgM (1B4B1); biotin-conjugated anti-mouse CD22 (2D6) (Southern Biotechnology Associates, AL); anti-mouse polyvalent Igs; HRP-conjugated PT-66 (an antiphosphotyrosine MAb; Sigma, St. Louis, MO); CD90 (Thy1.2) MicroBeads; anti-FITC MicroBeads (Miltenyi Biotec, Bergisch Gladbach, Germany); rabbit anti-mouse CD22 serum (Chemicon, Temecula, CA), HRP-conjugated donkey $F(ab')_2$ anti-rabbit Ig (Amersham Life Science, Buckinghamshire, United Kingdom); antiactin (Santa Cruz Biotechnology, Santa Cruz, CA); HRP-conjugated goat anti-mouse IgG; HRP-conjugated rabbit anti-goat IgG (ZYMED Lab, South San Francisco, CA). Anti-CD22 MAb (Cy34.1) was purified from the culture supernatant of hybridoma Cy34.1 (ATCC). Biotinylated *A. hypogaea* PNA was obtained from HONEN (Tokyo, Japan), and FITC-conjugated *Sambucus sieboldiana* agglutinin (SSA) was obtained from Seikagaku Kogyo (Tokyo, Japan).

Preparation of Fc fusion proteins of sialoadhesin and CD22. Recombinant soluble forms of the amino-terminal domains (domains 1 to 3) of mouse sialoadhesin/Siglec-1, mouse CD22/Siglec-2, and human CD22/Siglec-2 fused to the Fc region of human IgG1 (mSn-Fc, mCD22-Fc, and hCD22-Fc, respectively) were produced in stably transfected Lec2 cells, a cell line deficient in protein sialylation. The production of the Siglec (Sia-binding Ig superfamily lectin)-Fc fusion probe in the Lec2 cell line resulted in considerably enhanced binding to the ligand, which allowed the identification of changes in ligand expression. The Siglec-Fc probes were purified from the culture supernatant using protein A-Sepharose columns (Pierce, Rockford, IL).

Flow cytometry. Cell labeling was carried out in fluorescence-activated cell sorter buffer (1% bovine serum albumin [BSA] and 0.1% NaN_3 in phosphate-buffered saline [PBS]). Data were acquired using a FACScan (Becton Dickinson, Franklin Lakes, NJ) instrument and analyzed using FlowJo software (Tree Star, San Carlos, CA). For comparison with the microarray data, B lymphoma cells (1×10^5) were stained with FITC-conjugated GL7 (dilution, 1:100) for 1 h. This staining condition was determined using the criterion that the strongest staining did not reach a plateau. Mean fluorescence intensity (MFI) of GL7 staining was acquired using a FACScan at settings under which unstained control cells gave a signal of around 5 on the FL-1 channel. The mean FL-1 signal of each stained sample was divided by that of the unstained sample to produce the relative staining profiles on flow cytometry to be compared with the cDNA microarray profiles of relative gene expression. For mSn-Fc, mCD22-Fc, and hCD22-Fc staining, these Fc fusion proteins were precomplexed with R-PE-conjugated goat $F(ab')_2$ anti-human IgG.

Sialidase treatment. Sialidase treatment was carried out in 100 mM sodium acetate (pH 5.2) for 30 min at room temperature prior to the staining for flow cytometry. Sialidase from *Arthrobacter ureafaciens* (Calbiochem, San Diego, CA) and sialidase from *Salmonella enterica* serovar Typhimurium (Takara, Kusatsu, Japan) were used.

Immunoblotting with GL7. The cells were sonicated in detergent-free lysis buffer (25 mM Tris-HCl [pH 7.6], 1 mM dithiothreitol, protease inhibitor cocktail [Nacalai Tesque]). The pellets (membrane fractions) were collected by ultracentrifugation and solubilized in NP-40 lysis buffer (1% Nonidet P-40, 150 mM NaCl, 25 mM HEPES [pH 7.4], protease inhibitor cocktail). The extracts were subjected to immunoblotting with GL7 in the presence or absence of 100 mM Neu5Ac.

Development of cDNA microarray for glycan-related genes. The RIKEN Frontier Human Glyco-gene cDNA microarray, version 2, which was spotted by Takara, consisted of 888 genes, which included glycosyltransferase genes and genes related to sugar metabolism, glycan modification, glycan recognition, and lipid metabolism.

Use of cDNA microarray for identification of glycan-related genes. Poly(A)$^+$ RNA samples were isolated from mid-log-phase cells using the mTRAP system (Activemotif, Carlsbad, CA) and were quality checked using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). One microgram of poly(A)$^+$ RNA from the B-cell lines (rRNA contamination subtracted) and universal reference RNA (Clontech, Mountain View, CA) were labeled using a CyScribe first-strand cDNA labeling kit (Amersham). Competitive hybridization was performed on the microarray, and data were obtained using an Affymetrix 428 array scanner. To achieve a fair cross-cell line comparison, we fixed Cy3 as the signal for the universal reference RNA and Cy5 for the RNA from the B-cell lines. Microarray data were background corrected using a smoothing function and then Lowess normalized using linear models for microarray data. This readout was sigma normalized to avoid variation among microarray replicates. Then, the Cy5 signal from the B-cell lines was divided by the Cy3 signal to obtain the relative expression profile for each gene in the six cell lines as expression ratios relative to the universal reference RNA (1, 16, 29, 40). The gene expression profiles were compared with the GL7 staining profiles from flow cytometry. The similarity between the profiles was evaluated with Pearson's correlation coefficient, and probability values (P values) were calculated by the correlation coefficient test. For the correlation coefficient test of a sample size of six, a coefficient of 0.81 indicates a statistical significance level of 5%.

Transfection. CHO-K1 cells were stably transfected with pIRES (where IRES is internal ribosome entry site) vector (Clontech), either with or without rat cDNA for *St6gal1*. Transfected cells were selected with G418 (1 mg/ml), and multiple stable clones were established.

Enzyme-linked immunosorbent assay (ELISA). In 96-well assay plates, GL7 antibody was immobilized in wells coated with the capturing antibody, purified

anti-rat IgM. The wells were washed and incubated with streptavidin-conjugated sugar chain probes (50 μM), prepared as previously reported (65). The captured probes were detected with biotinylated alkaline phosphatase (Vector Laboratories, Burlingame, CA) and *p*-nitrophenyl phosphate by measuring the absorbance at 405 nm.

Spleen sectioning and immunohistochemistry. Mice were immunized intraperitoneally with 3×10^8 sheep red blood cells (SRBC) in 100 μl of saline. Spleens were removed 8 or 10 days after immunization and embedded in Tissue-Tek OCT (22-oxyacalcitriol) compound (Sakura Finetechnical, Tokyo, Japan). Spleen sections were cut at a 6-μm thickness on a cryostat microtome (Leica Geosystems, Heerbrugg, Switzerland), thaw-mounted onto Matsunari adhesive silane-coated slides, and fixed in acetone. After rehydration in Tris-buffered saline and blocking in Tris-buffered saline with 5% BSA and 0.05% Tween 20, the sections were stained with GL7, PNA, or mCD22-Fc precomplexed with R-PE-conjugated anti-human IgG. The stained sections were analyzed under a confocal laser-scanning microscope (Olympus, Tokyo, Japan).

Magnetic sorting preparation of splenic B-cell-enriched fraction. B-cell-enriched fractions were obtained by Thy1.2 depletion of splenocytes on a MACS (magnetic cell sorter) depletion column (Miltenyi Biotec). Thy1.2-depleted fractions were stained with B220 to confirm B-cell enrichment. To avoid Neu5Gc contamination in the experimental systems, RPMI 1640 medium (Invitrogen, Carlsbad, CA) containing 10% human serum (Uniglobe, Reseda, CA) or chicken serum (JRH Biosciences, Lenaxa, KS), rather than fetal bovine serum (FBS) (JRH), was used in most of the experiments. In addition, sodium pyruvate (Invitrogen), nonessential amino acids solution (Invitrogen), L-glutamine, and 2-mercaptoethanol were added to the medium.

Germinal center B-cell analyses. Splenic B cells from SRBC-immunized mice were incubated with FITC-conjugated GL7 and then with anti-FITC MicroBeads. The labeled cells were collected as germinal center B cells using a MACS LS column (Miltenyi Biotec). The germinal center, nongerminal center, and control (untreated) B cells were lysed by sonication in detergent-free lysis buffer (described above), and the lysates were separated by ultracentrifugation. The supernatant (cytosolic fraction) was used for immunoblotting with anti-Cmah antibody, and the pellet (membrane fraction) was used for the analysis of Sia species by high-pressure liquid chromatography (HPLC). Immunoblotting was performed using rabbit N8 antiserum against mouse Cmah, as previously reported (27). The ratios of Neu5Gc were determined by derivatizing Sia with 1,2-diamino-4,5-methylenedioxybenzene (DMB), a fluorescent compound for α-keto acids, as previously described (27). In brief, Sia was released by incubating the pellet in 2 M acetic acid at 80°C, derivatized with DMB (Dojindo, Mashiki, Japan), and analyzed on a reverse-phase column (TSK-gel ODS-80Tm; Tosoh, Tokyo, Japan) using a Shimadzu LC10 HPLC system.

Detection of Sia in tissues. The ratios of Neu5Ac and Neu5Gc were determined as above. Sia was released by incubating tissues in 100 mM sulfuric acid (which also destroys the *O*-acetyl group often found on the C-7 to C-9 positions of Sia molecules), derivatized with DMB, and analyzed by HPLC.

Real-time RT-PCR analysis. Real time reverse transcription-PCR (RT-PCR) experiments were performed using a QuantiTect SYBR Green PCR kit (QIAGEN Japan, Tokyo, Japan) and an ABI 7700 sequence detection system (Applied Biosystems Japan, Tokyo, Japan). Total RNA was purified from untreated or lipopolysaccharide (LPS)-stimulated mouse splenic B cells, and 2 μg was used for reverse transcription. The amplification cycle was as follows: 15 min at 95°C, followed by up to 40 cycles of 15 s at 94°C, 30 s at 58°C/50°C, and 30 s at 72°C. The PCR primers used for amplification were: ZP-5, 5′-AGATTTAC AAGGATTCC-3′; ZP-E, 5′-CTTAAATCCAGCCCA-3′ (*Cmah*); PS-mCD22-6, 5′-CCTCCACTCCTCAGGCCAGA-3′; PS-mCD22-E, 5′-GCCTATCCCATTG GTCCCT-3′ (*Cd22*); PS-ST6Gal-1, 5′-TCTTCGAGAAGAATATGGTG-3′; PS-ST6Gal-A, 5′-GACTTATGGAGAAGGATGAG-3′ (*St6gal1*); PS-GAPDH-1 (where GAPDH is glyceraldehyde-3-phosphate dehydrogenase), 5′-GTGGAGATTGTTGCC ATCAACG-3′; PS-GAPDH-A, 5′-TCTCGTGGTTCACACCCATCAC-3′ (*Gapdh*); PS-BACTIN-1, 5′-ACGATATCGCTGCGCTGGTC-3′; and PS-BACTIN-A, 5′-CAT GAGGTAGTCTGTCAGGT C-3′ (*Actb*). Each sample was analyzed in more than three wells. Relative mRNA abundance was calculated using the comparative cycle threshold method and expressed as a ratio to the nonstimulated sample.

Retrovirus preparation and infection. *Cmah* cDNA was cloned into the modified mouse stem cell virus vector, which expresses *Cmah* and the extracellular domain of human *CD4* by means of an internal ribosome entry site. Plasmids were transiently transfected into Plat-E packaging cells (35), and retrovirus-containing supernatants were collected. After stimulation with LPS for 12 to 14 h, splenic B cells were spin infected (at 32°C for 90 min) with the retrovirus in the presence of *N*[1-(2,3-dioleoyloxy)propyl]-*N*,*N*,*N*-trimethylammonium methylsulfate (DOTAP; Roche Diagnostics, Mannheim, Germany). The retrovirus-infected B cells were cultured in the presence of 30 μg/ml LPS for 2 to 2.5 days, and then human CD4-positive cells were enriched with a MACS system using MACSelect 4 MicroBeads (Miltenyi Biotec). The sorted cells were subjected to flow cytometry or a proliferation assay (described below).

Targeting construct and embryonic stem (ES) cells. The *Cmah* targeting vector was assembled from a 129/Sv genomic clone containing exons 4 and 5 of this gene and a neomycin resistance gene driven by the phosphoglycerate kinase 1 promoter (PGK-neoR) as well as a diphtheria toxin A gene fragment driven by the MC1 promoter (DT-A) as positive and negative selection markers, respectively. The construct was created by inserting the PGK-neoR cassette into the NspV site of exon 5 of the *Cmah* gene. The DT-A cassette was then ligated adjacent to the 3′ terminus of the construct.

Generation of mutant mice. Gene targeting and generation of mutant mice were performed essentially as described previously (23). In brief, E14 cells were electroporated with a Bio-Rad Gene Pulser (0.8 kV; 3 μF) using 30 μg of NotI-linearized targeting vector. The electroporated cells were selected in medium containing G418 (125 μg/ml) and screened for homologous recombination by Southern blot analysis of genomic DNA digested with BglI, using both radiolabeled 5′ internal and 3′ external probes. The mutant cells were microinjected into 3.5-day-old C57BL/6J blastocysts, and the embryos were transferred into the uteri of pseudopregnant ICR mice. Mice were used for the determination of immunological features after more than seven backcrosses to the C57BL/6J strain. All mice examined in this study were housed in a specific-pathogen-free facility.

Serum isotype-specific antibody measurement. Serum samples from nonimmunized mice at 8 to 12 weeks of age were subjected to isotype-specific ELISAs. Isotype-specific capturing antibodies were coated onto 96-well ELISA plates, and nonspecific binding was blocked with 1% BSA-supplemented PBS. A serially diluted standard MAb of each isotype (Ancell, Bayport, MN) and diluted serum samples were captured on the wells. The captured Abs were detected with alkaline phosphatase-conjugated isotype-specific goat IgG using a 1420 ARVO SXc (Wallac, Turku, Finland) luminometer.

Determination of antibody production in immunized mice. Eight-week-old mice were immunized after preimmune serum was obtained. Freund's complete adjuvant containing 100 μg of dinitrophenyl (DNP)-keyhole limpet hemocyanin (KLH) was used for primary T-dependent immunization by intraperitoneal injection, and a second boost was performed with the antigen in incomplete adjuvant. For T-independent immunization, 10 μg of DNP-Ficoll in PBS was injected. The anti-DNP titer was measured essentially as above, except that DNP-BSA was used for antibody capture, and a mixed pool of DNP-KLH-immunized serum was used as the standard. The value relative to that of the pooled serum was used to normalize the values obtained from different plates.

B-cell proliferation analysis. In 96-well plates, 100-μl aliquots of B cells at 1×10^5 cells/ml were stimulated in RPMI 1640 medium containing the indicated concentrations of stimulation reagents. After 24 h of incubation, bromodeoxyuridine (BrdU) was added, and the incubation was continued overnight. Incorporated BrdU was detected using a chemiluminescent ELISA system (Roche Diagnostic GmbH) with an 1420 ARVO SXc luminometer.

Immunoblotting and immunoprecipitation of CD22. Splenic B cells were adjusted to 5×10^5 cells/50 μl in RPMI 1640 medium. After preincubation at 37°C, the B cells were stimulated with $F(ab')_2$ anti-mouse IgM (10 μg per 5×10^5 cells) at 37°C. To detect the pattern of tyrosine phosphorylation, cells were lysed in sodium dodecyl sulfate-polyacrylamide gel electrophoresis sample buffer (50 mM Tris-HCl [pH 7.6], 2% sodium dodecyl sulfate, 0.1% pyronin G, 10% glycerol, 2-mercaptoethanol). For immunoprecipitation studies, the stimulated B cells were lysed in NP-40 lysis buffer (1% Nonidet P-40, 150 mM NaCl, 25 mM HEPES [pH 7.4], 5 mM NaF, 2 mM sodium orthovanadate, protease inhibitor cocktail [Nacalai Tesque]), and CD22 was immunoprecipitated with anti-CD22 (Cy34.1) antibody and protein G-Sepharose beads (Amersham Biosciences). In the CD22 immunoprecipitation studies, after a probing step with PT-66, the membrane was reprobed with anti-CD22 polyclonal antibody.

Experimental animals. The studies presented here were performed in accordance with animal care guidelines and were approved by the animal experimental committee of Kyoto University Graduate School of Biostudies.

Microarray data accession numbers. The GEO platform (GPL3465) and experimental results are registered in the Gene Expression Omnibus database under accession number GSE4407.

RESULTS

Sia involvement in GL7 staining of B-cell lines. During B-cell development in mice, the epitope of the MAb GL7 appears and disappears in multiple maturation steps (5, 18, 32,

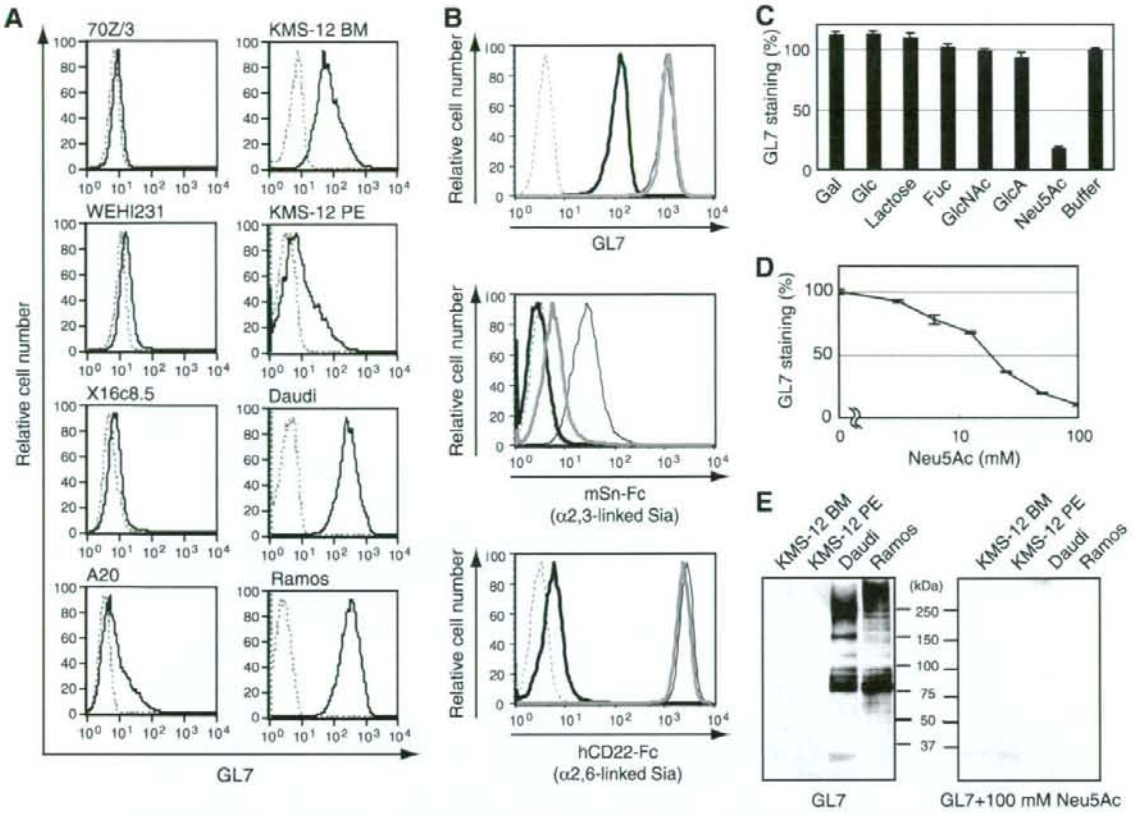

FIG. 1. Involvement of Sia in the GL7 epitope. (A) GL7 staining in flow cytometry. Mouse B-cell lines (70Z/3, WEHI231, X16c8.5, and A20) and human B-cell lines (KMS-12 BM, KMS-12 PE, Daudi, and Ramos) were stained with FITC-conjugated GL7. Black solid lines indicate staining with GL7, and gray dashed lines indicate nonstaining controls. (B) The effect of sialidase treatment on GL7 staining. Daudi cells were treated with sialidase before staining with FITC-conjugated GL7, mSn-Fc, or hCD22-Fc. Gray dashed lines indicate negative controls (nonstaining for GL7 and R-PE-conjugated anti-human IgG for the others), and black thin lines indicate the results without sialidase treatment. Black bold lines indicate the results with *A. ureafaciens* sialidase treatment, and gray bold lines indicate results with *S. enterica* serovar Typhimurium sialidase treatment. Sialidase from *A. ureafaciens* releases α2-3,6,8-linked Sia, whereas sialidase from *S. enterica* serovar Typhimurium is specific to the α2-3 linkage. To confirm the effect of sialidase treatment, changes in cell surface expression of α2,3-linked Sia and α2,6-linked Sia were detected with mSn-Fc and hCD22-Fc chimeric probes precomplexed with R-PE-conjugated anti-human IgG, respectively. (C and D) Effect of free sugars on GL7 binding. Daudi cells were stained with FITC-conjugated GL7 in the presence of 50 mM free sugars (C) or the indicated concentrations of Neu5Ac (D). The data are shown as the relative MFI of each staining. Gal, galactose; Glc, glucose; Fuc, fucose; GlcNAc, *N*-acetylglucosamine; GlcA, glucuronic acid. (E) GL7 blotting of human B-cell lines. Membrane fractions of human B-cell lines (KMS-12 BM, KMS-12 PE, Daudi, and Ramos) were analyzed by GL7 immunoblotting. The addition of 100 mM Neu5Ac during incubation with GL7 reduced most of the staining on blotted membranes.

38). We were interested in the change of GL7 epitope expression, and thus we first assessed the reactivity of this antibody with various B-cell lines, including human germinal center-like Burkitt lymphomas. GL7 showed stronger reactivity toward human B-cell lines than toward mouse B-cell lines (Fig. 1A). The GL7 epitope has been shown to be sensitive to sialidase treatment, although the type of sialidase used in the study reporting this finding was not specified (19). To understand the relationship of GL7 epitopes present on human B-cell lines and mouse activated B cells, we further characterized the determinant on human B-cell lines. The GL7 epitope on Daudi cells was similar to that on mouse activated B cells, as GL7 staining of Daudi cells was also inhibited by sialidase treatment when a broad-range sialidase, *A. ureafaciens* sialidase, was used (Fig. 1B). In contrast, *S. enterica* serovar Typhimurium sialidase, which is specific to α2,3-linked Sia, had no effect (Fig. 1B). To assess the role of Sia and other sugars in GL7 reactivity, we analyzed the inhibitory effects of sugar on GL7 binding. The results clearly showed specificity of Neu5Ac for inhibition (Fig. 1C), and the inhibition was dependent on the Neu5Ac concentration (Fig. 1D). Neu5Ac is a major form of Sia in human cells. GL7 binding was decreased with a metabolic N-glycosylation inhibitor, tunicamycin (see Fig. S1 in the supplemental material). Multiple bands were detected in immunoblotting experiments using the membrane fraction of Daudi cells (Fig. 1E). Thus, it is likely that GL7 recognizes some glycan epitopes, including Sia, rather than some specific protein(s).

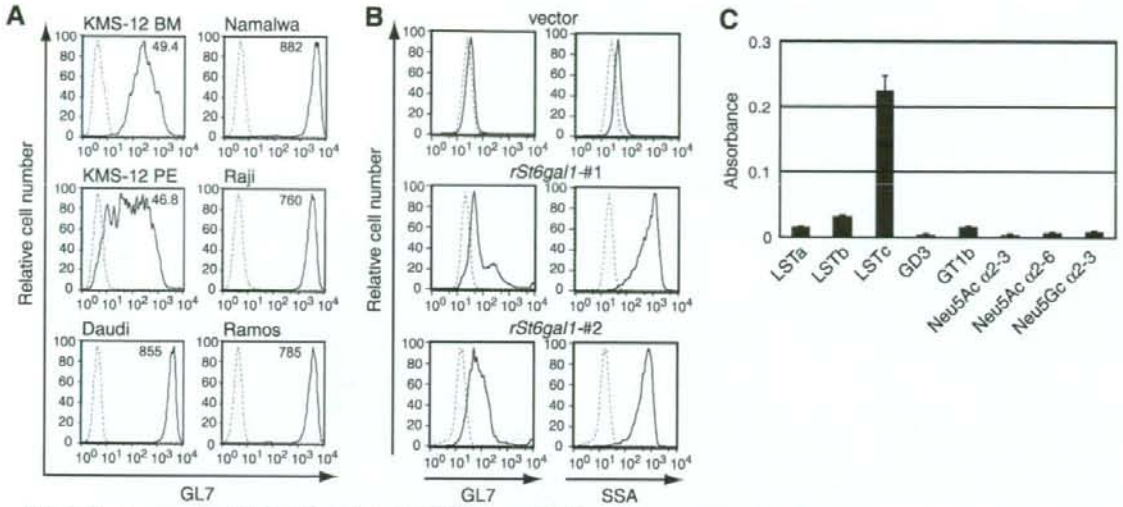

FIG. 2. Involvement of α2,6-linked Neu5Ac in the GL7 epitope. (A) Numerical comparison of GL7 staining among human B-cell lines. The results of GL7 staining of human B-cell lines were numerically compared using MFI values in flow cytometry. To normalize the binding in different cells, the endogenous fluorescence of sample cells (gray dashed lines) was adjusted to an MFI of around 5. For comparison with the gene expression profile, GL7-stained MFI values were divided by the background value. The relative values indicated on the top of each staining were used as the GL7 determinant expression profile. (B) Appearance of the GL7 determinant by ST6GAL1 expression. CHO-K1 clones stably transfected with rat *St6gal1* or an empty vector (as a control) were stained with FITC-conjugated GL7 or FITC-conjugated SSA. The results from two such clones are shown. (C) Carbohydrate binding assay of GL7. Carbohydrate binding was measured using ELISA. Data are shown as the means of triplicate samples, and the bars represent standard errors of the mean. LSTa, Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc; LSTb, Galβ1-3(Neu5Acα2-6)GlcNAcβ1-3Galβ1-4Glc; LSTc, Neu5Acα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc; GD3, Neu5Acα2-8Neu5Acα2-3Galβ1-4Glc; GT1b, Neu5Acα2-3Galβ1-3GalNAcβ1-4(Neu5Acα2-8Neu5Acα2-3)Galβ1-4Glc; Neu5Ac α2-3, Neu5Acα2-3Galβ1-4Glc; Neu5Ac α2-6, Neu5Acα2-6Galβ1-4Glc; Neu5Gc α2-3, Neu5Gcα2-3Galβ1-4Glc.

Strong correlation between expression of the GL7 epitope and expression of the *ST6GAL1* gene in human B-cell lines. Sia clearly plays an important role in GL7 epitope expression. Interestingly, the GL7 staining of a panel of human B-cell lines was not uniform but, instead, exhibited different intensities (Fig. 1A). Given that a number of bands were detected in immunoblotting experiments, the differences in GL7 epitope expression seemed to be caused by differences in the expression level of an enzyme(s) involved in the biosynthesis of the GL7 epitope glycan rather than differences in carrier protein expression. Therefore, we analyzed the correlation of GL7 epitope expression with the relative level of Sia-related gene expression. The reason to expect such a correlation was that glycosyltransferase activity tends to be regulated through the control of gene expression and substrate accessibility rather than through posttranslational modifications. Six human B-cell lines were stained with GL7 (Fig. 2A), and the relative MFI from flow cytometry was compared with the gene expression profile of the same set of B-cell lines obtained from a newly developed cDNA microarray that can be used to analyze the expression of glycan-related genes. To perform cross-sample comparisons of gene expression among cell lines, we compared poly(A)$^+$ RNA from each B-cell line and commercially available universal reference RNA. The relative gene expression was obtained by dividing the cDNA microarray fluorescence signal from cellular RNA by that of the universal reference (see Table S1 in the supplemental material). From among the genes spotted on the microarray, various genes for sialyltransferases and Sia-metabolizing enzymes were picked to examine their possible relationships to the degree of GL7 staining, because it has been shown that sialyltransferase gene expression might correlate with the surface phenotype of lectin binding (2). We calculated the Pearson's correlation coefficient. Among the sialyltransferase and other Sia-metabolizing enzyme genes, *ST6GAL1* showed the strongest correlation between its expression profile and the GL7 staining profile (Table 1). This result indicates that *ST6GAL1* expression could be responsible for the biosynthesis of the GL7 epitope in these human B-cell lines. ST6GAL1 transfers Sia onto a Gal residue of terminal *N*-acetyllactosamine (LacNAc; Gal β1-4GlcNAc) with an α2,6 linkage (42), and B cells have been shown to express this enzyme (20, 64). This indicates that the terminal transferase reaction by ST6GAL1, but not the supply of the substrate, is the rate-limiting step in GL7 epitope biosynthesis in these cells. Interestingly, a negative correlation was found between GL7 staining and the expression of *SIAE*, a gene encoding Sia 9-*O*-acetylesterase (Table 1). Although Sia 9-*O*-acetylesterase cleaves the *O*-acetyl group of Sia, *SIAE* is expressed in cell types expressing its substrate, 9-O-acetylated Sia (57). If the degree of 9-O acetylation were to correspond with the level of *SIAE* expression, GL7 binding might be negatively affected by 9-*O*-acetyl modification, similar to CD22 (56).

Effect of ST6GAL1 overexpression on GL7 epitope expression. Data from the correlation index calculation suggest that GL7 recognizes α2,6-linked Sia on N-glycan and that the expression of the GL7 epitope on human B cells depends mainly

TABLE 1. Pearson's correlation index analysis of Sia-related genes[a]

Index	*P* value	Gene name	Encoded enzyme
0.937	5.87E-3	*ST6GAL1*	ST6Gal I
0.806	5.30E-2	*ST3GAL3*	ST3Gal III
0.551	2.57E-1	*CMAH*	Pseudogene for CMP-Neu5Ac hydroxylase
0.473	3.44E-1	*ST3GAL2*	ST3Gal II
0.215	6.82E-1	*SLC35A1*	CMP-Sia transporter
0.173	7.43E-1	*ST8SIA1*	ST8Sia I
0.142	7.89E-1	*ST3GAL6*	ST3Gal VI
0.137	7.96E-1	*PGM3*	GlcNAc-6-P mutase
0.096	8.56E-1	*GMPPB*	GDP-Man pyrophosphorylase
0.052	9.22E-1	*ST6GALNAC2*	ST6GalNAc II
−0.103	8.47E-1	*ST8SIA3*	ST8Sia III
−0.196	7.10E-1	*ST8SIA4*	ST8Sia IV/PST
−0.210	6.89E-1	*GNE*	UDP-GlcNAc-2-epimerase/ ManNAc kinase
−0.283	5.87E-1	*ST6GALNAC6*	ST6GalNAc VI
−0.442	3.80E-1	*ST3GAL5*	ST3Gal V
−0.448	3.72E-1	*ST6GALNAC1*	ST6GalNAc I
−0.452	3.68E-1	*ST8SIA5*	ST8Sia V
−0.508	3.04E-1	*SAS*	Neu5Ac-9-P synthase
−0.639	1.72E-1	*ST6GALNAC4*	ST6GalNAc IV
−0.678	1.39E-1	*NEU3*	Membrane sialidase
−0.696	1.25E-1	*NEU1*	Lysosomal sialidase
−0.739	9.30E-2	*ST8SIA2*	ST8Sia II/STX
−0.742	9.12E-2	*ST3GAL4*	ST3Gal IV
−0.898	1.52E-2	*ST3GAL1*	ST3Gal I
−0.938	5.62E-3	*SIAE*	Sia-9-*O*-acetylesterase

[a] Pearson's correlation coefficient (index) values of relative gene expression in the microarray against relative GL7 staining MFI among six B-cell lines were calculated for sialyltransferase genes and Sia metabolism-related genes. A positive value indicates the presence of a correlation between gene expression and staining. A negative value indicates the presence of a negative correlation. Index values are also expressed as *P* values.

on *ST6GAL1* expression. To evaluate these findings, we explored the ST6GAL1 expression dependence of GL7 epitope expression. CHO-K1 cells are known to lack α2,6-linked Sia on their cell surfaces. As expected, the parental CHO-K1 cells were GL7 negative (data not shown), as were vector-transfected CHO-K1 cells (Fig. 2B). In contrast, rat *ST6GAL1* (r*St6gal1*)-transfected CHO-K1 cells showed a marked increase in GL7 staining (Fig. 2B). The increase in GL7 staining upon r*St6gal1* expression coincided with the increase in staining by SSA, a plant lectin which reacts with Sia α2,6-Gal/GalNAc on glycans. As CHO-K1 cells are nonimmune cells, GL7 seemed to recognize α2,6-linked Neu5Ac-containing sugar chains on various proteins. Immunoblotting analysis of these stable clones further clarified that the introduction of r*St6gal1* was sufficient to give rise to bands on the blot. The membrane fractions of both CHO-K1 stable clones and human B-cell lines resulted in multiple bands (data not shown).

Glycan-binding assay of GL7. To confirm that GL7 is an antiglycan antibody that recognizes α2,6-linked Sia and also to determine the fine specificity of the epitope, we examined GL7 binding to various glycan probes (65) by ELISA. GL7 bound to LSTc (Neu5Acα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc) but not to its structural isomer with α2–3 linked Neu5Ac, LSTa (Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc) (Fig. 2C). Interestingly, GL7 did not bind to Neu5Acα2-6Galβ1-4Glc (sialyllactose) in spite of the existence of α2,6-linked Neu5Ac in the probe. The glucose (Glc) of the reducing terminal was destroyed during probe preparation for coupling with streptavidin. Thus, it is likely that the structure of Neu5Acα2-6Gal is not sufficient for GL7 binding but that the binding requires at least a trisaccharide for optimal recognition or GlcNAc in the underlying lactosamine. Taking all of the results into consideration, we concluded that GL7 recognizes α2,6-linked Sia-containing glycan chains that are often found on N-glycans of various proteins.

A shift in the major Sia species, Neu5Gc to Neu5Ac, in the mouse germinal center reaction. It was still not clear why GL7 failed to react with mouse mature B cells, given that these cells abundantly express α2,6-linked sialoglycans, as St6gal1 is also expressed in these cells (20, 64). The dominant difference in sialylation between mice and humans occurs in the Sia modification at the C-5 position (60). Humans predominantly express Neu5Ac, whereas the major Sia in mice is Neu5Gc (Fig. 3A). It is possible that the change in GL7 reactivity could be a consequence of the change in sia modification. Neu5Gc modification in biosynthesis is regulated by the Cmah reaction in the cytosol, which metabolically gives rise to the donor, CMP-Neu5Gc, for a subsequent sialyltransferase reaction(s) in the Golgi apparatus (Fig. 3B) (24, 25). We therefore asked whether mouse B cells undergo a change in Sia species, from Neu5Gc to Neu5Ac, in GL7-positive cells. We first stained the germinal centers with GL7 and the lectin domain of mouse CD22 (mCD22-Fc), because mouse CD22 demonstrates a marked preference for Neu5Gc-bearing over Neu5Ac-bearing α2,6-linked sialoglycan ligands (26, 44, 50). As shown in Fig. 3C, in the SRBC-immunized mouse spleen, GL7-positive germinal centers were specifically excluded by mCD22-Fc recognition. This complementarity of staining appeared to be the result of the probe preferences, Neu5Ac for GL7 and Neu5Gc for mCD22-Fc, respectively. We then assessed Cmah expression and the Neu5Ac-Neu5Gc ratio in GL7-positive germinal center B cells. Germinal center (GL7-bound) cells showed severely reduced expression of Cmah, and this reduction coincided with the loss of Neu5Gc in the membrane fraction of the cells (Fig. 3D). In contrast, GL7-negative SRBC-immunized B cells were not significantly different from nonimmunized splenic B cells. Thus, the gain of GL7 staining reflected the loss of the CD22 ligand in germinal center B cells due to the repression of Cmah.

Real-time PCR analysis during mouse B cell activation. LPS stimulation induces the GL7 epitope in B cells (28). Therefore, we adopted this system to assess the enzyme (gene) responsible for GL7 epitope expression. Cmah is responsible for Sia species change, and St6Gal1 is responsible for Sia linkage biosynthesis. We examined the expression of *Cmah* and *St6gal1* to determine whether changes in the expression of these genes could account for the GL7 epitope induction detected in B-cell activation events. In real-time RT-PCR experiments, *Cmah* expression showed an 80% reduction in LPS-stimulated B cells compared with unstimulated splenic B cells after 48 h of incubation (Fig. 4A). This reduction was already detectable after 3 h of culture. Despite the slightly enhanced expression level of α2,6-linked Sia-containing glycan probed with SSA, *St6gal1* expression showed a subtle reduction in activated B cells after 48 h (Fig. 4A and B). *Cmah* reduction appears to play a prominent role in the appearance of the GL7 epitope in activated B cells. Retrovirus-mediated ectopic *Cmah* expression consistently reduced the expression of the GL7 epitope in

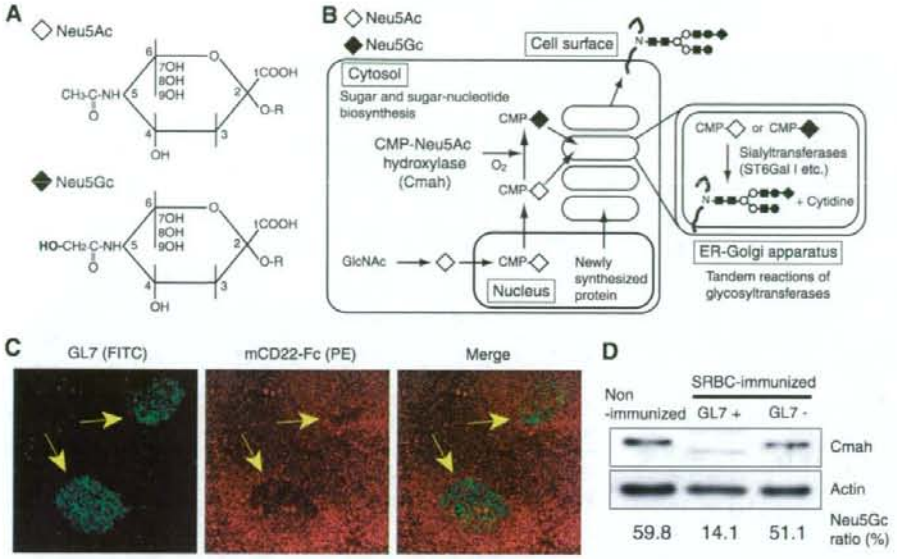

FIG. 3. Change in Sia species in germinal centers. (A) Structural differences between two major molecular species of Sia. The metabolic precursor Neu5Ac and its modified form Neu5Gc differ only by an oxygen atom at the C-5 position. The conversion of CMP-Neu5Ac to CMP-Neu5Gc is catalyzed by the enzyme 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

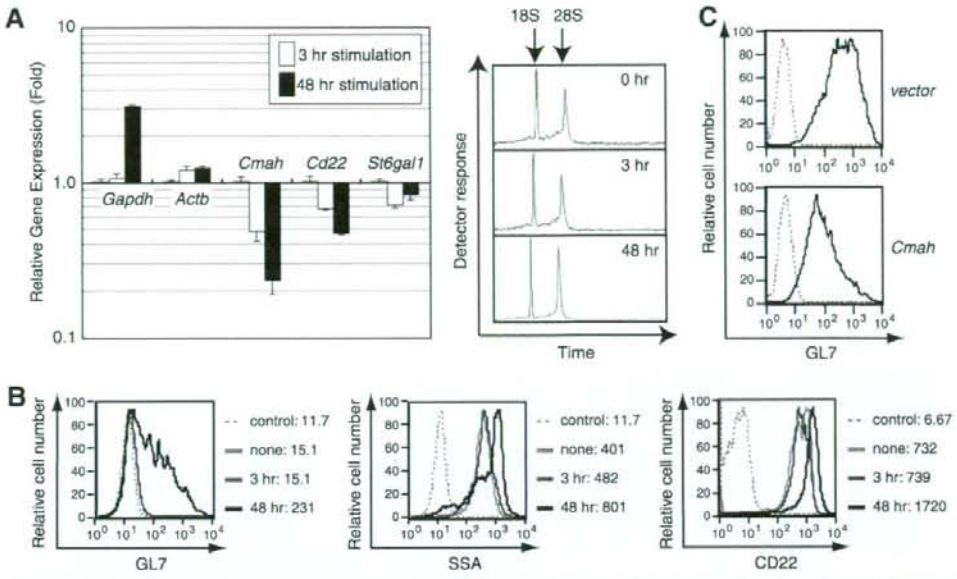

FIG. 4. Downregulation of *Cmah* mRNA in primary cultured B cell blasts, causing GL7 epitope expression. (A and B) Cmah repression caused by in vitro B-cell activation. Splenic B cells were stimulated with 30 μg/ml LPS for the indicated times. Reverse-transcribed cDNAs prepared from total RNA of these 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

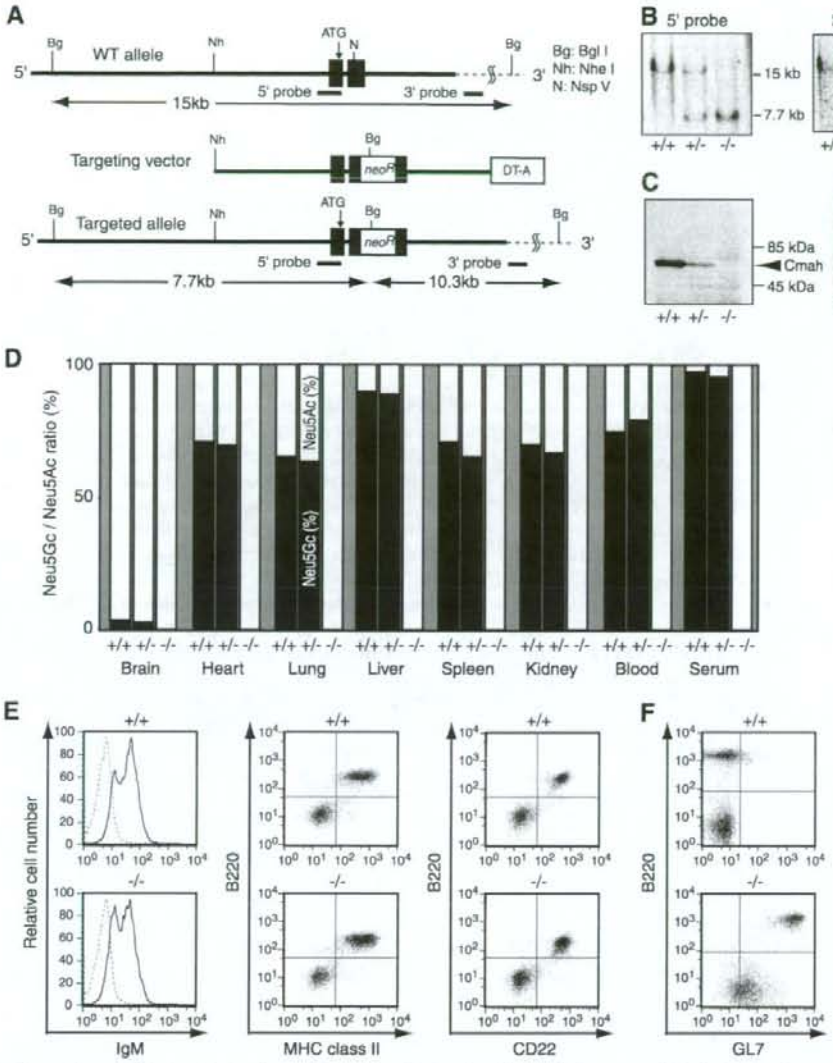

FIG. 5. Generation and biochemical analyses of *Cmah* knockout mice. (A) Allele for targeted *Cmah*. A targeting vector was created by inserting the *PGK-neoR* cassette into the NspV site of the second coding exon (exon 5) of the *Cmah* gene. (B) Genotype of homologous recombination of selected ES cell 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

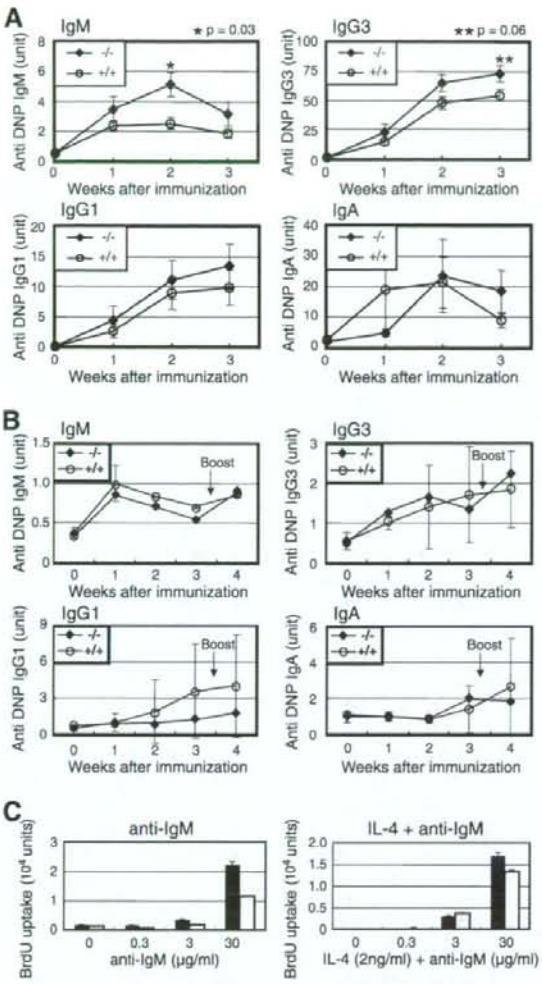

FIG. 6. Hyperresponsive phenotypes of *Cmah*-null mice. (A) T-independent hyperresponse of *Cmah*-null mice. DNP-Ficoll was used to immunize 8-week-old mice. Serum was collected each week and analyzed for reactivity with DNP-conjugated BSA coated on ELISA plates. The titer of hapten-reacting mouse Igs 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 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.

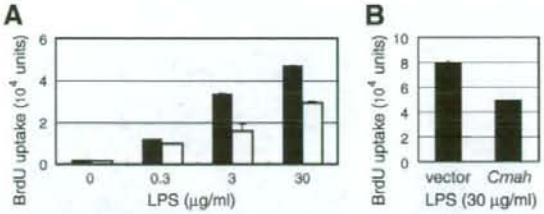

FIG. 7. Rescue of augmented proliferation of *Cmah*-null B cells by Cmah expression. (A) In vitro hyperproliferation response of *Cmah*-null B cells to LPS. Splenic B cells from wild-type (open columns) and *Cmah*-null mice (filled columns) were assessed for proliferation using LPS from *S. enterica* 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

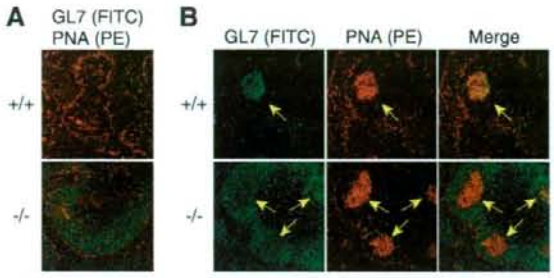

FIG. 8. Changes in staining of germinal center markers in normal or SRBC-immunized *Cmah*-null mice. (A) Histochemical analyses of spleen sections without immunization. Spleen sections from wild-type and *Cmah*-null mice were costained with FITC-conjugated GL7 and biotin-conjugated PNA visualized by 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

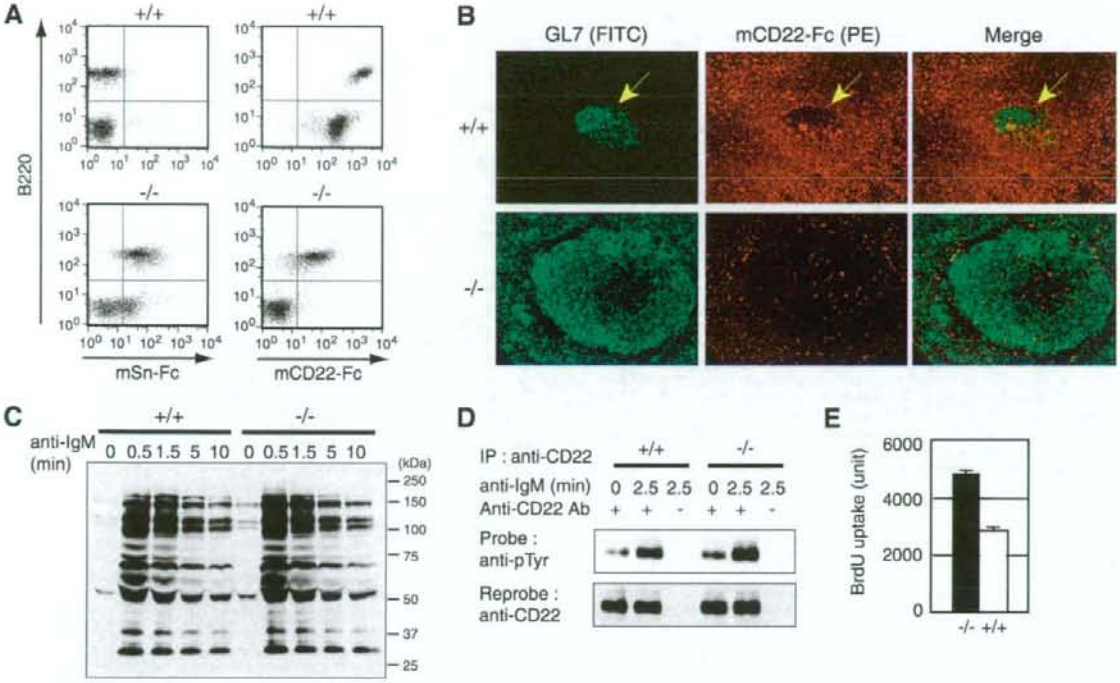

FIG. 9. Loss of optimal CD22 ligand and normal immediate response upon BCR cross-linking in *Cmah*-null mice. (A) Loss of optimal ligand for CD22 in *Cmah*-null mice. The expression of surface ligands for sialoadhesin and CD22 was detected by flow cytometry. Splenocytes from wild-type and *Cmah*-null 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. Hashimoto.** 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:

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

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

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