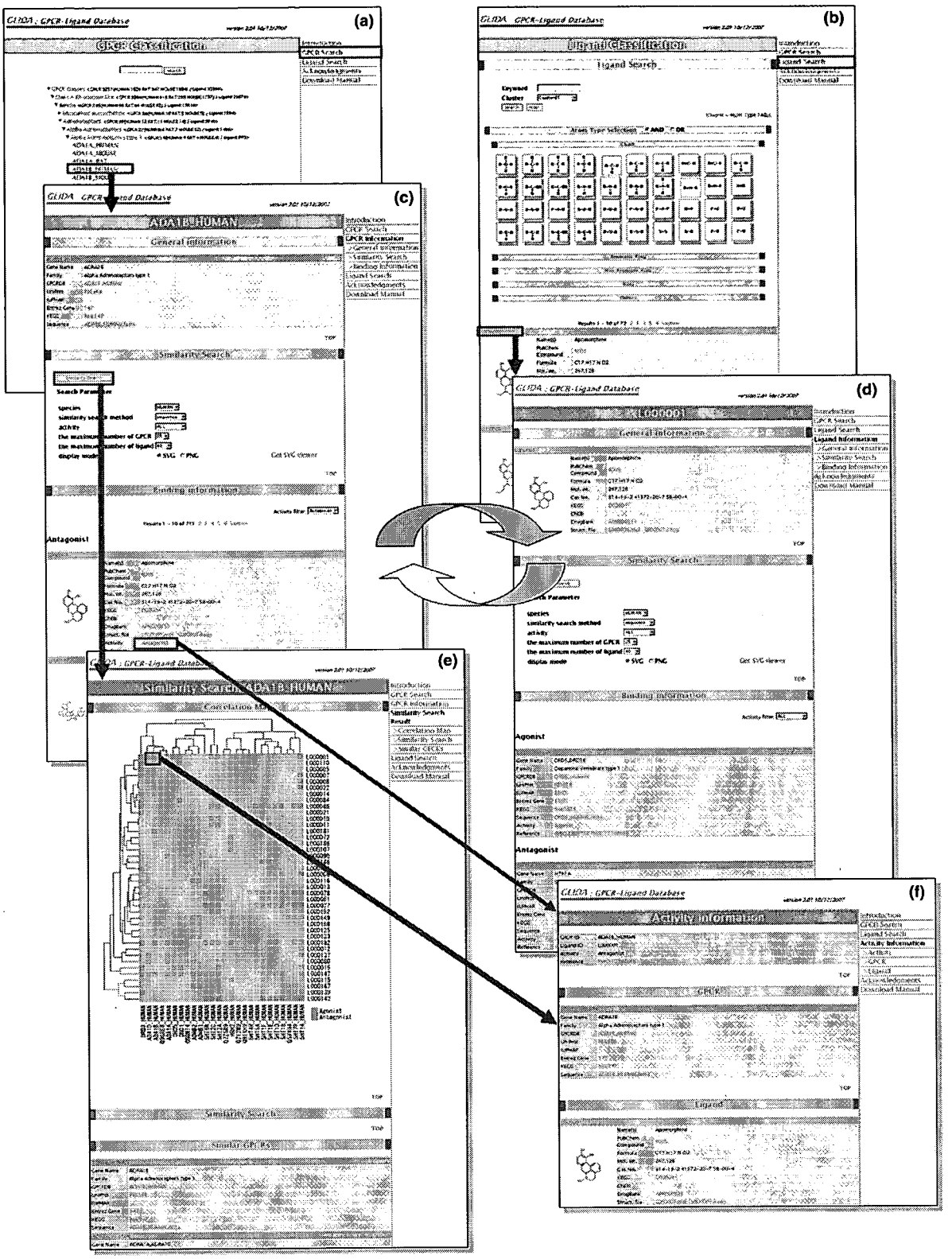


Figure 1. A screenshot of GLIDA showing linked relations among search pages (**a** and **b**), result pages (**c** and **d**), an analytical report page (**e**), and a binding information page (**f**). The analytical report page consists of a correlation map and a list resulting from a similarity search. Red and blue colors of the spots on the correlation map indicate the ligand activities of antagonists including inverse agonist and agonists including full/partial agonist, respectively.

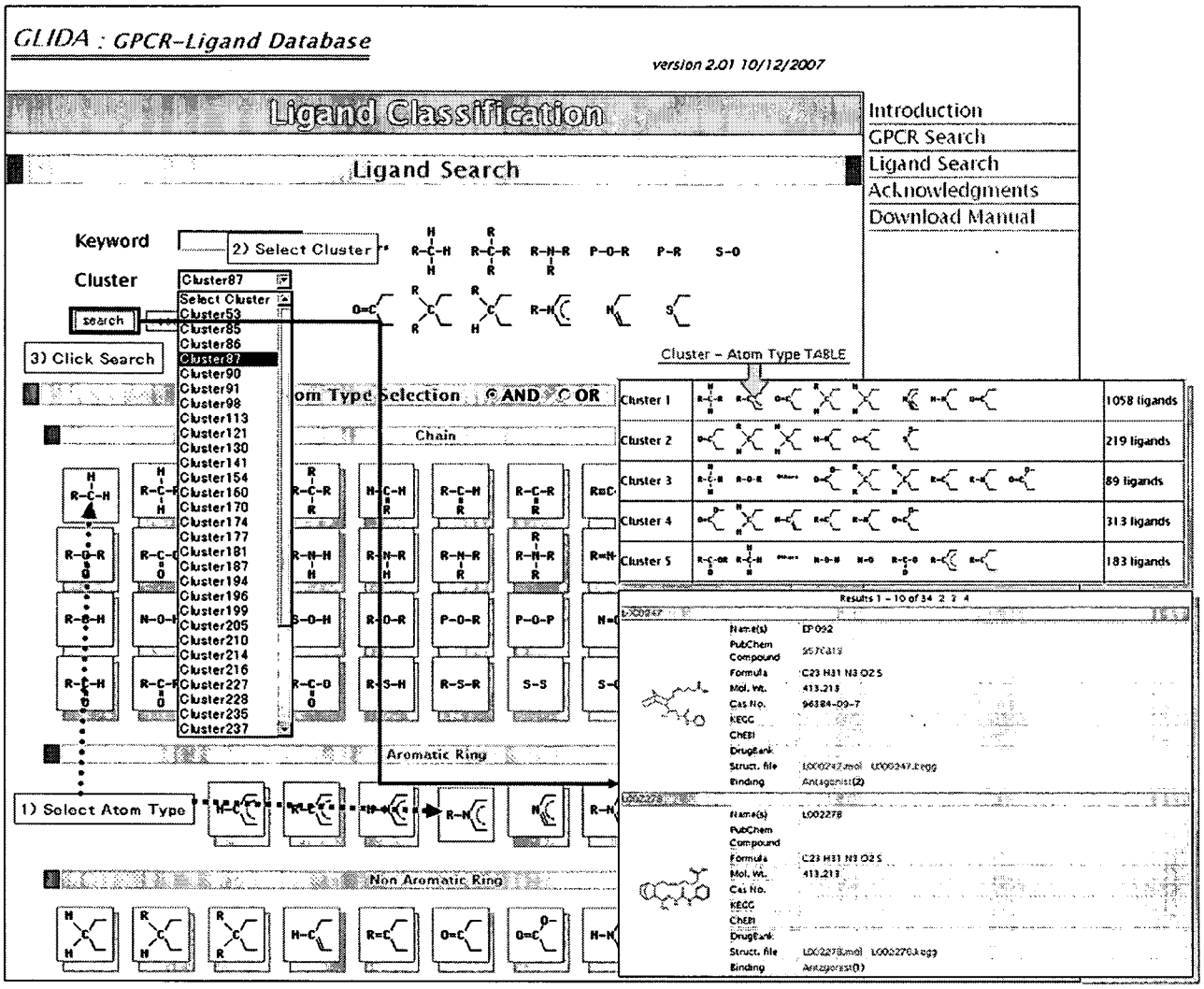


Figure 2. A screenshot of the ligand search process on the ligand classification page. Users can search the ligands from two starting points: keyword search and cluster selection. If they have a chemical structure of their query compound, the ligand search is performed using the cluster selection tool as follows. Selecting a set of atom types (step 1) that the query compound contains, the pull-down menu of cluster selection displays the list of the only clusters that include selected atom types as the principal components (step 2). By selecting a cluster from the list, users can check the principal component's atoms on the upper right section of the page. Finally, upon clicking the search button, GLIDA displays the list of all ligands classified in the selected cluster (step 3). The 'Atom Type TABLE' button links the user to the page showing the cluster size and representative atom types for each cluster.

or automatically by using the keyword search function. Every GPCR (or ligand) has its own results page (Figure 1c or d) containing a general information table regarding a GPCR (or ligand), a table of its correlated ligands (or GPCRs) and a menu button to carry out a similarity search and correlation analysis.

Classification of GPCRs and ligands

The GPCR classification table on the search page was adapted from the phylogenetic tree within the GPCRDB information system (http://www.gpcr.org/7tm/phylo/phylo.html). The GPCR classification table displays the entries of the corresponding GPCRs at the tree branches, and these are hyperlinked with the corresponding result pages (Figure 1a). GLIDA also provides an original ligand classification (Figures 1b and 2). With the great increase in ligand entries, we have to improve our method of classifying all the ligands in GLIDA. Hierarchical clustering and its tree representation, which were used in the old version of GLIDA, are unsuitable for the data mining of huge chemical databases. We therefore have adopted principal component analysis (PCA) for clustering of 23 214 ligand structures in this new version, as follows. We generated frequency profiles of the atoms and the bonds converted into the KEGG atom types from MDL MOL files of ligand entries (19). The KEGG-type profile for each ligand is shown in 'Struct. file' item of

general information table of GLIDA. PCA was applied to the data matrix consisting of 700 KEGG-type features' columns and 23 214 ligand entries' rows. The resulting principal components (PCs) constitute a new set of linearly independent, orthogonal axes that capture the directions of maximum variance in the data. The samples (chemical compounds) were then projected onto these PC axes. Herein, we used the top 314 PCs as seeds of clusters that account for >80% (cumulative proportion) of the total variance. Finally, each compound was assigned to the PC cluster having the maximum score among the 314 PCs. In order to annotate the features of each cluster (PC), we selected for each PC the atom types and their bonds corresponding to the top 10 loadings having the largest magnitude. The ligand classification page displays a table of all the atom types selected by PCA (Figure 2). By clicking on some of the atoms in this table, users can search clusters that include the selected atom types. Consequently, the ligands relevant to users' interests are included in the retrieved cluster.

Similarity search and correlation map for GPCRs and ligands

The fact that similar proteins bind similar ligands is the underlying principle of the Chemical Genomics approach to drug discovery (11). GLIDA has a variety of similarity search functions for GPCRs and ligands, respectively, on its result pages (Figure 1c or d). Alignment scores for protein sequences generated by the BLAST algorithm provide similarity measures for GPCRs. In addition to sequence similarity, gene expression patterns in tissue origins and developmental stages were used as similarity measures. The expression data for each GPCR was generated from the EST sequences in different libraries served from NCBI/Unigene (http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi). We can thereby retrieve the GPCRs that present tissue-/stage-specific distribution similar to a query GPCR. For example, co-expression information on specific GPCRs enables us to speculate about GPCR-heterodimerization that might have an effect on their activity (1). Ligand similarity is defined by the dissimilarity (distance) of frequency profile patterns generated from the constitutive atoms and bonds of the chemical structure, using the KEGG atom types (19,20). From the similarity search, the most similar GPCRs (or ligands) within the users' selected parameters are retrieved and listed with their similarity scores on an analytical report page (Figure 1e). In the latest GLIDA version, various parameters have been added as search options, such as selections of species, ligand activities, displayed number of GPCRs/ligands and map graphical mode. As another result of similarity search calculations, GLIDA illustrates the correlation map (Figure 1e) showing homologous GPCRs (or ligands) and their ligands (or GPCRs) that are retrieved. This map shows spots that match the GPCRs and their ligands in a 2D matrix. The ordering along the x-axis and the y-axis are calculated respectively by two-way clustering of the GPCRs and the ligands, based on their similarities. In particular, the ordering along the x- and y-axes allows users to evaluate the sequence similarities among GPCRs and the correlation coefficients among ligands simultaneously. By analyzing the correlation patterns between GPCRs and ligands that are illustrated by these maps, we can gain detailed knowledge about their interactions. We can then utilize this information to infer possible candidates for development of GPCR-specific drugs. Furthermore, we have enhanced a graphical interface to display the correlation map between GPCRs and ligands. Graphics are an important tool to aid visualization and interpretation of high-dimensional data. The old version of GLIDA used only the PNG (Portable Network Graphics) format to display a GPCR–ligand correlation map. Due to the great increase in entries, the latest GLIDA version introduces the SVG (Scalable Vector Graphics) format, which is adaptable to an enormous correlation map size. The SVG vector image can be scaled indefinitely without loss of image quality, while the PNG bitmap image cannot. Users must install the free plug-in software on their computer in advance to use the SVG format (http://www.adobe.com/svg/viewer/install/). In the case of uninstalled devices, PNG representation should be selected as a graphical mode. Figure 1 shows an example of the GPCR–ligand search and analysis process starting from a GPCR query using GLIDA.

DISCUSSION AND FUTURE DIRECTIONS

GLIDA provides a unique database useful for GPCR-related Chemical Genomics research and drug discovery. GLIDA is distinct from other public Chemical Genomics databases because it contains original, GPCR-specific chemical entries and offers a common mining platform of bioinformatics and chemoinformatics. GLIDA provides several advantages over other databases, in that a search can be started either from a GPCR or from a ligand. Thus, searches can be carried out in a dynamic and user-friendly way. GLIDA's coverage of chemical and biological information simultaneously also provides an advantage to users by saving them the time and labor required to search multiple databases. The ligand search page is another distinct characteristic of GLIDA, in that it displays the structural distribution of ligands. It thereby facilitates research on GPCR-related drugs by incorporating structural aspects of the ligand compounds into the search. The analytical report pages resulting from the calculated structural similarities of GPCRs and ligands can give the user deep insights into the GPCR–ligand relationships. The lists of neighboring ligands (or GPCRs) and the correlation maps are useful visualization tools for analyzing correlations among the structural features and the GPCR–ligand-binding properties. Because this database system can be applied to proteins other than the GPCR family, it may also be considered as a promising database for other types of Chemical Genomics research. One critical issue is how to define similarity metrics for proteins and ligands, because the underlying principle of GLIDA is that similar receptors bind similar ligands. For example, ligand similarity can be defined by manifold representations such as graph, fingerprint and descriptors.

Protein similarity can be also measured in different ways such as overall sequence homology (phylogenetic relationships), consensus motifs, common binding sites, 3D structures and reported functional annotations. Therefore we will add new menus incorporating these various similarity metrics for GPCRs and ligands. GLIDA will be updated continuously. In particular, we are now planning to add the drawing tool of chemical structures and to expand the ligand-searching function for an arbitrary chemical query.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, from the JSPS, KAKENHI, Grant-in-Aid for Publication of Scientific Research Results and from the Ministry of Health, Labour and Welfare of Japan. Financial support from the SUNTORY INSTITUTE FOR BIOORGANIC RESEARCH, the TATEISI SCIENCE AND TECHNOLOGY FOUNDATION and the Okawa Foundation for Information and Telecommunications is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. George,S.R., O'Dowd,B.F. and Lee,S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nature Rev. Drug Discov.*, **1**, 808–820.
2. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
3. Strachan,R., Ferrara,G. and Roth,BL. (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov. Today*, **11**, 708–716.
4. Foord,S.M., Bonner,T.I., Neubig,R.R., Rosser,E.M., Pin,J.P., Davenport,A.P., Spedding,M. and Harmar,A.J. (2005) International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol. Rev.*, **57**, 279–288.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, 12.
6. Schreiber,SL. (2004) Stuart Schreiber: biology from a chemist's perspective. Interview by Joanna Owens. *Drug Discov. Today*, **9**, 299–303.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, D402–D404.
8. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
9. Zerhouni,E. (2003) The NIH Roadmap. *Science*, **302**, 63–72.
10. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
11. Klabunde,T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.
12. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, **34**, D673–D677.
13. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
14. Roth,B.L., Lopez,E., Beischel,S., Westkaemper,R.B. and Evans,J.M. (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **102**, 99–110.
15. Civelli,O. (2005) GPCR deorphanizations: the novel, the known and the unexpected transmitters. *Trends Pharmacol. Sci.*, **26**, 15–19.
16. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **35**, D193–D197.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
18. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
19. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
20. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Chem-Bio Informatics Journal, Vol. 7, No. 2, pp.24-34(2007)

Compound-Transporter Interaction Studies using Canonical Correlation Analysis

Masato Kitajima[1,2], Yohsuke Minowa[3], Hideo Matsuda[2], Yasushi Okuno[4*]

[1]*Current Address:Life Science Systems Dept. ,PLM Solutions Div. ,Fujitsu Kyushu System Engineering Limited, 2-2-1,Momochihama,Sawara-ku,Fukuoka,814-8589,Japan*
[2]*Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan,*
[3]*National Institute of Biomedical Innovation Toxicogenomics Informatics Project 7-6-8 Asagi Saito Ibaraki-City Osaka, 567-0085,Japan*
[4]*Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

**E-mail: okuno@pharm.kyoto-u.ac.jp*

(Received November 3, 2007; accepted November 15, 2007; published online December 10, 2007)

Abstract

The efficient screening of lead compounds or drug candidates for efficacy and safety is critically important during the early stage of drug development. Compounds are usually screened from a diverse 'chemical space' based only on its pharmacological effects, but this screening is not enough to guarantee drug safety. To solve this problem, we devised a chemical space that takes into account interaction information with proteins such as drug transporters. We also created and evaluated a mathematical model for predicting compound-transporter interactions. This was achieved by first generating an interaction correlation matrix based on drug transporters and their corresponding inhibitor compounds. To implement a screening scheme that takes into account interaction with drug transporters, we created a model using Canonical Correlation Analysis (CCA) that makes use of the known information on interaction between drug transporters and their corresponding inhibitors. Cross-validation of the model gave satisfactory test results and a physiologically relevant chemical space was constructed based on the model.

Key Words: Pharmacokinetics, Transporter, chemoinformatics, bioinformatics

Area of Interest: Molecular Recognition

Chem-Bio Informatics Journal, Vol. 7, No. 2, pp.24-34(2007)

1. Introduction

During the drug development process it is important to screen compounds for efficacy and safety at an early stage in order to prevent unnecessary and costly analysis later on. It is thought that the diverse chemical space may contain as much as 10^{60} chemical structures or more. Searching for a drug candidate with a good balance of efficacy and safety from this huge chemical space is obviously very difficult, as evidenced from the fall in the number of new drug applications in recent years [1][2].

The previously used screening approach involved sampling a diverse chemical library for those leads that display only promising pharmacological effects i.e. drug efficacy. It is important to select a compound with excellent pharmacokinetic properties (drug safety), not just pharmacological effects, when screening at the early stage of drug development. When analyzing the pharmacokinetic properties of drug candidates, ligand interactions with Phase I enzymes such as cytochromes P450, Phase II conjugation enzymes (e.g. GST and sulfotransferases), as well as transporter proteins that play a crucial role in Phase III, must be considered [3]. Of these, transporter proteins, which are important in facilitating absorption of compounds in the intestines as well as the degree of penetration across the blood-brain barrier, play a central role in determining the bioavailability of drugs.

This paper focuses on transporter proteins that play an important role in drug pharmacokinetics. The interaction studies were carried out based on information on the interactions of the compounds and the transporter proteins. To implement a screening scheme that takes into account interaction with drug transporters, we created a model using CCA that exploits the characterized interactions between drug transporters and their corresponding inhibitors. The model is evaluated and then used to create a physiologically relevant chemical space.

2. Method

2.1 Gathering of compound-transporter interaction data

The compound-transporter interaction data used in this study was extracted from the ADME Database (developed by Fujitsu Kyushu Systems Engineering Ltd., Fukuoka, Japan) [4][5][6]. The database is a collection of information on drug transporters as well as drug metabolizing enzymes found in the literature. Two kinds of transporter proteins were selected for this study; the ABC transporter family and the SLC transporter family. Compounds that interact with these transporter families were also extracted from the database. The compound-transporter interaction type available in the database includes substrates, inhibitors, inducers and activators. However, for the purpose of this study only the inhibitors were selected.

A total of 17 ABC transporter families and 110 different SLC transporter proteins were selected for this study. Data concerning the interaction of these transporter proteins with known compounds was extracted from the ADME Database. The database contains 5,860 compound-transporter interactions between the selected 117 transporter proteins and their interacting 3,275 compounds.

The selected transporter proteins are as follows.

Table 1. List of transporter proteins used in the study

ABCA1	SLC1A1	SLC5A6	SLC7A7	SLC19A1	SLC23A1
ABCA2	SLC1A2	SLC5A7	SLC7A8	SLC19A2	SLC23A2
ABCA9	SLC1A3	SLC5A8	SLC10A1	SLC21A11	SLC26A2
ABCA10	SLC1A4	SLC5A9	SLC10A2	SLC21A12	SLC26A3
ABCB1	SLC1A5	SLC6A1	SLC10A4	SLC21A14	SLC26A4
ABCB4	SLC1A6	SLC6A11	SLC13A1	SLC21A2	SLC26A6
ABCB5	SLC1A7	SLC6A12	SLC13A2	SLC21A20	SLC26A7
ABCB11	SLC2A1	SLC6A13	SLC13A3	SLC21A3	SLC26A8
ABCC1	SLC2A10	SLC6A14	SLC13A4	SLC21A6	SLC26A9
ABCC2	SLC2A11	SLC6A2	SLC13A5	SLC21A8	SLC27A4
ABCC3	SLC2A12	SLC6A3	SLC15A1	SLC21A9	SLC28A1
ABCC4	SLC2A13	SLC6A4	SLC15A2	SLC22A1	SLC28A2
ABCC5	SLC2A2	SLC6A5	SLC15A4	SLC22A11	SLC28A3
ABCC10	SLC2A3	SLC6A6	SLC16A1	SLC22A12	SLC29A1
ABCC11	SLC2A4	SLC6A9	SLC16A10	SLC22A16	SLC29A2
ABCG1	SLC2A6	SLC7A1	SLC16A3	SLC22A2	SLC29A4
ABCG2	SLC2A7	SLC7A10	SLC16A5	SLC22A3	SLC32A1
	SLC2A8	SLC7A11	SLC16A7	SLC22A4	SLC36A1
	SLC4A4	SLC7A2	SLC17A1	SLC22A5	SLC38A1
	SLC5A1	SLC7A3	SLC18A1	SLC22A6	SLC38A4
	SLC5A2	SLC7A5	SLC18A2	SLC22A7	SLC38A5
	SLC5A4	SLC7A6	SLC18A3	SLC22A8	SLC43A2

2.2 Organizing the collected data

A correlation matrix of compound-transporter interactions was constructed based on the collected compound-transporter interaction information. Here compounds that interact with a transporter protein are flagged '1', and those that do not interact are flagged '0' as shown in Table2.

The similarity between transporter proteins was calculated based on this interaction correlation matrix. The Tanimoto coefficient found below was used to evaluate similarity [7].

$$\text{Tanimoto}(X,Y) = \frac{C}{A + B - C}$$

A: No. of bits in X that were flagged '1'
B: No. of bits in Y that were flagged '1'
C: No. of flagged bits common to both X and Y

The interaction similarity of the compounds was also defined as Tanimoto coefficient based on the correlation matrix. We used the distance measure transformed from the Tanimoto coefficient as shown below:

$$\text{Distance}(X,Y) = 1 - \text{Tanimoto}(X,Y)$$

Table 2. Excerpt from the correlation matrix of compound-transporter interactions.

Compound Name \ Transporter Name	ABCA1	ABCA10	ABCA2	ABCA9	ABCB1	ABCB11	ABCB4	ABCB5	ABCC1	ABCC10	ABCC11	ABCC2	ABCC3	ABCC4	ABCC5	ABCG1	ABCG2
Verapamil	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0
Cholyltaurine, Taurocholic acid, Taurocholate	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1
4,4'-Diisothiocyanostilbene-2,2'-disulfonic acid, DIDS	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0
Phloretin	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
Bromosulfophthalein, Sulfobromophthalein, BSP	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
Cyclosporin A, Cyclosporine, Cyclosporin, Ciclosporin	0	0	0	0	1	1	1	0	1	1	0	1	1	1	0	1	1
Probenecid	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0
Indomethacin	0	0	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0
Progesterone	0	0	0	0	1	1	0	0	1	0	0	1	0	1	1	0	1
Quinidine	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
Cimetidine	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Phloridzin, Phlorizin	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Pravastatin, Pravastatin acid	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1
Rifampicin, Rifampin	1	0	0	0	1	1	0	0	1	0	0	1	1	0	0	0	1
1-Methyl-4-phenylpyridinium, MPP(+)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Estradiol 17beta-D-glucuronide, E(2)17betaG	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0
L-glutamine, Gln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L-leucine, L-Leu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methotrexate	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	0	1
MK571, MK-571	0	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	1
Chlorpromazine	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Desimipramine, Desipramine	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Diclofenac	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
Ketoprofen	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0

2.3 Canonical Correlation Analysis (CCA)

Next, CCA was performed using the collected compound-transporter interaction data. Dragon X was used to calculate the compound descriptors [8]. A total of 929 descriptors were calculated. Highly correlated descriptors were grouped together and then filtered to give a final total of 324 compound descriptors.

The transporter proteins were calculated as bigram (two amino acids) frequency in protein sequences, and were used to generate a total of 400 protein descriptors. CCA was then performed, which generated 324 components. CCA is a technique to extract common features from a pair of multivariate data (chemical and protein descriptors). CCA finds a linear transformation of the chemical and protein spaces such that the correlation coefficient is maximized. Therefore we can construct the chemical space with the higher correlation to the protein space by extraction of the some components with the higher correlation coefficients.

2. 4 Cross Validation

A 5-fold cross-validation test was performed using only the 44 CCA components with P value < 0.01 of correlation coefficient test. The whole training compound set was divided into five sets. The first set was left out for testing and the remaining four sets were used to train a model. The compounds from the first set were then used to evaluate the trained model. The model was evaluated by setting a Euclid distance threshold from a test compound and using the closest neighboring compounds within this threshold for prediction. The procedure was repeated for all five sets, each time leaving out one set for testing, until all compounds from all five sets had been evaluated.

3. Result

3.1 Analysis of compound-transporter interactions

A significant number of compounds were found to inhibit more than one transporter using the collected compound-transporter interaction data. Indeed, out of these compounds, 183 were identified as inhibiting five or more transporters. The list below shows the frequency for each "interaction count" of a compound.

Table 3. Frequency of Interaction count for a single compound

Interaction Count	Frequency
18	1
17	1
16	2
15	3
14	1
13	2
12	4
11	6
10	8
9	15
8	19
7	18
6	37
5	66
4	118
3	355
2	393
1	2226

Moreover, the similarity of each transporter protein was calculated from the interaction matrix profile by using the Tanimoto coefficient. The result of clustering based on this similarity measure is shown in Figure 1. As shown in the figure, ABCG1 and ABCB4 are similar by interaction pattern even though the sequence similarity between both proteins is very low. Analogous results were found for ABCG2, which was shown to be similar to ABCC1 and ABCC2.

Moreover, Cyclosporin A that interacts with 15 kinds of transporter proteins, and MK571 that interacts with 11 kinds of transporter proteins were examined and compared as shown in Figure 2 [9][10]. Even though the two compounds share a low degree of structural similarity, they were found to interact with the same 10 transporter proteins. Our results demonstrate that it is not possible to explain all the similarities in interaction by simply comparing the structure of the relevant compounds or by protein sequence alignments alone. We then performed CCA on the collected interaction data to construct a correlation model for building a chemical space that reflected the classification of the transporters.

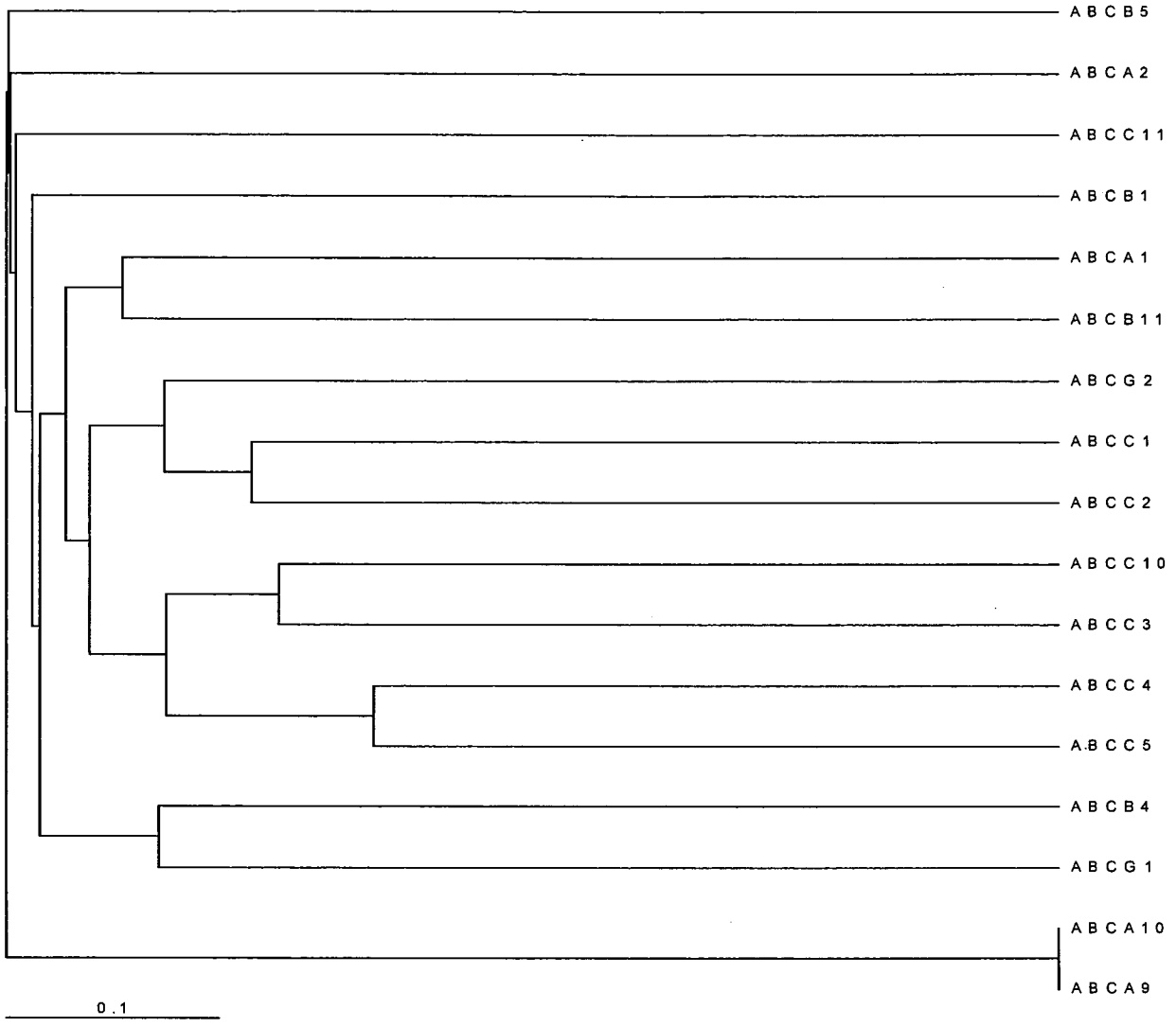
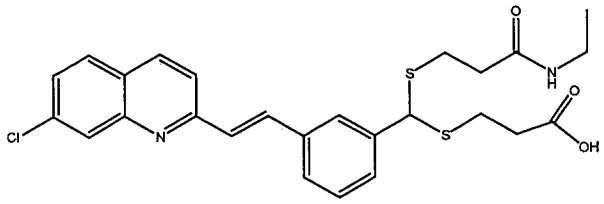
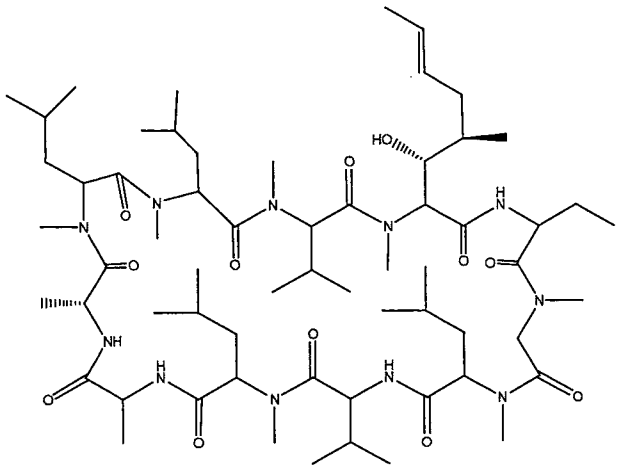


Figure 1. ABC transporters' cluster analysis based on similarity of interaction



MK571



Cyclosporin A

Figure2. Chemical structures of MK571 and CyclosporinA

3.2 CCA Result

The correlation model was constructed by using canonical correlation analysis (CCA). Performance of the model was evaluated using 5-fold cross-validation. CCA analysis and 5-fold cross-validation were performed using the collected compound-transporter interaction data. Below is the definition of the terms used in the evaluation of results.

Table 4. Definition of terms used in validation results
List of the total number of compounds (frequency) with corresponding numbers of transporter interactions (interaction count)

	Definition
TruePositive	The predicted transporter-interaction matches an observed transporter-interaction of the test compound
TrueNegative	The predicted NON- interaction matches an observed NON-interaction of the test compound
FalsePositive	The predicted transporter-interaction do not match any of the observed transporter-interaction of the test compound
FalseNegative	The predicted NON- interaction matches an observed transporter-interaction of the test compound

Sensitivity and Specificity are calculated as follows:

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

$$FalsePositiveRate = 1 - Specificity$$

To evaluate the performance of the model, the ROC plot (x-axis=FalsePositiveRate, y-axis=Sensitivity) is shown in Figure 3 below.

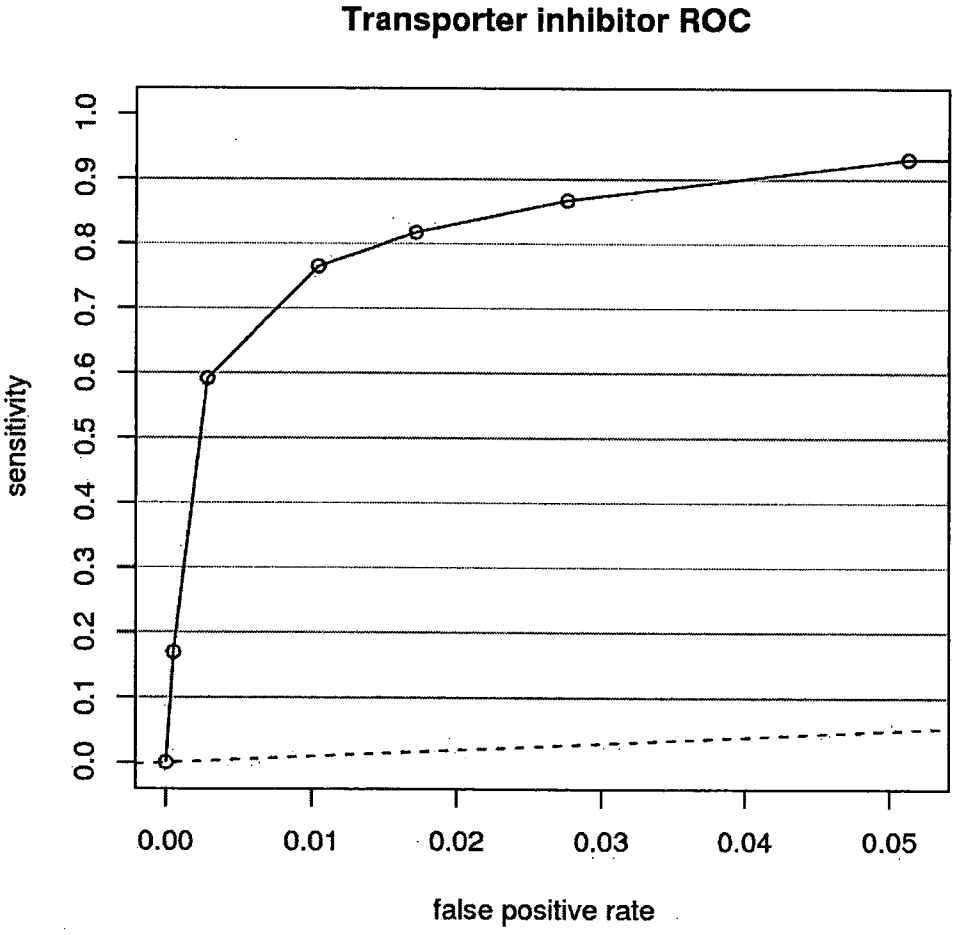


Figure 3. ROC of transporter inhibitor

ROC was plotted the average point of the sensitivity and false positive rate calculated under each condition of the maximum number of neighboring compounds (1,5,10,20) and the Euclid distance thresholds (1, 10, 20, 40, 80, 160 and 200)

This graph shows that the closer the curve inclines to the upper left corner the better is the model performance. The closer the curve declines to the dotted line the poorer is the performance, since the dotted line shows the performance curve of random models. The results show that our model has a very high performance, as evidenced by the curve's inclination to the upper left corner.

Then we identified 183 compounds that interact with more than 5 transporter proteins. A similarity map of the compounds was constructed as shown in Figure 4; where x-axis represents the similarity based on structural descriptors and y-axis represents similarity based on the interaction correlation matrix (Table 1). Our similarity map shows two distinct groups of compounds. Group B represents compounds with extremely low structural descriptor similarity, as exemplified by Cyclosporin A and Vinblastine. In this group, Cyclosporin A and MK571 show relatively high interaction similarity even though they have low structural descriptor similarity. For clarity, the similarity map of group A is enlarged by excluding group B as shown in Figure 5. The figure also shows that even though the compounds which interact with the subfamily of SLC22 and SLC28 have low structural similarity, they show high similarity with regards to interaction with SLC22 or SLC28.

Furthermore, the same similarity map is reconstructed by plotting in the x-axis the similarity of the compounds measured by CCA (instead of using the structural descriptors) as shown in Figure 6. The figure shows that compounds with low structural similarity may have high similarity by CCA, as shown by the Cyclosporin A and MK571 pair, as well as the SLC22 and the SLC28 interacting compounds. Thus, it was shown that constructing a chemical space using information on compound-transporter interaction is much better than simply using structural descriptors alone.

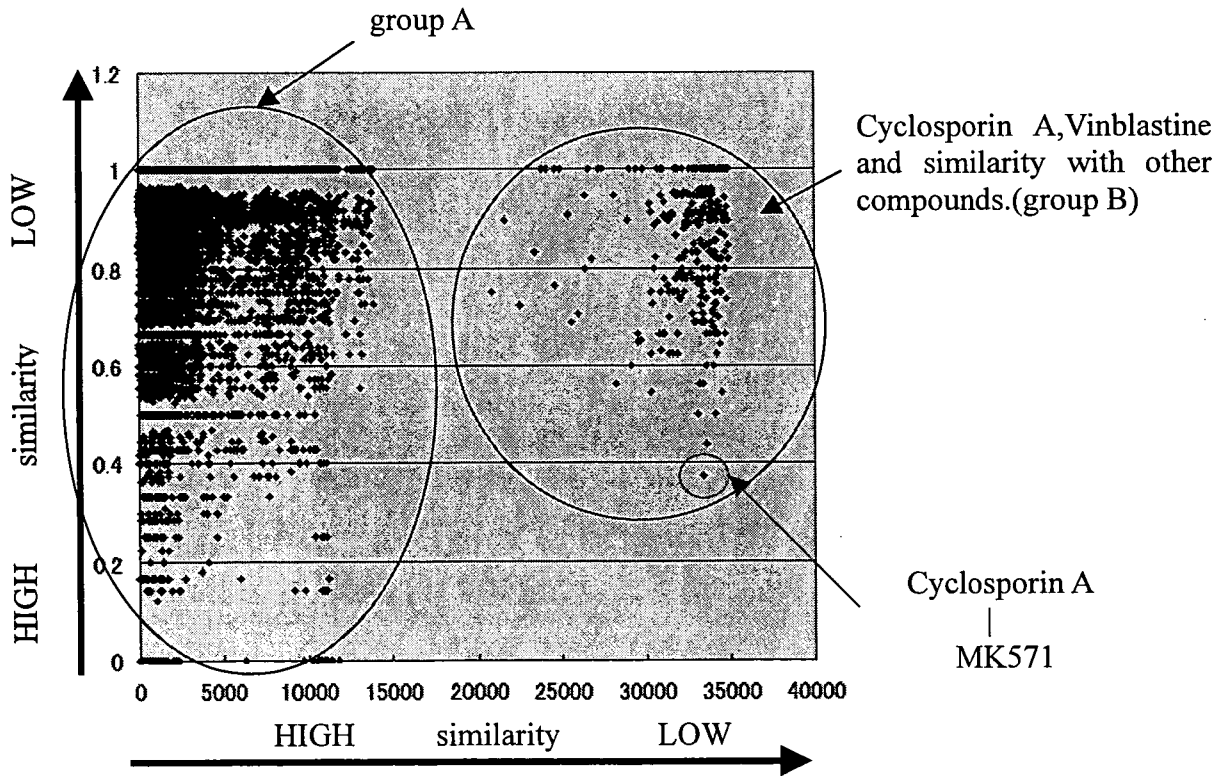


Figure 4. Two-dimensional map of structural similarity vs. interaction similarity
(Cyclosporin A and Vinblastine included in the map)
y-axis : Similarity in interaction pattern with a transporter protein
x-axis : Similarity in chemical structure

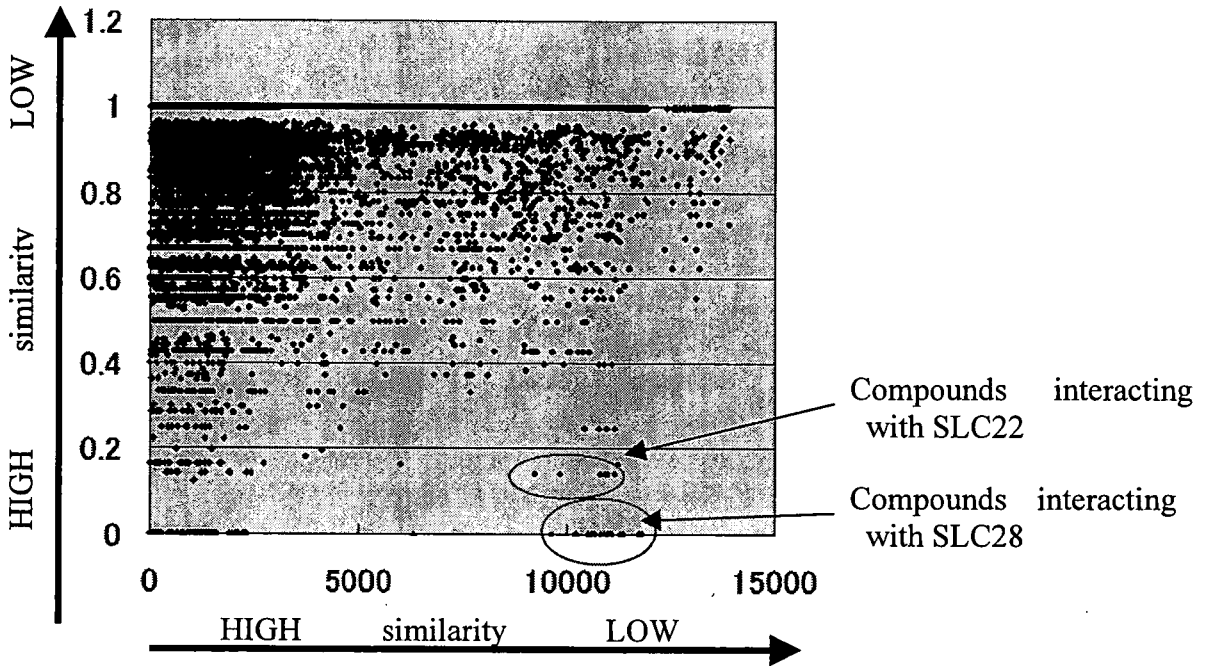


Figure 5. Two-dimensional map of structural similarity vs. interaction similarity
(Cyclosporin A and Vinblastine excluded from the map)
y-axis : Similarity in interaction pattern with a transporter protein
x-axis : Similarity in chemical structure

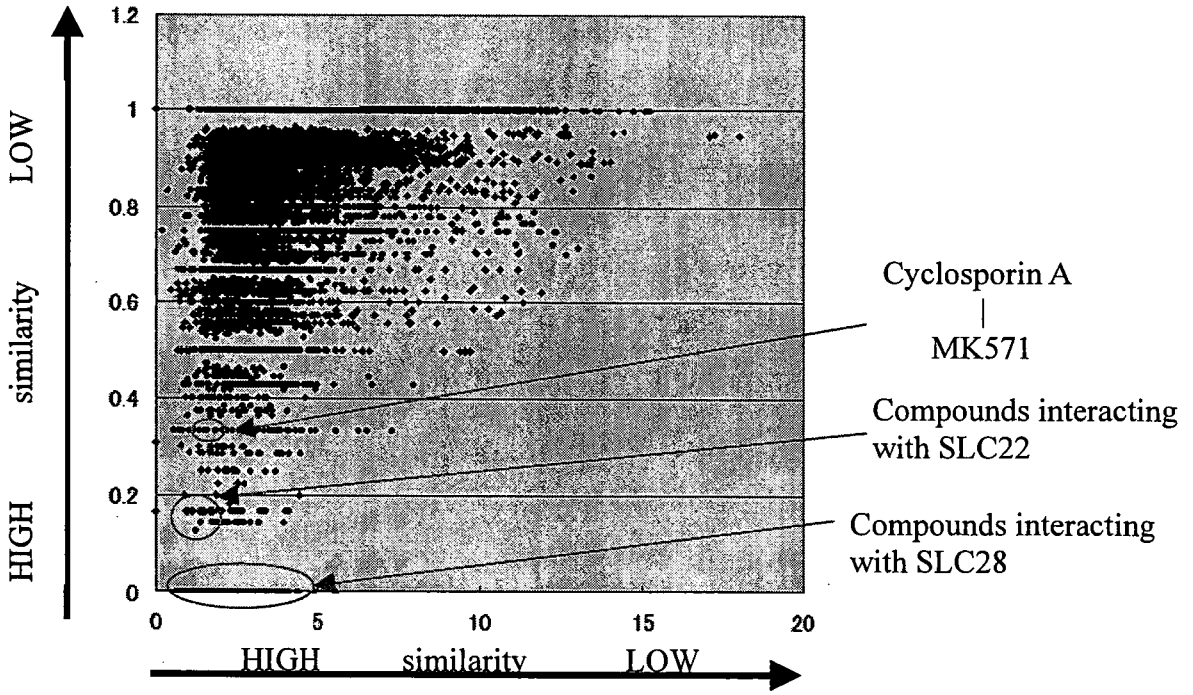


Figure 6. Two-dimensional map of similarity by CCA vs. interaction similarity
(Cyclosporin A and Vinblastine included in the training)
y-axis : Similarity in interaction with a transporter protein
x-axis : Similarity by CCA

5. Discussion

It was found that the results of classifying the transporter proteins by similarity of interaction pattern are different from the results obtained when classifying them by sequence similarity. Moreover, it was found that compounds showing similarity in interactions with more than one transporter protein may not necessarily have structural similarity at all.

These results show that it is difficult to predict the interaction between a compound and protein (i.e. related to pharmacological and pharmacokinetic effects) based on chemical structural similarity or protein sequence similarity alone. To create a physiologically relevant chemical space, information on compound–protein interactions is required. By utilizing CCA, we built an interaction model that was used to create a chemical space. Compounds that have high similarity in terms of interaction with proteins, but that are not necessarily similar in terms of structure, were clustered together. Thus a chemical space of transporter inhibitors was created, although this technique can also be applied to construct a chemical space (or a focused library) of compounds that interact with any specific target protein.

Moreover, a compound-transporter interaction model was constructed using CCA, which gave good evaluation results. This technique can be extended to develop chemical spaces of not only the inhibitors but also the substrates of transporter proteins. The resulting chemical spaces may be used for *in silico* screening of compounds with good pharmacological characteristics as well as good absorption, distribution and excretion properties.

The method described in this paper can also be extended to study toxicity related proteins. By building chemical spaces for such proteins, drug candidates with a good balance of efficacy and safety can be developed.

References

[1] Peter Kirkpatrick and Clare Ellis, Chemical space, *Nature*, **432**, 823 (2004).
[2] Christopher M. Dobson, Chemical space and biology, *Nature*, **432**, 824-828 (2004).
[3] Ishikawa T. The ATP-dependent glutathione S-conjugate export pump, *Trends Biochem. Sci.*, **17**, 463-468 (1992).
[4] http://jp.fujitsu.com/group/fqs/services/lifescience/asp/adme-database/index.html
[5] Rendic S., Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.*, **34**, 83-448 (2002).
[6] Rendic S, Di Carlo F.J., Human cytochrome P450 enzymes: a status report summarizing their reactions, substrates, inducers, and inhibitors. *Drug Metab. Rev.*, **29**, 413-580 (1997).
[7] Godden J.W., Xue L., Bajorath J., Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients, *Journal of Chemical Information and Computer Sciences*, **40**, 163-166 (2000).
[8] http://www.talete.mi.it/products/dragon_description.htm
[9] Byrne JA, Strautnieks SS, Mieli-Vergani G, Higgins CF, Linton KJ, Thompson RJ. The human bile salt export pump: characterization of substrate specificity and identification of inhibitors, *Gastroenterology*, **123**, 1649-1658 (2002).
[10] Leier I, Jedlitschky G, Buchholz U, Center M, Cole SP, Deeley RG, Keppler D. ATP-dependent glutathione disulphide transport mediated by the MRP gene-encoded conjugate export pump, *Biochem. J.*, **314**, 433-437 (1996).

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,

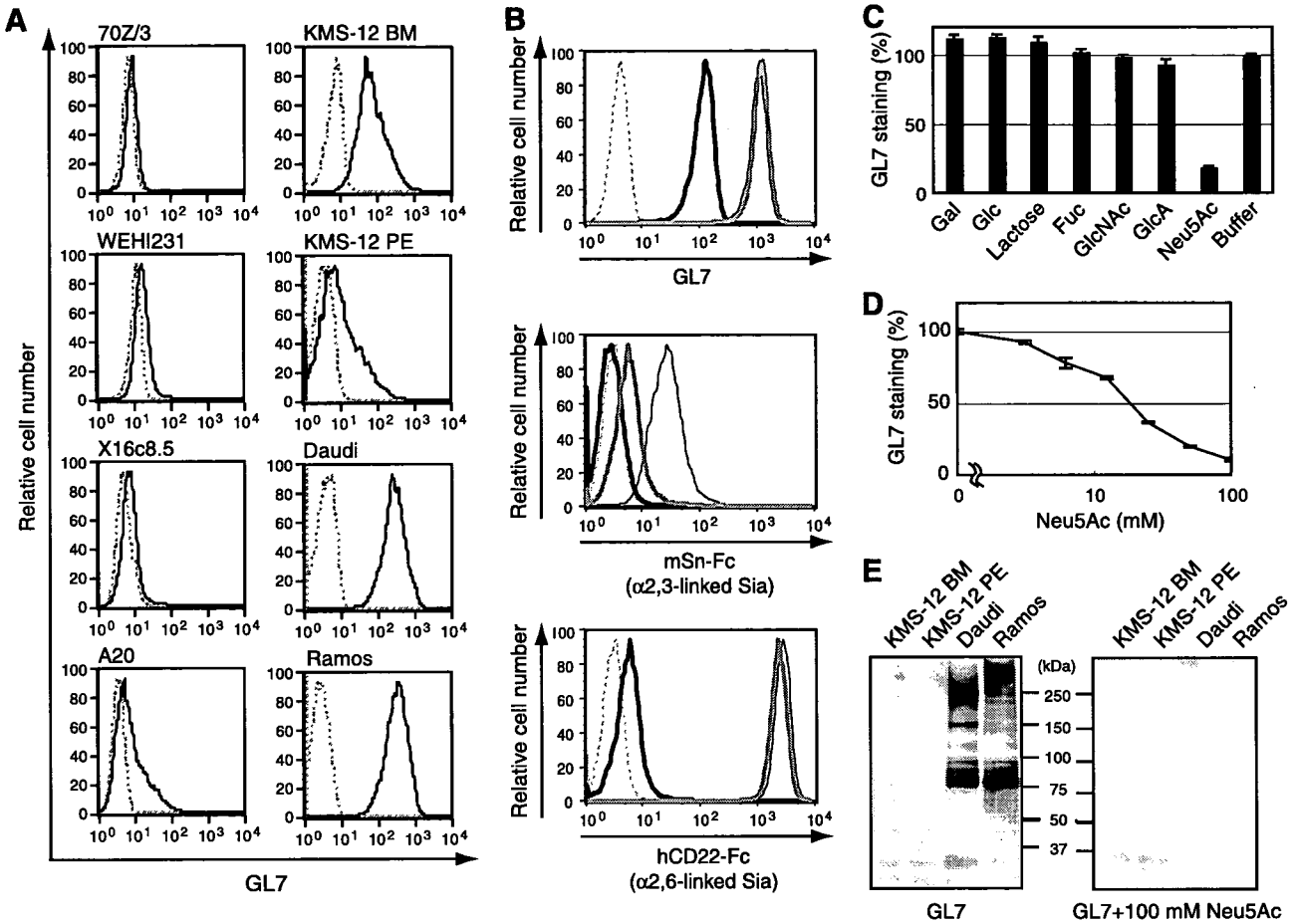


FIG. 1. Involvement of Sia in the GL7 epitope. (A) GL7 staining in flow cytometry. Mouse B-cell lines (70Z/3, WEHI231, X16c8.5, and A20) and human B-cell lines (KMS-12 BM, KMS-12 PE, Daudi, and Ramos) were stained with FITC-conjugated GL7. Black solid lines indicate staining with GL7, and gray dashed 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).

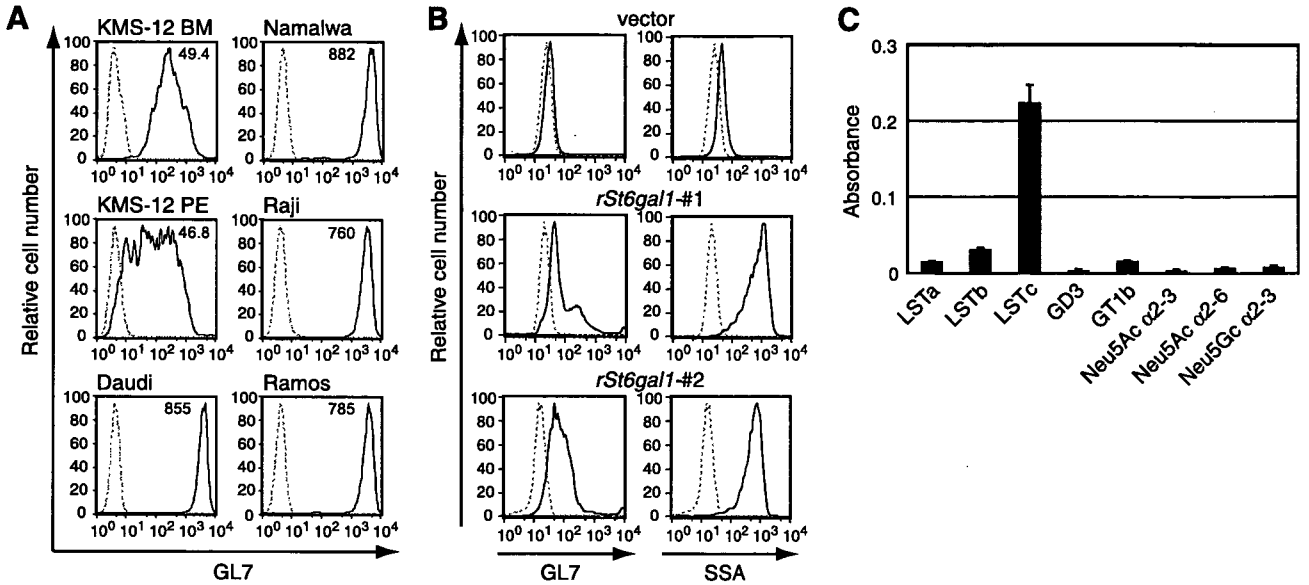


FIG. 2. Involvement of α2,6-linked Neu5Ac in the GL7 epitope. (A) Numerical comparison of GL7 staining among human B-cell lines. The results of GL7 staining of human B-cell lines were numerically compared using MFI values in flow cytometry. To normalize the binding in different cells, the endogenous 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