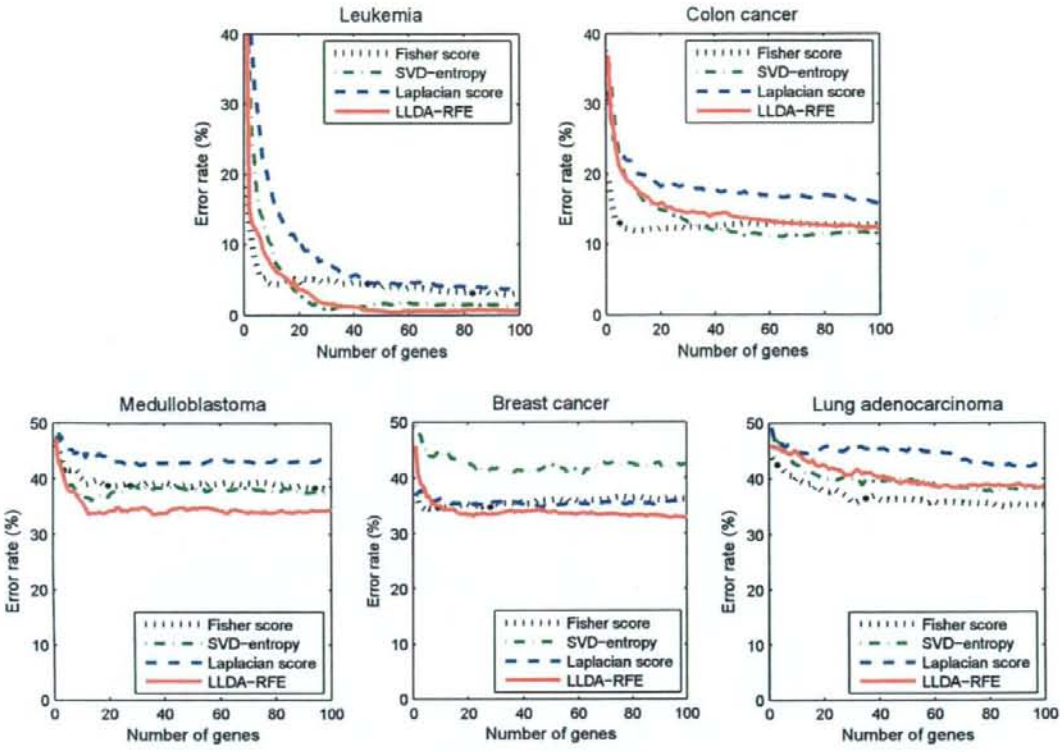


Fig. 3. Average error rates for Fisher score, SVD-entropy, Laplacian score and LLDA-RFE on binary-class datasets.

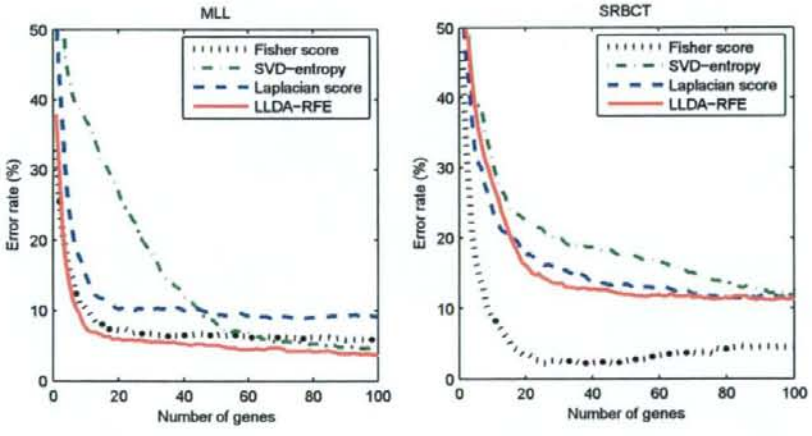


Fig. 4. Average error rates for Fisher score, SVD-entropy, Laplacian score and LLDA-RFE on multi-class datasets.

construct a graph such that samples with the same class are always connected, while those with different classes disconnected, and those with no class labels are adaptively connected or disconnected depending on the nearest neighbors.

APPENDIX
PROOF OF THEOREM 1.

It is straightforward to verify that $S_g - 2S_\ell$ can be decomposed as follows:

$$\begin{aligned} S_g - 2S_\ell &= \frac{1}{n} X(L_g - 2L_\ell)X^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell)(P\Lambda Q^T)^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell) Q\Lambda P^T \\ &= \frac{1}{n} PV\Delta V^T P^T \\ &= \frac{1}{n} (PV)\Delta (PV)^T. \end{aligned}$$

Since both P and Q are orthonormal,

$$(PV)^T(PV) = V^T P^T PV = V^T V = I.$$

Hence, the theorem holds. □

ACKNOWLEDGMENTS

This work was supported by grant from the 21st Century COE program "Knowledge Information Infrastructure for Genome Science". A part of this work was done while S. Niijima was in the Graduate School of Systems Life Sciences at Kyushu University, and also supported in part by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to Prof. S. Kuhara).

REFERENCES

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA*, vol. 96, pp. 6745-6750, 1999.

[2] O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling," *Proc. Nat'l Academy of Sciences USA*, vol. 97, pp. 10101-10106, 2000.

[3] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia," *Nature Genetics*, vol. 30, pp. 41-47, 2002.

[4] D.G. Beer, S.L.R. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M.G. Taylor, M.D. Iannettoni, M.B. Orringer, and S. Hanash, "Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816-824, 2002.

[5] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 12, pp. 1624-1637, 2005.

[6] F.R.K. Chung, *Spectral Graph Theory*, Regional Conference Series in Mathematics, no. 92, 1997.

[7] C.H.Q. Ding, "Unsupervised Feature Selection via Two-way Ordering in Gene Expression Analysis," *Bioinformatics*, vol. 19, no. 10, pp. 1259-1266, 2003.

[8] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Amer. Statist. Assoc.*, vol. 97, pp. 77-87, 2002.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Boston, MA: Academic Press, 1990.

[10] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[11] G.H. Golub and C.F. Van Loan, *Matrix Computations*, third ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.

[14] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P.O. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology*, vol. 1, no. 2, research0003, 2000.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," In Y. Weiss, B. Schölkopf and J. Platt (eds.), *Advances in Neural Information Processing Systems 18*, pp. 507-514, Cambridge, MA: MIT Press, 2006.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328-340, 2005.

[17] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[18] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 157-165, 2006.

[19] F. Li and Y. Yang, "Analysis of Recursive Gene Selection Approaches from Microarray Data," *Bioinformatics*, vol. 21, no. 19, pp. 3741-3747, 2005.

[20] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face Recognition Using Kernel Scatter-difference-based Discriminant Analysis," *IEEE Trans. Neural Networks*, vol. 17, no. 4, pp. 1081-1085, 2006.

[21] M. Loog, "On an Alternative Formulation of the Fisher Criterion That Overcomes the Small Sample Problem," *Pattern Recognition*, vol. 40, pp. 1753-1755, 2007.

[22] S. Michiels, S. Koscielny, and C. Hill, "Prediction of Cancer Outcome with Microarrays: a Multiple Random Validation Strategy," *Lancet*, vol. 365, pp. 488-492, 2005.

[23] S. Niijima and S. Kuhara, "Recursive Gene Selection Based on Maximum Margin Criterion: a Comparison with SVM-RFE," *BMC Bioinformatics*, vol. 7, 543, 2006.

[24] A.Y. Ng, M.I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," In T. Dietterich, S. Becker and Z. Ghahramani(eds.), *Advances in Neural Information Processing Systems 14*, pp. 849-856, Cambridge, MA: MIT Press, 2002.

[25] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, "Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression," *Nature*, vol. 415, pp. 436-442, 2002.

[26] H. Tang, T. Fang, and P.-F. Shi, "Laplacian Linear Discriminant Analysis," *Pattern Recognition*, vol. 39, pp. 136-139, 2006.

[27] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[28] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *Bioinformatics*, vol. 22, no. 14, pp. e507-e513, 2006.

[29] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veenman, H. Dai, Y.D. He, L.J. van't Veer, "A Protocol for Building and Evaluating Predictors of Disease State Based on Microarray Data," *Bioinformatics*, vol. 21, no. 19, pp. 3755-3762, 2005.

[30] L. Wolf and A. Shashua, "Feature Selection for Unsupervised and Supervised Inference: the Emergence of Sparsity in a Weight-based Approach," *J. Machine Learning Research*, vol. 6, pp. 1855-1887, 2005.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: a General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.

[32] J. Ye, "Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems," *J. Machine Learning Research*, vol. 6, pp. 483-502, 2005.
[33] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, no. 4, pp. 181-190, 2004.
[34] J. Zhu and T. Hastie, "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics*, vol. 5, no. 3, pp. 427-443, 2004.

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 5, 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

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.

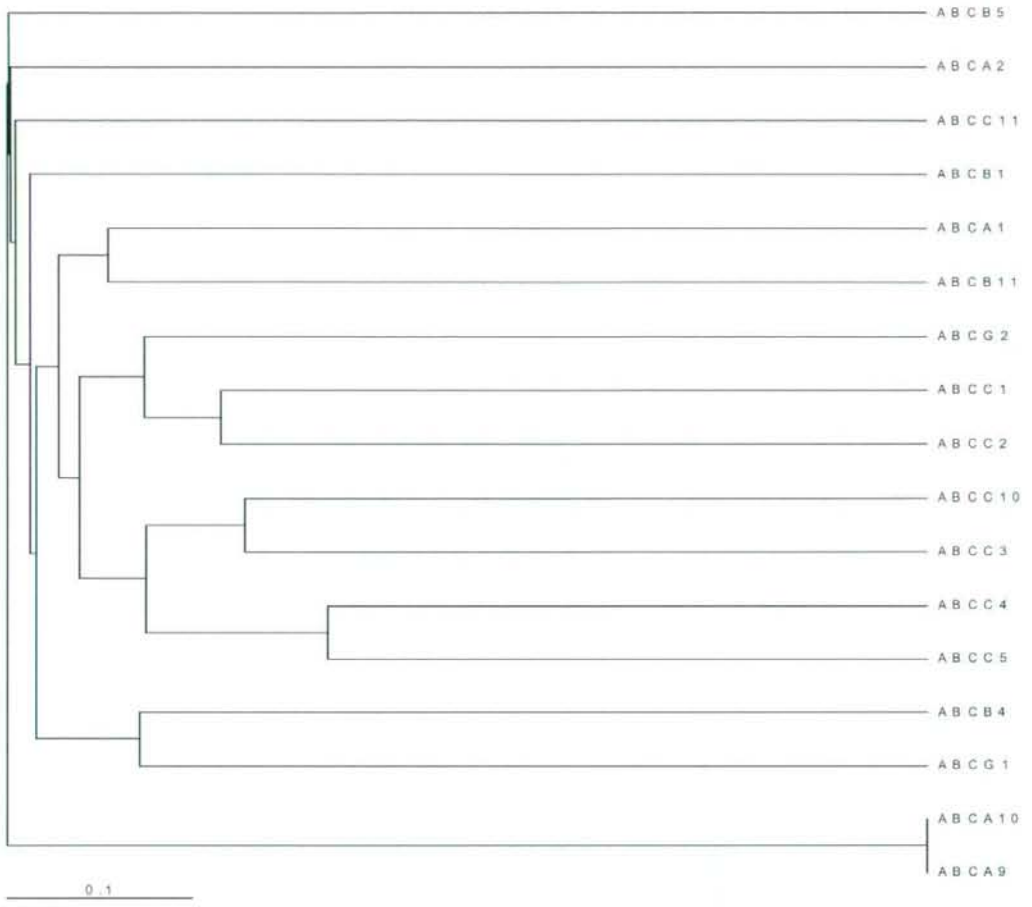
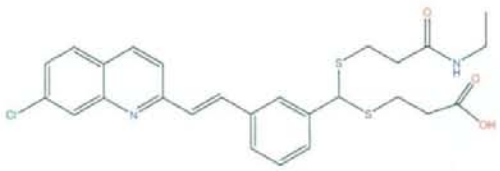
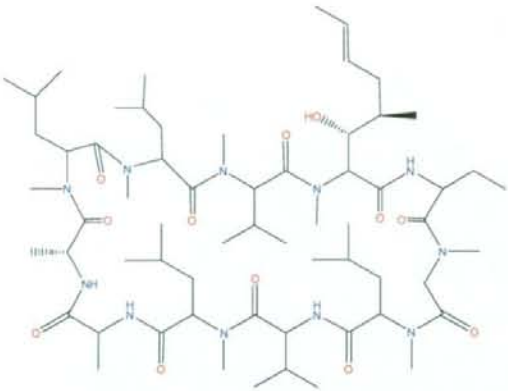


Figure 1. ABC transporters' cluster analysis based on similarity of interaction



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.

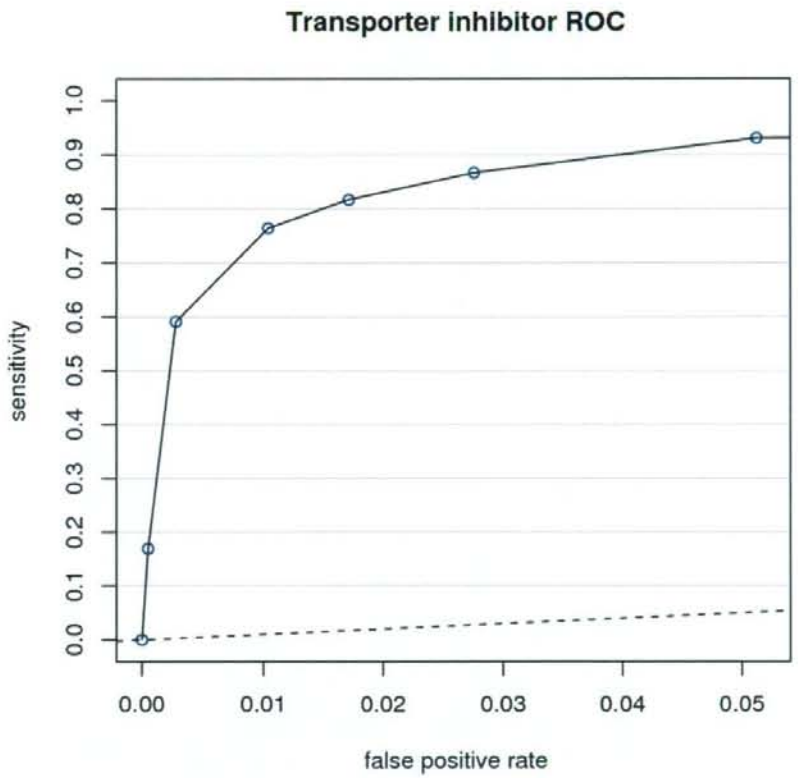


Figure 3. ROC of transporter inhibitor

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.

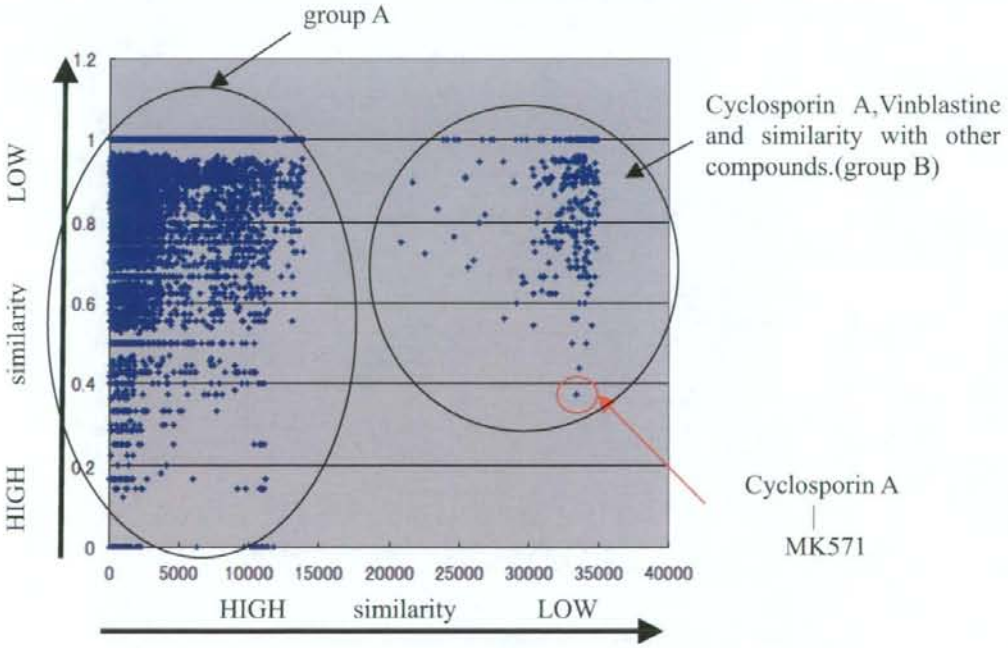


Figure 4. Two-dimensional map of structural similarity vs. interaction similarity
(Cyclosporin A and Vinblastine included in the map)
y-axis : Similarity in interaction pattern with a transporter protein
x-axis : Similarity in chemical structure

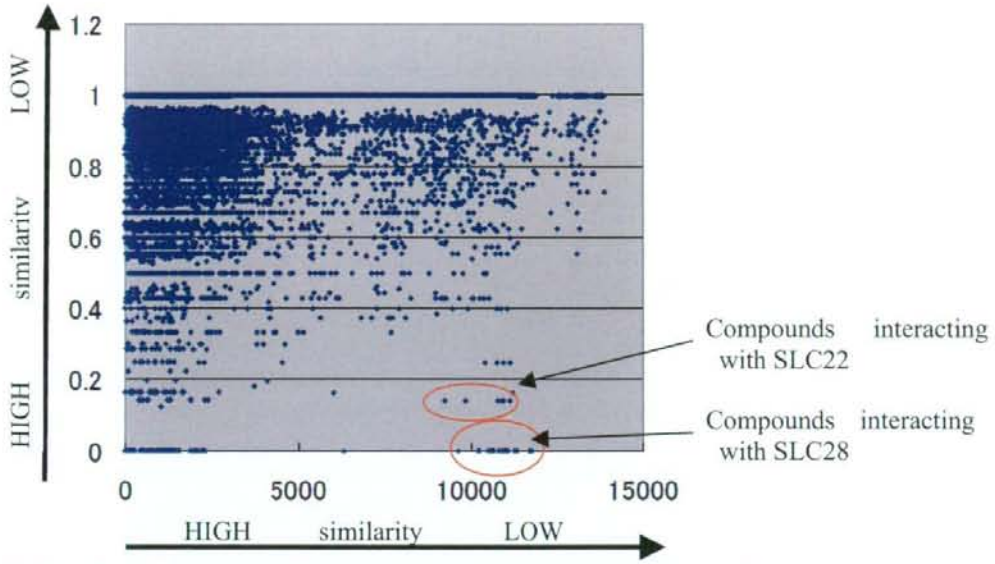


Figure 5. Two-dimensional map of structural similarity vs. interaction similarity
(Cyclosporin A and Vinblastine excluded from the map)
y-axis : Similarity in interaction pattern with a transporter protein
x-axis : Similarity in chemical structure

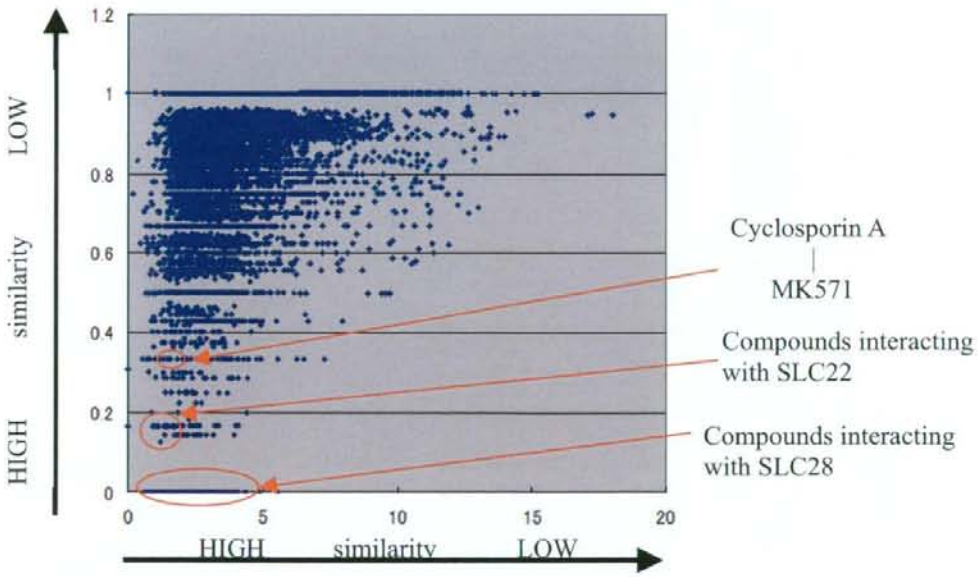


Figure 6. Two-dimensional map of similarity by CCA vs. interaction similarity
(Cyclosporin A and Vinblastine included in the training)
y-axis : Similarity in interaction with a transporter protein
x-axis : Similarity by CCA

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).

OPEN ACCESS Freely available online

PLoS one

Correlation Index-Based Responsible-Enzyme Gene Screening (CIRES), a Novel DNA Microarray-Based Method for Enzyme Gene Involved in Glycan Biosynthesis

Harumi Yamamoto[1,4], Hiromu Takematsu[1,5]*, Reiko Fujinawa[4], Yuko Naito[1,5], Yasushi Okuno[2], Gozoh Tsujimoto[3], Akemi Suzuki[4], Yasunori Kozutsumi[1,4,5]

1 Laboratory of Membrane Biochemistry and Biophysics, Graduate School of Biostudies, Kyoto University, Sakyo, Kyoto, Japan, 2 Department of PharmacoInformatics, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo, Kyoto, Japan, 3 Department of Genomic Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo, Kyoto, Japan, 4 Supra-Biomolecular System Research Group, RIKEN Frontier Research System, RIKEN, Wako, Saitama, Japan, 5 Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Corporation (JST), Kawaguchi, Saitama, Japan

Background. **Glycan biosynthesis occurs though a multi-step process that requires a variety of enzymes ranging from glycosyltransferases to those involved in cytosolic sugar metabolism. In many cases, glycan biosynthesis follows a glycan-specific, linear pathway. As glycosyltransferases are generally regulated at the level of transcription, assessing the overall transcriptional profile for glycan biosynthesis genes seems warranted. However, a systematic approach for assessing the correlation between glycan expression and glycan-related gene expression has not been reported previously.** ***Methodology.*** **To facilitate genetic analysis of glycan biosynthesis, we sought to correlate the expression of genes involved in cell-surface glycan formation with the expression of the glycans, as detected by glycan-recognizing probes. We performed cross-sample comparisons of gene expression profiles using a newly developed, glycan-focused cDNA microarray. Cell-surface glycan expression profiles were obtained using flow cytometry of cells stained with plant lectins. Pearson's correlation coefficients were calculated for these profiles and were used to identify enzyme genes correlated with glycan biosynthesis.** ***Conclusions.*** **This method, designated correlation index-based responsible-enzyme gene screening (CIRES), successfully identified genes already known to be involved in the biosynthesis of certain glycans. Our evaluation of CIRES indicates that it is useful for identifying genes involved in the biosynthesis of glycan chains that can be probed with lectins using flow cytometry.**

Citation: Yamamoto H, Takematsu H, Fujinawa R, Naito Y, Okuno Y, et al (2007) Correlation Index-Based Responsible-Enzyme Gene Screening (CIRES), a Novel DNA Microarray-Based Method for Enzyme Gene Involved in Glycan Biosynthesis. PLoS ONE 2(11): e1232. doi:10.1371/journal.pone.0001232

INTRODUCTION

The biosynthesis of glycan chains is a multi-step process. First, free sugars are biosynthesized by sugar-specific metabolic pathways. Then, these sugar molecules are further metabolized to nucleotide sugars, which serve as donors for glycosyltransferases [1]. Specific transporters move the nucleotide sugars to the endoplasmic reticulum (ER) or Golgi apparatus [2], where they are utilized by glycosyltransferases for the tandem addition of sugars to the termini of nascent glycan chains in a sugar- and linkage-specific manner [3]. This lengthy glycosylation process requires a great number of different enzymes operating at various levels of synthesis.

Thus far, more than 300 enzymes and transporter genes have been reported to be involved in the metabolism and biosynthesis of different glycans in diverse cell types and at various stages. Each glycan structure has its own specific biosynthetic pathway. The introduction of cloning expression methodology [4,5] has led to the successful cloning of a glycosyltransferase and to the demonstration that overexpression of a glycosyltransferase cDNA clone can confer the capability of glycan biosynthesis in over-expressing cells [6]. This mechanism is in contrast to that used by protein kinases, which also act via pathway-like processes but are often positively or negatively regulated by phosphorylation.

DNA microarray technology is very powerful because it can simultaneously detect changes in the expression levels of a large number of genes. In the field of glycobiology, extensive efforts have been made to identify the genes involved in glycan biosynthesis, and many have been shown to encode glycosyltransferases of the ER or Golgi apparatus. Many of these genes have been cloned, including those encoding large enzyme families [7]. Given the important role of gene transcription in the regulation of glycan biosynthesis, a glycan-focused cDNA microarray was developed to obtain the transcriptome of glycan-related genes [8,9]. As the presentation of glycomic information on a cell surface is likely to be regulated at the level of transcription of the enzymes in biosynthetic pathways, a glycan-focused DNA microarray may prove useful in elucidating glycan expression [8,10].

In the present study, we analyzed the glycan-related gene expression profiles for possible correlations with cellular glycan expression profiles in a cross-sample manner, using Pearson's

Academic Editor: Stefan Wölfl, Universität Heidelberg, Germany

Received March 19, 2007; **Accepted** November 4, 2007; **Published** November 28, 2007

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

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: htakema@pharm.kyoto-u.ac.jp

correlation coefficient. This analysis successfully identified specific genes encoding regulatory enzymes for the biosynthesis of specific glycans, from among the candidate genes of the glycan biosynthesis pathways. We designated this method correlation index-based responsible-enzyme gene screening, or CIRES (Figure 1).

RESULTS

Glycan-related gene-expression profiling using cDNA microarrays

The rat monoclonal antibody GL7 specifically stains germinal center B cells upon T cell-dependent antigen immunization. We recently demonstrated that GL7 recognizes the glycan Neu5Acα_{2-6}-Galβ_{1-4}-GlcNAc-R and that the sialyltransferase gene *ST6GAL1* is responsible for the biosynthesis of the glycan epitope recognized by GL7 [11]. By analyzing the correlation between the expression profiles for sialic acid (Sia) metabolism-related genes and the expression profiles for the GL7 epitope in a cross-sample manner, we showed that this type of correlation analysis was useful in screening for genes involved in the biosynthesis of the glycan epitope. In the present study, we further developed this systematic methodology by analyzing the correlations for cross-sample comparisons between the expression profiles for glycan-related genes and the expression profiles for various glycans, as determined by specific lectin binding (Figure 1).

The glycan-related gene expression profiles were obtained using total RNA isolated from six human B-cell lines cultured under optimal conditions, and cross-sample comparisons of these profiles were made in relation to a commercially available universal reference RNA consisting of a mixture of polyA(+) RNA from various organs. The gene expression profiles were determined as a ratio of the gene expression level to the universal reference cDNA expression level on the glycan-focused microarray ([11]; the complete relative gene expression profiles are shown in Table S1). Thus, the glycan-related gene expression profiles are expressed as the ratio of the gene expression signal at each spot on the microarray relative to the reference RNA signal.

After staining the cells with various anti-glycan probes (lectins) of known specificity, we determined the glycan expression profiles using flow cytometry. For each cell line, the transcriptional profile of glycan-related genes was used for cross-sample correlation analysis with the glycan expression profile.

Cell-surface glycan expression profiling using flow cytometric detection of lectin staining

Cell-surface glycan expression has been extensively studied using plant lectins that recognize specific glycan epitopes. To evaluate whether correlation analyses of lectin staining and glycan-related gene expression might provide useful information, we first performed lectin staining of a set of human B cells (Daudi, KMS-12BM, KMS-12PE, Namalwa, Raji, and Ramos) to obtain their cross-sample profiles of lectin epitope expression. We analyzed the strength of the correlations using Pearson's correlation coefficient, which is a standard, well-established method for assessing correlation. To prevent possible bias in the lectin choice, we used 15 plant lectins supplied in two commercially available sets.

To evaluate the efficacy of the calculations, the lectins were first divided into two groups based on the presence or absence of previous reports asserting a correlation between cell surface expression of a specific lectin epitope and expression of a certain glycosyltransferase gene. The lectins that lacked a reported correlation were divided into highly specific (or narrow) and broadly specific groups. The highly specific (narrow) lectins were further assigned to one of two subgroups according to the position of the epitope (terminal or interior) on the glycan chain.

CIRES correlation analyses of lectin staining profiles obtained using lectins with epitopes regulated by known biosynthetic enzyme genes

***Phaseolus vulgaris* leukoaggulutinin (PHA-L4)** PHA-L4 recognizes tri- or tetraantennary N-glycans with β_{1-6} branching of N-acetylglucosamine (GlcNAc), which often correlates with tumor progression [12]. Histochemical and immunoblot analyses have shown that PHA-L4 epitope expression correlates with the expression of the *MGAT5* (*GnT-V*) gene [13], and this lectin is commonly used as a marker for β_{1-6}-branched N-glycans. PHA-L4 epitope expression is diminished in *Mgat5*-null mice [14], and these mice exhibit enhanced rates of cytokine receptor internalization and subsequent cytokine signaling [15]. *MGAT5* expression was strongly correlated with the PHA-L4 staining profile, as shown in Figure 2A. The possible values of Pearson's correlation coefficient range between 1 and −1, where a value of 1 indicates complete correlation; therefore, the coefficient index between PHA-L4 staining and *MGAT5* expression (CI = 0.93) represents a highly significant correlation. Other correlated glycan-related genes were judged to be irrelevant to the biosynthesis of this epitope and are listed in Table S2, which contains the complete list of microarray-wide correlations for glycan-biosynthesizing genes.

The CIRES analysis correctly predicted that the *MGAT5* gene was responsible for expression of the PHA-L4 epitope. This prediction was confirmed by retrovirus-mediated gene expression in Namalwa B cells (Figure 2B). When a modified murine stem cell virus (MSCV) vector carrying genes for MGAT5 and enhanced green fluorescent protein (EGFP) divided with internal ribosomal entry site (IRES) *(MGAT5-IRES-EGFP)* was introduced into Namalwa cells, the level of PHA-L4 epitope expression was higher in the EGFP-positive population than in the EGFP-negative population. To rule out the possibility that viral infection somehow altered the cell surface glycan independently of glycosyltransferase expression, the vector carrying only *IRES-EGFP* was used as a negative control. In Namalwa cells expressing only EGFP, the EGFP-positive and EGFP-negative populations expressed identical levels of the PHA-L4 determinant (Figure 2B).

***Sambucus sieboldiana* agglutinin (SSA)** SSA recognizes α_{2-6}-linked Sia bound to galactose (Gal) or N-acetylgalactosamine (GalNAc). We previously showed that SSA epitope expression is induced in CHO cells by stable transfection with the rat *ST6GAL1* gene [11]. The deletion of *St6gal1* in mice eliminated the expression of the *Sambucus nigra* agglutinin (SNA) epitope [16], which is also recognized by SSA. In the present study, the correlation index was assessed to determine whether *ST6GAL1* gene expression correlated with SSA epitope expression, as determined by flow cytometry, in six B-cell lines. Although SSA staining in the six B-cell lines varied in intensity (Figure 2C), the staining profiles correlated with the gene expression profiles for *ST6GAL1* and a few other GlcNAc-transferase genes, including two β_{1-6} GlcNAc transferases and B3GNT5, which is involved in the biosynthesis of N-acetyllactosamine (LacNAc) units on glycan chains. These findings indicate that the SSA epitope detected by flow cytometry might be located at the terminus of poly-LacNAc units, which are often found on the β_{1-6} branch, and might extend beyond the glycocalyx of the cell surface.

***Arachis hypogaea* agglutinin (PNA)** PNA recognizes the Gal-exposed core-1 structure (Galβ_{1-3}GalNAc-Thr/Ser), and the capping of this epitope by sialylation severely reduces the affinity

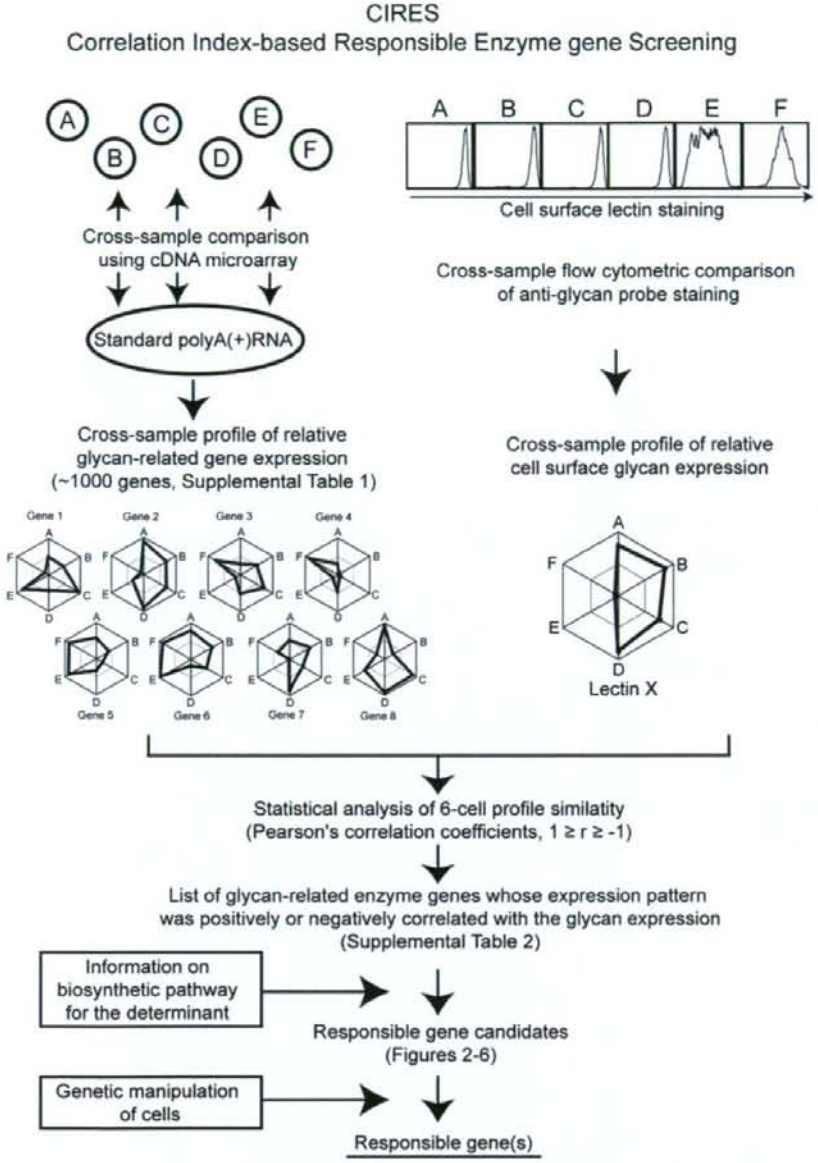


Figure 1. Schematic of the CIRES concept. The expression patterns of about 1000 glycan-related genes were profiled in a set of six different cell lines (A–F) by comparing the microarray binding of cellular cDNA and reference polyA(+) RNA and calculating the relative expression values (Table S1). The polygons in the left web graphs represent the relative gene expression profiles of eight glycan-related genes selected as examples. In these graphs, the difference in relative gene expression is expressed on a log scale, where the edge of the polygon corresponds to the strongest expression in each cell line (A–F). The same set of six cell lines were examined for cell-surface glycan expression using fluorescently labeled plant lectins and flow cytometry; the strength of the glycan expression is plotted as relative values among the six lines, where the edge of the polygon represents the strongest expression (web graph on top right). The glycan expression profiles were analyzed for correlations with the glycan-related gene expression profiles. Similarities and dissimilarities between the profiles were assessed using Pearson's correlation coefficient, which has values ranging from -1 (no correlation) to 1 (perfect correlation). A complete list of the genes found to be positively or negatively correlated with plant lectin staining patterns is presented in Table S2. Genes known to affect the biosynthesis of an epitope were selected from among the correlated genes (shown for each lectin in the tables on the right in Figures 2–6). A correlated gene identified by CIRES was confirmed as the gene responsible for regulating the biosynthesis of a particular glycan by transferring the gene into another cell line of the set, via gene transfer techniques such as retrovirus-mediated overexpression, and looking for a related change in epitope expression.
doi:10.1371/journal.pone.0001232.g001

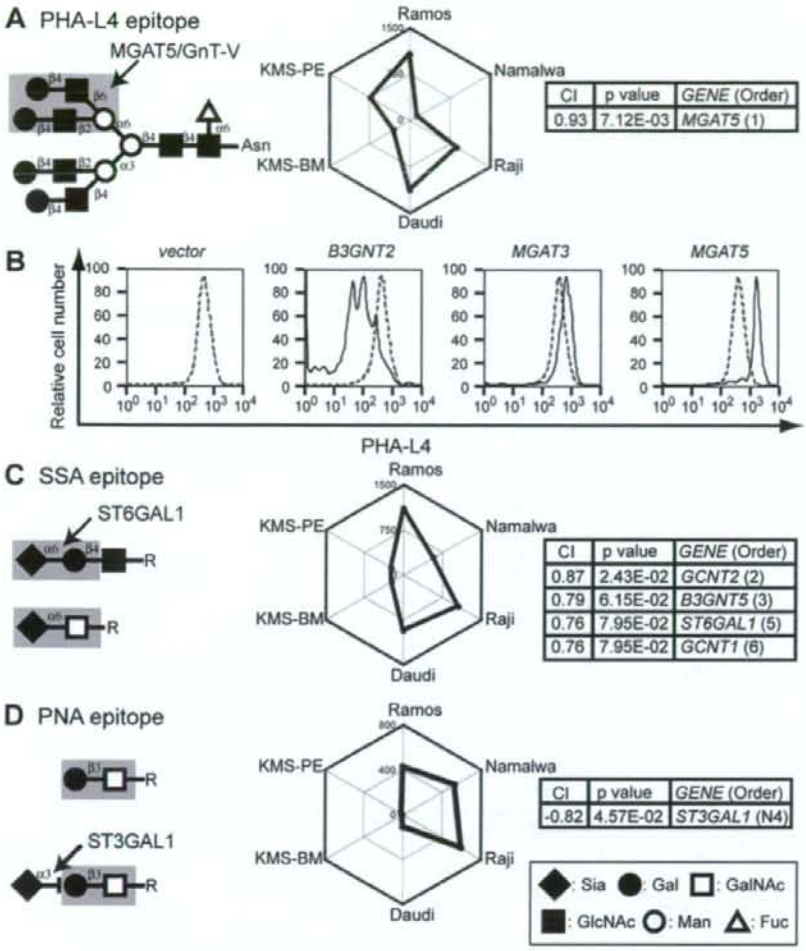


Figure 2. CIRES analyses of staining profiles obtained using lectins with known epitope expression-regulating enzymes. (A, C, D) Expected glycan structures for lectin recognition (left), web graphs of the lectin staining profiles (depicted as polygons) obtained using a set of six B-cell lines (middle), and the correlation indexes (CI, Pearson's correlation coefficient for profile matching) of the relevant genes that correlated with the plant lectin staining profiles and the *P* values of the correlations (right). The correlation orders of the glycan-related genes selected from the complete list of correlated genes (Table S2) are indicated as numbers in parentheses in the box for each gene, with a smaller number indicating a stronger correlation between gene expression and glycan expression profiles. Genes with a negative correlation are indicated by an N before the order number. The lectins used were (A) PHA-L4, (C) SSA, and (D) PNA. Lectin epitopes shown in the figures are taken from the literature unless otherwise specified [17,51]. (B) Namalwa cells were infected with MSCV harboring *MGAT5-IRES-EGFP*. Control cells were infected with empty vector (*IRES-EGFP*) or the same vector encoding *B3GNT2* or *MGAT3*. Flow cytometry results for PHA-L4 staining were compared between EGFP-positive cells (solid line) and EGFP-negative cells (dashed line).
doi:10.1371/journal.pone.0001232.g002

of the interaction [17]. Diminished PNA epitope expression reportedly coincides with an increase in α_{2-3} sialyltransferase activity, which sialylates the Gal residue [18]. This change was shown to occur during thymocyte maturation, in which PNA-positive cortex cells mature into PNA-negative medulla thymocytes. In a mouse model, the deletion of *St3gal1* caused a deficiency in the derepression of PNA reactivity during thymocyte development and eventually resulted in deficient $CD8^+$ T cell maturation [19].

The above findings suggest that the expression pattern of the PNA epitope might be positively affected by core-1 glycan biosynthesis and negatively affected by capping. Indeed, we found a negative correlation between PNA epitope expression and the *ST3GAL1* expression profile (Figure 2D). Thus, our correlation index analysis is able to not only identify a positive correlation but also reliably predict a negative correlation for a gene involved in the expression of a lectin glycan epitope.

Taken together, these results suggest that correlation indexing can be used to identify genes responsible for regulating cell surface expression of glycan epitopes, as determined by flow cytometry based on lectin binding. We designated this methodology as correlation index-based responsible-enzyme gene screening, or CIRES. After confirming that CIRES could be used to predict the genes involved in the biosynthesis of the glycan epitopes for these lectins (Figure 2), we used CIRES to assess the genes responsible

for the staining profiles of other plant lectins, as determined by flow cytometry.

CIRES correlation analyses of lectin staining profiles obtained using lectins that recognize specific terminal glycan structures and have unknown epitope expression-regulating enzyme genes

***Lens culinaris* agglutinin (LCA)** We assessed the staining profiles of lectins that recognize terminal structures of glycan chains. LCA recognizes the biantennary N-glycan chain with core α_{1-6} linked fucose (Fuc) attached to the chitobiose [20]. The presence of a core Fuc in the N-glycan of the Fc region of IgG severely represses the antibody-dependent cellular cytotoxicity activity of the antibody [21]. The expression of *FUT8* has been shown to be responsible for the biosynthesis of a core Fuc on N-glycans [22], but a correlation between *FUT8* gene expression and cell surface LCA staining has not been demonstrated in flow cytometry experiments.

Our analysis of the LCA staining profile and *FUT8* expression profile revealed a correlation (Figure 3A), although it was weaker than those for the three lectins described above (Figure 2). We also noted that *MGAT4b* gene expression negatively correlated with the LCA staining profile (Table S2). Considering that the presence of additional antennae on the N-glycan inhibits LCA binding [17], this type of negative correlation could be quite informative; however, in this case, no evidence was reported indicating that *MGAT4b* expression reduces the detection of the LCA epitope by flow cytometry. Nevertheless, CIRES is useful in predicting the genes involved positively or negatively in the biosynthesis of glycan epitopes.

***Ulex europaeus* agglutinin-I (UEA-I)** UEA-I recognizes α_{1-2}-linked Fuc on type-2 LacNAc, which is involved in forming the epitope of H-type human red blood cell antigen [23]. UEA-I staining did not reveal a significant positive correlation with the expression of the gene for α_{1-2} fucosyltransferase, which is involved in the biosynthesis of this linkage (Figure 3B). Instead, a prominent negative correlation was found with the expression profile of *ST3GAL6*, which has a preference for type-2 LacNAc substrates on both glycoproteins and glycolipids [24].

In theory, UEA-I binding should be affected by the expression of *FUT1* or *FUT2*, as they encode the proteins responsible for H antigen biosynthesis, and by the expression of A or B transferase, which can cap the H antigen to reduce the affinity [25]. However, the sequence similarity between the A and B (and also O) transferase genes prevented their differentiation in the microarray experiments. Redundant regulation by *FUT1* and *FUT2* in these cells may be the reason that no positive correlation with UEA-I epitope expression was observed. Alternatively, these data may suggest that negatively correlated ST3GAL6, which utilizes the same substrate as fucosyltransferases, may compete with the biosynthesis of this epitope by prior sialylation of the fucosyltransferase substrate(s).

***Ricinus communis* agglutinin (RCA120)** RCA120 preferentially recognizes terminal LacNAc structures found in various classes of glycans. These LacNAc structures are

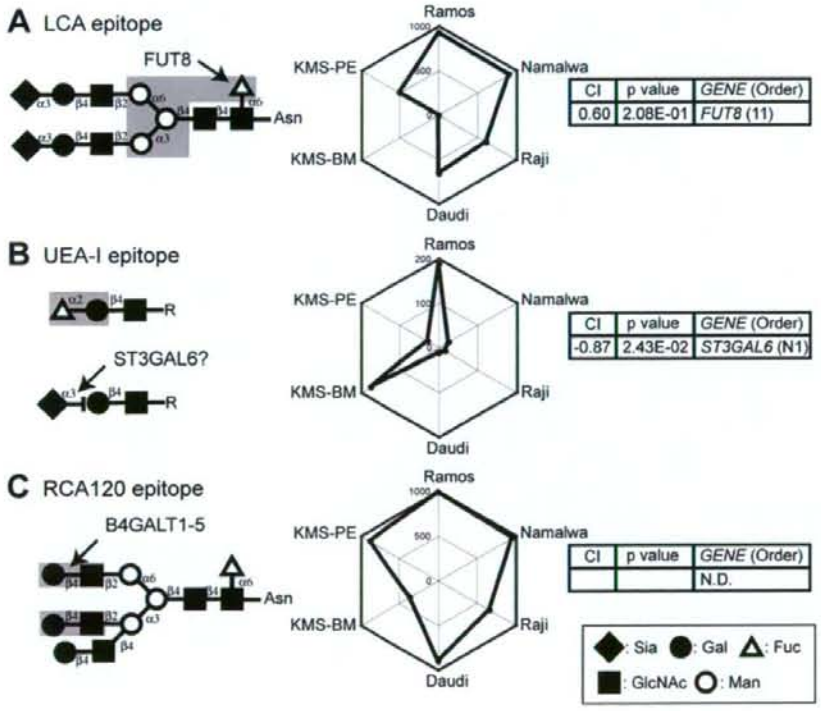


Figure 3. CIRES analyses of staining profiles obtained using lectins that recognize terminal glycan structures and have unknown epitope-expression-regulating enzymes. Presentation is the same as in Fig. 2 except that the plant lectins used were (**A**) LCA, (**B**) UEA-I, and (**C**) RCA 120. N.D. in the gene order list indicates that no gene was determined to have a correlation with the lectin staining.
doi:10.1371/journal.pone.0001232.g003

biosynthesized by a large group of B4GalT [26] and proximal GlcNAc-transferase family enzymes. The RCA120 staining profile revealed no obvious correlation with the genes for the enzymes known to be involved in this biosynthetic pathway (Figure 3C). Given our earlier correlation results identifying a terminal glycosyltransferase as being responsible for LacNAc expression (*i.e.*, sialyltransferase for SSA), this result was not surprising. It indicates that the abundant expression of LacNAc structures ensures the detection of a correlation between a terminal enzyme expression profile and the expression of a terminal glycan detected by flow cytometry. Moreover, capping of LacNAc should have a negative effect on its recognition by RCA120, which would make the detection of epitope expression more complex. It was clear that this procedure is not universally effective for lectin epitopes but that the effectiveness of the procedure depends on the type of glycan recognized by a lectin.

CIRES correlation analyses of lectin staining profiles obtained using lectins that recognize specific internal glycan structures and have unknown epitope expression-regulating enzyme genes

***Datura stramonium* agglutinin (DSA)** DSA recognizes tri- and tetraantennary N-glycans. It is specific for GlcNAc β_{1-4}-Man α_{1-3}-branched triantennary N-glycan [27,28], which is biosynthesized by MGAT4a and MGAT4b. In our experiments, the DSA staining profile resembled that of PHA-L4, and thus the two lectins correlated with similar genes, most prominently *MGAT5* (Figure 4A). This result could be explained by the fact that the addition of a β_{1-6} branch increases the preference of DSA for a ligand, even though MGAT4a/b activity is required. Ihara et al. have reported that DSA staining correlates with the expression level of *MGAT5* in *in vitro*-differentiated GOTO cells [29], suggesting that *MGAT5* may also be involved in the biosynthesis of the optimal DSA epitope, with a tetraanntenary glycan. Consistent with this idea, the introduction of *MGAT5* into Namalwa cells resulted in a 60% increase in DSA staining (Mean fluorescence intensity (MFI), 1986), compared with control (MFI, 1247) (Figure 4B). Interestingly, when *MGAT3* was introduced into Namalwa cells, DSA epitope expression was subtly suppressed (MFI, 961) compared with control expression (MFI, 1222), possibly due to the competitive relationship between MGAT3 and MGAT5 [30] (Figure 4B). These effects appeared to be specific, because no obvious shift was seen in cells with introduced *B3GNT2*.

***Phaseolus vulgaris* erythroagglutinin (PHA-E4)** The staining profile of PHA-E4 was similar to those of PHA-L4 and DSA. This result was unexpected because PHA-E4 recognizes bisecting GlcNAc-containing biantennary N-glycans, which comprise a type of glycan distinct from the PHA-L4 epitope. Owing to the similarity among the staining patterns of these three lectins, a correlation was also found between PHA-E4 staining and *MGAT5* (Table S2), but PHA-E4 staining did not correlate with *MGAT3* (*GnT-III*), which is the GlcNAc transferase gene expected to correlate by virtue of its known epitope specificity (Figure 4C). However, when we overexpressed *MGAT3* in Namalwa cells, the MFI value of PHA-E4 staining increased, from 873 in the control population to 1868 in the EGFP-positive population (Figure 4D). Thus, the expression level of *MGAT3* appears to be important for PHA-E4 epitope biosynthesis, as expected. The overexpression of *MGAT5* had no effect on PHA-E4 binding; the MFI value was 899 in the control population and 866 in the EGFP-positive population.

When we stained the membrane fractions from the six B-cell lines using PHA-E4 in lectin-blot analyses, the blot and FACS signal strengths differed, as seen in the shape of the staining profile (Figure 4E), and *MGAT3* expression did not correlate with the signal strength on the lectin blot. Somewhat consistent with our result, Miyoshi et al. reported that *MGAT3* expression levels did not necessarily correlate with cell-surface staining of the PHA-E4 ligand in flow cytometry experiments, although co-expression was found in lectin-blot experiments [31]. Thus, our results support the suggestion of Miyoshi et al. that the cell-surface expression level of the PHA-E4 ligand epitope may be regulated by factor(s) other than *MGAT3* expression. Consistent with this idea, they also reported that the presence of bisecting GlcNAc negatively affected the sorting of glycoproteins to the cell surface [32].

CIRES correlation analyses of lectin staining profiles obtained using lectins that recognize multiple glycan structures

Some of the lectins used in the present study had mixed or heterologous specificity. We assessed the correlation indexes for the staining profiles of these lectins.

***Maackia amurensis* lectin (MAM)** MAM is a mixture of two lectin subunits, MAL and MAH. MAL binds to Sia α_{2-3}-LacNAc structures [33], whereas MAH preferentially recognizes disialylated structures found in O-glycans [34]. Of the known sialyltransferases, ST3GAL3, ST3GAL4, or ST3GAL6 may synthesize the MAL epitope, and ST8s may synthesize disialylated glycans. Correlation-index analyses showed that *ST3GAL3* and *B3GNT2* may be responsible for the expression of the epitope in the six B-cell lines (Figure 5A). This is consistent with a previous report that repeating LacNAc units enhance MAL binding [35]. The MAL binding preference seemed to be more important than that of MAH in this CIRES prediction based on MAM staining and flow cytometry. As expected from its positive correlation with *B3GNT2* expression, the MAM epitope showed increased levels in Namalwa cells overexpressing *B3GNT2*, whereas overexpressed *MGAT3* was negatively correlated with the MAM staining profile, owing to the suppression of MAM epitope expression (Figure 5B). Thus, the MAM epitope may be preferentially biosynthesized on LacNAc units of the β_{1-6} branch of N-glycans. Alternatively, MGAT3 expression may change the sorting of the protein carrying the MAM epitope. Taken together, these results indicate that the expression of correlated genes can have an additive regulatory effect (positive or negative) on the cell-surface presentation of a lectin epitope.

***Triticum vulgaris* agglutinin, wheat germ agglutinin (WGA)** WGA is thought to preferentially recognize clustered N-acetyl groups found in N-acetylneuraminic acid (Neu5Ac), GlcNAc, and GalNAc. Neu5Ac is often a major WGA ligand because the Sia density on the termini of glycan chains tends to increase for the highly branched N-linked glycans [17]. The affinity of WGA for Sia was exploited in the isolation of Lec mutants in CHO cells [36]. The density of the N-acetyl group can also be high in the I-branched β_{1-6} GlcNAc-containing glycans [17].

The WGA staining profile correlated strongly with the expression profile of the *ST6GAL1* gene and weakly with that of the *ST3GAL3* gene (Figure 5C). (These genes were previously known as *ST6N* and *ST3N*, respectively [37].) Since WGA binding to sialylated glycans increases with the degree of sialylation, this correlation pattern seems to indicate that the supply of the substrate LacNAc is ample and that expression of the terminal sialyltransferases determines the expression level of the WGA epitope. Among these sialyltransferases, *ST6GAL1* appeared to play a more prominent role in biosynthesis in the B-cell lines used in this study. In addition to the I-branching β_{1-6} GlcNAc transferase [38], *GCNT2* expression also correlated with WGA