4GlcNAc-BSA (Dextra), 10 ng/ml Manα1–3(Manα1–6)Man-BSA (Dextra), 10 ng/ml αFuc-BSA (Dextra), 10 ng/ml αGalNAc-BSA (Dextra), and 10 ng/ml Siaα2–6Galβ1–4Glc-BSA (Dextra) dissolved in probing buffer.

Lectin Microarray Analysis—Hydrophobic fractions were prepared using CelLytic minimum essential medium protein extraction (Sigma-Aldrich) in accordance with the manufacturer's procedures (25, 27). After protein quantification using a BCA assay (Thermo Fisher Scientific), hydrophobic fractions were fluorescently labeled with Cy3 monoreactive dye (GE Healthcare), and excess Cy3 was removed with Sephadex G-25 desalting columns (GE Healthcare). After adjusting the protein concentration to 2 μg/ml with PBST (10 mM PBS, pH 7.4, 140 mM NaCl, 2.7 mM KCl, 1% Triton X-100), the hydrophobic fraction was labeled with Cy3 NHS ester (GE Healthcare). After dilution with probing buffer at 0.5 μg/ml, the Cy3-labeled hydrophobic fraction was applied to the lectin microarray and incubated at 20 °C overnight. After washing with probing buffer, fluorescence images were acquired using an evanescent field-activated fluorescence scanner (GlycoStation^{TM} reader 1200; GP BioSciences). The fluorescence signal of each spot was quantified using Array Pro Analyzer version 4.5 (Media Cybernetics, Bethesda, MD), and the background value was subtracted. The background value was obtained from the area without lectin immobilization. The lectin signals of triplicate spots were averaged and normalized to the mean value of 96 lectins immobilized on the array. An inhibition assay was performed by incubating Cy3-labeled cell membrane fractions of MEF(#1) or MRC5-iPS#25(P22)(#13) with a lectin microarray either in the absence or presence of 100 μg/ml of Galα1–3Galβ1–4GlcNAc-PAA (catalog no. 01-079, Glycotech) or a negative control PAA (catalog no. 01-000, Glycotech).

Gene Expression Analysis—Total RNA was extracted from each sample by using ISOGEN (NipponGene). The global gene expression patterns were monitored using Agilent whole human genome microarray chips (G4112F) with one-color (cyanine 3) dye. This microarray covers 41,000 well characterized human genes and transcripts. Of the 41,000 probes, 16,483 representative probes corresponding to the microarray quality control unique genes were used for the following analyses (37).

Statistics—Unsupervised clustering was performed by employing the average linkage method using Cluster 3.0 software. The heat map with clustering was acquired using Java Treeview. Differences between the two arbitrary data sets were evaluated by Student's *t* test to each lectin signal using SPSS Statistics 19 (SPSS). Significantly different lectin signals or the glycosyltransferase expression were selected if they satisfied a familywise error rate (FWER) by the Bonferroni method of <0.001.

Immunocytochemistry—Immunocytochemical analysis was performed as described previously (29, 33, 38). Human iPSCs were fixed with 4% paraformaldehyde in PBS for 10 min at 4 °C. After washing with 0.1% Triton X-100 in PBS (PBST), the cells were prehybridized in blocking buffer for 1–12 h at 4 °C and then incubated for 6–12 h at 4 °C with the following primary antibodies: anti-SSEA4 (1:300 dilution; Chemicon), anti-TRA-1-60 (1:300; Chemicon), anti-Oct4 (1:50; Santa Cruz Biotechnology, Inc.), anti-Nanog (1:300; ReproCELL), and anti-Sox2 (1:300; Chemicon). The cells were then incubated with anti-rabbit IgG, anti-mouse IgG, or anti-mouse IgM conjugated with Alexa Fluor 488 or Alexa Fluor 546 (1:500; Molecular Probes) in blocking buffer for 1 h at room temperature. The cells were counterstained with DAPI and then mounted using the SlowFade light antifade kit (Molecular Probes).

Teratomas—Teratoma formation was performed as described previously (1, 2). The 1:1 mixtures of the human iPSC suspension and basement membrane matrix (BD Biosciences) were implanted subcutaneously at 1.0×10^7 cells/site into immunodeficient, non-obese diabetic/severe combined immunodeficiency mice. Teratomas were surgically dissected out 8–12 weeks after implantation and were fixed with 4% paraformaldehyde in PBS and embedded in paraffin. Sections of 10-μm thickness were stained with hematoxylin-eosin.

Glycoconjugate Microarray Analysis—Glycoconjugate microarray production and analysis were performed as described previously (36). Briefly, glycoproteins and glycoside-polyacrylamide conjugates were dissolved in the Matsunami spotting solution at a final concentration of 0.5 and 0.1 mg/ml, respectively. After filtration, they were spotted on the Schott epoxy-coated glass slide using the Microsys non-contact microarray printing robot.

Cy3-labeled lectins dissolved in the probing solution (10 or 1 μg/ml) were applied to each chamber of the glycoconjugate microarray (100 μl/well) and were incubated at 20 °C overnight. After washing the chambers with the probing solution, fluorescent images were immediately acquired using an evanescent field-activated fluorescence scanner, the GlycoStation^{TM} Reader 1200, under Cy3 mode. Data were analyzed with the Array Pro analyzer version 4.5 (Media Cybernetics, Inc.). The net intensity value for each spot was determined by signal intensity minus background value. The lectin signals of triplicate spots were averaged and normalized to the highest signal intensity among 98 glycoconjugates immobilized on the array.

RESULTS

Development of High Density Lectin Microarray—In order to increase glycome coverage and the selection range of lectins suitable for stem cell evaluation, we first increased the number of immobilized lectins from 43 to 96, which is the largest number of immobilized lectins reported (39). For this purpose, lectins with defined structures were first categorized into lectin families with different protein scaffolds. We then selected lectins from various lectin families, intending to cover a wider range of glycan binding specificities. Especially, we increased lectins specific to terminal modifications, such as Sia and Fuc, which often change dramatically depending on cell properties. For production of recombinant lectins, the *E. coli* expression system was chosen to avoid glycosylation of the produced lectins, which might cause nonspecific binding to lectin-like molecules in the objective samples. The recombinant lectins thus produced were purified by affinity chromatography using the most appropriate sugar-immobilized Sepharose. The glycan-binding specificities of 96 lectins used in this study were analyzed by both glycoconjugate microarray (supplemental Fig. S1 and Table S1; also see "Experimental Procedures") (36) and,

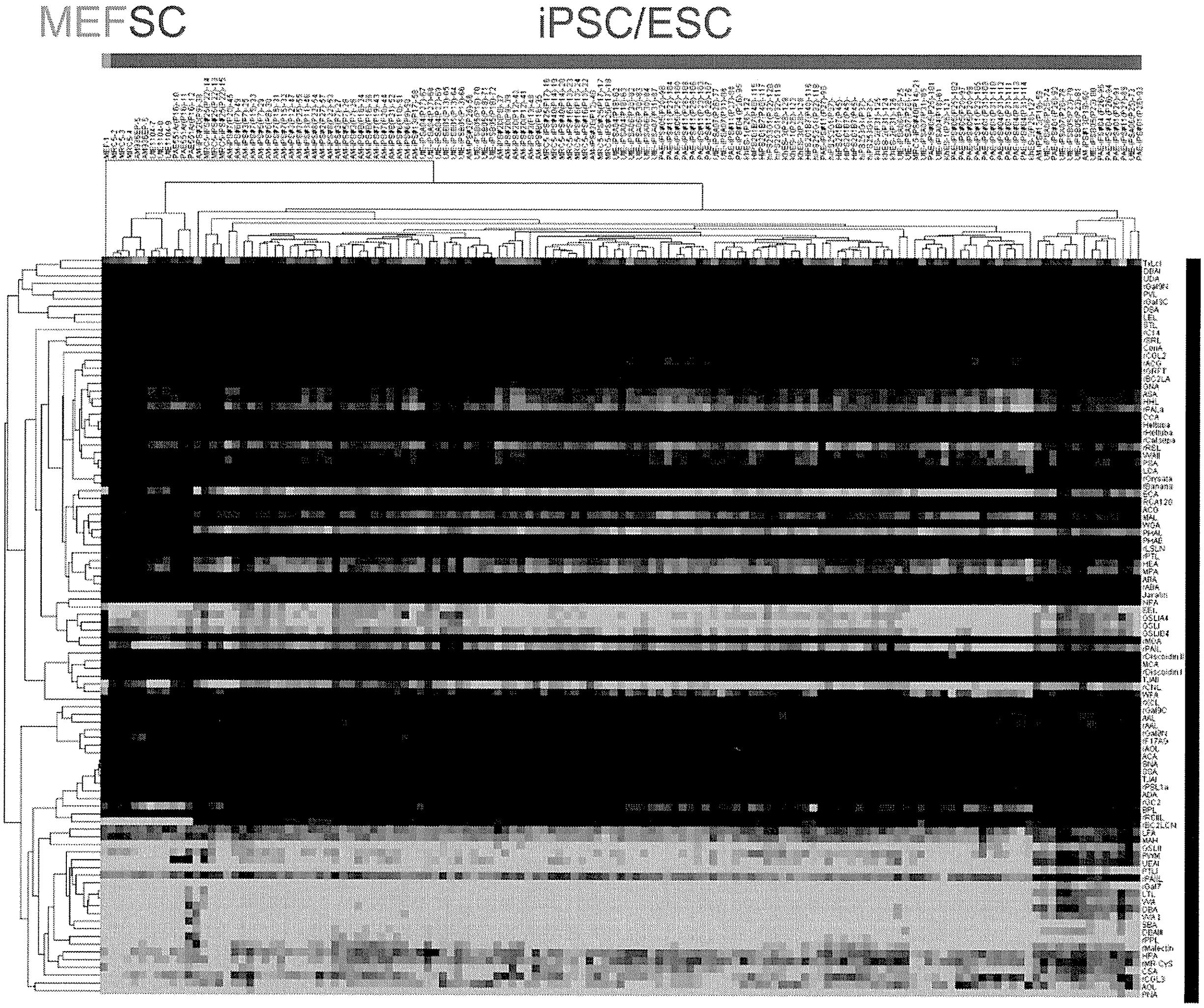


FIGURE 1. **Unsupervised cluster analysis.** Lectin microarray data of iPSCs (n = 123), their parental SCs (n = 11), ESCs (9), and MEF (n = 1) were mean-normalized and log-transformed and then analyzed by Cluster 3.0. The zero value of the lectin signal was converted to 1. *Yellow*, positive; *blue*, negative. Clustering method was average linkage. The heat map with clustering was acquired using Java Treeview.

more quantitatively, frontal affinity chromatography (see the Lectin Frontier Database Web page) (35, 40). Their basic specificities evaluated by the above two analytical methods are briefly summarized in supplemental Table S2. The 96 lectins were spotted onto epoxy-activated glass slides by a non-contact spotter (supplemental Fig. S2), and their quality was extensively assessed using a Cy3-labeled test probe (25). Lot-to-lot variance (coefficients of variation) of the developed high density lectin microarray was confirmed to be low (0.14) after mean normalization (25).

Transcription Factor-induced Reprogramming Leads to a Global Reversion Down to the Pluripotent State at a Cellular Glycome Level as Well—Using the developed lectin microarray, we have analyzed 135 cell samples in total, including 114 iPSCs, 11 SCs, and nine ESCs, all from human origins, as well as one mouse embryonic fibroblast (MEF). Human iPSCs were generated from four different SC lines: MRC5, AM, UtE, and PAE (supplemental Table S3) (28). We have also analyzed human iPSCs generated from human dermal fibroblasts with four (201B7) (1) and three transcription factors (253G1) (41) and three cell lines of human ESCs (42). All iPSCs used in this study were morphologically similar to ESCs, and their pluripotency was confirmed by staining with the established undifferentiation markers (SSEA4, Tra1–60, Oct4, Nanog, and Sox2) and DNA microarray (28).

Cell membrane hydrophobic fractions were prepared, and the extracted glycoproteins were then labeled with Cy3-*N*-hydroxysuccinimide ester and analyzed by lectin microarray (25). We have analyzed cell membrane fractions because they can be stored in a freezer until use and are easy to handle, allowing comprehensive analysis of a large number of samples (25, 26).

After being mean-normalized, the obtained data were first analyzed by unsupervised hierarchical clustering (Fig. 1). As a result, differentiated SCs and undifferentiated iPSCs/ESCs were clearly separated into two large clusters, whereas the four SCs were further separated according to their origins. This indicates that SCs (MRC5, AM, UtE, and PAE) with different glycan profiles have acquired profiles quite similar to one

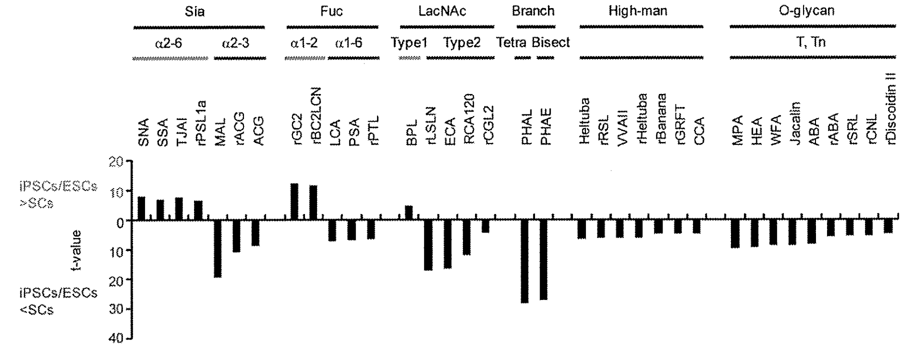


FIGURE 2. **Alterations of the lectin signals upon induction of pluripotency.** Lectin microarray data were mean-normalized and analyzed by Student's *t* test. Lectins with significantly different signals (FWER < 0.001) between undifferentiated iPSCs/ESCs (*n* = 123) and differentiated SCs (*n* = 11) were categorized into six groups based on the glycan binding specificities of lectins. Data are shown with *t* values. Also see supplemental Table S4.

another and even to ESCs upon induction of pluripotency. Thus, transcription factor-induced reprogramming was found to lead to a global reversion down to the pluripotent state at a cellular glycome level as well (27, 28).

Characteristic Features of Glycome Alteration upon Induction of Pluripotency—We then examined in more detail how glycan structures altered during the induction of pluripotency. The mean-normalized data were processed by Student's *t* test to select significant probe lectins discriminating between SCs and iPSCs/ESCs (supplemental Table S4). As a result, 38 lectins were selected with FWER of <0.001. Among them, nine gave higher signals in iPSCs than SCs, whereas 29 exhibited lower signals. Among the 38 lectins, 35 lectins were then categorized into six groups based on their glycan binding specificities, from which glycan alterations having occurred upon induction of pluripotency were estimated (Fig. 2), whereas the three lectins with broader specificities (wheat germ agglutinin (WGA), a Sia binder; rRSIIL and aleuria aurantia lectin (AAL), broad Fuc binders) were not included in this categorization. Here, the lectins with higher signals in iPSCs/ESCs than SCs are shown *below gray lines*, whereas lower signals are shown *below black lines*. The characteristic features of glycan structures of undifferentiated iPSCs/ESCs relative to differentiated SCs are summarized as follows. 1) The signals of α2–6Sia-binding lectins (SNA, SSA, TJAI, and rPSL1a) were increased, whereas those of α2–3Sia-binding lectins (MAL, rACG, and ACG) were decreased correspondingly (28). This agrees well with the previous report that α2–6-sialylated glycan expression is higher in undifferentiated (human ESCs) than differentiated cells (embryoid body) (43). 2) In terms of fucosylation, the signals of α1–2Fuc-specific lectins (rGC2 and rBC2LCN) were increased, whereas those of α1–6Fuc-specific lectins (LCA, PSA, and rPTL) were decreased. This is consistent with the recent report that human ESCs are stained with anti-Globo H (Fucα1–2Galβ1–3GalNAcβ1–3Galα1–4Galβ1–4Glc) and anti-H type 1 (Fucα1–2Galβ1–3GlcNAc), whose antigens contain α1–2Fuc (9). 3) The signals of type 1 LacNAc (Galβ1–3GlcNAc)-binding lectins (BPL) were increased, whereas those of type 2 LacNAc (Galβ1–4GlcNAc)-binding lectins (rLSLN, ECA, RCA120, and rCGL2) were decreased. This agrees well with the recent finding that type 1 LacNAc is the glycan epitope recognized by the well known pluripotency markers Tra-1-60 and Tra-1-81 (7). 4) The lectin signals specific to bisecting GlcNAc (PHAE), tetra-antennary *N*-glycans (PHAL), high mannose type *N*-glycans (Heltuba, rRSL, VVAII, rHeltuba, Heltuba, rBanana, rGRFT, and CCA), and *O*-glycans (MPA, HEA, WFA, Jacalin, ABA, rABA, rSRL, rCNL, and rDiscoidin II) were decreased.

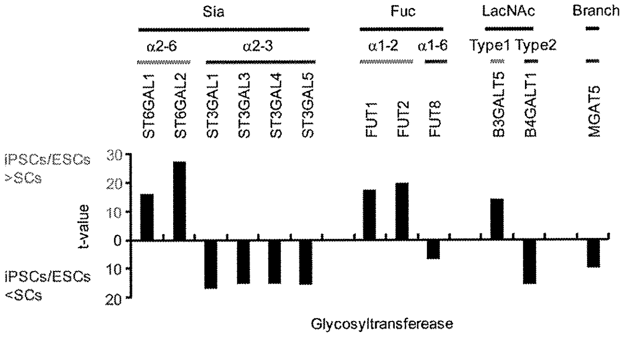


FIGURE 3. **Alterations in the expression of glycosyltransferases upon induction of pluripotency.** Glycosyltransferases related to the lectin signals in Fig. 2 are shown with *t* values. All data are shown in supplemental Table S5.

The expression profiles of glycosyltransferases synthesizing glycans agreed well with the results obtained by lectin microarray (Fig. 3 and supplemental Table S5); the expression of α2–6-sialyltransferases (ST6GAL1 and -2) (28), α1–2-fucosyltransferases (FUT1 and -2), the major glycosyltransferase involved in the synthesis of type 1 LacNAc (B3GALT5) (28), and MGAT5, a glycosyltransferase involved in the synthesis of tetra-antennary *N*-glycans, was increased, whereas that of α2–3-sialyltransferases (ST3GAL1, -3, -4, and -5), α1–6-fucosyltransferase (FUT8), and the major glycosyltransferase related to the synthesis of type 2 LacNAc (B4GalT1) was decreased correspondingly in iPSCs/ESCs relative to SCs. Based on the results obtained by lectin and DNA microarrays, it is conceivable that the expression of α2–6-sialylation, α1–2-fucosylation, and type 1 LacNAc is increased, whereas that of α2–3-sialylation and tetra-antennary *N*-glycans is decreased upon induction of pluripotency (Fig. 4).

Selection of the Best Lectin Probe to Discriminate Pluripotency—We then addressed the challenge to develop a lectin-based procedure to discriminate between differentiated SCs and undifferentiated iPSCs/ESCs, which could be utilized to monitor the state of differentiation. As described, rGC2, rBC2LCN, SNA, TJAI, SSA, rPSL1a, rRSIIL, BPL, and AAL gave

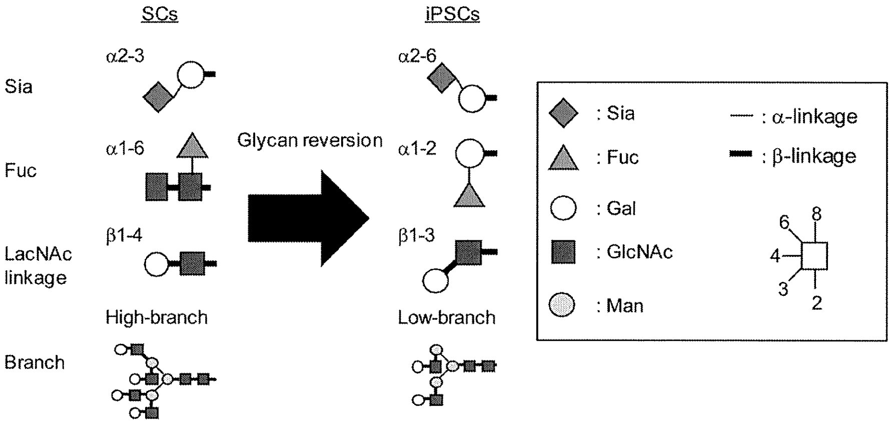


FIGURE 4. **Schematic representation of the putative glycan alterations upon induction of pluripotency.**

TABLE 1

Selection of the best lectin probe to evaluate pluripotency among 96 lectins

Lectins with significantly higher signals (FWER < 0.001) for iPSCs/ESCs than SCs are shown. Also see supplemental Fig. S4.

Lectin	SC (mean, $n = 11$)	SC (S.D.)	iPSC/ESC (mean, $n = 123$)	iPSC/ESC (S.D.)	FWER (Bonferroni)
rGC2	125	31	304	47	1E−21
rBC2LCN	1	1	23	6	2E−19
SNA	69	43	124	18	4E−11
TJAI	96	58	170	28	9E−10
SSA	97	55	156	22	9E−09
rPSL1a	61	32	107	21	2E−07
rRSIIL	78	23	109	20	3E−04
BPL	9	5	20	8	7E−04
AAL	518	77	677	114	1E−03

significantly higher signals in iPSCs/ESCs than SCs with FWER < 0.001 (Table 1). Among them, rBC2LCN showed the best performance as a probe to detect only undifferentiated iPSCs/ESCs but never reacted with differentiated SCs and MEF, whereas other lectins also reacted with SCs (Table 1). Namely, although rGC2 showed a better score in terms of FWER (1×10^{-21}) than rBC2LCN (2×10^{-19}), the former reacted strongly with MEF (Fig. 5). Similarly, SNA (4×10^{-11}), a representative α2–6Sia-binding lectin, showed significant cross-reactivity with a part of SCs derived from PAE in addition to MEF (Fig. 5).

Monitoring the Contamination of the Xenoantigen, αGal Epitope—From a practical viewpoint, monitoring possible contamination by xenotransplantation antigens in iPSCs/ESCs is essential for their safe use in regenerative medicine. A recombinant MOA (rMOA) recognizes the xenotransplantation antigen Galα1–3Galβ1–4GlcNAc (44) present in most cells from New World monkeys and non-primate mammals, including mice, but not in humans. Indeed, rMOA strongly bound to MEFs but not to any human SCs (Fig. 6). Therefore, rMOA signals should not be detected in human iPSCs. However, triplicate samples of the two cell lines MRC5-iPS#25(P22)(#13–15) and UtE-iPSB05(P13)(#64–66) exhibited significant signals on rMOA. In order to validate whether the binding of rMOA is mediated by a carbohydrate recognition domain of rMOA, we then performed inhibition assay. As shown in Fig. 6*B*, the binding of rMOA to cell membrane fractions of MEF and MRC5-iPS#25(P22,#13) were abolished in the presence of 100 μg/ml Galα1–3Galβ1–4GlcNAc-PAA (Fig. 6*B*), but no inhibitory effect was observed for 100 μg/ml of a negative control (PAA without sugar moiety), indicating that the binding is due to specific interactions via the rMOA carbohydrate recognition domain. As expected, no inhibitory effect of Galα1-3Galβ1–4GlcNAc-PAA on a Fuc-binding lectin, rAAL, was observed. These data unambiguously reflect contamination by the xenoantigen αGal epitope, in the above two cell lines, which were most probably contaminated with MEF.

DISCUSSION

Using the developed high density lectin microarray, we performed a systematic analysis of cell surface glycans of a large set of human iPSCs (114 cell types) and ESCs (nine cell types). As a result, a basis for a rational stem cell evaluation system was established, which can reveal both the state of undifferentiation and inclusion of αGal epitope (a representative xenoantigen). Such a comprehensive glycome analysis targeting iPSCs and ESCs has never been carried out so far. There are at least three key advantages in using a lectin microarray. 1) An overall glycan profile of each cell type is readily obtained using a relatively small number of cells ($\sim 1 \times 10^3$), and thus, the method is widely applicable to stem cells. 2) The proposed evaluation system includes selection of the best probe by a statistical strategy among a number of lectins, which are immobilized on the array. As candidate probes, carbohydrate-binding antibodies developed so far could also be included. 3) Various properties of stem cells can be assessed simultaneously (*i.e.* with "one-chip" technology). Using the same strategy described in this study, lectin-based evaluation methods targeting tumorigenesis and the differentiation propensity of stem cells could also be developed.

Based on the lectin signals and the expression profiles of glycosyltransferases, we concluded that the expression of α2–6-sialylation, α1–2-fucosylation, and type 1 LacNAc increases, whereas that of α2–3-sialylation and tetra-antennary *N*-glycans decreases correspondingly upon the induction of pluripotency. These changes are consistent with the recent reports that relevant glycans (*i.e.* the expression of Globo H and H type 1 with an α1–2Fuc and α2–6Sia in human ESCs) are higher than that in differentiated embryoid body (9, 43). Interestingly, the increased expression of α1–2Fuc and type 1 LacNAc, which are synthesized by the action of FUT1/2 and B3GalT5, respectively, is closely related to the synthesis of the well known pluripo-

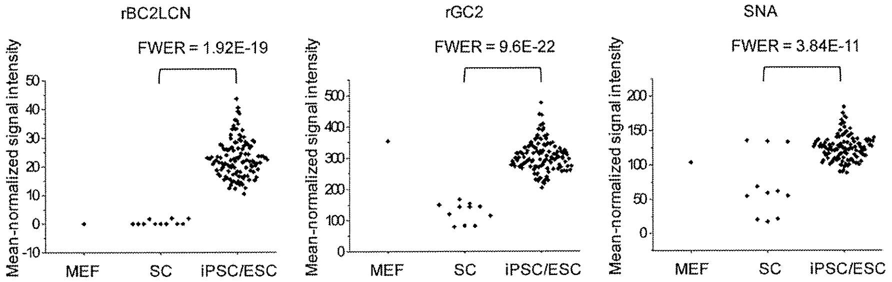


FIGURE 5. **Selection of the best lectin probe to discriminate pluripotency.** The mean-normalized signal intensities of rGC2, rBC2LCN, and SNA to MEF (n = 1), SCs (n = 11), and iPSCs/ESCs (n = 123) are shown.

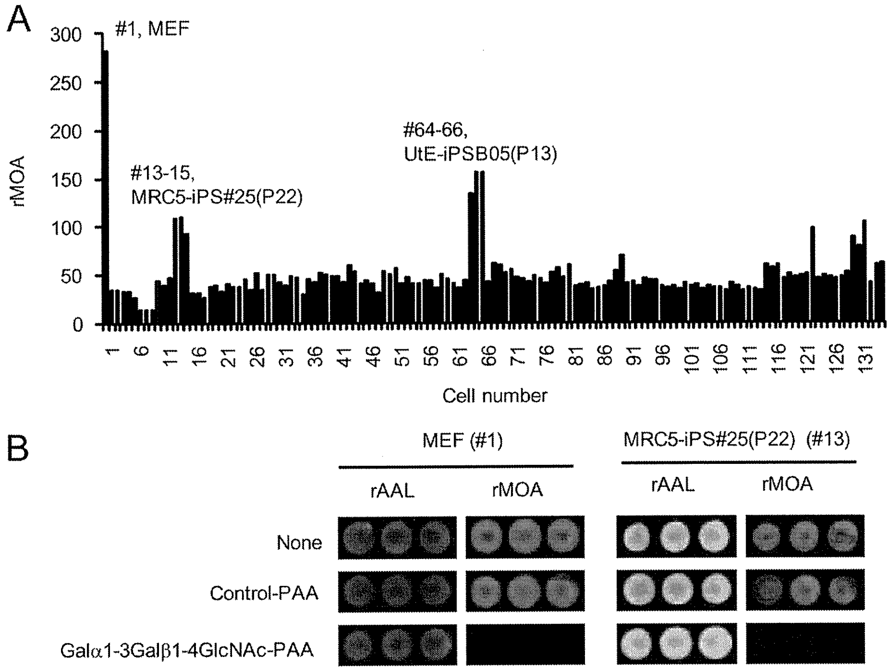


FIGURE 6. **Monitoring the contamination of the xenoantigen αGal epitope using the rMOA lectin.** *A*, mean-normalized lectin microarray data are represented by a *bar graph. Numbers* correspond to cell types described in supplemental Table S3. *B*, inhibition assay. Cy3-labaled cell membrane fractions of MEF(#1) or MRC5-iPS#25(P22)(#13) were incubated with lectin microarray either in the absence (*None*) or presence of 100 μg/ml Galα1-3Galβ1-4GlcNAc-PAA or negative control PAA without sugar moiety. Data shown were obtained at gain 110 for MEF and gain 120 for MRC5-iPS#25(P22).

tency markers, SSEA3/4 and Tra-1-60/81 (for a scheme, see supplemental Fig. S4).

In this study, rBC2LCN was selected as the best lectin probe to evaluate pluripotency among the 96 lectins. BC2LCN is a TNF-like lectin molecule identified from a Gram-negative bacterium *Burkholderia cenocepacia* (45). Glycoconjugate microarray analysis revealed that rBC2LCN binds specifically to Fucα1–2Galβ1–3GlcNAc (GalNAc)-containing glycans, such as H type 1 (Fucα1–2Galβ1–3GlcNAc), H type 3 (Fucα1–2Galβ1–3GalNAc), and Lewis b (Fucα1–2Galβ1–3(Fucα1-4)GlcNAc), which include the two structural characteristics related to the pluripotency (α1–2Fuc and type 1 LacNAc) as described above (supplemental Fig. S3). This observation is consistent with the previous report (45) in which Sulak *et al.* studied the glycan-binding specificity of BC2LCN in detail using glycan microarray and titration microcalorimetry. They also demonstrated that this lectin also binds to Globo H (Fucα1–2Galβ1–3GalNAcβ1–3Galα1–4Galβ1–4Glc), which was recently proposed as a glycosphingolipid type pluripotency marker (supplemental Fig. S4) (9, 45). These results explain the mechanism of how this lectin could be used as the probe to discriminate pluripotency. rBC2LCN could be used to probe glycoproteins and possibly all glycoconjugates carrying Fucα1–2Galβ1–3GlcNAc (GalNAc), whereas anti-SSEA3 and anti-SSEA4 specifically target glycosphingolipids. From a practical viewpoint, rBC2LCN is cost-effective because it can be produced in large amounts by the conventional *E. coli* expression system (84 mg/liter). Thus, this lectin could be a versatile probe to evaluate pluripotency.

In contrast, Globo H has also been reported to be overexpressed in epithelial cell tumors (46). Furthermore, α2–6Sia up-regulated in iPSCs/ESCs has been reported to be overexpressed in many types of human cancers, and its high expression positively correlates with tumor metastasis and poor prognosis (47). Thus, the glycan alterations upon induction of pluripotency observed in this study are apparently similar to those occurring during malignant transformation, as was implied recently (9). Although the reason for this similarity remains to be elucidated, the characteristic glycan changes should be related to the ability of eternal cell proliferation and maintenance, properties common to both cancer cells and pluripotent stem cells.

Glycans are located at the outermost cell surface, where various events take place on the basis of cell-to-cell recognition and interactions. Endogenous lectins, major counterpart molecules of glycans, should play crucial roles in the events (*e.g.* by regulating several signaling pathways). In this context, interactions occurring between cell surface glycans and endogenous lectins are considered to be essential for the maintenance of pluripotency, self-renewal, and differentiation of iPSCs/ESCs (48). Indeed, heparan sulfate proteoglycans were reported to regulate self-renewal and pluripotency of embryonic stem cells (49). Moreover, reduced sulfation on heparan sulfate and chondroitin sulfate were demonstrated to direct neural differentiation of mouse ESCs and human iPSCs (50). Recently, synthetic substrates recognizing cell surface glycans were reported to facilitate the long term culture of pluripotent stem cells (48). Thus, global analysis of the cellular glycomes of iPSCs and ESCs performed in this study will be necessary to provide the basis to explore the functions and applications of the stem cell glycobiology. They includes rational design of the effective substrates and culture conditions to support the long term propagation of ESCs and iPSCs (48). Of course, the results obtained in this study could also be readily applied to staining (specification of the place the event occurs), enrichment (*e.g.* lectin-aided capturing of necessary cells), and targeting of specific cells (*e.g.* elimination of unwanted undifferentiated cells). In this regard, stem cell glycoengineering with the aid of a lectin microarray is a key issue in realization of regenerative medicine in the near future.

Acknowledgments—We thank Yoshiko Kubo and Jinko Murakami for help in lectin microarray production, Dr. Atsushi Kuno and Dr. Yoko Itakura for advice, and Dr. Jun Iwaki for providing galectin expression vectors.

REFERENCES

1. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007) *Cell* **131,** 861–872
2. Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2007) *Science* **318,** 1917–1920
3. Gagneux, P., and Varki, A. (1999) *Glycobiology* **9,** 747–755
4. Varki, A. (1993) *Glycobiology* **3,** 97–130
5. Muramatsu, T., and Muramatsu, H. (2004) *Glycoconj. J.* **21,** 41–45
6. Schopperle, W. M., and DeWolf, W. C. (2007) *Stem Cells* **25,** 723–730
7. Natunen, S., Satomaa, T., Pitkanen, V., Salo, H., Mikkola, M., Natunen, J., Otonkoski, T., and Valmu, L. (2011) *Glycobiology*, in press
8. Lanctot, P. M., Gage, F. H., and Varki, A. P. (2007) *Curr. Opin. Chem. Biol.* **11,** 373–380
9. Liang, Y. J., Kuo, H. H., Lin, C. H., Chen, Y. Y., Yang, B. C., Cheng, Y. Y., Yu, A. L., Khoo, K. H., and Yu, J. (2010) *Proc. Natl. Acad. Sci. U.S.A.* **107,** 22564–22569
10. Hirabayashi, J. (2008) *J. Biochem.* **144,** 139–147
11. Wearne, K. A., Winter, H. C., and Goldstein, I. J. (2008) *Glycoconj. J.* **25,** 121–136
12. Wearne, K. A., Winter, H. C., O'Shea, K., and Goldstein, I. J. (2006) *Glycobiology* **16,** 981–990
13. Pilobello, K. T., Krishnamoorthy, L., Slawek, D., and Mahal, L. K. (2005) *Chembiochem* **6,** 985–989
14. Kuno, A., Uchiyama, N., Koseki-Kuno, S., Ebe, Y., Takashima, S., Yamada, M., and Hirabayashi, J. (2005) *Nat. Methods* **2,** 851–856
15. Uchiyama, N., Kuno, A., Tateno, H., Kubo, Y., Mizuno, M., Noguchi, M., and Hirabayashi, J. (2008) *Proteomics* **8,** 3042–3050
16. Matsuda, A., Kuno, A., Ishida, H., Kawamoto, T., Shoda, J., and Hirabayashi, J. (2008) *Biochem. Biophys. Res. Commun.* **370,** 259–263
17. Ebe, Y., Kuno, A., Uchiyama, N., Koseki-Kuno, S., Yamada, M., Sato, T., Narimatsu, H., and Hirabayashi, J. (2006) *J. Biochem.* **139,** 323–327
18. Pilobello, K. T., Slawek, D. E., and Mahal, L. K. (2007) *Proc. Natl. Acad. Sci. U.S.A.* **104,** 11534–11539
19. Hsu, K. L., Pilobello, K. T., and Mahal, L. K. (2006) *Nat. Chem. Biol.* **2,** 153–157
20. Tateno, H., Uchiyama, N., Kuno, A., Togayachi, A., Sato, T., Narimatsu, H., and Hirabayashi, J. (2007) *Glycobiology* **17,** 1138–1146
21. Krishnamoorthy, L., Bess, J. W., Jr., Preston, A. B., Nagashima, K., and Mahal, L. K. (2009) *Nat. Chem. Biol.* **5,** 244–250
22. Kuno, A., Kato, Y., Matsuda, A., Kaneko, M. K., Ito, H., Amano, K., Chiba, Y., Narimatsu, H., and Hirabayashi, J. (2009) *Mol. Cell Proteomics* **8,** 99–108
23. Narimatsu, H., Sawaki, H., Kuno, A., Kaji, H., Ito, H., and Ikehara, Y. (2010) *FEBS J.* **277,** 95–105
24. Matsuda, A., Kuno, A., Kawamoto, T., Matsuzaki, H., Irimura, T., Ikehara, Y., Zen, Y., Nakanuma, Y., Yamamoto, M., Ohkohchi, N., Shoda, J., Hirabayashi, J., and Narimatsu, H. (2010) *Hepatology* **52,** 174–182
25. Tateno, H., Kuno, A., Itakura, Y., and Hirabayashi, J. (2010) *Methods Enzymol.* **478,** 181–195
26. Kuno, A., Itakura, Y., Toyoda, M., Takahashi, Y., Yamada, M., Umezawa, A., and Hirabayashi, J. (2008) *J. Proteomics Bioinform.* **1,** 68–72
27. Toyoda, M., Yamazaki-Inoue, M., Itakura, Y., Kuno, A., Ogawa, T., Yamada, M., Akutsu, H., Takahashi, Y., Kanzaki, S., Narimatsu, H., Hirabayashi, J., and Umezawa, A. (2011) *Genes Cells* **16,** 1–11
28. Saito, S., Onuma, Y., Ito, Y., Tateno, H., Toyoda, M., Akutsu, H., Nishino, K., Chikazawa, E., Fukawatase, Y., Miyagawa, Y., Okita, H., Kiyokawa, N., Shimma, Y., Umezawa, A., Hirabayashi, J., Horimoto, K., and Asashima, M. (2010) *The Fourth International Conference on Computational Systems Biology (ISB2010), Suzhou, China, September 9–11, 2010*, pp. 381–388
29. Nishino, K., Toyoda, M., Yamazaki-Inoue, M., Makino, H., Fukawatase, Y., Chikazawa, E., Takahashi, Y., Miyagawa, Y., Okita, H., Kiyokawa, N., Akutsu, H., and Umezawa, A. (2010) *PLoS One* **5,** e13017
30. Cui, C. H., Miyoshi, S., Tsuji, H., Makino, H., Kanzaki, S., Kami, D., Terai, M., Suzuki, H., and Umezawa, A. (2011) *Hum. Mol. Genet* **20,** 235–244
31. Tsuji, H., Miyoshi, S., Ikegami, Y., Hida, N., Asada, H., Togashi, I., Suzuki, J., Satake, M., Nakamizo, H., Tanaka, M., Mori, T., Segawa, K., Nishiyama, N., Inoue, J., Makino, H., Miyado, K., Ogawa, S., Yoshimura, Y., and Umezawa, A. (2010) *Circ. Res.* **106,** 1613–1623
32. Kawamichi, Y., Cui, C. H., Toyoda, M., Makino, H., Horie, A., Takahashi, Y., Matsumoto, K., Saito, H., Ohta, H., Saito, K., and Umezawa, A. (2010) *J. Cell Physiol.* **223,** 695–702
33. Makino, H., Toyoda, M., Matsumoto, K., Saito, H., Nishino, K., Fukawatase, Y., Machida, M., Akutsu, H., Uyama, T., Miyagawa, Y., Okita, H., Kiyokawa, N., Fujino, T., Ishikawa, Y., Nakamura, T., and Umezawa, A. (2009) *Exp. Cell Res.* **315,** 2727–2740
34. Suemori, H., Yasuchika, K., Hasegawa, K., Fujioka, T., Tsuneyoshi, N., and Nakatsuji, N. (2006) *Biochem. Biophys. Res. Commun.* **345,** 926–932
35. Tateno, H., Nakamura-Tsuruta, S., and Hirabayashi, J. (2007) *Nat. Protoc.* **2,** 2529–2537
36. Tateno, H., Mori, A., Uchiyama, N., Yabe, R., Iwaki, J., Shikanai, T., Angata, T., Narimatsu, H., and Hirabayashi, J. (2008) *Glycobiology* **18,** 789–798
37. Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J.,

Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X. H., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q. Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W., Jr. (2006) *Nat. Biotechnol.* **24,** 1151–1161

38. Nagata, S., Toyoda, M., Yamaguchi, S., Hirano, K., Makino, H., Nishino, K., Miyagawa, Y., Okita, H., Kiyokawa, N., Nakagawa, M., Yamanaka, S., Akutsu, H., Umezawa, A., and Tada, T. (2009) *Genes Cells* **14,** 1395–1404
39. Gupta, G., Surolia, A., and Sampathkumar, S. G. (2010) *OMICS* **14,** 419–436
40. Nakamura-Tsuruta, S., Kominami, J., Kamei, M., Koyama, Y., Suzuki, T., Isemura, M., and Hirabayashi, J. (2006) *J. Biochem.* **140,** 285–291
41. Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008) *Nat. Biotechnol.* **26,** 101–106
42. Miyazaki, T., Futaki, S., Hasegawa, K., Kawasaki, M., Sanzen, N., Hayashi, M., Kawase, E., Sekiguchi, K., Nakatsuji, N., and Suemori, H. (2008) *Biochem. Biophys. Res. Commun.* **375,** 27–32
43. Satomaa, T., Heiskanen, A., Mikkola, M., Olsson, C., Blomqvist, M., Tiittanen, M., Jaatinen, T., Aitio, O., Olonen, A., Helin, J., Hiltunen, J., Natunen, J., Tuuri, T., Otonkoski, T., Saarinen, J., and Laine, J. (2009) *BMC Cell Biol.* **10,** 42
44. Grahn, E., Askarieh, G., Holmner, A., Tateno, H., Winter, H. C., Goldstein, I. J., and Krengel, U. (2007) *J. Mol. Biol.* **369,** 710–721
45. Sulák, O., Cioci, G., Delia, M., Lahmann, M., Varrot, A., Imberty, A., and Wimmerová, M. (2010) *Structure* **18,** 59–72
46. Bremer, E. G., Levery, S. B., Sonnino, S., Ghidoni, R., Canevari, S., Kannagi, R., and Hakomori, S. (1984) *J. Biol. Chem.* **259,** 14773–14777
47. Zhuo, Y., and Bellis, S. L. (2011) *J. Biol. Chem.* **286,** 5935–5941
48. Klim, J. R., Li, L., Wrighton, P. J., Piekarczyk, M. S., and Kiessling, L. L. (2010) *Nat. Methods* **7,** 989–994
49. Sasaki, N., Okishio, K., Ui-Tei, K., Saigo, K., Kinoshita-Toyoda, A., Toyoda, H., Nishimura, T., Suda, Y., Hayasaka, M., Hanaoka, K., Hitoshi, S., Ikenaka, K., and Nishihara, S. (2008) *J. Biol. Chem.* **283,** 3594–3606
50. Sasaki, N., Hirano, T., Kobayashi, K., Toyoda, M., Miyakawa, Y., Okita, H., Kiyokawa, N., Akutsu, H., Umezawa, A., and Nishihara, S. (2010) *Biochem. Biophys. Res. Commun.* **401,** 480–486

Discovery of Chemical Compound Groups with Common Structures by a Network Analysis Approach (Affinity Prediction Method)

Shigeru Saito,[†,§] Takatsugu Hirokawa,[†] and Katsuhisa Horimoto*[,†,‡]

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan, Chem & Bio Informatics Department, INFOCOM Corporation, Sumitomo Fudosan Harajuku Building, 2-34-17 Jingumae, Shibuya-ku, Tokyo, 150-0001, Japan, and and Institute of Systems Biology, Shanghai University, 99 Shangda Road, Shanghai 200444, China

Received July 9, 2010

We developed a method in which the relationship between chemical compounds, characterized by the secondary dimensional descriptors by a standard method, is first determined by network inference, and then the inferred network is divided into the compound groups by network clustering. We applied this method to 279 active inhibitors of factor Xa found by the first screening. A large network of 266 active compounds connected with 408 edges emerged and was divided into 10 clusters. Surprisingly, the chemical structures that were common within the clusters, but diverse between them, could be extracted. The activity differences between the clusters provide rational clues for the systematic synthesis of derivatives in the lead optimization process, instead of empirical and intuitive inspections. Thus, our method for automatically grouping the chemical compounds by a network approach is useful to improve the efficiency of the drug discovery process.

1. INTRODUCTION

Novel computational approaches and methodologies are increasing the efficiency of drug discovery, which involves numerous processes.[1] Indeed, various computational approaches in virtual screening are utilized to predict the activity of hypothetical compounds, based on the quantitative structure–activity relationship (QSAR).[2–5] In particular, the selection of compounds from a library or database of compounds is widely used to identify those that are likely to possess a given activity, when a single bioactive reference structure is available.[6–8] In this approach, fingerprint-based similarity searching is performed to identify the database molecules that are most similar to a user-defined reference structure.[9] Furthermore, the support vector machine is utilized to predict the activity of newly synthesized compounds with high accuracy.[10] The principal component analysis (PCA) also presents the relationship between the compounds, to allow a visual investigation of their activities in the principal component space. In particular, it generates a concept for the distribution of chemical compounds, named the chemical space, where different chemical compounds are reasonably distributed, depending on their corresponding origins.[11]

In spite of the popularity of computational approaches, empirical and intuitive approaches are still employed in drug discovery processes.[1] One reason for retaining the empirical and intuitive approaches is that after the first screening, the active compounds are usually compared in terms of the relationship between the chemical structure and its activity, before the next step of synthesizing the derivatives for selecting the ultimate lead. Unfortunately, this step partially depends on the empirical selection of the candidates for the chemical synthesis of the drug target, with reference to the chemical structures of the active and inactive compounds obtained by the first screening. Indeed, the structural information on the active compounds after the first screening is not fully utilized for selecting candidates of seeds for the derivative synthesis. Thus, the extraction of useful information about chemical structure and activity, in an automatic and visual manner, is desirable to systematically and efficiently synthesize derivatives for drug discovery.

We now propose an automatic method to visually group chemical compounds based on their structures, by using two types of network analysis methods. One is the network inference method. In the present study, we use the path consistency algorithm,[12] one of the graphical models from the family of probability models simplified by the conditional independences inherent in the graph,[13] which can visually infer the relationships between variables in a network form. Another is the network clustering method. This is a method to extract one property, named the "community structure", which indicates that the vertices in networks are often clustered into tightly knit groups, with a high density of within-group edges and a lower density of between-group edges.[14] This method is useful to automatically group the variables into some clusters from the connected network structure. Here we utilized the two network analysis methods to assess the relationship between chemical compounds. The utility of the present method is demonstrated by a set of chemical compounds after the first screening. The merits and pitfalls of the present method are also discussed, in terms of the previous computational methods.

* Corresponding author. E-mail: k.horimoto@aist.go.jp.. Telephone: +81 3 3599 8711.

[†] National Institute of Advanced Industrial Science and Technology (AIST).

[§] INFOCOM Corporation.

[‡] Shanghai University.

10.1021/ci100262s
Published on Web 12/09/2010

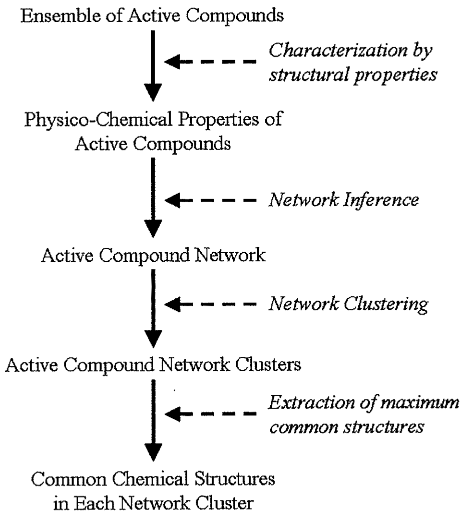

Ensemble of Active Compounds
Characterization by structural properties
Physico-Chemical Properties of Active Compounds
Network Inference
Active Compound Network
Network Clustering
Active Compound Network Clusters
Extraction of maximum common structures
Common Chemical Structures in Each Network Cluster

Figure 1. Workflow of the present method. The present method is schematically described in four steps.

2. MATERIALS AND METHODS

2.1. Overview of the Present Method. An overview of our method is schematically described in Figure 1. First, the chemical compounds selected by the first screening, in terms of drug activity, are characterized by their secondary structure properties, by a standard procedure. Second, the relationships between the compounds are investigated by a network inference method, the path consistency algorithm.[12] Third, the inferred network structures are divided into groups by a network clustering method, the Newman algorithm.[14] Fourth, the maximum common structures of the compounds are extracted in each cluster by a standard method. Thus, the characteristic features of the chemical structures hidden in the active chemical compounds are revealed visually and automatically by network analysis methods. The details of each step are described below.

2.2. Data Set. The data set contains a wide series of inhibitors of factor Xa extracted from the literature, all sharing a benzamidine moiety.[15] The considered data set contains 279 very active compounds (K_i lower than 10 nM) among a total of 435 chemical compounds, also including 156 low-activity compounds (K_i higher than 1 μM).

2.3. Descriptors. The calculated 2D descriptors were derived from the commercially available software, MOE, by Chemical Computing Group Inc. (http://www.chemcomp.com/). As a preprocessing step for the following analyses, the values of each descriptor were standardized by their averages and standard deviations. In this step, the number of descriptors was reduced, by leaving only the continuous values of the descriptors. Finally, 158 descriptors were used.

2.4. Network Inference by Path Consistency (PC) Algorithm with Modifications. The path consistency (PC) algorithm is a network inference method based on the graphical model.[12] The original PC algorithm is composed of two parts: the undirected graph inference by the partial correlation coefficient and the following directed graph generated by using the orientation rule. The present method partially exploits the first part of the PC algorithm, because the aim of the present application of the network inference method is to scrutinize the relationships between the chemical compounds, without the causality.

The algorithm for the first part is simple. The relationship between two variables is tested from the lower partial correlation coefficient to the higher one. For example, the relationship between the two variables is first tested by the zero-th partial correlation coefficient. If the null hypothesis is accepted, i.e., no association between the two variables, then no further test is performed for the higher order of the partial correlation coefficient. If it is rejected, then the relationship between the two variables is tested by the first partial correlation coefficient. In general, the $(m - 2)$-th order of the partial correlation coefficient is calculated between two variables, given $(m - 2)$ variables, i.e., $r_{ij,\mathrm{rest}}$, between X_i and X_j, given the 'rest' of the variables, $\{X_k\}$ for $k = 1, 2, ..., m$, and $k \neq i, j$, and after calculating the $(m - 2)$-th order of the partial correlation coefficient, the algorithm naturally stops. However, the algorithm does not usually request the $(m - 2)$-th order of the partial correlation coefficient for the natural stop. This is because no adjacent variables will be found after excluding the variables, even in the calculation of the lower order of the partial correlation coefficient. We provide the pseudocode of the algorithm in Figure 2.

In the sample data, the zero-th order (i.e., the condition where subset S is empty) of the partial correlation coefficient is calculated by Pearson's correlation coefficient, $r_{ij|S = \phi}$, expressed by

$$r_{ij|S=\varphi} = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)}}$$

where $\mathrm{cov}(X_i, X_j)$ and $\mathrm{var}(X_i)$ are the covariance between X_i and X_j and the variance of X_i. The higher order of the partial correlation coefficients, $r_{ij|S}$, expressed by

$$r_{ij|S} = \frac{-r^{ij}}{\sqrt{r^{ii} \cdot r^{jj}}}$$

where $ij|S$ means $S=\{1, 2, ..., p\}\backslash\{i,j\}$, and r^{ij} is the i-j element of the inverse correlation coefficient matrix.[13] Note that the dimensions of the correlation coefficient matrix are related to the orders of the partial correlation coefficients. The m-th order partial correlation coefficient is calculated from the $(m + 2)$ dimension of the correlation coefficient matrix. The partial correlation coefficient is statistically tested by using the Z-statistic.[16] First, z-transforms of the partial correlation coefficients are calculated, by the following equation:

$$z_{ij} = \frac{1}{2}\ln\left(\frac{1 + |r_{ij|S}|}{1 - |r_{ij|S}|}\right)$$

Then, the z-statistic is obtained from the following equation:

$$Z = \frac{z_{ij}}{\sqrt{1/n - 3 - p}}$$

where n is the number of samples and $p = |S|$ is the conditioning order of the partial correlation coefficient. The z-statistic follows the standard normal distribution, $N(0,1)$, and the significance probability can be set according to this distribution; i.e., we reject the null hypothesis H_0:$r_{ij|s} = 0$, if $Z > Z_{\alpha/2}$with significance level α. If H_0 is not rejected, then

```
Let $Adj(G,X_i) \setminus \{X_j\}$ be the set of nodes (variables) adjacent to $X_i$, except for $X_j$, in the undirected graph $G$.
Let $p$ be the degree of conditioning.
1: $G \leftarrow$ complete undirected graph
2: $p = 0$
3: repeat
4:   for all $X_i$ such that $|Adj(G,X_i)| - 1 \geq p$ do
5:       for all $X_j \in Adj(G,X_i)$ do
6:           for all subset $S \subseteq Adj(G,X_i) \setminus \{X_j\}$ such that $|S| = p$ do
7:               if $X_i \_||\_ X_j \mid S$ then
8:                   delete edge between $X_i$ and $X_j$ in $G$
9:               end if
10:          end for
11:      end for
12:  end for
13:  $p = p + 1$
14: until $|Adj(G,X_i)| - 1 \leq p$, $\forall X$
15: return $G$
Where "$X_i \_||\_ X_j \mid S$" means $X_i$ and $X_j$ are conditionally independent on $S$; i.e., there is no edge between $X_i$ and $X_j$.
```

Figure 2. Pseudocode of the modified path consistency algorithm. A pseudocode of the modified PC algorithm is described. In line seven, statistical hypothesis testing for the partial correlation between X_i and X_j conditioning on S is used to determine whether X_i and X_j are conditionally independent (for details, see text). If the partial correlation cannot be calculated, due to the multicollinearity, then we consider that X_i and X_j are always conditionally dependent on any other variables.

we consider $r_{ij|s} = 0$, and we judge the i-th and j-th nodes as being conditionally independent of S.

The key point in the present network inference is the two modifications of the original PC algorithm, for application to the chemical compounds. The first modification is the correction of the algorithm in the calculation of the partial correlation coefficient. Since many compounds frequently show very similar descriptor values, the difficulty increases in the numerical calculation of the partial correlation coefficients, due to the multicolinearity between the variables. The original PC algorithm accidentally stops if only one partial correlation between a pair of variables violates the numerical calculation, against the high similarity of the descriptors. To avoid the accidental stops by the highly associated compound pairs, the original PC algorithm is modified as follows: If the calculation of any order of the partial correlation coefficient between the variables is violated, then the corresponding pair of variables is regarded as being dependent. The second modification is the correction of the output by the algorithm. The network inference outputs the edges with positive and negative correlations. The edge with a positive correlation in the network can be interpreted as a relationship with direct similarity between the properties of the chemical compound structures, while the edge with a negative correlation indicates a relationship with dissimilarity in a linear fashion. Thus, the edges with the positive correlation are adopted, and those with the negative correlation are excluded from the inferred network.

2.5. Grouping of Chemical Compounds by Network Clustering. In networks, the vertices are often clustered into tightly knit groups, with a high density of within-group edges and a lower density of between-group edges. This property is called a "community structure", and the computer algorithms for identifying the community structure are based on the iterative removal of edges with high "betweenness" scores, which identify such structures with some sensitivity. Here, we applied one of these algorithms to group the chemical compounds in the inferred network.[14]

This method is based on the modularity that is measured by a parameter, the Q-value. The Q-value is defined as follows:

$$Q = \sum (e_{ii} - a_i^2)$$

where e_{ii} means the fraction of edges in cluster i with respect to all edges in the network, and a_i means the fraction of the number of edges that end in cluster i. First, this method considers each node as a cluster. In each subsequent step, two clusters are combined to maximize the increment of the Q-value, ΔQ. ΔQ is calculated as follows:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

where e_{ij} means the half of the fraction of edges between clusters i and j with respect to all edges in the network. In addition to the above definition, e_{ij} is commutative, $e_{ij} = e_{ji}$, in the undirected graph. The complexity of the calculation is on the order $O(N)$, where N is the number of nodes in the network, and we combine two clusters at most $(N - 1)$ times; therefore, in sparse networks, the clustering is complete after $O(N^2)$ times.

2.6. Maximum Common Structures of Clusters. The maximum common structure within the constituent compounds belonging one cluster was obtained by using ChemAxon JKlustor libMCS.[17]

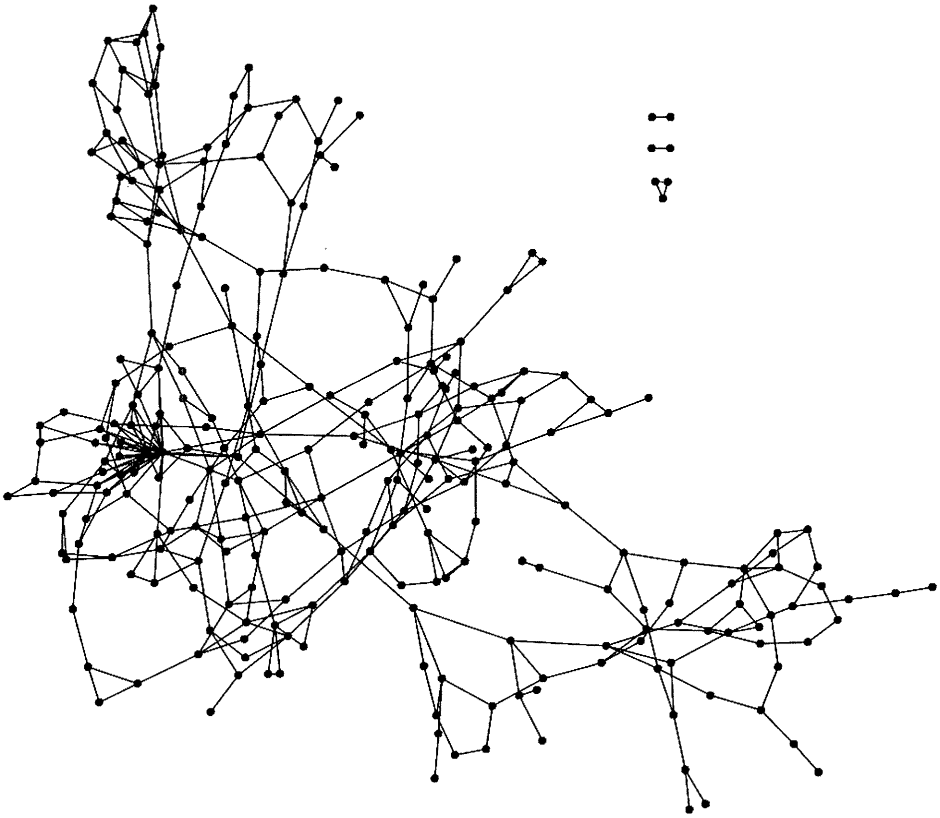


Figure 3. Chemical compound network inferred by path consistency algorithm. A large network of 266 compounds, inferred by the path consistency algorithm with 5% significance probability,[12] is described. The compounds and the established edges between compounds are denoted by open circles and straight lines, respectively.

3. RESULTS AND DISCUSSION

3.1. Chemical Compound Network. The relationships between the 279 active compounds were inferred by the PC algorithm. By the network inference, a large network containing 266 of the 279 active compounds emerged, as shown in Figure 3. Only seven compounds remained apart from the large network, and among them, five edges of the seven compounds were established. The emergence of a large network seems natural, because all of the compounds analyzed in this study share similar physicochemical properties, in terms of drug activity.

The large network contained 408 edges between compounds, and the average connectivity ([number of edges]/[$n(n - 1)/2$], where n is the number of nodes) was about 0.0116. As shown in Figure 3, the inferred network was relatively sparse, in terms of edge connectivity. Although several hubs were observed in the network, it seems difficult to identify clear relationships between the compounds by visual inspection.

3.2. Chemical Compound Network Clusters. To scrutinize the compound relationships, we applied a network clustering method to rationally rearrange the connectivity in the inferred network of Figure 3. In Figure 4, 10 clusters naturally emerged from the entire connectivity in the inferred large network. Thus, the large, complicated network was transformed into distinctive clusters, with the number of compounds in each cluster ranging from 14 to 41. The emergence of the clusters indicates that some distinctive compound groups with similar structural properties exist in the network. The following step involves the investigation of the constituent compounds of each cluster that emerged by two network analyses, in terms of chemical structure and activity.

3.3. Common Structures of Chemical Compound Network Clusters. We surveyed the structural relationship between the constituent chemical compounds that belong to each cluster in the active network. Interestingly, the structures of the constituent compounds were common within each cluster, and they were diverse between the clusters.

The common structures of the member compounds in the clusters of the active network are shown in Figure 5A. It is readily apparent that common structures were found for all of the clusters, and high densities of the constituent compound structures were present in all of the clusters. Indeed, on average, ca. 63.7% of the compounds shared common structures: the highest and lowest share rates were 100.0% in cluster 6 and 35.5% in cluster 3. In addition, the average density of heavy atoms over all constituent compounds in each cluster was high: 9 of the 10 clusters showed more than 50% of the average density, and the exceptional cases were found in cluster 8. Furthermore, the common structures of each cluster were distinctive between them, as seen in Figure 5A. To estimate the differences between the common structures, the Tanimoto coefficients were calculated between them, as follows:

$$T_{ij} = \frac{\sum_k (X_{ik}X_{jk})}{\sum_k X_{ik}^2 + \sum_k X_{jk}^2 - \sum_k (X_{ik}X_{jk})}$$

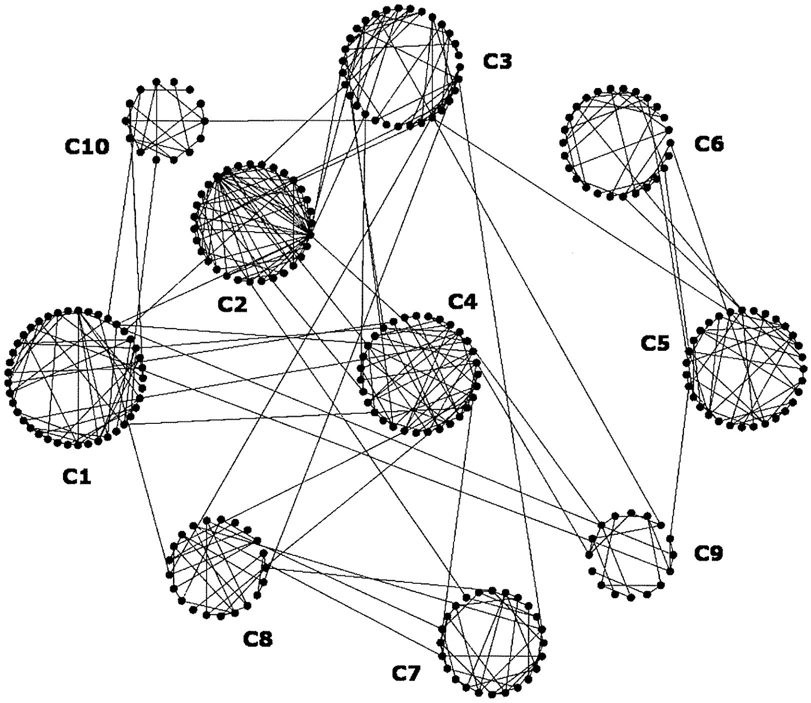


Figure 4. Network clusters of a large network of active chemical compounds. The network clusters estimated for the large network of active compounds in Figure 3 are depicted. Ten clusters emerged, and they are numbered in the order of the numbers of constituent compounds within the clusters.

As shown in Table 1, the Tanimoto coefficients for all pairs of common structures were much less than 0.85, a value that is generally considered to reflect similarity to each other. All of the coefficients were less than 0.4, except for only 0.688 between the common structures of clusters 5 and 6.

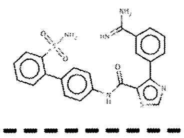
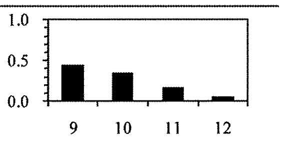
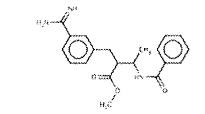
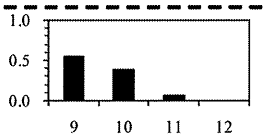
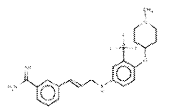
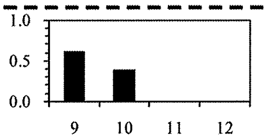
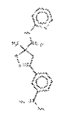
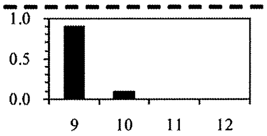
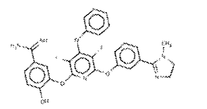
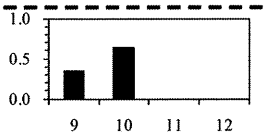
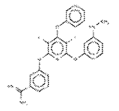
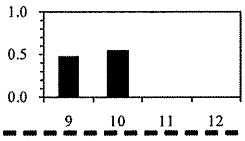
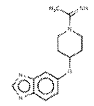
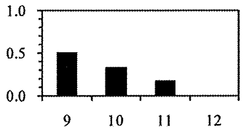
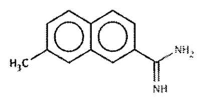
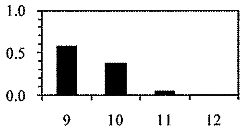
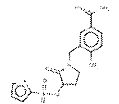
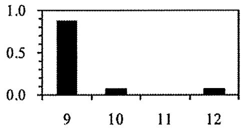
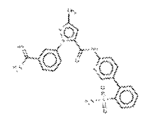
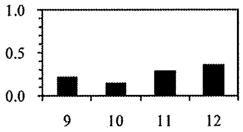
The structures common within clusters and diverse between clusters were further investigated in terms of the activity distribution, expressed by the $-\log(IC_{50})$ histogram of the constituent compounds. For reference, the histogram of all compounds was also drawn in Figure 5B, and in the histogram, the compounds with a $-\log(IC_{50})$ value of less than 9 (10 nM), which is generally regarded as the lead compound, were frequently included (29.3% of compounds). Subsequently, the compounds with IC_{50} values less than 10 nM were frequently observed in the histograms of each cluster in Figure 5A. Interestingly, some exceptions were also observed. A statistical difference between the total IC_{50} distribution in Figure 5B and the distributions in Figure 5A was found in several clusters. In the distributions of clusters 5 and 10, the frequency of observing an IC_{50} value than 10nM was relatively high, in comparison with the total distribution. This indicates that the common structures in clusters 5 and 10 may show a robust IC_{50} for any chemical modification. Thus, the common structure may be a candidate for lead optimization. In contrast, the frequencies of IC_{50} values less than 10 nM in clusters 4 and 9 were much lower than that in the total IC_{50} distribution. This indicates the possibility that many compounds with an IC_{50} activity of less than 10 nM can be synthesized from the common structures of the two clusters. Thus, the correspondence between the common structures and the IC_{50} distributions of each cluster provides some clues for the synthesis of new compounds in the lead optimization process.

3.4. Related Methods. For comparison with the performance of the present method, the PCA was performed for the same data. Figure 6 shows the projection of the cluster members of the active network in Figure 4 into the principal component space. As easily seen in the figure, the cluster members with each common structure are scattered in the space. Indeed, the constituent compounds in each cluster were projected into some duplicated spaces, while the compounds of clusters 5 and 6 were relatively separated from the other clusters in the projected space. As indicated in the preceding subsection, the Tanimoto coefficient between the common structures of clusters 5 and 6 was exceptionally large, and this similarity reflects the common configuration of the constituent compounds in the two clusters in the principal component space. In contrast, the Tanimoto coefficients between the common structures of the other clusters were small, and therefore the compounds were not clearly discriminated in the space. Thus, the PCA may be a low-resolution method to clearly detect the groups with common chemical structures in the data, followed by the first screening.

The fingerprint approach is well-known as another method to detect common structures in an ensemble of compounds.[9] Actually, we used this approach to identify the common structure of the constituent compounds in the respective clusters. As a trial, we applied the fingerprint approach to all of the compounds but were unable to find the common structure (data not shown). We expected this failure from the fact that the common structures of each cluster show much less similarity, as depicted in Figure 5A. In contrast to the PCA, therefore, the fingerprint method may be a high-resolution technique to detect the distinctive groups with common chemical structures.

Note that there are two reasons why we use the Newman method, instead of the standard hierarchical clustering by using the partial correlation matrix as a distance measure. One reason is that the edges in the inferred network are established by considering the higher order of correlation between multiple variables, instead of the distance between

A

Cluster No.	MCS	No. ofCompounds	Density	Distribution
1.		41(18).	0.969.	
2.		31(23).	0.767	
3		31(11).	0.880	
4**		31(17).	0.648	
5*		31(18).	0.953	
6.		26(26).	0.739	
7.		24(21).	0.542	
8		21(16).	0.376	
9*		16(8).	0.809	
10**		14(8).	0.967	

B

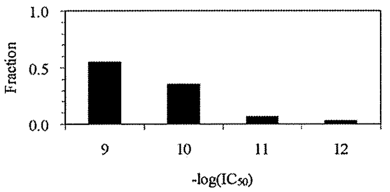


Figure 5. Maximum common structures of the active compound network clusters, together with IC_{50} histograms of constituent compounds. (A) The numbers of clusters in the first column are those described in Figure 4. The common structures of the 10 clusters in the second column were extracted by using ChemAxon JKlustor libMCS.[17] The total number of constituent compounds in each cluster is denoted in the third column, and the number of compounds sharing the corresponding common structures is also denoted in parentheses. In the fourth column, the average densities of heavy atoms in the common structures over the structures of all compounds are denoted. In the fifth column, the histograms of the IC_{50} values of the constituent compounds are depicted: the vertical and horizontal axes are the frequency of the compounds and the $-\log(IC_{50})$ values, respectively. In addition, the differences between each histogram of IC_{50} values for the respective clusters and that for the total active compounds (B) were tested by Fisher's exact test. The significance of the differences between the histograms is indicated at the cluster number in the first column: 5%, '**'; and 10%, '*'.

Table 1. Tanimoto Coefficients between Maximum Common Structures in Respective Active Compound Clusters

cluster no.	1	2	3	4	5	6	7	8	9	10
1	–									
2	0.357	–								
3	0.243	0.254	–							
4	0.261	0.304	0.244	–						
5	0.179	0.197	0.199	0.202	–					
6	0.160	0.166	0.226	0.232	0.686	–				
7	0.187	0.193	0.192	0.123	0.132	0.124	–			
8	0.137	0.224	0.176	0.152	0.126	0.150	0.073	–		
9	0.254	0.271	0.234	0.229	0.213	0.198	0.151	0.157	–	
10	0.213	0.165	0.228	0.274	0.222	0.246	0.142	0.095	0.246	–

pairs of variables in the clustering, The clustering technique in the present study is therefore suitable for keeping the inferred relationships between variables. The other reason is that the Newman method can automatically determine the number of clusters in terms of the network structure. In contrast, the number of clusters is determined by setting a threshold, as in hierarchical clustering, or the cluster number is done before the clustering, as in a self-organization map (SOM).

In summary, the PCA provides a coarse-grinning relationship between compounds from the macroscopic resources, and the fingerprint approach provides a fine relationship between limited ensembles of compounds. With these situations in mind, our procedure provides a medium relationship between compounds, to enrich the selection of molecules with a desired activity. Thus, it bridges the gap between the two methods, by finding the groups of common structures in the step after the first screening, during the process of the lead optimization.

3.5. Merits and Pitfalls of the Present Method. One of the merits of the present method is that it simply detects the structural similarity relationships between active compounds. Indeed, only one parameter, the significance probability in the path consistency algorithm, is set in the network analyses. Thus, the present method is highly automatic and visual, to help reveal a rational synthesis route of chemical compounds for new drug discovery.

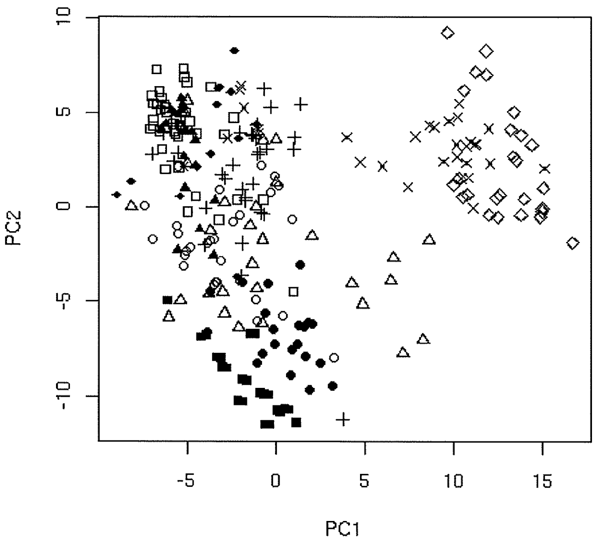


Figure 6. Distribution of members of network clusters in principle component space. The 266 active compounds in the network of Figure 3, which were characterized by the same number of descriptors (158 descriptors) as in the present analysis, were subjected to the PCA. The inertias of the first and second principal components (PC1 and PC2 in the figure) were 0.309 and 0.198, respectively. The constituent compounds of the 10 clusters in Figure 4 are indicated by the following symbols: cluster 1, □; 2, O; 3, Δ; 4, +; 5, ×; 6, ◇; 7, ■; 8, ●; 9, ▲; and 10, ◆.

One of the key points of our method is the application of network inference, based on the graphical model, to the chemical compounds. Among the similar chemical structures, the present network inference detects the 'well-balanced' similarity, by using the partial correlation coefficient. In general, the graphical model distinguishes between real correlation and pseudocorrelation, based on the calculation of a partial correlation coefficient that realizes the concept of conditional independence.[13] The merit of this graphical model is that it only establishes the connection between the compounds with common structures and not between those lacking common structures. This discriminative ability is useful for classifying a large number of active compounds into various groups with different common structures in a rational manner.

In the present analysis, one large network was inferred, and 10 clusters emerged. The numbers of networks and clusters naturally depend on the user-defined descriptors and one parameter in the network inference. In the present analysis, the chemical compounds were characterized by as many secondary structure descriptors as possible. In general, the kinds of descriptors in the analysis may be changed, according to the analyzed data and the analysis aim. Fortunately, the quantification of chemical compounds by descriptors can be easily and quickly performed, due to recent advances in high-performance computing. Although the heuristic choice of descriptors is important to characterize the compound set, the descriptor optimization responsible for the compound set can be included as a preprocessing step in the present work. Furthermore, the size of the network and the following cluster numbers can be controlled by the user-defined significance probability in the network inference. For example, if one chooses a more significant probability than that of the present study, then a smaller network and fewer clusters will be obtained, in which more similar common structures will be found. In addition, the computational time for the present data in the two network analyses was about 5 s, using a personal computer (one CPU with a 2.4 GHz Pentium IV processor and 1GB of memory, under the Linux system). At any rate, the easy manipulation of the data, using only one user-defined parameter, may promote the use of the present method in applications to discriminate between various active compounds in drug discovery.

4. CONCLUSIONS

We have proposed a novel method to group active chemical compounds, by first screening with a combination of two network analysis methods. The scrutinization of active inhibitors of factor Xa by our method revealed reasonable grouping in terms of chemical structure and significant differences between each group in terms of activity. The present results illustrate the possibility that our method will bridge the gap between the compound activity test by the first screening and the following synthesis of lead derivatives.

ACKNOWLEDGMENT

K.H. was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "ystems Genomics" (grant

20016028) and by a Grant-in-Aid for Scientific Research (A) (grant 19201039), from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

REFERENCES AND NOTES

(1) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.

(2) Todeschini, R.; Consonni, V. The Handbook of Molecular Descriptors, in the Series of Methods and Principles in Medicinal Chemistry; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: New York, 2000; Vol. 11.

(3) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835–857.

(4) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043.

(5) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, The Netherlands, 1999.

(6) Bajorath, J. Integration of virtual and high throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

(7) Bajorath, J. Virtual screening: methods, expectations, and reality. *Curr. Drug Discovery* **2002**, 2, 24–28.

(8) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.

(9) Hert, J.; Willett, P.; Wilton, D. J. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(10) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

(11) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.

(12) Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction, and Search (Springer Lecture Notes in Statistics)*, 2nd ed., revised; MIT Press, Cambridge, MA, 2001.

(13) Whittaker, J. *Graphical Models in Applied Multivariate Statistics*; Wiley: New York, 1990.

(14) Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E.* **2004**, *69*, 066133.

(15) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687–2694.

(16) Sokal, R. R.; Rohlf, F. J. *Biometry: The Principles and Practices of Statistics in Biological Research*, 3rd ed.; W. H. Freeman: 1994.

(17) *JChem JKlustor LibMCS*, version 5.3.8; ChemAxon: Budapest, Hungary, 2010.

CI100262S