Development of a Data-mining System for Differential Profiling of Cell Glycoproteins Based on Lectin Microarray

Atsushi Kuno1#, Yoko Itakura1#, Masashi Toyoda2, Yoriko Takahashi3, Masao Yamada1, Akihiro Umezawa2, and Jun Hirabayashi1*

1Research Center for Medical Glycoscience (RCMG), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
2Department of Reproductive Biology and Pathology, National Research Institute for Child Health and Development, 2-10-1, Okura, Setagaya, Tokyo, 157-8535, Japan
3Bioscience Group, Mitsui Knowledge Industry Co., Ltd., Hitotsubashi SI bldg., 3-26, Kandanishikicho, Chiyoda-ku, Tokyo 101-0054, Japan
4Glycomics Research Laboratory, Moritex Corporation, 1-3-3, Azamino-Minami, Aoba-ku, Yokohama City, Kanagawa 225-0012, Japan

# These authors contributed equally to this study

*Corresponding author: Jun Hirabayashi, Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan, Tel: +81-29-861-3124; Fax: +81-29-861-3125; E-mail: jun-hirabayashi@aist.go.jp

Received April 15, 2008; Accepted May 14, 2008; Published May 20, 2008


Copyright: © 2008 Atsushi K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Lectin microarray is an emerging technique enabling multiplex glycan profiling in a direct, rapid and sensitive manner. So far, there has been no robust system available for efficient data-mining to realize differential profiling, which is an effective approach to biomarker investigation. In the present paper, we describe a practical strategy for proteomics-based glycan-related biomarker discovery, with an example of mice embryonal carcinoma and embryonic stem cells and their differentiated forms with retinoic acid. Data were processed by the microarray system using a max-normalization procedure after a gain-merging process, followed by principal component analysis.

Keywords: Differential glycan profiling; Biomarker discovery; Lectin microarray; Principal component analysis

Abbreviations

EC cells: Embryonal Carcinoma cells; ES cells: Embryonic Stem cells; MS: Mass Spectrometry; PCA: Principal Component Analysis; TBSTx: Tris-buffered Saline containing 0.1% Triton X-100.

Introduction

Cell surface dynamics are characterized by altered glycosylation in the development and differentiation stages. Drastic glycosylation change has also been proposed for tumor progression and metastasis. For instance, cell surface sialylation and 1-6 branching of N-linked oligosaccharides are strongly correlated with differentiation of embryonal carcinoma cells and metastatic potential of cancer cells (Dennis et al., 1982; 1987; Hefferman et al., 1993). Therefore, it is highly likely that finding of novel cell differentiation-related or tumor-specific glycoproteins with significant structural changes will become reliable biomarkers. From these points of view, proteomics-based biomarker discoveries have now been complemented by extensive glyco-technologies, such as chemical capturing targeting N-linked glycoproteins (Zhang et al., 2003; Nishimura et al., 2005) and affinity capturing with the use of various glycan-binding proteins, i.e., lectins (dashed arrows in Fig. 1A) (Kaji et al., 2003).
One of the successful reports involving the concept of glycoproteomics includes the discovery of GP73, a novel glycoprotein discovered as a serological biomarker candidate for liver cancer (Block et al., 2005; Drake et al., 2006). Traditionally, serial lectin affinity chromatography (Cummings et al., 1982) has been a procedure for enrichment of particular glycoproteins with a target glycan structure of either N- or O-glycosylation (Madera et al., 2005; Qu et al., 2005). In this case, selection of a highly-effective set of lectins is essential for success in the biomarker discovery process (bold arrows in Fig. 1A). If a systematic data-mining procedure which follows differential glycan analysis were to be available, it would facilitate the design of an optimal set of lectins (bold arrows in Fig. 1A).

Lectin microarray is an emerging technology enabling an ultrasensitive measuring of multiplex lectin-glycan interaction analysis (Angeloni et al., 2005; Pilobello et al., 2005; Kuno et al., 2005). Taking advantage of the merits of this technology, i.e., sensitive detection and simple manipulation, an increasing number of studies using lectin microarray report that cell-surface glycans are closely associated with the functions, states and relation to diseases of individual cells (Ebe et al., 2006; Pilobello et al., 2007; Tateno et al., 2007). Among biological interests in glycans, a current trend is the focus on glycan-related biomarkers. However, there is no established strategy and optimized protocols for cell glycoprotein profiling, in particular regarding data-mining procedures. In this study, we describe logistic processes for differential cell glycoprotein profiling including data-mining as an alternative approach to conventional proteomics-based biomarker discovery. Key points of the strategy for cell glycoprotein profiling include: (1) fitting the protein concentrations in the appropriate range (0.2 and 0.5 µg/ml) to obtain robust and reproducible signal patterns, (2) a gain-merging technique to expand the dynamic range of the lectin-glycan interaction signals, and (3) the max-normalization procedure using the merged data for normalization. The data thus processed were found to be useful for systematic determination of the best set of lectins among more than 40 probe candidate lectins immobilized on the microarray (bold arrows in Fig. 1A). A model study focused on regenerative medicine is described for mice embryonal carcinoma and embryonic stem cells as well as their differentiated forms with retinoic acid.

Results and Discussion

Optimization of lectin microarray manipulations. For improved proteomics-based biomarker discovery, cell glycoproteins are proposed as targets. Glycosylation change is analyzed by a highsensitivity, robust, and reproducible method using lectin microarray, if the evanescent-field fluorescence-assisted detection method is adopted. However, the previous protocol for cell glycoprotein analysis has not fulfilled the recent requirements for detailed cell profiling and biomarker discovery (Ebe et al., 2006). To address these issues, we first established a strict protocol for differential analysis of cell glycoproteins using mouse embryonal carcinoma cells (mouse teratocarcinoma cell line F9) as a model. The analyte (i.e., glycoprotein) was focused on hydrophobic, raft-associated membrane-bound proteins isolated using a CelLytic MEM Protein Extraction kit (Sigma, St. Louis, MO), because we found the proteins to be analyzed showed the highest signal-to-noise ratio. A small aliquot of the obtained protein (200 ng from approximately 1 x 10⁵ cells) was labeled with Cy3-succimidyl ester (designated as Cy3-labeled glycoprotein). Various concentrations of the Cy3-labeled glycoprotein solution (60 µl, 0.02-1.0 µg/ml) were then subjected to the lectin microarray analysis. Due to the specificity of the CCD camera, a gain value should be set so that the observed fluorescence intensities of almost all positive-spots on the glass slide fall within the range 1,000 to 40,000, which provides a dynamic range with sufficient linearity. Each glass slide was successively scanned under different gain conditions. A dose-dependent increment of signal intensity was observed on most of the positive-spots (Fig. 1B). However, we could not confirm satisfactory linearity for all of the spots under a single gain condition. For instance, the signals of some positive-spots (e.g., GSL-I, ECA, SBA, LCA, ConA, TJA-II, and PSA) were kept below 1,000 under the lower gain (80) condition as shown in the top of Fig. 1B. Under the higher gain (100) condition, the intensities of four lectins (DSA, STL, WGA, and LEL) were above the upper limit 40,000, at protein concentrations of 0.2 µg/ml or more (the bottom of Fig. 1B). Such uneven linearity could cause inappropriate interpretation of the data. A useful data-optimization procedure needed to be introduced to solve this basic problem.
intensity (Int\(^i\)) ratios for lectin \(i\) should be almost the same value. To ensure high-reproducibility, the dynamic range was expanded by a “gain-merging” procedure. An outline of the procedure (Fig. 2A) is as follows: a slide glass is scanned under two different gain conditions; higher gain to “rescue” weak signals (e.g., lectin f in Fig. 2A) below 1,000 (Int\(^h\)\(i\)) and lower gain to “suppress” excessively strong signals (e.g., lectin d) over 40,000 (Int\(^l\)\(i\)). At this point, selection of appropriate “merging”-lectins is important (lectins a, b, and e in the case of Fig. 2A) for the signal intensities of which fall within the range 1,000 to 40,000 under both higher and lower gain conditions. With these selected merging lectins, a “Factor (\(F\))” is determined as the average of higher/lower ratios calculated for individual merging lectins by eq (1).

\[
F = \text{Average} (\text{Int}^h_i / \text{Int}^l_i) \ldots \text{eq (1)}
\]

The gain-merging procedure is completed by replacement of the over-range intensities (>40,000) obtained under the higher gain condition (e.g., Int\(^h\)\(i\)) with theoretical intensities (Int\(^T\)\(i\)) by eq (2).

\[
\text{Int}^T_i = \text{Int}^l_i \times F \ldots \text{eq(2)}
\]

For other lectins with no over-range under the higher gain condition, signal intensities obtained under the higher gain condition are used with no modification. During this process, all the resultant intensities of positive-spots were within the expanded dynamic range, from 1,000 to 40,000 x \(F\). When 1.0 \(\mu\)g/ml of F9 cell proteins were subjected to analysis (Fig. 1B), all 34 positive lectins fell within the merged dynamic range (1,000–132,000) after the gain-merging procedure (\(F=3.3\)), whereas 85% (29 lectins under the lower gain (80) condition) and 76% (26 lectins under the higher gain (100) condition) of positive lectins were within the original dynamic range (1,000–40,000), respectively.

Using the merged data, a normalization procedure was developed to simplify and stabilize the subsequent differential glycoprotein analysis. Considering the difficulty in selecting a universal lectin, to assure the same level of signal intensities, we selected a practical procedure to calculate the relative intensity in comparison with the strongest intensity among the positive-spots under the given conditions, i.e., max-normalization. The max-normalized data of F9 cells thus processed gave similar signal patterns provided that protein concentrations were maintained within the range 0.2 to 0.5 \(\mu\)g/ml (Fig. 2B).

A similar observation has also been made for the differentiated forms with retinoic acid (F9-RA) (Fig. 2B). These results suggest that the procedure of max-normalization following gain-merging contribute to the establishment of high-reproducible cell glycoprotein profiling with extremely simple and systematic manipulations.

**Principal component analysis:** We next examined whether or not a statistical analysis of the data could actually determine the best set of lectins, which should be useful for an efficient enrichment of relevant glycoproteins associated with glycosylation change induced by retinoic acid treatment. For this purpose, principal component analysis (PCA) using a web-based NIA array analysis tool (http://lgsun.grc.nia.nih.gov/ANOVA/; Chapman et al., 2001; Sharov et al., 2005), was chosen and applied to the above processed lectin microarray data of F9 cells (four different preparations) as well as F9-RA (three different preparations). For the sake of comparison, we also analyzed mouse embryonic stem cells (mES).
(four different preparations) and their differentiated forms (mES-RA) (two different preparations). The lectin microarray data processed according to the developed procedures gave two principal components (PCs). The 2D-biplot format thus obtained clearly divided the above 13 preparations into four independent clusters; i.e., F9, F9-RA, mES and mES-RA (the upper left of Fig. 2C). The result also revealed

![2D-biplot representation as a result of principal component analysis with gain-merging processing (upper). The data obtained for F9 and mouse embryonic stem cells were processed in comparison with those obtained for their retinoic acid-induced forms. Glycan alterations associated with cell line difference and differentiation induced by retinoic acid are depicted by PC1 and PC2, respectively (left). Lectins that showed dynamic enhancement with retinoic acid treatment were systematically selected as those showing strong double negative-correlation with respect to PC1 and PC2. Relative intensities of the two lectins thus selected (DBA and PHA(L)) toward glycoproteins from F9, mES and their differentiated forms with retinoic acid are represented by bar graphs (right). These data are compared with the principle component analysis using the raw data set without gain-merging processing (bottom).](image)

Conclusion

A lectin microarray-based data-mining system for differential profiling of cell glycoproteins has been developed by adopting max-normalization following gain-merging processes. This highly-reproducible analysis with simple and systematic manipulations should provide the basis of a robust and logistic strategy for the discovery of proteomics-based glycan-related biomarkers.

Acknowledgments

We thank N. Uchiyama, Y. Kubo, and J. Murakami for supplying the lectin microarray. We also thank A. Matsuda for critical discussion concerning the preparation of protein solution. This work was supported in part by a grant for New Energy and Industrial Technology Development Organization (NEDO) in Japan.
References


