Cover Story

Quantitative and focused proteomics for biomarker discovery

Efficient applications of mass spectrometric technology can lead to sensitive and high-throughput identification of clinically useful biomarkers.

Dr Koji Ueda, Laboratory for Biomarker Development, Riken, Japan

In this decade, laboratories worldwide have been equipped with a considerable number of high-end mass spectrometry systems for the purpose of biomarker discoveries. However, acceptable returns in investments have rarely been achieved in the field of serum/plasma protein biomarkers.

Despite the increasing rates of publications on cancer-related biomarkers, the number of US FDA-approved serum/plasma-protein tests is decreasing in the US. Furthermore, almost none of the FDA-approved biomarkers are used in standard clinical practice, and only two of them (α-Fetoprotein, human chorionic gonadotropin-β) have made it into the tumor, nodes, and metastases (TNM) staging guidelines. Disappointingly, none was discovered through the new high-throughput genomic or proteomic technologies.

In particular, neither FDA nor Ministry of Health, Labor and Welfare (MHLW) in Japan has approved any biomarkers for the early diagnosis of cancer. This situation is due in large part to a lack of strategies for efficiently focusing on the specific sub-proteome where the targeted biomarkers are enriched for precise comparison of hundreds of mass spectrometric datasets quantitatively, to increase the detection power of mass spectrometers.

These difficulties can be addressed and this article presents four key technologies to overcome these difficulties based on practical examples.

To date, two major protocols are used for quantitative analysis of mass spectrometric (MS) data sets. One is termed stable isotope labeling technology that uses, for example, a combination of amine-coupling tags incorporated with ^{13}C, ^{18}O, and/or ^{15}N. Although this technology enables simultaneous MS analysis of up to eight samples with relatively uncomplicated computing tools, it is difficult to apply to quantification analysis with hundreds of clinical samples that are necessary for biomarker screening. A second strategy is the label-free quantification technology addressed here. In principle, there is no limitation to the sample number, providing a great advantage especially in the proteomic analysis where many samples or multiple groups must be statistically assessed.

The Genedata Expressionist software platform provides considerable performance in accurate and high-throughput label-free quantification analysis even when handling huge data sets. Particularly, Genedata Expressionist's processing speed and scalability are much better than any other commercial or open source software developed for stand-alone PCs. The Genedata Expressionist platform consists of two modules: Refiner MS for MS data processing and Genedata Analyst for statistical analysis.

In this example workflow, the Genedata

Expressionist Refiner MS module initially constructs the MS chromatogram planes as shown in Figure 1, and subtracts the instrument specific noises and chemical noises effectively. At the fourth step of the workflow in Figure 2, the retention time (RT) grids on each MS chromatogram plane were perfectly aligned among 92 samples (Figure 3), which allowed the solid quantification analysis of multiple samples. Subsequently, peaks were detected from temporarily averaged m/z-RT planes by the Chromatogram Summed Peak Detection Activity to avoid missing peak-location information even if the peaks were not detectable in particular planes. The detected isotopic peaks belonging to the same peptide signals were grouped into individual clusters that were displayed as identically colored rectangles in Figure 1. In this case, a cluster is equivalent to a single peptide. The ion count of each cluster was finally calculated for the next statistical analysis.

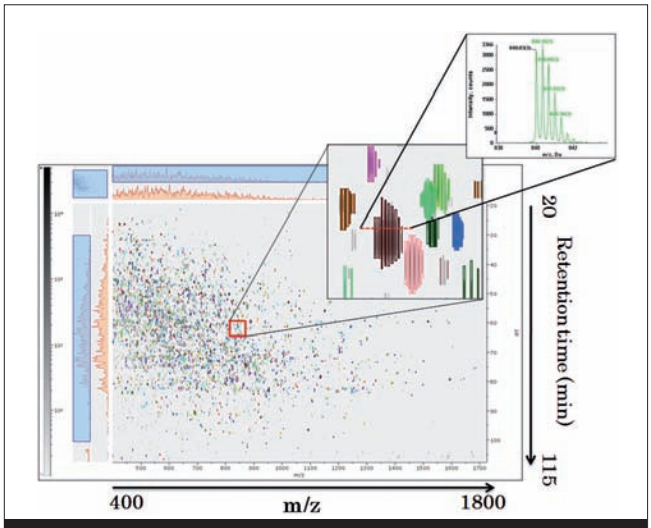


Figure 1: A representative 2D MS chromatogram plane in Genedata Expressionist Refiner MS module.

The quantified proteome information was seamlessly transferred into the Genedata Analyst module and submitted for various statistical analyses, including t-test, ANOVA, multiple testing corrections, principal component analysis (PCA), ROC curve analysis, leave-one-out cross validation test, etc. Genedata Analyst combines user-friendly interfaces with sophisticated graphics and flexibility of output results. These features contribute enormously to rapid and careful consideration of complicated MS data, which enables reliable identification of clinically useful biomarker candidates.

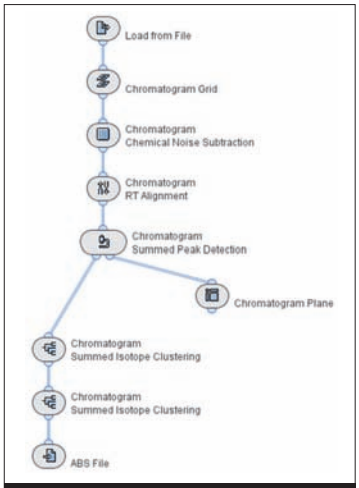


Figure 2: A typical MS data processing workflow in Genedata Expressionist Refiner MS module. The raw MS data from any type of mass spectrometer can be directly loaded.

Glycosylation-targeting strategy: IGEL method

Glycosylation is one of the most important and abundant post-translational modifications in human proteome. Glycoproteins have been used as therapeutic targets and biomarkers for cancer prognosis, diagnosis, and monitoring. As MS-based Systems Biology begins to revolutionize our understanding of biomedical sciences, acquiring glycoproteome profiles efficiently and comprehensively in biological samples (ie body fluids, cell surface) is critical to many biological and clinical researchers.

We recently reported a new approach to identify carbohydrate-targeting serum biomarkers, termed Isotopic Glycosidase Elution and labeling on Lectin column chromatography (IGEL). It is based on glycan structure-specific enrichment of glycopeptides by lectin-column chromatography and site-directed tagging of N-glycosylation sites by ^{18}O during the elution with N-glycosidase. The N-glycosidase elution in the presence of $H2^{18}O$ results in the specific amino acid substitution, asparagine to ^{18}O-asparatic acid residue, with 3-Da increase of the peptide molecular weight.

The mass spectrometric identification of this 3-Da increase enables absolutely accurate determination of N-glycosylation sites in complicated samples such as serum. Furthermore, the combination of IGEL with label-free quantification tools provided by Genedata Expressionist enabled us

Figure 3: A representative area of m/z - retention time planes after RT alignment of 92 LC/MS/MS data. In each panel, three isotopic clusters and grid lines are displayed, showing highly exact alignments.

Cover Story

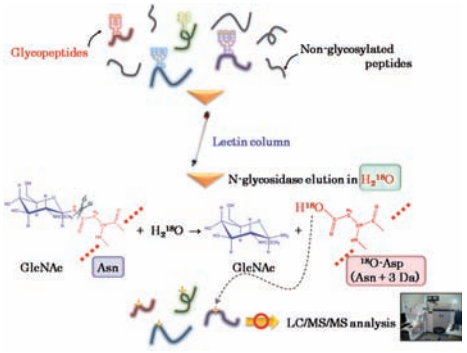
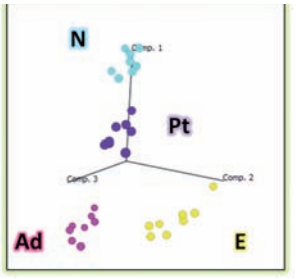


Figure 4: The concept of IGEL technology.



N : Normal control
Pt : Pancreatitis
E : Early stage pancreatic cancer
Ad : Advanced stage pancreatic cancer

Figure 5: The result of principal component analysis using 122 candidate biomarker peptides (n = 32) on the Genedata Expressionist Analyst module.

not only to identify N-glycosylation sites accurately, but also to quantitatively compare glycan structures on each glycosylation site in a single LC-MS/MS analysis. We applied the IGEL method to the comparative analysis of 32 serum samples from eight healthy controls, eight chronic pancreatitis patients, eight stage-0 or I pancreatic cancer patients, and eight stage-III or IV pancreatic cancer patients on the Sambucus sieboldiana agglutinin (SSA) lectin column recognizing α2, 6-linked sialic acid residues.

After detection and quantification of around 10,000 clusters from 32 cases in the Genedata Expressionist Refiner MS module, principal component analysis with Genedata Analyst module revealed that the four groups above were clearly distinguishable using the intensity profiles of only 122 clusters (Figure 5). This result indicated that the glycan structures on these 122 glycosylation sites are potential carbohydrate-targeting biomarkers for pancreatic cancer. By focusing on the glycoproteome of human sera, it has been absolutely feasible to discover potential biomarkers that would allow the early diagnosis of cancer.

CSC technology

Although the cell surface subproteome is responsible for sensing the microenvironment of the cell, little is known about the complex array of proteins involved in these processes and their signaling networks. In recent years, research efforts to uncover the cell surface proteome and its signaling mechanisms revealed a sandbox full of individual proteins including a myriad of functional observations buried in scientific publications. These observations include that cell-surface proteins are important for cell-cell communication, interaction with pathogens, binding of chemical messengers, migration, adhesion, and cell survival. Most cell types known so far can be classified through the serial identification of cell surface epitopes, using a limited set of ~350 anti-CD (cluster of differentiation) antibodies. The cell-surface capturing (CSC) technology enables the multiplexed and relative quantification of hundreds of cell-surface exposed proteins and provides a bird's eye view of the cell-surface proteome without antibodies. The strategy for selective chemical tagging of cell-surface glycoproteins on the intact living cell, followed by high-affinity enrichment and LCMS/MS analysis of peptides derived from the tagged proteins is illustrated in Figure 6.

Specific steps of this tandem affinity-labeling strategy include:

1. gentle covalent chemical labeling of oxidized carbohydrate containing proteins on live cells using the bi-functional linker molecule biocytin hydrazide (BH);
2. affinity enrichment of BH-labeled peptides;
3. specific enzymatic peptide release that permits systematic and selective identification of N-linked glycosylation sites from the surface glycoproteins; and
4. subsequent peptide and protein identification by means of LC/MS/MS.

The key difference when compared with other published

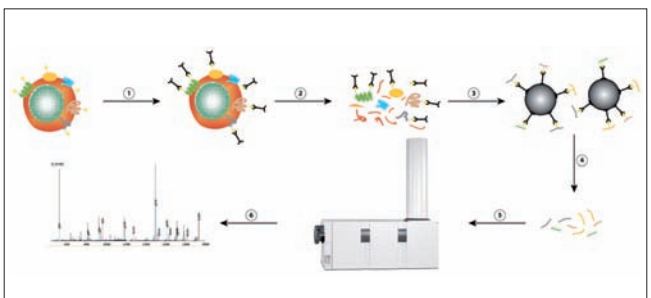


Figure 6: The CSC technology utilizes a three-step tandem affinity labeling strategy to confer the desired specificity for the proteins in the plasma membrane. Quantification of peptides can be achieved by combining the CSC technology with label or label-free peptide quantification strategies.

approaches, and which leads to the superior selectivity of the present method for cell surface proteins, is the oxidation of cell-surface polysaccharides on living cells combined with subsequent BH labeling. Simultaneous identification of the cell-surface subproteome provides a frozen snapshot of the functional capabilities of the cell for sensing the environment at any chosen experimental time-point. A comparison of several time-points using the CSC technology can reveal relative quantitative changes within the cell-surface proteome, for example, during carcinogenesis or metastatic transition of the cells.

Peptidome profiling technology

To date, more than 500 proteases/peptidases are known to be expressed by human cells. They function at almost all locations in the body including intracellular region, extracellular matrices, and in blood. They are involved in activation of other protein functions, degradation of cellular proteins, and notably tumor progression or suppression. Indeed many matrix metalloproteases are overexpressed on various types of tumor cells that facilitate construction of favorable micro-environment for tumor cells or promotion of metastasis. These protease/ peptidase activities definitely result in the production of digested peptide fragments reflecting the tumor progression or tumor-associated responses. Thus, peptidomic profiling of human serum or plasma is a promising tool for the discovery of novel tumor markers. However, methods allowing both detection of thousands of low- molecular weight components with simple procedures and quantitative differential analysis using around 100 serum samples have yet to be established.

Recently we developed a new approach to identify lung cancer specific serum peptide biomarkers. It is based on the one-step effective enrichment of peptidome fractions (1,000 < molecular weight < 5,000) with size exclusion chromatography and the precise label-free quantification analysis of LC/MS/MS data set with Genedata Expressionist proteomic data analysis platform (Figure 7). We applied this method to 92 sera (30 healthy controls and 62 lung adenocarcinoma patients), and finally identified 118 peptides that showed significantly altered serum levels between the control and lung cancer groups ($p < 0.01$, fold change > 5.0). Among them, we identified the peptide sequences of 19 peptides by MS/MS analysis and further confirmed their diagnostic powers by MRM-based relative quantification analysis on 96 additional serum samples. Hence, the peptidome profiling technology can provide simple, high-throughput, and reliable quantification of a large number of clinical samples, which is applicable for diverse peptidome-targeting biomarker discoveries using any types of biological specimens.

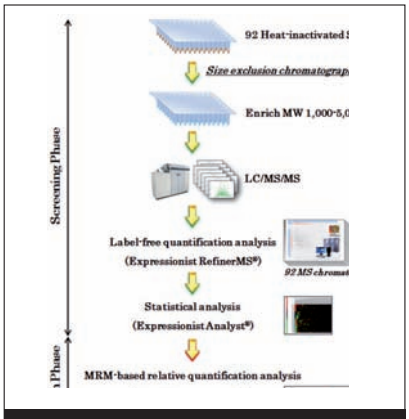


Figure 7: Schematic view of peptidome biomarker development workflow.

There are a wide range of diseases that could greatly benefit from the availability of effective protein biomarkers. Despite this need and potential patient benefit, the use of new protein biomarkers by regulatory agencies has been limited. Four key proteomics technologies described above seek to reverse this trend and enable more reasoned decision-making in drug development and in patient care. These technologies make use of varied approaches to creating a more complete biological description and understanding.

A common theme across the technologies is their collective ability to take advantage of the increasingly powerful array of high-end mass spectrometry systems on the market. The combination of improved mass spectrometry systems, focused proteomics technologies and sophisticated label-free quantification analysis tools deliver reproducible quantitative comparisons across large sample sizes that have previously proven elusive. This ability to accurately compare hundreds to thousands of biological samples enables in-depth screening for novel clinical biomarkers with an extraordinary dynamic range of concentration such as that seen in serum and plasma. With large population analyses this also provides for realistic ranges of variability and thus the key criteria for moving from biomarker discovery to further development and approved usage.

Finally, given the variety of approaches, these techniques could be tailored to many diagnostic and pharmaco-dynamic purposes to be used for comprehensive interpretations of systems biology in body fluids or tissues. **PA**