Label-free Quantitative Proteomics Using Large Peptide Data Sets Generated by Nanoflow Liquid Chromatography and Mass Spectrometry*

Masaya Ono‡, Miki Shitashige‡, Kazufumi Honda‡, Tomohiro Isobe§, Hideya Kuwabara§, Hirotaka Matsuzuki§, Setsuo Hirohashi‡, and Tesshi Yamada‡¶

We developed an integrated platform consisting of machinery and software modules that can apply vast amounts of data generated by nanoflow LC-MS to differential protein expression analyses. Unlabeled protein samples were completely digested with modified trypsin and separated by low speed (200 nl/min) one-dimensional HPLC. Mass spectra were obtained every second using the survey mode of a hybrid Q-TOF mass spectrometer and displayed in a two-dimensional plane with m/z values along the x axis, and retention time was displayed along the y axis. The time jitter of nano-LC was adjusted using newly developed software based on a dynamic programming algorithm. The comprehensiveness (60,000–160,000 peaks above the predetermined threshold detectable in 60–120 µg cell protein samples), reproducibility (average coefficient of variance of 0.35–0.39 and correlation coefficient of over 0.92 between duplicates), and accurate quantification with a wide dynamic range (over 10^3) of our platform warrant its application to various types of experimental and translational proteomics. Molecular & Cellular Proteomics 5:1338–1347, 2006.

Because of the large diversity in the physical and chemical characteristics of proteins no single platform capable of analyzing the entire protein content (or proteome) of complex biological specimens, such as tissue extracts, cell lysates, blood plasma/serum, and other body fluids, is currently available. For example, two-dimensional gel electrophoresis, a widely used proteome platform, is inadequate for analysis of high molecular weight, hydrophobic, or highly acidic/basic proteins (1, 2). So-called "shotgun" proteomics is an emerging concept that has been developed to cope with this problem (3, 4). Protein sample is enzymatically digested into a large array of small peptide fragments (or peptide array) (5).

Although each peptide fragment of peptide arrays represents the relative abundance of its source protein, MS/MS may not be powerful enough as a means of quantification. Because selection of precursor ions for MS/MS may not be constant from experiment to experiment, low abundance proteins can be easily overlooked (10). Several types of in vivo metabolic and in vitro chemical and enzymatic isotope labeling methods including ICAT, SILAC (stable isotope labeling by amino acids in cell culture), and iTRAQ (isobaric tagging for relative and absolute quantitation) have been developed to add a quantitative dimension to MS/MS (11–13). However, in vivo labeling cannot be used for clinical samples, and the efficiency of labeling cannot be matched completely in a large number of samples.

Several attempts have made to quantify peptides generated from unlabeled protein samples by LC-MS instead of LC-MS/MS (5, 14, 15) because there is a linear correlation between MS signal intensities and the relative quantity of peptides (14, 16). However, comparison of different LC-MS data sets is still challenging because of the unsteady flow of LC (17). Li et al. (10) recently tried to overcome this problem by developing a new software suit, namely SpecArray, that is capable of comparing multiple LC-MS data by aligning LC flows. They purified N-glycosylated proteins prior to LC-MS to reduce the sample complexity to a level capable of being managed by their software (5).

The dynamic programming algorithm has been used for comparing large, similar DNA sequences with high accuracy and speed (18); it may be applicable to the alignment of large LC-MS data sets obtained at slightly different LC flows. We also developed a series of software modules suitable for the detection, visualization, quantification, and comparison of LC-MS data generated from unlabeled and unpurified protein samples. The refinement of the nanoflow HPLC system was also necessary to achieve high sensitivity and reproducibility. The high mass accuracy and constancy of Q-TOF MS instru-
ments can eliminate mismatching by narrowing the mass tolerance (16). We integrated all of the machinery and software elements into a new proteomic platform called 2-dimensional image-converted analysis of liquid chromatography and mass spectrometry (2DICAL).1

EXPERIMENTAL PROCEDURES
Cell Culture and Sample Preparation

The colorectal cancer cell clone capable of inducing an actin-binding protein, actinin-4, under the strict control of the tetracycline-regulatory promoter system (DLD1 Tet-Off ACTN4) has been described previously (19). Induction of actinin-4 has been found to significantly increase cell motility and cause lymph node metastasis in experimental animals. Pancreatic cancer cell lines BxPC3 and Capan-1 have been described previously (20). BxPC3 is a cell line with high cell motility, and Capan-1 is a cell line with low cell motility. Cell lysates were prepared with buffer containing 0.01 M Tris-HCl, pH 7.4, 0.14 M NaCl, 1% Triton X-100, and a protease inhibitor mixture (Sigma). A blood specimen was collected from a healthy volunteer.

Plasma was obtained by centrifugation at 3000 rpm for 30 min and cryopreserved at −80 °C until analyzed (21).

To each 100 μl of cell lysate (3 mg/ml) or 5 μl of plasma, 900 μl of cold acetone (−20 °C) was added, and the samples were maintained at −20 °C for 20 min. After centrifugation at 17,400 × g for 10 min, the pellet was dissolved in 250 μl of distilled water and reprecipitated with cold acetone. Then 10 μl of 5 M urea, 2.5 μl of 1 m NH4HCO3, and 3.3 μg of sequencing grade modified trypsin (Promega, Madison, WI) were added, and a final volume of 50 μl was achieved by adding distilled water. After digesting at 37 °C for 20 h, peptides were extracted with 50 μl of acetonitrile, dried with a SpeedVac concentrator (Thermo Electron, Holbrook, NY), and then dissolved in 50 μl of 0.1% formic acid.

LC-MS

The splitless nanoflow HPLC system equipped with reversed-phase columns (inner diameter, 0.15 mm; 50 mm long) was constructed in collaboration with KYA (Tokyo, Japan). The columns were packed with the finest grade spherical silica gel chemically bonded with octadecyl groups. The average particle size of the material was 3 μm, and pore size was 120 Å. A 10-μl protein sample was separated at a speed of 200 nl/min with a linear gradient from 0–80% acetonitrile, 0.1% formic acid for 60 min. Mass spectra were acquired with an ESI-Q-TOF mass spectrometer (QTOF Ultima, Waters) in the 250–1600 m/z range every second for 60 min.

Peak Detection

The peak detection software was a modification of MassNavigatorTM (Mitsui Knowledge Industry, Tokyo, Japan), and performs the following steps (Steps 1–3).

Step 1—After base-line compensation for every spectrum, signals with a signal to noise ratio greater than 2 are selected.

Step 2—Mass signals are fitted to the isotope distribution model below, and the monoisotopic molecular weight and ion charge number are calculated (22, 23).

The intensity of isotopic mass was added to the monoisotopic molecular weight, and the summed monoisotopic intensity was the represented value in the next step.

Step 3—When the signals had the same ion charge number within 0.2 m/z in the consecutive spectrum before or after, the signals were grouped. The peak intensity was defined as the sum of ion intensity of the grouped signals, m/z was defined as the monoisotopic molecular weight, and retention time (RT) was defined as the time corresponding to the central gravity of the ion intensity. To eliminate noise, signals with the same m/z appearing in at least two sequential spectra were selected.

Adjustment of LC Time Jitter

To rectify the change in RT a compensatory function was calculated to maximize the correlation coefficient (CC) between the reference (A) and target (B) data (Fig. 1A). To increase the calculation speed and to provide robustness, the maximal value of every 1 m/z in each RT was substituted for the ionic strength of peaks. A dynamic programming algorithm (18, 24) was used to find the path that would yield the optimal correspondence position by using two-dimensional lattice coordinates at each cycle number (ascending RT order) of A and B. We defined coordinate values as

\[ (A) \text{ and target (B) data (Fig. 1B).} \]

Alignment of Peptide Data Generated by LC-MS

To make total ion intensity equal in different sets of data, peak intensity was normalized by multiplying it by the following normalizing coefficient,

\[ V_i = \frac{i}{\bar{i}} \]  

where \( V_i \) is normalizing coefficient, \( i \) is the total ion intensity indicated by the i number, and \( \bar{i} \) is the average ion intensity. The normalization value in this study was set between 0.8 and 1.2.
Alignment of Peptide Data Generated by LC-MS

Fig. 1. Adjustment of LC time jitter with a dynamic algorithm. A, concept underlying compensation of RT changes along the y axis. This algorithm finds a function that compensates for time jitter y’ by calculating y = f(y’) from target to reference. B, dynamic algorithm. The function is the path giving the optimal correspondence position (yellow box in two-dimensional lattice coordinates) in each cycle number (ascending RT order) of A and B. When the gap penalty is 0.5, the coefficient for the correlation between the mass spectra in the i cycle and the j cycle is 0.8, and the coordinate values are L(i - 1, j - 1) = 5, L(i - 1, j) = 3, and L(i, j - 1) = 7; then L(i, j) becomes 6.5. The coordinate values of L in all lattices are calculated in this manner. C, optimal correspondence position. The optimal correspondence position starts at the lattice coordinates (k, l) where

\[
(k, l) = \arg \max L(i, j) \text{ subject to: } \begin{cases}
  i = N, j = 1, ..., M \\
  i = 1, ..., N, j = M
\end{cases}
\]  

(Eq. 6)

When (k, l) is defined as V_o, the lattice coordinate is traced back. Thus

\[
V_i = \arg \max L(V_{prev}), V_{prev} = \begin{cases}
  V_{i-1} - (1, 1) \\
  V_{i-1} - (0, 1) \\
  V_{i-1} - (1, 0)
\end{cases}
\]  

(Eq. 7)

As shown in the example (ex), the next box to be selected has the maximum value among the upper side, left side, and left upper side of the former box. D, spline interpolation. To give the natural curve for the optimal correspondence coordinates, the coordinates arrangement of V’, which gives V_i = V_i + (1, 1), is extracted (the combination in yellow boxes). The function is the spline interpolation to the arrangement of V’.

Peak Matching and Protein Identification

Peaks in different LC-MS runs were matched with a tolerance of ±0.25 m/z and ±0.5 min after alignment of the RT. Matching was confirmed by visual inspection of enlarged two-dimensional views. MS/MS was performed on peaks having an RT and m/z within that range in the preparatory LC run for protein identification. Peptide fragmentation data were analyzed with Mascot software (Matrix Sciences, London, UK).
Immunoblot Analysis

Cell lysates were separated by SDS-PAGE and electroblotted onto polyvinylidene difluoride membranes (Millipore). Anti-hemagglutinin monoclonal antibody was purchased from Roche Applied Science, anti-breast cancer-specific gene-1 (BCSG1) goat polyclonal antibody (C-20) was from Santa Cruz Biotechnology (Santa Cruz, CA), anti-cytokeratin 18 (RCK16) and anti-cytokeratin 19 (BA17) monoclonal antibodies were from Chemicon (Temecula, CA), and anti-β-actin monoclonal antibody (AAN02) was from Cytoskeleton (Denver, CO). The membranes were incubated with horseradish peroxidase-conjugated secondary antibodies and visualized with an enhanced chemiluminescence kit (Amersham Biosciences) (25).

RESULTS

Strategy for Quantitative Proteomics Using 2DICAL—To improve the reproducibility of LC without reducing its separation capacity we developed the nanoflow HPLC system utilizing a splitless direct gradient pump. Concentration of protein samples into small volumes and separation by a low flow rate HPLC with small sized reversed-phase columns significantly increase the sensitivity of ESI-MS (6, 26). In our HPLC system the dead volume was minimized, and the flow rate was reduced to the 50–200 nl/min range, but the elution gradient did not fluctuate significantly during runs of over 3 h (data not shown). Multidimensional LC does not always separate peptides in the same manner among experiments. The complexity of protein samples can be sufficiently reduced by such long runs even by one-dimensional LC separation.

We also eliminated MS/MS and tried to detect all the peptides present in a given sample by high speed survey scanning as frequently as every 1 s. MS/MS scanning is a time-consuming procedure and often overlooks minor protein fragments (10). The automatic selection of precursor ions for MS/MS is not always constant, and such uncontrolled selection is likely to reduce the reproducibility.

Three thousand six hundred mass spectrum data (Fig. 2A) were obtained from a 1-h LC-MS run and converted into a single two-dimensional image with m/z values along the x axis and RT along the y axis (Fig. 2B). Fig. 2C shows a flowchart of the experimental procedures and subsequent data processing (Fig. 2, D–I). Signals with the same m/z appearing in at least two sequential RTs were grouped and considered as a peak (described under “Experimental Procedures”) (Fig. 2, F and G). After normalizing to the total ion intensity of all of the peaks, the intensity of each peak was converted to a log value and expressed as a digital spot image (Fig. 2, H and I). We were able to detect 68,243 independent peaks (spots) in a 60-µg lysate of DLD1 Tet-Off ACTN4 cells cultured in the presence of doxycycline (Dox) (19).

Detection of Actinin-4 as a Differentially Expressed Protein in the Entire Proteome of DLD1 Tet-Off ACTN-4 Cells Cultured in the Absence of Dox—We previously established a colorectal cancer cell line that is capable of inducing an actin-binding protein, actinin-4, under the strict control of the tetracycline-regulatory system (designated DLD1 Tet-Off ACTN-4) (19). Upon removal of Dox from the culture medium, DLD1 Tet-Off ACTN-4 cells showed expression of actinin-4 (Fig. 3A), extended filopodia, and increased their motility (19). In a model experiment we investigated whether 2DICAL could pinpoint actinin-4 as a differentially expressed protein among the entire protein content of DLD1 Tet-Off ACTN-4 cells. Whole-cell lysates of DLD1 Tet-Off ACTN-4 cultured in the presence and absence of Dox were analyzed by 2DICAL in duplicate. We noted LC time jitter among the four runs (Fig. 3B) with a maximum difference in RT of 296 s and an average RT difference among the four runs of 36.0 s (Fig. 3, B and C). However, it was possible to adjust the LC time jitter with the newly developed software based on the dynamic programming algorithm (Figs. 1 and 3D), which had been developed to search sequences for regions of similarity and to align large DNA sequences (18). Because the m/z value (x axis) of each peak did not fluctuate among runs because of the high mass accuracy and constancy of Q-TOF MS, only adjustment of RT (y axis) was necessary.

After the RT adjustment, the average coefficient of variance (CV) of peak intensities between duplicates reached 0.37 (Dox (+)) (Fig. 3E) and 0.37 (Dox (−)), and the dynamic range of peak intensity calculated by our system exceeded 10^3 (Fig. 3E). There were 106 peaks that were more than 10-fold more highly expressed in DLD1 Tet-Off ACTN4 cells in the absence of Dox and 68 peaks that were more than 10-fold more highly expressed in the presence of Dox (Fig. 3F). We further confirmed the differential expression by visual inspection (Fig. 3F) and by repeating the same LC-MS experiment several times (data not shown). MS/MS analysis identified 15 (Fig. 3G) of the 106 peaks as having been derived from actinin-4. The biological significance of the other proteins whose expression was affected by induction of actinin-4 will be described elsewhere.

Detection of Differentially Expressed Proteins in Poorly Motile Capan-1 and Highly Motile BxPC3 Pancreatic Cancer Cells—Finally we compared the protein expression profiles of two pancreatic cancer cell lines, Capan-1 and BxPC3 (20), to determine whether 2DICAL is applicable to comparisons of proteomes with large differences because most of the protein (Fig. 3) as well as mRNA (data not shown) content of DLD1 Tet-Off ACTN-4 cells was unchanged after removing Dox, and that may have simplified the adjustment of the sample-to-sample time jitter. We noted that a significant number of spots were detected equally in Capan-1 cells and BxPC3 cells (Fig. 4A, yellow spots), making it possible to perform time adjustment of similar quality to that for DLD1 Tet-Off ACTN4 cells. In addition, a total of 15,407 spots that were differentially expressed between Capan-1 and BxPC3 were detected (9692 spots that were expressed more abundantly in Capan-1 than in BxPC3 and 5715 spots that were expressed more abundantly in BxPC3 than in Capan-1, p < 0.01, Student’s t test, between triplicates) (Fig. 4B, red and green spots) after subtracting spots that were equally expressed by both (Fig. 4A, yellow spots). A representative differentially expressed pep-
tide is shown in Fig. 4C (red arrow). Peaks were tagged by their intrinsic RT and m/z values, and MS/MS was performed on peaks having the same RT and m/z in the preparatory LC run for protein identification (Fig. 2C). Protein identification and differential expression were confirmed by immunoblotting with available antibodies. Fig. 4D shows representative data: BCSG1 (27), cytokeratin 19, and cytokeratin 18 were expressed more abundantly in the poorly motile Capan-1 cells than in the highly motile pancreatic cancer BxPC3 cells.

**DISCUSSION**

We reviewed various aspects of LC-MS for application to large scale quantitative proteomics and eliminated factors that reduce reproducibility and/or comprehensiveness, such as...
as sample labeling, multidimensional LC separation, and MS/MS (Fig. 2C). However, the key technology of our platform is the application of the dynamic algorithm developed to align large DNA sequence data to the alignment of large peptide peak data generated by nano-LC-MS. Aebersold and colleagues (10, 28) also claimed that RT alignment is necessary to compare different LC-MS data and reported a new software suite. They first reduced the complexity of serum/plasma samples by extracting N-glycoproteins and identified 1000–5000 peaks from one set of LC-MS data. They reported a mean CV of 0.31 over four runs, and ~12% of the peptides were missed by the procedure (10). Wang et al. (14) detected ~3400 molecular ions in 25 human serum samples with a median CV of 25.7%. Our alignment method is capable of analyzing a much larger number of peaks with reasonable computing speed (2 h for the comparative analysis of two
LC-MS experiments). We were able to detect more than 100,000 peaks in unlabeled and unfraccionated protein samples and obtained equivalent but slightly higher CV values (0.35–0.39) than in their studies (Fig. 3E). We confirmed the efficacy of 2DICAL for quantitative protein analysis using two model experiments: an experiment that compared proteomes with small differences (Fig. 3) and another that compared proteomes with large differences (Fig. 4).

Bogdanov and Smith (29) reported two-dimensional display of capillary LC-FTICR analysis in which it was possible to detect more than 100,000 peaks and 1000 proteins in a single run and yielded a dynamic range of peak intensities of $10^3$. We were able to obtain a comparable level of comprehensiveness using an easy-to-use and common MS instrument. Our alignment software seems ideal for comparative analysis of large data sets generated by LC-FTICR-MS. However, reduction of sample complexity in both platforms still seems necessary for detection of low abundance serum or plasma proteins because the concentration ranges of serum/plasma proteins span an estimated >10 orders of magnitude (30). Furthermore, decreased sample complexity significantly reduces computing time and improves the accuracy of peak matching.

We consider the development of 2DICAL to still be in the early stage. Its reliability in regard to matching and quantification of low intensity peaks is not expected to be as high as for high intensity peaks. Because mismatching of peaks has a significant adverse effect on quantification, deliberate effort must be made to eliminate mismatching by visual inspection (Figs. 3F and 4C) and by recalculation. Differential protein expression cannot be identified based on statistical data alone. At a p value of <0.01, one would expect 1% of measured values to appear regulated due to variance in the data alone. Confirmatory reruns of different lots of samples were always necessary before targeted LC-MS/MS. Protein identification and differential expression also need to be confirmed by Western blotting whenever antibodies are available (Fig. 4D). We are now accumulating 2DICAL data to construct a two-dimensional map linking the m/z and RT of peptides to their amino acid sequences.

The amount of data obtained in one LC-MS experiment reached 2 gigabytes. The optimal number of samples for comparing seems to be determined by computer capacity.
Fig. 4. Detection of differentially expressed proteins in poorly motile Capan-1 and highly motile BxPC3 pancreatic cells. A, overlap of spots appearing in Capan-1 cells (red) and BxPC3 cells (green) in the 400–16,000 m/z and 0–40-min range. B, differentially expressed spots in Capan-1 cells (red) and BxPC3 cells (green) in the 400–16,000 m/z and 0–40-min range. The levels of expression of 15,407 peaks (spots) were concluded to differ significantly between the two cell lines ($p < 0.01$, Student’s $t$ test, between triplicates). C, a peptide differentially expressed in Capan-1 cells (left) and BxPC3 cells (right) at 614.0 m/z and 13.0 min. This peptide was subsequently identified as being derived from the product of the BCSG1 by MS/MS. D, immunoblot confirmation of the differential expression of BCSG1, cytokeratin 19, and cytokeratin 18 in Capan-1 cells (left) and BxPC3 cells (right).
We currently use a high performance computer cluster consisting of four processors with a clock speed of 3.6 GHz connected in parallel to overcome this limitation. We conclude that 2DICAL is a simple, high throughput, and highly reproducible platform that will provide a new paradigm of quantitative proteomics. The software will be made available to the scientific community.

* This work was supported by the “Program for Promotion of Fundamental Studies in Health Sciences” conducted by the National Institute of Biomedical Innovation of Japan and the “Third-Term Comprehensive Control Research for Cancer” conducted by the Ministry of Health, Labor and Welfare of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article therefore must be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence should be addressed: Chemotherapy Division, National Cancer Center Research Inst., Tokyo 104-0045, Japan. Tel.: 81-3-3542-2511; Fax: 81-3-3547-6045; E-mail: tyamada@gan2.res.ncc.go.jp.

REFERENCES

1. Herbert, B. (1999) Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis* 20, 660–663.

2. Oh-Ishi, M., Satoh, M., and Maeda, T. (2000) Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins. *Electrophoresis* 21, 1653–1669.

3. Swanson, S. K., and Washburn, M. P. (2005) The continuing evolution of comprehensive Control Research for Cancer” conducted by the Ministry of Biomedical Innovation of Japan and the “Third-Term Comprehensive Studies in Health Sciences” conducted by the National Institute of Biomedical Innovation of Japan and the “Third-Term Comprehensive Control Research for Cancer” conducted by the Ministry of Health, Labor and Welfare of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article therefore must be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence should be addressed: Chemotherapy Division, National Cancer Center Research Inst., Tokyo 104-0045, Japan. Tel.: 81-3-3542-2511; Fax: 81-3-3547-6045; E-mail: tyamada@gan2.res.ncc.go.jp.

We currently use a high performance computer cluster consisting of four processors with a clock speed of 3.6 GHz connected in parallel to overcome this limitation. We conclude that 2DICAL is a simple, high throughput, and highly reproducible platform that will provide a new paradigm of quantitative proteomics. The software will be made available to the scientific community.

* This work was supported by the “Program for Promotion of Fundamental Studies in Health Sciences” conducted by the National Institute of Biomedical Innovation of Japan and the “Third-Term Comprehensive Control Research for Cancer” conducted by the Ministry of Health, Labor and Welfare of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article therefore must be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence should be addressed: Chemotherapy Division, National Cancer Center Research Inst., Tokyo 104-0045, Japan. Tel.: 81-3-3542-2511; Fax: 81-3-3547-6045; E-mail: tyamada@gan2.res.ncc.go.jp.