Label-free Semiquantitative Peptide Feature Profiling of Human Breast Cancer and Breast Disease Sera via Two-dimensional Liquid Chromatography-Mass Spectrometry*

Qinhua Cindy Ru‡§, Luwang Andy Zhu‡, Jordan Silberman‡, and Craig D. Shriver¶

A label-free semiquantitative peptide feature profiling method was developed in response to challenges associated with analysis of two-dimensional liquid chromatography-tandem mass spectrometry data. One hundred twenty human sera (49 from invasive breast carcinoma patients, 26 from non-invasive breast carcinoma patients, 35 from benign breast disease patients, and 10 from normal controls) were repeatedly analyzed using a standardized two-dimensional liquid chromatography-mass spectrometry method. Data were extracted using the novel semiquantitative peptide feature profiling method, which is based on comparisons of normalized relative ion intensities. Hierarchical cluster analyses and principle component analyses were used to evaluate the predicative capability of the extracted data, and results were promising. Extracted data were also randomly assigned to either a training group (65%) or to a test group (35%) for artificial neural network modeling. Models best identified invasive breast carcinomas (212 predictions, 94% accurate) and benign non-neoplastic breast disease (96 predictions, 81.3% accurate). These results suggest that, after further development, the novel method may be useful for large scale clinical proteomic profiling. *Molecular & Cellular Proteomics 5:1095–1104, 2006.

Serum and plasma proteomics may uncover diagnostically useful biomarkers (1). Identification of diagnostic signatures from human fluids via high resolution two-dimensional gel electrophoresis (2-D gel) was first proposed more than 2 decades ago, but the idea was initially given little attention (2–4). Fortunately recent advances have rekindled the interest of researchers (5–7). Although useful (8, 9), the capabilities of 2-D gel are limited. Most proteins detected using this method are high abundance maintenance enzymes. Low abundance proteins, membrane proteins, and proteins with extreme isoelectric points or molecular weights are less frequently identified (10, 11). Relatively low throughput capacity and poor reproducibility also limit the utility of 2-D gel. Several novel approaches have been developed recently in response to these limitations, including SELDI-TOF-MS (12, 13), LC-MS/MS (14–17), and multidimensional liquid chromatography tandem mass spectrometry (18–23) (also referred to as multidimensional protein identification technology (MudPIT)). Proteomic tools have been developed so extensively that high throughput analysis of sample sets is no longer the primary concern when proteomically profiling human serum. The wide dynamic ranges of protein concentrations and the masking of low abundance proteins by high abundance proteins (1) have also been addressed.
through chemical depletion and fraction techniques (15, 18, 24, 25).

Having achieved substantial methodological developments in other areas, computational analysis and interpretation of enormous data sets from LC-MS-based technologies (e.g. 2-D LC-MS/MS or MudPIT) have become prominent concerns (26–31). Protein identification is typically conducted immediately following 2-D LC-MS/MS analyses via protein database searches that match tandem mass spectra to peptide sequences (32–36). Commonly used protein databases include the National Center for Biotechnology Information non-redundant (NCBI) database (37), Swiss-Prot (38), and International Protein Index (IPI) (39), and popular algorithms include those of SEQUEST (40) or Mascot software (41). Although these strategies simplify interpretation of large scale data and although they allow users to focus on identified proteins (42), they do have drawbacks. As noted by Nesvizhskii and Aebersold (43), protein inference problems inherent in current strategies require the attention of researchers (32, 44, 45). These problems include limitations of identification based on a single peptide, difficulties assigning exact peptide sequences to MS/MS spectra, difficulties comparing identification results acquired through different algorithms and/or databases, the absence of post-translation information in current protein databases, resultant difficulties in ascertaining differences between protein isoforms, difficulties integrating protein identification results and transcription data (i.e. DNA microarray data or RNA sequencing), and difficulties conducting quantitative proteomic profiling without isotope labeling.

Several approaches to overcoming these obstacles have been investigated recently. Clustering the multiple tandem mass spectra of peptides into one spectrum has been attempted (58), but this technique has yet to be perfected. One intensive peptide profiling study utilized combinations of liquid chromatography elution times and mass values as profiling signatures (58). Li et al. (59) recently developed an LC-MS-based method and adjunct data analysis software. They have focused on semiquantitative profiling of low abundance peptides that are ignored in tandem scans due to intensity discrimination, and their method has been successfully used to profile 10 mice sera.

Despite this progress, there is still a need for a method that can be applied to larger data sets for clinical profiling and that is capable of revealing clinically relevant features. We have developed a label-free semiquantitative feature peptide profiling method that may offer these capabilities. The method was evaluated through analysis of 2-D LC-MS data collected from 120 human sera of breast disease patients, breast carcinoma patients, and normal persons.

### EXPERIMENTAL PROCEDURES

#### Sample Collection

Human blood specimens were collected from volunteer patients who were enrolled in the Clinical Breast Care Project (founded by the United States Department of Defense). All patients were treated at Walter Reed Army Medical Center. Fresh blood was placed in vials marked with patient numbers and barcodes and placed on ice immediately thereafter. Plasma and serum fractions were then prepared according to the Clinical Breast Care Project standard operation protocol, which was approved by institutional review boards of Walter Reed Army Medical Center and the United States Department of Defense. Fresh blood, serum, and plasma samples were shipped in dry ice and delivered overnight to Windber Research Institute (Windber, PA). Upon arrival, specimens were immediately divided into aliquots of each sample and marked with patient numbers and barcodes. All operations were performed on ice, and all vials were labeled in advance with sample numbers and barcodes. 49 invasive breast carcinoma sera, 13 ductal carcinoma in situ (DCIS) sera, 13 atypical hyperplasia sera, 35 benign breast disease sera, and 10 healthy control sera were collected and processed randomly (shown in Table I and Supplemental Table 1S). No thawed sera were utilized.

#### Sample Digestion

All specimens were digested in accordance with the aforementioned standard operation protocol (60). 10–µl aliquots of each specimen (totaling ~0.5 mg of proteins) were first denatured with 25 µl of 2,2,2-trifluoroethanol (Catalog Number T-8132, Sigma) and 15 µl of 166 mM ammonium bicarbonate (Catalog Number A6141, Sigma) at 90 °C for 1 h. 2.5 µl of 200 mM DL-1,4-dithiothreitol (99%, Catalog...
Semiquantitative Peptide Feature Profiling of Sera

Fig. 1. HCAs of SPFP-extracted data. Different disease stages are differently colored. Atypical hyperplasia (AT) is in orange, benign non-neoplastic breast disease (BN) is in light green, benign neoplastic breast disease (BP) is in dark green, ductal carcinoma in situ (IS) is in pink, invasive breast carcinoma (IV) is in blue, and normal control (NR) is in green.

Number 165680050, ACROS Organics, Geel, Belgium) were then added, and the mixture was kept at room temperature for 1 h. Then 12.5 μl of 200 mM iodoacetamide (98%, Catalog Number 122270250, ACROS Organics) were added. The reaction was run at room temperature in the dark for an additional hour. 2.5 μl of 200 mM DL-1,4-dithiothreitol were again added to react with the remaining iodoacetamide. The mixture was allowed to react for an additional hour at room temperature. 300 μl of distilled water and 100 μl of 100 mM ammonium bicarbonate were then added to achieve a pH of 7.5–8.0. Finally 10 μg/25 μl modified trypsin (Catalog Number V511A, Promega, Madison, WI) was added. Digestion was conducted at 58 °C for 1 h or at 37 °C overnight. After digestion, 2 μl of formic acid were added to halt the reaction. The digested sample was diluted 20 times in preparation for MudPIT analyses.

2-D LC-MS Analysis

HPLC grade acetonitrile (part number 9017-02, J. T. Baker Inc.), HPLC grade water (Part Number 4218-02, J. T. Baker Inc.), and high purity formic acid (Catalog Number 11670-1, EMD Chemicals Inc., Gibbstown, NJ) were used to prepare the mobile phase. BioBasic strong cation exchange 100 × 0.32-mm (Part Number 73205-100365, Thermo Hyperil-Keystone, Bellefonte, PA) and C18 100 × 0.18-mm columns (Part Number 72105-100266, Thermo Hyperil-Keystone) were used for 2-D LC separation. Finally the LCQ ProteomeX work station (Serial Number LDP00482, Thermo Finnigan, San Jose, CA) was used to conduct 2-D LC-MS/MS analyses.

The salt step applied to the first dimensional strong cation exchange column was 10% mobile phase D (0.1% formic acid in 1 M ammonium chloride, 5% acetonitrile) and 0.1% formic acid, 5% acetonitrile as mobile phase C. This was followed by reverse phase separation on the second dimensional C18 column. Two gradients were utilized: 5–65% mobile phase B (0.1% formic acid in acetonitrile) was run for 30 min, and 65–80% mobile phase B was run for 10 min. Mobile phase A was 0.1% formic acid in water. The sample injection value was 10 μl, the flow rate of four pumps was 200 μl/min, and the flow rate on the spray needle after the splitting T was 1 μl/min.

The LCQ DECA XP PLUS ion trap mass spectrometer was tuned weekly with 5 pmol/μl angiotensin I to maintain an intensity level of high e+8 or low e+9. The voltage of the electrospray ion source was 3.80 kV, the capillary voltage was 37 V, and the capillary temperature was 150 °C. The full automatic gain control target was −2e+7, and the automatic gain control off ion time was 5 ms. Multiplier voltage of the detector was 850 V. All data collection was performed using the updated tune file (61).

Data Analyses

Data Export—All .raw data initially generated from 2-D LC-MS were first transformed into .txt files via Rawfile Version 1 (in-house software). This software can index raw data individually or in groups, and data files (averaging ~150 megabytes) can be transformed within 1 min. The resultant text files are peak lists containing three columns: mass scan number, m/z value, and ion peak intensity (shown in Supplemental Table 2S).

Hierarchical Clustering Analysis (HCA) — As a form of cluster analysis, HCA involves grouping similar items. This method is used for smaller sample sets (typically less than 250) (62) and was therefore well suited for analysis of 120 samples (Supplemental Table 1S). Spotfire 8.0 (Spotfire, Somerville, MA) was used to conduct HCA of 300 raw data sets from the 120 sera (Fig. 1). The export data were first summarized with the mass interval as 0.5 amu, and only the highest peak within each mass unit was selected for further normalization. Then taking the highest peak in the entire run as 100%, all the other peaks in the same run were normalized correspondingly. Finally HCA was conducted on the summarized and normalized data set. Results of HCA were presented in a dendrogram, which represents the similarity of two samples by distances between two columns (the smaller the distances, the more similar the samples) (62).

Principal Component Analysis (PCA) — As another kind of cluster, PCA identifies variance in principle components. These components are orthogonal (and therefore uncorrelated) to previous principle components of the same data set. The n-dimensional data set was transformed into a three-dimensional data set in preparation for PCA (63). These transformations reduce the original data to its most important dimensions, filter out noise, and facilitate more distinct cluster formation.

PCA was conducted using Clementine 8.0 (SPSS, Inc., Chicago, IL). The export data were first summarized with the mass interval as 1.0 amu, and only the highest peak within each mass unit was selected for the further normalization. Then for the normalization, taking the average peak intensity of each run as 1000, all the peak intensities in each run were adjusted correspondingly. Finally PCA was conducted on the summarized and normalized data set. Because input nodes were limited to 650, peptides detected in 57 min of 2-D LC-MS analysis were separated into two m/z value-based groups. The m/z range of the first group was 300–950, and that of the second group was 950–2000. A three-dimensional diagram of the results was created in which single samples are represented as individual spots (Fig. 2).

Artificial Neural Network (ANN) Modeling — ANNs are mathematical models with connection geometry analogous to neurons (64). These
tools identify arbitrary nonlinear multiparametric functions from experimental data (64). Through trial and error and by using different connection weight combinations, ANNs are “trained” to recognize complex relationships between input and output. These tools are used for diagnosis and prognosis (65, 66), for pattern recognition (67), for compound detection (68), and for biological functioning assessment (69). ANNs are currently the premier bioinformatic modeling tool because of their applicability to complex relationships and mechanisms (70).

ANN was conducted using Clementine 8.0 (SPSS, Inc.). The same summarized and normalized data set used for PCA was also assessed using ANNs. ANN modeling generates predictions in a manner similar to the human brain. The procedure involves entering sample data (input layers), artificial reasoning (hidden layers), and relating samples to pathological categories (output layers). Based on signal-to-noise ratios, the top 20 peptide features and corresponding normalized relative intensities were selected from each pathological category and designated as input nodes, and six disease stages (invasive breast carcinoma, ductal carcinoma in situ, atypical hyperplasia, benign neoplastic disease, benign non-neoplastic disease, and normal) were designated as output nodes. Hidden layers are not related to biology or to data; they are simply paths that the computer uses to “think.” One hundred twenty samples were randomly divided into two groups: a training group (65%) and a test group (35%). Training was randomly performed 12 times (Fig. 3), yielding 12 models. Corresponding input layers, hidden layers, and output layers are listed in Table II.

RESULTS

To test the predictive value of the data extracted via the semiquantitative peptide feature profiling (SPFP) method and the reproducibility and stability of 2D LC-MS platform, HCA was first conducted on 300 raw data sets of 120 sera (Fig. 1). Columns represent samples, and each pathological stage is designated by a different color. HCA results clearly reveal delineations between different disease stages. Most samples of the same pathological stage were clustered, and the distribution of stage groups in the dendrogram was reasonable.

The left end of the dendrogram was characterized by invasive samples. Middle and right dendrogram areas were characterized by in situ or atypical samples and benign or normal samples, respectively. In addition, data from multiple runs of the same sample were always clustered first, and this supported the validity of data extracted by the SPFP method as well as the reproducibility and stability of our 2-D LC-MS platform.

PCA was conducted to independently confirm HCA results. Fig. 2 displays PCA results. Each sample is represented by a single spot, and samples from different pathological stages are differently colored. It appears obvious that those invasive samples are clustered and that they are delineated from non-invasive samples and benign and normal samples as well. PCA clearly confirmed HCA results, further supporting the predictive capability of SPFP-extracted data.

The overall predictive accuracy of 12 random ANN models for 501 tests was 77.8% (Table III), and the accuracy of individual models ranged from 70.6 to 84.8% (Table IV). These models performed best when applied to invasive samples (201 tests, 94.81% accurate) followed by benign non-neoplastic samples (96 tests, 81.25% accurate), normal samples (38 tests, 78.95% accurate), benign neoplastic samples (51 tests, 66.67% accurate), atypical samples (44 tests, 50% accurate), and in situ samples (60 tests, 41.67% accurate).

DISCUSSION

Development of SPFP Method—HPLC has been used for many years to quantitatively analyze proteins, peptides, and many types of metabolic molecules (71–74). Combined with modern mass spectrometry, 2-D LC-MS produces quality data that contain rich quantitative information. Some researchers, however, have noted that it is difficult to conduct
quantitative analyses without isotopic labeling. This perception may be a response to the incompleteness of all peptide tandem mass spectrometry elution peaks. Tandem mass spectrometry involves a standardized sequence of events: a full scan followed by a tandem scan. Because tandem scans focus on the collision of a single target peptide, elution behavior of all other peptides cannot be recorded simultaneously. It is, of course, difficult to conduct quantitative proteomic analyses without complete peak detection.

This argument is reasonable although not entirely correct. Two considerations are noteworthy: time and quantitative analysis criteria. Tandem mass spectrometry scan times are generally 200 ms or less for full and tandem scans. Average peptide elution times, on the other hand, are ~0.5 min for 2-D LC separation. The ratio of scan times to elution times is ~1:150. Although some elution points cannot be recorded when tandem scan events occur, the elution peaks of most peptides can be ascertained based on information from multiple full scans.

Using criteria other than peak area for quantitative analyses moreover may facilitate proteomic profiling without isotopic labeling. Peak asymmetry (a ratio of two half-peak widths) and peak height (the maximum height detected in an entire elution peak) may also be analyzed fruitfully. Peak asymmetry is usually used for purity testing. Peak height, like peak area, is commonly used in quantitative analyses. Peak area is the primary focus of most quantitative analyses because it provides more information than peak height. However, incomplete peptide elution peaks and the potential overlay of elution peaks with similar retention times and mass values both may render peak area less practical than peak height.

Therefore, an SPFP method was developed to explore this and other methodological possibilities. Through the comparison of normalized peak intensity of the summarized peptide features extracted from the 2D LC-MS results of various samples, we hope to mine out the peptide features that are significant and pathologically relevant from the enormous data set. The proposed signal processing approach appears to work well, suggesting it may be helpful for routine compar-

---

**TABLE II**

| Model no. | Input layer | Hidden layer 1 | Hidden layer 2 | Output layer | Estimated accuracy |
|-----------|-------------|----------------|----------------|--------------|--------------------|
| 965,452   | 106         | 17             | 6              | 6            | 95.83              |
| 1,731,552 | 108         | 22             | 18             | 6            | 92.00              |
| 1,917,815 | 107         | 22             | 22             | 6            | 92.59              |
| 2,142,754 | 105         | 22             | 18             | 6            | 90.00              |
| 3,332,948 | 109         | 11             | 6              | 6            | 84.85              |
| 4,041,737 | 111         | 5              | 6              | 6            | 91.30              |
| 4,296,339 | 101         | 27             | 22             | 6            | 92.31              |
| 5,043,191 | 107         | 14             | 6              | 6            | 83.33              |
| 6,043,964 | 103         | 14             | 6              | 6            | 83.87              |
| 6,793,679 | 101         | 8              | 6              | 6            | 82.76              |
| 7,461,205 | 113         | 27             | 22             | 6            | 94.12              |
| 9,172,249 | 106         | 22             | 18             | 6            | 78.26              |
ison of complex mixtures for the purpose of differential expression analysis and biomarker detection.

Performance Inconsistency of ANN Modeling—Poor pathological characterization of the two non-invasive breast carcinomas is attributable to two factors. The first factor is disease stage distribution. Utilizing 120 samples and 300 extracted raw data sets, the present investigation is larger than any previous 2-D LC-MS breast cancer serum profiling study. However, the sample set was not well balanced. Again the set included 49 invasive breast carcinoma sera, 13 DCIS sera, 13 atypical hyperplasia sera, 24 benign non-neoplastic sera, 11 benign neoplastic sera, and 10 normal control sera. Although clinical sample collection provided reliably diagnosed sera, this procedure also yielded a suboptimal pathological distribution. “Cut-one-off” modeling may have facilitated accurate pathological characterization despite this imbalance, but a more stringent ANN modeling approach was utilized to prevent “overfitting.” With less data available for ANN training, it is unsurprising that modeling of underrepresented disease stages yielded less accurate models (Table IV).

The second reason for poor pathological characterization of non-invasive breast carcinomas is related to their distinctiveness. Differences between samples of opposite pathological extreme, of course, are better defined than differences between samples that are pathologically similar. Like ANN modeling, PCA and HCA clearly delineated invasive and normal/benign non-neoplastic groups. However, they failed to

| Overall | Correct | 390 | 77.84% |
|---------|---------|-----|--------|
|         | Wrong   | 111 | 22.16% |
|         | Total   | 501 |        |

| Breakdown: Randomization |
|--------------------------|
| 4,296,339                |
| Correct: 36 | 70.59% |
| Wrong: 15   | 29.41% |
| Total: 51   |        |
| 5,043,191                |
| Correct: 31 | 72.50% |
| Wrong: 11   | 27.50% |
| Total: 42   |        |
| 6,043,964                |
| Correct: 29 | 72.50% |
| Wrong: 11   | 27.50% |
| Total: 40   |        |
| 6,793,679                |
| Correct: 30 | 78.95% |
| Wrong: 8    | 21.05% |
| Total: 38   |        |
| 7,461,205                |
| Correct: 37 | 82.22% |
| Wrong: 8    | 17.78% |
| Total: 45   |        |
| 9,172,249                |
| Correct: 39 | 79.59% |
| Wrong: 10   | 20.41% |
| Total: 49   |        |

| Breakdown: Group |
|------------------|
| AT               |
| Correct: 22 | 50% |
| Wrong: 22     | 50% |
| Total: 44      |        |
| BN               |
| Correct: 78 | 81.25% |
| Wrong: 18      | 18.75% |
| Total: 96      |        |
| BP               |
| Correct: 34 | 66.67% |
| Wrong: 17     | 33.33% |
| Total: 51      |        |
| IS               |
| Correct: 25 | 41.67% |
| Wrong: 35      | 58.33% |
| Total: 60      |        |
| IV               |
| Correct: 201 | 94.81% |
| Wrong: 11     | 5.19% |
| Total: 212     |        |
| NR               |
| Correct: 30 | 78.95% |
| Wrong: 8     | 21.05% |
| Total: 38      |        |
delineate other stages with comparable precision. This also suggests that the extracted data from DCIS and atypical hyperplasia samples may not be representative of these pathologies.

Comparison with Previous Methods—Important clinical profiling issues include reproducibility, quantitative differences, and clinical relevance. Typical protein database search strategies may be poorly suited for addressing these issues. To overcome limitations of database search strategies, researchers recently developed new LC-MS data mining tools. These researchers include Aebersold and co-workers (59, 75–78), Rexach and co-workers (50, 79), and Zubarev and co-workers (48, 80). The developing semiquantitative profiling methods of these researchers can broadly be divided into two categories: MS-based peptide fingerprinting (59, 77) and MS/MS-based peptide sequencing (48, 50, 75, 76, 78–80).

Similar to the peptide array method recently published (59), SPFP is designed to computationally identify biomarker signals via analysis of raw 2-D LC MS data. Proposed identification methods all involve raw data exportation, summarization, normalization, extraction of features (e.g., peak apex, m/z value, or retention time), and alignment of corresponding features from different samples. Compared with the peptide array method (59), SPFP has two limitations. First, SPFP bypasses the monoisotopic test. This step was foregone because, unlike the ESI-Q-TOF mass spectrometers used in previous studies, the mass resolution of the LCQ ion trap mass spectrometer (approximately 0.5 amu) was insufficient for detection of double and triple charged ions in full mass scan. Second, previous approaches have used software to fix retention time shifts, and the adjusted retention times were then used as the third criterion for data alignment of peptide arrays. SPFP bypasses the retention time in initial data alignment, focusing primarily on m/z values and peak intensities. The disadvantage of this bypass is that the different ions with similar m/z values may occasionally be identified as a single profiling target.

Although data analysis methods utilized for SPFP may be less thorough, their simplicity may also be advantageous. Every data processing step introduces opportunities for error. Potential errors related to retention time are a good example. Our observations and a previous investigation (58) suggest that retention times are difficult to manually modify because the retention times of the peptide ions were inconsistent even in the same run of the same sample. In some cases, these retention times had opposite signs (one positive, the other negative). Retention time is certainly an important criterion for...
accurate alignment, but it is also troublesome. This exemplifies how errors can arise when adjusting retention time shift. The SPFP method may bypass some risk of such errors due to the simple data process.

The approach described herein and the other previously developed approaches all have advantages and limitations. It is not yet clear which of these approaches will be useful, but it is clear that all are worthy of further exploration. The proposed SPFP signal processing approach appears to work well on a large clinical data set, suggesting it may be helpful for routine comparison of complex mixtures for the purpose of differential expression analysis and biomarker detection. Future research will focus on improving the data extraction algorithms (i.e. may involve the retention time) and utilizing the larger and well balanced sample set to improve the predictive accuracy.

Acknowledgments—We sincerely appreciate the kind support from Dr. Yonghong Zhang on data analyses. We also appreciate Dr. Richard Mural for the kind review and valuable feedback. We thank all colleagues at Winder Research Institute and Walter Reed Army Medical Center. We are grateful for the support of participating patients and their families.

* This work was supported by the United States Department of Defense (Clinical Breast Care Project) through the United States Army Medical Research and Materiel Command/Telemedicine, the Advanced Technology Research Center (TATRC) in Fort Detrick, MD, and the Henry M. Jackson Foundation for the Advancement of Military Medicine in Rockville, MD. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The online version of this article (available at http://www.ncpmonline.org) contains supplemental material.

§ To whom correspondence should be addressed: Windber Research Inst., 620 Seventh St., Windber, PA 15963. Tel.: 814-386-3039; Fax: 814-282-0388; E-mail: ruqh@hotmail.com.

REFERENCES

1. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell. Proteomics 1, 845–867

2. Anderson, L., and Anderson, N. G. (1977) High resolution two-dimensional electrophoresis of human plasma proteins. Proc. Natl. Acad. Sci. U. S. A. 74, 5421–5425

3. Merrill, C. R., Goldman, D., Sedman, S. A., and Ebert, M. H. (1981) Ultra-sensitive stain for proteins in polyacrylamide gels shows regional variation in cerebrospinal fluid proteins. Science 211, 1437–1438

4. Merrill, C. R., Switzer, R. C., and Van Keuren, M. L. (1979) Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. Proc. Natl. Acad. Sci. U. S. A. 76, 4335–4339

5. Wulffkuhle, J. D., Liotta, L. A., and Petricoin, E. F. (2003) Proteomic applications for the early detection of cancer. Nat. Rev. Cancer 3, 267–275

6. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207

7. Brimond, E. P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol. Cell. Proteomics 3, 367–378

8. Varnum, S. M., Covington, C. C., Woodbury, R. L., Petritis, K., Kangas, L. J., Abdullah, J. S., Pound, J. G., Smith, R. D., and Zangar, R. C. (2003) Proteomic characterization of nipple aspirate fluid: identification of potential biomarkers of breast cancer. Breast Cancer Res. Treat. 80, 87–97

9. Celis, J. E., Gromov, P., Cabezon, T., Moreira, J. M., Ambartsouman, N., Sandelin, K., Rank, F., and Gromova, I. (2004) Proteomic characterization of the interstitial fluid perfusing the breast tumor microenvironment: a novel resource for biomarker and therapeutic target discovery. Mol. Cell. Proteomics 3, 327–344

10. Gorg, A., Weiss, W., and Dunn, M. J. (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4, 3665–3685

11. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. U. S. A. 97, 9390–9395

12. Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y., and Chan, D. W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin. Chem. 48, 1299–1304

13. Vlahou, A., Laronga, C., Wilson, L., Gregory, B., Fournier, K., McGaughey, D., Perry, R. R., Wright, G. L., Jr., and Semmes, O. J. (2003) A novel approach toward development of a rapid blood test for breast cancer. Clin. Breast Cancer 4, 203–209

14. Adkins, J. N., Varmus, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L., and Pounds, J. G. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. Mol. Cell. Proteomics 1, 947–955

15. Trumalai, R. S., Chan, K. C., Prieto, D. A., Issac, H. J., Jr., Conrads, T. P., and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. Mol. Cell. Proteomics 2, 1096–1103

16. Shen, Y., Jacobs, J. M., Camp, D. G., II, Fang, R., Moore, R. J., Smith, R. D., Ciao, W., Davis, R. W., and Tompkins, R. G. (2004) Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. Anal. Chem. 76, 1134–1144

17. Zhang, X. H., Li, E. G., Li, X., Mallick, P., Kely-Spratt, K. S., Masselon, C. D., Camp, D. G., II, Smith, R. D., Kemp, C. J., and Aebersold, R. (2005) High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. Mol. Cell. Proteomics 4, 144–155

18. Hood, B. L., Lucas, D. A., Kim, G., Chan, K. C., Blonder, J., Issac, H. J., Veenstra, T. D., and Conrads, T. P. (2005) Quantitative analysis of the low molecular weight serum proteome using 18O stable isotope labeling in a lung tumor xenograft mouse model. J. Am. Soc. Mass Spectrom. 16, 1221–1230

19. Kislinger, T., Gramolini, A. O., MacLennan, D. H., and Emili, A. (2005) Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. J. Am. Soc. Mass Spectrom. 16, 1207–1220

20. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247

21. Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., Haynes, P. A., Hays, L., Schietz, D., Ulaszek, R., Wei, J., Wolters, D., and Yates, J. R., III (2002) Proteomic survey of metabolic pathways in rice. Proc. Natl. Acad. Sci. U. S. A. 99, 11989–11994

22. Le Roch, K. G., Johnson, J. R., Flores, L., Zhou, Y., Santoroysan, A., Grainger, M., Yan, S. F., Williamson, K. C., Holder, A. A., Carucci, D. J., Yates, J. R., III, and Winzeler, E. A. (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res. 99, 11969–11974

23. Schirmer, E. C., Flores, L., Guan, T., Yates, J. R., III, and Gerace, L. (2003) Nuclear membrane proteins with potential disease links found by subtractive proteomics. Science 301, 1380–1382

24. Rothemund, D. L., Locke, V. L., Liew, A., Thomas, T. M., Wasinger, V., and Ryllat, D. B. (2003) Depletion of the highly abundant protein albumin from human plasma using the GradientFlow. Proteomics 3, 279–287

25. Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L., and Steiner, S. (2003) Multi-component immunofluorescence subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. Proteomics 3, 422–432

26. Patterson, S. D. (2003) Data analysis—the Achilles heel of proteomics. Nat. Biotechnol. 21, 221–222

27. Boguski, M. S., and McIntosh, M. W. (2003) Biomedical informatics for proteomics. Nature 422, 233–237

28. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tan-
Semiquantitative Peptide Feature Profiling of Sera

29. Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) Informatics for protein identification by mass spectrometry. Methods 35, 223–236

30. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic informatics. Rev. Neurobiol. 61, 129–157

31. Baldwin, M. A. (2004) Protein identification by mass spectrometry: issue to be considered. Mol. Cell. Proteomics 3, 1–9

32. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658

33. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) New database-based identified peptides and assesses the quality of MS/MS techniques. Mol. Cell. Proteomics 4, 38–48.

34. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Kirstensen, D. B., Brond, J. C., Andersen, J. R., Sorensen, M., Rahman, K., Radulovic, D., Cox, B., Tossant, J., and Emili, A. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal. Chem. 76, 5556–5568

35. Yang, X., Dondetti, V., Dezube, R., Maynard, D. M., Geer, L. Y., Epstein, J., Chen, X., Markey, S. P., and Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. J. Proteome Res. 3, 1002–1008

36. Kislinger, T., Rahmann, K., Radulovic, D., Cox, B., Tossant, J., and Emili, A. (2004) PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics 3, 96–106

37. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 3, 1023–1038

38. Wheeler, D. L., Church, D. M., Edgar, R., Federhjem, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, D. G., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center of Biotechnology Information: update. Nucleic Acids Res. 32, D35–D40

39. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O’Donovan, C., Phan, I., PILibout, S., and Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370

40. Perkinns, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

41. Liebman, H., and Shriver, C. D. (2006) Exploring human plasma proteome: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol. Cell. Proteomics 5, 1194–1204

42. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Lipton, M. S., Auberry, K. J., Shtrittmatter, E. F., Chen, X., Markey, S. P., and Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. J. Proteome Res. 3, 1002–1008

43. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) New database-based identified peptides and assesses the quality of MS/MS techniques. Mol. Cell. Proteomics 4, 38–48.

44. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 3, 1023–1038

45. Wheeler, D. L., Church, D. M., Edgar, R., Federhjem, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, D. G., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center of Biotechnology Information: update. Nucleic Acids Res. 32, D35–D40

46. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O’Donovan, C., Phan, I., PILibout, S., and Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370

47. Kersey, J. P., Duarte, J., Williams, A., Karavidopoulou, Y., Binney, E., and Apweiler, R. (2004) The international protein index: an integrated database for shotgun proteomics experiments. Proteomics 4, 1955–1988

48. Eng, J. K., McCormack, A. L., and Yates, J. R. III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989

49. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

50. Liebman, H., and Shriver, C. D. (2006) Exploring human plasma proteome: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol. Cell. Proteomics 5, 1194–1204

51. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Lipton, M. S., Auberry, K. J., Shtrittmatter, E. F., Chen, X., Markey, S. P., and Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. J. Proteome Res. 3, 1002–1008

52. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) New database-based identified peptides and assesses the quality of MS/MS techniques. Mol. Cell. Proteomics 4, 38–48.

53. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 3, 1023–1038

54. Wheeler, D. L., Church, D. M., Edgar, R., Federhjem, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, D. G., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center of Biotechnology Information: update. Nucleic Acids Res. 32, D35–D40

55. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O’Donovan, C., Phan, I., PILibout, S., and Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370

56. Kersey, J. P., Duarte, J., Williams, A., Karavidopoulou, Y., Binney, E., and Apweiler, R. (2004) The international protein index: an integrated database for shotgun proteomics experiments. Proteomics 4, 1955–1988

57. Eng, J. K., McCormack, A. L., and Yates, J. R. III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989

58. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

59. Liebman, H., and Shriver, C. D. (2006) Exploring human plasma proteome: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol. Cell. Proteomics 5, 1194–1204

60. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Lipton, M. S., Auberry, K. J., Shtrittmatter, E. F., Chen, X., Markey, S. P., and Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. J. Proteome Res. 3, 1002–1008
910, 173–179
68. Agatonovic-Kustrin, S., and Beresford, R. (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 22, 717–727
69. Sparks, T. C., Anzeveno, P. B., Martynow, J. G., Gifford, J. M., Hertlein, M. B., Worden, T. C., and Kirst, H. A. (2000) The application of artificial neural networks to the identification of new spinosoids with improved biological activity toward larvae of Heliothis virescens. Pestic. Biochem. Physiol. 67, 187–197
70. Spining, M. T., Darsey, J. A., Sumpter, B. G., and Noid, D. W. (1994) Opening up the black box of artificial neural networks. J. Chem. Educ. 71, 406–411
71. Coulais, Y., Campistron, G., Caillard, C., and Houin, G. (1986) Quantitative determination of alizapride in human plasma by high-performance liquid chromatography. J. Chromatogr. 374, 425–429
72. Alton, K. B., Desrivieres, D., and Patrick, J. E. (1986) High-performance liquid chromatographic assay for hydrochlorothiazide in human urine. J. Chromatogr. 374, 103–110
73. Poirier, J. M., Jaillon, P., and Cheymol, G. (1986) Quantitative liquid chromatographic determination of sotalol in human plasma. Ther. Drug Monit. 8, 474–477
74. Broquaire, M., Rovei, V., and Braithwaite, R. (1981) Quantitative determination of naproxen in plasma by a simple high-performance liquid chromatographic method. J. Chromatogr. 224, 43–49
75. Zhang, H., Yi, E. C., Li, X. J., Mallick, P., Kelly-Spratt, K. S., Masselon, C. D., Camp, D. G., II, Smith, R. D., Kemp, C. J., and Aebersold, R. (2005) High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. Mol. Cell. Proteomics 4, 44–55
76. Pan, S., Zhang, H., Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M. J., and Aebersold, R. (2005) High throughput proteome screening for biomarker detection. Mol. Cell. Proteomics 4, 182–190
77. Prakash, A., Mallick, P., Whiteaker, J., Zhang, H., Paulovich, A., Flory, M., Lee, H., Aebersold, R., and Schwikowski, B. (2006) Signal maps for mass spectrometry-based comparative proteomics. Mol. Cell. Proteomics 5, 423–432
78. Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S., and Aebersold, R. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol. Cell Proteomics 5, 652–670
79. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzihradszky, K. F., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: I. How much of the data is theoretically interpretable by search engines? Mol. Cell. Proteomics 4, 1189–1193
80. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (Jan 25, 2006) ModifiComb: new proteomics tool for mapping substoichiometric post-translational modifications, finding novel types of modifications and fingerprinting complex protein mixtures. Mol. Cell. Proteomics 10.1074/mcp.T500034-MCP200