A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry*

Xiao-jun Li‡§, Eugene C. Yi‡, Christopher J. Kemp¶, Hui Zhang‡, and Ruedi Aebersold‡

There is an increasing interest in the quantitative proteomic measurement of the protein contents of substantially similar biological samples, e.g. for the analysis of cellular response to perturbations over time or for the discovery of protein biomarkers from clinical samples. Technical limitations of current proteomic platforms such as limited reproducibility and low throughput make this a challenging task. A new LC-MS-based platform is able to generate complex peptide patterns from the analysis of proteolyzed protein samples at high throughput and represents a promising approach for quantitative proteomics. A crucial component of the LC-MS approach is the accurate evaluation of the abundance of detected peptides over many samples and the identification of peptide features that can stratify samples with respect to their genetic, physiological, or environmental origins. We present here a new software suite, SpecArray, that generates a peptide versus sample array from a set of LC-MS data. A peptide array stores the relative abundance of thousands of peptide features in many samples and is in a format identical to that of a gene expression microarray. A peptide array can be subjected to an unsupervised clustering analysis to stratify samples or to a discriminant analysis to identify discriminatory peptide features. We applied the SpecArray to analyze two sets of LC-MS data: one was from four repeat LC-MS analyses of the same glycopeptide sample, and another was from LC-MS analysis of serum samples of five male and five female mice. We demonstrate through these two study cases that the SpecArray software suite can serve as an effective software platform in the LC-MS approach for quantitative proteomics. Molecular & Cellular Proteomics 4: 1328–1340, 2005.

The identification and quantification of the protein contents of biological samples plays a crucial role in biological and biomedical research (1–4). Due to the large dynamic range and the high complexity of most proteomes, it is very challenging to identify and accurately quantify the majority of proteins from such samples. LC-MS/MS-based methods are currently most efficient for the identification of a large number of proteins and have been widely applied in biological and biomedical research (5–7). If combined with stable isotope labeling, such methods can also accurately quantify proteins (7–13). In a typical LC-MS/MS-based quantitative proteomic experiment, samples to be compared are differentially and isotopically labeled, combined, and enzymatically digested into peptides. The obtained peptide samples are then separated by a multidimensional LC system and analyzed by MS/MS. Peptides are ionized either by ESI (14) or by MALDI (15), and peptide ions are selected, usually in the order of decreasing signal intensity, for fragmentation by CID (16, 17). Peptides are identified by using an automated search engine, such as SEQUEST (18) or Mascot (19), to match their fragment ions against a designated protein database. Peptides are quantified by using a quantification software tool, such as XPRESS (8) or ASAPRatio (20), that uses the relative MS signal intensities of the different isotopic forms to calculate the relative abundance of each identified peptide. The identification and quantification of proteins is then achieved by combining the information obtained from the peptides that associate with the particular protein (20, 21). The quantitative LC-MS/MS approach can routinely identify and quantify hundreds to thousands of proteins from a single biological sample but is generally unable to comprehensively analyze proteomes.

There is an increasing interest in the quantitative proteomic measurement of the protein contents of numerous, substantially similar samples. Typical examples include the discovery of protein biomarkers from clinical samples (3, 22, 23) and the measurement of the response of cells and tissues to perturbations. For biomarker discovery, large numbers of samples need to be processed to achieve sufficient statistical power to distinguish disease-specific markers from coincidental pro-
teome fluctuations within the human population. For the study of cellular response to perturbations, temporal and dose-dependent changes are usually particularly informative for the identification of patterns of proteins that are specifically affected by the treatment (24). Therefore, these and similar applications require high sample throughput and highly reproducible coverage of the proteome. Quantitative LC-MS/MS is of limited use for such large scale studies because of significant undersampling of complex proteomic samples. Even after extensive peptide fractionation, a significant fraction of the peptides present in a sample is not selected by the mass spectrometer for CID (25). These peptides are neither identified nor quantified. Therefore, if multiple, substantially similar samples are being analyzed, the fraction of those peptides that is measured in every sample rapidly decreases with increasing sample number. Furthermore the peptides that are consistently detected tend to be the high abundance peptides that generate intense MS signals, whereas many lower abundance, biologically interesting peptides are inconsistently sampled. As a result, it is very difficult to consistently obtain quantitative information on low abundance proteins across multiple samples by the LC-MS/MS.

As an alternative to LC-MS/MS-based methods, peptide mapping by LC-MS has been gaining momentum as a method for quantitative proteomics (23, 26, 27). The LC-MS approach is based on the principle that the MS signal intensity of each peptide in a substantially similar sample analyzed under identical conditions is proportional to the abundance of the peptide within the dynamic range of the instrument and is at least monotonic to the abundance beyond the dynamic range. Therefore one may evaluate the relative abundance of a peptide in different, related samples by analyzing the samples under identical LC-MS conditions and by comparing MS signal intensity of the same peptide in different LC-MS runs (26–28). The obtained relative abundance is quantitative within MS dynamic range and semiquantitative beyond the range. The LC-MS approach for quantitative proteomics can be summarized in the following three steps. 1) Proteins are extracted and purified from related samples and enzymatically digested into peptides. Peptide samples are then analyzed by an LC-ESI-MS system preferably using a high resolution, high accuracy Q-TOF or similarly performing analyzer under identical conditions. Optionally peptides can be labeled with stable isotope reagents to achieve more accurate quantification, and a limited number of CID attempts can be carried out to identify a subset of the detected and quantified peptides. 2) Software tools are applied to compare peptide patterns, extract peptide relative abundance from LC-MS data, and identify a list of peptides that stratify samples with respect to their genetic, physiological, or environmental characters. Despite the fact that the amino acid sequence of the detected peptides is generally not known, identical peptides in different samples can be unambiguously matched by their m/z, their charge state, and their chromatographic retention time. 3) Information (m/z, charge, and retention time) of discriminatory peptides is fed into a mass spectrometer to identify the amino acid sequence of such peptides by targeted MS/MS and database searching (23). By first measuring the relative abundance of thousands of peptides and then focusing the power of MS/MS selectively on discriminatory peptides, this method has a better chance of identifying discriminatory proteins than the standard LC-MS/MS shotgun approach.

Here we present a new software suite, SpecArray, that fulfills the important second step of the LC-MS approach for quantitative proteomics. The software suite takes a set of LC-MS data as input and outputs a peptide versus sample array that stores the relative abundance of thousands of peptide features matched in all samples. The format of a peptide array is identical to that of a gene expression microarray except that peptide features replace gene names in a peptide array. Just like microarrays (29), peptide arrays can be subjected to unsupervised clustering analyses (such as hierarchical or k-means) to classify sample types and/or to discriminate analyses (such as Student’s t test or linear discriminant function) to identify peptides discriminating between samples of different characters. The SpecArray software suite contains five distinct but integrated software tools. 1) The Pep3D module visualizes LC-MS data in graphic images to ensure data quality (25); 2) the mzXML2dat module extracts high quality MS signals from raw noisy MS signals; 3) the PepList module applies pattern matching to extract peptide features from MS signals; 4) the PepMatch module aligns peptide features between different samples; and 5) the PepArray module generates a peptide array from aligned peptide features in all samples. To minimize the adverse effects of possible retention time shift in different LC runs, PepMatch first evaluates a retention time calibration curve (RTCC) between any two samples and then uses the RTCCs to align peptides of all samples. To correct any systematic errors due to uneven sample loading or uneven ionization efficiency, PepArray performs sample-dependent ratio normalization (20) before reporting peptide relative abundance. We applied the SpecArray software suite to analyze two sets of LC-MS data: one was from four repeat LC-MS analyses of the same glycopeptide sample, and another was from LC-MS analysis of serum samples of five male and five female mice. As demonstrated by these two study cases, the SpecArray software suite is very useful for analyzing LC-MS data and can serve as an effective software platform in the LC-MS approach for quantitative proteomics.

**EXPERIMENTAL PROCEDURES**

**Serum Collection and Sample Preparation**—Ten serum samples were collected from five male and five female mice of the same litter

---

1 The abbreviations used are: RTCC, retention time calibration curve; CV, coefficient of variation; S/N, signal-to-noise ratio; SIC, single ion chromatogram; r.m.s., root mean square.
Quantitative LC-MS Approach for Peptide Profiling

at the age of 22 weeks using a procedure described previously (23). A plasma sample was collected from a male mouse at the age of 22 weeks using a similar protocol except that blood was placed in a K$_2$EDTA-coated 1.5-mi microcentrifuge tube and centrifuged at 4 °C for 5 min at 3000 rpm. Formerly N-linked glycosylated peptides were isolated from 50 μl of each sample using the N-linked glycopeptide capture procedure as described previously (23, 30).

**LC-MS Analysis**—Peptide isolates from 5 μl of original serum or plasma sample were analyzed by a reversed phase LC-ESI-Q-TOF-MS system as described previously (23). An ESI-Q-TOF mass spectrometer (Waters, Beverly, MA) was used. Experimental settings were identical in all sample analyses. Raw MS data were converted into the mzXML common file format using the MassWolf file converter (31).

**Initial MS Signal Processing**—The software tool mzXML2dat processes raw LC-MS data. The process is illustrated in Supplemental Fig. S1 and includes the following five steps. 1) A smooth MS spectrum is generated from each raw MS spectrum by using the one-dimensional translation-invariant wavelet transformation filtering method (Symmlet8) (32, 33). 2) A centroid MS spectrum is then generated from each smoothed MS spectrum that consists of all peak apexes (as specified by m/z and intensity) in the smoothed spectrum. 3) Local background noise within each centroid MS spectrum is estimated as the median intensity of all signals that are within a ±50 m/z window of the target m/z value. 4) Low intensity signals are removed from each centroid MS spectrum using a signal-to-noise ratio (S/N) cutoff of 2, and local background is subtracted from the retained MS signals. 5) The remaining MS signals, as specified by m/z, intensity, and S/N, constitute a denoised, centroid spectrum. Thus a denoised, centroid MS spectrum is derived from each raw MS spectrum.

Because the current mzXML schema (31) does not store S/N for MS signals, we designed an mzDAT binary file format to store all denoised, centroid MS spectra of an LC-MS analysis. Details of the mzDAT format are described in Supplemental Fig. S2. The mzDAT format allows for quick access to individual MS spectra and also decreases the file size. The mzXML2dat program also allows the user to specify the m/z range and the retention time range within which raw data are processed and converted. Hence information-poor spectra, such as those collected when the reversed phase column is washed, can be removed without further analysis.

**Extraction of Peptide Features**—The software tool PepList extracts peptide features from denoised, centroid LC-MS data as stored in our mzDAT file format. The extraction process contains the following three steps. 1) Signals in each MS spectrum are compared with peptide isotopic distributions to identify potential peptide features from the spectrum. Starting with the most intense signal in the spectrum, the software first determines the charge state of the signal by examining the m/z difference between neighboring MS signals. Assuming the most intense signal is produced by the monoisotopic mass (M) of a peptide, the software then calculates the expected peptide isotopic distribution (34, 35) and assesses how well the distribution fits neighboring MS signals. Considering that the monoisotopic peak of a peptide may not be the most intense among all isotopic peaks of the peptide, the software repeats the pattern matching process by assuming that the most intense signal is produced by M + 1 or M + 2 isotopic mass of a peptide. If at least one of the three expected isotopic distributions fits the MS signals well, a peptide feature is identified from the MS spectrum; the best fitting assumption determines the monoisotopic mass of the peptide. A peptide feature is specified by its monoisotopic mass, its charge, and the retention time of the MS spectrum. The intensity and the S/N of the peptide feature are given by the corresponding values of the monoisotopic peak. In some cases the charge state of the most intense signal cannot be unambiguously determined. In such cases all possible charge states are examined, but only the one that provides the best fitting of nearby signals is eventually accepted. MS signals that are matched to the identified peptide feature are then removed from the MS spectrum. If no peptide feature can be identified, only the most intense signal is removed. The feature identification process is repeated for the remaining signals in the MS spectrum until no MS signal has a S/N above 5. In this way the PepList identifies a list of peptide features from each MS spectrum. Some peptide features may be generated by the same peptide but in different charge states. 2) Peptide features of nearby MS spectra (within 0.5 min in retention time), similar mass (within a m/z tolerance of 0.05), and the same charge state are compared with each other, and those of a lower intensity are discarded. 3) A single ion chromatogram (SIC) is reconstructed for each remaining peptide feature by summing MS signals of the first three isotopic masses of the peptide and tracing the summed MS signals against time. Applying methodologies developed in the ASAPRatio program (20), the raw SIC is smoothed, and the background noise in the SIC is estimated. Peptide features whose intensity at SIC peak apex is less than twice the corresponding SIC background are discarded. The remaining peptide features are collected as the output of the software. Each peptide feature is specified by its monoisotopic mass, its charge, and the retention time at the corresponding SIC peak apex, while the area of the SIC peak determines the abundance of the feature, and the S/N at the SIC peak apex indicates the quality of the feature.

**Peptide Alignment across Multiple Samples**—The software tool PepMatch aligns peptide features from multiple samples. The whole alignment process can be described in the following seven steps. 1) Peptide features of two samples are first paired together if their charges are identical and their masses are close (within a m/z tolerance of 0.5). At this initial stage, information on peptide retention time is ignored, and a peptide feature in one sample may pair with several peptide features in another sample. 2) The software then learns from all peptide pairs the RTCC between the two LC-MS analyses. The RTCC is determined in the retention time versus retention time space under the following two constraints. (a) The RTCC minimizes the root mean square (r.m.s.) distance between peptide pairs and itself, and (b) the RTCC is monotonic. 3) The software then evaluates the distance between each peptide pair and the RTCC, computes the ratio between the distance and the r.m.s. distance, and applies the formula of a normal distribution to calculate a p value for the peptide pair. The p value assesses the significance of a peptide pair. Peptide pairs having a very low p value (using a cutoff of 10$^{-3}$) are removed. Among peptide pairs that share at least one common peptide feature, the one having the lowest p value is also removed. 4) Steps 2 and 3 are repeated until all peptide pairs have a p value above the cutoff of 10$^{-3}$ and each peptide feature pairs with only one peptide feature in another analysis. At this point, the RTCC between the two LC-MS analyses is determined. 5) To collect the likely peptide pairs from the two analyses, the software evaluates a p value for all original peptide pairs in step 1. The calculation is similar to that in step 3 except that the p value of a peptide pair now depends on both its distance to the RTCC and the m/z difference between the two corresponding peptide features. The final peptide pairs are selected in the same way as in step 4. In this way, peptide features between the two LC-MS analyses are aligned to pairs. 6) Steps 1–5 are repeated for all samples until peptide features are aligned between every two samples. 7) Peptide features of all samples are combined into a peptide “super” list. At the beginning, every peptide feature of every sample enters the peptide super list as a separate entry. Based on pairing information and pairing strength (as evaluated by the p value), entries linked by pairs are iteratively combined in the order of strength. Weak conflicting pairs are treated as false positive and discarded in the process. In the
end, the peptide super list stores all alignment information between peptide features of all samples.

Generation of a Peptide Array—The software tool PepArray generates a user-customized peptide array from a peptide super list (the output of PepMatch). There are five steps in the process. 1) Based on user-specified criteria, the software first selects a list of peptide features from the peptide super list as entries to the peptide array. The software allows a user to list samples by group and specify the required minimal number of appearances for each group. To enter the peptide array, a peptide feature must satisfy the appearance requirement of at least one group. The software also allows a user to specify a m/z range and a retention time range for valid peptide features. Out-of-range peptide features are not selected for the peptide array. 2) The software then generates an initial peptide versus sample array from selected peptide features. The first dimension of the array is the list of all samples, the second dimension is the list of all selected peptide features, and an array element specifies the abundance of the corresponding peptide feature in the corresponding sample. If a peptide feature is missed in a sample, the corresponding array element is left blank. 3) For each blank element in the initial peptide array, the software then searches the corresponding peptide feature in the corresponding denoised, centroid LC-MS data to check whether the feature is truly absent. The retention time of the missed feature in the sample is estimated from retention times of the feature in other samples using the RTCCs between the samples to correct any retention time drift. If valid peptide signals are found, the abundance of the feature is evaluated, as in the PepList program, and stored in the initial peptide array; otherwise the array element is left blank. 4) The software then carries out a sample-dependent normalization (20) to correct any systematic errors on peptide abundance that may arise from variations in sample loading or ionization efficiency between LC-MS analyses. The software collects a list of “common” peptide features that appears in at least two-thirds of all samples in the initial peptide array. If a common peptide feature appears in a sample, the ratio between its abundance in the sample and its averaged abundance across all samples is evaluated. The software then collects all abundance ratios of common peptide features in each sample and applies a methodology developed for the ASAPRatio program (20) to calculate a sample-dependent normalization factor for the sample. Peptide abundance in the initial peptide array is then normalized by the corresponding normalization factor. 5) In the final step, the software evaluates the average abundance of each peptide feature across all samples, divides peptide abundance in individual samples by the averaged abundance, and obtains the relative abundance of the peptide feature in individual samples. The final output of the software is a peptide array describing the relative abundance of individual peptide features in individual samples. Peptide features in the peptide array are specified by their m/z, retention time (in minutes), monoisotopic mass, and charge state. Depending on the nature of the samples, a user may decide whether to keep blank elements in a peptide array or to replace them with a very low abundance ratio. The peptide array is so formatted that it can be analyzed directly by the clustering program Cluster (rana.lbl.gov/EisenSoftware.htm) (29). To add flexibility, the software allows a user to skip steps 4 and/or 5.

Unsupervised Hierarchical Clustering Analysis—The Cluster program (rana.lbl.gov/EisenSoftware.htm) (29) was used for unsupervised hierarchical clustering analysis of peptide arrays. Peptide array data were unfiltered and logarithmically transformed. Both samples and peptide features were clustered using single linkage clustering with centered correlation similarity metrics. The TreeView program (rana.lbl.gov/EisenSoftware.htm) (29) was used to visualize the clustering results.

RESULTS

Generating a Peptide Array from a Set of LC-MS Data

As depicted in Fig. 1, the software suite SpecArray generates a peptide array from a set of LC-MS data in five distinct functional steps, and each step is achieved by one of the five software tools: Pep3D, mzXML2dat, PepList, PepMatch, and PepArray. We describe in the following these five functional steps in detail and illustrate their main features with data collected from four repeat LC-MS analyses of an N-linked glycopeptide sample that was derived from plasma of a male mouse (23, 30).

Step 1—LC-MS data in mzXML common file format (31) is first visualized by the Pep3D software tool to assess data quality (25); see Fig. 2A for an example of a Pep3D image. Because the LC-MS approach relies exclusively on MS signal intensity for peptide quantification, it is crucial to obtain high quality, reproducible LC-MS data. Pep3D is used for assessing data quality due to the easy-to-use nature of the tool and easy-to-interpret nature of the results. Common problems such as low peptide concentration, chemical or polymer contaminations, non-optimal LC-MS performance, insufficient LC separation, etc. can be easily diagnosed by visually inspecting Pep3D images (25). LC-MS data are assessed almost immediately after collection. Low quality data are rejected without further processing. Samples are repeatedly analyzed until high quality data are obtained from each sample.

Step 2—Raw LC-MS data stored in the profiling mode are processed into denoised, centroid data by the software tool mzXML2dat. Raw LC-MS spectra consist of mainly noisy data. Mining peptide signals from such noisy data is an important step in the analysis of LC-MS data. The software tool mzXML2dat applies advanced signal processing technologies, such as the translation-invariant wavelet transformation (32, 33), to extract peptide signals from raw LC-MS data as described under “Experimental Procedures.” To illustrate the effect of this step, we plotted two Pep3D images side by side.
in Fig. 2: Fig. 2A was generated from raw LC-MS data in profiling mode, whereas Fig. 2B was from the corresponding denoised, centroid data. It is apparent that the denoised, centroid data retained most peptide signals but removed the majority of noise. Due to centroiding, peptide MS signals are much focused along the m/z axis in Fig. 2B. Peptide MS signals in Fig. 2B are slightly less intense than those in Fig. 2A due to the subtraction of background noise. The file size of the denoised, centroid data is significantly smaller than that of the raw data: the average file size from the four repeat LC-MS analyses was 2.0 GB in raw vendor file format and 3.5 GB in the mzXML file format but only 5.8 MB in the mzDAT file format. As a by-product, one may use the mzXML2dat software tool to significantly reduce the file size of LC-MS data, and hence the cost of data storage, without losing a significant amount of peptide information.

Step 3—Peptide features are extracted from denoised, centroid LC-MS data by the software tool PepList. The LC-MS approach aims to compare MS signals of real peptides and to identify the underlying proteins. Hence it is crucial for the
LC-MS approach to reliably extract peptide features from LC-MS data. PepList applies rather restrictive conditions to ensure that most extracted peptide features are in fact derived from peptides as described under “Experimental Procedures.” PepList specifies a peptide feature by peptide monoisotopic mass, charge, and retention time and uses MS signals to determine the abundance and the S/N of the feature. A total of, respectively, 2770, 3105, 3064, and 3162 peptide features were extracted from data of the four repeat LC-MS analyses. In Fig. 2C, we overlayed peptide features from one of the analyses onto the corresponding Pep3D image where identified peptide features are marked with boxes of colors reflecting their S/N values. Clearly most peptide features were supported by the underlying MS signals. The number of peptide features extracted from the four LC-MS data had an average of 3025, a S.D. of 175, and a coefficient of variation (CV) of 5.8%. The consistency in the number of peptide features provides an indication that LC-MS data are reproducible. Some peptide features were produced by the same peptides but in different charge states. Peptide features of the four LC-MS analyses can be combined to, respectively, 2559, 2880, 2872, and 2957 unique peptides. Only 7% of peptide features were produced in more than one charge state. Currently the software considers only charge states up to \([M + 5H]^5^+\) because higher charge states were rarely observed and may be beyond the resolution of the mass spectrometer used. Among peptide features of all four LC-MS analyses, a total of 6744 peptide features were found in the \([M + H]^+\) charge state, 3477 were found in \([M + 2H]^2^+\), 1426 were found in \([M + 3H]^3^+\), 379 were found in \([M + 4H]^4^+\), and 75 were found in \([M + 5H]^5^+\). More than half of all peptide features were found in the \([M + H]^+\) charge state, whereas less than 4% were found in charge states higher than \([M + 3H]^3^+\). In comparison, a small percentage of peptides were identified in the \([M + H]^+\) charge state in a typical LC-MS/MS experiment (20). The discrepancy in the peptide charge distribution between the two experiments may arise from differences in LC and/or MS settings (36). The discrepancy may also reflect the fact that singly charged peptide ions are poorly fragmented and thus under-represented among peptides identified by an LC-MS/MS experiment. In addition, non-peptide components such as chemical contaminants and polymers may also be detected as peptide features (25). Depending on the amount of samples and the sensitivity of the LC-MS system used, PepList normally extracts 1000–5000 reliable peptide features from one LC-MS data set.

Step 4—Peptide features from different samples are aligned against each other by the software tool PepMatch so that the abundance of the same peptide present in different samples can be compared. In most LC-MS analyses, the peptide m/z value is highly reproducible from run to run. Peptide retention times, however, may shift up to a few minutes between two runs even though peptide elution order is rather reproducible. PepMatch evaluates a RTCC to correct any retention time shifts between two LC-MS analyses and aligns peptide features based on their charge, mass, and calibrated retention time as described under “Experimental Procedures.” In Fig. 3A, we plotted the retention time of aligned peptide features in two LC-MS analyses along with the corresponding RTCC. The r.m.s. distance between aligned peptide features and the RTCC was 0.94 min, and the r.m.s. m/z difference between aligned peptide features was 0.025. To illustrate the quality of the alignment, we plotted the intensity of aligned peptide features in the two analyses in Fig. 3B where peptide features with alignment \(p < 0.5\) were plotted in green, whereas those with \(p > 0.5\) were plotted in red. The collapse of most data, including those with \(p < 0.5\), into a diagonal line in the intensity scatter plot indicates that the alignment is quite successful. The scatter plot has the same characters as that of two microarrays. It is apparent that peptide features of high intensity and high alignment \(p\) values are more reproducible than those of low intensity or low \(p\) values. The few outliers in Fig. 3B are most likely due to misalignment as indicated by their low \(p\) values.

Step 5—A peptide versus sample array is generated from aligned peptide features by the software tool PepArray. The peptide array describes the relative abundance of aligned peptide features in different samples and is the output of the SpecArray software suite. The generation of a peptide array from aligned peptide features is described under “Experimental Procedures.” We plotted in Fig. 3C a section of the peptide array that was generated from the four LC-MS analyses. Peptide features in the peptide array were selected if they were aligned to at least two analyses. Although peptide features that were aligned across all analyses are always selected in a peptide array, peptide features that were missed in some analyses may be filtered out by some selection criteria. Because not all peptide features can be aligned across all analyses, a different selection criterion may thus select a slightly different set of peptide features for the peptide array. There were a total of 3188 peptide features in the peptide array among which 2078 had peptide relative abundance across all four analyses, 752 had peptide relative abundance across three, and 358 had peptide relative abundance across two. About 12% of all array elements were blank, which may be caused by peptide misalignment, peptide overlapping, or failure in detecting low abundance peptides in some analyses by MS or by software. No blank element exists if one selects only peptide features that are aligned across all samples when designing a peptide array. Features of the same peptide ionized in different charge states were treated as separated entries in the peptide array. Alternatively one may opt to combine features of the same peptide into a single entry in the peptide array. This alternative is not adopted in PepArray due to the consideration that features in different charge states may have different alignment configurations. Although features in charge states having high MS signal intensities may be aligned across all samples, those in charge states having...
low MS signal intensities may be missed in some or many samples. As a result, it is rather difficult to combine features having different alignment configurations. The 3188 peptide features in the peptide array corresponded to 2913 unique peptides: 2662 peptides (91.4%) had only one feature in the peptide array, 229 (7.9%) had two, 20 (0.7%) had three, and only two (0.1%) had four. The CV distribution of the relative abundance of the same peptide ionized in different charge states had a mean of 0.14, a median of 0.10, and a mode of about 0.03. Hence the relative abundance of the same peptide ionized in different charge states was consistent. The CV distribution of peptide relative abundance evaluated in differ-
ent analyses is an important criterion in assessing the quality of the peptide array. For the four repeat analyses, the data are plotted in Fig. 3D. The CV distribution had a mean of 0.31, a median of 0.24, and a mode of about 0.1. In other words, the LC-MS approach is able to evaluate peptide relative abundance with an accuracy of about ±20–30%, i.e. smaller changes than the accuracy may not be detected by the approach. Such accuracy makes it possible to apply the LC-MS approach to discover discriminatory peptides with about 50% abundance change in different samples.

**Applying LC-MS Approach to the Quantitative Profiling of Mouse Sera**

To demonstrate the application of the LC-MS approach for large scale sample profiling, we collected serum samples from five male and five female mice of the same litter at the age of 22 weeks. Peptides that are N-glycosylated in the intact proteins were isolated in their deglycosylated form using protocols described previously (23, 30). It has been demonstrated that the sample preparation procedure is highly reproducible (23). N-Linked glycopeptide samples were analyzed by an LC-ESI-Q-TOF-MS system under identical settings. We applied the SpecArray software suite to analyze the LC-MS data. In Fig. 4 we plotted four Pep3D images that were generated from these data. Pep3D images of other samples are very similar. We used identical parameters to generate these images so that the gray scale of a peptide spot reflects directly the MS signal intensity of the corresponding peptide. The overall pattern of these Pep3D images closely resembles each other, indicating that the samples had similar peptide contents. The signal intensity from the “male 1” sample is noticeably weaker than that from other samples probably due to lower peptide concentration in the sample. The occurrence of intensity variations between different LC-MS analyses illus-
Quantitative LC-MS Approach for Peptide Profiling

TABLE I

| Sample | 1 | 2 | 3 | 4 | 5 | Average | S.D. | CV |
|--------|---|---|---|---|---|---------|------|----|
| Male   | 2442 | 3621 | 3030 | 3371 | 3723 | 3237 | 519 | 16.0 |
| Female | 2986 | 3621 | 2816 | 2245 | 2700 | 2874 | 500 | 17.4 |

| Sample | 1 | 2 | 3 | 4 | 5 | Average | S.D. | CV |
|--------|---|---|---|---|---|---------|------|----|
| Male   | 2442 | 3621 | 3030 | 3371 | 3723 | 3237 | 519 | 16.0 |
| Female | 2986 | 3621 | 2816 | 2245 | 2700 | 2874 | 500 | 17.4 |

The number of peptide features that were detected from individual samples is listed in Table I. The number ranges from 2245 (“female 4”) to 3723 (“male 5”); the largest difference is 1478. The Pep3D images revealed a strong correlation between the number of peptide features and the overall MS signal intensity as illustrated by Pep3D images of male 1 and male 5 in Fig. 4. On average, about 363 more peptide features were detected from samples of male mice than from samples of female mice. This difference however is statistically insignificant \( (p = 0.29) \). The average number (3056) of peptide features detected from the serum samples is close to that (3025) from the four repeat LC-MS analyses of the plasma sample. But the corresponding CV (17.0%) is about 3 times larger than that (5.8%) of the four repeat analyses. The increase in CV was likely due to biological difference between different mice or due to experimental handling of different samples.

We tested four different feature selection criteria when generating a peptide array from the mouse serum data. The results are listed in Table II. At the lowest stringency, peptide features were selected if they were aligned in at least two of the 10 samples (the “2/10” criterion). The obtained peptide array contained a total of 5319 peptide features. But only 687 of the selected features were aligned across all 10 samples, and 38.9% of all array elements were blank. Some blank elements may be due to the absence (or low abundance) of some peptides in the corresponding serum samples and hence carry biological information. Others may be due to experimental and software artifacts. The latter is likely the cause for most blank elements in this peptide array. At the highest stringency, peptide features were selected if they were aligned in either five of the five samples of male mice or five of the five samples of female mice (the “5/5 or 5/5” criterion). The obtained peptide array contained a total of 897 peptide features. Most (547) of the selected features were aligned across all 10 samples, and only 6.4% of all array elements were blank. Peptide relative abundance stored in this peptide array is much more reliable than that in the 2/10 peptide array. But the more restrictive criteria may also cause some loss of useful information. Although the total number of peptide features varied significantly in peptide arrays that were generated from different feature selection criteria, the number of peptide features that were aligned across all 10 samples was more consistent. Indeed mostly the same peptide features were fully aligned in all peptide arrays. Despite the large variation of array size, the same core information was contained in all peptide arrays that were generated from the same set of LC-MS data. Because the software searched and recovered some missing features, more peptide features were aligned across all 10 samples at a low stringency than at a high stringency.

It is straightforward to evaluate the CV distribution of the relative abundance of the same peptide measured in different samples from a peptide array. We plotted in Fig. 5A the CV distribution evaluated from the 5/5 or 5/5 peptide array. The CV distribution among samples of male mice had a mean of 0.51 and a median of 0.49, whereas the corresponding values among samples of female mice were 0.48 and 0.45, respectively. In comparison with the CV distribution among the four repeat LC-MS analyses of the same plasma sample (see Fig. 3D), CV distributions among samples of different mice shifted significantly toward higher values. Despite the same genetic background of all mice, mouse-to-mouse variations were prevalent, which makes it difficult to discover sex-specific features. (Variations arising from sample collection and sample preparation may also contribute to the larger CV values. It has been demonstrated that variations from sample preparation are small (23).) The CV distribution among samples of all mice of both sexes had a mean of 0.55 and a median of 0.53. There was a small but noticeable shift toward higher values in the CV distribution of all mice in comparison with those of unisex mice. This shift was likely due to biological difference between male and female mice. But it appears that sex-specific variations in peptide abundance were submissive to variations between individual mice. More thorough investigation is needed to verify this finding. The CV distributions evaluated from all four peptide arrays listed in Table II are plotted in Supplemental Fig. S3. Despite the large discrepancy in the number of peptide features between different peptide arrays, all CV distributions were very similar to each other. This reflects the robustness of peptide arrays against different feature selection criteria.

We carried out an unsupervised clustering analysis of the 5/5 or 5/5 peptide array using the Clustering program (rana.lbl.gov/EisenSoftware.htm) (23). A section of the clustering results was plotted in Fig. 5B. Although mice were arranged in the right order with respect to their sex, mice female 4 and “female 5” were clustered with male mice. Similar results were obtained from all four peptide arrays listed in Table II; see Supplemental Fig. S4. Despite the apparent difference be-
between male and female mice, other unknown factors dominate the difference between these samples. To investigate this result further, we recall that blank elements in a peptide array may be traced to two different origins: 1) the corresponding peptides are truly absent (or beyond detection) in the corresponding samples, or 2) experimental and software artifacts cause the corresponding peptides not to be detected or aligned properly in the corresponding samples. Without repeat experiments, it is rather difficult to distinguish between the two origins. The Clustering program treats blank elements as artifacts and completely ignores such data in the unsupervised clustering analysis. Assuming that blank elements in the 5/5 or 5/5 peptide array arose from the first origin, we replaced those blank elements with one-tenth of the lowest relative abundance in the array (0.03) in a separated unsupervised clustering analysis. Mice were properly clustered in the new clustering analysis; see Fig. 5.

### Table II

| Number of samples | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | Total | Blank |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|-------|
| 2/10<sup>a</sup>  | 422 | 626 | 599 | 633 | 664 | 646 | 562 | 480 | 687 | 5319  | 38.9  |
| 5/10<sup>a</sup>  | 146 | 342 | 425 | 456 | 424 | 668 | 2461| 4690| 35.6|       |       |
| 2/5 or 2/5<sup>b</sup> | 221 | 472 | 483 | 557 | 618 | 627 | 549 | 477 | 686 | 4690  | 35.6  |
| 5/5 or 5/5<sup>b</sup> | 12  | 46  | 92  | 204 | 543 | 897 | 6.4 |    |    |       |       |

<sup>a</sup> Two (five) out of the 10 samples of both male and female mice.

<sup>b</sup> Two (five) out of the five samples of male mice or two (five) out of the five samples of female mice.

A total of 566 (63.1%) peptide features in the 5/5 or 5/5 peptide array had a $S/N$ above 10. To discover specific peptide features that can distinguish male mice from female mice, we applied Student’s $t$ test to evaluate the discriminatory $p$
value of the 566 peptide features from peptide relative abundance stored in the peptide array. The complete results are listed in Supplemental Table S1. The five most discriminatory peptides had a $p$ value of $8.6 \times 10^{-6}$, $2.0 \times 10^{-3}$, $2.2 \times 10^{-3}$, $3.0 \times 10^{-3}$, and $3.1 \times 10^{-3}$. As an example, we plotted the zoomed Pep3D images and the relative abundance of the most discriminatory peptide in Fig. 6. The Pep3D images were generated with exactly the same parameters so that the gray scale of the circled peptide feature in individual samples directly reflects the corresponding peptide MS signal intensity. Despite the large variation in the relative abundance of the peptide among different mice of the same sex, the relative abundance of male mice was consistently higher than that of female mice in this example. Another example containing the second and the third most discriminatory peptides is given in Supplemental Fig. S5. The two features had a mass difference of 0.994 and a retention time difference of 1.7 min. It is not clear whether the two features were chemically related, e.g. by deamidation (37). In this study, we did not find any peptide features that were present in only one sex; see Table II. A total of 15 peptides had a $p$ value less than 0.01. Due to the small sample size and the prevalent mouse-to-mouse variations in peptide abundance, very few peptide features had a statistically significant $p$ value in distinguishing male mice from female mice. Nevertheless analyzing the peptide array allowed us to discover such discriminatory peptides, which were specified by their monoisotopic mass, charge, and retention time and can be selected for targeted MS/MS identification (23).

**DISCUSSION**

We present here a new software suite, SpecArray, that generates peptide arrays from sets of LC-MS data. We used data collected from four repeat LC-MS analyses of a glyco-
advances will enhance the quality of LC-MS data. On data
tools for transcriptomics, we anticipate a broad application of the LC-MS approach and the SpecArray software suite in quantitative proteomics. This new platform of quantitative proteomics can be especially useful when one needs to measure the protein contents of a large number of substantially similar samples as in the case of biomarker discovery, time course studies, knock-out experiments, etc. By first collecting information on discriminatory peptides and then using targeted MS/MS to identify the corresponding proteins, the platform may also provide researchers a better chance than most current proteomic platforms of discovering proteins of biological or physiological interest. The output format also makes it natural to adopt some existing microarray analysis tools to the downstream analysis of SpecArray output results, further simplifying the extraction of biological information from LC-MS data.

There is no doubt that the LC-MS approach and the SpecArray software suite are still at their early development stage. Experimental and software artifacts are present in peptide arrays. Improvement on sample source, LC-MS analysis, and data analysis is underway. More and more attention is paid to minimize potential artifacts that may be introduced during sample collection, sample storage, and/or sample preparation (4). On LC-MS analysis, ongoing efforts are aimed to reduce run-to-run variations in peptide retention time and peptide MS signal intensity and to cut down the overlap of peptide features by simplifying sample complexity, increasing MS resolution, and improving LC separation. We have demonstrated recently (data not shown) that one may achieve a consistent and reproducible LC-MS analysis of about a hundred samples after optimizing LC-MS components such as LC column, LC gradient, and MS settings. New developments such as chip-based LC columns and improved nanoflow chromatography systems also significantly increase the resolution power and the reproducibility of peptide separation (38). New MS instruments such as the linear ion trap Fourier transform (LTQ-FT) mass spectrometer offer much improved accuracy and sensitivity in measuring peptide ions (39). All these experimental advances will enhance the quality of LC-MS data. On data analysis, some low abundance or overlapping peptide features may be difficult to extract reliably. A small percentage of peptide features may be misaligned by the software tools. Sophisticated analytic methods are under development to detect low abundance peptide signals, decompose overlapping peptide features, and unambiguously determine peptide features that are absent from a sample. These new methods will definitely make information stored in peptide arrays more reliable. In short, the prowess of the LC-MS approach and the usefulness of peptide arrays are expected to improve along with the constant improvement on sample source, LC-MS analysis, and data analysis.

The SpecArray software suite is written in C language. The current version runs on the Linux operating systems. A new version for the Windows operating systems is planned. Just like other software tools developed by our group, the SpecArray software suite will be freely distributed under an open source license upon publication and will be available at tools.proteomecenter.org/software.php.
Quantitative LC-MS Approach for Peptide Profiling

12. Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fenselau, C. (2001) Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal. Chem. 73, 2836–2842

13. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics 3, 1154–1169

14. Fenn, J. B., Mann, M., Meng, C. K., Wong, E. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64–71

15. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. 60, 2299–2301

16. Hunt, D. F., Yates, J. R., III, Shabanowitz, J., Winston, S., and Hauer, C. R. (1996) Protein sequencing by tandem mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 93, 6233–6237

17. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. Chem. Rev. 101, 269–295

18. 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

19. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3651–3667

20. Li, X. J., Zhang, H., Ranish, J. A., and Aebersold, R. (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal. Chem. 75, 6648–6657

21. 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

22. 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

23. 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, 144–155

24. Blagoev, B., Ong, S. E., Kratchmarova, I., and Mann, M. (2004) Temporal analysis of phosphosyntome-dependent signaling networks by quantitative proteomics. Nat. Biotechnol. 22, 1139–1145

25. Li, X. J., Pedrioli, P. G., Eng, J., Martin, D., Yi, E. C., Lee, H., and Aebersold, R. (2004) A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. Anal. Chem. 76, 3856–3860

26. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal. Chem. 75, 4818–4826

27. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G., Foss, E., Mao, Y., and Emili, A. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography–tandem mass spectrometry. Mol. Cell. Proteomics 3, 984–997

28. Bondarenko, P. V., Chelius, D., and Shaler, T. A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography–tandem mass spectrometry. Anal. Chem. 74, 4741–4749

29. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. 95, 14863–14868

30. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat. Biotechnol. 21, 660–666

31. Pedrioli, P. G., Eng, J. K., Hubyler, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. Nat. Biotechnol. 22, 1459–1466

32. Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11, 674–693

33. Cowman, R. R., and Donoho, D. L. (1995) Translation-Invariant De-Noiseing, in Wavelets and Statistics (Antoniadis, A., and Oppenheim, G., eds) p. 125–150. Springer-Verlag, New York

34. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 6, 229–233

35. Wehofs, M., Hoffmann, R., Hubert, M., and Spengler, B. (2001) Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples. Eur. J. Mass Spectrom. 7, 39–46

36. Li, Y., and Cole, R. B. (2003) Shifts in peptide and protein charge state distributions with varying spray tip orifice diameter in nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. 75, 5739–5746

37. Teshima, G., Porter, J., Yim, K., Ling, V., and Guzzetta, A. (1991) Deamidation of soluble CD4 at asparagine-52 results in reduced binding capacity for the HIV-1 envelope glycoprotein gp120. Biochemistry 30, 3916–3922

38. Yrin, A., Killeen, K., Brennen, R., Sobek, D., Werlich, M., and van de Goor, T. (2005) Microfluidic chip for peptide analysis with an integrated HPLC column, sample enrichment column, and nanoelectrospray tip. Anal. Chem. 77, 527–533

39. Syka, J. E., Marto, J. A., Bai, D. L., Horning, S., Senko, M. W., Schwartz, J. C., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., Shabanowitz, J., and Hunt, D. F. (2004) Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. J. Proteome Res. 3, 621–626