Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry[*][S]

Dragan Radulovic[‡§¶], Salomeh Jelveh[‖], Soyoung Ryu[§], T. Guy Hamilton[‡‡], Eric Foss[**], Yongyi Mao[§§], and Andrew Emili[‖]**

We have developed an integrated suite of algorithms, statistical methods, and computer applications to support large-scale LC-MS-based gel-free shotgun profiling of complex protein mixtures using basic experimental procedures. The programs automatically detect and quantify large numbers of peptide peaks in feature-rich ion mass chromatograms, compensate for spurious fluctuations in peptide signal intensities and retention times, and reliably match related peaks across many different datasets. Application of this toolkit markedly facilitates pattern recognition and biomarker discovery in global comparative proteomic studies, simplifying mechanistic investigation of physiological responses and the detection of proteomic signatures of disease. ***Molecular & Cellular Proteomics 3:984–997, 2004.***

Protein expression profiling is the study of the key functional molecules—the proteins—present in a biological system. In practice, it involves determining the identities, abundance, and post-translational states of the myriad of proteins present at specific time points within the life cycle of an organism (1). Because cells respond to physiological cues and environmental perturbations, the proteome serves as a unique and informative "readout" of phenotypic state (2). By providing an overview of entire biochemical pathways, expression profiling complements and extends traditional single molecule analyses in the generation of testable hypotheses regarding the biological roles of proteins in health and disease (3).

Many of the successes of therapeutic intervention have evolved from improvements in the ability to diagnose, stage, and stratify subgroups of patients who may respond differently to various management strategies (4). Despite these advances, the treatment of many diseases, such as cancer and cardiovascular disease, still suffers from the fact that most patients present at late stages of illness. Earlier detection of pathology is highly beneficial to patient outcomes (5), yet there are few effective diagnostic tools for recognizing early stage disease or prognostic tools for identifying those at high risk of dying or being nonresponsiveness to therapy (6). Development of satisfactory therapeutics is also hampered by a lack of informative bioassays (6). As a result, biological markers are urgently needed to improve the efficacy of clinical intervention, the reliability of clinical trials, and the validation of leads and targets (3, 6).

DNA microarrays are commonly used to detect differences in gene expression between different physiological states (7), including global changes in mRNA abundance across repeat experiments, distinct experimental perturbations, discrete time points, or large patient cohorts (7). Pattern recognition algorithms can then be applied to sort and classify samples based on their expression profiles (8). Nonetheless, it is likely that pathophysiological mal-adaptations associated with common pathologies, such as diabetes and cancer, are more accurately reflected in the proteomic patterns of disease-affected tissues (3), especially in samples with little messenger RNA (*e.g.* serum) (9).

To date, most clinically useful protein biomarkers have either been found serendipitously or through limited candidate evaluation based on hypotheses regarding disease action (3, 10). The lack of effective generic procedures for routinely detecting differences in global protein patterns across many different samples hinders the discovery of new biomarkers (3, 5). This constraint is particularly apparent in a clinical setting (5, 11), where specialized analytical procedures are often required to derive useful qualitative and quantitative information from the minuscule amounts of protein typically found in patient specimens, such as a biopsy. Furthermore, while sensitive immunoassays can be used to prospectively validate a biomarker, they are generally not well suited to the discovery of new biomarkers (3).

From the ‡Department of Statistics, Yale University, New Haven, CT; §Department of Mathematics, Florida Atlantic University, Boca Raton, FL; ‖Program in Proteomics and Bioinformatics, Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada; **Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA; and §§School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada

Received May 10, 2004, and in revised form July 14, 2004

Published, MCP Papers in Press, July 21, 2004, DOI 10.1074/mcp.M400061-MCP200

MS has emerged as a key enabling technology for protein expression profiling (1, 12). Recent ground-breaking studies have demonstrated the utility of combining MS-based profiling and computer-based pattern recognition as a means of detecting proteomic signatures of cancer in blood (13, 14). However, the relatively simple MS instrumentation used in these pioneering studies was biased toward the detection of low molecular mass proteins (13). Moreover, it did not allow for ready protein identification, which is critical if such biomarkers are to form the basis of a simplified, widely adopted diagnostic (5, 10). Reliable methods for determining both the identity and quantity of large numbers of proteins across many different clinical samples are therefore urgently needed to test and prospectively validate the hypothesis that compensatory responses to disease are reflected by changes in the proteomic patterns of blood or tissue (10, 15). Moreover, there is a parallel need to develop rigorous statistical methods to evaluate the significance of any differences detected (16). This is particularly true of global proteomic studies subject to numerous sources of variation, both experimental and biological in origin.

Gel-free protein profiling procedures coupling capillary-scale HPLC to data-dependent MS/MS (LC-MS) present an exciting new paradigm for proteomic screening (1, 12). In particular, multidimensional protein identification technology (17, 18) and isotope-coded affinity reagents (19) now allow for the "shotgun" profiling of hundreds of proteins in a single experiment, albeit with a significant expenditure of time and effort. The clinical impact of these methods has been limited to date (3), however, in part due to problems associated with the reproducibility of LC-MS (20, 21), as well as to difficulties in extracting clinically relevant information from the limited number of samples that can practicably analyzed using these specialized methods (16).

In an effort to improve the reliability of LC-MS-based profiling studies, Smith and colleagues (22) have reported the utility of advanced equipment, FT-ICR MS and HPLC pumps capable of sustained performance at >10,000 psi. While this strategy circumvents many of the problems associated with traditional profiling procedures, it relies on technologies that are not widely available to the broader biomedical community. Moreover, it does not address fundamental issues concerning the statistical evaluation of multivariate proteomic datasets for the purpose of biomarker discovery (16).

Becker and colleagues (20) recently introduced an alternative computational method for detecting differential protein abundance by LC-MS without the need for isotopic labeling or advanced instrumentation. Their approach relies on the roughly linear relationship of MS signal as a function of peptide ion concentration. Proprietary data processing algorithms were then used to track quantitative variation in peptide signal across different LC-MS datasets. A key step in minimizing sample dispersion was the use of a "time wrapping" alignment algorithm to correct for spurious deviations in recorded ion maps, resulting in modest (~25%) coefficients of variation across integrated peak intensities. Significant computational cost was observed with increasing sample complexity (20), restricting the effectiveness of this first-generation platform for pattern recognition across multiple complex proteomic datasets (10, 15, 20, 23, 24).

Experimental repetition, pattern recognition, and mathematical algorithms can minimize the effects of unwanted noise and spurious signal fluctuation (15, 25). Here, we report the development of a more advanced generation of computer algorithms, statistical data-mining procedures, and software built upon these principles that greatly facilitate large-scale protein expression profiling of mammalian tissue samples using basic gel-free shotgun profiling procedures and standard LC-MS instrumentation. We show that this informatics toolkit allows for systematic global comparison and classification of complex tissue proteomic samples, speeding discovery of biologically relevant proteomic biomarkers.

EXPERIMENTAL PROCEDURES

Programming and Data Pre-processing

The software suite encompasses a set of integrated modules (described below) written in Fortran that produce interlinked data tables. A high-level schematic of the program workflow, highlighting key functionalities (algorithms and scripts) of the platform, is provided in Supplemental Fig. S9. Plots were prepared using Microsoft Excel.

Signal Filtering Algorithm

After conversion of the raw LC-MS data files to text format, a Perl script is used to parse out all irrelevant MS/MS scan data. Peak *m/z* ratios in the retained scans are rounded off to the closest integer and binned (±0.5 *Th*).

Applying the M-N rule—The program processes individual nominal *m/z* ion traces $\{Z_i\}$ (where Z_i is the intensity on the i^{th} scan header, and *m/z* is fixed) and computes a robust center, *C*. We suggest $C = 30\%$ of the trimmed mean of $\{Z_i\}$, although $C =$ median of $\{Z_i\}$ produces reliable results. The data are smoothed using moving averages. That is, for a given fixed *m/z* slice, the feature intensities are transformed by averaging over a fixed window of five consecutive scans. Next, for predefined constants *M* and *N*, the algorithm extracts only those features of Z_i greater than M*C for *N* observations in a row. For example, if $C = 3{,}000$, and the *M-N* rule is set to 5–3, we would declare Z_i a pixel if Z_{i-1}, Z_i, and Z_{i+1} are all larger than 15,000 (*i.e.* the ion signal intensity was >5*C for at least three scans in a row). Finally, for a declared constant, $L_i = 2^{i-1}1{,}000$, $i = 1, \ldots 5$, a set of M-N constants (rules) producing L_i pixels are chosen.

The *M-N* computation scales linearly with the number of experiments, and application of the algorithm (five levels per analysis, with each level taking ~1 min of CPU time) is generally not a limiting factor.

Normalization—A basic form of data normalization is carried out by a two-step mechanism. First, before application of the *M-N* rule, the feature intensities of individual datasets are transformed into an integer (ranging between 1–10,000 arbitrary units) by dividing all feature intensities with a constant, *K*. The constant, *K*, is chosen as the minimum value such that the total number of features with intensities larger than *K**10,000 equals 100. In other words, there are only 100 features with intensities above the cut-off value of 10,000. The second mechanism is designed to detect a potential normalization problem. It monitors both the *K* constants and the *M-N* rules produced for each corresponding pamphlet, and issues an alarm (error message) if these (feature intensities and pixel count) differ for more than 20%. No

TABLE I
Evaluation of the number of features (pixels) extracted using a given M-N rule

M-N	Exp.[a]										Noise[b]
	1	2	3	4	5	6	7	8	9	10	
9–6	724	763	786	941	656	647	486	680	574	414	0
6–9	461	537	619	655	480	439	336	428	394	270	0
6–4	3,608	3,730	3,926	3,995	3,015	2,830	2,301	3,057	2,818	2,708	0
3–3	15,595	16,606	16,769	16,570	13,247	11,250	13,691	14,122	11,572	13,040	34

[a] Exp. = genuine protein sample experimental LC-MS dataset.
[b] Noise = control sample.

error messages were generated for any of the datasets reported in this study.

Evaluating System Stringency—Generally, the first scans acquired during LC-MS consist of noise (assuming no sample bleed-through). Thus, by concatenating the first ~10% of total acquired scans obtained from 10 independent LC-MS analyses of a yeast cell tryptic digest, we constructed a virtual experimental dataset consisting entirely of nonpeptide ion noise. We then evaluated the performance and sensitivity of various elaborations of the *M-N* rule with each of the genuine LC-MS datasets *versus* the noise dataset alone. As seen in Table I below, the *M-N* rule approach proved to be highly stringent. Even a liberal Level 3–3 threshold, which extracts ~11–16,000 features on average from each of the peptide profiles, resulted in detection of a few spurious peaks in the control dataset (*i.e.* only 34 false-positives satisfying the rule detected).

Peak Detection Algorithm

Basic Extraction Principle—All adjacent neighboring (touching) features (pixels) are grouped and assigned the same peak ID.

Peak Extraction Algorithm—The algorithm proceeds in an iterative stepwise manner, starting from Level 1 (1,000 pixels) using the basic extraction principle. Each distinct peak (defined by discrete scan headers and *m/z*) is assigned a unique ID. Next, the algorithm progressively adds features from successive levels (Level 2, then 3, and so on). If these added features overlap with more than one peak, the groupings are split (IDs reassigned) such that, at most, only a single peak is retained from a previous level. That is, the overlapping features are bisected at each additional level by computing a discriminating line that separates (and hence preserves) the original peaks.

Peak Alignment Algorithm

Because a pamphlet can be represented as a collection of pixels with *X* and *Y* coordinates, where *X* is the scan number and *Y* the nominal *m/z* (for the alignment problem, we need not consider intensity values), we let $P_1 = \{X_{i,1},Y_{i,1}\}_{i=1}^{L1}$ and $P_2 = \{X_{i,2},Y_{i,2}\}_{i=1}^{L2}$ formally represent two different pamphlets. A relatively simple but robust measure of similarity between two datasets (Matching) is then calculated based on the percentage feature (pixel) overlap, defined as:

$$\mathit{Matching}\ \ (P_1, P_2) = \max(L_1, L_2)^{-1} \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} 1_{X_{i,1}=X_{j,2}} 1_{Y_{i,1}=Y_{j,2}}$$

Given a smooth increasing function, $F(X,Y)$, we let $\tilde{P}_2 = \{F(X_{i,2},Y_{i,2}),Y_{i,2}\}_{i=1}^{L_2}$ serve as a time-transformed second pamphlet. The alignment problem now reduces to finding the function, *F,* that maximizes Matching$(P_1,\tilde{P}_2)$. We considered only functions of the form

$$F(x, y) = \sum_{j=1}^{D} \left(\sum_{i=1}^{E} (A_{i,j}x + C_{i,j}) 1_{a_{i-1} \le x \le a_i} \right) 1_{b_{j-1} \le y \le b_j}$$

where $A_{i,j}$ and $C_{i,j}$ are chosen such that $|F(x_1,y) - F(x_2,y)| \le K_1|x_1 - x_2|$ and $|F(x,y_1) - F(x,y_2)| \le K_2|y_1 - y_2|$, with *Lipschitz* constants $0.9 \le K_1 \le 1.1$ and $-0.05 \le K_2 \le 0.05$. The partitions $\{a_i\}_{i=1}^{E}$ and $\{b_j\}_{j=1}^{D}$were uniform, while constants *D* and *E* were set as 5 and 6, respectively. The optimization now reduces to finding the optimal values for $A_{i,j}$ and $C_{i,j}$. Accelerated Random Search (26) was the optimization schema of choice, because it is robust and easy to implement. As a further measure of peak matching, a final "wobble" function is applied wherein a peak is allowed to move (±1–2% of total scan headers) in order to find the nearest adjacent peak in a different experimental dataset. Generally, even for complex mixtures and higher level pamphlets, there is an extremely low probability that there will be two (or more) peaks exactly equidistant. If that happens, the software will pick one (randomly) and an alarm (error message) is produced (we have rarely seen this form of error).

The alignment algorithm is computationally intensive and scales with the square of the number of experiments (*e.g.* pair-wise dataset matchings). Sufficient RAM is therefore suggested to carry out the most demanding calculations in memory.

Peptide Quantitation

A peptide quantitation module processes input LC-MS datasets and outputs signature expression profiles, along with a measure of statistical variation. Peak integration is performed by summing the intensities of grouped features across adjacent MS scans recorded in full scan mode.

Protein Sample Preparation

Human serum was prepared according to standard practice. Purified human heart troponin complex was obtained from a commercial source. For the mouse protocol, we removed chow from the fasted mice in the morning and sacrificed all mice 24 h later. The strain used (27) was an inbred cross of C57BL6 × 129. Liver extract were prepared as reported by Kislinger *et al.* (28). Protein fractions were precipitated, solubilized in urea, resuspended in 100 mM NH_4HCO_3 with 1 mM $CaCl_2$ (pH 8.5), and digested with Poroszyme trypsin beads (Applied Biosystems, Foster City, CA). The resulting peptide mixtures were solid phase extracted with SPEC-Plus PT C18 cartridges (Ansys Diagnostics, Lake Forest, CA) and stored at −80 °C until further use. Synthetic peptides were obtained from Sigma Aldrich (St. Louis, MO).

LC-MS Analysis

Peptide mixtures were subjected to capillary-scale LC-MS using a quaternary HPLC pump coupled online to an LCQ DECA ion trap MS (Thermo Finnigan, San Jose, CA) essentially as described (29). Briefly, a fused-silica microcolumn (100 μm i.d. × 365 μm o.d.) was pulled with a Model P-2000 laser puller (Sutter Instrument Co., Novato, CA) and packed with ~5 cm of 5-μm C18 reverse-phase material (Zorbax XDB-C18; Agilent, Palo Alto, CA). After loading, the column was placed in-line with the ion source and the peptides eluted with a linear gradient [100% buffer A (5% ACN, 0.02% heptafluorobutyric acid, 0.5% acetic acid) to 80% solvent B (100% ACN) over 45, 60, or 90 min] at a tip flow rate of ~0.3 μl/min using a split line. Eluting peptide ions were analyzed with alternating MS modes, using a full scan mass range of 400–1,600 *m/z* followed by data-dependent CID. A dynamic exclusion list was used to limit collection of redundant CID spectra.

Protein Identification

Peptide fragmentation product ion spectra were sequence-mapped against a database of nonredundant protein sequences (Swiss-Prot) using the SEQUEST software algorithm (30) running on a multiprocessor computer. The probability-based evaluation algorithm STATQUEST (28) was used to filter all putative matches based on a ≥95% likelihood of predicted accuracy. Functional annotation was obtained from Swiss-Prot.

RESULTS

In a typical LC-MS-based profiling experiment, peptide mixtures derived from protein digests are fractionated using a chromatography column packed with reverse-phase media and electrosprayed into an online MS/MS instrument. In data-dependent experiments, the instrument alternatively records the signal intensities, *m/z* ratios, and retention times of all of the detectable eluting peptides, as well as the fragmentation pattern of individual peptides subject to CID. The recorded ion peak intensities reflect intrinsic electrochemical properties of a peptide (31) and its relative concentration (20, 32). Peptide sequence can often be deduced by database searching of the daughter ion spectra (1, 12).

Modern LC-MS systems can resolve hundreds of peptides (1, 12). A typical dataset, consisting of a multiplexed stream of co-eluting ion peaks (or ion map) acquired on a quadrupole ion-trap, is shown in Supplemental Fig. S1. We refer to such data as an empirical profile. Inspection of representative total ion chromatograms demonstrates the general reproducibility of LC-MS (Supplemental Fig. S2). Nonetheless, even under controlled conditions (20, 22), both stochastic system performance variation and chemical and electronic noise can affect the relative position, width, amplitude, and shape of individual peaks (Supplemental Fig. S3). We refer to this peak artifact as drift and distortion.

Extracting Quantitative Information from LC-MS Datasets

An effective way for enhancing signal-to-noise is by performing repeat analyses (25). The challenge, then, in profiling experiments is to detect related peaks across different datasets, despite peak drift and distortion. Sophisticated filtering techniques, such as time series, Fourier transform, expectation-maximization, and certain pattern recognition algorithms (25), can be challenging to implement in an effective and practical manner (D.R. and Y.M., unpublished observations). Hence, we chose to develop robust, assumption-free, threshold-like data filtering algorithms for detecting real differences in peak number and intensity, yet that would not be overly sensitive to the effects of spurious noise (25).

Step 1: Data Filtering and Signal Extraction—The key task is to reliably extract genuine signal, representing individual peptides, from large collections of interpolated MS full scans. To reduce noise, we created a filtering algorithm (see "Experimental Procedures") to pre-process the spectra, binning (by nominal *m/z*) and smoothing the data using "moving averages." Any signal above a fixed threshold, $M{\cdot}Tm$ (where Tm represents the centroid "trimmed mean" intensity and M a pre-defined coefficient), for N consecutive scans is recorded as a feature in a data matrix, wherein the x coordinate stores the scan number, the y coordinate the nominal *m/z*, and the z coordinate the recorded ion intensity. We refer to this matrix as a data pamphlet.

A feature extraction algorithm is then used to select an optimal set of $(M,N)_i$ rules to acquire a predefined series of L_i features for a geometrically increasing sequence, $L_i = 2^{i-1}$ 1,000. [Pamphlets with L_i features are referred to as Level i pamphlets.] The algorithm starts conservatively, extracting the most prominent ion features first, and then progressively adding features until the cutoff is met. Statistical analysis (see "Experimental Procedures") suggests [$M = 3,n = 3$] as a generally acceptable lower threshold, resulting in many discrete features (depending on sample complexity) with little specious background. Examples of Level 5 and 2 pamphlets (16,000 and 2,000 features), generated by LC-MS analysis of a yeast cell extract, are shown in Fig. 1, *a* and *b*, respectively. Despite evidence of crowding, higher resolution "zoom in" reveals good peak discrimination (Fig. 1*c*).

Theory and empirical evidence suggest that peak intensities are not independent, but rather can be negatively correlated due to ion-ion interactions leading to signal suppression (33). This effect is often pronounced with contaminants such as detergents or polymers that perturb ionization efficiency and are manifested by prominent vertical "drop-out" strips in a pamphlet (Supplemental Fig. S4*a*). Because such suppression can lower median scan feature intensities (Fig. S4*b*), such artifacts can be corrected using adjacent scan median values (Fig. S4*c*). However, since gross ion suppression is nearly always related to sample contamination, it may be preferable to declare an affected pamphlet invalid and repeat the analysis.

Step 2: Peak Definition—Global proteomic studies depend on the comprehensive accounting of peptide patterns. We therefore developed a contour detection algorithm*,* based on established boundary detection and integration techniques (see "Experimental Procedures"), to automate peak definition. By first converting feature intensities to unity, pamphlets can be treated as a bitmap (collection of pixels) to simplify data processing and visualization. The routine then subgroups (assigns a unique ID to) nearest neighbor features (pixels), starting off with a Level 1 pamphlet and then iteratively importing additional features from higher-level pamphlets (see "Experimental Procedures"). By maintaining peak independence, the algorithm reliably sorts large numbers of closely spaced fea-

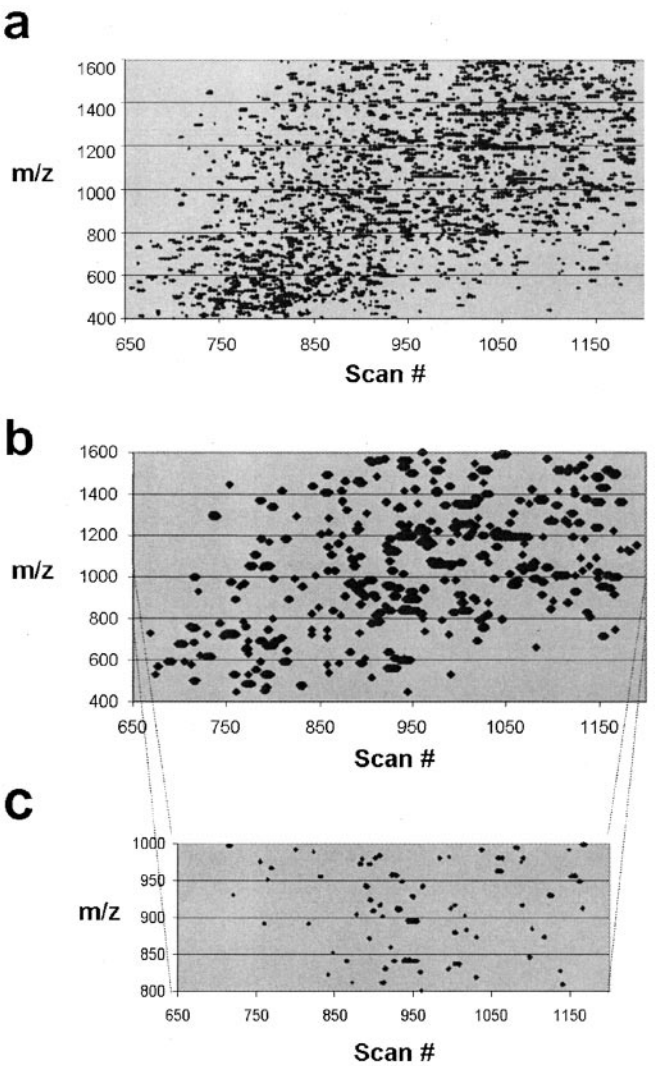

FIG. 1. **Feature detection in LC-MS-based proteomic profiles.** Representative examples of (*a*) Level 5 and (*b*) Level 2 data pamphlets, harboring 16,000 and 2,000 peptide features (*black pixels*), respectively. The datasets were generated by extracting ion peak signal above a specified threshold from an LC-MS dataset obtained for a yeast tryptic digest. *c*, higher resolution "zoom in" of the Level 2 pamphlet reveals good peak dispersion. *x*-axis, scan number; *y*-axis, *m/z* ratio.

tures without arbitrarily fusing unrelated adjacent peaks (Fig. 2; compare *panels a* and *b*).

As a measure of the reliability of peak detection, we mapped high-confidence (*p* value <0.05) peptide sequences, derived by searching CID spectra against a comprehensive protein-sequence database using the SEQUEST (30) and STATQUEST (28) algorithms, on to a data pamphlet. As expected, peak-to-peptide overlap was extensive, with few outliers (Fig. 2*c*). This analysis also affirms a well-known fact that considerably more peptide peaks are detectable in full scan MS mode than can be identified in the same time frame using the quasi-stochastic CID process.

Step 3: Correction of Peak Drift and Distortion and Peak Alignment—Biomarker discovery depends on the careful examination of protein abundance across multiple samples (5, 10). Although repeat LC-MS analyses are generally highly similar, an obvious alignment problem presents itself following attempts to overlay related data pamphlets (Fig. 3*a*). This deviation can be nonlinear along both peptide retention time and (albeit less prominently) *m/z*. In the example provided, peak drift and distortion are more pronounced at the beginning of the analyses.

To surmount this problem, we devised an efficient pamphlet alignment algorithm that uses self-optimizing 2D smooth spline transformation (see "Experimental Procedures") to correct for nonlinear deviation in peak patterns across both the *m/z* and time axes. The algorithm optimizes feature overlap between an input pamphlet and a second reference pamphlet. Exhaustive pair-wise alignments are performed to find a global optimum with larger sample sets.

Last, to compensate for residual random (nonsmooth) variation, each of the peaks detected by contour mapping is "wobbled" to maximize peak overlap (see "Experimental Procedures"). Although this local optimization is limited in scope (±1% total scans), it provides an added measure of peak matching. Fig. 3*b* illustrates the considerably improved peak matching achieved by this multistage procedure.

We note that, just as data normalization is often used to correct for systemic signal discrepancies in microarray studies (34), global peak intensities of different datasets can likewise first be normalized by adjusting median feature intensities to unity prior to matching. However, many substantive issues are raised by normalization procedures (34). In our experience, well-controlled sample preparation and LC-MS procedures serve sufficiently well in most instances such that data normalization is not a major concern. Nevertheless, normalization may improve the inferences that can be drawn from comparisons of proteomic datasets generated by different sources and locations.

Computational time is another obvious constraint here. We have worked under the general guiding principle that the routine application of our informatics platform should not exceed the time necessary to complete the LC-MS analyses themselves (that is, the rate of data production should not exceed data processing capacity). In fact, running the software on a basic single Pentium CPU Win/PC workstation is generally more than sufficient to keep up with the data output of a dedicated LC-MS system collecting spectra more or less around-the-clock.

The alignment algorithm is by far the most computationally intensive and scales with the square of the number of experiments (*e.g.* pair-wise dataset matchings). While the inherent computational difficulties (multidimensional optimization generally requires 5–10 min of CPU time per matching) will be hard to speed up, the concept of "Mother pamphlet" was specifically designed to tackle the quadratic increase.

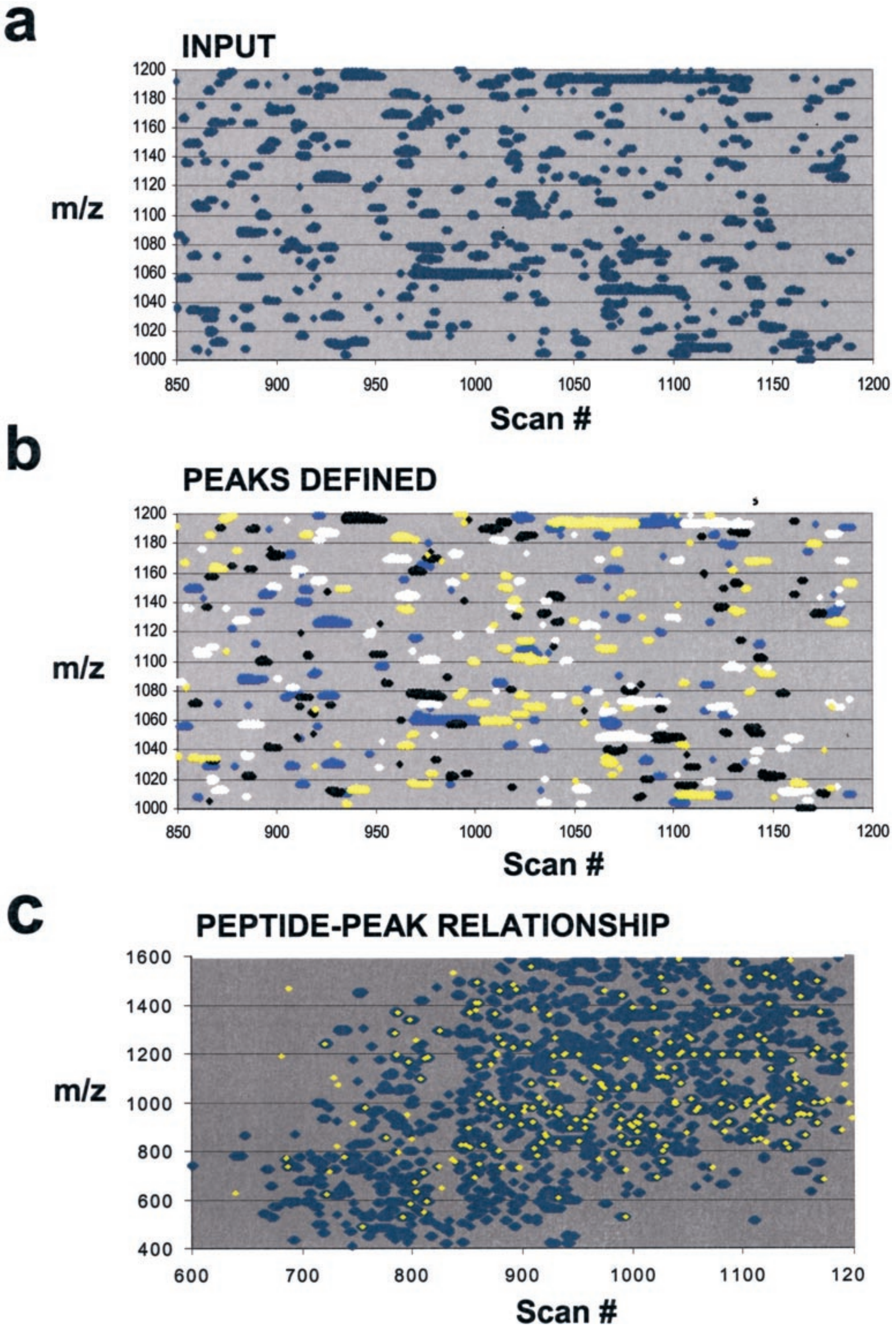

FIG. 2. **Automated peak detection.** Automated detection of discrete peaks using boundary detection and integration (contour mapping) techniques. Examples of an (*a*) input pamphlet and (*b*) post-analysis of this same profile, with individual peaks highlighted in alternating colors to enhance visual discrimination. *c*, example of the close correspondence of high-confidence peptide sequences identified by CID (*yellow*) to pamphlet peaks detected by the software (*blue*).

Step 4: Quantitative and Qualitative Proteomic Comparisons—Once aligned, LC-MS datasets can be directly compared in a systematic manner. As a simple first-pass measure of similarity, feature overlap, [Matching (*i*,*j*)], is calculated (see "Experimental Procedures"). The higher the ratio, the more closely related two data pamphlets are deemed to be.

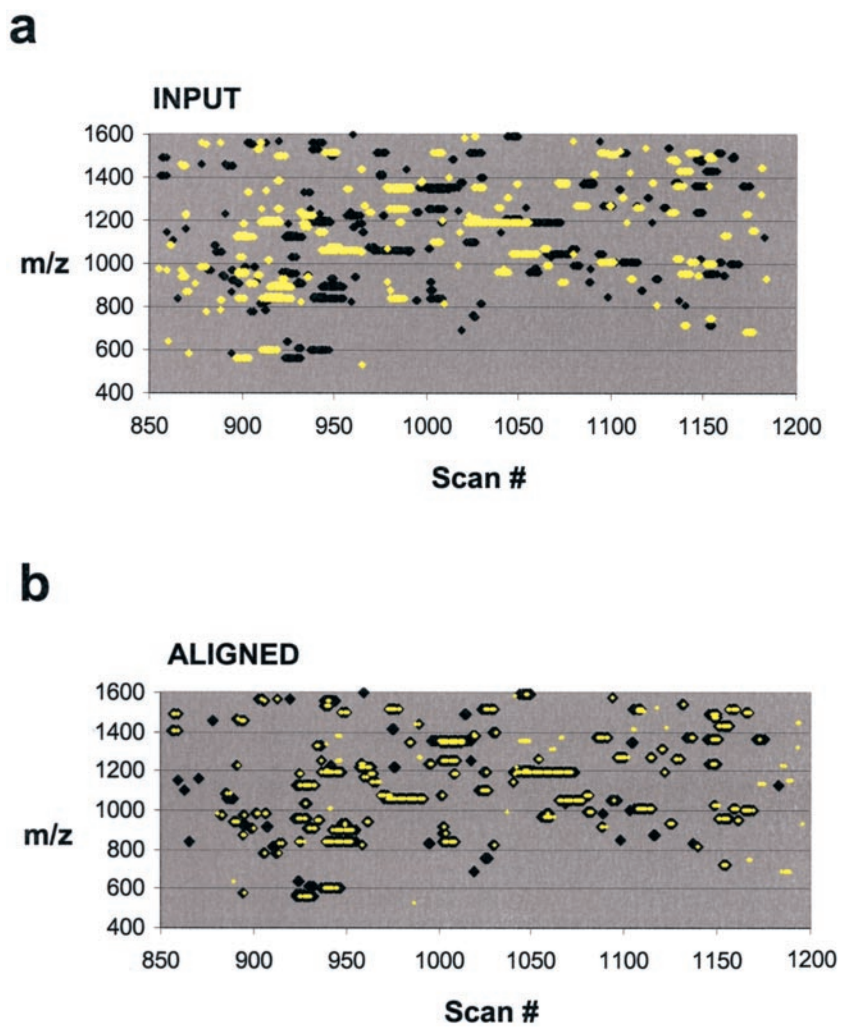

FIG. 3. **Profile alignment and peak mapping.** Nonlinear deviation in peak patterns recorded in repeat LC-MS datasets. *a*, a misalignment problem presents itself following attempts to overlay two related data pamphlets. Peak drift and distortion are more pronounced at the beginning of the analyses. *b*, data overlap is considerably improved after application of the peak alignment algorithm.

Whereas the overlap of the input datasets shown in Fig. 3*a* is only 12.3%, this improves to 77.2% post-alignment (Fig. 3*b*).

For quantitative peak comparisons, grouped feature intensities are summed. Consistent with earlier reports (20), standard titration curves recorded for model peptides exhibited linear signal responses after peak processing and quantification, with residual variation mitigated by repeat analyses and signal averaging (Fig. 4*a*). Moreover, a good correlation ($R^2 = 0.84$) was observed in scatter $\log_{10}$ plots of peak intensities measured for >400 peptides reproducibly detected in the two aligned datasets reported in Fig. 3*b*, with relatively few outliers and only modest dispersion at lower signal-to-noise ratios (Fig. 4*b*). Importantly, >93% of the peaks exhibited 2-fold or less deviation in observed signal intensities, an established benchmark of reproducibility used in microarray studies (34), across a 3–4 order of magnitude dynamic range.

As a general test of sensitivity, we evaluated the ability of the software to detect modest differences in sample composition. To this end, we compared the data profiles of a set of closely related peptide mixtures that differed only in the concentration of a single spiked peptide, angiotensin. As expected, the software revealed both quantitative (Fig. 4*c*) and qualitative (*inset*) differences in sample composition over the effective dynamic range of the MS instrumentation used in this study, allowing for detection of 2-fold changes in peptide abundance in an otherwise highly complex proteomic mixture.

We next tested the sensitivity of the platform to spurious experimental variations stemming from fluctuations in sample work up. Duplicate aliquots of a protein mixture were processed in parallel and analyzed by repeat LC-MS. The resulting profiles were found to be highly similar (Supplemental Fig. S5). We concluded that the platform is relatively robust to artifacts stemming from standard sample handing procedures.

Sample Classification

A clinically important end-goal of expression profiling is sample classification (14). We therefore tested if the software could highlight differences in the proteomic patterns of liver extracts derived from two physiologically distinct groups of inbred mice. The two groups consisted of two fasted mice (C and D) and three control fed mice (A, B, and F). Fasting is

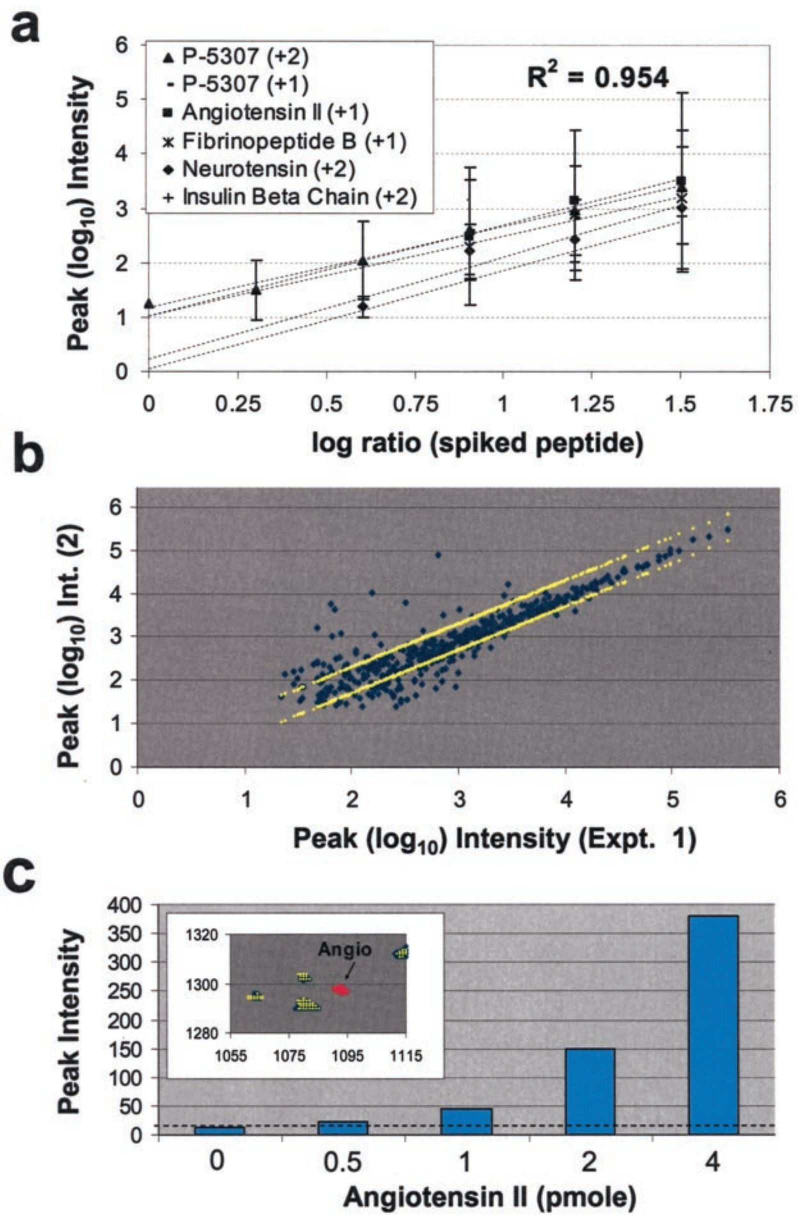

FIG. 4. **Peak quantification.** *a*, linear-fit standard curves of average measured peak intensities in triplicate analyses of five different synthetic peptides across a dilution series. Peptide identity is indicated in the *inset*, with ion charge state shown in *brackets*. *b*, scatter ($\log_{10}$) plot analysis of the intensities of >400 peaks detected in two repeat LC-MS analyses, showing the reproducibility of peak quantification over a range of peak intensities. Two-fold error bars are highlighted in *yellow*. *c*, detection and quantification of 2-fold titrations of a single marker peptide, angiotensin (*m/z* = 1,297), spiked into serum. Average measured peak intensity is shown, with trace background signal indicated by a *dashed line*. *Inset*, Pamphlet overlay reveals the spiked angiotensin peak (*red pixels*) amid peptides common to both the spiked (*yellow*) and control (*blue*) samples.

known to induce substantive biochemical reorganization of liver metabolism in order to sustain circulating blood glucose levels. These adaptations are reflected at the protein level (35). As a general test of reproducibility, each tissue sample was analyzed in duplicate on two separate days, for a total of four datasets per sample. Using our software, the abundance of several hundred peptide peaks was then tracked across each dataset. By combining data-mining algorithms with statistical scoring procedures, we aimed to identify robust proteomic signatures that would enable sample classification. To this end, mouse F was treated as an unknown, with a key task being to determine its physiological status based solely on the proteomic patterns alone. All 20 datasets were converted to pamphlets, aligned, and compared in a comprehensive pairwise manner. The typical data overlap observed for the 10 pairs of same day repeat analyses was ~75% (Supplemental Fig. S6*a*), and only negligibly lower (~72%) for the 20 pairs of different day repeat datasets.

In principle, one could use either quantitative patterns or qualitative (present/absent) differences in peptide abundance to discriminate between the samples. For the purpose of diagnostic development, it may be preferential to focus on the latter (5, 6, 15, 36). A good way to calibrate the software was to see if it could reveal modest differences in protein abundance between individual mice within the two groups, as some biological variation is expected (37). Indeed, pair-wise comparisons of all intra-group datasets revealed greater differences in the proteomic patterns of individual mice (Supplemental Fig. S6*b*) than could be explained by experimental variation alone. With this added confidence, we then tested

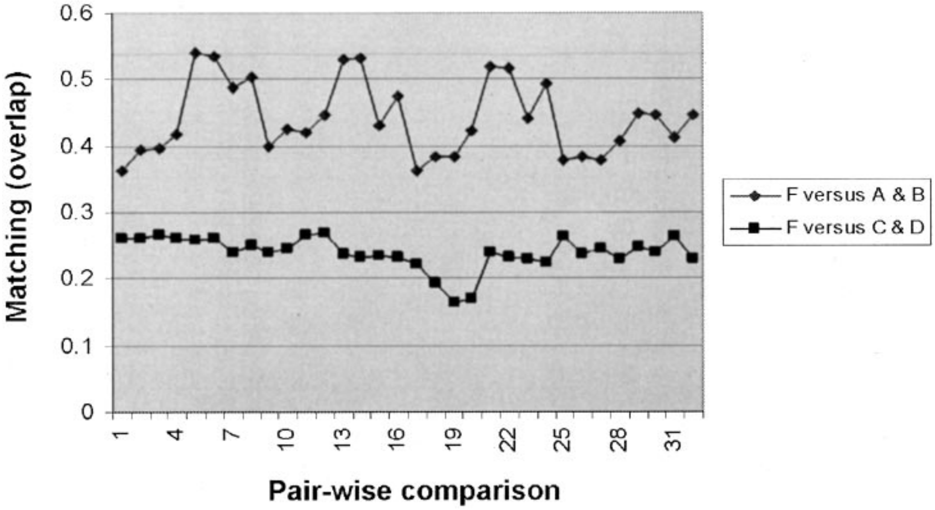

FIG. 5. **Sample classification.** Comprehensive pair-wise comparisons of peak overlap (*Matching*) between the data pamphlets derived for mouse F and each of the two defined groups of mice (fed, mice A and B; fasted, mice C and D).

whether the software generated patterns could be used to correctly classify mouse F. Indeed, the profile separation was pronounced (Fig. 5), allowing for its unambiguous assignment to the control (fed) group. Not only were each of the mouse F data profiles more similar "on average" to the control mice (A and B) than to the fasted mice (C and D), they were more alike in every single comparison.

In these experiments, the separation between the two classes of mice was obvious. In the more realistic setting wherein the separation of profiles is not so clear-cut, we suggest the application of suitable statistical criteria to make the distinction. In this regard, the most powerful statistical test is likely the *t* test, although different methods (*e.g.* ANOVA) could also be applied. The key issue of consideration in this setting is whether the "average" proteomic pattern for any given sample is significantly distinct to allow discrimination between the respective classes. In this case, application of the *t* test indicates a highly significant distinction to be inferred ($p < 10^{-6}$; <0.000001).

We made use of this relatively straight-forward procedure for its simplicity, speed and ability to handle multiway classifications, but alternative classification procedures and algorithms may be more effective in certain data-mining scenarios. Again, these algorithms can be readily incorporated as stand-alone modules within the platform. When the phenotypic difference (in terms of proteomic profiles) between classes is less pronounced, it is reasonable to expect that the two curves plotted in Fig. 5 could sometimes cross. However, our software can still handle this scenario, provided that the intra-group variations are sufficiently small to detect statistically significant differences in comparisons between the two respective classes. The reason for this lies in the fact that the classification algorithm was used not only to process the data shown in Fig. 5, but also to take into account additional possible sources of experimental and biological variability as reported in Supplemental Fig. S6.

We note that the matching score of an unknown test dataset against all other datasets obtained for a particular class (for example, all profiles acquired for the group of fed mice) can likewise be regarded as a "point" in high-dimensional space (typically known as the "feature space" in pattern recognition literature, where the dimension is the number of samples within a group). Generating such a point for each dataset (that is, for both the fed and starved mice) gives rise to two clusters of points in the space, one for each class. If the two clusters are sufficiently separable in space or exhibit relatively confined covariance structure, robust classifiers can be readily obtained using established, rigorous criteria, allowing ready classification of the unknown sample F even when the two curves shown in Fig. 5 cross. Of course, if these clusters are not sufficiently distinct, for instance due to severe intra-class profile variations, virtually all classification schema would be expected to fail. So, in the end, any data-mining approach will be data driven.

Sequence Validation

To validate our peak matching procedures, we evaluated the peptide sequence identities deduced for the peaks matching across different datasets. In a sense, this is an ultimate test of our methodology, because failure of any step (*e.g.* peak detection, alignment, matching, or even the database search itself) would result in nonuniform identifications. Of 647 peaks matched across the mouse samples (Level 2 pamphlets), ~200 were sequence identified ($p < 0.05$) in more than half the samples. Encouragingly, the vast majority (>93%) of the matched peaks were assigned the same sequence identity across the different datasets, confirming the reliability of the software to accurately track identical peaks between samples. The few exceptions consisted of either an occasional mismatch (<2%), possibly due to errors by the database search algorithm, or to dual sequence assignments (~5%), most likely because the peak detection or alignment algorithms had artifactually fused two adjacent but distinct peaks.

Repetition Improves Data Consistency

One might have expected greater similarity between repeat LC-MS pamphlets (ideally 100%), suggesting other deficiencies in our methodology. Conceptually, this lack of reproducibility represents a barrier to biomarker discovery, because it seems unlikely that one could distinguish between closely related samples (*e.g.* >95% peak overlap), when experimental reproducibility is <80%. Clearly the issue is statistical in nature, because it is impossible to repeat an LC-MS experiment with ~100% reproducibility (16, 38).

To gauge the extent to which experimental repetition might compensate for this problem, we developed the concept of a Mother Pamphlet (*J*,*K*,*L*), a data matrix combining key elements (scan number, *m/z*, and signal intensity) that define all the features reproducibly detected at least *K* times in an Level *L* pamphlet generated from *J* repeat datasets. For example, a Mother Pamphlet (4,2,3) contains only those peaks detected in common in any two of four related input Level 3 pamphlets. Based on this new measure, each of the four profiles recorded for mouse F did, in fact, exhibit considerably better peak overlap (~99%) to a Mother Pamphlet (3,1,1) constructed from the remaining three other datasets (Supplemental Fig. S7*a*). That is, virtually all of the peaks detected in one pamphlet were likewise found in at least one of the other datasets. Hence, even limited repetition reveals essentially all of the principle peptide peaks that define a sample.

We expanded on this concept by establishing the commonality of proteomic patterns within a group of related mice. We compared a Mother Pamphlet (4,4,1) encompassing all peaks reproducibly detected in all four mouse F datasets to a more inclusive Mother Pamphlet (8,2,3) created for all eight datasets derived for the other two fed mice. As expected, a high degree of similarity (90.5%) was still detected (Supplemental Fig. S7*b*), with the modest residual variation due, in part, to biological variation between individuals (37).

Biomarker Identification

By logical extension, one should be able to detect peptides that differ reproducibly between sample groups by comparing Mother Pamphlets. To pinpoint reproducible differences in the proteomic patterns of the fed and fasted mice classes, we prepared and compared a Mother Pamphlet (12,5,3) from all 12 fed mouse datasets, considering only those peaks reproducibly detected at least five times, to a more stringent Mother Pamphlet (4,4,1) prepared from each of the two fasted mice datasets (Fig. 6, *a* and *b*). As expected, unique peptides were reproducibly detected in the fasted mouse states (Fig. 6*c*). A subset of these putative biomarkers were identified by CID sequencing and found to map to enzymes belonging to metabolic functions known to be elevated in the fasting state (35). For instance, three peptides mapped to betaine-homocysteine *S*-methyltransferase, an enzyme involved in homocysteine metabolism; two matched to the fatty-acid binding protein L-FABP, which is linked to the transport of lipids in liver; and two mapped to 10-formyltetrahydrofolate dehydrogenase, which mediates *de novo* biosynthesis of purine. Hence, the profiling procedures revealed biologically relevant changes in the abundance of physiological significant biomarkers (albeit relatively abundant enzymes) with absolute specificity and sensitivity.

Case Study: Blood Profiling

The use of blood-borne markers is widespread in clinical practice (3, 39). Although imposing severe limitations in dynamic range due to the overabundance of albumin, blood represents an attractive resource for clinical biomarker discovery as it is readily accessible using relatively noninvasive procedures (3, 9). Indeed, Liotta and colleagues have established the potential diagnostic value of proteomic serum profiling (39). We therefore evaluated the ability of our platform to detect modest differences in the proteomic patterns of duplicate serum samples, one of which had been spiked with troponin (a well-studied marker of myocardial infarction) prior to analysis. As expected, several unique peaks were reproducibly detected in the spiked serum and could readily be assigned to troponin as they were likewise detected in analyses of troponin alone (Supplemental Fig. S8). Although limited in scope, this pilot study attests to the generality of our profiling platform for biomarker screening.

DISCUSSION

Proteomics has the potential to provide unprecedented insight into the molecular changes that accompany physiological transitions, including those preceding clinical presentation (14, 15, 24, 36, 40, 41). Global proteomic surveys offer particularly powerful classification value because each profile is composed of hundreds to thousands of data points (10, 40). While gene expression profiling has been fruitful in this regard (42), protein profiling holds more promise because of the intrinsic advantages of proteins in clinical pharmacology. Protein profiling is also preferable to candidate testing in biomarker discovery because it is not restricted by limited prior knowledge about disease or drug action. Furthermore, it endows researchers with the ability to investigate biochemical adaptations from a systems perspective (43) and can accelerate the validation and annotation of ongoing genome sequencing projects (44).

While gel-free LC-MS-based profiling methods offer remarkable analytical speed and sensitivity (17), its variability has limited its general suitability for biomarker discovery (16). In an attempt to overcome this, several research groups have developed innovative chemical labeling strategies designed to improve the reliability of quantitative inferences that can be made by LC-MS (19). In addition, an "accurate mass and elution time" profiling method based on ultra-high performance LC-MS systems has been reported (22). While effective,

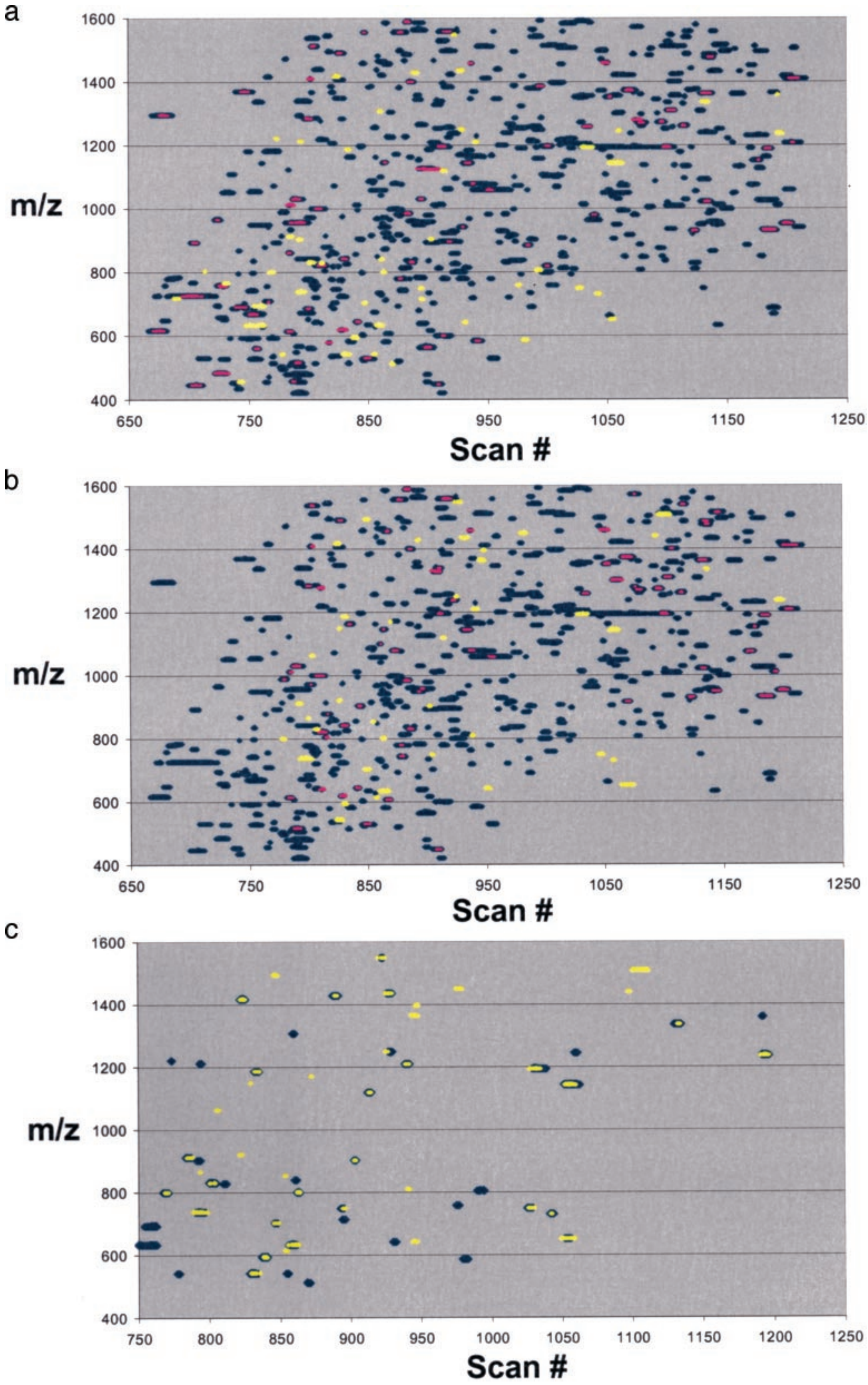

FIG. 6. **Biomarker discovery.** Comparison of mouse liver proteomic patterns. *a*, overlap of a Mother Pamphlet (12,5,3) derived for the datasets of the fed mice (A, B, and F; *blue peaks*) with a Mother Pamphlet (4,4,1) generated for mouse C (*pink*). *Yellow peaks* represent candidate biomarkers detected exclusively (and reproducibly) in the mouse C profiles. *b*, as in *a*, but showing peaks detected in the mouse D datasets. *c*, strong similarity in the biomarker patterns detected for mouse C (*yellow*) and mouse D (*blue*).

the impact of these approaches on the clinical domain has been restricted to date, in part due to the considerable dedicated instrument time, technical expertise, and costs associated with multisample analyses (38).

To overcome these limitations, we have developed and validated a complementary informatics strategy designed to derive reliable qualitative and quantitative protein profiling data using established, broadly applicable LC-MS procedures. The software described here corrects for spurious deviation between experiments, permitting meaningful comparisons of proteomic datasets for the purpose of identifying differential protein expression between samples. It also automates large-scale pattern recognition and mining of proteomic datasets for the purpose of sample classification and biomarker discovery. The Mother Pamphlet strategy, in particular, allows detection of reproducible differences in proteomic patterns, improving proteome coverage and dynamic range.

Using this approach, we have shown that informatics methods can reveal biologically significant changes in tissue protein patterns, allowing for sample classification without being overly sensitive to experimental noise (*e.g.* the particular day an LC-MS experiment was performed). The data strongly suggest that the software can serve as the basis for systematic molecular investigation of disease or therapeutic action. Many experimental variations can be envisaged, including strategies to monitor dynamic changes in the levels of protein post-translational modifications in response to stimuli. Importantly, this informatics platform can also accommodate the use of isotope-based chemical labeling methods (19, 29) to further enhance the accuracy of the quantitative measurements.

The threshold-like data filtering criteria used here can be implemented as a semi-quantitative measure of peptide abundance (for instance, by comparing the presence/absence of peak detection at different pamphlet filter levels). In this study, we performed a series of control experiments showing that integrated peptide peak intensities nicely correlate with abundance (Fig. 4). Thus, in order to compare the levels of peptides between classes, one only needs to revisit the appropriate Mother Pamphlets and compare the intensities corresponding to peptides of interest. Indeed, we set forth with the long-term aim of developing the software as the basis for routinely quantifying differences (relative ratios or fold-changes in protein abundance) in peak intensities across even the most complex proteomic patterns. Implementation of this feature is now a relatively straightforward programming issue because there are no substantive mathematical difficulties, and we hope to address this desirable functionality in the next generation of the software. The utility of various methods of data normalization on the reliability of the inferences made from proteomic comparisons also needs to be more extensively evaluated. Of course, one is still unlikely to be able to detect all possible biomarkers due to limitations in protein and peptide extraction and variations inherent to the experiments themselves—namely, dynamic range confines and extremes in biological complexity and/or variability.

The alignment process can be computationally demanding, particularly for larger sample sizes. In the mouse tissue profiling example provided, instead of a larger full-scale set of 400 alignments (5 mice*4 Pamphlets = 20 datasets, resulting in $20^2 = 400$ matched pairs), the alignment problem can be stratified by first creating 5 Mother Pamphlets for each individual mouse (*i.e.* 4 mice*$4^2 = 64$ alignments) and then aligning the 5 Mother Pamphlets ($5^2 = 25$ matchings), totaling 89 alignments, with a corresponding increase in processing speed. Nevertheless, we have found that comparisons of upward of 100 individual datasets are quite manageable and are taxing only for the highest feature extraction pamphlet levels. Given the modular design of the software and depending on the design of experiments and available hardware, different and more appropriate computational methods to address this possible constraint can also be implemented if needed. Likewise, the brute force *M-N* algorithm could also be sped up, but since this computation scales linearly with the number of experiments and since the application of the algorithm (5 levels per analysis, with each taking ~1 min of CPU time, as compared with the 60–90 min typically required for most LC-MS analyses) is not a limiting factor, there is no pressing need at this point to optimize it further.

It should be noted that a key aspect of the profiling strategy outlined here is the ability to detect and evaluate candidate biomarkers without the need for carrying out time-consuming CID. Proteins of interest, such as those whose levels change reproducibly as a result of an experimental perturbation or which help differentiate between clinical samples, can then be identified in targeted follow-up sequencing experiments. While similar concepts have recently been introduced by others (18, 22, 45), our approach has major advantages in that it builds on established experimental techniques and existing instrumentation that are broadly available throughout the biomedical research community. Nonetheless, our toolkit can exploit the improved dynamic range, resolution, and mass accuracy of newer generation MS instrumentation. Moreover, although high-abundance proteins were preferentially detected in the mouse profiling experiments reported here (in part due to the limited dynamic range of the instrumentation used), we expect that proteome coverage can be significantly improved by using basic subcellular fractionation and affinity enrichment techniques prior to LC-MS (1, 2). By uncoupling sequence identification from peptide quantitation, a markedly expanded number of samples can be analyzed in a single day, increasing the precision and throughput of quantitative proteomic measurements, resulting in a better accounting of biological variation.

Acknowledgments—We thank Mike Davey, Andrei Starostine, and Thomas Kislinger for expert technical support; Alexandr Ignatchenko, Dean Payelev, Slaven Stricevic, Gordana Derado, and Josip Derado for data analysis and programming assistance; and Toni Bedalov, Brian Cox, and Ata Ghavidel for critical feedback.

* This work was supported in part by funding from the National Science and Engineering Research Council of Canada (NSERC), Genome Canada and the Canadian Protein Engineering Network Centre of Excellence (PENCE) to A. E. 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.

S The on-line version of this manuscript (available at http://www.mcponline.org) contains supplemental material.

¶ To whom correspondence should be addressed: Dragan Radulovic, Department of Mathematics, Florida Atlantic University, Boca Raton, FL. E-mail: radulovi@fau.edu. Andrew Emili, CH Best Institute, 112 College Street, Toronto, ON, Canada M5G 1L6. Tel.: 416-946-7281; Fax: 416-978-8528; E-mail: andrew.emili@utoronto.ca.

‡‡ Current address: Shoreline Community College, North Shoreline, WA 98133.

REFERENCES

1. Kislinger, T., and Emili, A. (2003) Going global: Protein expression profiling using shotgun mass spectrometry. *Curr. Opin. Mol. Ther.* **5,** 285–293
2. Tyers, M., and Mann, M. (2003) From genomics to proteomics. *Nature* **422,** 193–197
3. Hanash, S. (2003) Disease proteomics. *Nature* **422,** 226–232
4. Berkow, R., Beers, M. H., and Merck Research Laboratories. (1999) *The Merck Manual of Diagnosis and Therapy*, 17th Ed., Merck Research Laboratories, Whitehouse Station, N.J.
5. Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003) The case for early detection. *Nat. Rev. Cancer* **3,** 243–252
6. Frank, R., and Hargreaves, R. (2003) Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.* **2,** 566–580
7. Sevenet, N., and Cussenot, O. (2003) DNA microarrays in clinical practice: Past, present, and future. *Clin. Exp. Med.* **3,** 1–3
8. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* **2,** 418–427
9. Rosenblatt, K. P., Bryant-Greenwood, P., Killian, J. K., Mehta, A., Geho, D., Espina, V., Petricoin, E. F., and Liotta, L. A. (2004) Serum proteomics in cancer diagnosis and management. *Annu. Rev. Med.* **55,** 97–112
10. Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations. *Mol. Cell. Proteomics* **3,** 367–378
11. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C., and Liotta, L. A. (2002) Clinical proteomics: Translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* **1,** 683–695
12. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
13. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
14. Petricoin, E. F., and Liotta, L. A. (2003) Clinical applications of proteomics. *J. Nutr.* **133,** 2476S–2484S
15. Diamandis, E. P. (2004) Analysis of serum proteomic patterns for early cancer diagnosis: Drawing attention to potential problems. *J. Natl. Cancer Inst.* **96,** 353–356
16. Boguski, M. S., and McIntosh, M. W. (2003) Biomedical informatics for proteomics. *Nature* **422,** 233–237
17. Wu, C. C., and MacCoss, M. J. (2002) Shotgun proteomics: Tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther.* **4,** 242–250
18. Washburn, M. P., Ulaszek, R., Deciu, C., Schieltz, D. M., and Yates, J. R., 3rd. (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* **74,** 1650–1657
19. Aebersold, R. (2003) Quantitative proteome analysis: methods and applications. *J Infect Dis* **187,** Suppl. 2, S315–S320
20. 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
21. Washburn, M. P., Ulaszek, R. R., and Yates, J. R., 3rd. (2003) Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. *Anal. Chem.* **75,** 5054–5061
22. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2,** 513–523
23. Petricoin, E. F., 3rd, Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C. B., Levine, P. J., Linehan, W. M., Emmert-Buck, M. R., Steinberg, S. M., Kohn, E. C., and Liotta, L. A. (2002) Serum proteomic patterns for detection of prostate cancer. *J. Natl. Cancer Inst.* **94,** 1576–1578
24. Petricoin, E. F., and Liotta, L. A. (2003) Mass spectrometry-based diagnostics: The upcoming revolution in disease detection. *Clin. Chem.* **49,** 533–534
25. Han, J., and Micheline Kamber, M. (2000) Data mining: Concepts and techniques. *The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann Publishers, San Francisco, CA
26. Radulovic, D. (2003) On accelerated random search. *SIAM J. Optimization* **14,** 703–731
27. Hamilton, T. G., Klinghoffer, R. A., Corrin, P. D., and Soriano, P. (2003) Evolutionary divergence of platelet-derived growth factor alpha receptor signaling mechanisms. *Mol. Cell. Biol.* **23,** 4013–4025
28. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2,** 96–106
29. Cagney, G., and Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20,** 163–170
30. Eng, J. K., McCormack, A. L., and Yates, J. R. I. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **11,** 976–989
31. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: A framework for understanding peptide dissociation. *J. Mass Spectrom.* **35,** 1399–1406
32. 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
33. Annesley, T. M. (2003) Ion suppression in mass spectrometry. *Clin. Chem.* **49,** 1041–1044
34. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32,** (suppl.) 496–501
35. Edvardsson, U., von Lowenhielm, H. B., Panfilov, O., Nystrom, A. C., Nilsson, F., and Dahllof, B. (2003) Hepatic protein expression of lean mice and obese diabetic mice treated with peroxisome proliferator-activated receptor activators. *Proteomics* **3,** 468–478
36. Diamandis, E. P. (2003) Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin. Chem.* **49,** 1272–1275
37. Parker, L. T., and Leamy, L. (1991) Fluctuating asymmetry of morphometric characters in house mice: The effects of age, sex, and phenotypical extremeness in a randombred population. *J. Hered.* **82,** 145–150
38. Aebersold, R., and Cravatt, B. F. (2002) Proteomics—Advances, applications and the challenges that remain. *Trends Biotechnol.* **20,** S1–2
39. Liotta, L. A., Ferrari, M., and Petricoin, E. (2003) Clinical proteomics: Written in blood. *Nature* **425,** 905
40. Liotta, L. A., Kohn, E. C., and Petricoin, E. F. (2001) Clinical proteomics: Personalized molecular medicine. *J. Am. Med. Assoc.* **286,** 2211–2214
41. Patterson, S. D., and Aebersold, R. H. (2003) Proteomics: The first decade and beyond. *Nat. Genet.* **33,** (suppl.) 311–323
42. Michener, C. M., Ardekani, A. M., Petricoin, E. F., 3rd, Liotta, L. A., and Kohn, E. C. (2002) Genomics and proteomics: Application of novel technology to early detection and prevention of cancer. *Cancer Detect.*

Prev. **26,** 249–255

43. Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., Haynes, P. A., Hays, L., Schieltz, D., Ulaszek, R., Wei, J., Wolters, D., and Yates, J. R., 3rd. (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11969–11974
44. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526
45. Griffin, T. J., Lock, C. M., Li, X. J., Patel, A., Chervetsova, I., Lee, H., Wright, M. E., Ranish, J. A., Chen, S. S., and Aebersold, R. (2003) Abundance ratio-dependent proteomic analysis by mass spectrometry. *Anal. Chem.* **75,** 867–874