The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry

II. EVALUATION OF TANDEM MASS SPECTROMETRY METHODOLOGIES FOR LARGE-SCALE PROTEIN ANALYSIS, AND THE APPLICATION OF STATISTICAL TOOLS FOR DATA ANALYSIS AND INTERPRETATION*

Priska D. von Haller‡, Eugene Yi, Samuel Donohoe, Kelly Vaughn, Andrew Keller, Alexey I. Nesvizhskii, Jimmy Eng, Xiao-jun Li, David R. Goodlett, Ruedi Aebersold, and Julian D. Watts§

Proteomic approaches to biological research that will prove the most useful and productive require robust, sensitive, and reproducible technologies for both the qualitative and quantitative analysis of complex protein mixtures. Here we applied the isotope-coded affinity tag (ICAT) approach to quantitative protein profiling, in this case proteins that copurified with lipid raft plasma membrane domains isolated from control and stimulated Jurkat human T cells. With the ICAT approach, cysteine residues of the two related protein isolates were covalently labeled with isotopically normal and heavy versions of the same reagent, respectively. Following proteolytic cleavage of combined labeled proteins, peptides were fractionated by multidimensional chromatography and subsequently analyzed via automated tandem mass spectrometry. Individual tandem mass spectrometry spectra were searched against a human sequence database, and a variety of recently developed, publicly available software applications were used to sort, filter, analyze, and compare the results of two repetitions of the same experiment. In particular, robust statistical modeling algorithms were used to assign measures of confidence to both peptide sequences and the proteins from which they were likely derived, identified via the database searches. We show that by applying such statistical tools to the identification of T cell lipid raft-associated proteins, we were able to estimate the accuracy of peptide and protein identifications made. These tools also allow for determination of the false positive rate as a function of user-defined data filtering parameters, thus giving the user significant control over and information about the final output of large-scale proteomic experiments. With the ability to assign probabilities to all identifications, the need for manual verification of results is substantially reduced, thus making the rapid evaluation of large proteomic datasets possible. Finally, by repeating the experiment, information relating to the general reproducibility and validity of this approach to large-scale proteomic analyses was also obtained. ***Molecular & Cellular Proteomics 2:428–442, 2003.***

A main objective of proteomics research is the systematic identification and quantification of the proteins expressed in a cell, or contained within a cell compartment or other protein complex. The common approach to quantitative protein analysis to date has been the combination of protein separation, most commonly high-resolution two-dimensional polyacrylamide gel electrophoresis (2DE)[1] and tandem mass spectrometry (MS/MS). For this approach, protein identification is accomplished by individual spot excision, in-gel-digestion, and sequence identification by MS/MS. When desired, relative protein quantification is achieved by visualizing differences in the 2DE patterns from related samples via silver staining or radiolabeling (1–6).

This method has proven quite successful for the cataloguing of large numbers of proteins in complex samples. However, the approach is highly repetitive, labor intensive, and difficult to automate. In addition, it necessarily selects only for proteins that can be resolved by 2DE, missing many larger and smaller proteins, in addition to proteins with lower solubility, such as membrane proteins. Also, due to sample loading limitations for 2DE, it generally selects for only the most abundant proteins in a biological sample (4, 7), thus missing

From the Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103

Received, May 8, 2003, and in revised form, June 24, 2003
Published, MCP Papers in Press, June 25, 2003, DOI 10.1074/mcp.M300041-MCP200

[1] The abbreviations used are: 2DE, two-dimensional polyacrylamide gel electrophoresis; EM, expectation maximization; IADIFF, INTERACT differential; ICAT, isotope-coded affinity tag; LC, liquid chromatography; μLC-MS/MS, microcapillary-liquid chromatography tandem mass spectrometry; MIF, macrophage inhibitory factor; MS, mass spectrometry; MS/MS, tandem mass spectrometry; p_{comp}, computed probability that the given peptide sequence assignment is correct.; P_{comp}, computed probability that the given protein identification is correct.; TCR, T cell receptor.

many lower abundance, regulatory proteins, rarely detected when complex mixtures are analyzed. 2DE also typically resolves different posttranslationally modified forms of the same proteins. Given the high degree and variety of post-translational modifications occurring on the proteins of eukaryotic organisms, this results in great difficulties in obtaining accurate quantitative data on the many proteins that separate into multiple spots, as well as multiple proteins that co-migrate to the same spot, during 2DE. However, because the *in vivo* activities of many proteins are regulated by post-translational modification, the ability to readily resolve differentially modified forms of protein allows for the use of 2DE to monitor changes in the known "active" and "inactive" forms of many proteins.

The recently developed isotope-coded affinity tag (ICAT) technology instead allows for quantitative proteomic analysis based on differential isotopic tagging of related protein mixtures (8–11) and is summarized schematically in Fig. 1. ICAT reagents consist of three functional elements: a thiol-reactive group for the selective labeling of reduced Cys residues, a biotin affinity tag to allow for selective isolation of labeled peptides, and a linker synthesized in either an isotopically normal ("light") or "heavy" form (utilizing ^{2}H or ^{13}C) that allows for the incorporation of the stable isotope tags. In a typical experiment, protein disulfide bridges are reduced under denaturing conditions, and the free sulfhydryl groups of the proteins from the two related samples to be compared are labeled respectively with the isotopically "light" or "heavy" forms of the reagent. The samples are then combined, proteolyzed with trypsin, and the resulting peptides can be separated by any number of optional fractionation steps, including the removal of untagged peptides (*i.e.* not containing a Cys residue) via avidin-affinity chromatography. Peptide/protein identifications are made by MS/MS analyses of the individual fractions, followed by protein sequence database searching of the observed MS/MS spectra. Finally, the observed ratio between the signal intensities for the unfragmented isotopically "light" and "heavy" forms of the same peptide yields the relative abundances of that peptide, and hence the protein from which it was derived, in the original samples.

We have applied the ICAT approach to the investigation of the role of detergent-resistant lipid raft membrane microdomains in T cell receptor (TCR) signaling in the human cell line, Jurkat. We also sought to evaluate the reproducibility, performance, and reliability of the method by comparing the results of two repetitions of the same experiment. In this paper, we present in-depth and systematic technical analyses and discussions of the identifications made within each dataset, as well as comparisons between various datasets. In particular, we show that the application of new, automated, statistical modeling algorithms greatly improved the accuracy of and confidence in both peptide and protein identifications made by assigning probability scores to each peptide and protein matched. While our general approach performed well in analyzing what would normally be challenging protein samples to approaches such as 2DE, the protein identification overlap between the two repetitions of the experiment, along with a number of observations made during the data processing, raised a number of caveats that should be kept in mind when performing and interpreting proteomic data. Furthermore, the use of statistical data analysis removed much of the need for manual verification of both peptide and protein identifications.

These experiments thus illustrated how statistical tools of this nature will greatly facilitate the timely processing of large proteomic datasets, currently a time-consuming and frequently manual process. Also, the application of such tools for assigning measures of confidence to each peptide and protein identified should offer some form of standardization for the interpretation of, in particular, large proteomic datasets. In turn, this should enable researchers to perform any experiment, interpret their results consistently, and then compare the results to those from any other related experiment. Finally, the general application of statistical tools such as these should allow, for the first time, the transparent comparison of related datasets from multiple laboratories.

EXPERIMENTAL PROCEDURES

Purification of Lipid Rafts from Jurkat T Cells—A total of 5×10^8 exponentially growing Jurkat T cells were resuspended at $\sim 2 \times 10^7$/ml in RPMI 1640 medium supplemented to 10% fetal calf serum, split into two equal aliquots, and chilled on ice for 15 min. Cells were simultaneously treated with anti-TCR (OKT3) and anti-CD28 monoclonal antibodies, which were cross-linked with a secondary antibody for 2 min to simulate costimulation, essentially according to standard laboratory protocols (12–14). Detergent resistant membranes (rafts) were purified essentially as described elsewhere (15). Cells were lysed on ice in 25 mM Tris, pH 7.5, 150 mM NaCl, 10 mM β-glycerophosphate, 5 mM EDTA, 1 mM Na_3VO_4, 1 mM phenylmethanesulfonyl fluoride, 10 μg/ml soybean trypsin inhibitor, 2 μg/ml leupeptin, 1 μg/ml aprotinin, 0.1% Triton X-100, dounce homogenized (10 strokes), and mixed with an equal volume of 80% sucrose in MNE buffer (25 mM 2-morpholinoethanesulfonic acid, 150 mM NaCl, 5 mM EDTA, pH 6.5). Rafts were then isolated by sucrose density step gradient ultracentrifugation (16–18 h, 200,000 $\times$ *g*, 4 °C). The low-density raft-containing fraction was further diluted with MNE buffer, and the rafts were pelleted by centrifugation (5 h, 200,000 $\times$ *g*, 4 °C). The lipid raft-containing pellet was then dissolved in 50 mM Tris, pH 8, 5 mM EDTA, 6 M urea, 0.05% SDS.

Protein Labeling and Digestion—ICAT labeling and analysis was performed essentially according to the manufacturer's protocol (ICAT Kit for Protein Labeling; Applied Biosystems, Foster City, CA), with optimized conditions known to result in quantitative labeling (16). In short, following reduction of cysteines and labeling of control (d0-ICAT) and stimulated (d8-ICAT) samples, the samples were pooled and then diluted to $\leq$1 M urea, $\leq$ 0.01% SDS for proteolysis, using an excess of trypsin (Promega, Madison, WI).

Peptide Separation and Purification—The peptides were separated by cation exchange chromatography using a 4.6 $\times$ 200 mm Polysulfoethyl A column (5 μm particles, 300 Å pore size; Poly LC, Columbia, MD) at a flow rate of 800 μl/min. Peptides were eluted by a gradient of 0–25% B over 30 min, followed by 25–100% B over 20 min (buffer A: 5 mM K_2HPO_4, 25% CH_3CN, pH 3.0; buffer B: 5 mM K_2HPO_4, 25% CH_3CN, 600 mM KCl, pH 3.0). The elution profile of the cation exchange chromatography (Fig. 8*A*) determined which fractions were further analyzed. Forty-three (fractions 10–52) cation exchange frac-

tions were individually processed over avidin cartridges (Applied Biosystems) according to the manufacturer's protocol (ICAT Kit for Protein Labeling; Applied Biosystems), to isolate the labeled Cys-containing peptides. Both the avidin column eluate and flow-through fractions were retained. To increase the peptide concentration of Cys-containing peptides for microcapillary-liquid chromatography MS/MS (μLC-MS/MS) analysis, avidin column eluates were pooled in pairs combined (except fraction 52), making a total of 22 fractions for μLC-MS/MS. Because the flow-through fractions contained higher peptide concentrations, these were analyzed individually by μLC-MS/MS. Three sets of samples were generated for subsequent μLC-MS/MS analysis: the avidin-affinity elutes (*i.e.* mostly Cys-containing ICAT-labeled peptides) from the two iterations of the biological experiment and the avidin-affinity flow-through samples (*i.e.* unlabeled peptides) from the first iteration of the biological experiment. The resultant three data subsets generated from the analysis of these samples were termed ICAT 1, ICAT 2, and Flow-through 1, respectively.

μLC-MS/MS Analysis—Fifty to 100% of each sample was loaded using an autosampler and sequentially analyzed by automated data-dependent μLC-MS/MS (17). Injections were made on 10 cm × 100 μm capillary column packed in-house (Magic C_{18}; Michrom BioResources, Auburn, CA). Peptides were eluted with a linear gradient of 10–40% B over 50 min at ~200–300 nl/min (buffer A: 0.4% acetic acid, 0.005% heptafluorobutyric acid in H_2O; buffer B: 100% acetonitrile). A HP1100 solvent delivery system (Hewlett Packard, Palo Alto, CA) was used with precolumn flow splitting. An LCQ-DEKA ion-trap mass spectrometer (ThermoFinnigan, San Jose, CA) with an in-house built micro-spray device was used for all analyses. Peptide fragmentation by collision-induced dissociation was carried out in an automated fashion using the dynamic-exclusion option, and the resultant MS/MS spectra were recorded. The uninterpreted MS/MS data were finally submitted to a suite of software tools for automated database searching and statistical interpretation of the search results. This process, summarized in Fig. 2, is described below, and more extensively under "Results and Discussion."

Database Searching of Observed MS/MS Spectra Using SEQUEST™—Automated database searching using SEQUEST™ software (18) was performed to identify peptide and protein sequence matches for each recorded MS/MS spectrum. Uninterpreted MS/MS spectra were searched against a locally maintained human protein sequence database (version dated 9/8/2002) with typical contaminants such as porcine trypsin (used for proteolysis) and bovine serum albumin (a major component of cell culture medium) additionally included. SEQUEST™ search parameters for ICAT-labeled samples were set as follows: static modification for d0-ICAT-labeled Cys was set to +442.22, with a +8 differential modification for d8-ICAT-labeled Cys; +16 for oxidized Met; mass tolerance ± 3 Da; no proteolytic enzyme specified. SEQUEST™ search parameters for flow-through fractions were the same, but without the modifications for Cys. SEQUEST™ database search software is available from ThermoFinnigan.

Statistical Analysis of Peptide Sequence Matches Using PeptideProphet™—SEQUEST™ output files were automatically submitted to PeptideProphet™ (19) for computation of the probability that each peptide sequence assignment is correct (p_{comp}). The resultant outputs from SEQUEST™ and PeptideProphet™ were displayed using INTERACT (9), a software tool that allows for web/intranet-based data display, and data filtering and sorting via a range of user-definable parameters. INTERACT was used to restrict the datasets by filtering at different p_{comp} cut-offs, and its sorting functions were used to determine the number of "single hit" peptides and proteins (*i.e.* database entries identified via only one peptide with a p_{comp} above the predetermined threshold) that were contained within each filtered version of the data. The in-house software tool, INTERACT differential (IADIFF) was used for side-by-side comparison of identified peptide sequences contained within multiple INTERACT files. This allowed for determination of the overlap between the three datasets for both the peptide sequence matches made and the proteins (*i.e.* database entries) to which they corresponded. INTERACT also generates an Excel spreadsheet version of any filtered and/or sorted dataset for distribution and publication purposes.

Statistical Analysis of Protein Sequence Matches Using ProteinProphet™—The INTERACT data files for all three datasets (ICAT 1, ICAT 2, and Flow-through 1) were submitted to ProteinProphet™. ProteinProphet™ utilizes the list of peptide sequences and their respective p_{comp} scores to determine a minimal list of proteins (database entries) that can explain the observed data and to compute a probability (P_{comp}) that each protein was indeed present in the original sample(s) (20). The ProteinProphet™ output groups together all peptides that (potentially) match a given protein (*i.e.* database entry). It deals with indistinguishable database entries by grouping them as one "protein." This commonly occurs when multiple sequences (mRNAs) and fragments of the same sequence are represented as multiple database entries. Highly homologous gene families are dealt with by formation of related "protein groups," again as single output results. ProteinProphet™ then generates a computed probability (P_{comp}) for each protein or protein group match. These functions are discussed in detail below under "Results and Discussion." The ProteinProphet™ output is also web-based and can be readily exported to an Excel spreadsheet for sorting, distribution, and publication purposes.

More information on PeptideProphet™, ProteinProphet™, and INTERACT can also be found on the Proteomics pages at www.systemsbiology.org/. These applications are available upon request and are open source.

RESULTS AND DISCUSSION

Sample Preparation and μLC-MS/MS Analysis—The general experimental strategy employed for this study is summarized in Fig. 1. Briefly, lipid rafts were isolated from both control and stimulated Jurkat human T cells via standard protocols (15) with a few variations. Cell stimulation was via cross-linking of the TCR with the coreceptor CD28 (12–14). Proteins copurifying with Jurkat T cell lipid rafts were isolated via conventional detergent insolubility (in 0.1% Triton X-100) at 4 °C, followed by sucrose density ultracentrifugation (15, 21). Proteins from control cells were labeled with isotopically normal ("light") ICAT reagent and from stimulated cells with isotopically heavy reagent. The two ICAT reagents differed by 8 mass units and are referred to as the d0- and d8-ICAT reagents, respectively. Samples were combined, proteolyzed with trypsin, and the resultant peptides fractionated by cation exchange chromatography, and individual fractions further processed by avidin-affinity chromatography to enrich for ICAT-labeled peptides. Both the avidin-affinity eluate (ICAT-labeled peptides) and flow-through fractions (unlabeled peptides) were retained for subsequent μLC-MS/MS analyses, as described under "Experimental Procedures." This protocol was repeated a second time to allow assessment of the reproducibility and reliability of the approach.

From the two iterations of the experiment described above, the following fractions were carried forward for μLC-MS/MS analysis: all pooled avidin eluate fractions (*i.e.* Cys-containing, ICAT-labeled peptides) from both experiments, which will be

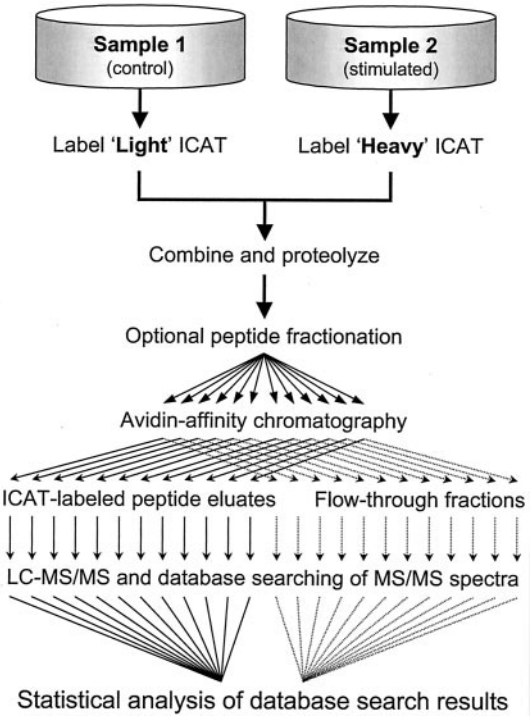

FIG. 1. **Schematic summary of the generic ICAT approach.** The ICAT approach to quantitative proteomics breaks down into three essential steps: ICAT-labeling and proteolytic cleavage of protein samples/mixtures; avidin-affinity enrichment of (labeled) Cys-containing peptides; peptide/protein identification and quantification by MS. This approach allows for additional, optional separation/fractionation of samples at almost any stage of the procedure for the purposes of further enrichment and sample complexity reduction prior to MS. Because heavy and light ICAT-labeled peptide pairs are chemically identical, they will copurify, thus preserving the encoded ICAT ratio for relative protein quantification at the end of the procedure. Subsequent statistical analysis of data generated allows for more accurate and transparent determination of positive peptide and protein identifications.

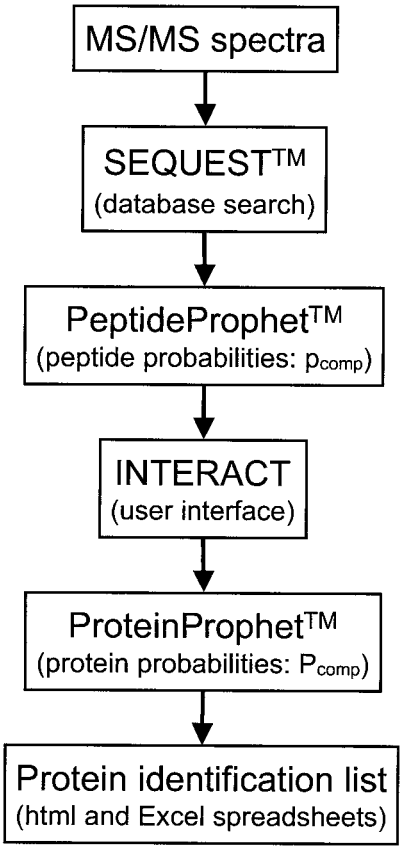

FIG. 2. **Schematic summary of the data flow for automated database searching and statistical data analysis.** Acquired MS/MS spectra are submitted to SEQUEST™ (18) for searching protein sequence database(s). The complied SEQUEST™ search results are then submitted to PeptideProphet™ (19), and the combined SEQUEST™/PeptideProphet™ outputs displayed via the html interface INTERACT (9). INTERACT lists, among other things, all MS/MS scan file locations with their assigned peptide sequences (according to SEQUEST™), and their corresponding SEQUEST™ score and PeptideProphet™ p_{comp} values. INTERACT serves as a user interface that additionally allows for filtering and sorting of the data at this stage of the analysis via a wide range of user-definable parameters. INTERACT also writes an Excel spreadsheet file of the user sorted/filtered data (or entire dataset), as desired, for export. In the final step, ProteinProphet™ (20) takes the INTERACT data file and derives a list of protein identifications and their corresponding P_{comp} scores from the observed peptide data. The ProteinProphet™ output is also in the form of a viewable html spreadsheet, which can similarly be exported to an Excel file.

referred to as the ICAT 1 and ICAT 2 datasets, respectively; the avidin flow-through fractions (*i.e.* non-Cys-containing peptides) from the first (ICAT 1) experiment, which will be referred to as the Flow-through 1 dataset. All recorded MS/MS spectra were searched against a human protein sequence database using SEQUEST™ software (18). Peptide and protein identifications inferred from these search results were determined using PeptideProphet™ (19) and ProteinProphet™ (20) software tools, respectively, summarized in Fig. 2, and further described below and under "Experimental Procedures."

The Need for Statistical Data Analysis for Validation of Peptide and Protein Identifications from Large Datasets—Currently, MS/MS data are searched via a range of database search tools that generate scores relating in some way to the quality of the peptide sequence assigned to each spectrum. To date, determination of the final list of "correct" peptide identifications has typically been based on a "threshold approach," where data is filtered on the basis of these scores alone, with everything below the threshold being discarded. Protein identifications are subsequently determined from the database entries from which the peptide sequences were derived. Typically, visual inspection of spectra is performed by the user to verify spectral quality, and hence the "correctness" of peptide/protein identifications. This is particularly the case when scores are close to the preset threshold, or in cases of "single hits," whereby a protein is identified via only a single peptide sequence identification.

This process is necessarily highly variable. Furthermore,

each user/laboratory has their own opinion of a suitable minimum threshold score to set. This problem is compounded by the fact that the various laboratories use both a range of database search engines, each with their own unique scoring system, and different types of mass spectrometers, each producing MS/MS spectra with their own unique characteristics. In fact, due to a range of variable factors, MS/MS spectral quality can dramatically affect scores obtained for spectra derived from the same peptide. This means that, even if using the same filtering threshold for all experiments, direct comparison of experiments, whether in the same laboratory or another, is most problematic. This difficulty is further compounded by the fact that visual inspection of data is a matter of individual opinion, and thus varies greatly from one individual to the next. Indeed, it is highly unlikely that the same, experienced, user would make precisely the same judgment calls for every spectrum viewed in a large dataset upon a second visual inspection.

Thus there is a clear and recognized need for alternative methods of data analysis to help obviate the time-consuming and vacillatory nature of visual data interpretation. These are needed to help provide the consistency required for comparison of results generated in different experiments, and by different laboratories, using different machines and different database search engines (22). An obvious approach to addressing these issues is the application of statistical methods to the interpretation of proteomic data. Such approaches would replace the threshold method of determining which proteins have been identified by instead assigning confidence levels to potential identifications. The next few sections below describe the application of two new statistical tools, designed for such a purpose, applied to peptide and protein identifications, respectively. Fig. 2 summarizes the data flow for this process. Also discussed below are some of the limitations and pitfalls inherent in the interpretation of any such large proteomic dataset.

Statistical Analysis and Validation of Peptide Identifications—Following SEQUEST™ searching of recorded MS/MS spectra, rather than interpret the data solely on the basis of filtering by database search engine output scores (threshold approach) as in the past (18, 23), SEQUEST™ output files were submitted to a recently developed statistical data modeling algorithm, PeptideProphet™. This algorithm generates its own discriminant score for the peptide sequence assigned to each MS/MS spectrum, based on weighting of a number of parameters for the peptide, including the various SEQUEST™ scores, the mass differential between the observed and calculated mass for the sequence in question, etc. (19). PeptideProphet™ then calculates the population distribution for the discriminant scores for all peptide matches. Next it learns the underlying distributions of "positive" (*i.e.* correct) and "negative" (*i.e.* incorrect) identifications that explain this observed distribution. PeptideProphet™ then employs an expectation maximization (EM) algorithm to perform an iterative process of refining the model to better fit the observed data. A detailed account of how this process works has been published elsewhere (19). PeptideProphet™ performs this modeling process separately for +1, +2, and +3 peptide ion distributions.

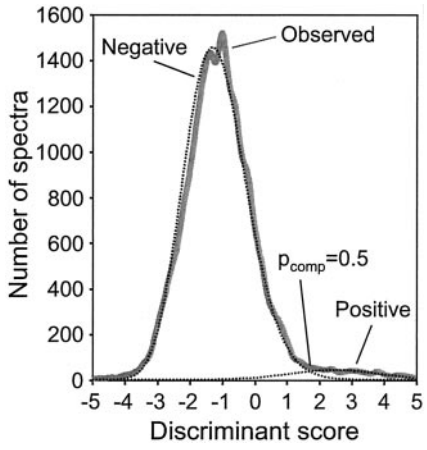

FIG. 3. **Statistical modeling of recorded MS/MS data by PeptideProphet™.** PeptideProphet™ uses an EM algorithm to perform an iterative modeling process on observed data in order to identify the "positive" and "negative" peptide assignments (19). The example given shows the observed distribution (*solid gray line*) of peptide discriminatory scores generated from the SEQUEST™ output files relating to the +2 peptide ion subset of the ICAT 1 dataset (18,109 out of a total of 38,881 MS/MS spectra). The *dotted lines* represent the (final) fitted "positive" and "negative" discriminatory score distributions generated by the iterative modeling of these data. From these two distributions, the p_{comp} reported by PeptideProphet™ for each peptide in the +2 ion data subset represents the calculated probability that the given peptide assignment belongs to the population of positive identifications (on a scale of 0 to 1, where 0 is "incorrect" and 1 is "correct"). A p_{comp} score of 0.5 occurs at the indicated point, where the positive and negative modeled distributions intersect. PeptideProphet™ generates p_{comp} values for the +1 and +3 ion populations, in a similar fashion, separately.

Fig. 3 shows the final modeled positive and negative discriminant score distributions, generated by PeptideProphet™, for the 18,109 SEQUEST™ output files that comprised the +2 peptide ion subset of the ICAT 1 data subset. One thing immediately apparent is that, in this case, the model learned that only a small fraction of the peptide assignments made by SEQUEST™ were, in fact, correct. The final step PeptideProphet™ performs is to use these positive and negative distributions to compute a probability (P_{comp}), for each of the 18,109 peptide assignments for being a member of the positive identification distribution. This computed probability is on a scale of 0 to 1, where 0 is "incorrect" and 1 is "correct," with $p_{comp} = 0.5$ occurring at the point at which the two distributions intersect. These p_{comp} values, along with SEQUEST™ output scores, peptide sequences, database entries assigned, etc., are then exported to a software application called INTERACT. INTERACT is a web-based application that allows the user to view the data, as well as sort and/or filter it according to a range of user-definable parameters, including peptide sequence, SEQUEST™ scores, p_{comp}, database accession, etc. (9).

TABLE I

Summary of potential peptide and protein identifications electronically filtered at varying degrees of confidence

Lipid rafts were isolated from control and stimulated Jurkat T cells and proteins subjected to ICAT labeling and ultimately μLC-MS/MS analysis, as summarized in Fig. 1 and described under "Experimental Procedures." Two iterations of this procedure were performed. MS/MS data were searched using SEQUEST™ software against a human protein sequence database. Three separate datasets were compiled: the avidin-affinity eluate fractions (ICAT-labeled Cys-containing peptides) from the two iterations of the experiment (ICAT 1 and ICAT 2), and the avidin-affinity flow-through fractions of the first ICAT experiment (Flow-through 1). The resultant SEQUEST™ output files generated within each dataset were automatically modeled using PeptideProphet™, then manually sorted and filtered. Different computed probability (p_{comp}) thresholds were set to filter each dataset using INTERACT, which reported the total number of peptide identifications (which includes redundant identifications of the same sequence), the total number of unique peptide sequences identified, the total number of proteins (database entries) these corresponded to, and the number of these which were identified by only a single peptide (single hits). This table lists these numbers for minimum peptide p_{comp} thresholds of 0.95, 0.9, 0.7, and 0.5 for each of the three datasets. Additional filtering with INTERACT was performed to obtain the same numbers for only Cys-containing peptides, given in parentheses (+C) for the ICAT 1 and 2 datasets. Finally, the number of unique peptides identified is also given as a percentage of total peptide matches, and the number of single hit proteins as a percentage of total proteins identified.

		$p_{comp} \geq 0.95$ (+C)	$p_{comp} \geq 0.9$ (+C)	$p_{comp} \geq 0.7$ (+C)	$p_{comp} \geq 0.5$ (+C)
ICAT 1 SEQUEST output files: 38,881	Peptides	1,322 (1226)	1,378 (1.270)	1,486 (1,356)	1,532 (1,387)
	Unique	593 (529)	610 (538)	654 (568)	679 (583)
	% Unique	44.8% (43.1%)	44.3% (42.4%)	44.0%(41.9%)	44.3% (42.0%)
	Proteins	312 (294)	321 (299)	343 (311)	362 (322)
	Single hits	110 (103)	114 (104)	126 (107)	143 (117)
	% Single hits	35.3% (35.0%)	35.5% (34.8%)	36.7% (34.4%)	39.5% (36.3%)
ICAT 2 SEQUEST output files: 14,087	Peptides	869 (852)	912 (893)	961 (936)	983 (955)
	Unique	335 (320)	348 (332)	370 (350)	382 (360)
	% Unique	38.6% (37.6%)	38.2% (37.2%)	38.5% (37.4%)	38.9% (37.7%)
	Proteins	228 (221)	236 (228)	250 (240)	259 (248)
	Single hits	97 (94)	98 (94)	108 (103)	115 (109)
	% Single hits	42.5% (42.5%)	41.5% (41.2%)	43.2% (42.9%)	44.4% (44.0%)
Flow-through 1 SEQUEST output files: 48,831	Peptides	4,113	4,390	4,851	5,152
	Unique	1,532	1,607	1,748	1,843
	% Unique	37.2%	36.3%	36.0%	35.8%
	Proteins	565	592	644	702
	Single hits	230	247	282	334
	% Single hits	40.7%	41.7%	43.8%	47.6%

The ICAT 1, ICAT 2, and Flow-through 1 datasets were thus separately curated and analyzed within INTERACT, using peptide p_{comp} as the basis for restricting the datasets. For example, a p_{comp} of 0.5 means that, according to the statistical model, the sequence match given is 50% likely to be correct, whereas a peptide match with a p_{comp} of 0.95 is 95% likely to be correct. Table I shows how filtering of the three datasets within INTERACT at different minimum p_{comp} values affects the output. Protein matches given are the number of unique database entries that the filtered peptide list represents, as reported by INTERACT, and does not necessarily reflect the actual number of proteins finally identified. This is addressed by using ProteinProphet™ and is discussed separately below. However, the numbers in Table I do illustrate some aspects of filtering large datasets via p_{comp} alone, making several observations apparent.

First, the numbers indicated in Table I for peptides and "proteins" (*i.e.* the database entries the peptides were assigned to) retained after filtering does not decrease dramatically when filtering at higher values for p_{comp}. This illustrates the effectiveness of PeptideProphet™ at discriminating between the positive and negative distributions (only ~13% fewer assignments when filtering at p_{comp} of ≥ 0.95 *versus* ≥ 0.5). Second, most of the assigned database entries ("proteins") eliminated by filtering at a higher p_{comp} are "single hits" (*i.e.* database entries identified via only one peptide assignment). ProteinProphet™ effectively filters out many such single hit protein identifications by penalizing the p_{comp} values for their single peptides when calculating its own probabilities for protein identifications, as will be seen below. Third, while there was a little variation between the three data subsets, the percentage of unique peptide sequences remaining after filtering changed very little when filtering at different values of p_{comp}. These observations combined suggested that a value of $p_{comp} \geq 0.5$ was an acceptable starting point for generating a final list of peptide assignments made in these experiments. This full list of peptide assignments at $p_{comp} \geq 0.5$ for all three datasets combined, derived from 101,799 initial SEQUEST™ output files, containing 7,667 peptides and representing 2,669 unique peptide sequences, is given separately elsewhere (24).

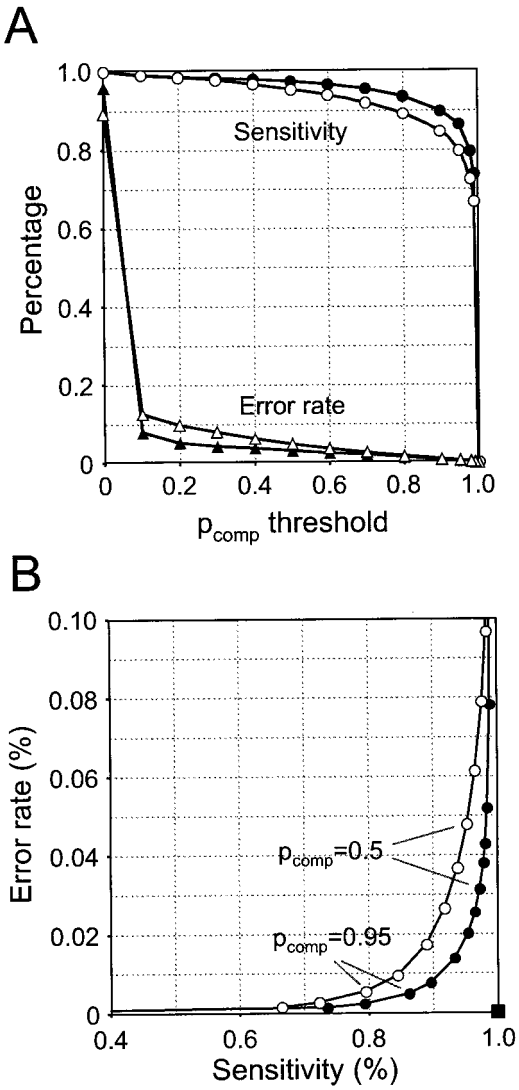

FIG. 4. **Peptide identification error and sensitivity rates for ICAT 1 and Flow-through 1 datasets determined by PeptideProphet™.** *A*, error rates (*triangles*, percentage of total identifications which are false) and sensitivity (*circles*, percentage of total correct identifications remaining after restricting data at a given p_{comp} threshold) as a function of minimum peptide p_{comp} threshold set to restrict the dataset. The avidin-affinity eluate (mostly Cys-containing, ICAT-labeled peptides) and flow-through (unlabeled peptides) fractions are compared, *i.e.* the ICAT 1 (*filled*) and Flow-through 1 (*open*) datasets. *B*, error rate *versus* sensitivity plots for both the ICAT 1 (*filled circles*) and Flow-through 1 (*open circles*) datasets. The relationship between sensitivity and error rate is fixed for each dataset modeled with PeptideProphet™, such that the consequences of selecting a desired p_{comp}, error rate, or sensitivity for data filtering and presentation are known. The "ideal point" (*i.e.* 100% sensitivity and 0% error rate) is indicated with a *filled square*. The curves in *A* and *B* also illustrate how the additional cysteine constraint used to model the ICAT 1 dataset (*filled*) increases the sensitivity and decreases the error rate over those modeled for the Flow-through 1 dataset.

Error versus Sensitivity: Compromising Maximal Return with False Positives—Another significant benefit of using the PeptideProphet™ data modeling algorithm is that the computed probabilities generated for all peptide identifications allow the user to know what is referred to as the "sensitivity" and "error rate" for the entire dataset. Sensitivity is defined as the percentage of the actual correct identifications contained in the restricted (filtered) dataset. Error rate is defined as the percentage of the identifications contained in the restricted (filtered) dataset that are incorrect (*i.e.* false positives). The sensitivity and error rate are directly related and are dependent on the p_{comp} threshold set to filter the dataset. Also, sensitivity and error rates vary for each dataset thus analyzed because, as described above, the p_{comp} values depend upon the calculated positive and negative assignment distributions, which vary from one dataset to another.

Fig. 4 shows plots of the peptide sensitivity and error rates for the ICAT 1 and Flow-through 1 data subsets. These were both derived from the same initial set of samples, separated after avidin-affinity chromatography: ICAT 1 being mostly the ICAT-labeled Cys-containing peptides, and Flow-through 1 being the unlabeled peptides. Fig. 4*A* shows how the error and sensitivity rates are affected as the p_{comp} threshold set to filter the datasets is altered. It is important to reiterate that these curves are not fixed and can vary substantially from one dataset to another.

As mentioned above, the sensitivity and error rates for any dataset modeled are directly related, which can be illustrated by plotting them against each other, as shown in Fig. 4*B*. These plots show that by setting a more stringent p_{comp} threshold, very few false positives are included (low error rate) with the sacrifice being the loss of some of the correct identifications (lower sensitivity). However, as can be seen in Fig. 4*B*, this information allows the user to know the sensitivity and error rates for any p_{comp} filtering threshold used, or if desired, to set the p_{comp} threshold so as to yield a desired sensitivity or error rate. For example, filtering the ICAT 1 dataset at $p_{comp} \geq 0.95$ captured 86.4% of all the correct peptide matches, but with just 0.47% of the filtered list being false positives. On the other hand, filtering at $p_{comp} \geq 0.5$ captured 97.4% of the correct matches, but at a cost of a 3.1% false positive rate.

Fig. 4*B* also illustrates the flexibility and power of a statistical data modeling approach. One can readily see that, compared with Flow-through 1, the ICAT 1 dataset yielded a curve are closer to the ideal point: *i.e.* 100% sensitivity with a 0% error rate (indicated in Fig. 4*B* with a *filled square*). This is not a coincidence. This occurred because ICAT labeling targets cysteine residues. Thus we were able to include the presence of (labeled) cysteine in the peptide sequences assigned by SEQUEST™, for the ICAT 1 (and ICAT 2) data subset, as an additional factor for PeptideProphet™ to model for its calculation of final p_{comp} values. This ultimately lead to better discrimination between "correct" and "incorrect" identifications for ICAT 1 *versus* Flow-through 1, for which the additional constraint did not apply (and was thus not used for p_{comp} calculations for Flow-through 1). Indeed, this inherent flexibility of PeptideProphet™ makes it able to utilize the output results generated by almost any database search pro-

gram, as well as improving its performance for "specialized" applications other than ICAT. For example, if one were searching for phosphorylated peptides, the presence of (phosphorylated) serine, threonine, and/or tyrosine in the matched sequence could be included, as appropriate, as additional contributory factors for p_{comp} calculation for that particular dataset.

False Positives Relating to Database Issues—An important caveat to bear in mind when interpreting proteomic data, even when applying statistical tools such as PeptideProphet™ (and ProteinProphet™) to improve confidence in identifications, arises when studying higher eukaryotic organisms (including humans) where the sequence databases searched are incomplete and/or not fully annotated. Indeed, at the time of writing, only a few genomes have been fully completed, even fewer being eukaryotic. Furthermore, for most genomic sequences, it is not yet clear which sequences represent those that code for protein, nor is it yet clear what, in fact, constitutes one gene. This means that the sequence databases searched for a proteomic investigation of most organisms, particularly for humans, are *de facto* also incomplete. Any search algorithm used to search proteomic data, including SEQUEST™, will only report the "best" match from the searched database, which is what they are designed to do. However, if a peptide/protein in the original sample is not represented in the database searched, or the sequence in the database is incorrect (due to a sequencing error or polymorphism, for example), then the "best" match reported will also be incorrect.

False positives of this nature are hard to identify by their very nature. This is because good MS/MS spectra can randomly yield "acceptable" matches to the wrong sequence. The "correct" match would of course yield a better search result, but is not represented in the database. It is difficult to know how frequently this occurs, though this is likely related to how much of the database is "missing" (also not known). However, if the organism of study has little sequence information available on it, then such events would likely be frequent. While not completely immune from this effect, because it is a data modeling algorithm rather than a database search engine, PeptideProphet™ evaluates multiple parameters to model the false identification population, thus does not rely solely on scores generated by the search engines and visual data inspection. These parameters include the SEQUEST™-generated cross-correlation (Xcorr) score (an indication of the number of peaks of common mass between observed and expected spectra) and preliminary SpRank (a preliminary indication of how well the assigned peptide scored relative to those of similar mass) (18), and for experiments where trypsin was used for proteolysis, the number of tryptic termini for the assigned peptide (19). This enables PeptideProphet™ to identify many false positives of this nature by assigning them a low p_{comp}. Furthermore, PeptideProphet™ is also impartial when it evaluates potential identifications, unlike even an experienced human user, who may make different judgment calls for the same data point on different days and be biased toward potential identifications that fit with the biology of the experiment in question.

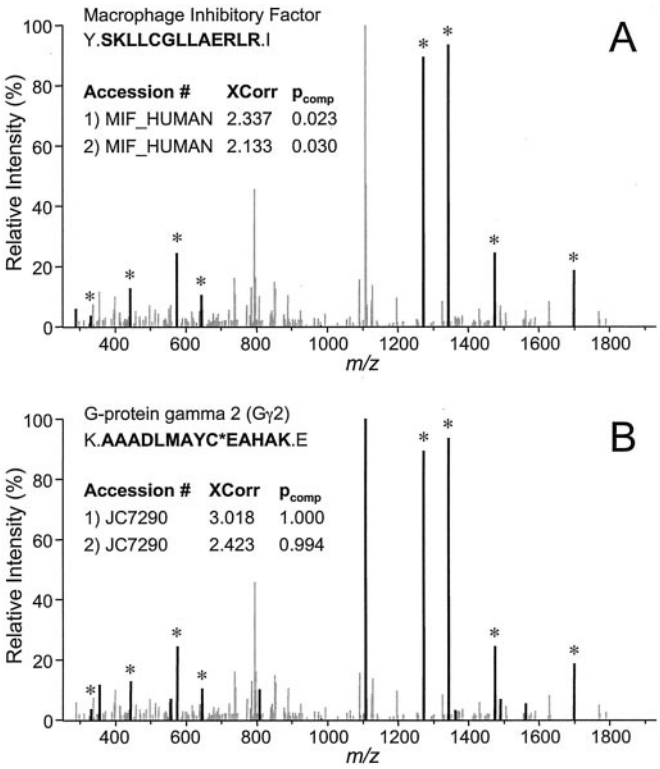

FIG. 5. **Example of discrimination between real and false positive peptide identification using SEQUEST™ and PeptideProphet™.** *A*, observed MS/MS spectrum assigned to the given peptide sequence from the human protein sequence database entry for macrophage inhibitory factor (MIF_HUMAN) in the ICAT 1 dataset. SEQUEST™ cross-correlation scores (XCorr) and PeptideProphet™ p_{comp} for this same database assignment made in both the ICAT 1 (*1*) and ICAT 2 (*2*) datasets are given. *B*, additional searching of a nonredundant protein sequence database matched the same MS/MS spectrum to a different peptide sequence from the bovine heterotrimeric G protein Gγ2 (C* = d8-ICAT-labeled Cys). This sequence is conserved in the human homolog of Gγ2 (JC7290), an entry missing from the human sequence database originally searched. This protein sequence entry was added to the human protein sequence database and the searches rerun, yielding new SEQUEST™ cross-correlation scores and PeptideProphet™ p_{comp} values for this same MS/MS spectrum. The *m/z* peaks highlighted in *bold* in *A* and *B* represent those that were matched by SEQUEST™ to the respective peptide sequences. Those fragment ion peaks that matched both sequences given in *A* and *B* are additionally indicated with an *asterisk*. The scores originally obtained with SEQUEST™ in *A* might be considered a positive identification, according to SEQUEST™ filtering parameters employed in the past (18, 23). PeptideProphet™, however, was able to clearly distinguish between the false positive identification of MIF in *A versus* the likely correct identification of Gγ2 in *B* for the same MS/MS scan.

Fig. 5 illustrates this point well. In two iterations of the same ICAT experiment, the same peptide from the same protein, macrophage inhibitory factor (MIF), was identified with SEQUEST™ scores that passed commonly used filtering param-

eters used to date (18, 23). Fig. 5*A* shows one such MS/MS scan, along with the search results for the given peptide sequence from both the ICAT 1 and ICAT 2 datasets. While the peptide sequence was only partially tryptic, the biology of MIF was in keeping with the biology of the experiment performed and made acceptance of the identification tempting. However, when PeptideProphet™ interpreted the data, it reported low values for p_{comp}, indicating that MIF was not identified. However, as shown in Fig. 5*B*, when the same data were searched against a nonredundant database, the same MS/MS scan (better) matched a peptide sequence for the bovine heterotrimeric Gγ2 protein. The human homologue of this gene, though known, was not in the human database searched for some reason, resulting in the errant MIF identification. Because the human and bovine Gγ2 amino acid sequences are conserved for the region spanning the assigned peptide, when the human Gγ2 sequence was added to the human database and the data re-searched, the p_{comp} values for this new match were now very high (*i.e.* most likely correct). Indeed as discussed elsewhere,[2] heterotrimeric G proteins are highly abundant in lipid rafts (the source of our initial sample), thus this result was also in keeping with the biology.

Remarkably, as can be seen in Fig. 5, many of the MS/MS fragment ion peaks matched potential fragment ions from both peptide sequences (peaks in *bold*), eight of which were common to both sequences (indicated with *asterisks*). Thus, using SEQUEST™ alone, even an experienced user could be forgiven for assuming MIF to be the correct identification, whereas PeptideProphet™ yielded an unequivocal result. Indeed, immunoblotting confirmed that MIF was not present in the original samples (data not shown). This being said, even if statistical tools such as PeptideProphet™ are used, it is still likely that some incorrect identifications will result from searching incomplete databases.

Statistical Analysis and Validation of Protein Identifications—When interpreting proteomic MS/MS data, there are two related but entirely separate steps to the process of identifying the protein(s) in the original sample. The first, as has been discussed above, is assigning individual MS/MS spectra to peptide sequences in a database. For this purpose, PeptideProphet™ was developed to calculate a level of confidence (computed probability) for each sequence so assigned. The second step is the determination of the proteins that these peptides, collectively, represent. This process is very different from the process of assigning peptide sequence to MS/MS spectra, but is also by no means simple, especially when dealing with complex higher eukaryotic organisms such as human. ProteinProphet™ was thus developed to assist in the deconvolution of the complexities inherent in protein identification, again by calculating a probability (P_{comp}) for each protein potentially identified (20)

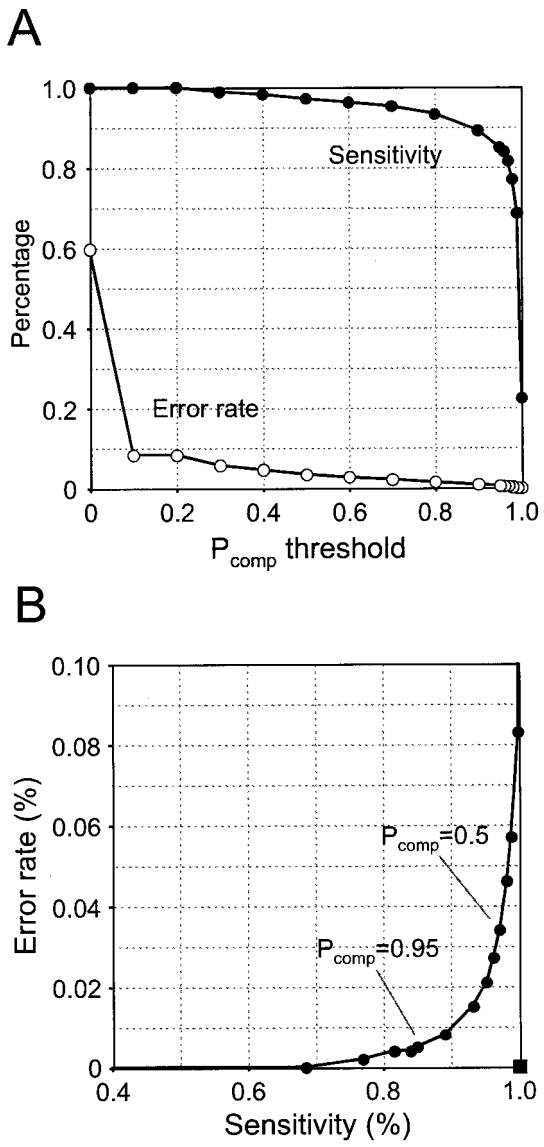

FIG. 6. **Protein identification error and sensitivity rates determined by ProteinProphet™.** *A*, Error rates (*open*, percentage of total identifications which are false) and sensitivity (*filled*, percentage of total correct identifications remaining after restricting data at a given P_{comp} threshold) as a function of minimum protein P_{comp} threshold set to restrict the data (a compilation of the ICAT 1, ICAT 2, and Flow-through 1 datasets). *B*, Error rate *versus* sensitivity plot for the same data. The relationship between sensitivity and error rate is fixed for each dataset modeled with ProteinProphet™, such that the consequences of selecting a desired P_{comp}, error rate, or sensitivity for data filtering and presentation are known. The "ideal point" (*i.e.* 100% sensitivity and 0% error rate) is indicated with a *filled square*.

ProteinProphet™ also uses an EM algorithm to derive the simplest list of proteins (*i.e.* database entries) that can explain the observed peptide data. ProteinProphet™ uses features of the observed peptides assigned to each database entry in question pertinent to the likelihood that this protein was ac-

[2] P. D. von Haller, E. Yi, S. Donohoe, K. Vaughn, A. Keller, A. I. Nesvizhskii, J. Eng, X. Li, B. Wollscheid, D. R. Goodlett, R. Aebersold, and J. D. Watts, manuscript in preparation.

tually present in the original sample(s), including number of sibling peptides (different peptide sequences matching the same database entry) and the p_{comp} values for each of these peptides, etc. In this way, ProteinProphet™ assigns each potential "protein" identification its own P_{comp} value: *i.e.* the probability that the given protein identification is correct, on a scale of 0 (incorrect) to 1 (correct). A detailed description of how ProteinProphet™ models the peptide data and calculates P_{comp} values can be found elsewhere (20).

In order to generate a "final list" of the proteins (database entries) most likely copurifying with the T cell lipid rafts, the three datasets (ICAT 1, ICAT 2, and Flow-through 1) were combined and submitted to ProteinProphet™ (in this case, ProteinProphet™ included all peptides with a $p_{comp} \geq 0.05$ for its calculations in order to speed up the process, without sacrificing data of any significance). As with the final peptide list, the resultant protein identification dataset was filtered at $P_{comp} \geq 0.5$. This final protein list, along with the P_{comp} values for each protein identification match and its matching peptide sequences with their respective p_{comp} values, among other things, are given separately elsewhere (24). Similarly to the peptide identification results, the ProteinProphet™ output allows for the determination of error rate and sensitivity plots for these data (see Fig. 6). Again, ProteinProphet™ models each set of peptide data separately, thus the results are data-dependent and will be different for the same protein(s) identified in separate experiments. In this case, as shown in Fig. 6, when we restricted our protein data to list only the most confident identifications ($P_{comp} \geq 0.95$), we got a false positive (error) rate of 0.5%, but at the price of retrieving only 85.0% of the actual correct identifications. However, when we filtered the data at $P_{comp} \geq 0.5$, we instead retrieved 97.2% of the actual correct matches, but at a price of a 3.4% false positive rate. Again, the power of statistical data analysis with tools such as ProteinProphet™ is the control it gives the user in making informed and transparent decisions when interpreting data, allowing them to accurately know the likelihood that any specific potential protein identified was present in the original sample(s).

Dealing with Protein Redundancy at Both the Database and Biological Level—One of the drawbacks of studying higher eukaryotic organisms is the increased occurrence of somewhat functionally redundant families of proteins that are highly conserved at the primary sequence level. Frequently this results in MS/MS spectra that are assigned to peptide sequences that are absolutely conserved between multiple species and/or gene family members. In such cases, while we are able to assign the most likely peptide sequence, we are unable to ascribe it with certainty to any single database entry. This problem is (unnecessarily) compounded by the chaos currently existing in many sequence databases, often with multiple entries (cDNA, RNA, partial coding sequences, etc.) for what is undoubtedly the same protein. These issues are particularly prevalent in human sequence databases. The best solution to this problem is to fix the database(s), condensing the redundancies into single entries, and making accession numbers and annotations more systematic. This would lessen the confusion when interpreting protein identification data. However, dealing with highly related and conserved protein families and structural domains will remain a challenge to the correct identification of proteins belonging to such groups when studying higher eukaryotes.

ProteinProphet™ deals with these problems, when necessary, by grouping proteins (database entries) in one of two ways, examples of which can be found within the full list of T cell lipid raft proteins identified (24). On occasion a peptide, or set of peptides, can be assigned to a single database entry. Other times, two or more database entries are essentially identical. This commonly occurs when one or more mRNA/cDNA/partial coding sequences for the same protein (or fragment thereof) have separate entries in the database being searched. Typically, when interpreting just SEQUEST™ output results, one gets multiple protein "identifications" from such cases, and the onus is on the user to rationalize the results. This process is very time consuming and difficult for large datasets, typically resulting in a higher number of proteins "identified" than are actually present in the dataset. In cases where a peptide, or set of peptides, match two or more essentially identical database entries, ProteinProphet™ groups these entries together to form one "protein," collectively making a single entry in its protein identification output. In effect, the software reports that this "protein" has been identified at a given P_{comp}, but that it cannot distinguish between the two or more database entries listed for it, on the basis of the available peptide data.

On other occasions, a high degree of protein sequence homology makes it difficult to distinguish between conserved gene family members. ProteinProphet™ similarly deals with these scenarios via the formation of "protein groups," which again form single entries in its output file. ProteinProphet™ again assigns a P_{comp} for the entire group, *i.e.* the probability that one *or more* of the family members were present in the original sample(s). The protein family members from whom ProteinProphet™ has assigned one or more peptides are then listed under the group heading, along with the peptide(s) matched to each entry, in the same way that it is done for other proteins. Finally, ProteinProphet™ assigns P_{comp} values to indicate which protein group members were most likely present in the original sample(s), based on the preponderance of the evidence (typically, though not necessarily, those with the highest number of unique peptide sequences).

One thing to note about such "protein groups" is that none of the assigned peptides will be exclusive to any one database entry. Proteins for which unique identifying peptides have been assigned by PeptideProphet™ (at high enough p_{comp}), will automatically be assigned their own individual output lines with corresponding P_{comp} by ProteinProphet™. Thus it is fair to say that it is not possible to say with absolute

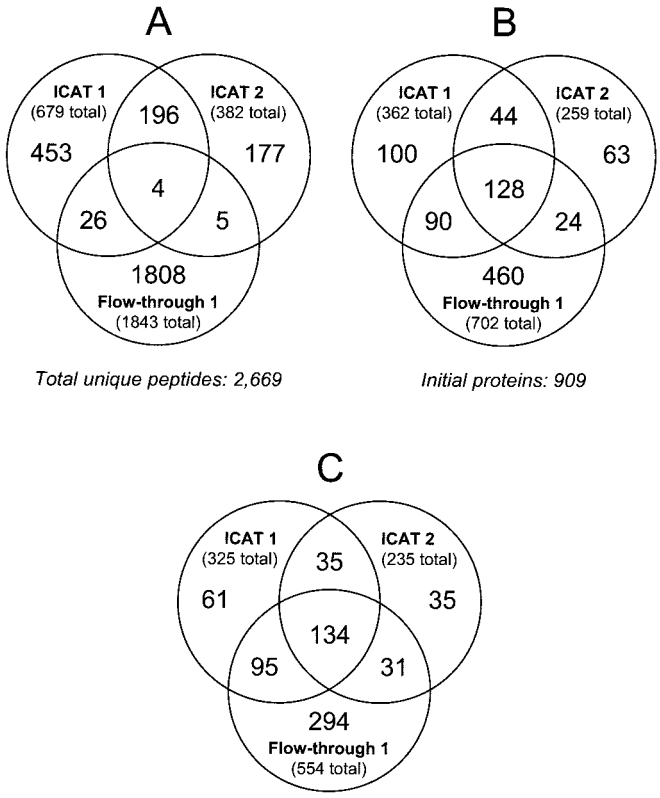

FIG. 7. **Overlap of peptide and protein identifications determined by PeptideProphet™ and ProteinProphet™ for all three datasets.** A total of 2,669 unique peptide sequences were identified with a PeptideProphet™ $p_{comp} \geq 0.5$ in the combined ICAT 1, ICAT 2, and Flow-through 1 datasets. Using INTERACT to sort the data into protein matches (*i.e.* database entries), the 2,669 peptides corresponded to a total of 909 separate database entries. ProteinProphet™ was then used to refine and condense these into 685 proteins or protein groups with a $P_{comp} \geq 0.5$. *A*, Overlap of unique peptide sequences identified separately in the three datasets, determined using IADIFF. *B*, Overlap of proteins (database entries) identified via peptide identifications from *A*, determined using IADIFF. *C*, Overlap of final protein/protein group identifications determined by ProteinProphet™, again via peptide identifications from *A*.

certainty that any of the specific protein group family members were indeed present in the original sample(s), only that one or more were at the given P_{comp} for the group. Also, while many group members are assigned a P_{comp} of zero, they similarly cannot be ruled out with any certainty. Finally, on occasions when the gene family is large and/or has a very high degree of sequence homology, ProteinProphet™ will assign a P_{comp} of 1 to the group, but zero to all group members. This is an indication that while this class of protein was clearly present in the original sample(s), the peptides observed were shared by too many separate database entries to calculate which were most likely present. Four examples of this occurred when studying T cell lipid rafts: tubulin α and β chains, stomatin, and spectrin α chain.[2]

Overlap of Peptide and Protein Identifications from Related Datasets—One of the observations made when performing large-scale proteomic μLC-MS/MS-based investigations on complex protein mixtures is that reanalysis of the same sample under essentially identical conditions leads to the identification of a somewhat different set of proteins than was observed in previous analyses (17, 25–27). The overlap between consecutive μLC-MS/MS runs of the same sample typically depends on the sample complexity. This can range from close to 100% overlap for a very small set of abundant proteins to as low as 20% overlap for complex mixtures spanning a wide range of abundances. When dealing with highly complex protein mixtures, multidimensional chromatography is required to simplify the peptide mixture for separate μLC-MS/MS analyses to allow for increased peptide/protein identification, as was performed in this study. Even when performing such additional upstream prefractionation, the overlap between the dataset obtained for the whole experiment and that obtained from further repetitions of the same protocol can likewise vary tremendously, again depending on the complexity of the original sample and the prefractionation protocol employed.

This variability in the overlap in the set of peptides/proteins identified when complex samples are repeatedly analyzed has led, indirectly, to an unfortunate and unforeseen trend in proteomics, whereby the proteins in two samples, related through a biological experiment, are separately determined. The absence (or presence) of a given protein in one sample *versus* the other is then, incorrectly, interpreted as being a consequence of the biological experiment. While it is not entirely clear why this often poor overlap occurs, it is most likely due in large part to the mass spectrometer's rate of sampling from the large set of overlapping peptide peaks eluting from the various chromatography columns employed. From this we can infer that when the overlap between multiple experiments is not 100% (100% overlap is very rare) then not all of the proteins in the original sample(s) were identified. Given this, it is thus not appropriate to draw the conclusion that a given protein was not in a given sample, simply on the grounds that it was not identified, even if the same protein was identified in a separate analysis of a highly related, or even identical, sample. It was to address this problem that stable isotope-tagging approaches, such as ICAT, were devised. With such an approach, the related samples are labeled with different isotopic versions of the same chemical and then combined. The original samples are then analyzed as one. Once a peptide is identified, reconstructing the ion chromatograms for the different isotopic versions of the same peptide determines the relative abundance of the peptide (hence protein) in the original samples. Thus if only one isotopic version is observed, it now is valid to assume that the peptide/protein was not present in the other original sample(s), or that its level was reduced sufficiently so as to be indistinguishable from the observed level of signal noise for the given experiment.

In order to look at this overlap effect more closely, and the effect, if any, that subsequent statistical data analyses had on it, we performed the ICAT experiment twice in its entirely. From these, we generated two sets of peptide/protein identification data from the avidin-affinity eluates (*i.e.* Cys-containing peptides) to compare the ICAT 1 and ICAT 2 datasets. We chose to focus our attention on the ICAT-labeled peptides because we were also interested in the reproducibility of the observed ICAT ratios for proteins identified in common between the two experiments. This second, equally important, aspect of the comparison is discussed in detail elsewhere.[2] Finally, we analyzed the avidin-affinity flow-through fractions for one of the iterations of the experiment. This was done to assess the benefit, in terms of increased protein identifications and our confidence in them, *versus* the cost through additional machine and data processing time (*i.e.* the overlap between the ICAT 1 and Flow-through 1 datasets).

Fig. 7 shows the overlaps for both peptide and protein identifications made for all three datasets. These peptides represent the 2,669 unique peptide sequences derived from the list of total peptide identifications at $p_{comp} \geq 0.5$, listed separately elsewhere (24). As would be expected, the overlap at the peptide level between the avidin-affinity eluate samples (ICAT-labeled Cys-containing peptides) and flow-through samples (unlabeled peptides) was very low (Fig. 7*A*). Indeed, only 29 of 5,152 assigned peptides (23 of the 1,843 unique peptides) in the flow-through fractions contained (unmodified) Cys (24), and only 17 peptides contained ICAT-labeled Cys upon re-searching of the data for ICAT modifications (data not shown). Also, all but one of the overlapping peptide sequences between the ICAT and flow-through samples were non-Cys-containing peptides, coming from the ~9.5% (at $p_{comp} \geq 0.5$) of unlabeled peptides nonspecifically binding and eluting from the avidin cartridges (see Table I). These observations confirmed that both the ICAT-labeling process and the avidin-affinity step to enrich for ICAT-labeled peptides worked most efficiently.

Fig. 7*B* shows the initial overlap at the protein (database entry) level, represented by the peptides identified in Fig. 7*A*. The 909 database entries were simply those assigned by SEQUEST™ sequence database searching, reported using INTERACT (with no human data curation) prior to the implementation of ProteinProphet™. We would thus expect this number to be higher than the final number of actual identifications, because it does not take into account multiple database entries for essentially the same protein. However, even with this caveat, we observed that 60.2% of the database assignments from the ICAT 1 dataset were confirmed in the Flow-through 1 dataset. Interestingly, even though it was a separate iteration of the same experiment, a similar number (58.7%) of the identifications in the ICAT 2 dataset were also confirmed in the Flow-through 1 dataset. When comparing the ICAT 1 and 2 datasets, we observed that 47.5% of the ICAT 1 identifications were confirmed by repeating the experiment (ICAT 2), in keeping with the observation that complex samples yield different results when such analyses are repeated. We also observed that a much higher number (66.4%) of the ICAT 2 identifications were confirmed in ICAT 1. We believe that this effect (and the lower number of total identifications in ICAT 2 *versus* ICAT 1) was due to a smaller amount of starting protein material in the ICAT 2 experiment, likely due to some losses incurred during sample preparation.

Another reason why we expected that the number of identifications shown in Fig. 7*B* would be an overestimate of the actual number of unique proteins present was that because the peptide data was initially filtered at $p_{comp} \geq 0.5$, many lower-confidence "single hit" proteins (those to which only a single peptide was assigned) would likely be included in the final list. We would also expect many such "single hits" to be eliminated upon further processing using ProteinProphet™ and subsequent data filtering. Prior to the availability of tools such as ProteinProphet™, data reduction and simplification of such a list of identifications has been a manual process, typically involving numerous BLAST searches, and the "weeding out" of poor quality hits, based on raw data inspection, one MS/MS scan at a time. For large datasets, this process is necessarily very slow. Also, because the manual process involves frequent and nonreproducible judgment calls by the user, assessing the confidence in each final curated list is almost impossible, making the comparison of results between different individuals and laboratories difficult at best. ProteinProphet™ was developed, in part, to try and address these problems.

Fig. 7*C* shows the overlap at the protein level determined solely by ProteinProphet™, with the only human input being setting the cut-off for protein inclusion in the list, again at $P_{comp} \geq 0.5$. Comparing Fig. 7, *B* and *C*, several things become apparent. We observed a reduction in total protein identifications, particularly in the Flow-through 1 dataset. As can also be seen from Table I, much of this was due, as expected, to the loss of "single hit" proteins, because they are frequently incorrect. ProteinProphet™ "penalizes" single hit identifications in a data-dependent fashion, based upon the learned number of sibling peptides distribution generated by the EM algorithm (20). In order to obtain a protein $P_{comp} \geq 0.5$, a "single hit" peptide score must typically be high, in these datasets requiring a peptide p_{comp} of ~0.95 or higher. We also observed that the overlap between ICAT 1 and 2 was a little higher, but close to that in Fig. 7*B*: 52.0% of ICAT 1 confirmed in ICAT 2 and 71.9% of ICAT 2 confirmed in ICAT 1. However, the overlaps between the ICAT datasets and Flow-through 1 dataset were increased; now more than 70% of ICAT-identified proteins were confirmed by additionally analyzing the avidin flow-through fractions. We believe this increase may be due in part to the loss of single hits, but also because ProteinProphet™ condenses identical and related database entry matches into single "proteins," an effect that would also contribute to the reduction in total identifications

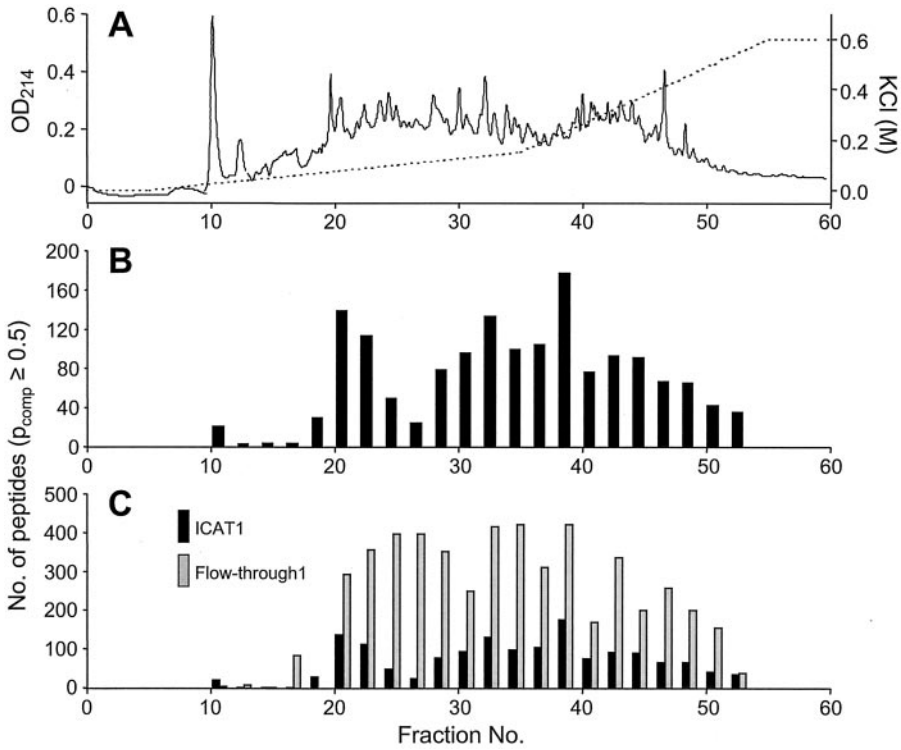

FIG. 8. **Distribution of peptide assignments with respect to cation exchange elution time for avidin-affinity eluate and flow-through fractions.** Following ICAT-labeling of protein samples, their combination and proteolysis, peptide mixtures were separated via cation exchange chromatography. *A*, ultraviolet absorbance trace recorded at 214 nm for the initial experiment, with the salt gradient (KCl) used for elution overlain (*dotted line*). Fractions 10–52 were retained, individually subjected to avidin-affinity chromatography, retaining both eluate and flow-through fractions from each. Eluate fractions (ICAT-labeled peptides) were pooled in pairs and, along with individual flow-through fractions (not pooled), subjected to peptide/protein identification and quantification via automated μLC-MS/MS. *B*, distribution of peptide assignments (at $p_{comp} \geq 0.5$) that could be made (according to criteria given in the text) from avidin-affinity eluate fractions (*i.e.* mostly Cys-containing, ICAT-labeled peptides), with respect to cation exchange fraction number. *C*, distribution of peptide assignments that could be made from avidin-affinity flow-through fractions (*i.e.* unlabeled peptides), with respect to cation exchange fraction number, superimposed on the data shown in *B*. Because flow-through fractions were analyzed separately, identifications from both fractions pooled in *B* were summed to generate the appropriate numbers for comparison in *C*. Note: For various reasons, flow-through fractions 18, 19, 30, and 40 were not analyzed, thus numbers are reduced in *C* where appropriate.

made. This potential benefit of additionally analyzing avidin-affinity flow-through fractions should thus should be considered when performing a quantitative ICAT-type experiment, balanced against the increased machine time and data interpretation time required, when one wishes to be as sure as possible of the identity of the proteins regulated in a biological experiment.

Sample Complexity Reduction Via Multidimensional Chromatography—As the data presented above demonstrate, proteomic analysis of complex biological samples still present significant challenges. While statistical analysis of database search results clearly holds great promise for improving the speed, accuracy, and transparency of proteomic data interpretation, the reproducibility (overlap) of related experiments and the maximization of protein identifications for such experiments are more difficult to address. One thing that does seem clear from our work and that of others is that the simplification (*i.e.* fractionation) of peptide samples is a requirement for any attempt at optimal data return for complex protein mixtures (9, 28). This is also critical if one is to have any chance of identifying the lower abundance proteins in any such mixture (which often turn out to be the more interesting proteins biologically).

In the experiments presented here, we used a three-step peptide separation protocol, previously applied with some success to the identification (and quantification) of membrane proteins (9). The first step was an ion exchange fractionation, which separates peptides roughly according to charge state. The second step was an avidin-affinity column, which enriches for the ICAT-labeled (*i.e.* Cys-containing) peptides. This should simplify the peptide mixture, which in turn should help increase the number of protein identification possible from a complex starting material. The final step was reversed-phase liquid chromatography, which is performed with online MS and MS/MS for both peptide identification and quantification. While it is almost impossible to compare different separation strategies in an attempt to determine an "ideal" approach for maximal data return, we were able to draw some conclusions

about the effectiveness of the separation steps used here from our data. As mentioned above (see also Table I and Fig. 7*A*), the recovery and identification of essentially only Cys-containing peptides in the ICAT 1 and 2 datasets (avidin-affinity eluate) and non-Cys-containing peptides in the Flow-through 1 dataset confirmed the effectiveness of the avidin-affinity step in an ICAT experiment. The effectiveness of μLC for peptide separation prior to MS and MS/MS is also well established and widely documented. Because ion exchange fractions (both avidin-affinity eluates and flow-through fractions) were separately analyzed by μLC-MS/MS, we were also able to assess from our data the effectiveness of the ion exchange peptide fractionation step for a large-scale quantitative proteomic experiment.

We were interested to see whether there was a relationship between peptides that produced MS/MS data of sufficient quality for peptide sequence identifications (in this case yielding $p_{comp} \geq 0.5$) and where they eluted from the ion exchange column. We did this by sorting the peptide identification datasets by cation exchange fraction number. Fig. 8 shows the ion exchange ultraviolet trace (Fig. 8*A*) aligned with the number of peptides that were subsequently identified via μLC-MS/MS from that portion of the profile in the first ICAT experiment, where both the avidin-affinity eluate (ICAT 1) and flow-through (Flow-through 1) fractions were analyzed. These data showed that subsequently useful peptide assignments were obtained throughout the ion exchange gradient for ICAT-labeled peptides (Fig. 8*B*), peptides also capable of yielding quantitative information. The fact that the labeled peptides did elute throughout the gradient used also suggested that the ICAT modification itself did not adversely affect the chromatographic properties of the peptides. When the flow-through peptide identifications were superimposed (Fig. 8*C*), we observed a similar distribution of hits (even though a few flow-through fractions were not successfully analyzed for various reasons). Thus we could conclude that the use of ion exchange as a preliminary fractionation step for large-scale proteomic experiments is an effective strategy for simplification of both ICAT-labeled and unlabeled peptide mixtures.

CONCLUSIONS

With the advent of large-scale mass spectrometry-based proteomics and quantitative proteomics has come the problem of how to interpret, present, disseminate (publish), and compare the large datasets generated. A major hurdle to overcome has been the disparate ways in which the raw MS data are interpreted and the lists of proteins identified in any one experiment decided upon. To achieve this, a range of different protein sequence database search programs have been used to interpret data generated on different types of mass spectrometer. Determining what has in fact been identified in each experiment has subsequently relied upon simple threshold filtering approaches, based upon scores generated by the different search engines, often followed by manual verification of many of the less clear protein identifications. Apart from being very time consuming, the range of "acceptable" filtering parameters used by different laboratories, the incompatibility of the different search engine scoring systems, and the vacillatory nature of user-based manual verification has essentially made the comparison of results from different experiments and between different laboratories difficult at best. There has thus been a clear need for some system of standardization, which will allow for consistency and transparency in data interpretation, and facilitate comparison of one dataset to any other, regardless of how the data is generated (20, 22).

A logical solution to this problem is the use of statistical data analysis. By using statistical algorithms to interpret the results of protein sequence database searches, it should be possible to assign confidence (or probability) to each individual peptide and protein identification. One of the benefits of probability-based statistical analysis is that it also allows the user to know the likely error (false positive) rate of any large dataset restricted on the basis of calculated probability. This is, of course, far more realistic than the current method of reporting results simply as a list of proteins "successfully" identified, at the exclusion of all else. Thus the adoption of suitable statistical approaches to the interpretation of MS-based proteomic data should, for the first time, allow the investigator to compare results from completely separate experiments. Furthermore, if common statistical approaches are applied to the datasets in question, they should allow for the comparison of any one dataset to those generated in other laboratories, even using different machines and search algorithms. Additionally, datasets already published could be reprocessed using the latest versions of these new tools in order to facilitate such comparisons.

Fortunately, this urgent need has been recognized by a number of groups working in proteomics, and several early attempts providing statistical tools for the interpretation of (in particular MS/MS) proteomic data have recently emerged and are beginning to be used. One such attempt has been to generate statistical significances for each peptide assignment in an experiment, based upon the database search engine output scores generated (29). Other recent attempts have used training datasets to determine an algorithm that calculates distributions of "correct" and "incorrect" peptide assignments for any given dataset (of search engine output results), based on the training dataset (30, 31). While such an approach can allow for the calculation of probabilities from these distributions (30), they lack the ability to take data quality into account by relying exclusively on the training data, rather than "learning" the distributions from the observed data, as PeptideProphet™ and ProteinProphet™ are capable of (19, 20). Nevertheless, all of these attempts at applying statistical methods to the interpretation of proteomic MS/MS data hold considerable promise and represent steps in the right direction.

It is thus hoped that the application of new, statistically validated, methodological approaches such as these will soon alleviate much of the confusion and complexity currently in the MS-based proteomics field. This, in turn, will allow for a common platform for the presentation and dissemination (*i.e.* publication) of such proteomic data, allowing for the extraction of more and clearer information by the research community as a whole, and thus accelerate the already significant inroads MS-based proteomics is making into the study and understanding of human biology and disease.

* This work was supported in part by grants from the National Institutes of Health (RO1-AI-41109-01 and RO1-AI-51344-01 to R. A. and J. W., respectively), the National Heart, Lung, and Blood Institute Proteomics Center at the Institute for Systems Biology (N01-HV-28179), and a fellowship awarded by the Swiss National Science Foundation to P.D.H. We thank Oxford GlycoSciences (UK) for additional generous financial support. 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.

‡ Current address: MacroGenics, 1441 North 34th Street, Seattle, WA 98103.

§ To whom correspondence should be addressed. Tel.: 206-732-1283; Fax: 206-732-1299; E-mail: jwatts@systemsbiology.org.

REFERENCES

1. Boucherie, H., Sagliocco, F., Joubert, R., Maillet, I., Labarre, J., and Perrot, M. (1996) Two-dimensional gel protein database of Saccharomyces cerevisiae. *Electrophoresis* **17,** 1683–1699
2. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730
3. Link, A. J., Hays, L. G., Carmack, E. B., and Yates, J. R. 3rd (1997) Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143. *Electrophoresis* **18,** 1314–1334
4. Garrels, J. I., McLaughlin, C. S., Warner, J. R., Futcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., and Payne, W. E. (1997) Proteome studies of Saccharomyces cerevisiae: Identification and characterization of abundant proteins. *Electrophoresis* **18,** 1347–1360
5. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 14440–14445
6. Bardel, J., Louwagie, M., Jaquinod, M., Jourdain, A., Luche, S., Rabilloud, T., Macherel, D., Garin, J., and Bourguignon, J. (2002) A survey of the plant mitochondrial proteome in relation to development. *Proteomics* **2,** 880–98
7. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9390–9395
8. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
9. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19,** 946–951
10. Smolka, M. B., Zhou, H., and Aebersold, R. (2002) Quantitative protein profiling using two-dimensional gel electrophoresis, isotope-coded affinity tag labeling, and mass spectrometry. *Mol. Cell. Proteomics* **1,** 19–29
11. Flory, M. R., Griffin, T. J., Martin, D., and Aebersold, R. (2002) Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol.* **20,** S23–S29
12. Heller, M., Goodlett, D. R., Watts, J. D., and Aebersold, R. (2000) A comprehensive characterization of the T-cell antigen receptor complex composition by microcapillary liquid chromatography-tandem mass spectrometry. *Electrophoresis* **21,** 2180–2195
13. Heller, M., Watts, J. D., and Aebersold, R. (2001) CD28 stimulation regulates its association with N-ethylmaleimide-sensitive fusion protein and other proteins involved in vesicle sorting. *Proteomics* **1,** 70–78
14. Watts, J. D., Sanghera, J. S., Pelech, S. L., and Aebersold, R. (1993) Phosphorylation of serine 59 of p56lck in activated T cells. *J. Biol. Chem.* **268,** 23275–23282
15. Zhang, W., Trible, R. P., and Samelson L. E. (1998) LAT palmitoylation: its essential role in membrane microdomain targeting and tyrosine phosphorylation during T cell activation. *Immunity* **9,** 239–246
16. Smolka, M. B., Zhou, H., Purkayastha, S., and Aebersold, R. (2001) Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal. Biochem.* **297,** 25–31
17. Yi, E. C., Marelli, M., Lee, H., Purvine, S. O., Aebersold, R., Aitchison, J. D., and Goodlett, D. R. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23,** 3205–3216
18. Eng, J., McCormack, A. L., and Yates J. R. 3rd (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. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
20. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, in press.
21. Brown, D. A., and Rose, J. K. (1992) Sorting of GPI-anchored proteins to glycolipid-enriched membrane subdomains during transport to the apical cell surface. *Cell* **68,** 533–544
22. Hancock, W. S., Wu, S. L., Stanley, R. R., and Gombocz, E. A. (2002) Publishing large proteome datasets: Scientific policy meets emerging technologies. *Trends Biotechnol.* **20,** S39–S44
23. Haynes, P. A., Fripp, N., and Aebersold, R. (1998) Identification of gel-separated proteins by liquid chromatography-electrospray tandem mass spectrometry: comparison of methods and their limitations. *Electrophoresis* **19,** 939–945
24. von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via ICAT and tandem mass spectrometry: I. Statistically annotated datasets for peptide sequences and proteins identified via the application of ICAT and tandem mass spectrometry to proteins co-purifying with T cell lipid rafts. *Mol. Cell. Proteomics* **2,** 426–427
25. Spahr, C. S., Davis, M. T., McGinley, M. D., Robinson, J. H., Bures, E. J., Beierle, J., Mort, J., Courchesne, P. L., Chen, K., Wahl, R. C., Yu, W., Luethy, R., and Patterson, S. D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. *Proteomics* **1,** 93–107
26. Koc, E. C., Burkhart, W., Blackburn, K., Moyer, M. B., Schlatzer, D. M., Moseley, A., and Spremulli, L. L. (2001) The large subunit of the mammalian mitochondrial ribosome. Analysis of the complement of ribosomal proteins present. *J. Biol. Chem.* **276,** 43958–43969
27. Fejes, A., Yi, E. C., Goodlett, D. R., and Beatty, J. T. (2003) Shotgun proteomic analysis of chromatophores from the purple phototrophic bacterium *Rhodopseudomonas palustris. Photosynth. Res.*, in press.
28. Washburn, M. P., Wolters, D., and Yates, J. R. 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247
29. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768–774
30. MacCoss, M. J., Wu, C. C., and Yates, J. R. 3rd (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74,** 5593–5599
31. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2,** 137–146