iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates*[S]

David Shteynberg‡, Eric W. Deutsch‡, Henry Lam§, Jimmy K. Eng¶, Zhi Sun‡, Natalie Tasman‡, Luis Mendoza‡, Robert L. Moritz‡, Ruedi Aebersold‖‡‡, and Alexey I. Nesvizhskii§§¶¶‖‖**

The combination of tandem mass spectrometry and sequence database searching is the method of choice for the identification of peptides and the mapping of proteomes. Over the last several years, the volume of data generated in proteomic studies has increased dramatically, which challenges the computational approaches previously developed for these data. Furthermore, a multitude of search engines have been developed that identify different, overlapping subsets of the sample peptides from a particular set of tandem mass spectrometry spectra. We present iProphet, the new addition to the widely used open-source suite of proteomic data analysis tools Trans-Proteomics Pipeline. Applied in tandem with PeptideProphet, it provides more accurate representation of the multilevel nature of shotgun proteomic data. iProphet combines the evidence from multiple identifications of the same peptide sequences across different spectra, experiments, precursor ion charge states, and modified states. It also allows accurate and effective integration of the results from multiple database search engines applied to the same data. The use of iProphet in the Trans-Proteomics Pipeline increases the number of correctly identified peptides at a constant false discovery rate as compared with both PeptideProphet and another state-of-the-art tool Percolator. As the main outcome, iProphet permits the calculation of accurate posterior probabilities and false discovery rate estimates at the level of sequence identical peptide identifications, which in turn leads to more accurate probability estimates at the protein level. Fully integrated with the Trans-Proteomics Pipeline , it supports all commonly used MS instruments, search engines, and computer platforms. The performance of iProphet is demonstrated on two publicly available data sets: data from a human whole cell lysate proteome profiling experiment representative of typical proteomic data sets, and from a set of *Streptococcus pyogenes* experiments more representative of organism-specific composite data sets. *Molecular & Cellular Proteomics 10: 10.1074/mcp.M111.007690, 1–15, 2011.*

A combination of protein digestion, liquid chromatography and tandem mass spectrometry (LC-MS/MS)[1], often referred to as shotgun proteomics, has become a robust and powerful proteomics technology. Protein samples are digested into peptides, typically using trypsin. The resulting peptides are then separated and subjected to mass spectrometric (MS) analysis, whereby a subset of the available precursor ions are sampled by the MS instrument, isolated and further fragmented in the gas phase to generate fragment ion spectra (MS/MS spectra). From these spectra, the peptides and then the proteins present in the sample and, in conjunction with quantification strategies, their relative or absolute quantities can be determined (1).

The volume of data generated in proteomic experiments has been growing steadily over the past decade. This has been aided by the rapid progress made in several facets of proteomics technology, including improved sample preparation and labeling techniques and faster, more sensitive mass spectrometers (2). The resulting explosion in the number and

From the ‡Institute for Systems Biology, Seattle, WA; §Department of Chemical and Biomolecular Engineering, the Hong Kong University of Science and Technology, Hong Kong; ¶Department of Genome Sciences, University of Washington, Seattle, WA; ‖Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland; **Faculty of Sciences, University of Zurich, Zurich, Switzerland; ‡‡Center for Systems Physiology and Metabolic Diseases, Zurich Switzerland; §§Department of Pathology, University of Michigan, Ann Arbor, MI; ¶¶Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

Received January 6, 2011, and in revised form, August 3, 2011
Published, MCP Papers in Press, August 29, 2011, DOI 10.1074/mcp.M111.007690

[1] The abbreviations used are: LC-MS/MS, liquid chromatography-tandem MS; PSM, peptide to spectrum matches; FDR, false discovery rate; TPP, Trans-Proteomic Pipeline; FFE, free-flow electrophoresis; OGE, off-gel electrophoresis; NSS, number of sibling searches; NRS, number of replicate spectra; NSE, number of sibling experiments; NSI, number of sibling ions; NSM, number of sibling modifications; EM, expectation maximization.

size of data sets has necessitated computational tools that can analyze data from diverse types of experiments and instruments in a robust, consistent, and automated manner (3). In particular, the need to distinguish between true and false peptide to spectrum matches (PSMs) produced by automated database search engines became an essential task for the meaningful comparison of proteomic data sets. In early work this was accomplished by applying rigid score cutoffs, but this soon proved problematic because of unknown false discovery rates (FDR) in the filtered data sets. Without a uniformly applied confidence measure it was difficult and unreliable to combine and compare data sets (4, 5).

In recent years there has been substantial progress in developing bioinformatics and statistical tools in support of shotgun proteomic data. This includes the development of new and improved tandem MS (MS/MS) database search algorithms, as well as statistical data methods for estimating FDR and posterior peptide and protein probabilities (reviewed in (2, 6)). There is also ongoing work on improving other aspects of proteomic data analysis, including tools for isotope label-based and label-free quantification, data management systems, and data exchange mechanisms, as reviewed in (5, 7–9). One of our group's contributions to these efforts was the development of the computational tools PeptideProphet (10) (analysis of MS/MS database search results) and ProteinProphet (11) (protein-level analysis), which allowed faster and more transparent analysis of proteomic data. These tools constitute the core elements of the widely used Trans-Proteomic Pipeline (TPP) (12).

At the same time, the last few years witnessed a dramatic increase in the speed of data acquisition. As a result, many data sets are now collected in multiple replicates or involve the generation of otherwise highly overlapping data sets. This is the case in label-free quantitative measurements across multiple samples,(13). MS-based reconstruction of protein interaction networks,(14–16) or the comprehensive characterization of proteomes of model organisms such as *Drosophila* (17), *C. elegans* (18), and *S. cerevisiae* (19) via extensive fractionation of the proteome sample and the mass spectrometric measurement of each fraction. Although PeptideProphet and ProteinProphet have been shown to provide accurate estimates in the case of small to intermediate data sets, several simplifying assumptions in these tools limit their performance with increasing data set size (20). Furthermore, there is a growing interest in the analysis of MS/MS data using a combination of multiple search engines, with the intent to maximize the number and confidence of peptide and protein identifications. This approach has become computationally feasible with the availability of faster computers, the prevalence of computing clusters, and recent emergence of cloud computing. However, combining the results of multiple searches presents additional technical challenges, including the heterogeneity of search engine scores, the propagation of errors, and informatics challenge related to nonuniform data formats.

Here we present a computational method and software tool, iProphet, designed to address these challenges. Although the existing PeptideProphet/ProteinProphet workflow considers only two levels of information, PSMs and protein identifications, iProphet expands this modeling framework to more accurately reflect the nature of shotgun proteomic data by introducing additional levels such as peptide precursor ions (defined as precursor identical PSMs) and unique peptide sequences (defined as sequence identical PSMs). Peptide assignments to MS/MS spectra from different database search tools are naturally integrated within the same framework. As the main outcome, iProphet permits the calculation of more accurate posterior probabilities and FDR estimates at the level of unique peptide sequences, where one probability is used for all sequence identical PSMs observed, which is important for obtaining more accurate FDR estimates at the protein level. The semiparametric modeling approach of PeptideProphet allows fitting the data without prior distributional assumptions, making the entire workflow (PeptideProphet/iProphet/ProteinProphet) capable of handling a wide variety of data from many different search engines. We demonstrate the performance of iProphet on two publicly available data sets: data from a human whole cell lysate proteome profiling experiments representative of typical proteomic data sets, and from a set of *Streptococcus pyogenes* experiments more representative of organism-specific composite data sets.

MATERIALS AND METHODS

Experimental Data Sets—Two publicly available data sets were used in this work. The first was taken from a study on Human Jurkat A3 T leukemic cells(21) (referred to as human data set in the main text). The raw MS data (RAW files) were obtained from the Tranche (22) data exchange system (hash key provided in supplemental information for the original manuscript). Briefly, the cells were lysed, the lysate was separated using one dimensional SDS-PAGE, sections of gel were digested with trypsin, and analyzed using a linear ion trap mass spectrometer (LTQ, Thermo-Fisher). In this work, we use a single complete replicate of the whole cell lysate experiment (replicate 1; 19 gel bands). It comprises a set of 19 MS files containing 161,425 MS/MS spectra in total.

The second data set is a set of five samples from a *Streptococcus pyogenes* study (23) (referred to as *S. pyogenes* data set in the text). The data files are available in the PeptideAtlas raw data repositories (http://www.peptideatlas.org/repository/) as accessions PAe000283 - PAe000287. This data set was divided into five experiments comprising a total of 64 LC-MS/MS runs and 212,880 MS/MS spectra. One of the experiments was from peptide samples fractionated by Free-Flow Electrophoresis (FFE) (Weber Inc., now BD Diagnostics) and analyzed on a hybrid LTQ-FT-ICR (Thermo-Fischer) instrument. One of the experiments was from FFE fractionated samples analyzed on an LTQ instrument. One of the experiments was from peptide samples fractionated by off-gel electrophoresis (OGE) (GE Healthcare) and analyzed on an LTQ instrument. The final two experiments were from samples separated by a strong cation exchange (SCX) and collected on an LTQ instrument.

MS/MS Database Search and PeptideProphet Analysis—All raw MS data were converted to the mzXML file format (24) searched with six search engines (see below), then processed with PeptideProphet,

iProphet, and ProteinProphet, in that order. All mzXML files from the human data set were processed together. In the *S. pyogenes* data set, each of the five MS data subsets described above were processed separately by PeptideProphet. The individual PeptideProphet results were combined in ProteinProphet with or without using iProphet as an intermediate step. The comparison with Percolator was performed using the Human data set and the FFE-LTQ-FT subset of the *S. pyogenes* data set.

The protein sequence database used to search MS/MS spectra from the Human data set comprised of the human RefSeq protein sequences downloaded on April 19, 2010. The protein sequence database used to search the second data set included the *S. pyogenes* protein sequences extracted from RefSeq (version NC_002737) database and human IPI version 3.54 (the human protein sequences were appended to account for a small degree of human protein contamination observed in some of the *S. pyogenes* samples). Decoy sequences were included in both databases by randomizing the tryptic peptides of the target sequences. In each database, the decoys were divided into two distinct sets, with one of the decoy sets being used by PeptideProphet (required for semiparametric modeling, see below), and by extension, the iProphet model. The other decoy set was used to independently validate the performance of the computational tools. The same procedure was applied when running and evaluating the performance of Percolator. The two sets of decoys were created independently from each other by randomizing tryptic peptide sequences. In each decoy set, it was enforced that any identical peptides that originated from different forward proteins resulted in identical shuffled peptides in the decoy protein sequences. The order of the sequences in the final database files was then randomized to remove any bias or tendency of search engines to report equivalent hits based on their order in the database.

Six different search engines were used in this work: Mascot (25), SEQUEST (26), X! Tandem (27) (with k-score plug-in(28)), Inspect (29), MyriMatch (30), and OMSSA (31). A similar set of parameters was applied for all search engines, which can be summarized as follows; a precursor mass tolerance of ± 3 Daltons, fixed mass modification for iodoacetamide derivatives of cysteines (in the *S. pyogenes* data set), variable mass modification for methionine oxidation, and allowing partially tryptic peptides (except in Mascot, which performed significantly slower on our single Mascot license computer when allowing partially tryptic peptides). For an extended description of search engine specific parameters see supplemental Table S1.

For all search engines, PeptideProphet was run on the data in the semiparametric mode (32) specifying one set of the decoy sequences to be used in modeling to fix the shape of the negative distribution. The PeptideProphet accurate mass model (33) was used for the LTQ-FT measurements and the regular mass model for the LTQ data. The number of tolerable termini and mass models of PeptideProphet were disabled for the Inspect searches. Inspect uses this information in computing the score which resulted in a strong bias for peptides having correct mass and number of tolerable termini. Therefore, using these models in PeptideProphet was not statistically sound. The PeptideProphet pI model(34) was applied to all OGE and FFE experiments and the retention time model was applied to experiments showing high quality chromatographic separation.

iProphet Models—The iProphet program implements five models, in an iterative fashion, to refine an initial PeptideProphet analysis. The probability adjustments are based on the number of sibling searches (NSS), replicate spectra (NRS), sibling experiments (NSE), sibling ions (NSI, *i.e.* differently charged peptide precursor ions), and sibling modifications (NSM).

Number of Sibling Searches, NSS—The NSS model rewards or penalizes identifications based on the output of multiple search engines (in pepXML format) for the same set of spectra. No assumptions are made regarding the orthogonality of the search engines being combined. For each spectrum, there are one or more probabilities from the search engines used, as calculated by the individual PeptideProphet analysis on each search result. The probabilities are combined by summing the probabilities of PSMs that agree on the peptide sequence and dividing this value by the number of other searches performed on the spectrum. Thus, the range of possible values for NSS is [0, 1]. The NSS statistic is calculated as follows:

$$NSS_d = \frac{\sum_{\{d' | d \neq d' \wedge Pep_d = Pep_{d'}\}} P(Pep_{d'})}{\sum_{\{d' | d \neq d'\}} 1}$$

Number of Replicate Spectra, NRS—The NRS statistic models the intuition that in a typical data set, multiple high probability identifications of the same precursor ion should increase the confidence of that precursor ion being correctly identified. On the other hand, repeated observation of PSMs having low to intermediate probabilities and corresponding to the same peptide ion suggests that all those PSMs are false. As computed, NRS yields a positive value for precursor ions that are commonly identified with probabilities above 0.5; NRS yields a negative value for precursor ions that are commonly identified with probabilities below 0.5; NRS is 0 for precursor ions that are identified from one spectrum only. Thus, this method of computing NRS attempts to preserve the probabilities of precursor ions identified only once but with a high probability. Its influence is learned from each data set itself, and thus will vary between data sets. The range of possible values for NRS is enforced to [–15, 15]. The NRS statistic is computed according to the following formula:

$$NRS_i = \sum_{\{i' | i \neq i' \wedge Pep_i = Pep_{i'}\}} (P(Pep_{i'}) - 0.5)$$

Number of Sibling Experiments, NSE—NSE is a statistic that is used to model multiple identifications of the same precursor ion across different experiments under the assumption that precursor ions that are observed in multiple experiments and matched to the same peptide sequence are more likely to be correct. As computed, NSE yields a positive value for precursor ions that are commonly identified with probabilities above 0.5 across different experiments; NSE yields a negative value for precursor ions that are commonly identified with probabilities below 0.5 across different experiments; NSE is 0 for precursor ions that are identified from one experiment only. The range of possible values for NSE is enforced to [–15, 15]. The NSE statistic is computed by the following formula:

$$NSE_x = \sum_{\{x' | x \neq x' \wedge Pep_x = Pep_{x'}\}} (P(Pep_{x'}) - 0.5)$$

Number of Sibling Ions, NSI—The NSI model rewards peptides that are identified by precursors with different charges. The NSI statistic is calculated as follows:

$$NSI_z = \sum_{\{z' | z \neq z'\}} P(Pep_{z'})$$

Number of Sibling Modifications, NSM—The NSM model rewards peptides that are identified with different mass modifications. The NSM statistic is calculated as follows:

$$NSM_m = \sum_{\{m' | m \neq m'\}} P(Pep_{m'})$$

Estimation of FDR—As described above, each sequence database contained two sets of decoys. The first set of decoys was used at the

modeling stage (in PeptideProphet and also in Percolator) and thus was not available for independent decoy-based FDR estimation. Thus, unless specified, any reference to decoys in the main text in the context of FDR estimation and accuracy assessment refers to the second set of decoys for each data set.

In each experiment, the lowest probability PSMs (probability less than 0.002) were used to estimate the fraction r of decoy hits among all matches expected to be false and therefore randomly distributed through the database. This ratio (~38%) was then applied to compute the decoy-based FDR estimate at a probability threshold τ:

$$FDR_\tau = \frac{D_\tau / r}{N_\tau}$$

where N_τ and D_τ are the total number of matches and the number of decoy matches, respectively, passing a minimum probability threshold τ, and D_τ/r is the decoy-estimated number of false matches. The total number of correct matches above threshold τ is estimated as

$$NC_\tau = N_\tau - D_\tau / r$$

The same approach is used to compute decoy-based estimates at the unique peptide sequence level and at the protein level. The same analysis was applied based on Percolator's results, except that q-values computed by that algorithm were used for sorting the peptide and protein lists instead of the posterior probabilities.

The model-estimated (*i.e.* using computed posterior probabilities and not decoy counts) FDR for a minimum probability threshold τ is computed as (10)

$$FDR_\tau^{\mathrm{mod}} = \frac{\sum_{\{i, P(Pep_i) \geq \tau\}} (1 - P(Pep_i))}{N_\tau}$$

The model-estimated number of correct matches above a minimum probability threshold τ is computed by the formula

$$NC_\tau^{\mathrm{mod}} = \sum_{\{i, P(Pep_i) \geq \tau\}} P(Pep_i)$$

These equations are modified accordingly to compute similar metrics at the level of unique peptide sequences and proteins.

RESULTS

PeptideProphet and ProteinProphet—Before introducing the extended modeling framework provided by iProphet, it is informative to briefly summarize the conventional PeptideProphet and ProteinProphet approach to the analysis of shotgun proteomic data. The existing strategy considers information at two distinct levels: PSM (PeptideProphet) and protein identification (ProteinProphet). PeptideProphet takes as input all PSMs from the entire experiment (considering the top scoring PSM for each experimental MS/MS spectrum only). It then applies the expectation-maximization (EM) algorithm to derive a mixture model of correct and incorrect PSMs from the data using various sources of information (denoted here as D). These include the primary information such as the database search scores, but also auxiliary information based on various properties of the assigned peptides (*e.g.* number of missed cleavages, mass accuracy, etc.). PeptideProphet then computes the posterior probability for each PSM, denoted as *P(Pep)*, using Bayes' Law:

$$P(Pep) = P(+|D) = \frac{P(D|+)P(+)}{P(D|+)P(+) + P(D|-)P(-)} \quad \text{(Eq. 1)}$$

where $P(D|+)$ and $P(D|-)$ are the probabilities of observing a PSM having information D among correct and incorrect PSMs, respectively, and $P(+)$ and $P(-)$ are prior probabilities of a correct and incorrect PSM in the data set (*i.e.* the overall proportions of correct and incorrect PSMs). The prior probabilities and the parameters governing the $P(D|+)$ and $P(D|-)$ distribution are learned from the data itself. Importantly, while PeptideProphet considers each PSM in the context of the whole population of correct and incorrect PSMs, it does not take into account information about other PSMs in the data set that identify the same peptide sequence. The main statistical unit of PeptideProphet, therefore, is the posterior probability of PSM.

After computing PSM probabilities, the analysis continues at the protein level. This task is carried out by ProteinProphet, which takes as input the list of PSMs and computed posterior probabilities (the output from PeptideProphet) and uses this information to estimate the probability that a particular protein is present in the sample. The protein probability *P(Prot)* is computed as the probability that at least one PSM corresponding to the protein is correct:

$$P(Prot) = 1 - \prod_i (1 - P'(Pep_i)) \quad \text{(Eq. 2)}$$

In computing protein probabilities using Eq. 2, PSMs corresponding to the same peptide ion are represented by a single contribution having maximum probability. All remaining PSMs, however, are considered as independent evidence for their corresponding protein (index i in Eq. 2 labels distinct peptide precursor ions). In other words, the model considers related PSMs corresponding to the same unique peptide sequence but having different precursor ion charge state or modification status on equal footing with PSMs corresponding to completely different peptide sequences.

The protein-level model recognizes the nonrandom nature of peptide to protein grouping, *i.e.* the fact that correct peptides, more than the incorrect ones, tend to correspond to a small number of proteins (each identified by multiple peptides). It attempts to correct for this bias and prevent overestimation of protein level probabilities by adjusting the initial probabilities computed by PeptideProphet, *P(Pep)*, to account for the protein grouping information (the number of sibling peptides, NSP), indicated as *P'(Pep)* in Eq. 2. Intuitively, ProteinProphet rewards peptide precursor ions that have many sibling ions corresponding to the same protein, and punishes those that have few (many of which are "single hit" protein identifications).

Extended Multilevel Modeling in iProphet—iProphet extends the initial approach described above by introducing

additional levels more accurately reflecting the nature of shotgun proteomic data: peptide precursor ions (LC-MS/MS run specific as well as all runs combined), peptides (modification-specific), and unique peptide sequence (see Fig. 1). The PSM level and other levels of information within the multilevel model are linked via the corresponding grouping variables. For example, the existence of multiple PSMs corresponding to the same peptide precursor ion is taken into account by introducing a new variable, the number of replicate (or repeated) spectra (NRS), and using it in a manner similar to the NSP adjustment in ProteinProphet. The iProphet framework further extends it to include multiple search engines by introducing an additional level of information capturing the search engine-specific PSM. Specifically, iProphet implements five additional models to refine the initial PeptideProphet computed probabilities. Each model uses an additional type of information (Fig. 1) using the following grouping variable (see Methods for detail):

(*i*) *Number of Sibling Searches, NSS:* The NSS model rewards or penalizes identifications based on the output of multiple search engines for the same set of spectra.

(*ii*) *Number of Replicate Spectra, NRS*: The NRS statistic models the intuition that in a typical data set, multiple high probability identifications of the same precursor ion should increase the confidence of that precursor ion being correctly identified. On the other hand, repeated observation of PSMs having low to intermediate probabilities and corresponding to the same peptide ion suggests that all those PSMs are false.

(*iii*) *Number of Sibling Experiments, NSE:* The NSE statistic is used to model multiple identifications of the same peptide precursor ion across different "experiments" under the assumption that precursor ions that are observed in multiple experiments and matched to the same peptide sequence are more likely to be correct. In this context, experiments can be repeat analyses of the same sample, analyses of different fractions of the same sample or different biological samples from the same species. Sample origins play a significant role as far as having certain peptides or proteins appear in only some experiments. For instance, some samples may be enriched for particular proteins or be restricted to specific tissues and would contain proteins unique to samples of that type. It is up to the researcher to define reasonable boundaries between experiments, by assigning different experiment tags to each experiment being analyzed.

(*iv*) *Number of Sibling Ions, NSI:* The NSI model rewards peptides that are identified by two or more peptide precursor ions of different charge. This model is based on the empirical observation that correct peptide matches are often identified in more than one charge state, whereas incorrect identifications of the same peptide are less likely to be observed in multiple charge states (with high scores).

(*v*) *Number of Sibling Modifications, NSM:* The NSM model rewards peptides that are identified with different mass modifications. This model is based on the fact that incorrect identifications matching the same peptide sequence, but with two different mass modifications (with high scores in both cases), are less likely to be observed. Mass modifications could arise, *e.g.* from oxidation of methionine or by other intended (*e.g.* SILAC or ICAT labeling(6)) or artifactual modifications that are anticipated as variable modifications in search engines.

Probability Calculation—Each of the models described above is learned as a mixture of two distributions, representing the correct and incorrect PSMs. Starting initially with PeptideProphet probabilities, iProphet applies the EM algorithm to concurrently learn all of the new mixture models in an iterative fashion. On the first iteration, initial PeptideProphet probabilities (computed using Eq. 1) are used as an estimate of correctness to compute the distributions among correct and incorrect PSMs, for each of the new statistics being modeled by iProphet. The adjusted probability of each PSM, denoted here as *Pr(Pep)*, is computed using Bayes Law:

$$Pr(Pep) = \frac{P(G|+)P(Pep)}{P(G|+)P(Pep) + P(G|-)(1 - P(Pep))} \quad \text{(Eq. 3)}$$

Where $G = \{NSS, NRS, NSE, NSI, NSM\}$, $P(Pep)$ is the initial PeptideProphet probability, and $P(G|+)$ and $P(G|-)$ denote the joint probability distributions among correct and incorrect PSMs. Assuming conditional independence of these statistics, the distribution $P(G|+)$ is computed as the product of the individual distributions, $P(NSS|+)P(NRS|+)P(NSE|+)P(NSI|+)P(NSM|+)$, and similarly for $P(G|-)$. During subsequent iterations, the algorithm recalculates the probability of each PSM based on the initial PeptideProphet probability and the rewards and penalties awarded by the mixture models of the iProphet statistics determined by the ratio of the positive and negative distributions, $P(G|+)/P(G|-)$, learned in the previous iteration. The individual distributions are estimated using the kernel-density estimation procedure (32). The algorithm stops when all values have converged. Note that because all models are learned concurrently the order in which the models are applied does not affect the results.

Application of iProphet to the initial PeptideProphet results produces adjusted PSM probabilities (Eq. 3). The main outcome of iProphet is the identification probability at the unique peptide sequence level, taken as the maximum probability of all PSMs corresponding to that sequence. Hence, the input into ProteinProphet is the set of unique peptide sequences and their probabilities. ProteinProphet performs further adjustment of the peptide probabilities for the number of sibling peptides, NSP (with the definition of NSP modified compared with the conventional workflow to count unique peptide sequences). NSP adjusted probabilities are then used to calculate the final protein probability as in Eq. 2. In the presence of shared peptides (*i.e.* peptides whose sequence is present in multiple entries in the protein sequence database), Pro-

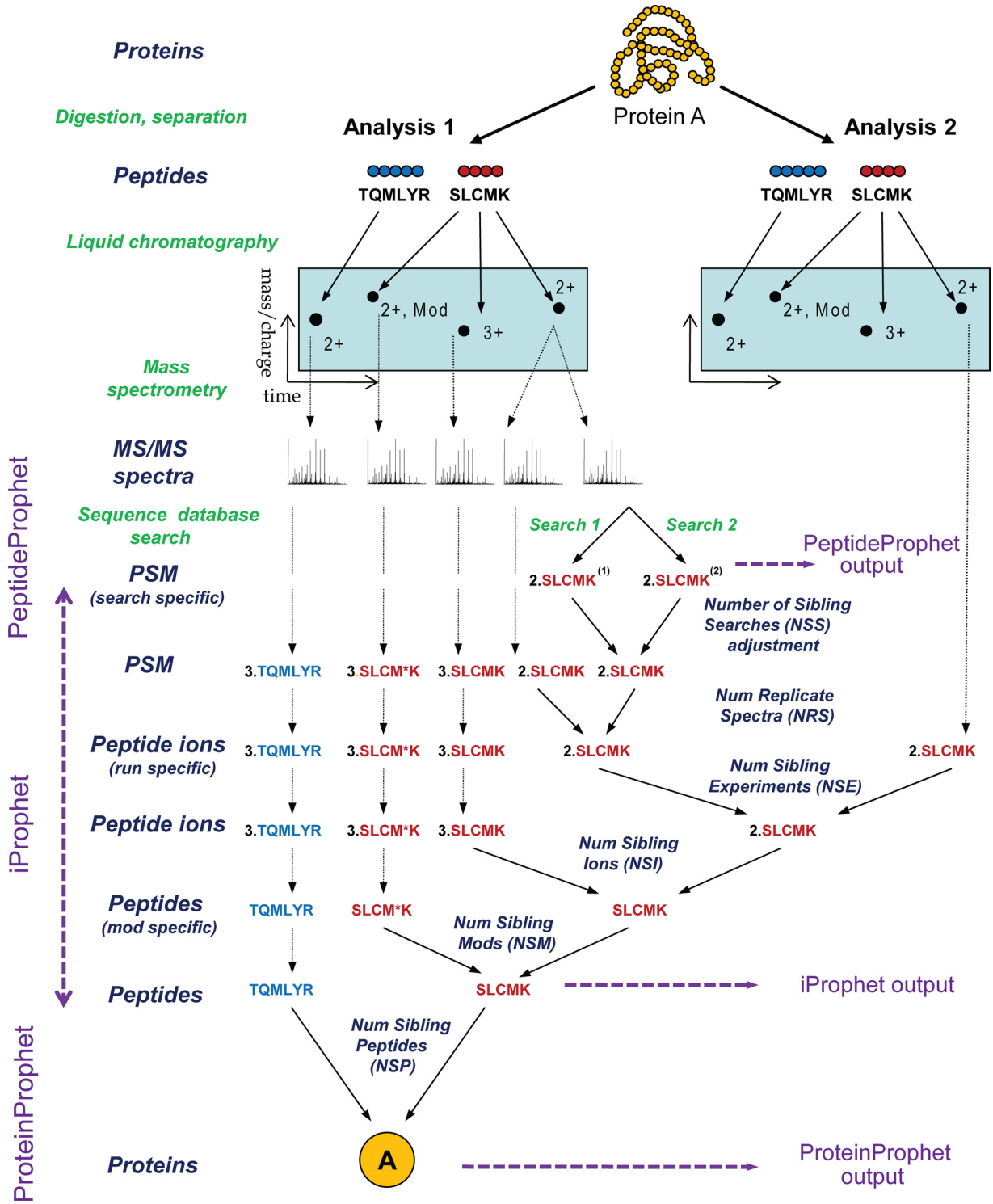

FIG. 1. **Overview of shotgun proteomic data and the computational strategy.** The protein sample is digested into peptides, with some peptides present in the unmodified and a modified (*e.g.* oxidized methionine) forms. The peptide sample is separated using liquid chromatography (LC) coupled online with a tandem mass spectrometer. The first stage of MS measures mass to charge ratios of peptide ions injected in the instrument at any given time. A peptide can be ionized into multiple peptide precursor ions having different charge state (*e.g.* 2+ and 3+). Selected peptide ions are subjected to MS/MS sequencing (some multiple times). Each acquired MS/MS spectrum is assigned a best matching peptide sequence using sequence database searching. When multiple search tools are applied in parallel (Search 1 and Search 2), each spectrum produces multiple peptide to spectrum matches (*search-specific PSM level*), which could be the same or different peptides summarized at the *PSM level*). Within the same LC-MS/MS run, the same peptide ion can be identified from multiple PSMs (*run-specific*

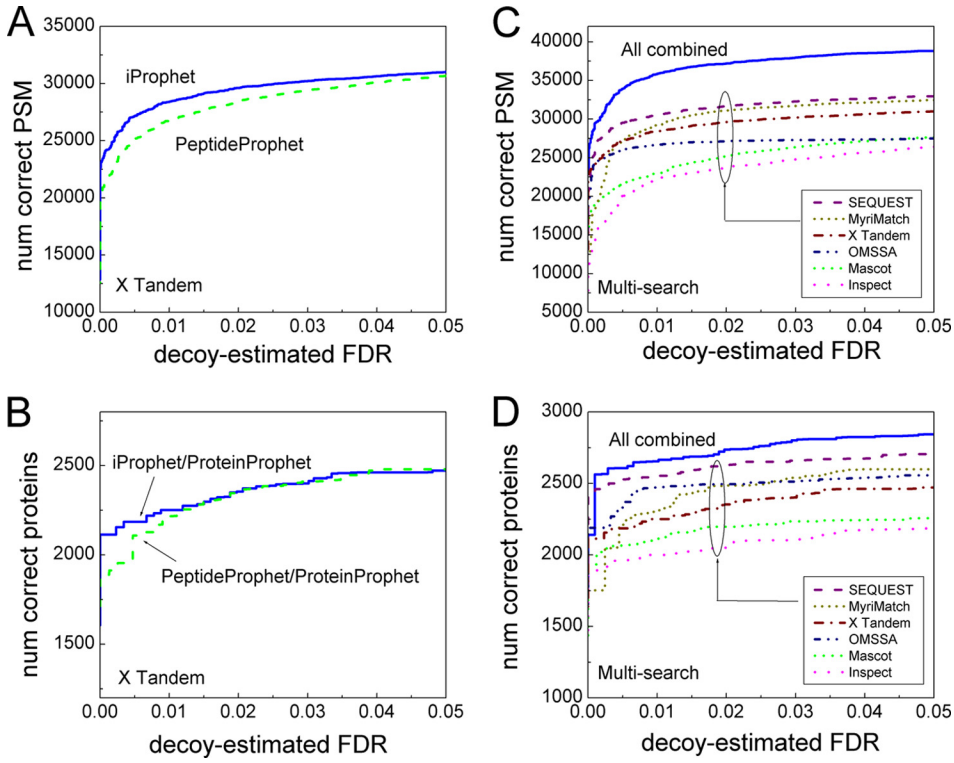

FIG. 2. **Discriminating power of computed probabilities.** *A*, The number of correct PSMs as a function of FDR obtained using iProphet (*solid blue line*) and PeptideProphet (*green dashes*). Human data set, X! Tandem search. *B*, Same as (*A*), at the protein level, after application of ProteinProphet. *C*, The number of correct PSMs as a function of FDR obtained using iProphet when analyzing individual search engine results (six search engines listed in the box), and all search engines combined (*solid blue curve*). *D*, Same as (*C*), at the protein level, after application of ProteinProphet.

teinProphet apportions such peptides across the corresponding protein entries and performs protein inference as described previously (11). The need to deal with shared peptides concurrently with NSP adjustment explains why the adjustment for NSP is still performed in ProteinProphet, *i.e.* not in iProphet, even though it would be appealing to perform the entire analysis, from PSM to protein level, within a single model.

Analysis of iProphet Performance—The performance of iProphet was first investigated using a human data set representative of data sets generated in a typical experiment (see Methods). The data set was searched with six different search engines (SEQUEST, X! Tandem, MyriMatch, OMSSA, Inspect, and Mascot), and the output from each search engine was processed using PeptideProphet (see Methods). ProteinProphet was then applied to PeptideProphet results for each search engine individually (producing search engine-specific summaries), as well as to all search engines combined. The analysis was repeated with an addition of iProphet applied prior to ProteinProphet. The results were compared at three levels (PSM, unique peptide sequence, and protein) in terms of the ability of computed probabilities to separate between correct and incorrect identifications (discriminating power), as well as their accuracy.

The use of iProphet resulted in a gain of about 10–15% correctly identified PSMs at an FDR of 1%, depending on the search engine. The results of the analysis based on X! Tandem search results are shown in Fig. 2*A* (see supplemental Figs. S1 for other search engines). The figure plots the number of correct PSMs as a function of FDR, estimated with the help of decoys (NC_τ *versus* FDR_τ , see Methods), in the most relevant range of FDR values below 5%. Overall, the most significant improvement (percent-wise) at the PSM level was achieved for the search engines that, in this particular

peptide ion level). The experiment may consist of several LC-MS/MS analyses (Analysis 1 and 2), in which case the same peptide ion can be identified in multiple runs (*peptide ion level*). Considering the modification status, the same peptide can be identified in multiple forms (*modification-specific peptide level*), which are then further collapsed into a single identification at the *unique peptide sequence level*. Multiple unique peptide sequences may correspond to the same protein (*protein level*). PeptideProphet calculates the posterior probability of a correct PSM, individually for each search engine output. iProphet combines multiple lines of evidence and computes accurate probabilities at the level of unique peptide sequences, assisted by the introduction of new grouping variables: NSS, NRS, NSE, NSI, and NSM. ProteinProphet combines peptide probabilities to compute the protein probability (with an additional adjustment for NSP).

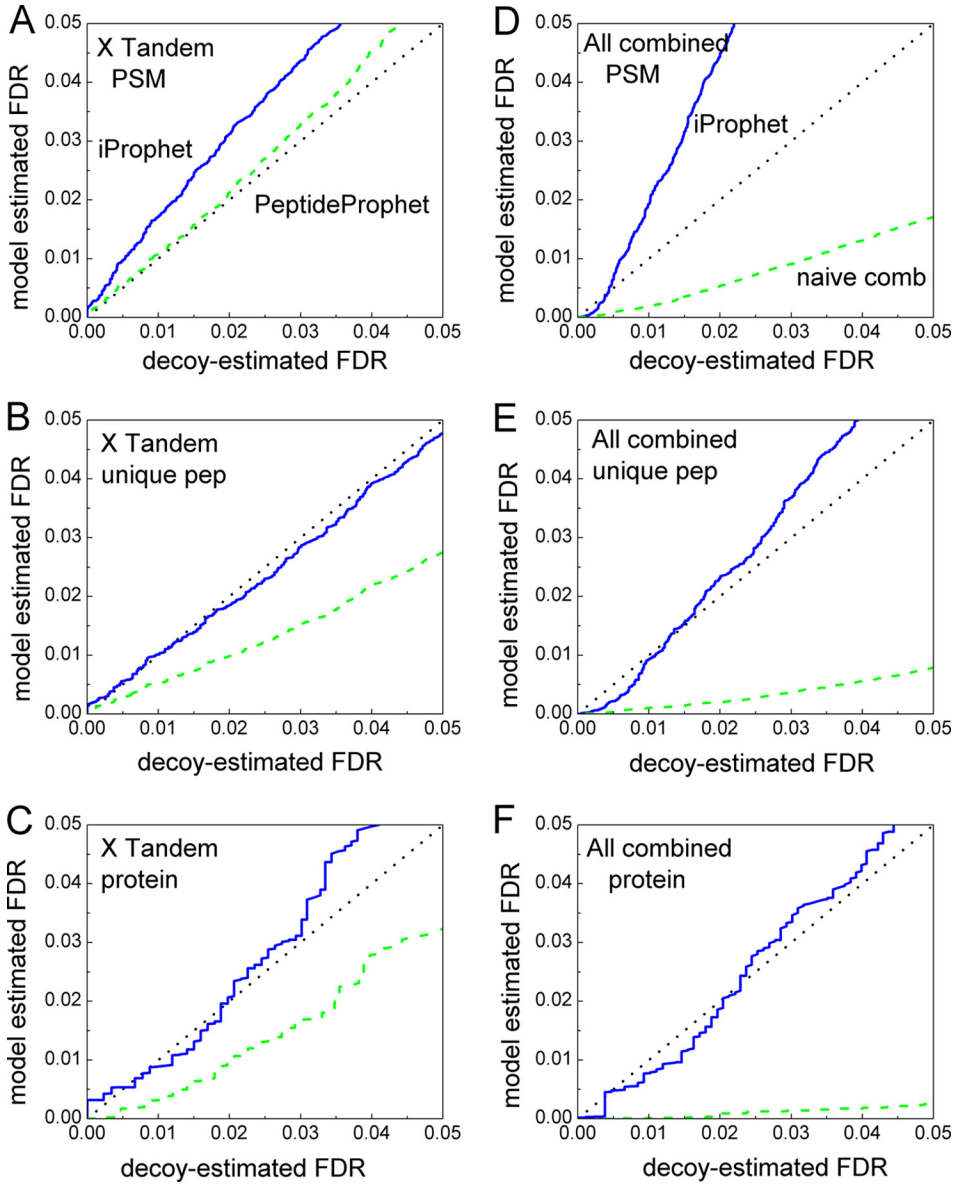

FIG. 3. **Accuracy of probability-based FDR estimates.** FDR estimated using probabilities computed by the iProphet model (*solid blue line*) and by PeptideProphet (*green dashes*) plotted as a function of FDR estimated using decoys. A perfect agreement between the two methods (probability-based and decoy-based) is indicated by a 45-degree dotted line. *A*, X! Tandem, PSM level. *B*, X! Tandem, unique peptide sequence level. *C*, X! Tandem, protein level. *D*, All six search engines combined using iProphet with NSS model enabled (*solid blue line*), or simply by selecting the identification having the highest PeptideProphet probability across the individual search results ("naïve combination"). FDR estimated at the PSM level. *E*, All search engines combined using iProphet or using the naïve approach, unique peptide sequence level. *F*, All search engines combined using iProphet or using the naïve approach, protein level (after application of ProteinProphet).

data set, had the worst performance (Inspect, Mascot). At the protein level, the use of iProphet in the pipeline either improved or did not significantly affect the results (see Fig. 2*B* for X! Tandem, and supplemental Fig. S2 for other search engines). As expected, combining all search engines with iProphet and inclusion of the NSS model yielded an additional improvement of ~4600 correctly identified PSMs (~15% gain) over iProphet's best single search engine performance (SEQUEST) at an error rate of 1% (Fig. 2*C*). A substantial improvement was also observed at the protein level (Fig. 2*D*).

Next, the accuracies of posterior probabilities were investigated by comparing the model-based FDR estimates with the decoy-based estimates at the PSM, unique peptide sequence, and protein levels. The FDR computed using PeptideProphet and PeptideProphet/iProphet probabilities (FDR_{τ}^{mod} see Methods) for a single search engine X! Tandem are shown in Fig. 3*A*–*C* (see supplemental Figs. S3–S5 for other search engines). At the PSM level, for which it was designed, PeptideProphet produced highly accurate probability estimates, as indicated by a very close agreement be-

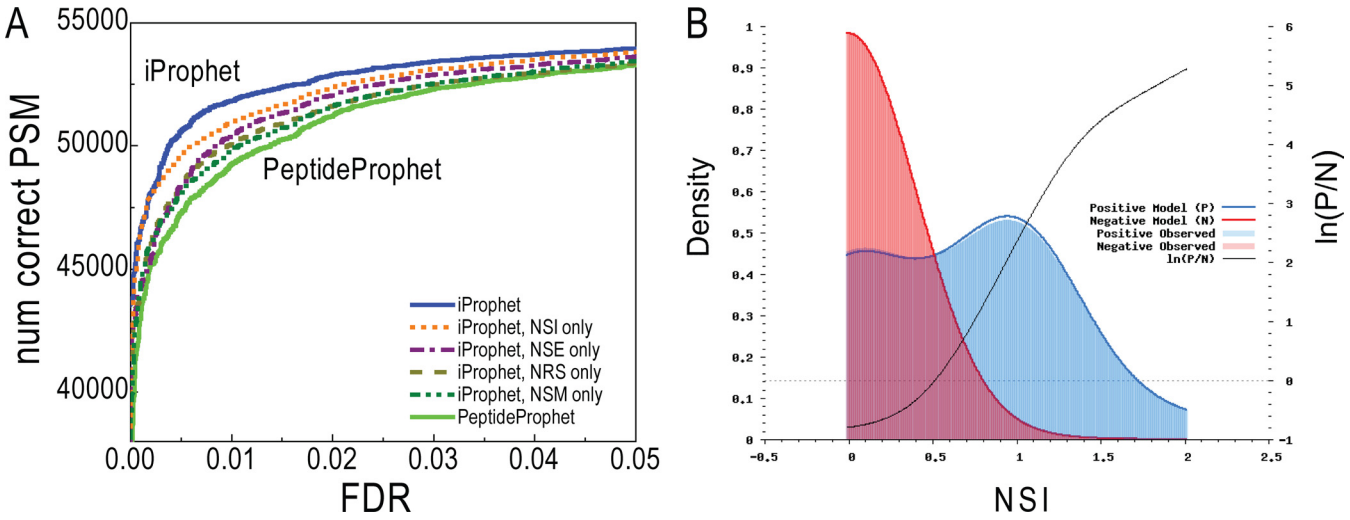

FIG. 4. **Contribution of different models in iProphet.** *A*, The number of correct PSMs as a function of FDR obtained using PeptideProphet (*green dashes*), iProphet (*solid blue line*), and using iProphet with only a single model enabled: NSM, NRS, NSE, or NSI. *B*, The distributions of the number of sibling ions, NSI, statistics among incorrect (*red*) and correct (*blue*) identifications. The shaded areas represent the actual distributions observed, $P(NSI|-)$ and $P(NSI|+)$, labeled as negative (N) and positive (P), respectively. The red and blue solid lines show the iProphet modeled distributions. The solid black curve represents the natural log of the ratio $P(NSI|+)/P(NSI|-)$. When the ratio of the distributions is above 1 (0 on the log scale, indicated by the dotted horizontal line), the model boosts the probability of a PSMs having NSI value in that range, and reduces the probability in the range of NSI values where the ratio drops below 1. *S. pyogenes* data set, SEQUEST search.

tween the probability-based and decoy-based FDR estimates (a perfect agreement is indicated by a 45-degree line). On the other hand, the iProphet computed probabilities at the PSM level tend to be conservative (*i.e.* lower than decoy predicted probabilities). This is not the case, however, at the level of unique peptide sequences and at the protein level, where the FDR estimates based on PeptideProphet computed probabilities become less accurate (Figs. 3*B* and 3*C*, and supplemental Fig. S4 for other search engines). Simply taking the maximum PeptideProphet probability among all PSMs identifying the same unique peptide sequence overestimates the probability at that level, resulting in underestimated FDR. In contrast, iProphet computes more accurate probabilities (as indicated by good agreement between the decoy-based and probability based FDR estimates, see Fig. 3*B*) at the unique peptide sequence level because it takes into account all information that goes into the identification of a unique peptide sequence. This trend continues at the protein level, as shown in Fig. 3*C* (see also supplemental Fig. S5). In all cases, the protein probabilities computed with the help of iProphet are more accurate than the standard PeptideProphet/ProteinProphet probabilities.

Importantly, iProphet allows computing accurate probabilities even when combining the results of multiple different database search tools (Fig. 3*D*–3*F*). The naïve way of combining the search results (*i.e.* taking the maximum probability assignment for each MS/MS spectrum across all search tools) generates, as expected,(35) a significant overestimation of the probabilities as compared with decoy predictions, even at the PSM level (Fig. 3*D*). In contrast, iProphet remains conservative at the PSM level even when multiple search engines are combined, and the probabilities and FDR estimates remain accurate at the unique peptide sequence level and the protein level (Figs. 3*E* and 3*F*, respectively).

The same trends observed above for the human data set are seen with the *S. pyogenes* data set. This multi-experiment data set, representative of composite organism-specific data sets, challenges the performance of the standard TPP pipeline and further highlights the utility of iProphet. On this data set, the improvements in the number of correct PSMs at FDR of 1% ranged from 10% to 30% (see supplemental Figs. S6 and S7 for PSM and protein-level results, respectively). Although the comparison of search engine performances is not the focus of this paper, it is interesting to note that the ranking of individual search engines (based on the estimated number of correct PSM or protein identifications at a fixed FDR) was different in this data set than in the human data set (*e.g.* Inspect has performed substantially better). Furthermore, there was a substantial variation in the ranking across individual data sets generated on different instruments. This indicates that it would be difficult to define an optimal search engine or search strategy for the analysis of a particular data set. Combining the results from all search engines with iProphet and inclusion of the NSS model yielded an additional improvement of ~11,000 correctly identified PSMs (~20% gain) over iProphet's best single search engine performance (see supplemental Fig. S8). Importantly, the probabilities computed when using iProphet in the pipeline were again significantly more accurate than those produced by the standards PeptideProphet/ProteinProphet workflow, both for the individual search engine results (see supplemental Figs. S9–S11 for PSM, unique peptide sequence,

and protein levels, respectively), and for all search engines combined (supplemental Fig. S12). Overall, the analysis presented above for both data sets demonstrates that the use of iProphet in the postprocessing analysis improves the number of correct results at a given FDR. Even more importantly, it improves the accuracy of the reported probabilities.

To investigate the contribution of individual iProphet models to the overall improvement in performance at the PSM level, iProphet was applied to the *S. pyogenes* data set searched with X! Tandem. When running iProphet, only one of iProphet's NSE, NSI, NSM, and NRS models was enabled at a time (NSS model is not applicable when using a single search engine). Fig. 5*A* shows an increase in the number of correct PSMs provided by each separate model in iProphet, with respect to the initial PeptideProphet analysis. The results of running iProphet implementing all applicable iProphet models are shown for the reference as well. In this data set, the number of sibling ions (NSI) model made the largest contribution. The NSI distributions among correct and incorrect PSMs learned by iProphet in this data set are shown in Fig. 5*B*. There is a substantial difference between correct and incorrect PSMs in terms of their NSI properties, with correct PSMs having on average higher NSI values (on the NSI scale from 0 to 2, see Methods). In other words, a peptide identification based on an MS/MS spectrum acquired on, *e.g.* a 2+ charged peptide precursor ion is more likely to be correct if the same peptide was also identified with high probability from a 3+ charge state MS/MS spectrum (NSI = 1; natural log of $P(NSI = 1|+)/P(NSI = 1|-) \sim 2$), and even more so if it is also identified from a 4+ MS/MS spectrum (NSI = 2; $ln(P(NSI = 2|+)/P(NSI = 2|-)) \sim 5$). Also note that, as described above, the identifications of peptides from precursor ions of different charge state are no longer treated as independent events (unlike the conventional PeptideProphet/ProteinProphet workflow). Instead, after rewarding (penalizing) PSMs having high (low) NSI values, and similarly with other variables, all PSMs corresponding to the same unique peptide sequence make a single contribution toward the protein probability (Eq. 2). As a result, application of Eq. 2 leads to more accurate protein probabilities.

Fig. 6 shows the distributions of NSE, NSI, NSM, NRS, and NSS variables learned by iProphet in the *S. pyogenes* data set, all search engines combined. It shows that the NSS model, when applicable, is also highly discriminative. It should also be noted that the importance of different models in iProphet varies depending, among other factors, on the experimental protocols used to generate the data. For example, the NSM model becomes a highly discriminating model in quantitative experiments based on stable isotopic labeling of peptides or proteins, such as SILAC or ICAT experiments (see supplemental Fig. S13 for an example of the NSM distributions in a SILAC data set). In such experiments, peptides observed in multiple modified forms (*e.g.* light and heavy SILAC-labeled peptides) receive a significant probability reward if both forms are identified with high scores.

Comparison to Percolator—The performance of PeptideProphet and iProphet was compared with that of Percolator(36) (version 1.14b), employed here as a state-of-the-art benchmark. Percolator employs a semisupervised machine learning approach to differentiate between correct and decoy PSMs. For a discussion on the differences and similarities between PeptideProphet and Percolator see (37, 38). Of the six search engines used in this work, the Percolator software was only able to process SEQUEST search results.

Combining data from all six search engines with iProphet showed a clear improvement over SEQUEST/Percolator in both data sets (see supplemental Fig. S14). To perform a more direct comparison, PeptideProphet, PeptideProphet/iProphet, and Percolator were run on SEQUEST search results only. On the Human data set (Fig. 6*A*), Percolator slightly outperformed PeptideProphet. The performance of iProphet and Percolator was essentially equivalent in the most relevant range of low FDR values (below 5%). Percolator slightly outperformed iProphet at higher error rates (see Fig. 6*A* inset). This can be explained by the fact that Percolator is able to consider non-top-hit matches for each MS/MS spectrum, whereas PeptideProphet and iProphet only consider the top hit. The option of using top ten best scoring peptides per MS/MS spectrum within the PeptideProphet framework was previously investigated as well,(38) and demonstrated a ~5–10% improvement in number of correct PSMs in the high FDR region. This option is being considered for implementation in the TPP. However, it is not expected to significantly increase the number of identifications at the unique peptide or protein levels.

The analysis was repeated using the LTQ-FT subset of the *S. pyogenes* data set (Fig. 6*B*). In this case, PeptideProphet was equivalent or even slightly better than Percolator in low FDR range. This is likely because of the ability of PeptideProphet to effectively use high mass accuracy of the MS measurement on this type of instrument in the model. As a result, iProphet outperformed both Percolator and PeptideProphet in that FDR range. As in the Human data set, Percolator was able to identify ~5% more PSMs in higher FDR range (see Fig. 6*B*, inset) by going beyond the top peptide assignment per spectrum.

Software Implementation of iProphet—iProphet has been conceived and implemented as an integral part of the TPP. A tutorial describing how to use the TPP has been published (12) and an online version of the tutorial is available at *http://tools.proteomecenter.org/wiki/index.php?title = TPP_Demo2009*. The workflow applies PeptideProphet to model each search of each data set (see Fig. 7). iProphet then takes as input one or more pepXML files that have been processed with PeptideProphet and thus contain PeptideProphet probabilities. iProphet can be applied to PeptideProphet results separately for each search engine, or used to combine the results of multiple searches and multiple experiments.

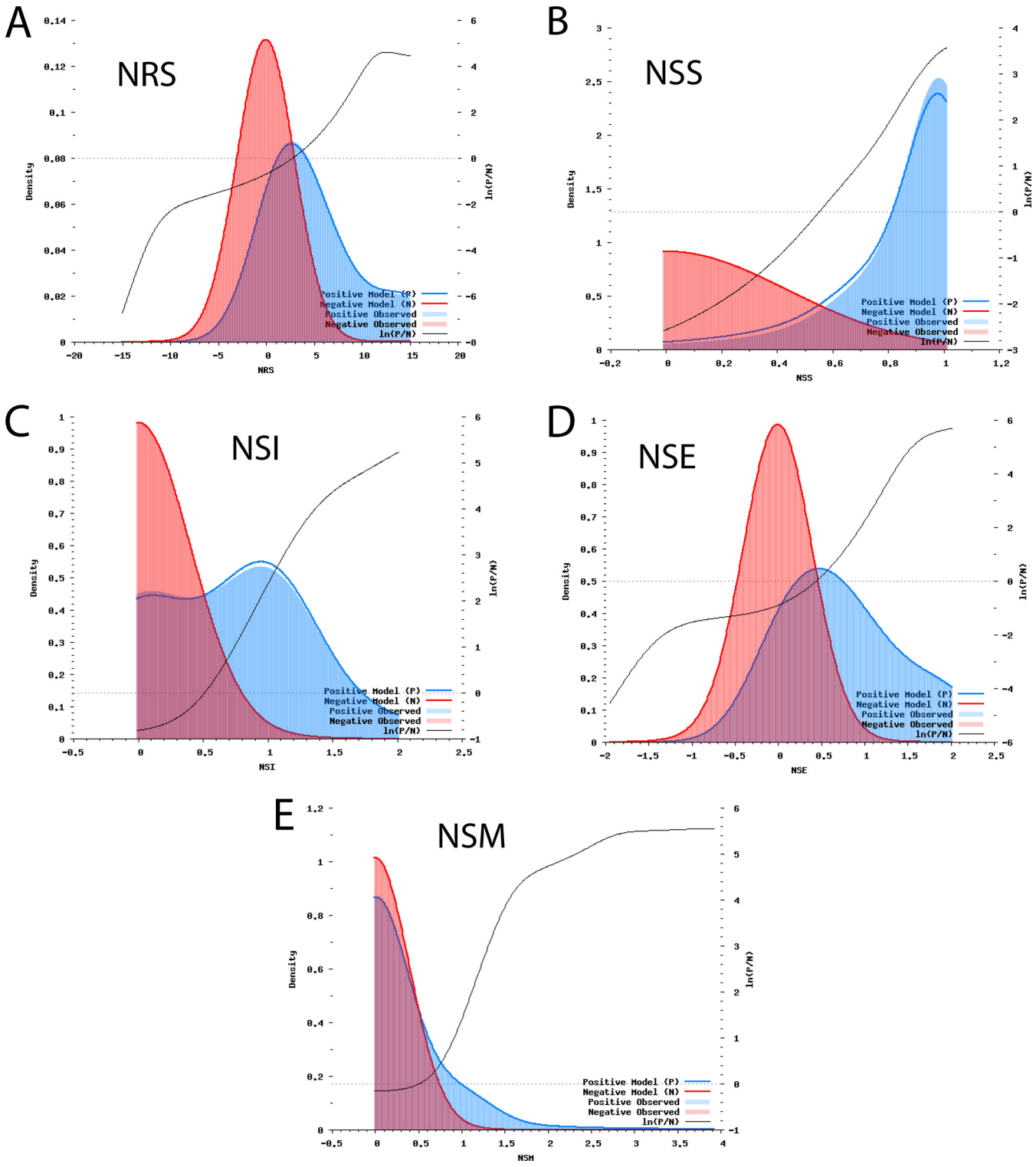

FIG. 5. **The distributions of grouping statistics learned by iProphet.** The negative (*red*) and positive (*blue*) distributions for all the five grouping variable used in iProphet. See Fig. 4*B* legend for detail. *A*, Number of replicate spectra, NRS. *B*, Number of sibling searches, NSS. *C*, Number of sibling ions, NSI. *D*, Number of sibling experiments, NSE. *E*, Number of sibling modifications, NSM. *S. pyogenes* data set, all search engine combined.

The output of iProphet is another pepXML file that contains the top scoring entry for each spectrum in the input file; iProphet writes the probabilities it computes to the resulting pepXML file. The statistics outlined above (NSS, NRS, NSE, NSI, and NSM) for each PSM are reported in the output file as well and also the global distributions learned for the

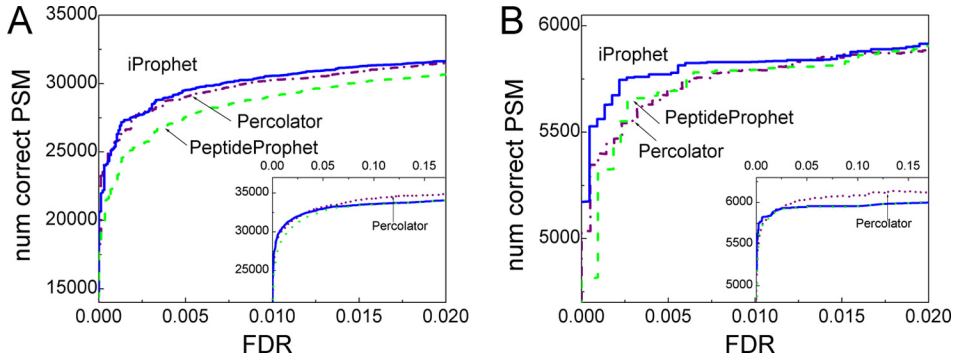

FIG. 6. **Comparison between iProphet, PeptideProphet, and Percolator.** The number of correct PSMs as a function of FDR obtained using iProphet (*solid blue line*), PeptideProphet (*green dashes*), and Percolator (*purple, dash dot*), applied to SEQUEST search results. Inset shows an extended range of FDR values (up to 20%). *A*, Human data set. *B*, FFE-LTQ-FT subset of the *S. pyogenes* data set.

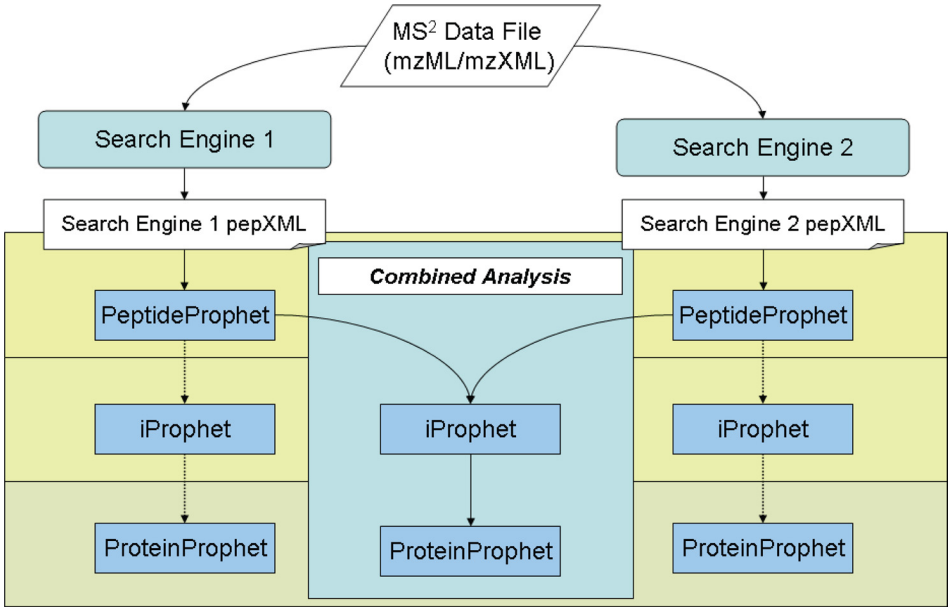

FIG. 7. **Overview of the possible TPP workflow.** Analysis with iProphet can be performed as an intermediate step between PeptideProphet and ProteinProphet in a single search analysis or a combined search analysis.

entire data set. Not all data sets analyzed by iProphet can use all models and unneeded models are automatically disabled. For example, the model for multiple sibling searches is not useful when only a single search is being processed and as a result, it is automatically disabled in this analysis scenario. Each of the iProphet models can also be manually disabled by the user. The output from iProphet is further processed using a modified version of ProteinProphet, which produces a protXML file with the protein level summary of the results.

iProphet is implemented in C++, like PeptideProphet and ProteinProphet, and is automatically available upon installation of the TPP software suite. Because it is a part of the TPP, it works on all three major operating system platforms: Windows, Linux/Unix, and OS X. It works with instruments from all major instrument vendors as converters from vendor-specific binary formats to the open XML-based formats (39) of the TPP are included. Further, all the visualization and quantification tools of the TPP work seamlessly with iProphet output. The TPP already has full support for many of the most popular search engines including Mascot (25), SEQUEST (26), X! Tandem (27), ProbID (40), and SpectraST (41). Beta support exists for Phenyx (42), Inspect (29), MyriMatch (30), and OMSSA (31). Therefore, iProphet can likely work with the engine of choice at most sites.

The iProphet program, as well as all other TPP programs, is free and open source software. The source code is available in a publicly accessible source code repository hosted on SourceForge. Current users of the TPP need only upgrade to the latest version to have immediate access to iProphet to process new and older data sets. For new users, the installation of the TPP is easy and is available on Windows, Linux, and OS X. Installation instructions are available at the Seattle Proteome Center web site: http://www.proteomecenter.org/software.php.

DISCUSSION

The main task of PeptideProphet and ProteinProphet, and the new iProphet tool described here is the statistical analysis of and integration of peptide and protein identifications in MS/MS-based proteomic data sets. In the last decade, several strategies have emerged that address this problem. Therefore, it should be informative to discuss the methods employed in this work with respect to other approaches (for an in-depth discussion on this subject see (37)). First, many search engines, including Mascot, X! Tandem, and OMSSA, convert the original search scores into expectation values (*E*-values (43)) and report them as confidence scores for individual PSMs. The steps used to compute *E*-values essentially represent the conventional *p* value computation as the tail probability in the distribution generated from random matches, with the random distribution estimated under certain parametric (*e.g.* Poisson) assumptions (31, 44), via theoretical derivation of the tail part of the random distribution (45), or using empirical fitting (43, 46). An alternative approach, which falls in the same category of single-spectrum statistical confidence scores as *E*-values, is based on the concept of generating functions (47). The advantage of these scores over the original search scores is that they are largely invariant under different scoring methods, allowing a clearer interpretation of goodness of the PSM across different data sets. However, *p* values/*E*-values and other single-spectrum statistical scores are not sufficient when the analysis involves simultaneous processing of multiple MS/MS spectra. Thus, additional modeling is necessary to calculate statistical measures more suitable for filtering of large collections of MS/MS database search results, such as FDR(48) and posterior peptide and protein probabilities (37, 49–51).

In shotgun proteomics, the methods for estimating FDR can be broadly divided into two categories. The simple target-decoy strategy(52) requires that MS/MS spectra are searched against a database containing target and decoy sequences and assumes that matches to decoy peptide sequences and false matches to target sequences follow the same distribution (52). The main advantage of this strategy is minimal distributional assumptions and the ease of implementation. At the same time, the simple target-decoy strategy does not directly provide a probability score for individual PSMs or proteins. Instead, a more elaborate statistical analysis can be carried out using a mixture model approach (53)—the strategy implemented in PeptideProphet and its extension iProphet. These tools calculate the posterior probability for each individual PSM as the baseline measure for distinguishing between true and false identifications. These probabilities are directly related (33) to another local error rate measure, the local FDR (sometimes referred to as peptide error probability), and can be used to estimate FDR for an entire filtered subset of PSMs (*e.g.* accepting PSMs with probability greater or equal than a 0.99 threshold). They can also be taken as input to the protein level analysis (ProteinProphet), which includes the calculation of protein level probabilities and FDR estimation. Another important advantage of PeptideProphet is that it is easily expandable to include additional information about the peptides being matched. Such types of information include mass accuracy, peptide separation coordinates (*e.g.* pI, retention time), digestion properties (the number of enzymatic termini expected from the specificity of the protease), and more. PSMs are rewarded or penalized based on their concordance with the expected values for each of these attributes; the incorporation of this information into the model further increases the discrimination between correct and incorrect identifications.

Recent improvements in PeptideProphet tools have substantially improved its robustness and general applicability. Although the original implementation of PeptideProphet was based on unsupervised mixture modeling, the semisupervised version (33) can incorporate decoys into the model for improved robustness of the modeling in the case of challenging data sets (*e.g.* where the model does not converge because of a very small population of correct PSMs). The parametric assumptions were relaxed with an introduction of the semiparametric model (32), which made it possible to apply PeptideProphet to a larger number of database search tools. These advances eliminate most practical disadvantages when compared with the simple decoy-based approach, while providing a number of significant advantages as described above.

Nevertheless, several important limitations of the existing PeptideProphet/ProteinProphet tools when applied to very large data sets have become apparent in recent years. As we noticed early on (11) one of the most critical challenges is the inflation of the error rates when going from PSM to unique peptide sequence to protein level. The NSP adjustment implemented in ProteinProphet was designed to address this problem, and it works well in the case of small to intermediate data sets. One of the main aims of the iProphet program described here is to calculate accurate probabilities at the level of unique peptide sequences starting with the PeptideProphet probabilities, which are accurate at the PSM level. This, in turn, leads to more accurate probabilities and FDR estimates at the protein level. Inclusion of the five additional models in iProphet to better reflect the multilevel structure of shotgun proteomic data also improves the separation between true and false PSMs, thus increasing the number of correct identifications. Explicit modeling of each factor (NRS, NSM, etc) also provides insights into the nature of the sample and ensures robustness against different experimental strategies and setups. Another important challenge—dealing with shared peptides and creation of proteins groups for final presentation (54)—continues to be handled by ProteinProphet and thus is not discussed here in detail.

In the absence of gold standard data sets of sufficiently large size and complexity, the assessment of the accuracy of computed probabilities is based on the comparison of prob-

ability-based and decoy count based FDR estimates. However, the decoy-based FDR estimates vary depending on the details of the target-decoy database search, *e.g.* two separate searches against the target and the decoy database or a single search against a concatenated target plus decoy database (55–57) (for a recent review see (37)). Furthermore, none of the existing decoy database creation methods capture all significant sources of false identifications (*e.g.* false positives arising because of sequence homology(49)). Thus, although the trend is clear in that iProphet improves the accuracy of peptide and protein-level estimates compared with the conventional PeptideProphet/ProteinProphet, one should not expect a perfect agreement between the probability-based and decoy-based FDR estimates.

The analysis presented here using two diverse data sets demonstrated that the use of iProphet provides an increased number of correct identifications at the same FDR, and more accurate probabilities at all levels. This has been verified on many other data sets processed as a part of the PeptideAtlas project (58, 59). This project aims to collect raw MS/MS data from many sources and experiments, process them through a single data analysis pipeline, and present the results as a compendium of all peptides and proteins observed in the publicly available data. The iProphet program has become a crucial component of the PeptideAtlas pipeline and serves to further expand the coverage and quality of the database (3).

The iProphet's multilevel modeling framework allows easy integration of multiple database search tools. A multisearch strategy is now fairly easy to apply given the availability of fast computers and a growing number of freely available open-source database search programs. However, despite previous efforts exploring this strategy (35, 60–62), it has not yet been widely used in practice. Integration of the output from multiple search engines into a single summary statement has been cumbersome and time-consuming, in part because of different output formats generated by each search engine. With the availability of iProphet, these multiple search results may be easily combined into one summary file, leveraging the strengths of different search engines to yield a significantly higher number of identifications at a constant FDR. Given iProphet's implementation as a part of the commonly used TPP pipeline, it should make the promising multi-search analysis option available to a large number of users.

Acknowledgments—We thank Moritz, Nesvizhskii, and Aebersold lab members and other early iProphet adopters who helped test and provided feedback to help develop this software.

* This work has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179, from NIH Grants R01 CA126239, R01 GM094231, R01 GM087221, from PM50 GMO76547/Center for Systems Biology, from NSF MRI No. 0923536, and the Systems Biology Initiative of the Duchy of Luxembourg.

S This article contains supplemental Figs. S1 to S14 and Table S1.

|||| To whom correspondence should be addressed: Department of Pathology, University of Michigan, Ann Arbor, 4237 MS1, 1301 Catherine St., Ann Arbor, MI 48105. Tel.: 734-764-3516; E-mail: nesvi@med.umich.edu.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009) Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annu. Rev. Biomed. Eng.* **11,** 49–79
3. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **33,** 18–25
4. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The Need for Guidelines in Publication of Peptide and Protein Identification Data: Working Group On Publication Guidelines For Peptide And Protein Identification Data. *Mol. Cell. Proteomics* **3,** 531–533
5. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–797
6. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–1031
7. Vitek, O. (2009) Getting started in computational mass spectrometry-based proteomics. *Plos Comput. Biol.* **5,** 4
8. Blackburn, K., and Goshe, M. B. (2009) Mass Spectrometry Bioinformatics: Tools for Navigating the Proteomics Landscape. *Curr. Anal. Chem.* **5,** 131–143
9. Wright, J. C., and Hubbard, S. J. (2009) Recent Developments in Proteome Informatics for Mass Spectrometry Analysis. *Comb. Chem. High Throughput Screen* **12,** 194–202
10. 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
11. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
12. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R. (2010) A guided tour of the TransProteomic Pipeline. *Proteomics* **10,** 1150–1159
13. Isserlin, R., and Emili, A. (2008) Interpretation of large-scale quantitative shotgun proteomic profiles for biomarker discovery. *Curr. Opin. Mol. Ther.* **10,** 231–242
14. Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., Lin, Z. Y., Breitkreutz, B. J., Stark, C., Liu, G., Ahn, J., Dewar-Darch, D., Reguly, T., Tang, X., Almeida, R., Qin, Z. S., Pawson, T., Gingras, A.-C., Nesvizhskii, A. I., and Tyers, M. (2010) A Global Protein Kinase and Phosphatase Interaction Network in Yeast. *Science* **328,** 1043–1046
15. Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. (2008) Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 1454–1459
16. Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138,** 389–403
17. Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J., Hafen, E., Schlapbach, R., and Aebersold, R. (2007) A high-quality catalog of the Drosophila melanogaster proteome. *Nat. Biotechnol.* **25,** 576–583
18. Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008) Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science* **320,** 938–941
19. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–U60

20. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417
21. Wu, L., Hwang, S. I., Rezaul, K., Lu, L. J., Mayya, V., Gerstein, M., Eng, J. K., Lundgren, D. H., and Han, D. K. (2007) Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling. *Mol. Cell. Proteomics* **6,** 1343–1353
22. Falkner, J., and Andrews, P. C. (2008) Tranche Project. *http://tranche.proteomecommons.org/*
23. Malmström, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E. W., and Aebersold, R. (2009) Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460,** 762-U112
24. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466
25. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
26. Eng, J., McCormack, A. L., and Yates, J. R. (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
27. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316
28. MacLean, B., Eng, J. K., Beavis, R. C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22,** 2830–2832
29. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: Identification of posttransiationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639
30. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6,** 654–661
31. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964
32. Choi, H., Ghosh, D., and Nesvizhskii, A. I. (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **7,** 286–292
33. Choi, H., and Nesvizhskii, A. I. (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **7,** 254–265
34. Malmström, J., Lee, H., Nesvizhskii, A. I., Shteynberg, D., Mohanty, S., Brunner, E., Ye, M., Weber, G., Eckerskorn, C., and Aebersold, R. (2006) Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **5,** 2241–2249
35. Searle, B. C., Turner, M., and Nesvizhskii, A. I. (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **7,** 245–253
36. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics data sets. *Nat. Methods* **4,** 923–925
37. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73,** 2092–2123
38. Ding, Y., Choi, H., and Nesvizhskii, A. I. (2008) Adaptive Discriminant Function Analysis and Reranking of MS/MS Database Search Results for Improved Peptide Identification in Shotgun Proteomics. *J. Proteome Res.* **7,** 4878–4889
39. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1,** 2005.0017
40. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2,** 1406–1412
41. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667
42. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–1463
43. Fenyö, 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
44. Sadygov, R. G., and Yates, J. R., 3rd (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75,** 3792–3798
45. Alves, G., Ogurtsov, A. Y., Wu, W. W., Wang, G., Shen, R. F., and Yu, Y. K. (2007) Calibrating e-values for MS2 database search methods. *Biol. Direct* **2**
46. Klammer, A. A., Park, C. Y., and Noble, W. S. (2009) Statistical Calibration of the SEQUEST XCorr Function. *J. Proteome Res.* **8,** 2106–2113
47. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.* **7,** 3354–3363
48. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Statistical Soc.* **57,** 289–300
49. Choi, H., and Nesvizhskii, A. I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **7,** 47–50
50. Fitzgibbon, M., Li, Q., and McIntosh, M. (2008) Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.* **7,** 35–39
51. Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior error probabilities and false discovery rates: Two sides of the same coin. *J. Proteome Res.* **7,** 40–44
52. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214
53. Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statistical Assoc.* **96,** 1151–1160
54. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440
55. Navarro, P., and Vazquez, J. (2009) A Refined Method To Calculate False Discovery Rates for Peptide Identification Using Decoy Databases. *J. Proteome Res.* **8,** 1792–1796
56. Blanco, L., Mead, J. A., and Bessant, C. (2009) Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *J. Proteome Res.* **8,** 1782–1791
57. Wang, G., Wu, W. W., Zhang, Z., Masilamani, S., and Shen, R. F. (2009) Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* **81,** 146–159
58. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H. K., Lin, B. Y., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L. H., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**
59. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9,** 429–434
60. Alves, G., Wu, W. W., Wang, G., Shen, R. F., Yu, Y. K. (2008) Enhancing peptide identification confidence by combining search methods. *J. Proteome Res.* **7,** 3102–3113
61. Jones, A. R., Siepen, J. A., Hubbard, S. J., and Paton, N. W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **9,** 1220–1229
62. Edwards, N., Wu, X., and Tseng, C. W. (2009) An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clin. Proteomics* **5,** 23–36