EBP, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets

Thomas S. Price‡, Margaret B. Lucitt‡, Weichen Wu‡, David J. Austin‡, Angel Pizarro‡, Anastasia K. Yocum§, Ian A. Blair§¶, Garret A. FitzGerald‡‖, and Tilo Grosser‡**

MS/MS combined with database search methods can identify the proteins present in complex mixtures. High throughput methods that infer probable peptide sequences from enzymatically digested protein samples create a challenge in how best to aggregate the evidence for candidate proteins. Typically the results of multiple technical and/or biological replicate experiments must be combined to maximize sensitivity. We present a statistical method for estimating probabilities of protein expression that integrates peptide sequence identifications from multiple search algorithms and replicate experimental runs. The method was applied to create a repository of 797 non-homologous zebrafish (Danio rerio) proteins, at an empirically validated false identification rate under 1%, as a resource for the development of targeted quantitative proteomics assays. We have implemented this statistical method as an analytic module that can be integrated with an existing suite of open-source proteomics software. Molecular & Cellular Proteomics 6:527–536, 2007.

Proteomics methods aim to identify the proteins expressed in biological samples (1) just like transcriptomics methods detect and quantify the abundance of RNA molecules (2). Both techniques can point to particular tissue correlates, such as disease biomarkers, and find networks of co-regulated genes and proteins that reveal the mechanisms underlying biological processes (3). As thousands of simultaneous measurements are taken, however, balancing sensitivity of detection against rates of false positive error is a major challenge. With rapid advances in large scale proteomics technology, there is an urgent need for methods that maximize sensitivity and control of false protein identifications by integrating the results of multiple experiments.

In “shotgun” proteomics studies, a protein sample is digested with a proteolytic enzyme, and the resulting peptide mixture is separated by LC prior to conducting collision-induced dissociation and MS/MS (1). This generates thousands of peptide ion fragmentation spectra containing diagnostic amino acid sequence information. Peptide sequences are inferred by matching the resulting product ion spectra to theoretical or empirical spectra in peptide sequence databases (4). No match is exact, so the peptide identifications are not certain. Database search algorithms either calculate a probability that the spectrum results from a random fragmentation (e.g. MASCOT, Ref. 5) or use heuristic scores (such as SEQUEST, Ref. 6), which can then be converted into probabilities of accurate peptide identification (7, 8). A variety of other statistical models have also been proposed to score peptide matches (9). These algorithms have to some extent complementary merits, and the sensitivity and specificity of peptide sequence identifications from MS/MS data can be improved by the use of more than one search algorithm (10–12). A recent comparison of five database search programs concluded with the recommendation to use consensus scoring from at least two different algorithms and noted the need for appropriate methods of combining scores from different algorithms (11). Database search methods using different assumptions and procedures can result in markedly different sets of protein identifications (e.g. Ref. 13).

Probable peptide sequences need to be assembled to proteins (14). A simple method for identifying proteins is to filter the peptide matches using empirically derived thresholds and then sort by parent protein. Variants of this approach do not attempt to estimate false identification rates, nor do they deal in a useful way with “degenerate” peptides whose sequence matches more than one protein in the database (15–17). More sophisticated methods use statistical models to calculate protein expression probabilities (12, 18–20) or ranking scores (21) based on the number and quality of peptide sequence matches and other criteria, such as protein length, the size of the database, and the degeneracy of the peptide sequence identifications. These models can be empirically validated by searching against databases containing random “distractors” such as reversed protein sequences (12, 21–23).

A major problem is that many protein identifications have low reproducibility (24): in general, only abundant proteins are reliably identified (25, 26). Sensitivity is improved by performing multiple technical replicates (27), perhaps using different
Significance Analysis for Protein Identification

instrumentation and technical procedures (10, 25). This creates a demand for analytical methods that can integrate data from multiple experiments, which will likely increase as common data standards emerge that facilitate the exchange of proteomic datasets from diverse sources.1,2

We present a statistical model that implements consensus scoring of peptide identifications from multiple search algorithms and combines information from independent replicate experiments in a single analysis while allowing sensitivity to be balanced against false identification rate. The model assumes that every peptide sequence that could theoretically result from enzymatic digestion of a protein in the database has a chance of being identified in the search results, whether correctly or incorrectly. The probabilities of correct identification are combined across multiple peptide searches using a function that returns the maximum probability from consensus identifications and penalizes non-consensus identifications. Both correct and incorrect peptide sequence identifications are assumed to occur at random in this “space” of peptides at rates that are governed by model parameters including protein length, estimated protein abundance, the size of the search database, and the number of peptide sequence identifications in the dataset. For each protein in the database, a likelihood ratio is calculated for the possibility that the peptide identifications whose sequence matches the protein are all incorrect. These likelihood ratios are used to estimate the expression probabilities from which updated parameter estimates are obtained. The procedure is iterated until convergence at the maximum likelihood estimates. Replicates are integrated by simultaneously estimating multiple sets of model parameters. Degenerate peptides whose sequence matches multiple proteins are treated using “Occam’s razor,” a principle by which the smallest set of probable proteins is chosen that is sufficient to explain the peptide sequence identifications (19). An outline of the statistical model used and its implementation in the program Empirical Bayes Protein identifier (EBP)3 is described under “Experimental Procedures.” The software is available upon request for download as an extension to the Windows release of the Institute of Systems Biology Trans-Proteomic Pipeline.

We applied the model to a set of three biological replicates of zebrafish proteins, searched by both SEQUEST and MAS-COT algorithms. The zebrafish has emerged as an informative model organism for studies of vertebrate biology (41), genetics (28), and pharmacology (29, 30). Its genome has been recently sequenced, and an acceptably complete and non-redundant protein sequence database is available (35). The aim was to identify as many reliably expressed proteins as possible in larvae 5 days postfertilization, a developmental stage at which most organ systems are functional. This would allow the selection of candidate proteins for targeted quantitation (31) in mutagenesis (28) and chemical screens (32).

EXPERIMENTAL PROCEDURES

Sample Collection

Zebrafish embryos were bred and raised from natural matings of wild-type tail long fin fish as described previously (33) and staged according to existing protocols (34). Embryos were collected at 120 h postfertilization, immediately flash frozen using liquid nitrogen, and stored at −80 °C. Frozen embryos were thawed and washed twice with PBS to remove any residual medium before lysis. Five tubes of 20–30 embryos were lysed in 7 M urea, 2 M thiourea, 4% CHAPS (GE Healthcare), 100 mM DTT (Bio-Rad), Phosphatase Inhibitor Mixture 11 (Sigma), and Complete protease inhibitor mixture tablet (Roche Applied Science) and disrupted using a Qiagen TissueLyser (2 × 3 min at 30 Hz). Samples were centrifuged at 12,000 × g for 30 min at 4 °C, and the supernatant was collected. Proteins were precipitated using methanol/chloroform and resuspended in 0.2% (w/v) RapigestSM SF (Waters, Milford, MA), in 50 mM ammonium bicarbonate (Sigma). Samples containing 2.0 mg of protein were reduced with 5 mM DTT for 60 min at 60 °C and alkylated with 5 mM iodoacetamide (Bio-Rad) for 30 min at room temperature. Trypsin (Promega, Madison, WI) digestion using a 1:20 enzyme:protein (w/w) ratio at 37 °C for 16 h was carried out in 0.2% (w/v) Rapigest SF (Waters), in 50 mM ammonium bicarbonate. The pH was then lowered to 3 with 1% formic acid solution followed by incubation at 37 °C for 1 h and centrifugation. The precipitant was removed, and an equal volume of strong cation exchange (SCX) mobile phase A (10 mM ammonium formate, 25% acetonitrile, pH 3) was added. The solution was incubated at 4 °C for 4 h before SCX separation.

Two-dimensional Chromatography and Mass Spectrometry

Protein digestion, off-line fractionation, and LC-MS/MS were performed as described previously (25). SCX separation of the peptides was carried out on a PolySufoethyl A column (100 × 4.6 mm, 5 μm, 300 Å; PolyLC, Columbia, MD) at a flow rate of 0.2 ml/min. The sample was loaded for 10 min with mobile phase A (10 mM ammonium formate, 25% acetonitrile, pH 3), and a gradient was run to 100% mobile phase B (500 mM ammonium formate, 25% acetonitrile, pH 6.8) to elute the peptides over 80 min. Forty fractions were collected for reversed phase separation on line to a Thermo Finnigan LTQ ion trap mass spectrometer (Thermo Electron, San Jose, CA) using ESI. Each fraction was injected onto a Vydac C18 (Everest, 150 × 1 mm, 300 Å, 5 μm; Bodmann, Aston, PA) column with 0.1% formic acid, 0.01% trichloroacetic acid in water for 10 min before a gradient (30 μl/min) was run over 120 min from 3 to 70% mobile phase B (0.1% formic acid, 0.01% trichloroacetic acid in acetonitrile). The mass spectrometer was operated in a data-dependent MS/MS mode (m/z 200–2000) in which the top seven ions were subjected to fragmentation. Dynamic mass exclusion was enabled with a repeat count of 2 every 45 s for a list size of 250.


3 The abbreviations used are: EBP, Empirical Bayes Protein identifier; SCX, strong cation exchange; EM, Expectation-Maximization.
Data Processing

Raw mass spectra were converted to DTA peak lists using BioWorks Browser 3.2 (Thermo Finnigan, San Jose, CA) with the following parameter settings: peptide mass range, 300–5000 Da; threshold, 10; precursor mass, ±1.4 Da; group scan, 1; minimum group count, 1; minimum ion count, 15. Searches were conducted using SEQUEST and MASCOT against a database containing both forward and reverse sequences of proteins contained in the International Protein Index (IPI) Danio rerio protein sequence database version 3.07 (35). It was specified that peptides should have a maximum of two internal cleavage sites with methionine oxidation and cysteine carbamidomethylation as possible modifications. SEQUEST searches specified that peptides should possess at least one tryptic terminus and used a peptide mass tolerance of ±1.4 Da and a fragment ion tolerance of ±0.1 Da. The search results were converted into pepXML format (42). Peptide identification probabilities for both SEQUEST and MASCOT searches were calculated by executing PepTideProphet (7). SEQUEST results were processed using the “O” tag, which uses ΔCn* values unchanged. Peptide sequence identifications with probabilities greater than 0.05 were exported for the protein probability analyses. EBProteinProphet (19) analyses were run using SEQUEST and MASCOT results for each sample both separately and together. EBP analyses were run using the default settings except for the calculation of number of tryptic digests per protein, which specified peptides with at least one tryptic terminus for the SEQUEST analyses and two tryptic termini for the analyses of MASCOT and combined SEQUEST plus MASCOT data. Combined EBP analyses were conducted in which the results for all three replicates were analyzed together with protein probabilities calculated for the hypothesis that expression was evident in at least one of the three samples.

Statistical Model

Assumptions

Peptide Probabilities—Table I lists and describes the quantities used or estimated as part of the model. We begin with a set of MS/MS spectra to which possible peptide sequence identifications have been assigned using one or more search algorithms. Each peptide sequence identification has an associated probability ranging from 0 for random matches to 1 for peptide sequences that can be identified with absolute certainty. The first step in our procedure summarizes these probabilities. The combined probability \( p_i \) for peptide identification \( i \) of spectrum \( s \) is computed as the proportion of algorithms (e.g. MASCOT and SEQUEST) that identify peptide sequence \( i \) multiplied by the maximum probability identification. Repeated identifications of the same peptide sequence in multiple spectra of the same search are not treated as independent events but as repeated fragmentations of the same peptide resulting in spectra of varying quality. A conservative strategy is adopted by which only the highest probability identification is retained for each peptide. The resulting dataset comprises \( n \) peptide identifications, each of which is either correct with peptide probability \( p_i \) or is a random match with probability \( 1 - p_i \). We assume that the accuracy of each peptide identification is an independent random event. For ease of computation, the peptide probabilities may be bounded at some low threshold by ignoring, say, all \( p_i \leq 0.05 \).

Protein Matches—We seek to identify a set of "true" protein matches \( T \) that includes every protein that is truly expressed in the sample and whose sequence matches at least one correctly identified peptide sequence. This requires first identifying a set of "true plus homologues" protein matches \( H \) that includes every protein whose sequence matches at least one correctly identified peptide whether or not these proteins are actually expressed in the sample. That is, \( H \) includes all truly expressed proteins plus proteins that are not expressed but whose sequence matches correct peptide hits in

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_i )</td>
<td>Probability that peptide identification ( i ) from spectrum ( s ) is correct</td>
</tr>
<tr>
<td>( p )</td>
<td>Overall probability that peptide identification ( i ) is correct</td>
</tr>
<tr>
<td>( \Omega )</td>
<td>Set of all proteins in the search database</td>
</tr>
<tr>
<td>( H )</td>
<td>Set of true plus homologue protein matches</td>
</tr>
<tr>
<td>( T )</td>
<td>Set of true protein matches</td>
</tr>
<tr>
<td>( N )</td>
<td>Number of proteins in ( \Omega )</td>
</tr>
<tr>
<td>( D_j )</td>
<td>Set of theoretical digest products for protein ( j )</td>
</tr>
<tr>
<td>( y_i )</td>
<td>Effective length of protein ( j )</td>
</tr>
<tr>
<td>( z_j )</td>
<td>Set of peptide identifications whose sequence matches protein ( j )</td>
</tr>
<tr>
<td>( w_{ij} )</td>
<td>Number of peptides in ( z_j )</td>
</tr>
<tr>
<td>( v_j )</td>
<td>Weight of peptide identification ( i ) to protein ( j )</td>
</tr>
<tr>
<td>( a_j )</td>
<td>Total probability of peptide identifications matching protein ( j )</td>
</tr>
<tr>
<td>( \theta_j )</td>
<td>Abundance category for protein ( j )</td>
</tr>
<tr>
<td>( N_a )</td>
<td>Number of proteins of abundance ( a )</td>
</tr>
<tr>
<td>( n_a )</td>
<td>Total length of proteins of abundance ( a )</td>
</tr>
<tr>
<td>( T )</td>
<td>Set of theoretical digest products for protein ( j )</td>
</tr>
<tr>
<td>( H )</td>
<td>Set of true plus homologue protein matches</td>
</tr>
<tr>
<td>( N_a )</td>
<td>Number of proteins of abundance ( a )</td>
</tr>
<tr>
<td>( n_a )</td>
<td>Total length of proteins of abundance ( a )</td>
</tr>
<tr>
<td>( H )</td>
<td>Set of true protein matches</td>
</tr>
<tr>
<td>( a )</td>
<td>True if protein ( j ) is in ( H ), false otherwise</td>
</tr>
<tr>
<td>( a )</td>
<td>True if protein ( j ) is in ( T ), false otherwise</td>
</tr>
<tr>
<td>( a )</td>
<td>Probability that ( j ) is in ( H )</td>
</tr>
<tr>
<td>( a )</td>
<td>Probability that ( j ) is in ( T )</td>
</tr>
</tbody>
</table>

Significance Analysis for Protein Identification

Table I

Notation of quantities used in algorithm
Significance Analysis for Protein Identification

homologous proteins that are truly expressed. In symbolic terms, $T$ equals or is a subset of $H$, and both $T$ and $H$ are subsets of the set of all proteins in the search database, $\Omega$.

The search database $\Omega$ contains $N$ proteins, each of which may be expressed in our sample. We know that protein $j \in \Omega$ when subjected to enzymatic digestion theoretically results in a set of proteolytic peptides, $D_j$, which may include peptides with missed cleavages or non-enzymatic termini and those with specified modifications. A function of the size of this set, $y_j = \exp[\left\{\log|D_j|\right\}]$, is a useful measure of protein length: assuming a log-normal null distribution of search scores (8), the probability that the highest scoring random match is to one of the $|D|$ theoretical digests from protein $j$ is proportional to $\exp\left[\left\{\log|D_j|\right\}\right]$. Let $M_j$ be the set of $z_j$ peptide identifications whose sequence matches protein $j$, i.e. those peptide identifications that among the theoretical digest products of protein $j$: $M_j = \{p_i : i \in D_j; i = 1, \ldots, n\}$.

Degeneracy—Our treatment of degenerate peptide identifications, whose sequence matches more than one protein in the database, follows that of ProteinProphet (19). In some instances, a group of proteins exclusively share a set of peptide sequence matches. These proteins are compiled as a “degenerate protein group,” and a single expression probability is estimated for the group of proteins because any or all of them may be present in the sample, and they cannot be distinguished. Whenever a peptide matches more than one protein but is not part of such a degenerate cluster (in other words, it is part of an overlapping but not identical subset of matches), it is assumed that the protein matches are the result of either one true match, the remainder being matches due to homology with the expressed protein, or random matches due to error in matching the mass spectrum to the peptide sequence. The choice of which protein match is the correct one is treated as a multinomial event with the class probabilities given by a set of “weights” $w_{ij} = \sum_{i \neq j} w_{ij} = 1$, whose values depend on the relative expression probabilities of the proteins $j$ to which the peptide matches. The probability that peptide $i$ is a true match to protein $j$ is given by $w_{ij}p_j$, the probability of a homology match is given by $(1 - w_{ij})p_j$, and the probability of a random match is $1 - p_j$. The initial estimates for the weights are chosen to be equal for all proteins that match peptide $i$.

Protein Abundance—Highly abundant proteins are likely to accumulate many more peptide matches than proteins of low abundance. Protein abundance has been shown to be strongly correlated with the number of spectra whose peptide identifications match the protein sequence at least for MS/MS using non-data-dependent acquisition (13, 26). Let us call this quantity $v_j$, estimated by the total of weighted peptide identification probabilities for all spectra matching peptides $j$. Proteins are assigned to ordinal abundance categories $a_j$ by binning these $v_j$ at preset thresholds. Most proteins have no peptide matches and are assigned to the lowest abundance category.

Parameters—Protein expression probabilities are calculated conditionally on a set of parameters that summarize the data and govern the rate of accumulation of true and random peptide matches to proteins. For proteins in abundance class $a$, the model parameters are $\theta_a = (N_a, r_a, y_a, k_a, \lambda_a)$. $N_a$ is the number of proteins of abundance $a$, and $r_a$ is their total length so that the entire protein space $\Omega$ contains $N = \sum_a N_a$ proteins with a total length of $\sum_a r_a$. $n_a$ is the total number of peptide matches to proteins of abundance $a$. Note that because of degeneracy $\sum_a n_a \geq n$. $y_a$ is the proportion of proteins of abundance $a$ that are also in $H$, and $k_a$ is their total length. Thus, $H$ contains $\sum_a N_a y_a$ proteins with a total length of $\sum_a k_a$. Finally $\lambda_a$ is the number of these peptide identifications that are correct.

Probability of Membership of $H$—The condition $H = j \in H$ is true if at least one of the peptide identifications in $M_j$ is a correct match. We assume that correct peptide identifications are independently and randomly distributed among the proteins in $H$ at a rate that is proportional to the effective length of the protein so that the number of correct identifications matching a random protein $j \in H$ is Poisson distributed with parameter $\lambda_j y_a/k_a$. We also assume that the $n_a - \lambda_a$ incorrect peptide identifications are independently and randomly distributed among the proteins so that the number of incorrect identifications matching a random protein $j \in \Omega$ is Poisson distributed with parameter $(n_a - \lambda_a)y_a/k_a$. Using Bayes’ theorem, the estimated odds of membership of $H$ given the data equal the prior odds $y_a/(1 - y_a)$ multiplied by the likelihood ratio of membership of $H$ given the peptide matches $M_j$ and the expected parameter values $\theta_a$. (The “hat” notation (”) indicates an estimated value.)

Estimation of the Protein Expression Probabilities—The condition $j = M \in T$ requires that at least one of the peptide identifications matching $j$ is correct (i.e. $j \in H$) and that this is so because the digest product is truly expressed in the sample. The estimated weights $w_{ij}$ together with $M_j$ and $\theta_a$ are used to calculate the probabilities of membership of $T$ conditional on membership of $H$, and hence the probabilities of membership of $T$ — the true protein expression probabilities.

Estimation Using the “Expectation-Maximization” (EM) Algorithm

Maximum likelihood values for the parameters are calculated using the EM algorithm. The posterior probabilities of homology set membership $h_i = Pr(H_j | \theta, M)$ are initialized using $h_i = \frac{y_a}{\lambda_a}$ where $y_a$ is an initial parameter value. (The “posterior” probability means the probability conditional on the data, $M$.) At the “Expectation” step, we calculate the expected values of the parameters $\theta$ given the estimated probabilities $h_i$. For the “Maximization” step, we compute the probabilities of membership of $H$ and $T$ given the matching peptide identifications $M_j$ and the expected parameter values. The peptide weights are then updated according to a function that assigns the highest weight to the protein or proteins with the greatest probability of expression. When iterated, this procedure converges to a maximum function that downweights all but the most likely proteins, thereby restricting the results set to a minimal set of proteins necessary to explain the peptide identifications. The updated weights are used to update $v_j$ and hence estimate the abundance categories $a_j$. The effect of the Maximization step is to successively maximize the profile likelihoods $L(H | M, \theta, a_j)$, $L(T | \theta, M, \theta, a_j, \hat{\theta})$, $L(w | T, M, \hat{\theta})$, and $L(a | T, \hat{\theta})$ under the simplifying assumption that all $H_i$ and $T_j$ are conditionally independent given $M, \theta, a_j$, and $\hat{\theta}$. The effect of the Expectation step is to maximize the profile likelihood $L(\theta | \hat{H})$. The Expectation and Maximization steps are repeated until convergence at the maximum likelihood estimates of $\theta, H, T, a_j$, and $w$ given $M$.

Expectation Step—The Expectation step proceeds as follows to calculate the expected values of the parameters $\hat{\theta}_a$ for each abundance category $a$. $\hat{N}_a$ and $\hat{r}_a$ are calculated, respectively, as the number of proteins of abundance $a$ and the number of peptide identifications matching these proteins. $\hat{y}_a$ is the total length of these proteins. $\hat{y}_a$ is calculated as the expected value of an unweighted sum of Bernoulli variables $H_j$ divided by $\hat{N}_a$. $\hat{\lambda}_a$ is calculated as the expected sum of the Bernoulli variables $H_j$ weighted by the protein “lengths.” Finally $\hat{\lambda}_a$ is computed as the expected sum of Bernoulli variables $k_a$ totated over all the peptide matches to each protein.

Maximization Step—The first part of the Maximization step consists in the calculation of maximum likelihood values for the probabilities $h_i$ given the data and the expected values of the parameters. According to Bayes’ theorem, the posterior odds that $j \in H$ are given by the prior odds $y_a/(1 - y_a)$ multiplied by the Bayes’ factor. The Bayes’ factor equals the likelihood ratio of $h_i$ given the estimated parameters $\hat{\theta}_a$ and

530 Molecular & Cellular Proteomics 6.3

ASMB & CELLULAR PROTEOMICS

MCP

MOLECULAR & CELLULAR PROTEOMICS

6.3

Molecular & Cellular Proteomics
the peptide matches to protein \( j \), \( M_j \). (Henceforward the \( j \) suffix to the parameter estimates are omitted.) Algebraically the posterior odds that \( j \in H \) are given by Equation 1.

\[
\frac{\hat{h}_j}{1 - \hat{h}_j} = \frac{\Pr(H|\hat{\theta}, M_j)}{\Pr(\neg H|\hat{\theta}, M_j)} = \frac{\Pr(H|\hat{\theta})L(H|M_j, \hat{\theta})}{\Pr(\neg H|\hat{\theta})L(\neg H|M_j, \hat{\theta})} \quad \text{(Bayes’ theorem)}
\]

(Eq. 1)

(The symbol \( \neg \) indicates logical negation, equivalent to the Boolean operator NOT.) The likelihood that \( j \in H \) is the probability that all \( z_j \) peptide matches in \( M_j \) have arisen by chance,

\[
L(\neg H|M_j, \hat{\theta}) = \text{Poisson} \left( z_j \left| \lambda \right. \right) \prod_{M} (1 - \rho_i)
\]

where \( 0 < \rho_i \leq 1 \) and \( \text{Poisson}(n | \theta) = e^{-\theta} \theta^n / n! \). The likelihood that \( j \in H \) equals the sum of the probabilities that there are \( s \) correct peptide identifications and \( z_j - s \) random matches in \( M_j \), \( 0 \leq s \leq z_j \),

\[
L(H|M_j, \hat{\theta}) = \sum_{s=0}^{z_j} \text{Poisson} \left( s \left| \frac{\lambda}{s} \right. \right) \text{Poisson} \left( z_j - s \left| \frac{\lambda}{z_j - s} \right. \right) \cdot \sum_{s=\text{set } S \subset M_j} \prod_{s} \rho_i \prod_{M \setminus S} (1 - \rho_i)
\]

\[
= \sum_{s=0}^{z_j} c_{\gamma}(s) \text{Poisson} \left( s \left| \frac{\lambda}{s} \right. \right) \text{Poisson} \left( z_j - s \left| \frac{\lambda}{z_j - s} \right. \right) \quad \text{(Eq. 3)}
\]

where \( 0 < \rho_i \leq 1 \) and \( c_{\gamma}(s) = \Pi_{M_j} (1 - \rho_i). \sum_{s=\text{set } S \subset M_j} \Pi_{M_j \setminus S} \text{Poisson}(n | \theta) \), summing over all possible \( S \), subsets of \( M_j \), with \( s \) members. This likelihood is undefined if \( \rho_i = 0 \) for any \( i \), indicating that \( H_j \) is true with probability 1. Note also that there remains the possibility that \( H_j \) is true, if all the peptide matches are random \( (s = 0) \) or if there are no peptide matching at all \( (z_j = 0) \). This corresponds to the small but non-zero probability that \( j \) is a protein that is truly expressed in the sample but to which no identified peptide matches. Knowing the quantities from Equations 2 and 3 we can now estimate the probabilities \( h_j \) by applying Bayes’ theorem using Equation 1.

The next part of the Maximization step estimates the probabilities of true protein expression. The posterior probability of \( T_j \) conditional on \( H_j \) is the probability that correct peptide identifications truly match protein \( j \) rather than a homologous sequence in a different protein,

\[
\Pr(T_j|H_j, M_j, \hat{\theta}, \hat{\psi}) = \sum_{s=0}^{z_j} c_{\gamma}(s) \text{Poisson} \left( s \left| \frac{\lambda}{s} \right. \right) \text{Poisson} \left( z_j - s \left| \frac{\lambda}{z_j - s} \right. \right) \quad \text{(Eq. 4)}
\]

where \( 0 < \rho_i \leq 1 \) and \( c_{\gamma}(s) = \Pi_{M_j} (1 - \rho_i). \sum_{s=\text{set } S \subset M_j} (1 - \Pi_{S}) \Pi_{M_j \setminus S} \text{Poisson}(n | \theta) \), summing over all possible \( S \), subsets of \( M_j \), with \( s \) members. Estimates of the true protein expression probabilities can be derived using Equation 5.

Significance Analysis for Protein Identification

\[
\hat{\epsilon}_j = \frac{\Pr(T_j|M_j, \hat{\theta}, \hat{\psi})}{\Pr(\neg T_j|M_j, \hat{\theta}, \hat{\psi})} = \frac{\Pr(T_j|H_j, M_j, \hat{\theta}, \hat{\psi}) \cdot \Pr(H_j|M_j, \hat{\theta}, \hat{\psi})}{\Pr(\neg T_j|H_j, M_j, \hat{\theta}, \hat{\psi}) \cdot \Pr(H_j|\hat{\theta}, M_j)} \quad \text{(Eq. 5)}
\]

The estimated protein expression probabilities \( \hat{\epsilon}_j \) are used to update the peptide weights according to a function that assigns the highest weight to the protein or proteins with the greatest probability of expression. These updated weights are used to calculate updated estimates \( \hat{v}_j \) and update the abundance classifications \( \delta_j \).

As an optional refinement to the model, the peptide probabilities \( \rho_i \) can be adjusted for sequence homology. Assuming that degeneracy is independent of all the other parameters used in calculating the probabilities that the database search results are correct, the odds that peptide identification \( i \) is correct given that its sequence matches \( k_i \) proteins in the search database is given by Equation 6.

\[
\frac{\Pr(i|k_i)}{\Pr(\neg i|k_i)} = \frac{\Pr(i)}{\Pr(\neg i)} \cdot \frac{\Pr(k_i|i)}{\Pr(k_i|\neg i)} \quad \text{(Bayes’ theorem)}
\]

(Eq. 6)

This process can be incorporated into the EM loop, updating the peptide probabilities to their posterior values at each iteration until convergence.

Extension to Multiple Replicate Experiments—The model allows integrated analysis of data from \( R \) independent replicate experiments. Separate sets of parameters, probabilities of membership of \( H \), and probabilities of membership of \( T \) conditional on membership of \( H \) are estimated for each replicate. The overall protein expression probabilities are calculated using a combinatorial function of these probabilities: typically evidence of true expression is required from at least \( x \) replicates where \( 1 \geq x \geq R \). The peptide weights are calculated as done previously using these overall expression probabilities.

EBP Implementation of the Algorithm

A program was developed to implement the algorithm using a slight refinement of the statistical model motivated by empirical Bayesian concepts, giving rise to the name: the Empirical Bayes Protein identifier. This modification uses binomial theory to estimate maximum likelihood hyperparameters for \( \lambda \), assuming an underlying Gamma distribution; Equations 2, 3, and 4 are then calculated by integrating over \( \lambda \). This procedure helps stabilize estimates of the parameter set \( \hat{\theta} \), avoiding a local maximum in the likelihood at \( \gamma = 0 \). Computational adaptations were used to optimize the speed of the algorithm for proteins with many peptide matches.

The default settings exclude peptide identifications with probabilities less than 0.5 for the purposes of calculating protein identification. When analyzing the combined results of two search algorithms, this is equivalent to including only those spectra that are matched to the same peptide by both algorithms. It is advantageous to exclude low probability peptide matches as far as possible because the likelihood function is dominated by the number of peptide hits to each protein. The default settings also specify two abundance categories, with a high threshold of \( 10 \) defining a category of abundant proteins, for the analysis of complex proteomes. These default values were chosen to be prudent by removing error associated with low probability peptide identifications and conservative in defining high abundance proteins.

To facilitate data integration, public data repositories need an
Analyses of a Test Protein Mixture

Mass spectra from the sample dataset, derived from electrospray LC-MS/MS of a mixture of 18 non-human proteins, were searched using SEQUEST and MASCOT against a sequence database containing human peptides plus the 18 non-human proteins and likely contaminants (36). The search results were preprocessed using PeptideProphet (7) to generate two sets of peptide sequence identifications and probabilities. Protein identifications were derived by applying the EBP and ProteinProphet (19) algorithms to the PeptideProphet outputs using the default settings. For the purposes of this study, an alteration was made to the PeptideProphet output to improve the comparability of the two algorithms. Specifically the lists of possible protein matches were merged for peptide sequences with indistinguishable mass (i.e. those with identical sequence except for leucine-isoleucine substitutions). This procedure is performed by default in the EBP software but is not performed by ProteinProphet. Proteins whose estimated expression probabilities exceeded cutoffs of \( p \approx 0.9 \) and \( p > 0.7 \) were reported. Sensitivity was calculated as the proportion of the 18 sample proteins identified. The empirical error rate was calculated as the proportion of proteins identified that were neither in the sample mixture nor known contaminants. The estimated error rate was calculated as the difference between the average expression probability and 1 for proteins passing the threshold.

Estimation and Empirical Validation of the False Positive Rate for Analyses of Complex Protein Samples

Electrospray ionization MS/MS spectra from three independent zebrafish protein samples, separated by two-dimensional LC, were searched against a concatenated database of forward and reverse zebrafish protein sequences using both SEQUEST and MASCOT. The results were preprocessed (8) to generate two sets of peptide sequence identifications and probabilities for each sample. The error rate was validated using the assumption that reversed sequence identifications are all false positives and that false positive identifications occur with equal probability to forward and reversed sequences. Hence, if the model identifies \( f \) forward sequence and \( r \) reversed sequence proteins at a given probability, the total number of false positives is empirically estimated as \( 2r \), and the error rate of all identified proteins (i.e. both forward and reversed sequence proteins) is given by \( x = 2r/(r + f) \). The rate of false identifications to forward sequence proteins, an empirical estimate of the false positive error rate, is given by \( r f = x/(2 - f) \). Accordingly the estimated probabilities for forward sequence identifications were adjusted by the function \( x \in (0, 1) \): \( x \rightarrow 2x/(x + 1) \), the inverse of \( x \in (0, 1) \): \( x \rightarrow x/(2 - x) \), to obtain the true expression probability estimates. This adjustment is calculated automatically in the software implementation of EBP.

RESULTS

Proof of Principle — Proof of principle that the EBP algorithm can correctly identify the constituents in a protein mixture was established by applying the EBP algorithm to a pre-existing dataset used to validate the ProteinProphet algorithm (19, 36). The dataset was derived from electrospray LC-MS/MS analysis of a mixture of 18 non-human proteins titrated in varying concentrations to simulate the range of concentrations found in complex samples (36). Both algorithms identified largely the same proteins. The error rates in proteins identified by the EBP algorithm were conservative both at the level of individual and integrated analysis of search results (Table II). In contrast, ProteinProphet identified a somewhat greater proportion of the sample proteins than EBP but at the expense of considerably elevated levels of false positive error (Table II, columns a–c). ProteinProphet only gave accurate error estimates when the MASCOT and SEQUEST search results were combined using the method applied in EBP (Table II, column d). Using this method of data integration, both ProteinProphet and EBP identified 10 of the 18 proteins in the sample with high probability (\( p \geq 0.9 \)) and no errors.

Estimation and Empirical Validation of the False Positive Rate on Replicated Dataset — Database searches of the three zebrafish samples followed by PeptideProphet (8) analysis resulted in two sets of peptide identifications for each sample that were thresholded at a minimum probability of 0.05. MASCOT identified 103,052 singly, doubly, and triply charged spectra with probabilities >0.05 from all three samples. SEQUEST identified 102,821 such spectra. Thus, a total of 205,872 possible peptide sequence identifications were submitted as input to EBP both as separate analyses per sample and algorithm and in a combined analysis in which proteins were called truly expressed when putatively identified in at least one of the three samples. The empirical error rates were well within the estimated error rates in analyses of both combined (Fig. 1, a and b) and individual samples (Supplemental Fig. 1). Quantile-quantile plots attesting to the appropriateness of the correction for protein length used in the combined sample SEQUEST and MASCOT analyses are shown in Supplemental Fig. 2. In contrast, analyses run without length correction (i.e. assuming that the rate of incorrect peptide identifications is independent of protein length) gave too many random matches to long proteins.

Individual Versus Integrated Analysis — In our analyses, SEQUEST identified more non-homologous proteins than MASCOT in each individual sample and in the combined analysis of all samples as exemplified by proteins whose estimated expression probabilities were \( p \geq 0.9 \) (Fig. 2, a–d). This discrepancy is likely to be caused at least in part by the different probability models applied to MASCOT and SEQUEST data during preprocessing (7). The number of false positives is estimated by the number of reversed sequence identifications, given in parentheses. No such false positives were identified in the overlap between SEQUEST and MASCOT results in any individual sample. In contrast, protein identifications from SEQUEST and MASCOT that were not confirmed in the other analysis were error-prone. More proteins were identified in each individual sample and in the combined analysis of all samples by integrating SEQUEST and MASCOT results than by calculating the overlap between separate SEQUEST and MASCOT analyses.
The biological replicate analyses identified partially overlapping sets of proteins (Fig. 2f). Combining the samples in a single analysis was slightly more sensitive than taking the overlap among the analyses of individual samples. For the analyses integrating SEQUEST and MASCOT data, all but 26 of the proteins identified in any of the three individual samples were also identified by the combined analysis including one estimated false positive. A total of 43 proteins were identified by the combined analysis but not by the analyses of the individual samples with no estimated false positives.

The omnibus analysis of MASCOT and SEQUEST search results for all three replicates (Fig. 1b) identified 797 forward sequence proteins with an estimated probability ≥0.82 at an empirical error rate of 0.8% compared with an estimated error rate of 1%. For the 874 proteins with an estimated probability ≥0.22, the empirical error rate was 2.5% compared with an estimated error rate of just under 5%. An interactive list of all proteins identified with probability ≥0.1 together with the supporting peptide identifications is available upon request.

**Comparison with ProteinProphet**—The replicated zebrafish datasets were submitted to both EBP and ProteinProphet to
compare the protein sets identified by the two algorithms. Whereas EBP estimated error rates conservatively for both SEQUEST and MASCOT search results, ProteinProphet estimated rates of error accurately only for the MASCOT datasets (Table III, column a). ProteinProphet analysis of the SEQUEST search results identified excessive numbers of false positives (Table III, column b). ProteinProphet also identified too many false positives in the combined analysis of SEQUEST and MASCOT results; the protein list with an estimated error rate of 1% had an empirical error rate of 6%, and the protein list with an estimated error rate of 5% had an empirical error rate greater than 15% (Table III, column c). ProteinProphet achieved greatest sensitivity and accurate rates of error when analyzing SEQUEST and MASCOT search sets using the data integration method implemented in EBP (Table III, column d). In contrast to ProteinProphet, EBP did not identify excessive numbers of false positives for the SEQUEST and MASCOT search sets either separately or combination.

**TABLE III**

Rates of sensitivity and empirical rates of false positive error for ProteinProphet and EBP analyses of MASCOT (column a) and SEQUEST (column b) search results for the combined *D. rerio* datasets

Combined analysis of SEQUEST and MASCOT search results were performed by either using standard EBP/ProteinProphet methods (column c) or the EBP method for integrating multiple search datasets prior to ProteinProphet analysis (column d).

<table>
<thead>
<tr>
<th>Estimated error</th>
<th>(a) MASCOT</th>
<th>(b) SEQUEST</th>
<th>(c) SEQUEST + MASCOT</th>
<th>(d) SEQUEST + MASCOT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proteins</td>
<td>%</td>
<td>Proteins</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Empirical error</td>
<td></td>
<td>Empirical error</td>
</tr>
<tr>
<td>ProteinProphet</td>
<td>1</td>
<td>589</td>
<td>1.5</td>
<td>974</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>731</td>
<td>5.3</td>
<td>1195</td>
</tr>
<tr>
<td>EBP</td>
<td>1</td>
<td>481</td>
<td>0.8</td>
<td>761</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>522</td>
<td>1.3</td>
<td>870</td>
</tr>
</tbody>
</table>

**FIG. 2.** Overlap among non-homologous forward sequence proteins with expression probability \(\geq 0.9\). Numbers in parentheses indicate estimated numbers of false positives as indicated by reversed sequence identifications. For each diagram, homology is calculated with respect to all the relevant analyses so that the numbers of proteins may not correspond to the totals from the individual analyses. a, sample 1. b, sample 2. c, sample 3. d, combined samples. e, legend for a–d. f, overlap among non-homologous proteins from different samples, for both SEQUEST and MASCOT search data, with expression probability \(\geq 0.9\).
DISCUSSION

Shotgun proteomics experiments, which identify proteolytic peptides, need statistical methods to infer the presence of the parent proteins. Probabilistic methods of protein identification allow sensitivity of detection to be balanced against the rate of false positive error, an improvement over traditional filtering approaches. We provide a statistical method that can integrate the results from replicate experiments and multiple searches, implemented as an analytical module that augments an existing suite of open source proteomics software (42). We used this technique to identify nearly 800 non-homologous zebrafish proteins at an empirically validated false identification rate under 1%, a useful resource for selecting protein targets for quantification in high throughput genetic (28) and pharmacological screens (32).

We performed a self-validating analysis on these complex replicate samples by searching against a database containing reversed sequences to obtain both empirical and model-based estimates of false positive error. This validation method suffers the limitation that many of the characteristics of the forward sequences are retained in the distractors, potentially resulting in an excessively conservative analysis (37). For example, spectra for which a correct match does not exist in the original search space, but that match with high probability to reversed sequences, may cause spurious false positives. Care should be taken to ensure a sufficiently complete search with regard to the specified proteins and possible peptide modifications, mutations, and nonspecific cleavages. One of the assumptions of the forward and reverse sequence database search validation method is that the search space includes all the peptides that are actually present in the sample. Specifying an inadequate search space may result in the search finding matches among the reversed sequences that are actually correct identifications rather than false positives due to random error. This violates the statistical model and may invalidate the results of the analysis.

Analysis of a standard protein mixture suggested that our model may achieve more accurate error estimates than the ProteinProphet algorithm (19) at the level of a single replicate. Moreover ProteinProphet is not designed to analyze multiple datasets and on the evidence of our analyses tends to accumulate false positive protein identifications as more data are added. Indeed in the asymptotic case every protein in the dataset would be identified with probability 1 (20). In contrast, EBP explicitly models biological replicates allowing flexible hypothesis testing. The EBP approach to integrating multiple search results is applicable independently of the statistical model and was successfully applied to ProteinProphet.

Consistent with other studies (10–12), consensus scoring from multiple searches improved sensitivity and accuracy. Sensitivity was increased further by combining the results of biological replicates in a single analysis. Future investigations may adapt this statistical approach to datasets acquired using a variety of technical methods, for example in application to large multisite projects (38, 39), and to support methods of protein quantitation (40).

* This work was supported in part by National Institutes of Health Grants ROI CA95856, P50 HL70128, P50 HL81012, MO1 RR00040, HL 54500, HL 62250, and HL 70128; American Heart Association National Scientist Development Grant 043014HN (to T. G.), and the Cardiovascular Institute of Philadelphia (to T. G.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

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

The A. N. Richards Professor of Pharmacology

The Elmer Bobst Professor of Pharmacology.

To whom correspondence should be addressed: The Inst. for Translational Medicine and Therapeutics, University of Pennsylvania, 809 BRB II/III, 421 Curie Blvd., Philadelphia, PA 19104. Tel.: 215-573-7600; Fax: 215-573-9004; E-mail: tilo@spirit.gcrc.upenn.edu.

REFERENCES
