A Heuristic Method for Assigning a False-discovery Rate for Protein Identifications from Mascot Database Search Results*

D. Brent Weatherly‡§, James A. Atwood III¶§, Todd A. Minning‡, Cameron Cavola¶, Rick L. Tarleton‡, and Ron Orlando¶¶

MS/MS and database searching has emerged as a valuable technology for rapidly analyzing protein expression, localization, and post-translational modifications. The probability-based search engine Mascot has found widespread use as a tool to correlate tandem mass spectra with peptides in a sequence database. Although the Mascot scoring algorithm provides a probability-based model for peptide identification, the independent peptide scores do not correlate with the significance of the proteins to which they match. Herein, we describe a heuristic method for organizing proteins identified at a specified false-discovery rate using Mascot-matched peptides. We call this method PROVALT, and it uses peptide matches from a random database to calculate false-discovery rates for protein identifications and reduces a complex list of peptide matches to a nonredundant list of homologous protein groups. This method was evaluated using Mascot-identified peptides from a Trypanosoma cruzi epimastigote whole-cell lysate, which was separated by multidimensional LC and analyzed by MS/MS. PROVALT was then compared with the two traditional methods of protein identification when using Mascot, the single peptide score and cumulative protein score methods, and was shown to be superior to both in regards to the number of proteins identified and the inclusion of lower scoring nonrandom peptide matches. Molecular & Cellular Proteomics 4:762–772, 2005.

As a complement to gene expression profiling, proteomics is the analysis of gene and cellular function at the protein level. The definition of proteomics has expanded to include not simply the proteins encoded by a genome but the analysis of protein isoforms, post-translational modifications, and protein-protein interactions (1–4). Although two-dimensional gel electrophoresis and MS have been the dominant techniques used in this field, LC-MS/MS has emerged as a valuable technology for the analysis of complex protein mixtures (5).

Typical LC-MS/MS experiments are subdivided into five stages. First, the proteins are extracted from the biological material followed by enzymatic digestions to produce peptides. The peptides are partitioned through multiple separations followed by analysis with LC-MS/MS. Two methods can then be employed to correlate an MS/MS spectrum with the peptide and protein from which it originated. The MS/MS spectrum is either sequenced de novo, yielding a complete amino acid sequence or sequence tag (6–13), which is then searched against a protein database, or software algorithms are used to directly match experimental with theoretical MS/MS spectra generated from peptides in a protein database (14–21).

All software designed for matching MS/MS spectra to peptide sequences function in a similar manner. The experimental peptide masses are first compared with theoretical peptide masses derived from in silico enzymatic digestion of the specified protein database. The subsets of theoretical peptides with similar masses (within a user-defined mass tolerance) to the experimental peptides are fragmented in silico according to specific cleavage rules. These theoretical fragment ion masses are then compared with the masses of fragment ions from the experimental MS/MS spectra. Although all search engines will match a theoretical peptide sequence to an experimental MS/MS spectrum, every match is not always correct. Thus, all search algorithms attempt to assign scores indicating the degree of similarity between the experimental and theoretical MS/MS spectra.

Probability-based scoring algorithms attempt to accurately reflect the probability of a match being random by gathering information about the database itself and using this information in the score calculation. Of the probability-based search engines described in the literature, Mascot (14) is one of the most widely used. The Mascot probability model is based on the MOWSE (22) algorithm. The MOWSE algorithm functions by creating a matrix of weighting factors for the specified enzymatic cleavage, and sequence entries are clustered into cells formed by a distribution of intact protein and peptide fragment molecular masses. Therefore, each cell contains the frequency at which peptide molecular masses occur for a distribution of protein masses. The frequency factors are normalized and used to calculate the final score. Mascot reports a probability-based ion score for each peptide match, which...
indicates the statistical significance of that MS/MS spectral assignment. Following clustering of peptides to proteins, Mascot combines individual ion scores and reports a cumulative protein score. Although the Mascot algorithm (www.matrixscience.com/help/results_help.html) establishes a threshold ion score under which there is less than 95% confidence in any individual peptide match, the algorithm attributes no such statistical significance to the cumulative protein scores.

Published proteomics initiatives using Mascot have established a variety of criteria necessary to distinguish between correct and random protein assignments. A general approach has been to report a protein identified if at least one peptide from that protein is matched at or above the threshold ion score (23–25). We refer to this as the single peptide score method (SPM). 1 Another approach, the cumulative protein score method (CPM), has been to eliminate peptide matches below a given ion score and utilize an empirically derived protein score to identify proteins (26, 27). These strategies hinge on the ability of Mascot to accurately reflect true probabilities associated with matching MS/MS spectra to peptide sequences. With this in mind, evaluations of the Mascot scoring algorithm have demonstrated that the ion score thresholds established for a match to be considered significant accurately reflect a 95% confidence level when searching data from peptide mass fingerprints (28) and tandem mass spectra generated on high precision Q-TOF instruments (17). However, in large scale proteomics projects, when thousands of peptides are identified, this level of confidence may be unsatisfactory, resulting in several hundred false-positive peptide and/or protein identifications. To discriminate between correct and random peptide assignments from a Mascot search, several methods have been proposed. These include statistical models to evaluate peptide assignments using information gathered during the database search (29, 30) and filters that eliminate random matches based on the properties of the assigned peptides and experimental conditions (31–33). These methods, although allowing the removal of a large portion of the false-positive peptide identifications, either ignore potentially valid lower scoring peptide matches or do not account for the false-discovery rate (FDR) associated with the resulting protein identifications.

For most proteomics projects, the eventual goal is to establish the repertoire of expressed proteins in the sample of interest (34). As noted by Nesvizhskii et al. (35, 36), this is not a straightforward process when analyzing proteins at the peptide level. The statistical significance associated with matching MS/MS spectra to peptide sequences does not correlate with the likelihood that the consequent proteins are identified (35–37). This is ascribed to real peptide matches clustering to individual proteins, whereas random matches occur with an equal distribution across the entire database (35, 36). To calculate an overall probability or expectation value for a protein assignment, Nesvizhskii et al. (35, 36), Sadygov et al. (37), and Fenyo and Beavis (30) have proposed combining individual probabilities or expectation values for potential peptide matches. Such methods allow the specification of expectation values (30) or probabilities (35–37) to the proteins rather than peptide assignments and enable the inclusion of potentially valid low scoring peptide matches in the final protein assignment (35–37). However, these methods require the recalculation of expectation values or probabilities of peptide assignments made by the database search software and provide no information about the overall proportion of incorrect matches in a subset of identified proteins.

Herein we detail PROVALT, a tool that organizes large proteomic datasets and calculates protein FDRs (PRO-FDR) using peptides identified by Mascot. PROVALT extracts peptide matches from multiple Mascot results files, eliminates peptide redundancy, and clusters peptides to their corresponding proteins. In addition, homologous proteins for which no distinguishing peptide has been identified are organized into protein groups. Pertinent information such as sample origin and experimental conditions are linked to each peptide and protein match as well. We also report the implementation of a statistical model within PROVALT, an algorithm that determines PRO-FDRs when using Mascot. This model is based on the implementation of a random database (32, 38, 39) and Mascot peptide probability scores to calculate expected PRO-FDRs for a minimal number of expressed proteins identified by Mascot-matched peptides. As in prior reports, the random database served as the null hypothesis. Identifications resulting from the null hypothesis are considered to be random, and the scores assigned to the random matches should follow a quasi-normal distribution with a false-positive rate related to each score. PROVALT compares the score distributions obtained from searching the normal and random databases to calculate the FDRs associated with each score threshold. The goal is to set score thresholds that identify as many real proteins as possible while encountering a minimal number of false-positive protein identifications. It is important to note the difference between the false-positive rate and FDR as it applies to protein identification. The false-positive rate is the rate at which truly null protein identifications are treated as significant, whereas the FDR is the rate at which significant protein identifications are actually null (40–42). For example, Peng et al. (38) used a random database to identify 7,537 peptides at a FDR of 1%. Thus, 75 of the identified peptides were likely random matches.

Rather than calculate error rates at the peptide level, we demonstrate that the PRO-FDR calculations employed by PROVALT provide a reasonable balance between the number
PROVALT: Protein Validation Technology

| Table I | Peptide fractions and RP-LC |
|---------|---------------------------|
| Fraction number | 1st Dimension RP-LC percentage | 3rd Dimension RP-LC percentage |
| 1 | 5–10% | 2–12% |
| 2 | 10–15% | 6–17% |
| 3 | 15–20% | 11–22% |
| 4 | 20–23% | 16–25% |
| 5 | 23–26% | 19–28% |
| 6 | 26–29% | 22–31% |
| 7 | 29–32% | 25–34% |
| 8 | 32–35% | 28–37% |
| 9 | 35–40% | 31–42% |
| 10 | 40–45% | 36–47% |

a Range RP-B at which each fraction was collected during the first-dimension reverse phase chromatography.

b For the RP-LC-MS/MS analysis, each fraction was subjected to a 1-h linear gradient over the corresponding range of RP-B.

of correct and incorrect protein assignments. Here we evaluate this tool using a dataset of ~50,000 MS/MS spectra generated from 100 LC-MS/MS analyses of peptides from Trypanosoma cruzi epimastigote whole-cell lysates, which were separated by multidimensional LC. Using this dataset, we evaluate the two predominant modes of protein identification when using Mascot, the SPM and CPM. We then compare the PRO-FDRs resulting from these traditional methods with the PRO-FDRs generated from the PROVALT analysis.

EXPERIMENTAL PROCEDURES

Cell Culture and Peptide Preparation—T. cruzi (Brazil) epimastigotes were grown as described previously (43). Proteins were extracted from 5 × 10⁸ parasites using Tri-Reagent (Sigma) per the manufacturer’s instructions. Proteins were reduced (8 M urea, 200 mM Tris-HCl, 40 mM DTT, pH 8.5) for 1 h at 55 °C followed by carboxymethylation with iodoacetamide (80 mM) for 30 min at room temperature. The protein solution was diluted to 6 mM urea by the addition of H₂O and digested overnight at 37 °C with endoprotease Lys-C (1:100; Sigma). Following dilution to 1 M urea with 30 mM ammonium bicarbonate, proteins were digested overnight at 37 °C with sequencing-grade porcine trypsin (1:50; Promega, Madison, WI). The digest was lyophilized to dryness, reconstituted in 500 μl of 0.1% TFA, and filtered prior to separation.

First-dimension Reverse Phase Chromatography (RP)—First-dimension separation was performed on an Agilent 1100 series work station (Palo Alto, CA) configured with a 4.6 × 150-mm Jupiter C₁₈ column (Phenomenex, Torrance, CA). Buffer A was H₂O/0.1% TFA, and buffer B was ACN/0.1% TFA. The peptides were loaded onto the column (flow rate 0.75 ml/min), desalted for 10 min at 5% buffer B, and then sequentially eluted during a 40-min linear gradient from 5 to 45% buffer B. Ten fractions were collected (Table I), frozen, and lyophilized to dryness.

Second-dimension Strong Cation Exchange Chromatography (SCX)—SCX was performed on a Hewlett Packard (Palo Alto, CA) 1100 series workstation equipped with a 1.0 × 150-mm polysulfonethyl A column (5 μm, 300 Å; PolyLC, Columbia, MD). The buffer solutions were 5% ACN/0.5% acetic acid (pH 3.0, SCX-A) and 175 mM ammonium acetate/5% ACN/0.5% acetic acid (pH 4.0, SCX-B). Each RP fraction was reconstituted in 100 μl of SCX-A, loaded at a flow rate of 75 μl/min, and washed with 100% SCX-A for 10 min. Ten fractions were collected at 5-min intervals during a 50-min linear gradient from 10 to 60% SCX-B. The fractions were dried by vacuum centrifugation and stored at −20 °C until analysis by LC-MS/MS.

**TABLE I**

| Fraction number | 1st Dimension RP-LC percentage | 3rd Dimension RP-LC percentage |
|-----------------|-------------------------------|------------------------------|
| 1               | 5–10%                         | 2–12%                        |
| 2               | 10–15%                        | 6–17%                        |
| 3               | 15–20%                        | 11–22%                       |
| 4               | 20–23%                        | 16–25%                       |
| 5               | 23–26%                        | 19–28%                       |
| 6               | 26–29%                        | 22–31%                       |
| 7               | 29–32%                        | 25–34%                       |
| 8               | 32–35%                        | 28–37%                       |
| 9               | 35–40%                        | 31–42%                       |
| 10              | 40–45%                        | 36–47%                       |

**Protein Sequence Database**—Two sequence databases were constructed for our analyses. First, we created a representative database (normal) consisting of the ~25,000 T. cruzi gene annotations provided by the Trypanosoma cruzi Sequencing Consortium as well as possible contaminating proteins from Bos taurus, Equus caballus, Homo sapien, and proteases. A randomized database was constructed by reversing the sequences in the normal database. The random database was used to establish accurate scoring thresholds for protein identification in the normal database.

**Database Searching**—Mascot (version 1.8) searches were performed with the following parameters: a specified trypsin enzymatic cleavage with one possible missed cleavage, peptide tolerance of 0.2 Da, variable modifications caused by carbamylation (+43 Da), and carboxyamidomethylation (+57 Da).

**PROVALT, Protein Organization and FDR Calculations**—The PROVALT method for identifying proteins at a specified PRO-FDR is outlined in Fig. 1. Protein identification begins by extracting all Mascot-matched peptides and corresponding ion scores from the normal and random Mascot search results. The results from each dataset are combined and filtered yielding nonredundant lists of peptides. The peptides are grouped as functions of score thus forming score bins (B) containing all peptides equal to or exceeding each Mascot ion score (i), where i represents a specific ion score S ranging from M to N.

\[
M = \min(S) \leq PEP(S) \leq S
\]

\[
N = \min(S) \leq PEP(S) \leq 0
\]

\[
rPEP(S) = \text{no. of peptides in normal database} = S
\]

\[
rPEP(S) = \text{no. of peptides in random database} \geq S
\]

Peptides in each score bin are then clustered to their corresponding proteins. Proteins are selected based on the degree of peptide coverage c.

\[
c = \{C, C-1, \ldots , 1\}
\]

C = user-defined maximum degree of peptide coverage

Starting with C, a histogram is formed based on the frequency of protein identifications within each B for both the random and normal peptide matches (Fig. 2). Protein identification FDRs are then calculated for each score bin as follows.
Fig. 1. The PROVALT method for protein identification and statistical validation. Protein identification begins with the extraction of peptide matches from both normal and random Mascot database search results. Peptides are binned as a function of ion score and clustered to proteins to yield a minimal protein list. Proteins are selected based on degree of peptide coverage. The number of identified proteins between the random and normal database searches is compared as a function of peptide ion score to determine the minimal ion score and peptide coverage level necessary to achieve a 1% PRO-FDR. Matching proteins and peptides are output, and the process is iteratively repeated for each coverage level.

\[
\text{PRO-FDR}_c(S) = \left[ \frac{r\text{PRO}_c(S)}{n\text{PRO}_c(S)} \right] \times 100\%
\]

\(n\text{PRO}_c(S)\) = no. of proteins identified in normal with peptide coverage \(c\) in \(B\)

\(r\text{PRO}_c(S)\) = no. of proteins identified in random with peptide coverage \(c\) in \(B\)

PROVALT then determines threshold \(S_c\) (for \(c = C\)) \hspace{1cm} (Eq. 3)

\[
S_c = \min (S) \text{PRO-FDR}_c(S) \leq \text{Max PRO-FDR}, \quad \text{for } S = \{M, M+1, \ldots, N\}
\]

Max PRO-FDR = user-defined maximum PRO-FDR \hspace{1cm} (Eq. 4)

and thus the minimal Mascot ion score threshold necessary to achieve the specified PRO-FDR for the given degree of peptide coverage \(c\). Peptides meeting these criteria are stored, and the remaining peptides that were not matched to proteins are then grouped as a function of ion score \(S\), thus forming new score bins \(B\) containing all peptides equal to or exceeding each Mascot ion score \(i\). For the next degree of coverage \(c = C - 1\), the distribution of ion scores will have a minimum \(M\) that is dependent on score threshold \(S_c\) determined for the previous degree of peptide coverage as follows.

\[
M = S_{c+1} + 1 \quad \text{for } c < C
\]

\[
N = \min (S) r\text{PEP}(S) = 0
\]

\(r\text{PEP}(S)\) = no. of peptides in random database \(\geq S\) \hspace{1cm} (Eq. 5)

Previously unmatched peptides in each bin are again clustered to their corresponding proteins along with the peptides matched at the previous levels. For the next degree of peptide coverage \(C - 1\), another histogram is formed, and the PRO-FDR is then calculated for each score bin. \(S_c\) (for \(c = C - 1\)) is calculated, and the peptides meeting these criteria are stored. This process is repeated until \(c = 1\).

For this dataset, the ion score thresholds for each coverage level are displayed in Table II.
Our proteomic analysis of epimastigotes of *T. cruzi* generated ~50,000 spectra from 100 LC-MS/MS analyses of peptides from epimastigote whole-cell lysates. Whole-cell lysates similar to those used in this study may contain greater than 10,000 proteins expressed at a $10^6$ dynamic range of concentrations. Also, such complex samples are often comprised of large protein families and post-translationally modified proteins, which complicate protein identification because many peptides are identified multiple times and/or match to several homologous proteins. Although Mascot, which is one of the most widely used probability-based search engines, groups peptides to proteins and ranks protein identifications as a function of cumulative protein score, it does not combine results from multiple LC-MS/MS analyses. Our analysis of the epimastigote proteome generated 100 results files containing peptide identifications that needed to be combined to identify proteins, which made using Mascot in its native form impractical. Although a single file could have been generated by concatenating all 100 peak-list files, we chose to search each file individually using the parameters described above. The reasoning for this was 2-fold; not only would this amount of data require robust and expensive computational resources if searched as a single file, but it would also make identifying peptides present in different fractions impossible because the visibility to the source fraction for each spectrum would be lost. These problems were eliminated by the integration of a Mascot results parser and peptide clustering algorithm into PROVALT.

**Mascot Parsing Tools**—The integration of the Mascot parser allows the extraction of relevant peptide information from multiple Mascot results files. The parsing tool functions with both dat and html versions of Mascot results. The inclusion of an html parser allows users to access the nonlicensed version of Mascot (www.matrixscience.com), save multiple results files as html, and extract the peptide matches for later protein identification. The extracted information includes sample source, peptide sequence, ion score, precursor mass error, and fragment mass errors. The information is stored along with the peptide sequences and is later integrated with the protein identification. Following parsing, PROVALT eliminates redundant peptide identifications (but tracks the information associated with each valid occurrence), matches peptides to proteins, and then determines a list of unique protein groups. Although redundant peptides are removed and only the highest probability peptide match is used for the subsequent protein identification, all peptide matches are reported. Peptide matches are considered redundant if they have identical sequences plus any modifications or resulted from precursor ions of multiple charge states. If a peptide was identified with and without a modification, PROVALT does not consider these redundant. All matches to a given peptide regardless of precursor ion charge state are reduced to a single identification, and the highest ion score ($S$) is used in the PRO-FDR calculation. Although spectra from multiple precursor ion charge states may exhibit distinct fragmentation patterns, these peptides are not considered as independent peptide matches by PROVALT and are not allowed to contribute to the peptide coverage ($c$) (44). We have adopted this approach because of the fact that during database searching precursor ions of multiple charge states will be converted to their singly charged precursor masses prior to comparison with the *in silico* derived peptides. The subset of *in silico* derived peptides passing within the precursor ion mass tolerance will essentially be identical for the singly charged and multiply charged species. Thus identical assignments made from peptides with different precursor ion charge states are not unequivocally independent events. Conversely, peptide assignments with alternate modifications are matched through different parent mass filters and are treated as independent.

**Peptide Clustering Tools**—Correlating peptides to proteins

---

**FIG. 2.** Distribution of FDRs as a function of minimal ion score for proteins identified by six or more peptides. PROVALT constructs a histogram that compares the number of proteins identified in both the normal and random database searches. This comparison is iteratively performed for each peptide coverage level. The above histogram compares the frequency of protein identifications for proteins containing at least six or more peptides at or above specific Mascot ion scores. Using Eq. 3, the expected percentage of false discoveries is calculated for each peptide bin, and a 1% FDR is achieved if a protein is identified by six or more peptides with at least a minimal Mascot ion score of 11.

**TABLE II**

| Peptide coverage | PROVALT calculated ion score thresholds |
|------------------|----------------------------------------|
|                  | Minimum ion score                      |
| 1                | 44                                     |
| 2                | 35                                     |
| 3                | 27                                     |
| 4                | 18                                     |
| 5                | 14                                     |
| 6                | 11                                     |

*a Minimum number of peptides that match to a protein following PROVALT clustering.

*b Minimum Mascot reported ion score for the peptide assignment to achieve a 1% PRO-FDR.*
in a large scale proteomics project is not a straightforward process; Nesvizhskii and Aebersold (36) and Rappsilber and Mann (45) noted that eukaryotic protein databases contain many homologous proteins that are often indistinguishable by a single peptide identification. Therefore, if a group of peptides corresponds to more than one protein, it is impossible to determine which protein is actually expressed without performing additional experiments. In a manner similar to the organization described in Nesvizhskii and Aebersold (36), PROVALT accounts for the presence of protein families by determining protein groups that represent all peptides used to identify a group of homologous proteins. The protein groups were determined as follows: If protein “A” and protein “B” both contained the peptide sequences “x” and “y,” then the identification of peptides x and y would not allow proteins A and B to be distinguished from each other. Thus, protein group 1 would be defined as peptides x and y and would contain both proteins A and B. If protein “C” contained the peptide sequences x and y and “z,” then the identification of peptide z would allow protein C to be differentiated from both proteins A and B. Thus protein C would be placed in protein group 2. Following peptide clustering, a protein that contains all the peptides in the protein group is identified as PRO-FDR. For example, protein B would represent protein group 1 with a peptide coverage of c = 2, and the ion scores (S) of both peptides x and y would be used in the PRO-FDR calculation. Although proteins A and B may both be expressed, for statistical reasons we considered the protein group as a single protein identification. If the representative protein (B) is determined to be identified above the defined PRO-FDR, then the final report contains all of the homologous proteins (A and B), not just the representative protein. Our method assumes that if a protein is identified above a given error rate, the constituent peptides are also identified with the same confidence. Therefore, this method makes no statistical distinction between proteins in the same group that are identified by differing degrees of peptide coverage. Although it may be assumed that a protein is more likely to be expressed if it contains a larger number of matched peptides, it is indeterminate whether a subset of those peptides results from the presence of a homologous protein. As a result, the total number of protein groups identified represents the minimal number of expressed proteins.

**Single Peptide Score Method**—The most common approach for separating random from real protein identifications is the SPM, in which a protein is identified if it contains at least a single peptide at or above the Mascot-derived ion score threshold. Mascot defines a peptide match as significant if the probability of that event occurs by chance with a frequency of <5%. Thus a peptide with an ion score corresponding to at least an absolute probability ≤0.05 is considered a real match and used in the protein identification. To model this method, the ion score thresholds that corresponded to each absolute probability needed to be determined over the entire dataset.

This was accomplished by searching both the normal and random databases, forming a histogram of the frequency of peptide matches as a function of ion score, and then applying the equation below to calculate the distribution of peptide false-discovery rates (PEP-FDR).

$$PEP-FDR(S) = \left[\frac{rPEP(S)}{nPEP(S)}\right] \times 100\% \quad (\text{Eq. 6})$$

Fig. 3 shows the distribution of peptide matches that occurred at or above a range of ion scores for the normal and random database search results as well as the calculated PEP-FDRs. At an ion score of 27 or above, 5% of the peptides matching in the normal database would be random identifications. Typically, when the SPM is applied, only peptide matches with ion scores equal to or exceeding 27 would be used to identify proteins.

**Comparison of PROVALT and SPM**—To compare the PRO-FDRs generated from PROVALT with the SPM, peptide matches above discrete ion scores were extracted from the normal and random database search results and clustered to proteins using the PROVALT clustering tool. A histogram was formed based on the frequency of protein identifications by at least one peptide at or above ion scores in the range of 22–50. PRO-FDRs were then calculated according to Equation 3 and plotted along with the PEP-FDRs as a function ion score (Fig. 4). Fig. 4 demonstrates that an acceptable PEP-FDR does not necessarily correlate with an acceptable PRO-FDR, which is a serious drawback of using the PEP-FDR to identify proteins. In fact, the PRO-FDR grows much more rapidly than the PEP-FDR as the minimum ion score threshold decreases. For example, if peptides are selected for protein identification based on a PEP-FDR of 5%, the resulting PRO-FDR is 24%.
This occurs because the PEP-FDR only represents the rate at which a peptide is randomly matched during a database search and does not account for the fact that random peptide matches do not cluster to individual proteins at the same rate as real peptide matches. This trend is shown in Fig. 5 in which the ratio of peptides to proteins was plotted versus peptide score for both the normal and random database searches using the SPM. From Fig. 5 it is evident that the ratio of peptides to proteins is much larger in the normal database search than it is in the random database search, proving the improbability that multiple random peptide matches will occur on a single protein. Thus a set of randomly identified peptides will result in more protein identifications than an equal number of real peptide identifications.

To utilize the SPM with high confidence in protein identifications, one must be highly stringent when choosing a minimum ion score threshold. To achieve a PRO-FDR of 1% for this dataset, a minimum ion score of 42 is necessary (a PEP-FDR of 0.2%). This threshold was chosen, and the peptides meeting this criteria were clustered to proteins. These results, along with those of using PROVALT at a 1% PRO-FDR, are displayed in Fig. 6, A and B. Using the PROVALT method, 1,935 unique peptides were used to identify 444 unique proteins. To achieve the same PRO-FDR using the SPM, 1,064 unique peptides would match to 386 unique proteins. Although our method results in over 15% more proteins, the increase in the number of peptides is nearly double. Clearly, the SPM limits the number of peptides that can match to each protein by discarding all lower scoring peptides. However, this is not advantageous. As Fig. 3 indicates, nonrandom peptide matches exist at lower scores, but the use of the single score threshold does not facilitate discriminatory inclusion of these peptides.

**Cumulative Protein Score Method**—The CPM involves the extraction of all peptide matches above a given ion score and utilizes an empirically derived protein score (sum of peptide ion scores) to identify proteins. This method has two advantages. First, potentially nonrandom lower scoring peptides are used in the final protein identification. Second, individual peptide ion scores are summed to yield a nonprobabilistic cumulative protein score. However, to date there has been no statistical basis for differentiating between random and nonrandom peptide matches below the ion score threshold supplied by Mascot, and furthermore no statistical method for choosing a cumulative protein score has been published. In addition, for this method to be effectively applied the peptides must be correctly clustered to their corresponding proteins first; consequently for unconcatenated peak lists the Mascot-derived protein score is unusable.

**Comparison of PROVALT and CPM**—We evaluated the effectiveness of the CPM by identifying proteins in both the normal and random databases and calculating FDRs based on cumulative protein score thresholds. We first extracted all top ranking unique peptides with an ion score at or above 1 from both databases, clustered them to the corresponding proteins, and calculated a total protein score by summing the individual ion scores. Fig. 7 is the frequency distribution of proteins that match at a range of protein score thresholds (from 20 to 120) for the normal and random databases as well as the calculated PRO-FDRs. As stated above, determining a protein score threshold has previously been largely empirical, and for small datasets Mascot reports as “significant” hits all
proteins with a cumulative score above the ion score threshold. However, Fig. 7 indicates that employing this method for this dataset would result in a PRO-FDR of 97% at a score of 27 (ion score determined from Fig. 3). To achieve a PRO-FDR of 1%, a minimum protein score of 110 would be required.

Fig. 6, A and B, compares the number of proteins and peptides identified at a 1% PRO-FDR when using the CPM and PROVALT methods. Although a high confidence PRO-FDR for protein identifications is achieved using both, a vast number of nonrandom protein matches is discarded by the CPM (Fig. 7). On the other hand, if a lower protein score threshold is used, the PRO-FDR would be unacceptably high.

Because of the prevalence of lower scoring random peptide matches contributing to the cumulative protein score, the CPM has been employed following removal of these lower scoring peptides prior to protein identification (26, 27). Although this approach offers a good compromise, as one increases the peptide score threshold required for protein identification the method degenerates into a SPM where a large number of nonrandom peptide matches that are important for accurate protein identifications are removed.

Comparison of PROVALT and Modified CPM—Because the frequency of random peptide matches increases as the Mascot ion score decreases, the application of a low score cutoff to filter potentially random peptide matches has been proposed (26, 27). However, this method ignores higher scoring potentially random matches that will positively contribute to the final PRO-FDR calculation. For comparison of PROVALT with this version of the CPM, we applied several peptide ion score thresholds (1, 10, 20, 30, and 40) to filter peptides prior to protein identification with the CPM at a 1% PRO-FDR. Fig. 8 compares the distribution of peptides at various peptide ion score ranges using our method and the cumulative score methods to identify proteins. From Fig. 8 it is evident that our method is more discriminating in the inclusion of lower scoring peptides than the cumulative score method, which utilizes a higher number of lower scoring peptides for protein identification. In fact, when the cumulative score method is employed with peptide score thresholds <30, the predominant number of peptides is within the lowest ion score range. Thus, protein identification with this method is mostly facilitated by inclusion of the peptides that are most likely random matches. Using PROVALT, the largest number of peptides used for protein identification falls within the highest ion score range. The cumulative score method does not attribute any significance to the lower scoring peptides but rather assumes that
their presence with higher scoring peptides on the same proteins makes them significant. However, because so few high scoring peptides exist in the random database, lower scoring peptides that match to the normal database have an unfair statistical advantage. PROVALT utilizes only a stochastic peptide ion score threshold that is based on the assumption that random matches will occur with an even distribution across all proteins in the database, whereas real matches will cluster to the proteins that are actually expressed. Thus, score thresholds are calculated according to the probability that some number of peptides match above a given ion score to the same protein regardless of total protein score. This allows us to have high confidence in all of the peptides that are selected by our method and thus the proteins that they identify. Without high confidence in the lower scoring peptides, the cumulative score method provides essentially the same amount of quality information as the SPM.

**DISCUSSION**

Recent technological advancements in the field of proteomics have facilitated the analysis of complex protein mixtures resulting in a dramatic increase in the number of published proteomes. Many of these projects have utilized the Mascot algorithm to match tandem mass spectra to peptides in a protein database. Although its use has been widespread, it remains unclear how Mascot should be applied to large scale proteomics projects. To date, three approaches have been offered: 1) using Mascot in its native form and employing either the SPM or CPM approach, 2) calculating PRO-FDRs as is employed by PROVALT, or 3) calculating an overall probability for protein identifications using the probabilities of individual peptide matches.

The main goal in using external tools such as PROVALT is to separate the valid from incorrect protein identifications, the difficulty of which increases with the size of the dataset. As has been shown, the problem with using either the SPM or CPM is balancing the tradeoffs between choosing high stringency that discards potentially useful information and low stringency that does not discard the useless information. Thus it would be advantageous to determine a measure of confidence in protein identifications using all possible peptide contributions to that protein and select proteins based on a user-defined minimum confidence. Previous work has demonstrated the utility of the random database approach for calculating FDRs for peptide identifications (38). Such methods facilitate the calculation of the proportion of random peptide matches among all peptide matches deemed significant. However, the calculation of a PEP-FDR provides no information about the error rate of a specific protein identification. With the ultimate goal being the identification of proteins rather than peptides, a more thorough approach would be to define a PRO-FDR. With this in mind, we have demonstrated the use of PROVALT, a computational tool designed to work in conjunction with Mascot results files to provide a method for organizing large proteomic datasets and calculating FDRs for protein identifications.

PROVALT assigns a PRO-FDR based on the distribution of proteins with minimum peptide coverage \((c)\) identified in peptide ion score bins \((B_s)\), information that can be determined for any proteomics project. However, special consideration should be given to the dataset size when applying this analysis because the statistical significance of the calculated error rates is diminished for small datasets. As has been the case with the use of PEP-FDRs (38), the calculation of PRO-FDRs as performed by PROVALT is best suited for large datasets exceeding several hundred protein identifications. Otherwise, Mascot is best used in its native form due to the fact that manual verification of questionable identifications is feasible. The utility of PROVALT is as a high throughput tool for large datasets because it assigns a measure of statistical significance to all identified proteins, thus diminishing the need for manual verification.

To use PROVALT, the user must determine the maximum degree of peptide coverage \((C)\) to use in the calculation of FDRs. The concept of a FDR is based on the assumption that the null hypotheses (proteins identified in the random database) follow a quasi-normal distribution of minimal ion scores.
(S). Indeed, score distributions at increasing c may not necessarily look normal themselves, especially when the sample size is small. Thus, as the sample size (number of protein identifications) increases, for each c the score distribution will approach normality, and the FDR calculations will become more accurate. In practice, determining C prior to PROVALT analysis will be dependent on the distribution of minimal ion scores S in the normal and random protein assignments. Upon assignment of beginning c = C by the user, a version of Fig. 2 can be generated and visual inspection of the distribution will suggest whether the assumption of normality is reasonable. The user can increase or decrease the value of C until a statistically significant level of peptide coverage is determined for the normal database. It is sensible to assume that as the size of the dataset decreases, C will also decrease until the method degenerates into a SPM at C = 1. Accordingly, the native form of Mascot may be best suited for these situations.

Another approach, used in tools like Protein Prophet or PROT_PROBE, is to combine the conditional probabilities of individual peptide matches to calculate an overall probability for the protein identification (35–37). A protein identification is then considered significant if its individual probability exceeds a chosen threshold. The major difference between this method and that of PROVALT is that PROVALT calculates a PRO-FDR based on minimum ion score thresholds and peptide coverage but does not calculate individual probabilities for each protein identification. It is important to note that the FDR does not represent the probability that a feature is significant but rather the expected proportion of random matches resulting when proteins are identified using peptide score and coverage criteria. On the other hand, an individual protein probability does not represent the proportion of false discoveries that may result when a probability threshold is employed. Thus probability thresholds alone may not provide an assessment of overall significance for datasets that include a large number of protein identifications. Sadygov et al. (37) observed that protein probabilities always approach 1 when calculated through a binomial distribution of peptide probabilities that do not account for database or dataset size. This is crucial because as the size of the dataset increases the frequency of peptide matches to a given protein will increase regardless of the quality of the peptide match, resulting in a number of proteins being identified by a subset of random peptides. In contrast, the exploitation of a random database allows the calculation of FDRs regardless of database search parameters, database size, or dataset size. Another apparent benefit of this approach is that the entire distribution of random peptide matches is observed and factored into the final PRO-FDR calculation. As the number of random matches increases with dataset size, the significance thresholds necessary to achieve a specified PRO-FDR will also increase, thus compensating for the frequency of higher-scoring random matches. Ideally, however, one would prefer to have both protein probabilities and PRO-FDR estimations. This would allow researchers to select a set of proteins based on a minimum FDR but remove those that have an unacceptable individual protein probability, which is especially applicable for low peptide coverage members of high confidence protein groups.

In this work, the traditional methods of protein identification when using Mascot have been examined. Both the SPM and CPM strategies mentioned above propose the filtering of peptides in some manner prior to protein identification. We feel this is incorrect because such strategies ignore the phenomenon of peptide clustering and fail to assign any statistical significance to the thresholds used for the protein identifications. In addition, an accurate treatment of redundant protein and peptide identifications is crucial if a legitimate FDR is to be determined, and careful consideration must also be given to how peptides are partitioned among homologous proteins. The PROVALT algorithm, although designed to function in conjunction with Mascot, is independent of the scoring algorithm used for the peptide identification. This method reduces a complex list of peptide matches to a nonredundant list of proteins that have been identified at a user-specified FDR. Using peptide and protein identifications resulting from the proteomic analysis of T. cruzi epimastigotes, we compared the PROVALT method with the traditional methods of protein identification. The PROVALT method was shown to be superior to both methods by the number of proteins identified and in the inclusion of lower scoring nonrandom peptide matches.

All software used in this study to parse Mascot results files, cluster peptides to proteins, and select proteins for identification at a specified FDR is integrated into the software package PROVALT. This software will be made publicly available at kiwi.ccr.uga.edu/tpcprot.

Acknowledgments—We thank Kumar Kolli for technical assistance with the MS analysis. We also thank Lance Wells and Carl Bergmann for critical reading of this manuscript.

REFERENCES
1. Rep, M., Dekker, H. L., Vossen, J. H., de Boer, A. D., Houterman, P. M., Speijer, D., Back, J. W., de Koster, C. G., and Cornelissen, B. J. (2002) Mass spectrometric identification of isoforms of PR proteins in xylem sap of fungus-infected tomato. Plant Physiol. 130, 904–917
2. Bendt, A. K., Burkovski, A., Schaffer, S., Bott, M., Farwick, M., and Hermann, T. (2003) Towards a phosphoproteome map of Corynebacterium glutamicum. Proteomics 3, 1637–1646
3. Zhang, H., Li, X. J., Martin, D. B., and Aeboersold, R. (2003) Identification...
