Comprehensive Analysis of a Multidimensional Liquid Chromatography Mass Spectrometry Dataset Acquired on a Quadrupole Selecting, Quadrupole Collision Cell, Time-of-flight Mass Spectrometer

II. NEW DEVELOPMENTS IN PROTEIN PROSPECTOR ALLOW FOR RELIABLE AND COMPREHENSIVE AUTOMATIC ANALYSIS OF LARGE DATASETS

Robert J. Chalkley‡§, Peter R. Baker‡, Lan Huang‡¶, Kirk C. Hansen‡, Nadia P. Allen‖, Michael Rexach‖, and Alma L. Burlingame‡

A thorough analysis of the protein interaction partners of the yeast GTPase Gsp1p was carried out by a multidimensional chromatography strategy of strong cation exchange fractionation of peptides followed by reverse phase LC-ESI-MSMS using a QSTAR instrument. This dataset was then analyzed using the latest developmental version of Protein Prospector. The Prospector search results were also compared with results from the search engine "Mascot" using a new results comparison program within Prospector named "SearchCompare." The results from this study demonstrate that the high quality data produced on a quadrupole selecting, quadrupole collision cell, time-of-flight (QqTOF) geometry instrument allows for confident assignment of the vast majority of interpretable spectra by current search engines. Molecular & Cellular Proteomics 4:1194–1204, 2005.

Modern mass spectrometers are able to produce large amounts of information-rich data in relatively short periods of time. The bottleneck in mass spectrometry-based peptide and protein identification is now at the stage of data analysis and verification of results. There are several search engines available that can analyze large datasets in a batch fashion, most notably Mascot (www.matrixscience.com) and Sequest (1). Although it would be desirable to be able to quote results from such searches without a need to look at and evaluate the raw data, this is not without risk at the moment as although both use probability-based scoring systems, the reliability of results from Sequest are known to be problematic (2, 3), and no extensive study of the performance of Mascot on large datasets has been published. Hence a number of groups have developed statistical analysis programs for evaluating these search results to be able to better define the reliability of the reported matches (4–7). In addition, the data analyzed by these search engines are not the raw data but rather peak centroided mass lists extracted from the raw data that do not always fully represent the information content in the raw data. A summary of the complications arising from automated peptide and protein identification has recently been published (8).

Protein Prospector contains a suite of programs developed at University of California, San Francisco that is used for analysis of proteome data (www.prospector.ucsf.edu). Historically it has been one of the major programs in proteomic analysis; however, the current web version (version 4.0.5) does not have the ability to analyze multiple MSMS spectra simultaneously in a batch fashion. Thus, its current use in analyzing large datasets is limited. Hence we have developed new programs within the Prospector framework specifically designed for large dataset analysis and comparison. The first of these is “Batch Tag,” which is based on the well established MS-Tag program but is able to analyze files containing large numbers of spectra from one or multiple sample fractions.

A new program within Protein Prospector called “SearchCompare” has been developed that is able to summarize and filter large dataset results. It also converts the peptide scores from Batch Tag into a new discriminant score. The scoring system used by Batch Tag simply gives a certain score for every ion type matched with the weighting of the scoring based on the occurrence of a particular ion type (e.g. 3 points for every “y” ion, 0.25 points for every internal ion, . . . ). These weightings are separately defined for different instrument types. SearchCompare then uses multiple parameters about the Batch Tag results to produce a new discriminant probability-based scoring system.

SearchCompare can also combine, filter, and compare mul-
multiple search results and is able to perform quantitation analysis of differentially isotopically labeled samples. It can produce three different types of report: all peptides/proteins identified by any search (union), all peptide/proteins identified by every search (intersection), or peptides/proteins only identified in a particular search (difference). An added feature of this program is its ability to compare results from both Prospector and Mascot searches. If a peptide is identified by both search engines there is a much higher probability of it being a real match due to the same result being returned using different algorithms. Also there will generally be a few correct matches that are found by one search engine but not by the other, and SearchCompare can identify these difference matches, which provide a set of spectra worth examining manually by the researcher.

The dataset analyzed and presented here to evaluate these new Protein Prospector features is part of an ongoing study of protein trafficking into and out of the nucleus by analyzing cargo proteins binding to members of the nuclear pore complex (9–11). There are classes of proteins that specifically transport proteins into the nucleus (importins) and out of the nucleus (exportins). The interaction of importins and exportins with their cargo proteins is controlled by the small GTPase Gsp1p (known as Ran in vertebrates). In its GTP-bound state it promotes dissociation of importin-cargo complexes, whereas in its GDP-bound state it dissociates exportin-protein complexes. Therefore, by establishment of a GTP/GDP gradient between the nucleus and cytoplasm Gsp1p is able to regulate nucleocytoplasmic transport (12–14). In this particular experiment we sought to understand the changes in protein interactions at the nuclear pore as the yeast progresses through the cell cycle by arresting cells at cell cycle checkpoints and comparing proteins interacting with Gsp1p-GTP using the cleavable ICAT technology (15).

These data were acquired during the development of our techniques for quantitation of low level samples using the cleavable ICAT technology (15). In our strategy, in addition to analyzing the ICAT-labeled peptides for quantitative information, we also analyzed the non-labeled peptides to better characterize the sample and provide corollary peptide identifications to the one or two that are typically matched to a given protein from the ICAT-labeled peptides. Unfortunately only a few ICAT-labeled peptides were detected in the ICAT fraction, arising from abundant proteins in the sample. This was probably due to the very low levels of sample (ICAT is normally performed with orders of magnitude more sample) combined with unexpectedly high ICAT-labeled peptide loss possibly due to peptides not being efficiently eluted from the biotin column. Nevertheless a large amount of data was acquired on the unmodified peptides through which a comprehensive characterization of binding proteins was achieved. It is the data from these unlabeled peptides that are presented in this database report, which we use to assess the performance of the new Prospector analysis software and compare its performance to that of a leading commercially available search engine, Mascot.

EXPERIMENTAL PROCEDURES

His-tagged Gsp1p was expressed and purified from Escherichia coli as published previously (9). Yeast cells were arrested at the G1 stage of the cell cycle using 2.5 µg/ml a-factor exposure for 3 h or at M phase using 20 µg/ml nocodazole for 3 h, and then interacting proteins were isolated as published previously (10). Proteins from both cell states (about 5–10 µg/cell state) were labeled with the cleavable ICAT reagent (Applied Biosystems, Foster City, CA) to give a total of 10–15 µg of protein, and this was analyzed essentially as in our published protocol for ICAT of low level samples (15). Briefly proteins were denatured in 9 m urea and reduced with trichloroethyphosphine, and then cysteines of G1 phase-arrested proteins were alkylated with light ICAT reagent, while M phase proteins were alkylated with isotopically heavy reagent. After tryptic digestion peptides were separated by strong cation exchange using a Beckman Gold HPLC system equipped with an analytical flow upgrade. Separation was achieved using a 2.1 × 10-mm polysulfoethyl A column (PolyLC) where Buffer A was 30% ACN, 0.05% formic acid and Buffer B was buffer A containing 400 mM NH4Cl. Six fractions were collected, and each of these was successively passed through the biotin affinity cartridge (part of Applied Biosystems ICAT kit). Each flow-through was collected separately, and then all ICAT peptides were eluted into one fraction using 30%ACN, 0.4% TFA. ICAT tags were cleaved in 95% TFA.

Each fraction was cleaned up using Zip Tips to desalt the samples and then analyzed by reverse phase LC-MSMS. Reverse phase chromatography was performed using an Ultimate HPLC system and a Famos autosampler (both LC-Packings). Separation was using a 75-µM × 150-mm Pepmap column (LC-Packings) at a flow rate of 300 nl/min. Buffer A was 0.1% formic acid, while Buffer B was ACN, 0.1% formic acid. The gradient separation was 5–40% B over 105 min. As peptides eluted off the column they were introduced on line into an ESI-QqTOF® instrument (QSTAR, MDS Sciex/Applied Biosystems) and were analyzed using data-dependent switching between MS to MSMS modes: after a 1-s MS spectrum up to three multiply charged precursor ions could be selected for 2-s MSMS spectral acquisitions. After a given precursor was selected, dynamic exclusion was used for the next 60 s to prevent its subsequent reselection. Peak lists of MSMS spectra from the six LC-MS runs were created using the Mascot script within Analyst that “smoothed” the data by merging data points in the MS spectra within 0.02 Da of each other prior to centroiding, and data points within 0.05 Da of each other in the MSMS spectra were merged prior to centroiding. The peak lists from all six fractions were searched together with either Protein Prospector or Mascot (version 2.0). For searches on Prospector the mass range from the lowest m/z recorded to the highest observed m/z peak was split into two, and the 20 most intense peaks in each half of the spectrum were used for searching. Mascot uses the raw peak list and performs threshold filtering of the peak list during searching in an undocumented fashion. Searches were carried out allowing for 150 ppm mass accuracy for the parent ion and 300 ppm mass accuracy for fragment ions. Oxidation of methionine, protein N-terminal acetylation, and pyroglutamate formation when the N-terminal amino acid is a glutamine residue were all allowed as variable modifications. Results from each search were saved, and these were then analyzed and compared using SearchCompare.

The abbreviation used is: QqTOF, quadrupole selecting, quadrupole collision cell, time-of-flight.
**QqTOF LCMS Dataset: II. New Protein Prospector**

**RESULTS**

During analysis of the six cation exchange fractions of the Gsp1p-binding proteins isolated from yeast a total of 3269 MSMS spectra were acquired. These spectra were initially searched using a new program in Protein Prospector called Batch Tag against the Swiss-Prot database (April 3, 2004), allowing only yeast proteins plus one or two expected non-yeast proteins (GST and human keratins). Batch Tag is based on MS-Tag, but it can take as its input text files from LC-MSMS analyses in several different formats, including the Mascot Generic Format (.mgf). Peak lists from some spectra can contain large numbers of peaks due to a lack of noise filtering. Hence Protein Prospector filters the MSMS peak lists prior to searching. It deisotopes the spectrum and assigns a charge state to singly and doubly charged fragment ions. Based on the fractional mass of ions Protein Prospector is also able to exclude certain peaks as being derived from chemical noise rather than a peptide component. A disproportionate amount of the noise in QSTAR spectra is in the lower mass region of the spectrum due to the majority of the chemical noise being singly charged. Peaks at higher m/z, which are often significantly weaker than those at low m/z, are generally more informative. Hence Protein Prospector splits each spectrum in half and uses the 20 most intense peaks in each half of the spectrum to search with 40 peaks per spectrum. This produces higher scoring and more reliable protein scores. Indeed it may be beneficial to split the spectrum into more sections or assign different numbers of peaks to each section; however, this was not investigated in this study. Prospector uses a scoring system of different values for different matching ion types. These values are based on the frequency of observation and the specificity of different ion types in tryptic peptides acquired on our QSTAR instrument and are presented in Table I.

**Table I**

<table>
<thead>
<tr>
<th>Ion type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>y ion</td>
<td>3</td>
</tr>
<tr>
<td>y loss iona</td>
<td>1.5</td>
</tr>
<tr>
<td>b ion</td>
<td>1.5</td>
</tr>
<tr>
<td>b loss iona</td>
<td>0.5</td>
</tr>
<tr>
<td>b + H2O</td>
<td>1.0</td>
</tr>
<tr>
<td>a ion</td>
<td>0.5</td>
</tr>
<tr>
<td>a loss iona</td>
<td>0</td>
</tr>
<tr>
<td>Internal ion</td>
<td>0.25</td>
</tr>
<tr>
<td>Immonium ion</td>
<td>0.5</td>
</tr>
<tr>
<td>M-SOCH4b</td>
<td>5.0</td>
</tr>
<tr>
<td>Unmatched peak</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

*a* Loss of either H2O or NH3.

*b* Side chain loss from parent ion of oxidized methionine-containing peptide.

This search came back with ~2000 top matching peptides of relatively high confidence primarily from proteins that were expected/know to be in the sample. These spectra were all briefly inspected and generally contained extensive y and sometimes “b” ion series. We then manually analyzed all unmatched spectra and all spectra that gave low confidence matches to ascertain the reason why these spectra had not been matched to proteins. This gave us a list of manually curated assignments for all spectra. Full details of this analysis are presented in the accompanying study (16). After manual curation of the dataset we were able to produce a list of “correct” predicted tryptic peptides to 2368 of the 3269 spectra. Comparing our list of answers to the Protein Prospector Batch Tag search results of Swiss-Prot yeast and known contaminant proteins there were 2214 correct assignments made by Batch Tag if one allowed for Leu/Ile substitutions; *i.e.* it correctly identified 93% of the spectra we assigned as tryptic peptides. In low energy CID spectra such as those acquired on a QqTOF instrument there is no way to differentiate between leucine and isoleucine; they could only be differentiated between by high energy fragmentation “d” or “w” ions (17). Therefore, a peptide scoring system for low energy CID data has to score peptides with leucine/isoleucine substitutions the same. It is also not always possible to differentiate between lysine and glutamine or phenylalanine and oxidized methionine as these have very similar masses.

This dataset was then searched against the whole Swiss-Prot database (April 3, 2004) and also against the whole National Center for Biotechnology Information (NCBI) (March 29, 2004) database. The whole Swiss-Prot search returned 2118 correct answers, whereas the NCBI search returned 2045 correct top answers. The different numbers of matches in these three searches reflects the number of proteins being searched in the database; there were 4925 yeast proteins of a total of 141,381 protein entries in Swiss-Prot, whereas the NCBI database contained 2,715,099 entries. Thus, despite the fact that 53 of the spectra in this dataset correspond to peptides from proteins that were not in the Swiss-Prot database but were in the NCBI database, many fewer correct answers are reported in the NCBI search due to the presence of an order of magnitude more database entries, so the highest scores of false positives increases.

**SearchCompare**—The fact that a given peptide is the top scoring match to a spectrum obviously should not mean that the match is thought to be correct. A peptide identification as part of an analysis of a complex mixture is not an isolated event, and if other peptides have been identified from the same protein this assignment is more likely to be correct. Therefore, we sought to introduce this notion as a factor in our scoring. We decided to use the highest score for a peptide...
from a particular protein as a parameter in creating a more reliable scoring system. For example, if a spectrum returns a match with a score of 20, and this is the only spectrum matching to this protein, then 20 would be used as the highest scoring match to this protein, whereas if another spectrum in the dataset matched a different peptide in the same protein with a score of 40, then 40 would be used as a parameter for creating a new score for the spectrum that itself had only scored 20. Cursory analysis of the search results also suggested that a high difference in score between the top match and a random match was a more reliable parameter for determining a correct match than the absolute score because spectra with large numbers of peaks in general scored more than spectra with fewer peaks even if most of these peaks were assigned to predicted fragments from the matched peptide. We chose to use the sixth best match to a given spectrum to calculate the difference score because the second and third matches to a spectrum often shared significant homology to the top match and as such could not be considered random matches. It should be noted that as we cannot distinguish between leucine and isoleucine, peptides whose only difference was Leu/Ile substitutions were saved with the same rank, so in some cases more than six sequence matches were saved from a given spectrum.

So we sought to combine the best peptide score for a protein with the difference score for the particular match in relation to the sixth match to create a new score that is more discriminatory between correct and incorrect answers than the Batch Tag score. We input the five highest scoring matches for each spectrum from a search of the whole Swiss-Prot database (a total of over 16,000 results) into the statistical package SPSS (www.spss.com) and indicated which matches were correct according to our manual assignments. SPSS then calculated the optimal weighting of the two parameters to maximize the ability of the score to differentiate between correct and “incorrect” answers. SPSS returned the following formula as optimal weighting of these two parameters for differentiating between correct and incorrect answers: Discriminant Score = -2.852 + (0.105 × best peptide score) + (0.11 × score difference). This suggests that the two parameters are of similar importance for discriminating between correct and incorrect answers.

Fig. 1 shows a histogram of the correlation between the discriminant scoring system and the curated results at predicting whether the result is correct or incorrect. This shows that using a confidence probability of 0.5 as the distinguishing threshold for the 16,909 results, there were 273 that were incorrectly predicted as correct (1.6%) and 474 results that were wrongly reported as incorrect (2.8%).

The effect of the discriminant score on the matches to a spectrum is exemplified in Table II, which shows the results for the spectrum acquired at 17.4 min in fraction 1. In this case, the best peptide score for the protein is VEPESYPK with a score of 22.1, and the second best match was D(L/I)(L/I)SMEK with a score of 17.9. The discriminant score for this spectrum is calculated as follows:

Discriminant Score = -2.852 + (0.105 × 22.1) + (0.11 × (22.1 - 17.9)) = -2.852 + 2.325 + 0.414 = 0.981

This score is above the threshold of 0.5, so the spectrum is predicted to be correct. The table below shows the results for the top five matches to the spectrum:

**TABLE II**

Peptide scores and discriminant scores for the top five matches to the MSMS spectrum at 17.4 min in the first cation exchange fraction

<table>
<thead>
<tr>
<th>Accession no.</th>
<th>m/z</th>
<th>z</th>
<th>Peptide match</th>
<th>Correct sequence</th>
<th>Rank</th>
<th>Sequence correct?</th>
<th>Run time</th>
<th>Protein score</th>
<th>Best peptide score</th>
<th>Peptide score</th>
<th>Score difference</th>
<th>Discriminant score</th>
</tr>
</thead>
<tbody>
<tr>
<td>P40069</td>
<td>474.7352</td>
<td>2</td>
<td>VEPESYPK</td>
<td>VEPESYPK</td>
<td>1</td>
<td>True</td>
<td>1-17.4</td>
<td>1477.5</td>
<td>22.1</td>
<td>8.3</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>Q9Y705</td>
<td>474.7352</td>
<td>2</td>
<td>D(L/I)(L/I)SMEK</td>
<td>VEPESYPK</td>
<td>2</td>
<td>False</td>
<td>1-17.4</td>
<td>37.5</td>
<td>19.6</td>
<td>17.9</td>
<td>4.2</td>
<td>-0.34</td>
</tr>
<tr>
<td>O32619</td>
<td>474.7352</td>
<td>2</td>
<td>VEENSKDK</td>
<td>VEPESYPK</td>
<td>3</td>
<td>False</td>
<td>1-17.4</td>
<td>17.8</td>
<td>17.8</td>
<td>3.1</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>P36845</td>
<td>474.7352</td>
<td>2</td>
<td>(L/I)DQDKDSK</td>
<td>VEPESYPK</td>
<td>4</td>
<td>False</td>
<td>1-17.4</td>
<td>14.9</td>
<td>14.9</td>
<td>1.1</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>Q99W57</td>
<td>474.7352</td>
<td>2</td>
<td>(L/I)DDKMAQK</td>
<td>VEPESYPK</td>
<td>5</td>
<td>False</td>
<td>1-17.4</td>
<td>14.2</td>
<td>14.2</td>
<td>0.4</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>Q9SS44</td>
<td>474.7352</td>
<td>2</td>
<td>VEEDQKSSR</td>
<td>VEPESYPK</td>
<td>6</td>
<td>False</td>
<td>1-17.4</td>
<td>13.8</td>
<td>13.8</td>
<td>1.6</td>
<td>-1.32</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Predictive performance of the discriminant scoring system. Predicted correct results by the discriminant scoring system (assuming confidence probability >0.5 indicates correct; <0.5 indicates incorrect) from the top five matches to every spectrum are compared with our curated results for each spectrum. A value of 0 indicates agreement between the discriminant scoring prediction for being correct/incorrect and our manual assignment. A value of 1 indicates spectra that we manually assigned as correct but the discriminant scoring believed were incorrect. A value of -1 indicates spectra that the discriminant scoring system believed to be correct but manual assignment assigned as incorrect.
example the correct peptide identification was the top ranked match prior to the discriminant scoring, but the score was not dramatically better than the false matches (22.1 versus 17.9). However, by combining the difference score (score difference = 22.1 – 13.8 = 8.3) with the best peptide score for the same protein (another spectrum matching to the same protein as the top match had a score of 45) the new discriminant score reports the top match as being a dramatically more confident match than the other matches (2.79 versus –0.34).

Fig. 2A illustrates that the discriminant score differentiates much more effectively between correct and incorrect answers than the Batch Tag peptide score; a plot of peptide scores forms a continuous distribution where incorrect and correct answers merge, whereas the discriminant score plot shows two maxima: one corresponding to correct results and one corresponding to incorrect results. If you compare the discriminant score distribution to distributions purely on the basis of the highest peptide score for a protein or the difference score (the two components that make up the discriminant score) it shows that neither of these initial parameters is as reliable as the combination of the two parameters, which provides a scoring system from which a reliable threshold for correct and incorrect matches can be inferred (see Fig. 2B).

When the list of best discriminant scoring peptides for each spectrum after searching against Swiss-Prot was compared with our curated peptide assignment list there were 2215 correct peptide matches of a possible 2368 compared with 2118 that were correctly matched by peptide score alone. Thus, the discriminant scoring correctly matched 93.5% of the spectra assigned as identifiable tryptic peptides. The results of this search are presented in Supplemental Table 1.

score of 0.54) Protein Prospector predicted 2311 matches to be correct of which 144 were false positives and 48 were false negatives (4.4 and 1.5%, respectively, of all spectra).

When we searched this dataset against NCBI 2204 spectra of the 3268 were correctly assigned compared with 2045 by peptide score alone. 2205 spectra were reported as being correct at >0.5 probability threshold (discriminant score of >0.82) with 114 false positives and 113 false negatives.

Fig. 3 shows a plot of peptide score against discriminant score. This shows that there is a general correlation between peptide and discriminant score (as one would expect), but the correlation is looser as one moves toward lower peptide score. Also it shows that the distribution of peptide and dis-
criminant scores does not change significantly between doubly, triply, and quadruply charged precursor ions.

Looking at the false positive peptide results with the discriminant scoring there is a common reason for incorrect matches. For example, for the spectrum acquired at time 43.31 min in cation exchange fraction 6 SearchCompare reports RINMIEELEK as the top discriminant scoring peptide, with a discriminant score of 3.32, which it therefore believes with high confidence is correct. We assigned the actual correct answer as LVNHFIQEFK, which is the top protein scoring match in Batch Tag prior to the discriminant scoring analysis. The peptide matches for this spectrum to the two different peptides are shown in Fig. 4, which shows that the match to the correct peptide is better because it explains more of the peaks observed and all the immonium ions. The reason why the other peptide is being predicted as correct is due to the "highest scoring peptide from a given protein" parameter. Both peptides are from proteins that are abundant in the sample. The correct peptide, from heat shock protein SSA, has a highest scoring peptide of 50.6, whereas the highest discriminant scoring match is to a peptide from lysyl-tRNA synthetase, which has the highest scoring peptide match in the whole dataset, scoring 57.8. Thus, there is a trend in the false positives to matching extra peptides from proteins already identified in the sample. This is an effect that is also seen in Mascot but should not be as pronounced in Prospector because only the top five hits are stored compared with 10 in Mascot. This problem does not generally create false positives at the protein level.

One question that may be raised is whether this new scoring performs better at analyzing precursor ions of a given charge state as this is commonly observed for analysis of ion trap data. Of the 3269 spectra acquired, 43 were singly charged (the algorithm that determines charge state “on-the-fly” sometimes makes errors and accidentally selects singly charged precursors for fragmentation), 2418 spectra were of
doubly charged precursors, 743 were of triply charged precursors, and 65 were of quadruply charged precursor ions. The breakdown of results at each charge state (Table III) shows that there is not a significant difference between the reliability of results for doubly and triply charged precursor ions, and there are not enough results for quadruply charged precursor ions to draw any statistical significance. The similar reliability of results for different charge states is probably helped by the fact that Prospector is able to determine fragment ion charge states from the spectra, so it can assign multiply charged fragment ions with more certainty.

A common approach to estimate the false positive level of a database search is to search the same dataset against a reversed database (2, 18). The assumption is made that all matches are false positives, so by using the same scoring thresholds used for reporting results from the forward database a measure of the false positive level can be determined. Searching this dataset against a reversed version of the same Swiss-Prot database returned 133 matches above the 0.5 confidence level and 45 above the 0.95 confidence threshold, corresponding to 4.1 and 1.4% false positive levels at the two threshold levels. A quick glance through the false positives showed that, for many of the incorrectly assigned results, the peptide assigned was highly homologous or even identical to the correct answer. The highest scoring match against the reversed database was from the spectrum at 18.65 min in cation exchange fraction 5, which matched the sequence LAEVYKAEK. The correct sequence match for this spectrum is LAEVYKAEK. The reversed database result matched y1, y2, y5–y8, y7–y8, y7–y8, y8–y8, and b3–b4, which are all identical in the correct answer (the match to the correct answer also included y4 and b2). The 66th highest scoring peptide match (the median match of the 133 false positive matches) in the reversed database search was to a spectrum at 25.03 min in the sixth cation exchange fraction and returned the sequence (pyroE)-IQELRKO from the reversed sequence of mouse Intersectin 2, whereas the correct match in the forward database is the identical peptide from lysyl-tRNA synthetase.

If you combine the forward database and reversed database and search this dataset then the number of peptides reported from the reversed database is dramatically different from that found by searching only the reversed database: 21 above the 0.5 confidence level and nine above the 0.95 confidence level. This disparity demonstrates the effect of having correct answers in the database on the false positive rate and suggests a reversed (or randomized) database does not produce an accurate estimate of the false positive level using the discriminant scoring of Protein Prospector.

*Results at the Protein Level*—For most researchers the important information is to ascertain how the search engine performs at the protein level rather than at the peptide level because it is the protein results that are used for interpretation of the biological significance (although for quantitation analysis the reliability of the peptide identifications is more important).

Protein Prospector is not using a protein scoring *per se*. However, the peptide discriminant score is using information about better matching peptides to the same protein when these exist, so this score should also function fairly well at the protein level simply by filtering protein matches to have a peptide match above a certain probability. The choice of the appropriate minimum probability threshold is to some extent dependent on the user, whether the user is prepared to trade losing a few correct identifications in return for increasing the reliability of the results they report. Table IV presents the performance of the discriminant scoring at two different confidence thresholds: proteins containing a peptide match at higher than 0.5 and 0.95 confidence probabilities. Also Table IV contains a comparison with the performance of Mascot on this same dataset when searching the same Swiss-Prot database. The Mascot search was performed on an in-house version of Mascot, and two parameters were used that are not used on the web version of the software. First a minimum peptide score of 12 was required for a match to be reported. Second the filtering parameter “RedBoldOnly” was set to 1.

### Table III

**Distribution and reliability of peptide assignments at different charge states**

<table>
<thead>
<tr>
<th>Charge (z)</th>
<th>Number of spectra</th>
<th>Predicted correct (≥0.5)</th>
<th>Correct</th>
<th>False positives (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2418</td>
<td>1775</td>
<td>1674</td>
<td>5.7</td>
</tr>
<tr>
<td>3</td>
<td>743</td>
<td>518</td>
<td>478</td>
<td>7.7</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>18</td>
<td>15</td>
<td>16.6</td>
</tr>
</tbody>
</table>

### Table IV

**Protein Prospector and Mascot analyses of this dataset**

<table>
<thead>
<tr>
<th></th>
<th>Swiss-Prot</th>
<th>NCBI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mascot</td>
<td>PP 0.5</td>
</tr>
<tr>
<td>Proteins reported</td>
<td>229</td>
<td>256</td>
</tr>
<tr>
<td>Number correct</td>
<td>209</td>
<td>232</td>
</tr>
<tr>
<td>Percent correct</td>
<td>91.3</td>
<td>90.6</td>
</tr>
<tr>
<td>Homologous reported</td>
<td>57</td>
<td>48</td>
</tr>
<tr>
<td>Number homologous real</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Total correct proteins identified</td>
<td>209</td>
<td>247</td>
</tr>
</tbody>
</table>

ND, not determined; PP, Protein Prospector.
Activating this parameter means that only proteins that contain at least one peptide that is red and bold (i.e. it is the top match to the spectrum, and it is the first time this match has been reported in the search results) are returned. The implementation of a minimum peptide score was a requirement for reliable results using previous versions of Mascot (versions 1.9 and older). However, Mascot version 2.0 uses a new scoring system for large datasets that makes using this minimum score threshold less important in terms of getting reliable results. However, in our experience, the implementation of the requirement for proteins to contain a red bold peptide match is still a very important filter for increasing the reliability of results. Mascot uses a probability threshold of 0.95 for reporting protein results. For this dataset Mascot reported a score threshold of greater than 36 as indicating “identity or extensive homology.”

Table IV shows that 91% of the 229 proteins reported by Mascot are correct. This slight discrepancy between the predicted performance of Mascot and that reported can be explained by the confusion/difficulty in defining a correct answer. Eight of the 20 proteins that we have defined as incorrect for Mascot are proteins that are homologous to proteins that are present in the sample (i.e. some of the peptide matches to this protein are correct, but the unique peptide matches to this protein are incorrect). Although these matches are not correct, they do share “extensive homology” to correct matches. Adding these matches leads to the Mascot scoring performing as its scoring model predicts (94.8% correct).

Protein Prospector searching against the Swiss-Prot database and using a 0.5 peptide match probability threshold for reporting proteins returns 256 protein matches plus a further 57 homologous proteins. Protein Prospector tries to separate homologous proteins out of the main protein list, so proteins that contain peptides from a protein already reported, but at least one unique peptide, are listed separately at the end. The list of homologous proteins reported includes some protein matches that are clearly independent matches. Of the 57 reported homologous proteins, 15 of them are independent protein matches. Nearly all of the incorrect homologous matches were reported on the basis of one unique peptide match, whereas 10 of the 15 real proteins had multiple unique peptides identified. If you discount the homologous protein matches, Protein Prospector correctly identifies roughly 10% more proteins (232 versus 209) but with slightly worse reliability than Mascot if you discount the false homologous matches of Mascot. Using a 0.95 peptide confidence threshold Protein Prospector correctly matches slightly more proteins than Mascot with less than half the number of false positives. If the homologous protein list was filtered in Protein Prospector to report the proteins with more than one unique peptide in the main protein list then at both thresholds Protein Prospector significantly outperforms Mascot in terms of proteins correctly identified and still has a low level of false positives.

The results for searching against the NCBI database demonstrate that the discriminant scoring again does an effective job at making correct protein identifications but that slightly fewer protein assignments are made in comparison to searching against Swiss-Prot. It also shows that a very large number of homologous proteins are reported. The large amount of redundancy in protein entries in NCBI presents problems in identifying real homologous protein matches. We did not attempt to determine how many of these homologous protein assignments we believe are real, independent protein identifications.

It is recognized within the field of proteomics that protein identifications on the basis of one peptide match are less reliable (19). In the results from this dataset there are 80 protein identifications reported above the 0.5 confidence level by Protein Prospector on the basis of a single peptide match. Of the 22 incorrectly reported protein identifications from Protein Prospector every single assignment is on the basis of a single peptide match. Thus, we believe that removing “one-hit wonders” from these results, which is an approach used by some to improve reliability of data (20, 21), would, in this case, actually create a completely correct set of protein results. However, it would also lose 58 correct protein identifications, 25% of all the correct answers.

Using the same 0.5 and 0.95 confidence thresholds when searching against the reversed database would report 128 proteins and 42 proteins, respectively, alarmingly high values. However, searching the combined forward and reversed databases reports 18 proteins and 6 proteins from the reversed protein database at the two threshold levels. Also of the 128 proteins reported from the reversed database, 120 are one-hit wonders, and 36 of 42 are above the 0.95 confidence level. Thus, combining these numbers with the demonstrated flaw in the reversed database at calculating the false positive rate of Protein Prospector scoring, this confirms that removing single peptide protein assignments would produce exceptionally reliable results.

Looking at the false positives in the main protein list, partial justifications for many of these false positives can be given. For example at the 0.95 threshold in Protein Prospector there are five incorrectly identified proteins. Three of the five were matched by spectra for which the correct sequence was not in Swiss-Prot; for one the correct answer is a peptide formed by a nonspecific enzyme cleavage, and for the other the software had incorrectly labeled the second isotope as the monoisotopic mass (and Protein Prospector had matched a homologous peptide to the correct match). This result demonstrates that the false positive matches are generally not where an incorrect answer gets a better score; rather the correct answer is not an option.

SearchCompare has a flexible output format where one can choose which columns one desires in the results (see Fig. 5). The results can be reported as a web page or alternatively it can be saved as a tab-delimited file that allows easy import.
into spreadsheets or databases. It should also be highlighted in passing that SearchCompare is able to perform quantitative analysis of isotopic labeling experiments (22). The output of the results also plots the distribution of the discriminant scores allowing one to see how well the scoring discriminates between correct and incorrect answers. It is from these distributions that SearchCompare calculates the probability that an answer is correct. Optimal performance of this discriminant scoring system is reliant on a large number of data points such that it can accurately model the distributions of correct and incorrect answers. Fig. 6 shows the discriminant score distribution for a very different dataset for comparison. This dataset was acquired as part of a cleavable ICAT quantitation analysis of proteins in the urine of one of a set of patients with Dent’s disease (23). This dataset, also acquired on a QSTAR instrument, contained 2528 MSMS spectra. In this dataset a discriminant score of 0.03 corresponded to a 0.5 confidence for peptide assignment, and using this threshold 312 peptides from 66 proteins were assigned. The discriminant scoring effectively separated the correct from the incorrect matches. However, there are many fewer correct results probably because urine contains large numbers of proteolytic peptides that do not have tryptic specificity as well as containing many non-peptide species. The distribution of the incorrect answers can be reliably modeled, but the distribution for the correct answers is more difficult to model such that the probabilities for correct answers may be less accurate. In this type of situation it may be more reliable to quote a probability of an answer being incorrect. This situation will also be encountered when analyzing smaller datasets (less than 500 spectra). Visual inspection of the histogram of discriminant scores can give a reasonable estimate of the reliability of a given score.

**DISCUSSION**

Presented here is new software within Protein Prospector that allows the analysis of large datasets. It can analyze multiple LC-MS runs in one analysis and produces reliable summaries of the results. Its performance has been compared with the current benchmark search engine for LC-MSMS analysis, Mascot, and the results show that even using a very stringent threshold for reporting the protein results (protein
must contain a peptide at 0.95 confidence). Protein Prospector is able to identify more proteins than Mascot and at a lower false positive level. Using a less stringent threshold (confidence threshold of 0.5) Prospector matches significantly more proteins at a comparable reliability.

To assess the performance of a search engine it is necessary to create a dataset with which to test its performance. One approach has been to create samples consisting of mixtures of 10–20 proteins and then assume any match to one of these proteins is correct and all others are incorrect (24). However, this is often not completely representative of the type of sample that is being analyzed in multidimensional chromatography experiments where hundreds of proteins may be present in a sample. Hence the approach we took here was to take a dataset from an ongoing project within the laboratory that consisted of a complex mixture of over 200 proteins and then manually analyze all the data to determine correct answers for each spectrum before comparing these results with those returned by search engines.

While performing this analysis we became increasingly aware of the difficulty of defining a correct and incorrect result. Using low energy CID it is impossible to distinguish between leucine and isoleucine based on fragment ions. Hence peptide results where these residues are interchanged had to all be accepted as correct. Glutamine and lysine also nominally have the same mass, and at the mass accuracy that these data was searched the search engine is not going to be able to distinguish between these residues, although if one manually interpreted the data by looking at the mass difference between fragment ions, then mass accuracy of QqTOF data is sufficient to distinguish between the two residues. We chose not to accept interchange of these residues both as correct answers. This was not a major issue because this dataset is of tryptic peptides, so there were very few lysine residues other than at the C terminus of peptides. Also our interpretation of spectra is to some extent subjective. In this dataset there are a number of spectra that are very weak, and we categorized them as being too weak to be able to yield a confident answer. However, for some of these spectra the search engines gave answers that indicated a reasonable confidence in the results. Some of these may be correct, but because we were not convinced the spectra were left in a “too weak to yield an answer” category.

These results add further fuel to the discussion of the value and reliability of single peptide protein identifications. In the protein results reported above our 0.5 confidence level, we believe nearly 10% of the reported proteins are false positives. However, all of these are one-hit wonders. By removing the one-hit wonders the results become (infinitely) more reliable. However, doing this also removes a quarter of the correct assignments. Indeed this percentage of protein assignments on the basis of single peptide hits is significantly lower than some recently published datasets where 40–70% of protein identifications were through single peptide identifications (25). This presents a conundrum for researchers as to whether to pay attention to these single peptide protein assignments; we believe in our dataset 58 of 80 (73%) are correct answers.

Although the performance of the new Protein Prospector scoring is clearly impressive, there are still obvious ways of improving the discriminatory ability of the scoring. The Batch Tag scoring values of 3 for a y ion etc. are “ballpark” figures determined for the importance of different ion types. By statistically analyzing the ions observed in this and other large datasets acquired on a given instrument it should easily be possible to fine tune the initial scoring that is the basis for the discriminant score.

The weighting for different ion types will obviously be instrument type-dependent. We have a similar dataset of the same sample analyzed in this study that was acquired on a MALDI-TOF-TOF instrument (4700 Proteomics Analyzer, Applied Biosystems). Initial results show that by using a different set of weighting values for different ion types the discriminant scoring performs equally well on this dataset.

The majority of the large datasets published thus far from multidimensional chromatography experiments of complex mixtures have been acquired on ion trap instruments. The very high percentage of spectra correctly assigned by Prospector in this study (over two-thirds) is in contrast to most previously published dataset of high throughput ion trap data where between 5 and 15% of the acquired spectra could be interpreted (2, 3, 26), although one study has reported 40% identification (25). This is unlikely to be a reflection of the reliability of results searched with different search engines but rather a measure of the relative quality of data acquired on a QqTOF instrument in comparison to that acquired on an ion trap both in terms of mass accuracy and the presence of a full mass range in the fragmentations spectra. Also the selection of only multiply charged precursor ions for fragmentation drastically reduces the number of non-peptide species selected for fragmentation. Lastly yeast has a well annotated genome with relatively few post-translational modifications compared with mammalian samples. Ion traps typically acquire many more spectra in a multidimensional analysis than are acquired on a QSTAR instrument. However, the number of interpretable spectra acquired from the two approaches may be comparable; the larger number of spectra does not necessarily produce significantly more information. We think this is very important information for the proteomics community at large due to the rapid growth of proteomics and the widespread use of ion traps for data acquisitions added to the fact that there are many people new to the field, and there are publications that report de facto that 70–90% of spectra in proteomic experiments have no match in the database (27). We hope this database publication will exemplify that mass spectrometers can produce high quality data from which high fidelity matches can be made from the majority of the data.

We have presented here a new set of software tools that
allow analysis of large scale LC-MSMS analyses. Its performance has been shown to be comparable to if not better than the current market leader for searching non-ion trap data, Mascot. In the near future we intend to make these new software tools available to the research community.

Acknowledgments—We acknowledge the assistance of Brian Williamson, Steve Martin, and other members of the Applied Biosystems Proteomics Research Centre in the acquisition of these data.

* This work was supported by National Institutes of Health National Center for Research Resources Grants RR01614 and RR15804 and NHLBI Grant HL074005-03 and by the Vincent J. Coates Foundation. 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.

REFERENCES