Journal of Biomolecular Techniques
17:327–332 © 2006 ABRF

Verification of Single-Peptide Protein Identifications by the Application of Complementary Database Search Algorithms

James G. Rohrbough,[1,2] Linda Breci,[3] Nirav Merchant,[4] Susan Miller,[4] and Paul A. Haynes[1,5,6,7]

[1]Department of Biochemistry and Molecular Biophysics, The University of Arizona, Tucson, AZ; [2]Air Force Institute of Technology, Civilian Institutions Programs, Wright-Patterson Air Force Base, OH; [3]Department of Chemistry, The University of Arizona, Tucson, AZ; [4]Arizona Research Laboratories, Biotechnology Computing Facility, The University of Arizona, Tucson, AZ; [5]Bio5 Institute for Collaborative Bioresearch, The University of Arizona, Tucson, AZ; [6]Australian Proteome Analysis Facility, Macquarie University, North Ryde, NSW, Australia; [7]Department of Chemistry and Biomolecular Sciences, Macquarie University, North Ryde, NSW, Australia

Data produced from the MudPIT analysis of yeast (*S. cerevisiae*) and rice (*O. sativa*) were used to develop a technique to validate single-peptide protein identifications using complementary database search algorithms. This results in a considerable reduction of overall false-positive rates for protein identifications; the overall false discovery rates in yeast are reduced from near 25% to less than 1%, and the false discovery rate of yeast single-peptide protein identifications becomes negligible. This technique can be employed by laboratories utilizing a SEQUEST-based proteomic analysis platform, incorporating the XTandem algorithm as a complementary tool for verification of single-peptide protein identifications. We have achieved this using open-source software, including several data-manipulation software tools developed in our laboratory, which are freely available to download.

Key Words: Database search algorithms, protein identification, SEQUEST criteria, tandem mass spectrometry, XTandem, single-peptide verification.

Protein identification from complex biological mixtures often involves the application of tandem mass spectrometry techniques[1,2] such as MudPIT,[3,4] which involves digestion of the protein mixture with a protease such as trypsin, followed by two stages of liquid chromatography separation using strong cation exchange (SCX) and reversed-phase (RP) separation. Peptides eluting after these separations are subjected to ionization and fragmentation in the mass spectrometer. Database search algorithms are then used to match the acquired spectra to peptide sequences from a protein database. Examples of such programs include SEQUEST,[1,5] Mascot,[6] Spectrum Mill,[7] ProteinLynx,[8] XTandem,[9–11] and OMSSA.[12] When a protein is identified from several unique peptide spectra, the inherent redundancy of identification improves the confidence in protein identification, even if the confidence of some of the peptide identifications is low. As the number of peptides assigned to each protein sequence decreases, the confidence of protein identification drops correspondingly.

There are many examples in current literature of proteomic analyses performed by application of the MudPIT technique.[13–17] However, there is no consensus on the search parameters used for the database search algorithms, or the treatment of proteins identified from single peptides. It is not correct to simply disregard single-peptide matches; such peptides may be the only detectable peptide from an enzymatic digest, and therefore perfectly valid for identification purposes. It is equally incorrect to include all proteins identified from single peptides, because of the variability in protein identification from poor mass spectra, resulting in a high rate of false-positive identifications.[18–21]

There have been numerous attempts to validate protein identifications from current database search

Address correspondence and reprint requests to: Associate Professor Paul A. Haynes, Department of Chemistry and Biomolecular Sciences, Macquarie University, North Ryde, NSW 2109, Australia (fax: 61-2-9850 6200; email: paul.haynes@chem.mq.edu.au).

algorithms, including: linear discriminate analysis used to determine the accuracy of search algorithm assignments;[22] the Qscore algorithm using a probabilistic scoring system and analysis of false-positive identification rates using a reversed database;[23] the heuristic approach to assigning false discovery rates;[24] the normalization of peptide identification scoring systems based on the length of the peptide;[25] utilization of the tryptic status of peptides as an additional level of validation;[3,25–27] the application of a support vector machine (SVM) to distinguish between correct and incorrect peptide identifications by SEQUEST;[28] and the inclusion of orthogonal parameters such as exact mass measurements of selected peptides.[29] One published report describes a proteomic analysis in which the final results were in the form of a consensus between the output from two different search algorithms.[30] However, neither this report, nor any of those mentioned above, specifically addresses the issue of improving the confidence rate of assignment for proteins identified from a single peptide. Several authors, however, have noted that consensus analysis of dual algorithm searching programs has considerable merit in terms of protein identification confidence levels.[7,31]

Our aim in this study was to develop a basic set of software tools that would enable us to achieve 95% or greater confidence of assignment for both single- and multiple-peptide-based protein identifications, using only freely available, open-source software in addition to our existing SEQUEST analysis platform. As a consequence, all software tools developed and used in this project are made freely available via our laboratory Web site.

The data used in the development and testing of this approach were acquired from triplicate MudPIT analyses of yeast (*Saccharomyces cerevisiae*) mixed organelle lysate sample (designated Y1, Y2, and Y3), prepared and analyzed as described,[13] and rice (*Oryza sativa*) leaf, root, and seed organ lysate samples (designated R1seed, R2root and R3leaf), prepared[32] and analyzed[13] as described.

The entire set of tandem mass spectra collected from all 13 chromatographic steps in each experiment were searched using TurboSEQUEST (BioWorks version 3.1, Thermo Electron)[1,5] run on a 16-processor IBM Beowulf cluster; with dta files generated from peptide spectra meeting the following criteria: Peptide MW Range = 400–3500 Da; Threshold = 1000; Precursor Mass = 1.40; Group Scan = 1; Minimum Group Count = 1; and Minimum Ion Count = 35.

All SEQUEST searches were performed with no enzyme specificity indicated. The search parameters used were default settings except for: peptide mass tolerance = 1.5; max number of modified amino acids per differential modification in a peptide = 4; static modification mass of +57.0 for acetylated cysteine; differential residue modification mass of +16.0 for oxidized methionine; a maximum of two internal cleavage sites; one allowed error in matching auto-detected peaks, and a mass tolerance of 1.0 for matching auto-detected peaks. SEQUEST search results were filtered using DTA-select v 1.9[33] using our laboratory default cutoff parameters: Xcorr for a 1+ ion = 1.8, Xcorr for a 2+ ion = 2.5, Xcorr for a 3+ ion = 3.5, deltaXcorr = 0.1.[13,34–36]

The single-peptide matches from SEQUEST were re-searched against the same database by XTandem version 2005.10.01.5 (open source software, available from http://www.proteome.ca/opensource.html).[9–11] The default XTandem search parameters were used, except for the following: a maximum valid expectation value of 0.02; residue mass modification of +57.022 for carbamidomethylated cysteine; potential residue mass modification of +16.0 for oxidized methionine; enzyme specificity = none specified; spectrum parameters including a fragment monoisotopic mass error of 0.5 Da and a parent monoisotopic mass error of ± 2.5 Da; spectrum conditioning parameters of 100 .0 spectrum dynamic range, total spectrum peaks 50, a minimum parent M+H of 400.0, and a minimum fragment *m/z* of 150.0.

Tandem MS spectra from rice organ samples were searched against a database of rice (*Oryza sativa japonica*) protein sequences (36,318 sequences—April 2005 version), representing the complete rice genome, from NCBI (www.ncbi.nlm.nih.gov). The yeast samples were searched against a yeast genome protein sequence database (6882 sequences, March 2005) from the *Saccharomyces* Genome Database (www.yeastgenome.org). Both the rice and yeast databases were supplemented with common laboratory contaminants.[13] Manipulation of mass spectrometry data was assisted by the use of several perl script programs designed in-house, all of which are freely available for download from our laboratory Web site as part of the Wildcat Toolbox (http://proteomics.arizona.edu/toolbox.html). The first release of this set of perl scripts is described in detail in a previous report,[37] but the data manipulation in this study was performed using two additional perl scripts, which have now been added to the toolbox collection.

For the data analysis outlined in this report, six distinct sets of MudPIT data were acquired, and all six data sets were searched using SEQUEST against both a forward and reversed database.[23–25,38] False discovery rates (FDR) were calculated by determining the number of matches against the reversed database as a percentage of the number of matches against the forward database, which gives an estimate of random sequence matches to the database, in accordance with recently published pro-

TABLE 1

Protein Identifications and False Discovery Rates in SEQUEST Analysis of MudPIT Data

			FDR[c]		
Expt No	Total proteins identified[a]	Single peptide proteins identified[b]	Single peptides only	Overall	Two peptides minimum
Y1	532	248	50.4	23.9	1.1
Y2	604	295	51.2	25.5	2.9
Y3	517	262	47.7	25.5	5.7
R1seed	221	155	41.9	29.9	3.1
R2root	258	175	28.6	19.4	0.0
R3leaf	247	169	59.2	40.9	2.6

[a] Number of proteins identified in yeast and rice MudPIT protein identifications using SEQUEST cutoff scores of: Xcorr for a 1+ ion = 1.8, Xcorr for a 2+ ion = 2.5, Xcorr for a 3+ ion = 3.5, deltaXcorr = 0.1.

[b] Number of proteins identified from single peptides only using SEQUEST with cutoff parameters detailed in footnote a.

[c] False discovery rates assessed by searching against a reversed sequence database, calculated using FDR is FP/(TP + FP), where FP is false positives and TP is total positives,[24] expressed as a percentage.

teomics data guidelines.[19,20] In numerical terms, FDR is FP/(TP + FP), where FP is false positives and TP is total positives.[24] It is important to note that we have not addressed false-negative assignments in this report for two reasons: first, identification of false-negative assignments from a biological sample where the "correct" answer is not known is problematic; and second, the method presented here is simply intended to limit the false discovery rate using available search algorithms.

The number of proteins identified in each experiment, along with the protein false discovery rate in each experiment, is shown in Table 1. The salient features of these data are, first, that the largest contributor to the overall false-positive rate is very clearly those proteins identified from single peptides, and second, that by using a two-peptide minimum criterion, our currently used SEQUEST cutoff parameters would give us a satisfactory confidence of protein assignment. When a minimum of two peptides per protein is imposed, our current SEQUEST parameter cutoff scores produce a false discovery rate below the targeted 5% threshold. One data set out of six has an FDR of 5.7%, but the average for all six experiments is 3.1%.

The DTA_sorter.pl script was developed to extract those .dta files corresponding to SEQUEST single-peptide identifications. This script uses the DTASelect-filter.txt output file[33] and separates all .dta files from a MudPIT run into three newly created folders: singlexcel, which contains all .dta files that correspond to single-peptide identifications; inexcel, which contains all of the .dta files that correspond to multiple-peptide protein identifications; and notinexcel, which contains all of the remaining .dta files. The script then creates a concatenated .dta file from all of the individual .dta files contained in each newly created subdirectory, for use in further searching.

The CommonSingles.pl script was developed for data output comparison purposes. It compares a DTASelect output file (DTASelect-filter.txt) to an XTandem Excel table output (obtained using the Global Proteome Machine xml input upview page at http://www.thegpm.org). The CommonSingles script produces a modified DTASelect output file that includes all of the single peptides found by XTandem that are also found by SEQUEST.

Spectra corresponding to the single-peptide-based protein identifications from all six experiments were sorted using DTA-sorter.pl, re-searched using XTandem, and the single-peptide identifications common to both algorithms were combined with the multiple-based protein identifications using the Commonsingles.pl program. The same procedure was used for both forward and reversed databases to allow calculation of FDR.

Table 2 shows the revised numbers of proteins identified in each of the six MudPIT experiments. The false discovery rates of the overall data sets have dropped from approximately 25% in the initial SEQUEST searches to less than 1% in the dual algorithm search results, while the false discovery rates for the single peptides considered in isolation have dropped from around 50% to less than 1%, zero in some cases. This is a dramatic improvement

TABLE 2

Protein Identifications and False Discovery Rates Observed Using Dual Algorithm Searching

Expt No	Total Proteins Identified in SEQUEST Searches	Revised Total Proteins Identified Using Dual Algorithm Search	Overall FDR[a] Using Dual Algorithm Search	FDR of Single Peptides Retained in Dual Algorithm Approach
Y1	532	417	0.005	0.0
Y2	604	467	0.011	0.013
Y3	517	384	0.021	0.008
R1seed	221	141	0.71	0.0
R2root	258	174	0.00	0.0
R3leaf	247	153	0.65	0.0

[a] Protein false discovery rates, determined as explained in Table 1.

in overall data quality, and has been obtained without increasing the number of false-negative assignments by simply excluding all of the single-peptide-based matches.

Within the yeast samples, there is a high level of reproducibility in the results. When compared to samples prepared from rice organs, there is a clear difference in false discovery rates, as expected in samples from different biological sources.[25] The reanalysis of the yeast MudPIT datasets results in the retention of an average of 76.7% of all proteins identified by SEQUEST, which includes on average 52.1% percent of the single-peptide identifications. For the rice MudPIT datasets, an average of 64.4% of the total proteins are retained, which includes an average of 48.3% of the single-peptide identifications.

While none of the partially tryptic peptides contained in the SEQUEST analysis data sets were confirmed by XTandem searching, a large number of fully tryptic peptides were dropped from the final dataset as they were not confirmed using the second algorithm. This confirms that we are not simply filtering the single-peptide matches on the basis of tryptic status, which is essential as not all of our experiments involve solely trypsin digestion. When analyzing the common singles, none of the dual algorithm consensus matches are partially tryptic; all are fully tryptic. However, out of 115 single-peptide matches dropped from Y1, 58 (50.4%) are partially tryptic; for Y2, 91 of 137 (66.4%) are partially tryptic; and for Y3, 83 of 133 (62.4%) are partially tryptic. Further analysis of the forward and reversed database search results (data not shown) demonstrates that imposing a fully tryptic constraint on the single-peptide matches would improve the FDR compared to the original SEQUEST results, but would not bring it below our desired threshold rate of less than 5%.

In conclusion, we have presented a method for verifying proteins identified from a single unique peptide during nanoLC-MS/MS experiments such as MudPIT analysis of a complex biological mixture. For the analysis of yeast MudPIT datasets, we are able to produce a revised results output with an overall false-positive assignment rate of less than 1%, which still retains over 75% of the proteins initially identified. Similarly, for analysis of the rice organ MudPIT datasets, we are able to retain over 60% of the proteins initially identified, with a revised overall false discovery rate less than 1%. This indicates that application of this technique is highly reproducible for the analysis of similar samples, and likely to yield comparable, yet distinctly different, results for samples prepared from different biological sources.

We have developed a technique that can be employed by laboratories utilizing a SEQUEST-based proteomic analysis platform, incorporating the XTandem algorithm as a complementary tool for verification of single-peptide protein identifications. We have achieved this using open-source software, including several data-manipulation software tools developed in our laboratory, which we have made freely available for download. We make these programs available to other users in the spirit of open-source collaboration, and we hope and expect that users will modify them to fit their own needs. For example, it would be relatively simple to adapt these tools for use with Mascot rather than SEQUEST as the primary search engine, or Mascot rather than XTandem as the secondary search engine. We are currently investigating these approaches, and we have encountered numerous validation issues, such as the selection of different protein isoforms by different programs, and the mechanisms each program uses for filtering out of peptide identifications which have a closely related hit due to the presence of, for example, Asp-Asn or Gln/Glu isoforms in the protein sequence database.

ACKNOWLEDGMENTS

The authors would like to thank Tim Radabaugh, Mike Galligan, Ron Beavis, Fatimah Hickman, Judith Hornby, and Kris Orsborn for helpful discussions and technical assistance, and the Bio5 Institute for Collaborative Bioresearch for funding. P.H. would like to thank Vicki Chandler, Vicki Wysocki, Peter Seghesio, and James Squire for continued support and encouragement. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Air Force, Department of Defense, or the United States Government.

REFERENCES

1. Eng J, McCormack AL, Yates JR, III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Amer. Mass Spectrom.* 1994;5:976–989.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198–207.
3. Washburn MP, Wolters D, Yates JR, III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;19:242–247.
4. Wolters DA, Washburn MP, and Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001;73:5683–5690.
5. Yates JR, III, Eng JK, McCormack AL, Schieltz D. A method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* 1995;67:1426–1436.
6. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567.
7. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 2005;5:3475–3490.
8. Skipp P, Robinson J, O'Connor CD, Clarke IN. Shotgun proteomic analysis of *Chlamydia trachomatis*. *Proteomics* 2005;5:1558–1573.
9. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* 2003;17:2310–2316.
10. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467.
11. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004;3:1234–1242.
12. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. *J Proteome Res* 2004;3:958–964.
13. Breci L, Hattrup E, Keeler M, Letarte J, Johnson R, Haynes PA. Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing. *Proteomics* 2005;5:2018–2028.
14. Benzinger A, Muster N, Koch HB, Yates JR, 3rd, Hermeking H. Targeted proteomic analysis of 14-3-3 sigma, a p53 effector commonly silenced in cancer. *Mol Cell Proteomics* 2005;4:785–795.
15. Durr E, Yu J, Krasinska KM, Carver LA, Yates JR, Testa JE, et al. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat Biotechnol* 2004;22:985–992.
16. Phillips GR, Anderson TR, Florens L, Gudas C, Magda G, Yates JR, III, et al. Actin-binding proteins in a postsynaptic preparation: Lasp-1 is a component of central nervous system synapses and dendritic spines. *J Neurosci Res* 2004;78:38–48.
17. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR, III. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal Chem* 2002;74:1650–1657.
18. Swanson SK, Washburn MP. The continuing evolution of shotgun proteomics. *Drug Discov Today* 2005;10:719–725.
19. Taylor GK, Goodlett DR. Rules governing protein identification by mass spectrometry. *Rapid Commun Mass Spectrom* 2005;19:3420.
20. Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, et al. Guidelines for the next 10 years of proteomics. *Proteomics* 2006;6:4–8.
21. Klammer AA, MacCoss MJ. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J Proteome Res* 2006;5:695–700.
22. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74:5383–5392.
23. Moore RE, Young MK, Lee TD. Qscore: An algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom* 2002;13:378–386.
24. Weatherly DB, Astwood JA, 3rd, Minning TA, Cavola C, Tarleton RL, Orlando R. A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* 2005;4:762–772.
25. Qian WJ, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, et al. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: The human proteome. *J Proteome Res* 2005;4:53–62.
26. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, et al. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* 1999;17:676–682.
27. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 2001;19:946–951.
28. Anderson DC, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res* 2003;2:137–146.
29. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, et al. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002;2:513–523.
30. Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, et al. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* 2004;76:3556–3568.
31. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004;22:1459–1466.
32. Koller A, Washburn MP, Lange BM, Andon NL, Deciu C, Haynes PA, et al. Proteomic survey of metabolic pathways in rice. *Proc Natl Acad Sci USA* 2002;99:11,969–11,974.
33. Tabb DL, McDonald WH, Yates JR, 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 2002;1:21–26.
34. Frigeri LG, Radabaugh TR, Haynes PA, Hildebrand M. Identification of proteins from a cell wall fraction of the diatom *Thalassiosira pseudonana*: Insights into silica structure formation. *Mol Cell Proteomics* 2006;5:182–193.
35. Medina ML, Haynes PA, Breci L, Francisco WA. Analysis of secreted proteins from *Aspergillus flavus*. *Proteomics* 2005;5:3153–3161.
36. Orsborn KI, Shubitz LF, Peng T, Kellner EM, Orbach MJ, Haynes PA, et al. Protein expression profiling of *Coccidioides posadasii* by two-dimensional differential in-gel electrophoresis and evaluation of a newly recognized peroxisomal matrix protein as a recombinant vaccine candidate. *Infect Immun* 2006;74:1865–1872.

37. Haynes PA, Miller SJ, Radabaugh TR, Galligan M, Breci L, Rohrbough J, et al. The Wildcat Toolbox: A set of perl script utilities for use in peptide mass spectral database searching and proteomics experiments. *Journal of Biomolecular Techniques* 2006;17:97–102.

38. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2003;2:43–50.