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

I. HOW MUCH OF THE DATA IS THEORETICALLY INTERPRETABLE BY SEARCH ENGINES?*

Robert J. Chalkley‡§, Peter R. Baker‡, Kirk C. Hansen‡, Katalin F. Medzihradszky‡, Nadia P. Allen¶, Michael Rexach¶, and Alma L. Burlingame‡

An in-depth analysis of a multidimensional chromatography-mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight (QqTOF) geometry instrument was carried out. A total of 3269 CID spectra were acquired. Through manual verification of database search results and de novo interpretation of spectra 2368 spectra could be confidently determined as predicted tryptic peptides. A detailed analysis of the non-matching spectra was also carried out, highlighting what the non-matching spectra in a database search typically are composed of. The results of this comprehensive dataset study demonstrate that QqTOF instruments produce information-rich data of which a high percentage of the data is readily interpretable.

Molecular & Cellular Proteomics 4:1189–1193, 2005.

Mass spectrometers interfaced to chromatographic separation allow the acquisition of large amounts of data in a relatively short period of time. New high throughput technologies have thus been developed to utilize this ability (1–6). The quantity of data produced renders manual analysis of a significant amount of the data impractical. Scientists are therefore dependent on automated database search engines to summarize their results, the most popular being Mascot (www.matrixscience.com) and Sequest (8).

In database searches of large datasets there is always a long list of spectra that have not been matched to anything by the search engine. There are a number of reasons why these may not match, including poor quality spectra, spectra of peptides containing modifications that were not considered in the search, or peptides that were formed by non-specific cleavages when a certain enzyme cleavage specificity was defined in the search engine. Also the data analyzed by search engines are not the raw data but rather centroided peak list data, which are not always completely representative of the raw data.

These unmatched spectra are typically ignored despite the possibility they could contain important information. A summary of the complications in automated peptide and protein identification has been published recently (9). Hence a number of groups have developed statistical analysis programs of search results to better define the reliability of the reported matches (10–13).

There are many groups publishing results from large scale mass spectrometric analyses using different combinations of mass spectrometers and search engines. Unfortunately if a researcher uses one particular combination of tools it can be difficult to assess the quality of the data in studies using different instrument and search engine combinations. Hence there is a drive toward making the raw data itself available so that one can independently assess results and, if desired, reanalyze the results using an alternative searching strategy (14).

In this study we present data from a multidimensional LC-MSMS experiment where we analyzed all acquired spectra manually. From this we are able to report exactly what these unmatched spectra actually constitute. We think this information is important for understanding where there are currently problems with these automated search strategies and to indicate areas where with further refinement this list of unmatched spectra could be reduced. The dataset submitted here was acquired on a QqTOF† geometry instrument, a QSTAR Pulsar (MDS Sciex/Applied Biosystems). A dataset of a multidimensional LC-MSMS experiment created on an ion

From the ‡Mass Spectrometry Facility, University of California San Francisco, San Francisco, California 94143-0446 and the ¶Department of Biological Sciences, Stanford University, Stanford, California 94305-0155

Received, January 1, 2005, and in revised form, April 27, 2005
Published, MCP Papers in Press, May 27, 2005, DOI 10.1074/mcp.D500001-MCP200

† The abbreviation used is: QqTOF, quadrupole selecting, quadrupole collision cell, time-of-flight.
were initially searched using Batch Tag, a new program in total of 3269 MSMS spectra were acquired. These spectra were initially searched using Batch Tag, a new program in total of 3269 MSMS spectra were acquired. These spectra reduced with trichloroethylphosphine, and then cysteines of G1/H9262 M phase using 20% acid and Buffer B was buffer A containing 400 mM NH4Cl. Six fractions of each cell state were labeled with the cleavable ICAT reagent (Applied Biosystems, Foster City, CA) and analyzed essentially following our published protocol for ICAT of low level samples (18). Briefly proteins were denatured in 9 M urea and reduced with trichloroethylphosphine, 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 trap. Separation was achieved using a 2.1 × 10-mm polysulfonethyl 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 (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% trifluoroacetic acid. ICAT tags were cleaved in 95% trifluoroacetic acid. Each fraction was reverse phase cleaned up (Zip Tips, Millipore) 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 achieved 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 acetonitrile, 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) and were analyzed using data-dependent switching between MS and MSMS modes; after a 1-s MS spectrum up to three multiply charged precursor ions could be selected for 2-s CID spectra acquisition. 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 each LC-MS run were created using the Mascot.dll script (version 1.4) within Analyst. These were searched using “Batch Tag,” a new piece of software in the latest in-house developmental version of Protein Prospector (for further details see Ref. 19). The database search results were used to assist in the manual analysis of each spectrum, i.e. if a sequence tag of three or four amino acids was manually interpreted and this matched to the result Batch Tag returned and this result also explained the assignment of all the major peaks in the spectrum then the assignment was accepted.

Approximately 2000 of these spectra gave confident results, and these were verified only by a cursory look at plots of the ions observed and what they were matched to. The majority of these matches were on the basis of an extensive “y” ion series. The other ~1300 spectra were manually analyzed in more extensive detail to determine whether the peptides could be de novo interpreted and, if not, why a peptide could not be confidently assigned.

Following this comprehensive analysis of the dataset we could confidently assign 2368 spectra to predicted tryptic peptides that we felt a search engine should be able to identify when allowing for the modifications of oxidized methionines, protein N-terminal acetylation, and pyroglutamate formation from N-terminal glutamine residues. This left 901 spectra that for various reasons one would not expect the search engine to make a confident match. The reasons for this are summarized in Table I and reported graphically in Fig. 1. 226 of the spectra were not fragmentation spectra of peptides but were rather fragments of chemical contaminants, most commonly ICAT-related products (presumably either chemical side product impurities during synthesis of the reagent or produced by side reactions during the reagent cleav-
age step in 95% trifluoroacetic acid). An example of one of these spectra is shown in Fig. 2. These spectra do not produce any immonium ion masses and nearly always contain characteristic fragment ions at m/z 481.28, 515.28, and 556.29.

Some fragmentation spectra were of peptides less than 620 Da in mass, generally corresponding to a peptide only five amino acids in length. Many of the spectra of these short peptides do not contain enough ions to make a confident assignment, and even if the sequence could be determined, then a five-amino acid string is not sufficient to uniquely identify a protein. The selection window for the precursor ion corresponded to roughly −1.5 and +2 Da from the selected monoisotopic mass. For 43 spectra multiple precursor ions co-eluted within this mass range and were simultaneously fragmented. In some cases both of the peptides could be identified by manual analysis and in many cases at least one could be determined, but unless one component was present at a significantly higher level than the other it would be difficult for the search engine to produce a confident match. 42 spectra were of peptides that would not be formed by tryptic cleavage of proteins in the database so are either formed by non-specific tryptic cleavage or other protease cleavage during sample isolation, and a further eight were formed by in-source fragmentation of an abundant co-eluting peak to produce a peptide with no enzyme specificity. Hence searching the dataset specifying tryptic cleavage would not match these spectra, although searching with no enzyme specificity could potentially identify these.

A total of 51 spectra were of modified peptides. The majority of these were either peptides where an asparagine had become deamidated to an aspartic acid or were from the trypsin, which is methylated to reduce chymotryptic activity and minimize autolysis (20). However, there was also a peptide that had an internal disulfide intact, thus having a molecular mass 2 Da less than the peptide with free sulfhydryl groups. A peptide from elongation factor 1α was identified that had a methylated lysine. This lysine 30 is a known site of modification (7).

A number of spectra could not be assigned because of problems in the creation of the peak list used for searching. The data are acquired as profile data but become converted to centroid data for database searching. Errors in the assignment of the peak charge state and recognition of the monoisotopic peak after this centroiding process lead to incorrect information about the parent ion mass, and thus the peptide will not be identified. Both of these problems were most common in spectra of components of relatively high mass (2500 Da or higher) and were mainly caused by poor ion statistics on weak monoisotopic peaks. Jagged peak shapes lead to labeling of multiple spikes on one isotopic peak, leading to the software interpreting this as a part of a highly charged ion and not part of the same isotope profile as the second and third isotopes (Fig. 3).

Four peptides corresponded to sequences that were not present in either Swiss-Prot (04.03.2004) or National Center for Biotechnology Information (NCBI) (03.29.2004) Databases. 313 spectra did not contain enough information for a confident manual assignment of a peptide mainly because they were weak spectra with few ions. In some cases a more intense MSMS spectrum of the same precursor was acquired at a similar time in the same or a neighboring ion exchange...
Table 1. reason for lack of assignment is supplied in Supplemental full curated list of all the spectra and their assignments or fraction; this allowed assignment of the weaker spectrum. The full curated list of all the spectra and their assignments or reason for lack of assignment is supplied in Supplemental Table 1.

DISCUSSION

From reverse phase LC-MSMS analysis of six cation exchange fractions a total of 3269 CID spectra were acquired. Of these 2368 spectra (72%) can be confidently interpreted as tryptic peptides by a combination of database searching with manual verification or manual de novo interpretation. There were errors in assignment of parent ion mass of 181 spectra through incorrect charge state and/or monoisotopic mass determination. Monoisotopic peak recognition and charge state determination are often not straightforward. However, new software is improving at this task. For example, Matrix Science recently released new software, Mascot Distiller, that made fewer errors in parent ion determination on this dataset (data not shown). Also many peak centroiding scripts, including the recent Mascot.dll in the Analyst software, if they are not certain of the charge state will “create” spectra with different charge states and assume the highest scoring MSMS that results from the multiple assigned charge states is correct. With monoisotopic peak and charge state correctly determined several more spectra could be assigned. 7% of CID spectra were not fragmentation spectra of peptides. This figure is likely to be much higher in datasets acquired on ion trap instruments. Because of the higher resolution of data acquired by time-of-flight, on-the-fly charge state determination of precursor ions allows one to specify to only fragment multiply charged precursor ions. Charge state determination of precursor ions on an ion trap can be performed using a narrow m/z range “zoom scan.” However, many users choose not to perform this scan as it significantly increases the duty cycle of the analysis, reducing the number of precursor ions that are fragmented. Chemical contaminants are generally singly charged, whereas peptides usually are multiply charged species through capture of protons on the basic N terminus and the C-terminal basic residue (lysine or arginine). Hence QqTOF MSMS datasets will contain significantly fewer fragmentation spectra of non-peptide species.

This study is not reporting the results of a database search but a manual analysis of what we think a search engine could theoretically achieve on this dataset. For analysis of how search engines perform on this dataset, see the accompanying study (19). As these results are on the basis of manual assignments, there is inherently a subjectivity to the results. For example, 313 spectra were categorized as being unassignable fragmentation spectra of peptides. Their lack of assignment is due to an inability to determine with personal confidence an identity for the spectrum. This was in general due to there being very few ions in the spectrum, although some spectra contained several fragment ions of which many were clearly not derived from a peptide; i.e. the spectrum was a mixture of fragmentation of a peptide and a chemical contaminant.

Through the manual analysis of all the data we have been able to assess the quality of data acquired on a QSTAR mass spectrometer. This analysis has also highlighted some of the problems with the data produced. Although this dataset cannot be taken as completely representative of all data acquired on this type of instrument, it does show that the data are typically information-rich and that a high percentage of the data should be assignable.

Acknowledgments—We acknowledge the assistance of Brian Williamson, Steve Martin, and other members of the AB 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.

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

REFERENCES

1. Link, A. J., Eng, J., Schietz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17, 676–682
2. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247
3. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999
4. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids
in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–386

5. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. Nat. Biotechnol. 21, 315–318

6. Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G., and Mann, M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 419, 537–542

7. Cavallius, J., Zoll, W., Chakraburty, K., and Merrick, W. C. (1993) Characterization of yeast EF-1 alpha: non-conservation of post-translational modifications. Biochim. Biophys. Acta 1163, 75–80

8. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) J. Am. Soc. Mass Spectrom. 5, 976–989

9. Baldwin, M. A., (2004) Protein identification by mass spectrometry: issues to be considered. Mol. Cell. Proteomics 3, 1–9

10. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392

11. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) 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. 2, 137–146

12. MacCoss, M. J., Wu, C. C., and Yates, J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal. Chem. 74, 5593–5599

13. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. J. Am. Soc. Mass Spectrom. 13, 378–386

14. Burlingame, A. L. (2003) Toward deciphering the knowledge encrypted in large datasets. Mol. Cell. Proteomics 2, 425

15. Von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry: I. Statistically annotated datasets for peptide sequences and proteins identified via the application of ICAT and tandem mass spectrometry to proteins copurifying with T cell lipid rafts. Mol. Cell. Proteomic. 2, 426–427

16. Allen, N. P., Huang, L., Burlingame, A., and Rexach, M. (2001) Proteomic analysis of nucleoporin interacting proteins. J. Biol. Chem. 276, 29268–29274

17. Allen, N. P., Patel, S. S., Huang, L., Chalkley, R. J., Burlingame, A., Lutzmann, M., Hurt, E. C., and Rexach, M. (2002) Deciphering networks of protein interactions at the nuclear pore complex. Mol. Cell. Proteomics 1, 930–946

18. Hansen, K. C., Schmitt-Ulms, G., Chalkley, R. J., Hirsch, J., Baldwin, M. A., and Burlingame, A. L. (2003) Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography. Mol. Cell. Proteomics 2, 299–314

19. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a QqTOF mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol. Cell. Proteomics 4, 1194–1204

20. Kostka, V., and Carpenter, F. H. (1964) Inhibition of chymotrypsin activity in crystalline trypsin preparations. J. Biol. Chem. 239, 1799–1803