OMICS A Journal of Integrative Biology
Volume 6, Number 2, 2002

Short Communication

Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis

ANDREW KELLER, SAMUEL PURVINE, ALEXEY I. NESVIZHSKII, SERGEY STOLYAR, DAVID R. GOODLETT, and EUGENE KOLKER

ABSTRACT

Several methods have been used to identify peptides that correspond to tandem mass spectra. In this work, we describe a data set of low energy tandem mass spectra generated from a control mixture of known protein components that can be used to evaluate the accuracy of these methods. As an example, these spectra were searched by the SEQUEST application against a human peptide sequence database. The numbers of resulting correct and incorrect peptide assignments were then determined. We show how the sensitivity and error rate are affected by the use of various filtering criteria based upon SEQUEST scores and the number of tryptic termini of assigned peptides.

INTRODUCTION

Tandem mass spectrometry (MS/MS) of peptides has been increasingly used to identify protein components in complex samples (Aebersold and Goodlett, 2001; Smith et al., 2002). To extract meaningful sequence data by MS/MS, proteins are first digested into smaller peptides, usually by the enzyme trypsin. These peptides are then separated by reverse phase HPLC on-line with the mass spectrometer and are transferred to the gas phase by electrospray ionization (ESI). Inside the mass spectrometer individual peptide ions are selected in a serial fashion for collision-induced dissociation (CID) so that a tandem mass spectrum contains fragments specific to a single parent peptide.

Several methods have been employed to determine peptides corresponding to tandem mass spectra. For example, the SEQUEST (Eng et al., 1994), Mascot (Perkins et al., 1999), and Sonar (Field et al., 2002) applications compare a tandem mass spectrum against those expected for all possible peptides of identical mass within a given mass tolerance obtained from a sequence database. *De novo* sequencing methods derive a peptide sequence from tandem mass spectra without using a database, and are especially valuable for samples from organisms with polymorphic mutations or unsequenced genomes (Hunt et al., 1986; Taylor

---

Institute for Systems Biology, Seattle, Washington.

and Johnson, 1997; Goodlett et al., 2001). Once peptides have been satisfactorily assigned to individual tandem mass spectra, they are used to infer the original protein contents of the sample.

Accurate peptide assignments to tandem mass spectra are necessary for accurate protein identifications. Researchers with expertise can manually confirm peptide identifications, yet such a time-consuming approach is not feasible for high-throughput analysis. Alternatively, they can attempt to separate the correct from incorrect peptide assignments by applying filtering criteria based upon computed scores and properties of the assigned peptides, such as the number of termini consistent with cleavage by trypsin (Link et al., 1999; Washburn et al., 2001; Han et al., 2001; Smith et al., 2002). However, the rates of false identifications that result from applying such filters are not known, nor how those rates are affected by sample preparation, mass spectrometer, sequence database, or spectrum quality. In addition, researchers often use different filtering criteria, making it particularly difficult to compare their results to one another.

In this work, we describe a data set of peptide tandem mass spectra generated from a control mixture of 18 purified proteins that can be employed to evaluate strategies for tandem mass spectral analysis. Based on this data set, the numbers of correct and incorrect peptide assignments resulting from the use of any method to assign peptides to tandem mass spectra, and any criteria for filtering data, can be determined. As an example, the tandem mass spectra were used to calculate sensitivities and false identification error rates resulting from applying various filtering criteria to SEQUEST search results with a human peptide database.

MATERIALS AND METHODS

Two mixtures, A and B, were obtained by mixing together 18 purified proteins of different physicochemical properties (Sigma, St. Louis, MO; Prozyme, San Leandro, CA) in the indicated relative molar amounts (Table 1). Each mixture, at approximately 1 μg/μL concentration, was digested overnight at 37°C

TABLE 1. PROTEIN COMPONENTS OF CONTROL MIXTURES A AND B

Protein	*Source*	*Sequence accession no.*	*Concentration in A (nM)*	*Concentration in B (nM)*
1. Bovine β-casein	Sigma C6905	P02666	1,000	100
2. Bovine carbonic anydrase	Sigma C2522	P00921	1,000	100
3. Bovine cytochrome c	Sigma C2037	P00006	400	120
4. Bovine β-lactoglobulin	Sigma L0130	P02754	200	1,000
5. Bovine α-lactalbumin	Sigma L6010	P00711	100	300
6. Bovine serum albumin	Sigma A3059	P02769	400	120
7. Chick ovalbumin	Sigma A2512	P01012	4	12
8. Bovine transferrin	Sigma T0178	Q29443	100	300
9. Rabbit GAPDH	Sigma G2267	P46406	20	6
10. Rabbit phosphorylase b	Prozyme	P00489	10	100
11. *E. coli* β-galactosidase	Prozyme	P00722	4	12
12. Bovine γ-actin	Sigma A3653	ATBOG	2	20
13. Bovine catalase	Sigma C40	P00432	20	6
14. Rabbit myosin (heavy and light chains)	Prozyme	P02562[a]	2	20
15. *E. coli* alkaline phosphatase	Prozyme	P00634	200	1,000
16. Horse myoglobin	Sigma M0630	P02188	40	4
17. *B. lichenformis* α-amylase	Sigma A4551	Q04977	40	4
18. *S. cerevisiae* phosphomannose isomerase	Prozyme	P29952	10	100

[a] Additional accession numbers for rabbit myosin heavy and light chains: P02603, P02602, P24732, Q28641, P04460, P04461, P35748, Q99105.

with 1 μg porcine modified trypsin (Promega) per 100 μg protein in the mixture, and purified on a handmade reverse phase column using a pressure cell from Mass Evolution (Spring, TX). The mixtures were loaded across the column twice in the presence of 0.2% acetic acid (HOAc), washed with 20 column volumes of 0.2% HOAc, and then eluted from the column in the presence of 100% acetonitrile. The resultant eluent was dried to completion and resuspended in the original volume of 0.2% HOAc. The complex peptide mixtures were analyzed by μLC-MS on an ESI-ITMS (ThermoFinnigan, San Jose, CA) using a standard top-down data-dependent ion selection approach, wherein the most abundant peak above background level is selected and a concurrent 3 min of dynamic exclusion is employed to prevent re-selection of previously selected ions. Peptides were eluted by an acetonitrile gradient (10% to 35% over 60 min) across a 10 cm × 100 μm C18 column while the ITMS continuously selected peptides for CID via alternating MS and MS/MS modes. To increase duty cycle, the zoom scan function capable of determining charge state was not employed. This experimental design was aimed to mimic realistic MS/MS experiments on complex protein mixtures.

In total, 14 LC/MS/MS runs were performed on mixture A, using 10 μL (A1), 5 μL (A2), 1 μL (A3), or 2.5 μL (A4-14) of 1:5 dilute mixture. Eight LC/MS/MS runs were performed on mixture B, using 1 μL (B1-2), 2 μL (B3-4), 5 μL (B5-6), or 7.5 μl (B7-8) of 1:20 dilute mixture.

RESULTS

Combined tandem mass spectra from the 14 LC/MS/MS runs on control mixture A and the 8 LC/MS/MS runs on control mixture B were used to evaluate database search results using the SEQUEST analysis program (Eng et al., 1994). Since the control mixtures contain proteins that range in concentration, molecular weight, physicochemical properties, and sequence, they serve as a reasonable approximation to realistic samples subjected to tandem mass spectral analysis. The spectra should only correspond to peptide sequences of the proteins in the control mixture. Thus, provided that the peptide sequences of all contents of the control mixture are known, the validity of database search results can be determined.

Verification of control mixture components

The sequence of one of the proteins in the control mixture, bovine γ-actin, was reported as putative. As a means of confirming that sequence, the tandem mass spectra were initially searched using SEQUEST against the NCBI nonredundant protein database to identify all high-scoring peptide assignments corresponding to proteins similar to γ-actin. Interestingly, high scoring assignments of three peptides corresponding to human actin were observed. As illustrated in Table 2, these peptides differ from their homologues in the reported bovine γ-actin sequence at the highlighted amino acids. Since no assignments to spectra of the expected bovine γ-actin peptides were observed, the true bovine γ-actin sequence was assumed to coincide with that of human actin for those peptides. A similar correction was made to the reported sequence of rabbit GAPDH, in which an unknown amino acid (denoted as X) was inferred to be an alanine on the basis of an assigned peptide corresponding to the human homologue (Table 2).

The SEQUEST search results with the NCBI nonredundant protein database also revealed the presence of several probable contaminants in the control mixture. For example, peptides that correspond to human keratin, a frequently identified contaminant introduced during sample handling, as well as to bovine α-

TABLE 2. DISCREPANCIES BETWEEN EXPECTED AND OBSERVED PEPTIDE SEQUENCES (INDICATED IN BOLD TYPE)

Protein	*Expected peptide*	*Observed peptide*
Bovine γ-actin	VAPEEHP**V**LLTEAPLNPK	VAPEEHP**T**LLTEAPLNPK
Bovine γ-actin	TTGIV**M**DSGDGVTH**T**VPIYEGYALPH	TTGIV**L**DSGDGVTH**N**VPIYEGYALPH
Bovine γ-actin	GYSF**T**TTAER	GYSF**V**TTAER
Rabbit GAPDH	VIISAPS**X**DAPMFVMGVNHEK	VIISAPS**A**DAPMFVMGVNHEK

s1–casein, bovine α-s2–casein, and bovine κ-casein, yet have no homology with any control mixture proteins, were assigned with high scores to spectra. The last three proteins were likely introduced to the control mixture as impurities in the bovine β-casein preparation, reported as 90% pure, which was present at relatively high concentrations in the control mixture.

Evaluation of SEQUEST search results with human sequence database

In order to evaluate SEQUEST, it was used to search the tandem mass spectra against a human peptide database (extracted from ftp://ftp.ncicrf.gov/pub/nonredun/protein.nrdb.Z) appended with sequences of the 18 control mixture proteins, with no constraints on the number of tryptic termini of peptides. The database included the corrected sequences of bovine γ-actin and rabbit GAPDH, and sequences of the identified contaminants. Since $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions cannot be distinguished by our data acquisition process using low resolution ESI ion trap mass spectrometry, each tandem mass spectrum is searched by SEQUEST against the database and assigned a peptide separately for each precursor ion charge. This analysis produced a total of 18,496 peptide assignments to spectra of $[M + 2H]^{2+}$ ions, 18,044 to spectra of $[M + 3H]^{3+}$ ions, and 504 to spectra of $[M + H]^{+}$ ions.

SEQUEST peptide assignments corresponding to proteins other than the 18 in the control mixture and contaminants are inferred to be incorrect. Peptide assignments corresponding to the 18 control mixture proteins or contaminants could occur merely by chance, and hence must be manually scrutinized to determine whether or not they are correct. Figure 1 shows distributions of SEQUEST *Xcorr* score for spectra of $[M+2H]^{2+}$ ions among peptide assignments corresponding to, or not corresponding to, control mixture proteins. This score measures the number of peaks of common mass between observed and expected spectra, and is an indication of the quality of the assignments. Low-scoring assignments due to chance alone are responsible for the prominent left shoulder of the distribution among peptide assignments corresponding to control mixture proteins. When chance assignments contributing to this distribution were identified by manual scrutiny and reclassified as "incorrect," the resulting *Xcorr* distribution among "correct" peptide assignments exhibited a reduced shoulder (Fig. 1, inset). In total, 1656 peptide assignments to spectra of $[M+2H]^{2+}$ ions, 984 to spectra of $[M+3H]^{3+}$ ions, and 125 to spectra of $[M+H]^{+}$ ions, were determined to be correct.

With SEQUEST database search results of known validity, sensitivities and false identification error rates resulting from the use of various filtering criteria can be calculated. Table 3 shows that filters 1, 4, and 5 yielded similar sensitivities (fraction of all correct peptide assignments passing filter) and error rates (fraction of all peptide assignments passing filter that are incorrect) to one another, whereas filters 2 and 3 achieved higher sensitivity by accepting correctly assigned peptides with only 1 tryptic terminus. However,

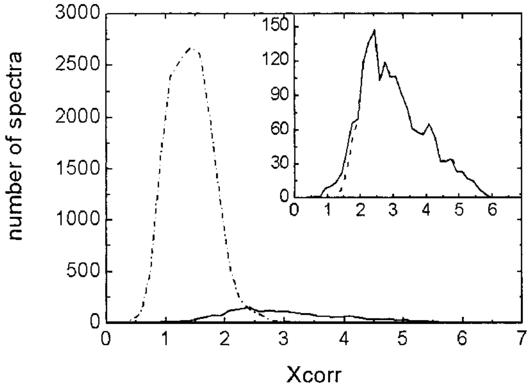

FIG. 1. SEQUEST *Xcorr* score distributions. Distributions for spectra of $[M + 2H]^{2+}$ ions among: peptide assignments corresponding to control mixture proteins [——]; peptide assignments not corresponding to control mixture proteins [– • – • –]; "correct" peptide assignments after reclassification of chance assignments as "incorrect" [– – – –]. The distributions were obtained by placing spectra in bins of width 0.17 according to *Xcorr*, and counting the resulting total number in each bin.

TABLE 3. SENSITIVITIES AND ERROR RATES FOR FILTERED SEQUEST DATABASE SEARCH RESULTS

Filtering criteria	*Sensitivity*	*Error rate*
(1) Washburn et al., 2001	0.67	0.03
(2) $X_1 \geq 1.5$, $X_2 \geq 2$, $X_3 \geq 2.5$, $\Delta C_n \geq 0.1$, $NTT \geq 1$	0.78	0.09
(3) $X_1 \geq 2$, $X_2 \geq 2$, $X_3 \geq 2$, $\Delta C_n \geq 0.1$, $NTT \geq 1$	0.77	0.15
(4) $X_1 \geq 1.5$, $X_2 \geq 2$, $X_3 \geq 2.5$, $\Delta C_n \geq 0.1$, $NTT = 2$	0.59	0.02
(5) $X_1 \geq 2$, $X_2 \geq 2$, $X_3 \geq 2$, $\Delta C_n \geq 0.1$, $NTT = 2$	0.59	0.03
None	1.0	0.93

Filtering criteria according to the number of tryptic termini of the assigned peptide (*NTT*) and SEQUEST scores: *Xcorr* for $[M + H]^+$ ions (X_1), *Xcorr* for $[M + 2H]^{2+}$ ions (X_2), *Xcorr* for $[M + 3H]^{3+}$ ions (X_3), and ΔC_n.

this comes at a cost of much increased error, reflecting a greater proportion of incorrectly assigned peptides with 1 tryptic terminus that pass the filter. One expects that the magnitude of increased sensitivity resulting from the use filter 2 or 3 will vary from sample to sample, depending on the efficiency of trypsinization and presence of protease contaminants. Employing no filter resulted in a very high false identification error rate (93%) since each spectrum is assigned a peptide separately for the $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ion cases, yet at most one such assignment can be correct. Additional incorrect peptide assignments likely correspond to very noisy spectra, or spectra produced by modified peptides or non-peptide substances. It should be emphasized that the sensitivities and error rates reported here for SEQUEST are valid for the control mixture spectra searched against the human peptide database, and may differ for other data sets, depending on sample preparation, mass spectrometer, sequence database, or spectrum quality.

CONCLUSION

The tandem mass spectra described here can be used to determine the accuracy of methods to assign peptides to tandem mass spectra. We report sensitivities and false identification error rates for various filtering criteria applied to SEQUEST search results with the human peptide database. It would also be interesting to use these spectra to evaluate additional filtering strategies for SEQUEST results, as well as the results of other MS/MS spectral analyses, including *de novo* sequencing. Moreover, this data set has recently been used to develop statistical models to distinguish correct from incorrect peptide assignments to MS/MS spectra (Keller et al., in preparation) and estimate the likelihood of their corresponding proteins in the original sample (Nesvizhskii et al., in preparation). The described protein mixture, along with other complex mixtures of known properties, could be employed in further statistical models of peptide and protein identification based upon MS/MS spectra.

DATA SETS

Spectra and correct peptide assignments identified from the SEQUEST search against the human peptide database are available upon request at *www.systemsbiology.org/protein_mixture.html.*

ACKNOWLEDGEMENTS

This work was supported in part by the DOE Microbial Cell Program's grant DE-FG08-01ER63218 to E.K. and by sponsored research agreement from Oxford GlycoSciences to A.K. We also thank Leroy Hood and Ruedi Aebersold for support, and Andrew Siegel, Leroy Hood, Ruedi Aebersold, Jimmy Eng, and Eugene Yi for helpful discussions.

REFERENCES

AEBERSOLD, R., and GOODLETT, D.R. (2001). Mass spectrometry in proteomics. Chem Rev **101,** 269–296.

ENG, J.K., McCORMACK, A.L., and YATES J.R. III. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom **5,** 976–989.

FIELD, H.I., FENYO, D., and BEAVIS, R.C. (2002). RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47.

GOODLETT, D.R., KELLER, A., WATTS, J.D., et al. (2001). Differential stable isotope labeling of peptides for quantitation and *de novo* sequencing. Rapid Comm Mass Spec **15,** 1214–1221.

HAN, D.K., ENG, J., ZHOU, H., et al. (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol **19,** 946–951.

LINK, A.J., ENG, J., SCHIELTZ, D.M., et al. (1999). Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol **17,** 676–682.

PERKINS, D.N., PAPPIN, D.J.C., CREASY, D.M., et al. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20,** 3551–3567.

SMITH, R.D., ANDERSON, G.A., LIPTON, M.S., et. al. (2002). The use of accurate mass tags for high-throughput microbial proteomics. OMICS **6,** 61–90.

TAYLOR, J.A., and JOHNSON, R.S. (1997). Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. Rapid Comm Mass Spec **11,** 1067–1075.

WASHBURN, M.P., WOLTERS, D., and YATES, J.R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol **19,** 242–247.

Address reprint requests to:
Dr. Andrew Keller
Institute for Systems Biology
1441 North 34th St.
Seattle, WA 98103

E-mail: akeller@systemsbiology.org

or:

Dr. Eugene Kolker
Institute for Systems Biology
1441 North 34th St.
Seattle, WA 98103

E-mail: ekolker@systemsbiology.org