technical notes Journal of proteome research

A Fast SEQUEST Cross Correlation Algorithm

Jimmy K. Eng,*[,†] Bernd Fischer,[‡] Jonas Grossmann,[§] and Michael J. MacCoss[†]

Department of Genome Sciences, University of Washington, Seattle, Washington, Institute of Computational Science, ETH Zurich, Zurich, Switzerland, and Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland

Received June 9, 2008

Abstract: The SEQUEST program was the first and remains one of the most widely used tools for assigning a peptide sequence within a database to a tandem mass spectrum. The cross correlation score is the primary score function implemented within SEQUEST and it is this score that makes the tool particularly sensitive. Unfortunately, this score is computationally expensive to calculate, and thus, to make the score manageable, SEQUEST uses a less sensitive but fast preliminary score and restricts the cross correlation to just the top 500 peptides returned by the preliminary score. Classically, the cross correlation score has been calculated using Fast Fourier Transforms (FFT) to generate the full correlation function. We describe an alternate method of calculating the cross correlation score that does not require FFTs and can be computed efficiently in a fraction of the time. The fast calculation allows all candidate peptides to be scored by the cross correlation function, potentially mitigating the need for the preliminary score, and enables an *E*-value significance calculation based on the cross correlation score distribution calculated on all candidate peptide sequences obtained from a sequence database.

Keywords: cross correlation • SEQUEST • tandem mass spectrometry • E-value

Introduction

The automated acquisition of tandem mass spectrometry data followed by the interpretation of the data by database searching is a widely used method for the analysis of proteins within complex mixtures. Advances in and the availability of mass spectrometry (MS) instrumentation, analysis software, and sequenced genomes have helped to spur the widespread growth of proteomics analysis to researchers in diverse fields and of diverse expertise. Development of software tools has played a major role in this growth. Users have the choice of a variety of commercial and open source MS/MS database search programs[1−8] at their disposal. One of these tools, SEQUEST,[9] was the first to implement automated database searching of uninterpreted tandem mass spectra and is still widely used today.

Each database search tool offers a variety of features and implements slightly different search strategies. However, one main differentiator between each MS/MS search tool is the score function which measures the closeness of fit between the acquired tandem mass spectrum and the candidate peptides retrieved from the sequence database. For SEQUEST, the primary score function is the cross correlation score (xcorr). One of the reasons the xcorr is so sensitive is because it involves a correction factor that assesses the background correlation for each acquired spectrum and the theoretically predicted spectrum from sequences within a database. To calculate this correction factor, a measure of similarity is calculated at different offsets between a preprocessed mass spectrum and a theoretical spectrum. The final xcorr is then the correlation at zero offset minus the mean correlation from all the individual offsets. The full cross correlation spectrum was calculated using Fast Fourier Transforms (FFT). The FFT-based xcorr algorithm is expensive to compute, so SEQUEST implemented an initial or preliminary score as a more rapid filter of candidate peptides; only the top 500 best preliminary scoring peptides were subjected to the xcorr analysis.

In this manuscript, we describe a faster, more direct way to calculate the xcorr with correction that is indistinguishable from the FFT-based calculation. This new method simplifies the score calculation and enables the xcorr to be calculated on every candidate peptide from the sequence database. The calculation of the xcorr for all candidate peptides enables an *E*-value calculation of match significance.[10] This *E*-value calculation increases the sensitivity of the cross correlation and simplifies the comparison of scores between spectra.

Materials and Methods

Software. The software is written in the C programming language and compiled with the GNU project C compiler 3.4.6. Analysis was performed on an Intel Core 2 Duo E6400 CPU (2.13 GHz) under the RedHat Enterprise Linux AS release 4 operating system.

Tandem Mass Spectrometry Data and Sequence Databases. Publicly available tandem mass spectrometry data from a yeast sample (11 Thermo LCQ Deca XP ion trap runs, opd00034_YEAST) and a human cell line sample (11 Thermo LCQ Deca XP runs, opd00036_HUMAN) were obtained from the Open Proteomics Database[11] and analyzed for this manuscript. Searches were performed against two sequence databases. The yeast sample was searched against a yeast ORF database obtained from the Saccharomyces Genome Database (SGD).[12] The corresponding reverse decoy forms of all sequences were appended to the

* To whom correspondence should be addressed. E-mail: engj@u.washington.edu.

[†] Department of Genome Sciences, University of Washington.

[‡] Institute of Computational Science, ETH Zurich.

[§] Institute of Plant Sciences, ETH Zurich.

10.1021/pr800420s CCC: $40.75

Published on Web 09/06/2008

original sequences to generate a database containing 11 766 sequence entries. The human cell line sample was searched against a forward plus reverse decoy database of human International Protein Index[13] (IPI) version 3.41. The human forward plus decoy database totaled 144 316 sequence entries; this number includes the addition of three forward and corresponding reverse trypsin sequences.

Derivation of the Fast Cross Correlation. The SEQUEST xcorr is a closeness of fit measure between an acquired experimental tandem mass spectrum and a theoretical spectrum representing a candidate peptide sequence obtained from a sequence database. As described in the original publication, the xcorr is calculated as follows:

$$\text{xcorr} = R_0 - \left(\sum_{\tau=-75}^{\tau=+75} R_\tau\right) \Big/ 151 \qquad (1)$$

where

$$R_\tau = \sum x[i] \cdot y[i+\tau] \qquad (2)$$

The xcorr is a scalar dot product between the acquired and theoretical spectrum with a correction factor.[14] Because of the correlation correction factor term used in the xcorr (the second term in eq 1), it has historically been faster to compute the full correlation function across all lags or offsets using FFTs as opposed to explicitly calculating the scalar dot products associated with just the 151 offsets used in eq 1.

Here, we present a method of calculating the xcorr rapidly without the use of FFTs. The method is derived as follows. Substituting eq 2 into eq 1, where x represents the theoretical spectrum and y is the acquired tandem mass spectrum, yields

$$\text{xcorr} = x_0 \cdot y_0 - \left(\sum_{\tau=-75}^{\tau=+75} x_0 \cdot y_\tau\right) \Big/ 151 \qquad (3)$$

As dot products possess the distributive property, eq 3 can be rewritten as

$$\text{xcorr} = x_0 \cdot \left(y_0 - \left(\sum_{\tau=-75}^{\tau=+75} y_\tau\right) \Big/ 151\right) \qquad (4)$$

The arrangement of the calculation shown in eq 4 suggests that the acquired input spectrum y can be preprocessed once at the beginning of a search, which is labeled as y' in eq 5.

$$\text{xcorr} = x_0 \cdot y' \quad \text{where} \quad y' = y_0 - \left(\sum_{\tau=-75}^{\tau=+75} y_\tau\right) \Big/ 151 \qquad (5)$$

Subsequently, each xcorr calculation is a scalar dot product between the theoretical spectrum x and the preprocessed input spectrum y'. Deviating from the original implementation, the current implementation does not make use of the y_0 value in determining the correction factor as that is the signal being measured. This results in the final xcorr calculation shown in eq 6.

$$\text{xcorr} = x_0 \cdot y' \quad \text{where} \quad y' = y_0 - \left(\sum_{\tau=-75, \tau\neq 0}^{\tau=+75} y_\tau\right) \Big/ 150 \qquad (6)$$

Results and Discussion

As described in the original publication, the acquired input spectrum already undergoes preprocessing steps for the xcorr analysis. The processing steps are listed below with associated plots shown in Figure 1. Figure 1a shows the tandem mass spectrum for a doubly charged yeast peptide SGVAVADESLTAFNDLK acquired in an ion trap mass spectrometer. When parsed into SEQUEST, the spectrum is read into near unit-dalton mass bins and with square root of the original intensities (Figure 1b). Intensities are further manipulated by normalizing the maximum intensity to be uniform within a fixed number of m/z windows across the entire spectrum (Figure 1c). The form of the spectrum represented in Figure 1c is what the xcorr is calculated against. The spectrum preprocessing described in eq 6 is applied to Figure 1c resulting in both positive and negative intensity values as shown in the y' spectrum (Figure 1d). The newly described xcorr calculation is generated by taking the scalar dot product of the theoretical spectrum against the spectrum depicted in Figure 1d. This dot product value is the cross correlation score with correction.

The fast xcorr algorithm was implemented in the SEQUEST program and compared against the classical FFT based calculation of its previous implementation. Figure 2a shows the result of a search where both the classical xcorr (XCorr column) and the newly described fast xcorr (fXCorr column) are calculated within the same search and output. The numbers in each column are identical, shown out to four significant digits in this example, which indicates that the fast xcorr implementation described here is exact and accurate. To further illustrate this point, a set of 1500 yeast tryptic searches were performed and the FFT versus dot product xcorrs for the top peptide hits were plotted against each other in Figure 2b. The plot is a perfect diagonal illustrating the fast xcorr implementation faithfully reproduces the FFT based calculation.

The xcorr has typically been calculated on just the top 500 candidate peptides that pass the preliminary score filter because the calculation using FFTs is too expensive to apply to every peptide. As the xcorr is applied to only a subset of the candidate peptides in any given search, it can be a fairly minor component in the overall search time depending on the complexity of the search. In fact, the xcorr is a more significant time component for quick searches on small sequence databases and less significant for complex, slower searches that are dominated by the preliminary score analysis.

Timing a large number of calculations on a 2.13 GHz Intel E6400 CPU resulted in a fast xcorr calculation that took approximately 4×10^{-6} s, while each FFT-based calculation took approximately 3×10^{-4} s. This benchmark shows that the fast xcorr calculation is ~75 times faster than the FFT-based calculation. The most obvious implementation of this new calculation method in the SEQUEST program is to replace the FFT-based calculation with the new calculation. However, with the use of this approach, SEQUEST search times would be incrementally improved by the time difference between the 500 (or less) FFT-based calculations versus new calculations in each search. This implementation was not pursued because the only tangible benefit would be to add minimal incremental gains to the overall search times.

More interestingly, the calculation method described here is fast enough that it is now possible to apply the xcorr to every candidate peptide from the sequence database being queried. By doing so, an expectation or *E*-value metric[10] can be calculated for each putative peptide identification based on the xcorr distribution of each spectrum. The *E*-value is a widely used statistical significance metric that was previously not feasible to calculate in SEQUEST because the slower FFT-based calculation made the computation of xcorr for every peptide candidate in a database prohibitive.

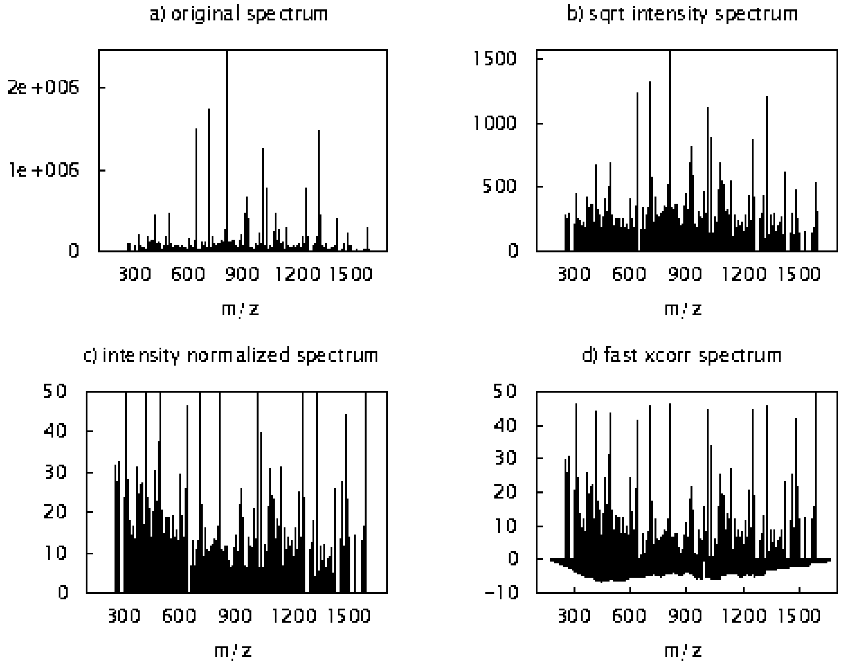
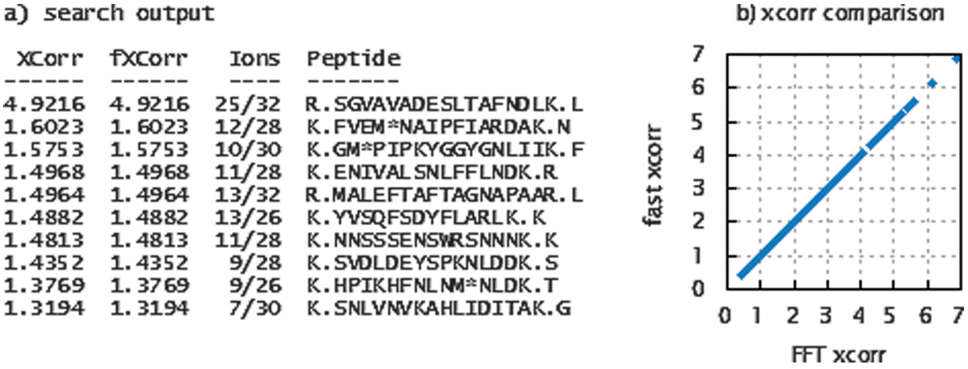


Figure 1. Cross correlation spectral processing of the input spectrum, including the final fast cross correlation processing. The original input spectrum (a), square root intensities of input spectrum (b), normalized intensities across *m/z* range (b), and fast cross correlation processing (d).



```
 XCorr   fXCorr   Ions   Peptide
------   ------   ----   -------
4.9216   4.9216  25/32   R.SGVAVADESLTAFNDLK.L
1.6023   1.6023  12/28   K.FVEM*NAIPFIARDAK.N
1.5753   1.5753  10/30   K.GM*PIPKYGGYGNLIIK.F
1.4968   1.4968  11/28   K.ENIVALSNLFFLNDK.R
1.4964   1.4964  13/32   R.MALEFTAFTAGNAPAAR.L
1.4882   1.4882  13/26   K.YVSQFSDYFLARLK.K
1.4813   1.4813  11/28   K.NNSSSENSWRSNNNK.K
1.4352   1.4352   9/28   K.SVDLDEYSPKNLDDK.S
1.3769   1.3769   9/26   K.HPIKHFNLNM*NLDK.T
1.3194   1.3194   7/30   K.SNLVNVKAHLIDITAK.G
```

Figure 2. SEQUEST search results comparing the classical FFT base score and the newly described fast cross correlation calculation. (a) Search example where the FFT-based score (XCorr) and fast dot product implementation (fXCorr) give identical values for the top 10 hits. (b) Plot of the FFT-based score versus the new implementation for the top peptide hits of 1500 random yeast tryptic searches. The calculated values are identical indicating that the fast derivation implements the cross correlation score faithfully and exactly.

The *E*-value is a statistical measure that can standardize the reporting of confidence of peptide identifications, and in fact, multiple MS/MS search tools currently report such a calculation.[1,4,5,15] In this implementation, the *E*-value is calculated as follows. A raw histogram of xcorr's is maintained for every peptide that undergoes a database search (Figure 3a). The histogram counts are log-transformed and a linear least-squares fit is applied to the underlying distribution (Figure 3b). Heuristics are implemented which exclude data points on the right most tail of the log-transformed distribution where positive identifications may be outliers. The top scoring peptide's xcorr value is projected down to intersect this least-squares fit line. The *E*-value is calculated as the inverse log of the *y*-axis value of this projection. The calculated value can be interpreted as the number of peptides that are expected to score as well as the top scoring peptide by chance in the search. Small *E*-values reflect matches that are likely to be nonrandom.

Various heuristics score filters for SEQUEST have been reported previously using xcorr, deltaCn, and so forth to set thresholds for acceptance of identifications.[16−19] This use of multiple heuristics as a criteria for a correct peptide identification is primarily because there is no single score from a SEQUEST search that can be used to effectively filter putative peptide identifications. The *E*-value is presumed to be a better single score metric than the existing SEQUEST scores and will provide a metric that can calibrate xcorr scores between spectra as a similar metric reported previously.[7] To test this, Figure 4 displays the results of a yeast tryptic search and a human semitryptic search where decoy based false discovery rates[20] are plotted as *q*-values.[21,22] The *q*-value plots are generated based on the *E*-value, raw cross correlation score (xcorr), normalized difference in the cross correlation score (dCn), and preliminary score rank (RSp). Two ad hoc combinations of SEQUEST score filters from previously reported work[17,19] are also plotted as data points for reference. The *E*-value outperforms each of the other SEQUEST scores at all useful *q*-value ranges in both searches. One of the ad hoc SEQUEST score cutoff values does produce more target hits in the human

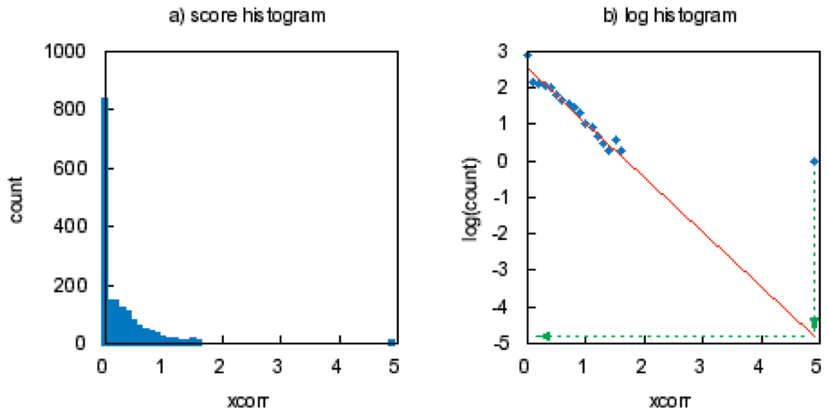


Figure 3. Calculation of *E*-values from the fast xcorr of peptide candidates within a sequence database. (a) Histogram of all xcorr scores for a search and (b) a linear least-squares fit (solid line) to the log histogram (dots). *E*-values are calculated by projecting the top hit along the least-squares line (projections shown by dashed lines).

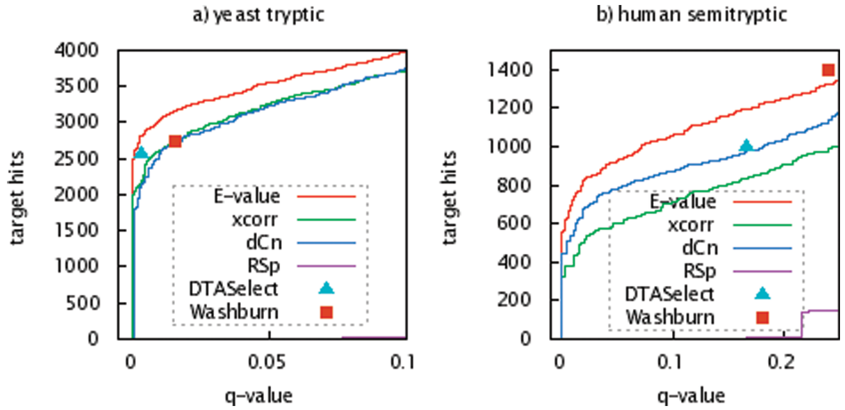


Figure 4. Target versus *q*-value plots for a yeast tryptic (a) and human semitryptic (b) search. *E*-value, cross-correlation score (xcorr), deltaCn (dCn), and preliminary score rank (RSp) are compared. The *E*-value gives the best performance across all useful *q*-value ranges.

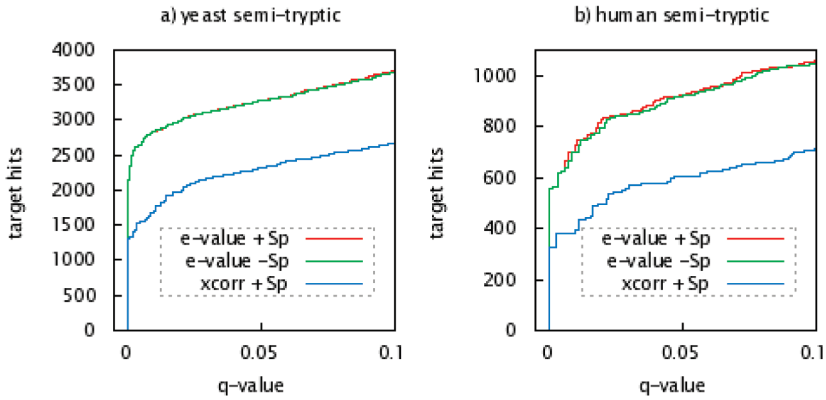


Figure 5. Target peptide versus *q*-value plots with and without the Sp score. The search was performed on a yeast (a) and a human (b) data set. Performance based on *E*-values with and without the preliminary score shows very little deviation in the *E*-value plots indicating that the preliminary score can be eliminated if desired. Xcorr plots are included for reference.

semitryptic search at its given false discovery rate (FDR), but that FDR is high (0.241) and greater than the useful FDR cutoff range that is typically applied (0.1 or less) in this form of analysis.

The preliminary score in SEQUEST was developed as a simple and quick score function because the xcorr was too expensive to calculate on all candidate peptides in a search. Given the new feasibility to calculate the xcorr on every peptide, is the preliminary score still needed and does it provide any ancillary benefits to the analysis? To address this question, the yeast and human data were searched with and without the preliminary score being applied. The data were searched through SEQUEST using a both a fully tryptic and a semitryptic constraint against databases composed of forward plus reverse decoy sequences. False discovery rates were calculated and reported as *q*-values. The semitryptic search analyses are shown in Figure 5.

The plots in Figure 5 show that identification performance is nearly identical irrespective of whether or not the preliminary score is applied. With a larger search space, as observed in the human semitryptic search, there seems to be a very small benefit to the preliminary score filtering. On the other hand, for fully tryptic searches (data not shown), the q-value plots are identical in the 0.0−0.1 q-value range. These results demonstrate that, if desired, the preliminary score can safely be left out for a cross correlation only based search.

Many postsearch analysis tools have been developed to reanalyze or rescore SEQUEST search results and improve on the analysis output.[23−25] Those tools consider more information than is available within a single database search, such as target-decoy hits or enzyme termini specificity, and do outperform *E*-value scores reported directly from a search tool. The *E*-value can be another parameter to pass to these software tools, potentially improving their performance as well.

The SEQUEST program is a commercial program that is currently licensed to Thermo Fischer Scientific. However, the fast cross correlation calculation has been implemented in the Crux[7] program that is available for free academic download (http://noble.gs.washington.edu/proj/crux). Additionally, the described spectral processing and resulting simple dot product calculation lends itself to being implemented in other tools, such as a score plug-in in the pluggable framework of the X!Tandem program.[26]

Conclusions

We describe a fast and precise implementation of the SEQUEST cross correlation score that is suitable to directly replace the more expensive FFT-based calculation that has been historically used. We show how the score function is derived and applications of that implementation. We also show that this implementation obviates the need for the preliminary score algorithm in SEQUEST if desired without a loss in sensitivity. Lastly, we show that the implementation allows all candidate peptides to be scored by the cross correlation algorithm, enabling an *E*-value to be calculated. The implementation of the cross correlation based *E*-value facilitates SEQUEST interpretation, allows standardize reporting of peptide identification confidence, and improves identifications over filtering based on existing classical SEQUEST scores.

Acknowledgment. We thank the Bill Noble laboratory for useful discussions on the xcorr implementation. We thank Veronika Glukhova and Michael R. Hoopmann for helpful comments and critical reading of the manuscript. This work is supported in part by the University of Washington's Proteomics Resource.

References

(1) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.
(2) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2* (10), 1406–1412.
(3) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3* (8), 1454–1463.
(4) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.
(5) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958–964.
(6) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–4639.
(7) Park, C. Y.; Klammer, A. A.; Kall, L.; Maccoss, M. J.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **2008**, *7* (7), 3022–3027.
(8) Roos, F. F.; Jacob, R.; Grossmann, J.; Fischer, B.; Buhmann, J. M.; Gruissem, W.; Baginsky, S.; Widmayer, P. PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics* **2007**, *23* (22), 3016–3023.
(9) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.* **1994**, *5*, 976–989.
(10) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75* (4), 768–774.
(11) Prince, J. T.; Carlson, M. W.; Wang, R.; Lu, P.; Marcotte, E. M. The need for a public proteomics repository. *Nat. Biotechnol.* **2004**, *22* (4), 471–472.
(12) Cherry, J. M.; Adler, C.; Ball, C.; Chervitz, S. A.; Dwight, S. S.; Hester, E. T.; Jia, Y.; Juvik, G.; Roe, T.; Schroeder, M.; Weng, S.; Botstein, D. SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **1998**, *26* (1), 73–79.
(13) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **2004**, *4*, 1985–1988.
(14) Powell, L. A.; Hieftje, G. M. Computer identification of infrared spectra by correlation-based file searching. *Anal. Chim. Acta* **1978**, *100*, 313–327.
(15) Havilio, M.; Haddad, Y.; Smilansky, Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (3), 435–444.
(16) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–682.
(17) Washburn, M. P.; Wolters, D.; Yates, J. R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–247.
(18) Han, D. K.; 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* (10), 946–951.
(19) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1* (1), 21–26.
(20) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.
(21) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (16), 9440–9445.
(22) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29–34.
(23) Moore, R. E.; Young, M. K.; Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass. Spectrom.* **2002**, *13*, 378–386.
(24) Keller, A.; Nesvizhskii, A. I.; 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* (20), 5383–5392.
(25) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923–925.
(26) MacLean, B.; Eng, J. K.; Beavis, R. C.; McIntosh, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **2006**, *22* (22), 2830–2832.

PR800420S