Precursor Charge State Prediction for Electron Transfer Dissociation Tandem Mass Spectra

Vagisha Sharma,[†,‡] Jimmy K. Eng,*[,‡] Sergey Feldman,[§] Priska D. von Haller,[‡] Michael J. MacCoss,[‡] and William S. Noble[‡,||]

Department of Biochemistry, Department of Genome Sciences, Department of Electrical Engineering, and Department of Computer Science and Engineering, University of Washington, Seattle, Washington

Received June 28, 2010

Electron-transfer dissociation (ETD) induces fragmentation along the peptide backbone by transferring an electron from a radical anion to a protonated peptide. In contrast with collision-induced dissociation, side chains and modifications such as phosphorylation are left intact through the ETD process. Because the precursor charge state is an important input to MS/MS sequence database search tools, the ability to accurately determine the precursor charge is helpful for the identification process. Furthermore, because ETD can be applied to large, highly charged peptides, the need for accurate precursor charge state determination is magnified. Otherwise, each spectrum must be searched repeatedly using a large range of possible precursor charge states. To address this problem, we have developed an ETD charge state prediction tool based on support vector machine classifiers that is demonstrated to exhibit superior classification accuracy while minimizing the overall number of predicted charge states. The tool is freely available, open source, cross platform compatible, and demonstrated to perform well when compared with an existing charge state prediction tool. The program is available from http://code.google.com/p/etdz/.

Keywords: electron transfer dissociation • charge state prediction • support vector machine • tandem mass spectrometry

Introduction

Electron-transfer dissociation (ETD) is a relatively new technique for analyzing peptides and proteins by tandem mass spectrometry. ETD induces fragmentation of positively charged peptides or proteins by transferring electrons to them. As an ion fragmentation technology, ETD has gained popularity particularly based on its ability to preserve labile post-translational modifications during the dissociation process.[1−6] ETD is also able to dissociate large peptides and whole proteins, thus providing a nice complement to collision-induced dissociation (CID) for general proteomics analysis.

Although commercially available instruments now exist that can measure the precursor mass and charge state accurately for ETD tandem mass spectra,[7] there still exists a class of instruments, specifically the Thermo Scientific LTQ-ETD, for which the charge states of the ETD spectra cannot be ascertained from the isotope distribution of the precursor ion. And because the precursor charge is a required parameter used by database search tools, it is important to determine ETD charge states; the absence of this information would require that all precursor charge states be analyzed. This issue is magnified for ETD data compared to typical tryptic CID data, as ETD samples are treated with a protease, such as Lys-C, with the specific intent of generating larger peptides that have a higher charge state distribution. And searching the large number of all possible charge states is both computationally expensive and presents unnecessary complications for downstream analysis.

To address this problem, we present an ETD charge state prediction tool that was developed using support vector machine (SVM) classifiers.[8,9] The specific goal of this tool is to maximize correct charge state determinations, thereby minimizing unnecessary sequence database searches due to spurious or unknown precursor charge states for a given spectrum. It should be noted that several tools, such as Charger[10] and the Charge Prediction Machine (CPM),[11] have been developed previously to address this specific problem. Charger, a commercial software tool from Thermo Scientific, performs spectral autocorrelations to identify complementary ions that sum to the precursor mass thus elucidating the precursor charge. Charger falls back to linear discriminant analysis when the charge state cannot be determined by the spectral processing. CPM is a classifier based on Bayesian decision theory and, unlike Charger, is freely available for academic users. CPM is written in C# and is available as a Windows executable that can be run on the Linux command line through the Mono project.[12] Because Charger is a commercial tool that we do not have access to, we evaluate our ETD precursor charge state classification performance relative to CPM. In summary, we demonstrate a novel implementation of an electron transfer

* To whom correspondence should be addressed. E-mail: engj@uw.edu.
[†] Department of Biochemistry.
[‡] Department of Genome Sciences.
[§] Department of Electrical Engineering.
[||] Department of Computer Science and Engineering.

10.1021/pr1006685

Published on Web 08/23/2010

dissociation charge prediction tool that shows favorable performance compared to an existing published tool. It is freely available to both academic and commercial groups and importantly is open source with a liberal license.

Materials and Methods

Data. To develop a charge state prediction tool for LTQ-ETD tandem mass spectra (MS/MS), the first requirement is to acquire a set of ETD MS/MS spectra where the precursor charge states are known. From this set we extract relevant features and train a classifier. MS/MS spectra from the Orbitrap-LTQ-ETD make good surrogates for LTQ-ETD data. The actual MS/MS scans are acquired in the same ion trap mass analyzer, having the same characteristics as LTQ-ETD MS/MS data, but the precursor scans are acquired at high resolution and mass accuracy in the Orbitrap analyzer. The Orbitrap precursor scans provide both an accurate precursor mass and charge state for a majority of the MS/MS scans. From a collection of Orbitrap-LTQ-ETD MS/MS spectra, we perform sequence database searches to identify peptide sequences and filter the results to isolate high confidence identifications with known precursor charge states. From this set of identified spectra, we remove redundancies to generate a unique set of peptides, where "unique" is defined as having a unique peptide sequence, modification state, and charge state combination. Finally, from the list of unique identifications, we reserve a subset of the data to serve as a test set for our classifier and use the remaining data as the training set.

A collection of 71 Thermo Orbitrap-LTQ-ETD LC−MS/MS runs of *S. cerevisiae* were acquired from the Coon Lab at the University of Wisconsin.[13] These data were searched using Mascot,[14] X!Tandem,[15] SEQUEST,[16] and OMSSA[17] to collate a list of known identifications with known charge states. The data were processed as follows. The raw data files were converted to mzXML[18] using the ReAdW program. The mzXML files were directly searched with SEQUEST and X!Tandem. MGF files were generated from the mzXMLs using MzXML2Search, and these were searched by OMSSA and Mascot. Common search parameters include alkylated cysteines as a static modification, oxidized methionine as a variable modification, ETD specific c and z• ion fragmentation, Lys-C enzyme (full digestion for Mascot and OMSSA, semi for X!Tandem and SEQUEST), and yeast ORFs database from SGD[19] (including reverse decoy sequences). For each search engine, search hits from all runs were converted to the pepXML format,[20] combined, and processed through the PeptideProphet[21] algorithm to compute peptide level posterior error probability assignments. To generate a conservative and accurate set of known training and test examples, a 0.95 minimum probability cutoff was applied to the search results. This minimum probability cutoff corresponds to a 0.6, 0.3, 0.3 and 0.3% PeptideProphet estimated false discovery rate for Mascot, X!Tandem, SEQUEST, and OMSSA, respectively. From the combined set of search results from all four search engines that pass the probability cutoff, a unique peptide set was extracted. As noted above, occurrences of the same peptide with different charge and modification states are treated as being unique. The result is a collection of 17 902 unique identifications used for analysis. A 20% subset of these spectra, 3580 entries, was selected randomly and reserved exclusively for testing. This test subset was selected from the overall set such that the proportion of entries for each charge state was maintained. The remaining 80% subset, 14 322 entries, was used for training. The summary of the Orbitrap-LTQ-ETD training and test sets are shown in Table 1.

Table 1. Summary of Orbitrap-LTQ-ETD Training and Test Data Sets

charge	combined	training count	test count
2+	2685	2148	537
3+	8261	6609	1652
4+	4967	3974	993
5+	1584	1267	317
6+	346	277	69
7+	59	47	12
Totals	17902	14322	3580

Table 2. Summary of LTQ-ETD Test Data Sets, Both Total and Unique Peptides

charge	test count nonunique	test count unique
2+	445	251
3+	2840	910
4+	1626	581
5+	374	144
6+	121	46
7+	16	5
Totals	5422	1937

A second independent test data set of *S. cerevisiae* were prepared and acquired on a Thermo LTQ-ETD XL as previously described (see data set 2 in ref 22). Because the precursor charge states of these tandem mass spectra were unknown, confident identifications needed to be obtained first to assign the correct precursor charge state to each spectrum. To obtain the identifications, the data were searched with SEQUEST using the same parameters as above. Each spectrum was analyzed assuming multiple precursor charge states 2+ through 7+ and processed through the PeptideProphet algorithm. The putative identifications were filtered by applying both 0.95 PeptideProphet probability and 0.05 SEQUEST E-value[23] cutoffs, which left no decoy matches passing the filter. After filtering, there were 10 peptide spectrum matches (PSMs) of the same spectrum interpreted as different charge states; a majority of these cases identify the same sequence where the smaller assumed charge state matches a subset of the identified peptide from the larger assumed charge state. These 10 entries were discarded to arrive at 5422 high confidence PSMs. From this set of PSMs, a unique set of peptides (unique peptide, modification, and charge state) were extracted from the full set with the goal of removing repeat/redundant identifications of the same peptide sequence. The full set of 5422 LTQ-ETD spectra and the unique set of 1937 spectra compose the second test set, summarized in Table 2.

Features. Each data point used to train or test a classifier consists of a class label and several features. In our application, the class labels are the precursor charge states and the features are extracted from the set of confidently identified ETD tandem mass spectra. As noted previously, significant diagnostic features from these spectra are the intense charge reduced precursor peaks. In contrast to the CPM tool, which sums and normalizes charge reduced precursor intensities and represents each as a single value for each assumed precursor, we extract the individual intensities associated with each charge reduced precursor and keep them as separate features. By extracting and keeping these features separate, it is hypothesized that the classifier will make use of the distribution of the intensities to improve discrimination. Specifically, the individual charge

Table 3. Twenty-seven Charge Reduced Precursor Features of Which 18 Represent Unique *m*/*z* Values[a]

2+	3+	4+	5+	6+	7+
2++	3+++	4++++	5+++++	6++++++	7+++++++
2+•	3++•	4+++•	5++++•	6+++++•	7++++++•
	3+••	4++••	5+++••	6++++••	7+++++••
		4+•••	5++•••	6+++•••	7++++•••
			5+••••	6++••••	7+++••••
				6+•••••	7++•••••
					7+••••••

[a] The following entries share the same charge reduced precursor *m*/*z* values and are represented only once in the feature table. 2++ same as 3+++, 4++++, 5+++++, 6++++++, and 7+++++++. 2+• same as 4++•• and 6+++•••. 3++• same as 6++++••. 3+•• same as 6++••••.

reduced precursor features are determined by taking the maximum intensity value around each computed *m*/*z*. Each charge reduced precursor intensity value is then normalized to a percentage of the base peak. There are 27 total features associated with charge reduced precursors for charge states 2+ through 7+, as shown in Table 3. However, 13 of the features share some redundancy. For example, 2+• is the same mass-to-charge (*m*/*z*) value as 4++•• and 6+++••• and would extract the same feature value. Removing the redundancies yields 18 unique charge reduced precursor features used for the analysis.

Potential water (H_2O) and ammonia (NH_3) neutral losses from the charge reduced precursor peaks were also used as input features. Similar to the charge reduced precursor features above, we extracted intensities at the expected *m*/*z* for each neutral loss feature, where a single maximum intensity value encapsulates both water and ammonia loss. The neutral loss intensities are also normalized to the base peak. The neutral loss features share no common *m*/*z*, unlike the charge reduced precursor features, so there are 27 neutral loss features. In total, 45 features (18 charge reduced precursor and 27 neutral loss) were extracted from each spectrum and used for classification. Please see Supporting Information for additional information on the feature extraction program.

Classifier. We predict the charge state of a given spectrum using a support vector machine (SVM) classifier. SVMs perform classifications by constructing an N-dimensional hyperplane that optimally separates data into two categories. The data vectors near the hyperplane are termed the support vectors. If the data are not linearly separable, then the original data are transformed into a new, higher dimensional space using a kernel function with the hope that the data are linearly separable in that new space. SVMs have been used previously to successfully predict charge state of CID LC−MS/MS spectra.[24]

The LIBSVM software package[25] was employed for this work and all analysis tools were written in Java. Both linear and radial basis function (RBF) kernels were evaluated, and the RBF kernel was selected for further investigation because of better classification performance. The accuracy of a trained SVM is dependent on the selection of the model parameters. The RBF kernel has two variables, C and gamma, that were optimized for the problem at hand; this procedure is necessary to improve the generalization versus overfitting behavior. The C is a cost parameter that controls the trade-off between minimizing training errors and maximizing the so-called "margin" between the hyperplane and the training data. The parameter gamma determines the RBF width. We used the LIBSVM tool grid.py to determine the optimal values for C and gamma for each trained classifier (see Supplemental Table 1 for a summary of these parameter values for each classifier, Supporting Information). C and gamma values that gave the best 5-fold cross validation accuracy on the training data were chosen for the SVM models.

Because SVM is inherently a binary classification algorithm, we generalize to support multiple classes by training multiple one-vs-all classifiers. Each classifier determines whether a particular spectrum belongs to a particular charge class or not. Six different single charge classifiers were trained, one per charge 2+ through 7+. In the case where none of the single charge classifiers report a class, we fall back to using additional SVMs. One SVM is trained with 2+ and 4+ charged spectra as positive examples and all other spectra as negative examples since these two charge states share a common charged reduced precursor feature (as noted in the footnote associated with Table 3). Another pairwise SVM is trained for charge pairs 3+ and 6+ as these two charge states share two common charge reduced precursor features. A third pairwise SVM is trained for charge pairs 5+ and 7+. Although the 5+/7+ charge states share no features in common, they are grouped together in a pairwise SVM as this classifier is also used to predict a four charge outcome as noted below. Thus, in addition to the six single charge SVMs, we trained three multiple charge SVMs to classify charge states 2+/4+, 3+/6+, and 5+/7+. If none of the single and pairwise charge SVMs confidently assigns a charge state to the spectrum, then the classifier checks if the alternate hypothesis of the 5+/7+ charge SVM passes the classification threshold. If it does, then charges 2+, 3+, 4+, and 6+ are reported because these represent the set of possible charges as the alternate hypothesis to the 5+/7+ classifier. Otherwise all six charge states are reported. The prediction work flow is summarized in Figure 1.

The classifiers were trained using probability estimates. The LIBSVM software estimates probabilities by fitting a sigmoid function that maps SVM outputs to posterior probabilities where the parameters of the sigmoid function are estimated by minimizing the negative log-likelihood function using 5-fold cross validation. A default probability cutoff of 0.98, which is a user setting that can be changed, is applied to the predictions used in the subsequent analysis. For each SVM model, charge state predictions with a probability score greater than the default cutoff are reported.

Results and Discussion

A correct precursor charge prediction is defined as the correct precursor charge being present in the set of one or more predicted charges for a spectrum. One could achieve ~100% prediction accuracy by always returning all six charge states for each query, but this comes at the expense of downstream analysis where each spectrum would require six separate database searches. So there is a trade-off between maximizing the ability to correctly predict the precursor charge for a spectrum (allowing multiple predictions) versus minimizing the number of charge predictions returned by a classifier. In an ideal scenario, one would correctly predict a single charge for each query spectrum. However, in order to maximize actual charge prediction sensitivity, multiple charge state predictions are usually required for at least a subset of the spectra. In fact, the CPM tool recommends that users set their relaxation parameter to 1.75, which generates a target of 1.75 charge state predictions for each input query. In practice, each additional

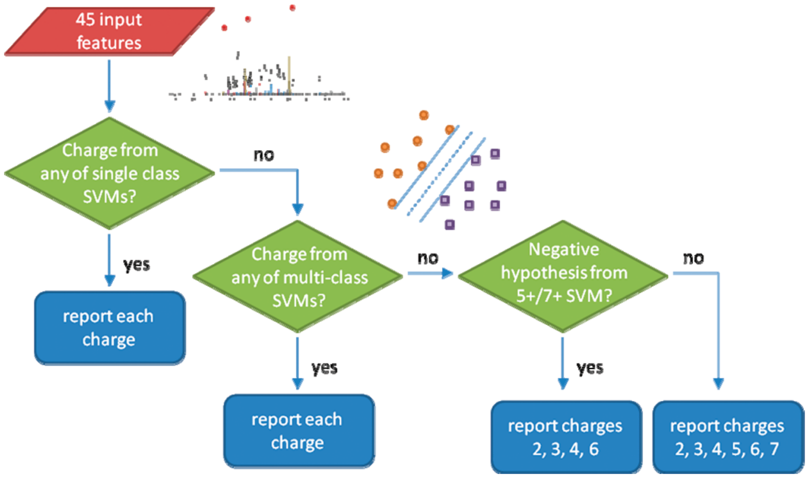
45 input features
Charge from any of single class SVMs?
no
yes
report each charge
Charge from any of multi-class SVMs?
no
yes
report each charge
Negative hypothesis from 5+/7+ SVM?
no
yes
report charges 2, 3, 4, 6
report charges 2, 3, 4, 5, 6, 7

Figure 1. Overall ETD precursor charge state classification scheme. Single charge classifiers are applied first. If no charge state is assigned by any of those classifiers, then the multiple charge classifiers are applied next. If all of those fail to report a charge, then the overall classifier defaults to reporting charges 2+, 3+, 4+, and 6+ based on the negative result from the 5+/7+ charge classifier in the second stage.

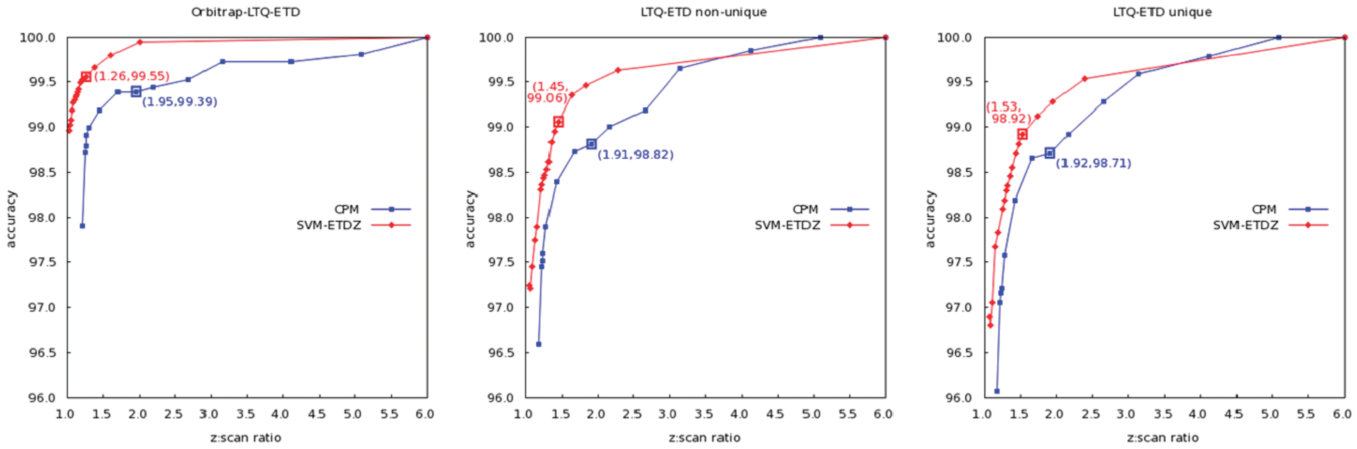

Figure 2. Accuracy versus charge:scan ratio performance comparison of the SVM classifier and CPM on the test data sets. Both tools were run with a range of input options, varying the relaxation parameter in CPM and varying the probability cutoff in the SVM classifier. Performance of the two classifiers at the default settings (1.75 relaxation parameter for CPM and 0.98 probability cutoff for SVM) are highlighted/labeled on the plots.

charge state prediction requires a separate database search, which is the main cost that we are trying to minimize while maintaining prediction sensitivity. Thus, the goal is to achieve accurate, high prediction sensitivity while minimizing the total number of charge state predictions.

To evaluate the performance of our classifier, we apply the SVM models to both the Orbitrap-LTQ-ETD and LTQ-ETD test data sets. These data sets are also input to CPM for comparison (CPM version 1.0.0.11 using model file Charges2−7yeastTrypsinLysC081027.etdc). For the SVM classifier, we vary the probability cutoff from 0.5 to 1.0. Similarly for CPM, we vary the relaxation parameter from 1.0 to 6.0. The performance of the two charge state classifiers is summarized in Figure 2. The plots chart classification accuracy versus “z:scan ratio” (i.e., charge-to-scan ratio or the average number of charge states predicted per spectrum). In this evaluation, optimal performance corresponds to 100% accuracy at a z:scan ratio of 1.0. The performance of the classifiers at their default settings (0.98 probability cutoff for SVM models, 1.75 relaxation parameter for CPM) are noted in the charts as the marked data points with listed coordinate values. Although both tools demonstrate very good performance, this analysis shows that the SVM classifier outperforms CPM with respect to both accuracy and z:scan ratio in all three data sets evaluated. For example, with respect to the Orbitrap-LTQ-ETD data set, the SVM classifier achieved 99.55% accuracy at a 1.26 z:scan ratio whereas CPM achieved 99.39% accuracy at a 1.95 z:scan ratio at default settings. On the LTQ-ETD nonunique data set at default settings, the performance was 99.06% accuracy at 1.45 z:scan ratio for the SVM classifier and 98.82% accuracy at 1.91 z:scan ratio for CPM. And for the LTQ-ETD unique peptide data set at default settings, the performance was 98.92% accuracy at 1.53 z:scan ratio for the SVM classifier and 98.71% accuracy and 1.92 z:scan ratio for CPM. The SVM classifier is able to achieve either higher accuracy at the same z:scan ratio or a lower number of z:scan predictions at the same classification accuracy compared with CPM across a broad range of settings on all test sets.

The performance of the SVM classifier as a function of precursor charge state are shown in Table 4 and Supplemental Tables 2 and 3 (Supporting Information) for the Orbitrap and two LTQ test sets, respectively. The results in these tables demonstrates that the SVM classifier performs well across the range of precursor charge states and is not biased toward superior performance for any specific precursor charge. The

Table 4. Accuracy by Precursor Charge State on the Orbitrap-LTQ-ETD Test Set Using a Probability Cutoff of 0.98

charge	test count	accurate predictions	% accurate	all charge predictions[a]
2+	537	537	100.0%	2
3+	1652	1650	99.9%	10
4+	993	989	99.6%	13
5+	317	312	98.4%	27
6+	69	67	97.1%	3
7+	12	9	75.0%	3

[a] Number of instances that were assigned all six (2+ through 7+) charge states. These instances were not classified at or above the probability threshold by any of the SVM models.

Table 5. Accuracy by Precursor Charge State on Orbitrap-LTQ-ETD Data for PSMs with PeptideProphet Probability Values of 0.5 to 0.8[a]

Charge	Test Count	Accurate Predictions	% Accurate	All Charge Predictions[b]
2+	4684	4677	99.9%	113
3+	6267	6244	99.6%	148
4+	3392	3358	99.0%	64
5+	1294	1250	96.6%	262
6+	286	272	95.1%	25
7+	36	28	77.8%	5

[a] These results reflect precursor charge prediction performance for spectra of lower confidence identifications. [b] Number of instances that were assigned all six (2+ through 7+) charge states. These instances were not classified at or above the probability threshold by any of the SVM models.

results also show that the classifier rarely resorts to reporting all seven charge states for a spectrum, indicating that the single and pairwise charge SVMs are effective in accurate charge predictions. Similarly, Table 5 demonstrates SVM classifier performance on spectra of lower confidence identifications. The SVM training and tests sets derived for the analysis presented in this paper used spectra of very confidently identified peptides in order to obtain accurate precursor charge state information but more importantly to generate training and test sets that did not contain the same duplicate peptide sequences. To test how the SVM classifier performs on less confidently identified peptides of presumably lower quality spectra, spectra of PSMs with PeptideProphet probability 0.5 to 0.8 from the Mascot searches of the Orbitrap runs were isolated and subjected to the SVM classifier. The predicted charge states were compared against the Orbitrap MS1 derived precursor charge states and summarized in Table 5. The z:scan ratio of the charge predictions for this data set is 1.58 with an overall classification accuracy of 99.19%. The results show that the prediction accuracy remains high for this very large set of less confidently identified spectra.

As the z:scan ratio of predicted charge states is lowered, this directly influences subsequent database search times due to having to search a smaller number of predicted charge states. However, there are intuitive but less obvious benefits of minimizing the z:scan ratio of predicted precursor charge states with respect to overall peptide identification performance. It is expected that as the z:scan ratio is reduced, there would be an associated reduction in "distracting" peptide identifications due to the analysis of fewer predicted charge states which in turn would improve overall peptide identifications. To test this hypothesis, the SVM classifier was applied to all spectra in two of the LTQ-ETD runs. For this analysis, ETD precursor charge states from the two runs were predicted using a range of probability cutoffs (0.5 to 1.0) that resulted in a range of z:scan ratios (1.38 to 5.99). These were searched by SEQUEST and both E-value based analysis and PeptideProphet analysis were performed using the same search conditions and analysis parameters as applied previously. A plot of the resulting PSM count vs z:scan ratio is shown in Figure 3 where PSM counts were determined at a 1% FDR. Both analyses exhibit a distinct trend toward a higher number of PSMs as the z:scan ratio is reduced. This shows the clear benefit that minimizing the number of charge predictions has on improving peptide identifications.

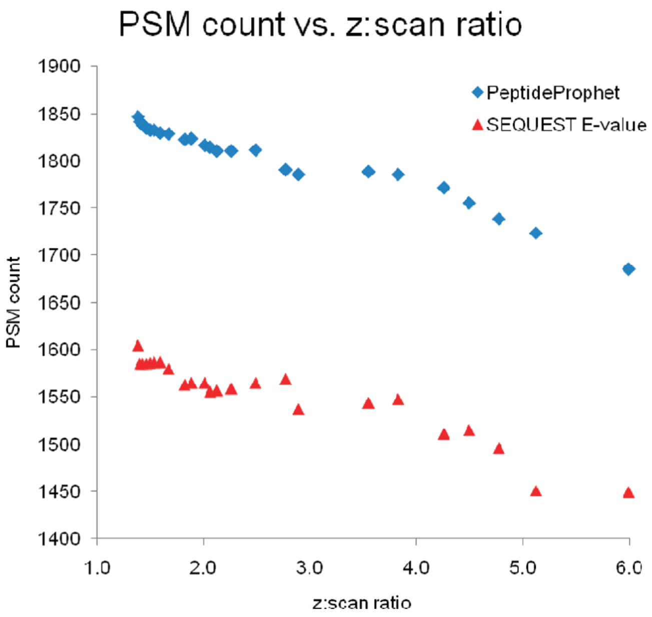

Figure 3. Plot of PSM as a function of z:scan ratio from two LTQ-ETD runs. PSM counts are determined based on SEQUEST E-values and PeptideProphet at a 1% FDR. In both analysis, the number of PSMs tends to increase as z:scan decreases indicating the benefits in minimizing superfluous charge predictions to the PSM identification rate.

There were two primary motivations behind developing a new ETD charge state prediction tool. The first motivation was that the existing tools that perform this function, Charger and CPM, both have various restrictions that did not lend them to be suitable tools for our analysis pipeline. Charger is a commercial Windows program available from Thermo Scientific while our primary analysis is performed on Linux systems using custom and open source tools. CPM is freely available to academic users. However, the source code to CPM is not distributed. It is primarily a Windows tool with Linux command line execution dependent on third party software support (Mono), and more importantly, the software has redistribution restrictions that do not allow it to be incorporated into and distributed with existing freely available and open source pipelines. Thus, we developed this ETD charge state prediction tool so that there is an unencumbered software application that is freely available, can run on multiple operating systems, and can be easily extended from the current Java implementation to other programming languages such as C++, Python, and Perl because of LIBSVM's cross language compatibility. The existing tool takes a collection of spectra in the dta, ms2[26] or mgf formats as input and returns the predicted charge states in a tab-delimited text file, a collection of .dta files, an .ms2 file, or an .mgf file.

The second motivation behind developing this tool was to evaluate the potential for better classification performance than previously demonstrated due to novel use of classification features. As noted above, the goal was to maximize accuracy while minimizing the z:scan ratio. Charger applies a signal processing approach with a fallback to linear discriminant analysis whereas CPM trains a classifier based on Bayesian decision theory and Bayesian discriminant analysis. With respect to the various types of classifiers available to apply to this problem, a major distinction between the performance of such tools is the discriminative quality of the actual underlying training/test features. Although our SVM classifier uses a subset of the features that CPM uses, we keep the individual features separate. The hypothesis is that the distribution of the charge reduced precursor peaks, specifically the set of intensities at each expected charge reduced m/z mass for each hypothesized precursor charge, is more discriminative than summing and then normalizing intensities to a single value for the charged reduced precursor peaks at each hypothesized precursor charge (which CPM does). We were not able to benchmark the Charger program but CPM was compared favorably against Charger in its publication. CPM itself performs extremely well, so demonstrating significant improvements over its performance is difficult. However, we do show that on the test data sets our classifier performs favorably compared to CPM across a broad range of settings. Thus, we have successfully developed a classifier that performs accurate ETD precursor charge state predictions while minimizing the total number of predictions, thereby directly minimizing the number of associated sequence database searches.

Conclusions

We developed an ETD precursor charge state classifier using a collection of SVMs that exhibits high accuracy and sensitivity while maintaining a low total number of overall charge state predictions. Although functional in its current stand-alone implementation, our classifier is meant to be incorporated into software tools and pipelines and work toward this goal has been initiated. Importantly, the tool is not only freely available but it is open sourced with an Apache 2.0 license. Current alternatives do perform very well, as demonstrated by the CPM analysis here, but are either commercial tools or are freely available to academic users but not open sourced and not amenable to being modified or redistributed. Thus we present an ETD precursor charge state classifier that will hopefully address an unmet need in the proteomics community.

Acknowledgment. We acknowledge Ning Zhang at the University of Washington for early prototype analysis, Mike Riffle at the University of Washington for manuscript editing, Henrik Molina at The Centre for Genomic Regulation for helpful discussions on ETD precursor charge states, and Joshua Coon, Danielle Swaney, and Douglas Phanstiel at the University of Wisconsin-Madison for supplying the Orbitrap-ETD data and helpful discussions. This work is supported in part by the University of Washington's Proteomics Resource (UWPR95794) and by NIH awards R01 EB007057 and P41 RR0011823.

Supporting Information Available: Software information and supplemental tables. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (7), 2193–8.

(2) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **2007**, *6* (11), 1942–51.

(3) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (26), 9528–33.

(4) Molina, H.; Matthiesen, R.; Kandasamy, K.; Pandey, A. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal. Chem.* **2008**, *80* (13), 4825–35.

(5) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (7), 2199–204.

(6) Chalkley, R. J.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (22), 8894–9.

(7) McAlister, G. C.; Phanstiel, D.; Good, D. M.; Berggren, W. T.; Coon, J. J. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal. Chem.* **2007**, *79* (10), 3525–34.

(8) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. *5th Annu. ACM Workshop on COLT* **1992**, *14*, 4–152.

(9) Noble, W. S. What is a support vector machine. *Nat. Biotechnol.* **2006**, *24* (12), 1565–7.

(10) Sadygov, R. G.; Hao, Z.; Huhmer, A. F. Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Anal. Chem.* **2008**, *80* (2), 376–86.

(11) Carvalho, P. C.; Cociorva, D.; Wong, C. C.; Carvalho Mda, G.; Barbosa, V. C.; Yates, J. R., 3rd. Charge prediction machine: tool for inferring precursor charge states of electron transfer dissociation tandem mass spectra. *Anal. Chem.* **2009**, *81* (5), 1996–2003.

(12) The Mono Project, http://www.mono-project.com.

(13) Swaney, D. L.; McAlister, G. C.; Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **2008**, *5* (11), 959–64.

(14) 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* (18), 3551–67.

(15) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–7.

(16) Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd. 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*, 976–89.

(17) 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), 958–64.

(18) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004**, *22* (11), 1459–66.

(19) 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), 73–9.

(20) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, 2005–17.

(21) 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), 5383–92.

(22) Deutsch, E. W.; Shteynberg, D.; Lam, H.; Sun, Z.; Eng, J. K.; Carapito, C.; von Haller, P. D.; Tasman, N.; Mendoza, L.; Farrah, T.; Aebersold, R. Trans-Proteomic Pipeline supports and improves

analysis of electron transfer dissociation data sets. *Proteomics* , *10* (6), 1190–5.

(23) Eng, J. K.; Fischer, B.; Grossmann, J.; Maccoss, M. J. A fast SEQUEST cross correlation algorithm. *J Proteome Res* **2008**, *7* (10), 4598–602.

(24) Klammer, A. A.; Wu, C. C.; MacCoss, M. J.; Noble, W. S. Peptide charge state determination for low-resolution tandem mass spectra. *Proc IEEE Comput Syst Bioinform Conf* **2005**, 175–85.

(25) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, accessed 11/1/2009.

(26) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., 3rd. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **2004**, *18* (18), 2162–8.

PR1006685