Anal. Chem. **2009,** *81,* 1996–2003

Charge Prediction Machine: Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra

Paulo C. Carvalho,*,†,‡ Daniel Cociorva,‡ Catherine C. L. Wong,‡ Maria da Gloria da C. Carvalho,§ Valmir C. Barbosa,† and John R. Yates III‡

Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Brazil, Biological Mass Spectrometry Laboratory, The Scripps Research Institute, La Jolla, California, and Laboratory for Control of Gene Expression, Biophysics Institute, Federal University of Rio de Janeiro, Brazil

Electron transfer dissociation (ETD) can dissociate highly charged ions. Efficient analysis of ions dissociated with ETD requires accurate determination of charge states for calculation of molecular weight. We created an algorithm to assign the charge state of ions often used for ETD. The program, Charge Prediction Machine (CPM), uses Bayesian decision theory to account for different charge reduction processes encountered in ETD and can also handle multiplex spectra. CPM correctly assigned charge states to 98% of the 13 097 MS2 spectra from a combined data set of four experiments. In a comparison between CPM and a competing program, Charger (ThermoFisher), CPM produced half the mistakes.

An important element of proteomics is the process of protein identification. A common approach to protein identification involves the use of tandem mass spectrometry (MS/MS). In this process peptide ions are isolated, fragmented using collision-activated dissociation (CAD), and the fragment ions are analyzed to determine their mass to charge ratios (m/z). Tandem mass spectra are analyzed through protein sequence database searching, sequence tag analysis, or de novo interpretation. All of these processes require calculation of an accurate molecular weight for the peptide.[1−3]

Peptides generated by trypsin digestion and ionized by electrospray ionization produce ions of predominately +2 and +3 charge states. Determining the charge state of peptide ions using CAD spectra and other features has been performed by several groups.[4−7] When the charge state cannot be determined it is common for a database search to be performed with a tandem mass spectrum for which a molecular weight has been calculated with both charge states. The correct answer and therefore the correct molecular weight is then determined by examining the sequence matches. This situation increases the computational burden for database searching and can increase the difficultly of correctly assigning tandem mass spectra to sequences, but in general this approach has worked reasonably well.

Electron capture dissociation (ECD) and electron transfer dissociation (ETD) are new methods to dissociate ions in mass spectrometers. A feature common to both of these methods is a preference for highly charged ions, which generally means larger polypeptides or even proteins can be dissociated with these methods. Both methods cleave randomly along the polypeptide backbone and work well when trying to pinpoint the location of posttranslational modifications. When ETD or ECD is used, the charge states of polypeptides observed can run from +2 to +7 and up; thus, the challenge of calculating molecular weight is increased and existing software for CAD cannot be applied. The alternative to accurate calculation of molecular weight of testing all possible calculations for each potential charge state ("all hypothesis search") greatly increases the computational overhead associated with the analysis of ETD/ECD spectra and thus is not desirable. Sadygov et al. developed a method (Charger) using linear discriminant analysis with autocorrelation to decipher charge states associated with ETD.[8] In comparison to analysis with the all hypothesis search, we observed 15% fewer identifications suggesting the method was not fully optimized. We developed a new algorithm, Charge Prediction Machine (CPM), to accurately classify the charge states for ETD spectra. This algorithm is based on Bayesian decision theory and introduces a relaxation parameter to solve cases with low confidence and can deal with multiplexed spectra (different cofragmented ion species). CPM is shown to efficiently classify charge states up to +7 while greatly reducing the number of searches as compared to the all hypothesis searches.

METHODS AND ALGORITHM

Preparation of the Testing (Unlabeled) Data Set. The yeast ETD MS2 testing data set consists of four LC−MS/MS experi-

* To whom correspondence should be addressed. E-mail: paulo@buscario.com.br.

† Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro.

‡ The Scripps Research Institute.

§ Biophysics Institute, Federal University of Rio de Janeiro.

(1) MacCoss, M. J.; Wu, C. C.; Yates, J. R., III. *Anal. Chem.* **2002**, *74*, 5593–5599.

(2) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.

(3) Tabb, D. L.; Saraf, A.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 6415–21.

(4) Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome. Res.* **2002**, *1*, 211–215.

(5) Na, S.; Paek, E.; Lee, C. *Anal. Chem.* **2008**, *80*, 1520–1528.

(6) Magnin, J.; Masselot, A.; Menzel, C.; Colinge, J. *J. Proteome. Res.* **2004**, *3*, 55–60.

(7) Klammer, A. A.; Wu, C. C.; MacCoss, M. J.; Noble, W. S. *Proc. IEEE Comput. Syst. Bioinf. Conf.* **2005**, 175–185.

(8) Sadygov, R. G.; Hao, Z.; Huhmer, A. F. *Anal. Chem.* **2008**, *80*, 376–386.

10.1021/ac8025288 CCC: $40.75
Published on Web 02/09/2009

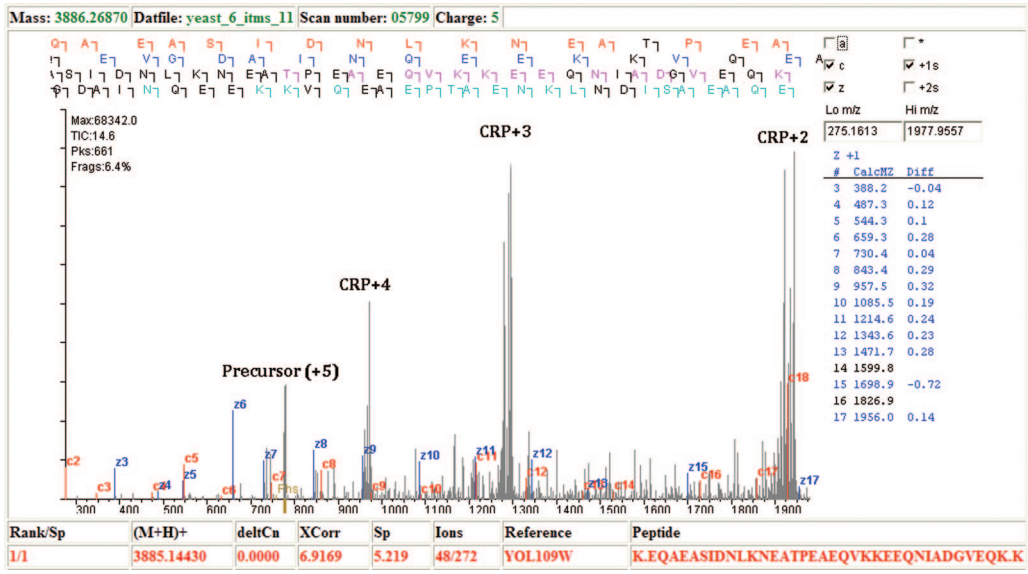

Figure 1. ETD spectrum for +5 charged peptide K.EQAEASIDNLKNEATPEAEQVKKEEQNIADGVEQK.K. The +2, +3, and +4 charge-reduced precursors (CRP), together with the +5 precursor ion, are the dominant peaks in the spectrum. Many spectral peaks can also be noted to the left of the CRPs; these are the neutral losses.

ments. Cells were harvested, lysed, and digested with endoproteinase Lys-C and trypsin as previously described.[9]

A. Chromatography. Capillary high-performance liquid chromatography (HPLC) was performed with an Agilent, Inc. model 1200 quaternary HPLC (Palo Alto, CA). Fused-silica capillary columns (50 μm i.d.) with a 2−3 μm opening and packed with reversed-phase C18 resin were prepared.[10] HPLC buffer solutions were water/acetonitrile/formic acid (95:5:0.1, v/v/v) as buffer A and water/acetonitrile/formic acid (20:80:0.1, v/v/v) as buffer B. The capillary columns were conditioned with buffer A; then digested proteins were pressure-loaded (~5 to 10 μg) onto the column and eluted with a 3 h linear gradient of buffer B from 10−100% into the ion source.

B. Mass Spectrometry. LC−MS/MS experiments were preformed on a linear ion-trap LTQXL-ETD mass spectrometer (ThermoFisher, San Jose, CA) having a chemical ionization source that generates fluoranthene anions[11] for ETD. Mass spectra were acquired using a data-dependent approach where each survey scan (300−2000 m/z) was followed by five ETD MS2 scans of the most intense precursor ions. The automatic gain control (AGC) target for the fluoranthene anion (m/z 202) was set at a value of 100 000. Ion/ion reaction duration was set at 50 ms. Supplemental activation function was applied for all the experiments in this study set. The mass spectrometer scan functions and HPLC solvent gradients were controlled by the Xcalibur data system (ThermoFisher, San Jose, CA).

Preparation of the Training (Labeled) Data Set. The yeast ETD MS2 training data set consists of 12 LC−MS/MS experiments and was acquired in the Coon laboratory; a detailed description is provided elsewhere.[12,13] These ETD tandem mass spectra were acquired on a hybrid linear ion trap-Orbitrap mass spectrometer (ThermoFisher, San Jose, CA). An Orbitrap MS scan was used to provide accurate precursor ion m/z assignment and high resolution to help assign charge state. Data-dependent acquisition of ETD was performed on six precursor ions which were then analyzed in the linear trap.

Precursor Charge Assignment in the Testing (Unlabeled) Data Set. MS2 spectra from the testing data set were extracted using RawExtract[14] and assigned charge states by using an all hypothesis search for charge states +2 through +9 for each spectrum. Since the testing data set was acquired on a low-resolution ion trap instrument, the only way to independently assign precursor charge states was to perform a protein database search in order to match the tandem mass spectra to yeast peptides. Once the peptide spectrum matches (PSMs) were determined, the testing data set was limited to contain only those peptides identified with extremely high confidence (false discovery rate of less than 0.1%).

The tandem mass spectra were searched against a *Saccharomyces cerevisiae* protein database containing 5873 protein sequences, containing the translations of all systematically named ORFs, downloaded as FASTA-formatted sequences from the *Saccharomyces* Genome Database (database released on December 16, 2005), and 123 common contaminant proteins, for a total of 5996 target database sequences. In order to calculate confidence levels and false discovery rates, a decoy database containing the reverse sequences of the 5996 proteins was appended to the target database,[15] and the SEQUEST algorithm was used to find the best matching sequences from the combined database.

(9) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242–247.
(10) Gatlin, C. L.; Kleemann, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R., III. *Anal. Biochem.* **1998**, *263*, 93–101.
(11) Schroeder, M. J.; Webb, D. J.; Shabanowitz, J.; Horwitz, A. F.; Hunt, D. F. *J. Proteome Res.* **2005**, *4*, 1832–1841.
(12) Hubler, S. L.; Jue, A.; Keith, J.; McAlister, G. C.; Craciun, G.; Coon, J. J. *J. Am. Chem. Soc.* **2008**, *130*, 6388–6394.
(13) McAlister, G. C.; Berggren, W. T.; Griep-Raming, J.; Horning, S.; Makarov, A.; Phanstiel, D.; Stafford, G.; Swaney, D. L.; Syka, J. E.; Zabrouskov, V.; Coon, J. J. *J. Proteome Res.* **2008**, *7*, 3127–3136.
(14) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., III. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2162–2168.
(15) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43–50.

SEQUEST searches were done on an Intel Xeon 80-processor cluster running under the Linux operating system. The peptide mass search tolerance was set to 3 Da. Average masses were used for predicted $(M + H)^+$ values in the search, and monoisotopic masses were used for the predicted fragment ions. The mass of the amino acid cysteine was statically modified by +57.0 Da, to take into account the carboxyamidomethylation of the sample. No enzymatic cleavage conditions were imposed on the database search, so the search space included all candidate peptides whose theoretical mass fell within the mass tolerance window, regardless of their tryptic status.

The validity of peptide/spectrum matches was assessed in DTASelect[16,17] using two SEQUEST-defined parameters, the cross-correlation score (XCorr) and normalized difference in cross-correlation scores (DeltaCN). The search results were grouped by charge state (+2 to +9) and tryptic status (fully tryptic, half-tryptic, and nontryptic), resulting in 24 distinct subgroups. In each one of these subgroups, the distribution of Xcorr and DeltaCN values for (a) direct and (b) decoy database hits was obtained, then the direct and decoy subsets were separated by quadratic discriminant analysis. Outlier points in the two distributions (for example, matches with very low Xcorr but very high DeltaCN) were discarded. Full separation of the direct and decoy subsets is not generally possible; therefore, the discriminant score was set such that a false discovery rate of 0.1% was determined based on the number of accepted decoy database peptides. This procedure was independently performed on each data subset, resulting in a false discovery rate independent of tryptic status or charge state.

In addition, a minimum sequence length of seven amino acid residues was required, and each protein on the list was supported by at least two peptide identifications, with a minimum sequence coverage of 5%. These additional requirements resulted in the elimination of most decoy database and false positive hits, as these tended to be overwhelmingly present as proteins identified by single peptide matches or with very low sequence coverage. After this last filtering step, the false discovery rate was estimated to have been reduced to below 0.1%.

Precursor Charge Assignment in the Training (Labeled) Data Set. MS2 spectra from the training data set were extracted using RawExtract[14] and assigned charge states using isotopic information present in the high-resolution full MS scans performed in the Orbitrap analyzer. The Xcalibur software (ThermoFisher, San Jose, CA) was used to assign precursor ion charge states to all spectra. The spectra for which XCalibur did not make an unambiguous charge assignment were removed from the training set. We thus obtained a total of 53 027 charge-assigned tandem mass spectra for the training data set.

The Charge Prediction Machine. CPM uses three MS2 attributes to predict a spectrum's charge state: the complementary fragment ions, the charge-reduced precursors, and the neutral losses. The classification problem is solved in an 18-dimensional feature space having each attribute unfold into six dimensions as described below.

Table 1. Charge State Distribution of the Training and Testing Data Sets[a]

charge	training data set	testing data set
+2	17 764 (33.5%)	7077 (54.2%)
+3	18 048 (34.0%)	4453 (34.1%)
+4	12 368 (23.3%)	1191 (9.1%)
+5	3971 (7.5%)	334 (2.6%)
+6	781 (1.5%)	3 (~0.0%)
+7	95 (0.2%)	0 (0.0%)
total	53 027	13 058

[a] The training data set and the testing data set are composed of 12 and 4 LC−MS assays, respectively. The table shows the charge state distribution accounting for all the runs together.

Before describing each attribute, the notation used throughout is introduced here. Let S be a mass spectrum, understood as a set of spectral "peaks", the ith of which characterized by I_i and m/z_i, respectively, its ion current and its measured mass to charge ratio. The total ion current of S (TIC) is defined as

$$\mathrm{TIC} = \sum_{i=1}^{|S|} I_i$$

Let $\mathrm{PPM}(x, y)$ be a parts per million indicator between two quantities x and y

$$\mathrm{PPM}(x, y) = \left| \frac{10^6 (x - y)}{y} \right|$$

and UPPM a user specifiable parts per million tolerance; such is required for computational purposes. Let H be the mass of a hydrogen atom.

A. The Complementary Ion Feature. CIF is derived from the expectation that fragmenting a +2 precursor ion in CID generates pairs of singly charged complementary product ions whose masses sum up to the mass of the precursor plus that of two hydrogens.[4,18] ETD favors the production of c- and z*-ions, and therefore the above expectation translates into $z_a(m/z_i) + z_b(m/z_j) = z_{\mathrm{precursor}}(m/z_{\mathrm{precursor}}) - H$, constrained by $z_a + z_b = z_{\mathrm{precursor}}$, $i, j \in \{1, 2, \ldots, |S|\}$, and $z_a, z_b \in \{2, \ldots, 7\}$. In practice, accounting for all z_a and z_b combinations that add up to $z_{\mathrm{precursor}}$ increases computational time without substantially increasing discriminatory power. In this regard, CPM focuses on an ordered subset (ς) of combinations that has proven to be effective during the cross-validation tests (data not shown). For the +2 precursor, this subset is $\varsigma_2 = \{+1, +1\}$; accordingly, $\varsigma_3 = \{+1, +2\}$, $\varsigma_4 = \{+2, +2\}$, $\varsigma_5 = \{+2, +3\}$, $\varsigma_6 = \{+3, +3\}$, and $\varsigma_7 = \{+3, +4\}$. In our notation, ς_k, the indexer (k) stands for a hypothesized precursor charge state and $\varsigma_k[j]$ stands for the jth member of the set.

CPM's CIF is computed for six dimensions, resulting in the features CIF_2, CIF_3,..., CIF_7. For the dimension k, let its expected complementary ion sum, similar to the above, be

$$\mathrm{ECS}_k = k(m/z_{\mathrm{precursor}}) - H$$

(16) Cociorva, D.; Tabb, L.; Yates, J. R. *Curr. Protoc. Bioinf.* **2007**, *Chapter 13*, Unit 13.4.

(17) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1*, 21–26.

(18) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327–342.

Table 2. Cross-Validation in Training Data Set Considering Charges 2 through 7 during the Learning Phase[a]

RP	+2	+3	+4	+5	+6	+7	total	RP effectiveness
1	562 (3.2%)	883 (4.9%)	742 (5.3%)	326 (8.2%)	148 (19.0%)	31 (32.6%)	2717 (5.1%)	NA
1.1	373 (2.1%)	511 (2.8%)	487 (3.9%)	159 (4.0%)	64 (8.2%)	17 (17.9%)	1611 (3.0%)	0.21
1.25	295 (1.7%)	372 (2.0%)	361 (2.9%)	112 (2.8%)	37 (4.7%)	11 (11.6%)	1188 (2.2%)	0.11
1.5	248 (1.4%)	268 (1.5%)	282 (2.2%)	92 (2.3%)	32 (4.1%)	7 (7.4%)	929 (1.7%)	0.07
1.75	215 (1.2%)	213 (1.2%)	240 (1.9%)	67 (1.7%)	28 (3.6%)	5 (5.3%)	768 (1.4%)	0.05
2	196 (1.1%)	173 (0.95%)	202 (1.6%)	57 (1.4%)	23 (2.9%)	5 (5.3%)	656 (1.2%)	0.04
2.5	151 (0.9%)	130 (0.7%)	149 (1.2%)	41 (1.0%)	20 (2.6%)	5 (5.3%)	496 (0.9%)	0.03
3	114 (0.6%)	97 (0.5%)	111 (0.9%)	32 (0.8%)	15 (1.9%)	3 (3.2%)	372 (0.7%)	0.02
4	75 (0.4%)	55 (0.3%)	50 (0.4%)	14 (0.3%)	10 (1.3%)	1 (1.1%)	205 (0.3%)	0.02
5	36 (0.2%)	21 (0.1%)	19 (0.2%)	4 (0.1%)	2 (0.3%)	0 (0%)	82 (0.2%)	0.01
6	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.01

[a] The number in each column indicates how many errors were detected for each charge state; the percentage is shown in parentheses. RP stands for the relaxation parameter.

Table 3. Cross-Validation Results Using the Training Data Set and Considering Charges 2 through 6 during the Learning Phase[a]

RP	+2	+3	+4	+5	+6	+7	total	RP effectiveness
1	507 (2.9%)	800 (4.4%)	710 (5.7%)	283 (7.1%)	344 (0.44%)	95 (100%)	2739 (5.1%)	NA
1.1	385 (2.2%)	600 (3.3%)	499 (4.0%)	167 (4.2%)	135 (17.3%)	95 (100%)	1881 (3.5%)	0.16
1.25	288 (1.6%)	349 (1.9%)	329 (2.7%)	109 (2.7%)	60 (7.9%)	95 (100%)	1230 (2.3%)	0.11
1.5	233 (1.3%)	241 (1.3%)	251 (2.0%)	84 (2.1%)	40 (5.1%)	95 (100%)	944 (1.8%)	0.07
1.75	206 (1.2%)	201 (1.1%)	207 (1.7%)	64 (1.6%)	33 (4.2%)	95 (100%)	806 (1.5%)	0.05
2	187 (1.1%)	166 (0.9%)	174 (1.4%)	54 (1.4%)	28 (3.6%)	95 (100%)	704 (1.3%)	0.04
2.5	141 (0.8%)	107 (0.6%)	114 (0.9%)	33 (0.8%)	21 (2.7%)	95 (100%)	511 (1.0%)	0.03
3	106 (0.6%)	74 (0.4%)	66 (0.5%)	18 (0.4%)	15 (1.9%)	95 (100%)	374 (0.7%)	0.02
4	51 (0.3%)	28 (0.2%)	17 (0.1%)	5 (0.1%)	9 (0.1%)	95 (100%)	205 (0.4%)	0.02
5	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	95 (100%)	95 (0.2%)	0.01

[a] The number in each column indicates how many errors were detected for each charge state; the percentage is shown in parentheses. RP stands for the relaxation parameter.

Let RS_k be the subset $\{l, m\}$ of peaks from S for which $PPM(\varsigma_k[1](m/z_l) + \varsigma_k[2](m/z_m), ECS_k) \leq UPPM$. Then

$$CIF_k = \frac{\sum_{j=1}^{|s|} I_j^{RS_k}}{TIC}$$

where $I_j^{RS_k} = I_j$ if $j \in RS_k$ ($I_j^{RS_k} = 0$ otherwise). The expectation is that the precursor's charge will equal the dimension indexer k of the dimension achieving the highest score.

B. The Charge-Reduced Precursor Feature. CRPF is based on evidence that intact charge-reduced precursors (CRPs) are frequently observed in MS2 spectra and could be effectively used for charge determination.[4] Figure 1 shows a typical ETD spectrum of a +5 peptide: the +2, +3, and +4 CRPs are present as dominant peaks in the spectrum, together with the +5 precursor ion itself. We present a methodology to identify CRPs, one that considers a subset (Ψ) of possible CRP charges for a given precursor charge state. Considering all possibilities increases the CRPF computation time without significantly increasing the model's discriminatory power (data not shown). Moreover, some CRPs could lie beyond the detectable m/z bounds. For the +2 precursor, CPM considers the +1 CRPs ($\Psi_2 = \{+1\}$); accordingly, $\Psi_3 = \{+1, +2\}$, $\Psi_4 = \{+1, +3\}$, $\Psi_5 = \{+3, +4\}$,$\Psi_6 = \{+3, +4, +5\}$, and $\Psi_7 = \{+4, +5, +6\}$.

The CRP sets were chosen to minimize the overlap among the expected CRP m/z values, denoted by $C\hat{R}P$, for which an estimate is

$$C\hat{R}P_j \approx \frac{k(m/z_{precursor}) - (k - \Psi_k[j])H}{\Psi_k[j]}$$

for a precursor charge k reduced to $\Psi_k[j]$. For a precursor of expected charge k, let $RS_k \subset S$ be such that $PPM(m/z_l, C\hat{R}P_j) \leq UPPM$ for all $j \in \Psi_k$ and all $l \in S$. Then $CRPF_k$ is given by

$$CRPF_k = \sum_{j=1}^{|S|} \frac{I_j^{RS_k}}{TIC}$$

C. The Neutral Loss Feature. NLF follows from the fact that, in ETD, CRPs frequently lose an uncharged or neutral fragment. CPM's algorithm searches for two common neutral losses, water (H_2O, ~18.02 amu) and ammonia (NH_3, ~17.03 amu), amounting to the value we call neutralLossMass, in subsets (ξ) of CRPs that expectedly lie within the m/z detection bounds. The chosen CRP subsets are $\xi_2 = \xi_3 = \{+1, +2\}$, $\xi_4 = \{+1, +3\}$, $\xi_5 = \{+3, +4\}$, $\xi_6 = \{+3, +4, +5\}$, and $\xi_7 = \{+4, +5, +6\}$. For a given k, let $RS_k \subset S$ be such that $PPM(m/z_l + (neutralLossMass)/(\xi_k[j]), C\hat{R}P_j) \leq UPPM$ for all $j \in \xi_k$ and all $l \in S$, with $C\hat{R}P_j$ defined as above (but on $\xi_k[j]$). Then

$$NLF_k = \sum_{j=1}^{|S|} \frac{I_j^{RS_k}}{TIC}.$$

D. Formalization of the Charge Prediction Machine. CPM uses a supervised learning strategy. The learning process begins

Table 4. CPM Benchmarks in the Testing Data Set Using the Classification Model That Considers Charges +2 through +7[a]

RP	+2	+3	+4	+5	+6	total	time	RP-corrected effectiveness
1	301 (4.3%)	222 (5.0%)	164 (13.8%)	79 (23.7%)	3/3	756 (5.8%)	73	NA
1.25	204 (2.9%)	122 (2.7%)	95 (8.0%)	45 (13.5%)	1/3	467 (3.6%)	75	0.05
1.5	175 (2.5%)	99 (2.2%)	76 (6.4%)	44 (13.2%)	1/3	395 (3.0%)	76	0.04
1.75	153 (2.2%)	75 (1.7%)	60 (5.0%)	39 (11.7%)	1/3	328 (2.5%)	81	0.03
2	133 (1.9%)	64 (1.4%)	42 (3.5%)	37 (11.1%)	1/3	277 (2.1%)	74	0.03
2.5	98 (1.4%)	41 (0.9%)	19 (1.6%)	26 (7.8%)	1/3	185 (1.4%)	76	0.03
3	65 (0.9%)	24 (0.5%)	14 (1.2%)	17 (5.1%)	0/3	120 (0.9%)	70	0.02
4	39 (0.6%)	7 (0.2%)	9 (0.8%)	6 (1.8%)	0/3	61 (0.5%)	77	0.02
5	16 (0.2%)	0 (0.0%)	1 (0.1%)	2 (0.6%)	0/3	19 (0.1%)	74	0.01
6	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	78	0.01

[a] The number in each column indicates the number of mistakes for each charge state, and their percentage. RP stands for the relaxation parameter. The time column displays how many seconds it took CPM to classify all spectra. RP-corrected considers the RP plus the summed values for the multiplex spectra correction.

Table 5. CPM Benchmarks in the Testing Data Set Using the Classification Model That Considers Charges +2 through +6[a]

RP	+2	+3	+4	+5	+6	total	time	RP-corrected effectiveness
1	256 (3.7%)	214 (4.8%)	149 (12.5%)	103 (30.8%)	3/3	725 (5.5%)	61	NA
1.25	170 (2.4%)	113 (2.5%)	90 (7.6%)	56 (16.8%)	2/3	431 (3.3%)	64	0.05
1.5	145 (2.0%)	87 (2.0%)	73 (6.1%)	43 (12.7%)	1/3	349 (2.7%)	65	0.04
1.75	131 (1.9%)	67 (1.5%)	54 (4.5%)	37 (11.1%)	1/3	290 (2.2%)	70	0.04
2	113 (1.6%)	53 (1.2%)	37 (3.1%)	35 (10.5%)	1/3	239 (1.8%)	72	0.03
2.5	83 (1.1%)	33 (0.7%)	16 (1.3%)	25 (7.5%)	0 (0.0%)	157 (1.2%)	65	0.03
3	52 (0.7%)	15 (0.3%)	11 (0.9%)	14 (4.2%)	0 (0.0%)	92 (0.7%)	70	0.02
4	22 (0.3%)	2 (0.0%)	4 (0.3%)	4 (1.2%)	0 (0.0%)	32 (0.2%)	65	0.02
5	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	65	0.01

[a] The number in each column indicates the number of mistakes for each charge state, and their percentage. RP stands for the relaxation parameter. The time column displays how many seconds took CPM to classify all spectra. RP-corrected considers the RP plus the summed values for the multiplex spectra correction.

by computing an input vector ($\vec{x}$) for each mass spectrum in the training data set following the schema $\langle\omega_i\rangle\langle\text{feature}_1\text{:value}_1\rangle...\langle\text{feature}_{18}\text{:value}_{18}\rangle$. In the latter, $\omega_i \in \{2, 3,..., 0.7\}$ stands for the class label (a precursor charge state assigned by a specialist; in our case, a combination of high-resolution Orbitrap MS1 and software); value_1 through value_6, value_7 through value_{12}, and value_{13} through value_{18} correspond to the computed scores for CIF_2-CIF_7, $CRPF_2-CRPF_7$, and NLF_2-NLF_7, respectively. Feature_1 through feature_{18} correspond to the numbers 1 through 18; this is done to comply with the sparse matrix representation schema that is widely adopted in the pattern recognition community and in the PatternLab for proteomics project, of which this tool is part.[19]

CPM adopts, for each class ω_i, the Bayesian discriminant function

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}) - \frac{1}{2}\ln|\Sigma_i| + \ln(P(\omega_i))$$

where $P(\omega_i)$ is the empirically obtained prior probability of class ω_i derived from the respective charge state frequency in the training data set, $\vec{\mu}$ is the mean vector, Σ is the covariance matrix, $|\Sigma|$ is its determinant, and Σ^{-1} is its inverse. CPM stores these variables to disk for quick retrieval and classification of future unseen examples. Classification is performed as follows: for each spectrum, the Bayesian scores are computed, and then remapped according to the following procedure:

$$h_i(\vec{x}) = \max_{\omega_j \in \{2,3,...,7\}}(g_j(\vec{x})) - g_i(\vec{x}), \quad \omega_i \in \{2, 3, \ldots, 7\}$$

All results are associated with their respective spectra and saved in a data structure referred to as the candidate solution array. For example, if there were 100 spectra in the classification data set, and 6 charge states being accounted for, the candidate solution array would have 600 elements. This array is then ordered in a nonincreasing order. Clearly, the most favorable solution for every input vector will obtain a score of 0; solutions with higher scores are associated with a lower expectation of class membership. CPM allows more than one output label per input vector on a case basis by making use of a user-specified relaxation parameter (RP). For example, for an RP of 1.5, the first 150 solutions in our example candidate solution array will be assigned to their corresponding input vectors. Clearly, every input vector will hold their highest expectation solution; the following 50 solutions that are also associated with a high degree of truth will be assigned to their respective spectra/input vector.

E. Accounting for Multiplexed Spectra. Previous work shows that in truly complex samples, closely situated m/z precursors can jointly be selected for fragmentation.[20] This event's

(19) Carvalho, P. C.; Fischer, J. S.; Chen, E. I.; Yates, J. R., III; Barbosa, V. C. *BMC Bioinf.* **2008**, *9*, 316.

(20) Hu, J.; Qian, J.; Borisov, O.; Pan, S.; Li, Y.; Liu, T.; Deng, L.; Wannemacher, K.; Kurnellas, M.; Patterson, C.; Elkabes, S.; Li, H. *Proteomics* **2006**, *6*, 4321–4334.

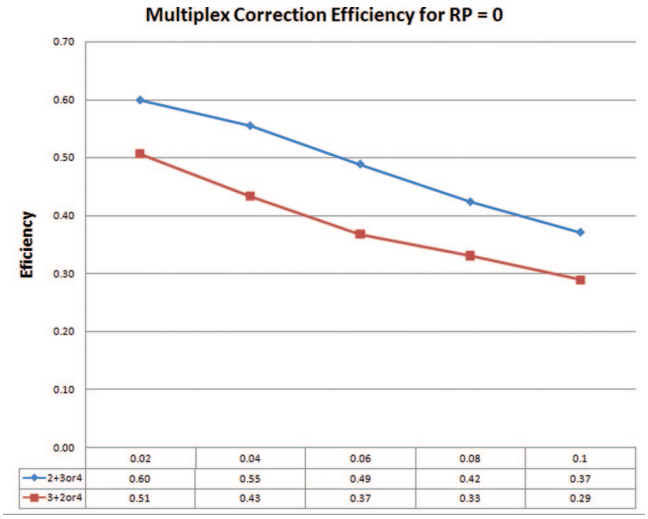

Figure 2. Multiplex correction efficiency for RP = 0. The *x*-axis values show the 2 + 3or4 and the 3 + 2or4 parameter applied to CPM on the testing data set when using a relaxation parameter of 0. The table below discriminates the efficiency values.

frequency can greatly vary from sample to sample (e.g., ~2% to ~10% or more in a single phase analysis of a digested yeast cell lysate; multiplexed spectra were observed in both the training and the testing data set).

Fragmentation usually prevails in one ion species, making the other leave more noticeable charge determination features in the mass spectrum (e.g., CRPs). While the search engine most likely identifies the more abundant peptide ion, CPM assigns a charge state according to the most evident charge features, possibly biased toward the more poorly fragmented ion species. To account for multiplexed spectra, CPM applies a postprocessing correction to potentially multiplexed spectra. The correction searches for +2 precursors that cofragmented with a +3 or +4 precursor (2 + 3or4) and for +3 precursors that cofragmented with a +2 or +4 precursor (3 + 2or4), and includes the extra charge state hypothesis.

The 2 + 3or4 procedure first selects spectra whose precursor ions are less than 850 m/z and have +2 as their most confident charge state. For each of these spectra, it sums the ion current of an expected +1 CRP peak, assuming the precursor charge is +2 (CRP2to1), and accordingly, CRP3to2, and CRP4to3. The CRP values are obtained as described in the Charge-Reduced Precursor Feature section. Afterward, CPM stores the multiplex charge state hypothesis by selecting between the CRP3to2 and CRP4to3 of highest value. Only the hypotheses of highest expectation will be included in the final output as described later.

In example, suppose the CRP3to2 was selected for a given spectrum. CPM then checks whether the +3 charge state hypothesis has not been included during the relaxation procedure and that CRP3to2 is greater than 5% of CRP2to1's value. If both hold true, CPM generates a data structure containing the spectrum's scan number, the +3 charge state hypothesis, the CRP3to2/CRP2to1 ratio, and stores the structure in an array.

If the CRP4to3 had a higher value, an analogous procedure would be performed, but considering the CRP4to3 instead of the CRP3to2 and the +4 charge state hypothesis; the result would be stored in the same array. After accounting for all spectra, the array is sorted in a nonincreasing order, according to the ratio

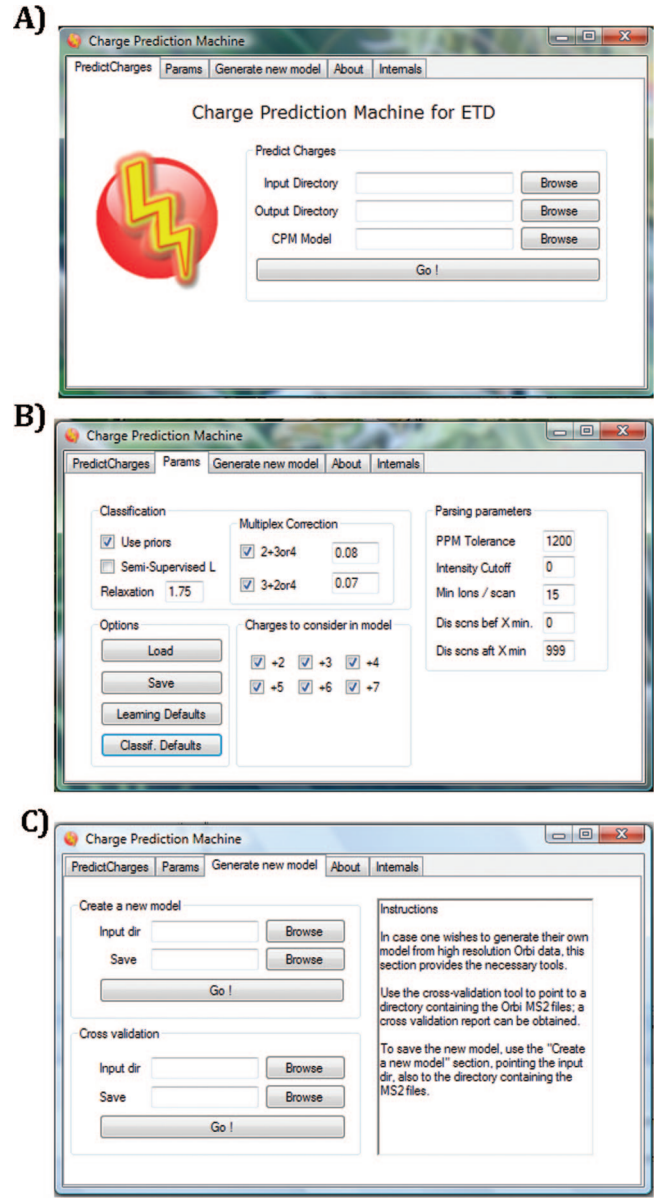

Figure 3. CPM's GUI. (A) CPM's main interface is used to submit a directory containing the MS2 files for charge prediction. (B) The user may control all aspects of CPM, including the relaxation parameter, the multiplex spectra corrections, and what charge states to considered in the Params tabs. The new settings may be saved to disk. (C) The GUI offers tools for creating new classification models and performing cross-validations.

score. Finally, CPM selects from the resulting array the first x hypotheses and includes them in the final charge state prediction output, where x equals is the product of the number of spectra and a user-specified variable named 2 + 3or4Correction. By default, 2 + 3or4Correction equals to 0.075. Clearly, this is analogous to allowing CPM to "relax" an extra 7.5% to account for multiplex spectra of this class.

For the 3 + 2or4Correction, CPM selects spectra derived from ions with an initial hypothesized charge state of +3 and having precursors of less than 1300 m/z. The procedure follows similarly as above, by selecting spectra that pass a minimum cutoff and storing the CRP2to1/CRP3to2 or CRP4to3/CRP2to1 ratios together with the new charge state hypothesis and scan number in an array. The results of highest expectation are selected and

included accordingly. The 3 + 2or4Correction has a default value of 0.075.

F. Computation. All computational procedures were carried out on an HP DV-5 notebook with a T9400 2.5 GHz microprocessor, 3 GB RAM, and Windows Vista. We recommend at least 1.5 GB RAM and using a microprocessor with two or more cores. This is because CPM was programmed using parallel computing libraries to take advantage of the latest multicore technology by distributing the jobs.

RESULTS AND DISCUSSION

The benchmarking was designed to account for laboratory to laboratory variability and tailored to reflect real operating conditions. In this regard, the training (12 LC−MS/MS experiments) and testing (four LC−MS/MS experiments) data sets were acquired in different laboratories (the Coon laboratory and the Yates laboratory, respectively). Their "true" charge state distribution and number of spectra are presented in Table 1. The charge states of peptides in the training data set were assigned using high-resolution Orbitrap MS1 data in the Coon laboratory, whereas the testing data set is a subset of spectra identified only with high confidence in the Yates laboratory. The use of high-confidence peptide identifications to assign charge states has been previously done elsewhere.[5,8]

Cross-Validation Results Using the Training Data Set. The cross-validation (CV) was performed by excluding one of the 12 LC−MS/MS data sets, and using the remaining 11 for training. CPM then predicts the precursor charge states in the spectra from the excluded data set to evaluate the empirical error. This process is repeated for all data sets of the set. Two CVs were performed: one accounting for charges +2 through +7 during the learning phase and the other excluding +7. No multiplex spectra correction was applied. The results from these approaches, tested with various RP setting, are presented in Tables 2 and 3, respectively. Assignment of charge states for spectra with charge states of +2 through +7 performed slightly better when only considering charges +2 through +6 and that the charge state prediction efficiency quickly drops as RP increases. As can be noted in both CVs, CPM achieved an error rate close to 1% while working with a RP of 1.75. The RP efficiency was measured by dividing how many newly and correctly assigned charge states were included by the number of charge state assignments resulting from the RP. Assigned charge states were taken as correct if they matched the charge state assigned using the high-resolution Orbitrap data.

CPM's Benchmarks on the Testing Data Set and Selection of Default Operating Parameters. CPM was benchmarked against the testing data set for various RP values and using two models obtained from the training data set: one trained with charge states +2 through +7 and the other without +7. These results are presented in Tables 4 and 5, respectively. In contrast to the previous analysis described above, the model accounting for charges +2 through +6 achieved a slightly better overall performance. This happens because there are no +7 spectra within the testing data set; therefore, there is one less false hypothesis to account for, and this makes the relaxation procedure seem more efficient. In light of the results from Tables 4 and 5, we empirically determined the default CPM RP to be 1.75 (with a 3 + 2or4Correction of 0.075, and a 2 + 3or4Correction of 0.075); however, we recall that the user can change these values and alter the compromise between search engine time and loss. With these settings, CPM achieved an error rate of 2.2% in the testing data set. The latter can be claimed to be conservative as the DTASelect software that was used to filter SEQUEST results was set to allow a 1% false positive rate in the data set.

The results from the training and testing data set suggest that CPM can adequately generalize between laboratories when operating with an RP of 1.5 or higher. Even though the classification efficiency quickly declines as the RP is increased, in our view it is better to allow a few more false charge state assignments rather than sacrifice the number of identified protein peptides (this is why we set the default RP to 1.75). The gist of CPM's relaxation procedure is to minimize the error rate through global decisions by ordering all charge state hypotheses in the solution array according to scores, so as to choose the most favorable ones within a user-specified relaxation bound.

The multiplex corrections can be interpreted as complementary relaxations that work on orthogonal premises to the RP procedure. An efficiency plot, evaluated similarly to the RP efficiency, is presented in Figure 2 for both multiplex corrections. Indeed, by manually verifying the spectra selected by these filters, we observed characteristics of multiplexed spectra (e.g., 3to2CRPs together with 2to1CRPs, etc.). However, these claims are bound to a specialist's interpretation; as far as we know, there is no reference software to account for such. Both correction efficiencies quickly declined as its user-specified value increased. These corrections were applied after the relaxation procedure; therefore, higher RPs yield even lower multiplex correction efficiency (not shown). The default parameters for both multiplex corrections were set to 0.075 so as to still be effective (∼0.1) when operating in conjunction with suggested RPs (1.75 or 2.0).

Comparison between CPM and Charger. Charger was executed on the same testing set as CPM and assigned 14 500 charge states to 13 058 spectra; 1926 spectra did not match the high-confidence SEQUEST results, yielding a 14.7% error. Since Charger assigned an extra ∼11% charge states, we set CPM's relaxation and multiplex corrections to allow up to an 11% global relaxation for a fair comparison. The RP was set to 1.06, the 3 + 2or4Correction to 0.0025, and the 2 + 3or4Correction to 0.0025. CPM produced a 7.2% mismatch for the same data set, thus, showing a better performance when operating under equivalent conditions for this data set. In our view, even CPM's 7.2% error was still high and this is why we suggest using a higher relaxation parameter of 1.75 as justified above, as to better handle data sets from different laboratories.

The CPM Software. The Charge Prediction Machine is a pattern recognition software programmed in C#. It can be installed with one single click of a mouse in a Windows (XP or Vista)-based PC and only requires the freely available.NET 3.5 or later. In case the user does not have the .NET, CPM can automatically update the computer by interfacing with Microsoft's Web site. Our software can also run on a Linux or Mac, thanks to the Mono project (http://www.mono-project.com). CPM can be executed in the command prompt to provide seamless integration in computational proteomic pipelines; however, it is also user-friendly because it can be executed using a graphic user interface (GUI) as shown in Figure 3. The GUI provides extra functionality such as creation and benchmarking of new classification models. For

example, a laboratory might wish to contribute by publishing a new model if they use different enzymes other than trypsin and Lys-C during sample preparation.

The CPM Windows version carries an automatic update to ensure the use of the latest version. Differently than web-based software, if there was an unpleasing software change, there is a roll-back option which allows CPM to return to its previous state. Taken together, CPM is an easy-to-use, powerful, and flexible tool for assigning charge states to precursors of ETD MS2 spectra.

Availability. CPM, together with the two classification models (charges +2 through +7 and +2 through +6) can be downloaded at the PatternLab for proteomics[19] project Web site or at the Yates Laboratory Web site (http://fields.scripps.edu/cpm); the license is free for academic use. Both classification models were generated using a trypsin and Lys-C digestion protocol followed by analysis in an ETD mass spectrometer. To ensure optimal performance, applying CPM to data using different enzymes or an ECD instrument requires that a model be generated accordingly.

ACKNOWLEDGMENT

The authors acknowledge CAPES, CNPq, a FAPERJ BBP Grant, NIH 5R01 MH067880, NIH P41 RR01, and the Genesis molecular biology laboratory for financial support. The authors thank Dr. Joshua Coon and Dr. Danielle Swaney at the Department of Biomolecular Chemistry, University of Wisconsin, Madison, for contributing with ETD data.

Received for review November 30, 2008. Accepted January 13, 2009.

AC8025288