Genomics Proteomics Bioinformatics 11 (2013) 86–95

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Basophile: Accurate Fragment Charge State Prediction Improves Peptide Identification Rates

Dong Wang [1,#], Surendra Dasari [1,2,#], Matthew C. Chambers [1], Jerry D. Holman [1], Kan Chen [3], Daniel C. Liebler [3], Daniel J. Orton [4], Samuel O. Purvine [4], Matthew E. Monroe [4], Chang Y. Chung [5], Kristie L. Rose [3], David L. Tabb [1,3,6,*]

[1] *Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[2] *Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA*
[3] *Department of Biochemistry, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[4] *Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99354, USA*
[5] *Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[6] *Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*

Received 12 September 2012; revised 3 November 2012; accepted 22 November 2012
Available online 8 March 2013

KEYWORDS

Fragmentation;
Basicity;
Fragment size;
Ordinal regression

Abstract In shotgun proteomics, database search algorithms rely on fragmentation models to predict fragment ions that should be observed for a given peptide sequence. The most widely used strategy (Naive model) is oversimplified, cleaving all peptide bonds with equal probability to produce fragments of all charges below that of the precursor ion. More accurate models, based on fragmentation simulation, are too computationally intensive for on-the-fly use in database search algorithms. We have created an ordinal-regression-based model called Basophile that takes fragment size and basic residue distribution into account when determining the charge retention during CID/higher-energy collision induced dissociation (HCD) of charged peptides. This model improves the accuracy of predictions by reducing the number of unnecessary fragments that are routinely predicted for highly-charged precursors. Basophile increased the identification rates by 26% (on average) over the Naive model, when analyzing triply-charged precursors from ion trap data. Basophile achieves simplicity and speed by solving the prediction problem with an ordinal regression equation, which can be incorporated into any database search software for shotgun proteomic identification.

# Equal contribution.
* Corresponding author.
E-mail: david.l.tabb@vanderbilt.edu (Tabb DL).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2012.11.004

Introduction

Shotgun proteomics relies heavily on database search software for identifying peptides from tandem mass spectra (MS/MS) [1,2]. In general, these algorithms enumerate peptides from a protein sequence database, predict their fragmentation spectra, and match them to the experimental MS/MS. Each peptide-spectrum match (PSM) is scored on the number of peak matches and mismatches between the predicted and experimental MS/MS. Peptides producing high scoring PSMs are assumed to be present in the sample. The accuracy of a search engine's peptide fragmentation prediction model plays a major role in the success of its scoring method. The Sequest search engine [3] introduced the Naive model, which is the most commonly used fragmentation model. This Naive model works under the assumption that all peptide bonds break with equal probability and that each resulting fragment will take on all charges below that of their precursor ion. This model, however, over-predicts the set of fragments expected for each peptide, especially for peptides carrying more than two protons. Therefore, peptide identification can benefit substantially from fragmentation models that generate a set of ions that are most likely to be observed for each sequence.

Several advanced fragmentation models were introduced to improve the prediction accuracy. Kapp et al. [4] and Schutz et al. [5] produced linear regression models for predicting fragment ion intensities. Elias et al. [6] and Arnold et al. [7] applied machine learning techniques to derive a decision tree from a number of peptide and fragment attributes to compute the probability of observing a fragment ion's intensity. In a similar fashion, Frank et al. [8] predicted the intensity ranks of observable peptide fragments. Zhang [9,10] and Sun et al. [11] constructed greatly improved methods that produce realistic MS/MS of a peptide sequence by modeling the gas-phase reaction kinetics and proton mobility. Both machine learning and kinetic models for predicting fragmentation spectra have been shown to be significantly more accurate than *ad hoc* models [12]. These models, however, tend to be too computationally intensive for routine use in database search algorithms that perform billions of PSMs per raw data file.

In this study, we have created a new fragmentation model, Basophile, for accurately predicting the charges of fragments from a peptide sequence. The observable fragmentation pattern depends on four key components: amino acid composition, size of the peptide, precursor charge state and the dissociation method employed [13]. Primary fragmentation of a peptide bond is either a charge-directed process, which involves a mobile proton migrating to the bond, or a charge-remote process, which is determined by the delicate balance between the total number of available protons and the number of proton sequestration sites (basic amino acids) [14–16]. Basophile predicts proton segregation by analyzing the basicity of the N- and C-terminal fragments surrounding a peptide bond. Consequently, Basophile reduces the overall number of fragments predicted for highly-charged ($> +2$) precursors. Basophile was trained and tested with large collections of PSMs aggregated from a variety of CID and higher-energy collision induced dissociation (HCD) data sources, and has been implemented in MyriMatch software [17]. In contrast with machine learning fragmentation models, Basophile is fast, effective and easily brought to bear in database search algorithms.

Results and discussion

MS/MS identification of highly-charged ($> +2$) precursors is problematic

The Naive model has a predilection to over-predict fragments expected for a peptide, especially if its precursor carries more than two protons. For instance, a Naive fragment table for a high-quality $+3$ PSM in **Figure 1** shows 43% of predicted ions unmatched. On an average, 57% of fragments predicted for CID Orbitrap PSMs never matched. Over-prediction rates are worse for HCD PSMs, with 74% of predicted fragments missing from the corresponding MS/MS scan. This increases the probability of peak matching by random chance for low-resolution or data-independent MS/MS of highly-charged peptides because they are often crowded with peaks. False peak matches in turn reduce the discrimination between correct and incorrect PSMs. Also, as the precursor charge increases, the set of peaks is squeezed into the low m/z area (Figure 1**A**), making it more likely that multiple predicted fragments may fall into single m/z bin, raising the possibility that the search engine matches the same observed peak to multiple predicted

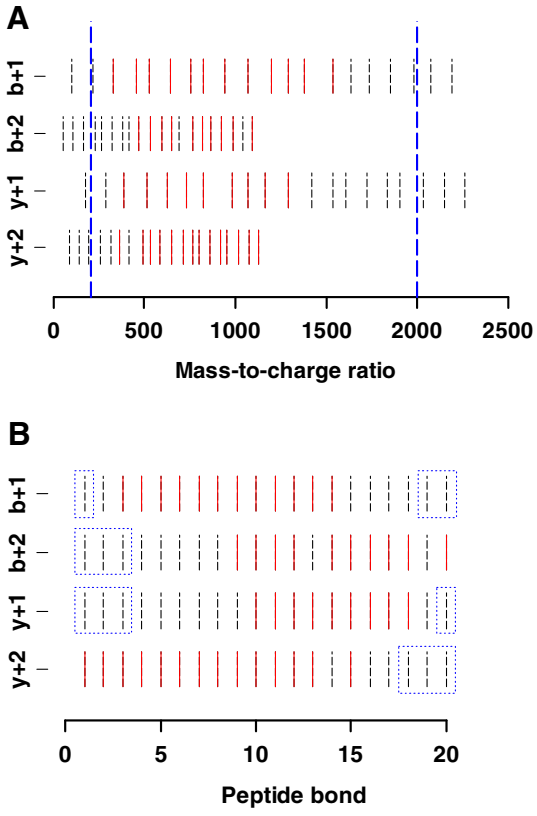

Figure 1 Fragment-peak-matches for a $+3$ CID PSM of "TLLEAIDAIEQPSRPTDKPLR"

Short vertical lines (including long dashed and solid lines) represent a predicted fragment ion in m/z (**A**) or per peptide bond (**B**) under the Naive model. Solid red lines indicate observed fragments. The long blue dashed lines in panel A indicate the scan range of the MS/MS spectrum; the rectangles in dotted lines on Panel B indicate that those ions are out of scan range.

peaks. Figure 1**B** shows patterns of charge segregation. At peptide bonds close to N-term, (b + 1; y + 2) is the dominant fragment pair; At bonds close to C-term, (b + 2; y + 1) is the dominant pair. Near the center of the peptide, the pattern of charge segregation is typically ambiguous. This gradual change is the target of the Basophile model.

Identification rates are correspondingly lower for highly charged peptides (**Figure 2**). Some of the reduced identification is attributable to less informative fragmentation patterns for triply and quadruply charged peptides. If a smaller fraction of peptide bonds is represented by fragment ions in the MS/MS, less information is available for discriminating between good and random matches. The use of fragmentation models that produce excessive fragment predictions, however, worsens matching further.

Constitution of charge segregation events

The Naive model predicts fragments that take on all the charges that are less than the precursor charge, but one fragment of the pair could possibly attract all the protons, leaving the other neutral [26,27]. For example, a + 3 precursor can take four unambiguous charge segregation events as (b + 3), (b + 2; y + 1), (b + 1; y + 2), (y + 3) and three ambiguous ones in between. Attempting to model all seven possible outcomes fails because some of these outcomes are more than ten times more common than others. The rare cases have too little information to establish their boundaries properly. Examinations of fragments from identified CID and HCD MS/MS scans revealed the most common charge segregation events for each precursor class. **Figure 3** summarizes the NIST-CID dataset and Figure S1 summarizes Yeast-Multi-Enzyme-CID and HCD-Orbitrap-Training datasets. Doubly-charged precursors fragment in a manner similar to how the Naive model would predict; with a high percentage of bonds producing two singly-charged fragments. Triply-charged precursors yield three main types of outcomes: doubly-charged N-terminus, doubly-charged C-terminus, or a mix of the two. Quadruply-charged peptides demonstrate that more charges imply more possible outcomes. Basophile training was limited to models of the three most common patterns for +3 (exemplified in **Figure 4**) and the five most common outcomes for +4 peptides.

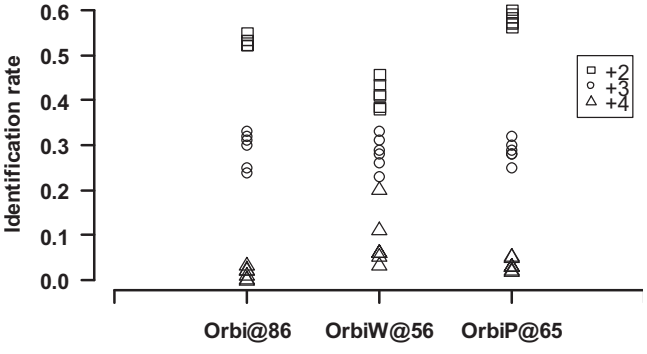

Figure 2 MS/MS of highly charged precursors suffers from low identification rates

MyriMatch identified peptides from the Yeast-CPTAC-CID (LTQ-Orbi) dataset, which featured 6 technical replicates for each of the three instruments. The Naive model was employed to predict fragments for matching. IDPicker filtered the PSMs at 2% *q*-value. Filtered PSMs were segregated by precursor charge state and normalized by the total number of MS/MS acquired with that charge state. MS/MS identification rates dropped dramatically at higher charge states.

Although all three training sources give similar patterns of charge segregation events, HCD-Orbitrap-Training was different from the others in that 36% of all bonds in +3 peptides produced only singly-charged y ions. Initially, these bonds were mapped to the event "b + 2; y + 1," leading to a strong bias toward this segregation event. These bonds, however, could also potentially be mapped to the "b + 1, b + 2; y + 1, y + 2" (ambiguous) or "b + 2; y + 1" categories. In order to associate these low-information bonds with appropriate categories, we developed an adjustment algorithm for +3 HCD peptides. In brief, ordinal labels were assigned, with "y + 1 only" bonds left blank for each peptide. The algorithm then fills the blanks by forcing the list of bonds to a non-decreasing order (*i.e.*, N-terminal basicity category can only increase or stay the same as one moves toward the C-terminus). The detailed algorithm is described in Supplementary File 1. Other fragment evidence sets such as "y + 2 only", "b + 1 only" and "b + 2 only" did not cause trouble during HCD-Orbitrap-Training as they did not trigger bias or comprise a significant fraction of events. A similar phenomenon was found for +4 peptides on HCD-Orbitrap-Training dataset, and a similar adjustment was applied.

Comparison of Basophile models

Three different Basophile models were trained with three diverse collections of PSMs: Basophile-NIST with NIST-CID, Basophile-Yeast with Yeast-Multi-Enzyme-CID and Basophile-HCD with HCD-Orbitrap-Training. Peptides in these three datasets differ in the relative distribution of basic residues and also the dissociation method employed to acquire their MS/MS. Peptides in the NIST-CID and HCD-Orbitrap-Training sets are primarily tryptic, whereas Yeast-Multi-Enzyme-CID contains peptides derived from a variety of digestion enzymes (including Proteinase-K). Also, the first two datasets have low-resolution ion trap CID MS/MS, whereas the last dataset contains high-resolution Orbitrap HCD MS/MS. All models contain two ordinal regression functions, tailored to predict fragmentation spectra for +3 and +4 precursors, respectively.

The standard error (SE) of regression coefficients for all +3 models was all $\leqslant$0.01. However, SEs for +4 Basophile-Yeast and Basophile-HCD models were larger than the corresponding Basophile-NIST model, reflecting the use of much larger spectral library for training with Basophile-NIST. We therefore chose the Basophile-NIST model as the preferred variant. However, it is important to note that the values of coefficients derived from all three training sets followed the same order. For instance, all three +3 regression functions have coefficient magnitudes of Arg > His > Lys > L_N at the N-terminus and Arg > Lys > His > L_C at the C-terminus, indicating that coefficients of all models are similar but on a different scale.

We compared the three Basophile models to the Naive model for peptide identification. To accomplish this, all trained models were implemented in the MyriMatch database

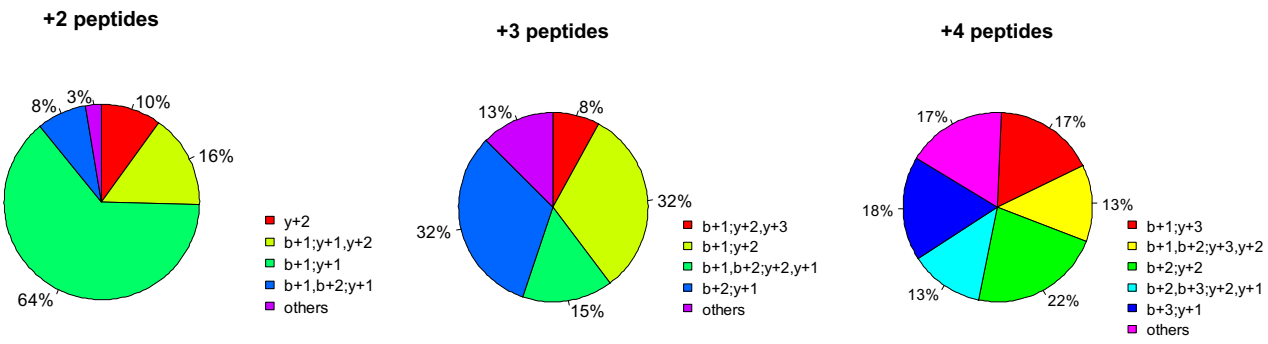

Figure 3 Precursor charge segregation events observed for NIST-CID peptides
PSMs in the NIST human ion trap spectral library were segregated by charge state. Charge states of the observed N- and C-terminal fragments were assessed for all peptide bonds. Frequencies of precursor charge segregation events are summarized here. Label "others" include all ordinal categories which are less than 5% and any other fragment patterns that did not fit a category.

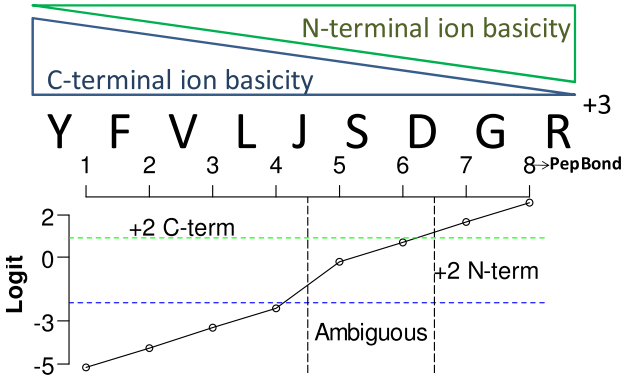

Figure 4 Fragment ion basicity and peptide charge segregation
Progressing from the N-terminus to the C-terminus, the basicity of the N-terminal fragment increases and the basicity of the C-terminal fragment decreases. Below the sequence is the ordinal logit calculated from regression function, and the dashed lines are the two cutoff values to distinguish "+2 C-term", "+2 N-term" and "Ambiguous charge" regions.

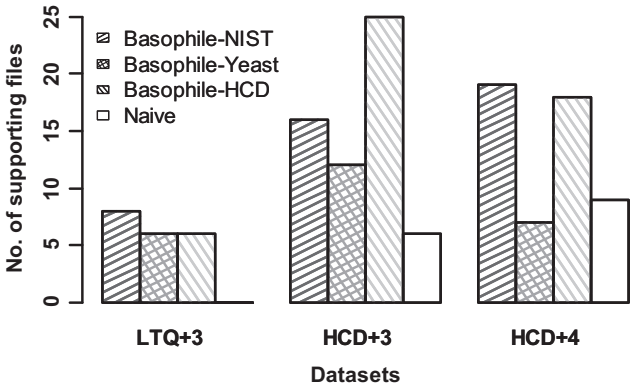

Figure 5 Comparison of Basophile models and Naive model
For each LC–MS/MS experiment, the prediction model that produced the most identifications was given a "vote." Though the HCD-trained Basophile performed well in HCD data, Basophile-NIST performed well across the samples. The Naive model was competitive only in HCD data, reflecting that false positive matching is a smaller detriment in such data.

search engine alongside the Naive prediction model. Searches for each of the four prediction models were run separately on two LTQ datasets (Yeast-CPTAC-CID (LTQ) and Dicty-LTQ) and one HCD dataset (HCD-Orbitrap-Testing) with the standard Multi-Variate Hypergeometrics (MVH) scorer. Figure 5 shows the number of files from the test datasets that "vote" for a particular prediction model by producing the most identified spectra at the same q-value. Basophile-NIST performed slightly better than Basophile-HCD, and both were significantly better than Basophile-Yeast. These results suggested that Basophile-NIST was reasonably robust for modeling HCD fragmentation, even though it was trained on CID spectra.

Basophile reduces fragment peak list size

The ability of Basophile-NIST to reduce the number of fragment predictions was compared to that of the Naive model. Figure 6 shows the number of fragments predicted and matched by the Naive and Basophile-NIST models, grouped by the fragment charge state. Compared to the Naive model, Basophile-NIST reduced the number of fragment predictions by an average of 42% with only slight reductions in the number of matched peaks. A majority of predicted y + 1 fragments (70%) were observed, whereas only a small minority of the predicted b + 2 fragments were matched (13%). This is not surprising because the HCD-Orbitrap-Testing dataset was rich in tryptic peptides that do not produce large numbers of b + 2 fragments; a dataset that enriches peptides with N-terminal basic residues might have matched more of these ions.

In contrast to the SQID model [28], Basophile produces a Boolean output, stating whether a peak is present or absent, rather than a probability associated with matching an experimental fragment. However, it is completely possible to combine the orthogonal SQID and Basophile models into a hybrid system that will not only assess the precursor charge segregation for a peptide bond, but also the likelihood of observing any fragments produced by dissociation of that bond. This method may also reduce the over-prediction further by erasing peptide bonds from the prediction.

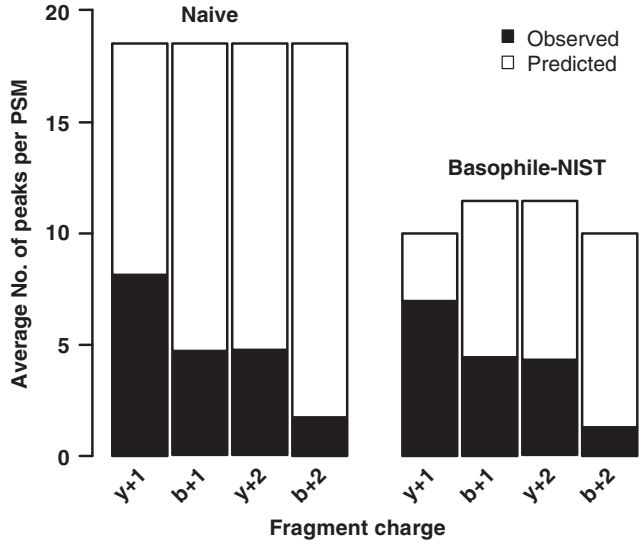

Figure 6 Basophile improves peak prediction accuracy
Basophile reduces the number of fragments predicted for peptide sequences. This reduction has a minimal impact on the number of matched ions for identified peptides, however. For +3 tryptic peptides, the number of matched b + 2 fragments lags behind other classes of fragments.

The reduction of predicted fragments may also prove beneficial to selected reaction monitoring (SRM) experiments. When an SRM is initially designed for an unobserved peptide, a researcher may attempt to monitor all possible fragments that would be produced for it, then reduce the set of fragments screened in further iterations of the SRM assay [29]. The use of Basophile can reduce the size of the initial set of transitions, enabling fewer mass spectral experiments for the first iteration or enabling the screening of a broader collection of peptides in the same number of experiments.

Processing times are frequently substantial since search algorithms process millions of potential peptide sequences, especially when protein databases come from a big proteome, even though this requirement is compromised nowadays by taking use of modern computational technologies such as multi-threading and computer clusters. Basophile naturally reduces the number of fragment ions by predicting a subset of Naive model, thus reducing the number of peaks compared between experimental and theoretical MS/MS. As a result, Basophile reduces search time. We recorded the time used for searches of Yeast-CPTAC-CID (LTQ) dataset with MVH scorer. Searches were performed on 25 cluster nodes, each with two processor cores. In the ten LTQ files, searches using the Naive model took 42 min on average, while searches using Basophile took 30 min. Over-prediction of fragments for peptides can contribute to the additional time required to search datasets.

Effect of the small, but more accurate peak lists on PSM scoring systems

We tested whether the trained Basophile-NIST models could improve peptide identification using the MVH and HGT + RST score systems. By reducing the number of predicted fragments, Basophile could lose identifications; by improving prediction accuracy, Basophile might reduce false positive matching and gain identifications. **Figure 7** compares the number of +3 and Figure S2 compares the number of +4 peptides identified in four testing datasets when MyriMatch employed the Basophile-NIST and Naive models for the search. For LTQ-CID datasets, Basophile-NIST consistently improved the +3 peptide identification over Naive models ($P < 0.01$). However, the Basophile-NIST model failed to improve the peptide identifications when analyzing HCD-Orbitrap spectra. It appears that the high-resolution precursor and fragment masses of HCD MS/MS neutralize any advantage gained from accurate fragment prediction, reducing the number of candidates compared to the MS/MS and hence false-positive matching. We tested this hypothesis by comparing the performance of the Basophile-NIST model on +4 precursor MS/MS present in the HCD-Orbitrap-Testing and Yeast-Multi-Enzyme-CID-trypsin datasets. All spectra were searched using the above mentioned protocol. Basophile-NIST did not significantly outperform Naive on +4 MS/MS in both datasets ($P > 0.05$). Because Basophile attempts to model charge segregation as a function of the full peptide sequence, it may fail to recognize cases in which secondary fragmentation occurs (as in HCD, which resembles triple quadrupole CID more than it does ion trap CID). Also, as Figure 3 shows, +4 peptides typically have more charge segregation events, which are determined by observable b and y ions. This could lead to misclassifications in the training set, however. For example, if the spectrum is really an ambiguous event (b + 1, b + 2, y + 3, y + 2), but we only observed (y + 2, b + 2), then we would mistakenly code it an "unambiguous event". The more the protons carried by a peptide, the higher the chance we erroneously classify it. What is more, we have far fewer training samples for +4 peptides than +3s. These are the two main reasons that +4 identifications are unexpectedly low.

We also tested Basophile-NIST on multiple LTQ–Orbitrap datasets. It turned out that the number of identifications did not consistently gain. For example, the Basophile gained +3 identifications on Yeast-Multi-chymo by 6%, on Yeast-Multi-trypsin by 2%, but lost +3 identifications on Yeast-CPTAC-CID (ORBI) by 3% when we use MVH scorer. As a result, Basophile-NIST did not perform consistently better than the Naive model on LTQ–Orbitrap datasets.

Both MVH and the hypergeometric and rank sum tests (HGT + RST) benefited from Basophile in LTQ dataset for +3 peptide identifications. The average improvement was 30% under HGT + RST system, and 20% under MVH system, indicating that HGT + RST system benefited more from reduced but more accurate predicted fragment list. These findings indicate the interdependence of fragment prediction and PSM scoring systems; a change in one frequently alters performance in the other. Models like Basophile may result in a spectrum being compared to some predictions that are dense with peaks and others that contain relatively few peaks. If a scorer is designed to normalize away these differences by taking into account the density of the spectrum prediction (as is the case for the HGT model), it can benefit from more accurate predictions. In contrast, when a scorer tends to give higher scores on average to predictions that are denser in peaks (as is true for MVH), more accurate predictions may provide less benefit.

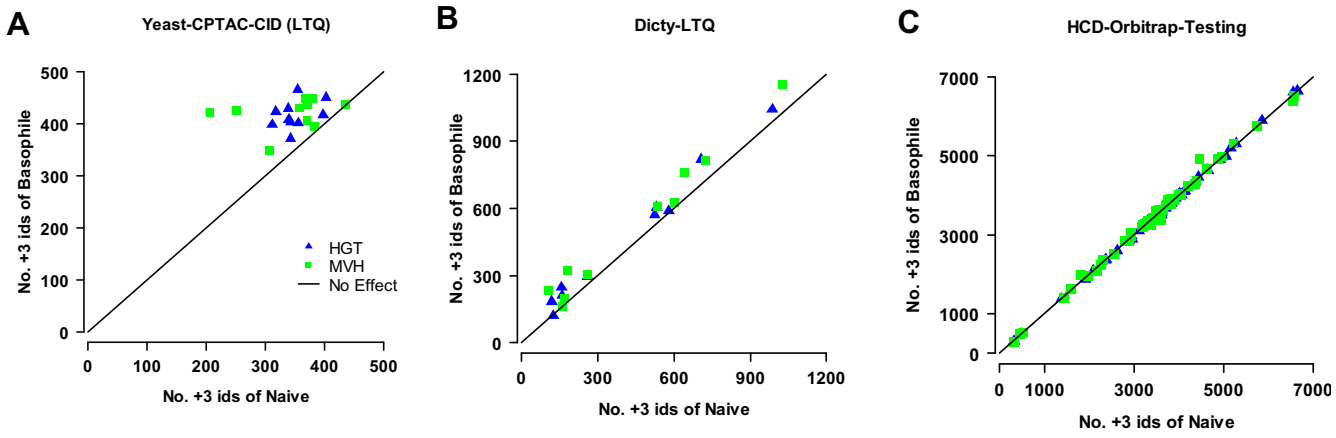

Figure 7 Comparison of Basophile and Naive models on +3 peptides identified
MyriMatch employed Basophile-NIST and Naive models for the search of +3 peptides in two LTQ datasets, Yeast-CPTAC-CID (LTQ) (**A**) and Dicty-LTQ (**B**), and one HCD Orbitrap dataset, HCD-Orbitrap-Testing (**C**). Reduced but more accurate peak list benefits both scorers by improved peptide identifications in low resolution data, but not in high resolution ones. IDs on axes indicate identification of peptides. Legend in panel A applies to all panels.

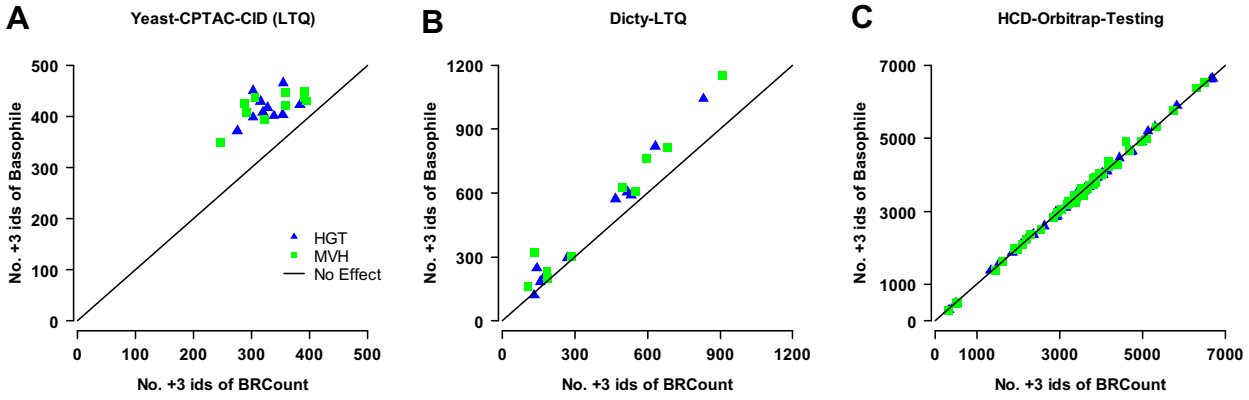

Figure 8 Comparison of Basophile and BRCount model on +3 peptides identified
MyriMatch employed Basophile-NIST and BRCount model for the search. Basophile-NIST outperformed the latter by improved peptide identifications on low resolution data, but failed on high resolution ones.

Comparison of Basophile model and BRCount model

The BRCount model, adapted from a description on the Protein Prospector website (prospector.ucsf.edu/) and communications with Robert Chalkley, predicts maximum fragment charge based on the count of basic sites (including both basic side chains and the N-terminus) contained in a fragment. Both Basophile and BRCount models count basic residues for prediction, but Basophile is trained from identified spectra and incorporates overall fragment length, while BRCount is a simple heuristic.

Basophile-NIST model performed better than the BRCount model when searching +3 precursors (**Figure 8**). Compared to the BRCount model, Basophile-NIST increased the +3 identification rates by 27% ($P < 0.001$) and 36% ($P < 0.01$) when using Yeast-CPTAC-CID (LTQ) and Dicty-LTQ datasets, respectively. However, Basophile-NIST did not outperform the BRCount model when using +3 precursors from HCD-Orbitrap-Testing dataset and +4 precursors from all datasets ($P > 0.05$). Comparisons in quadruply charged peptides are illustrated in Figure S3.

Conclusion

Basophile was designed to rapidly predict peptide fragmentation spectra (m/z values) from sequences that are matched to MS/MS of +3 and +4 precursors. The model improves the accuracy of predictions by reducing the number of unnecessary fragments that are routinely predicted for high charge state precursors. By predicting fewer fragments, Basophile potentially could fail to match observed fragments; by increasing prediction accuracy, Basophile gains identifications by reducing false positive matching. Basophile balances the two forces, making significant improvements for +3 identifications and achieving equivalent performance for +4 identifications compared with the Naive model. Basophile noticeably outperforms the BRCount model consistently in +3 identifications. Basophile also achieves simplicity and speed by solving the prediction problem with an ordinal regression equation that can be easily incorporated into the existing database search software for shotgun proteomic identification.

Table 1 Datasets used in this study

	Data	Species	Instrument	Enzyme	Experiments
Baso-NIST	NIST-CID	*H. sapiens*	Various ion trap	Principally trypsin	703
Baso-Yeast	Yeast-Multi-trypsin[a]	*S. cerevisiae*	Orbitrap	Trypsin	6
	Yeast-Muti-chymo	*S. cerevisiae*	Orbitrap	Chymotrypsin	6
	Yeast-Multi-lysC	*S. cerevisiae*	Orbitrap	Lys-C	45
	Yeast-Multi-proK	*S. cerevisiae*	Orbitrap	Proteinase K	18
Baso-HCD	HCD-Orbitrap-Training	*M. musculus*	Orbitrap Velos	Trypsin	19
	HCD-Orbitrap-Training	*C. elegans*	Orbitrap Velos	Trypsin	12
	HCD-Orbitrap-Training	*E. coli*	Orbitrap Velos	Trypsin	5
	HCD-Orbitrap-Training	*C. griseus*	Orbitrap Velos	Trypsin	94
Testing	Yeast-CPTAC-CID(LTQ)	*S. cerevisiae*	LTQ	Trypsin	10
	Dicty-LTQ	*D. discoideum*	LTQ	Trypsin	10
	HCD-Orbitrap-Testing	*S. oneidensis*	Orbitrap Velos	Trypsin	59
Other	Yeast-CPTAC-CID(ORBI)	*S. cerevisiae*	Orbitrap	Trypsin	18

Note: The "Experiments" column reports the numbers of LC–MS/MS experiments included in each dataset. [a] These data were used for training Basophile–Yeast and testing other Basophile models.

Materials and methods

Datasets

We gathered a diverse collection of peptide fragmentation spectra (MS/MS) for training and testing the Basophile model. **Table 1** summarizes the datasets used in this study. Detailed sample processing protocols are included in Supplementary File 2. The RAW data files are available from the EDRN Catalog and Archive Service (http://cancer.jpl.nasa.gov/ecas/).

NIST-CID

For this dataset, we used the November 29, 2011 version of the human ion trap spectral library from the National Institute of Standards and Technology (NIST). This library contains representative CID-MS/MS spectra for more than 190,539 distinct peptides collected from human samples [18]. A majority of the candidates (68%) in the library are tryptic peptides, which include the following numbers of precursors by charge state: +2: 165 K, +3: 85 K and +4: 30 K.

Yeast-CPTAC-CID

Yeast whole cell lysates were previously analyzed at Vanderbilt University (Nashville, TN) as part of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) initiative [19,20]. Proteins from the lysates were reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAA) and digested with trypsin. Peptide mixtures were subjected to replicate LC–MS/MS analyses using either an LTQ or an LTQ–Orbitrap mass spectrometer (Thermo-Fisher, Waltham, MA). A total of 262 and 42 K CID-MS/MS were collected from LTQ and LTQ–Orbitrap analyses, respectively.

Yeast-Multi-Enzyme-CID

Proteins from yeast whole cell lysates were reduced with DTT and alkylated with IAA. The protein mixture was apportioned into four aliquots, each of which was digested with trypsin, chymotrypsin, lys-C or proteinase-K, respectively (the individual dataset was then named as Yeast-Multi-trypsin, Yeast-Multi-chymo, Yeast-Multi-lysC and Yeast-Multi-proK, respectively). Resulting peptide mixtures were analyzed independently in replicates on an LTQ–Orbitrap mass spectrometer using LC–MS/MS at Vanderbilt University. A total of 664 K CID-MS/MS spectra were collected from all analyses.

Dicty-LTQ-CID

Membrane proteins were extracted from cultured *Dictyostelium discoideum* cells, reduced with DTT, alkylated with IAA and digested using porcine trypsin. The 10 different peptide mixtures were analyzed on an LTQ-XL mass spectrometer in LC–MS/MS analyses at Vanderbilt University. A total of 169 K CID MS/MS spectra were collected.

HCD-Orbitrap

A diverse collection of HCD MS/MS spectra was assembled from 5 different samples: *Mus musculus* brain tissue, *Caenorhabditis elegans* cells, *Escherichia coli* cells, *Cricetulus griseus* cells, and *Shewanella oneidenesis* MR-1 (formerly *Shewanella putrefaciens*) cells. Chi et al. [21] analyzed the *C. elegans* and *M. musculus* samples at the National Institute of Biological Sciences (Beijing, China). *E. coli* cells were analyzed at Vanderbilt University's Mass Spectrometry Research Center. Baycin et al. [22] analyzed the *C. griseus* cells at Johns Hopkins University (Baltimore, MD). *S. oneidenesis* MR-1 cells were analyzed at Pacific Northwest National Laboratory (Richmond, WA). Sample processing protocols are detailed in Supplementary File 2 and summarized here. In brief, proteins from these samples were reduced with DTT, alkylated with IAA and digested with trypsin. Peptide mixtures were subjected to replicate LC–MS/MS analyses using LTQ–Orbitrap mass spectrometers located at the respective institutions. A total of 211, 105, 16, 855 K and 1.19 million HCD MS/MS spectra were collected from *M. musculus*, *C. elegans*, *E. coli*, *C. griseus* and *S. oneidenesis* MR-1 samples, respectively. Data from the first four samples

were used to train the Basophile-HCD model (''HCD-Orbitrap-Training''), whereas *S. oneidenesis* MR-1 data were reserved for testing Basophile (''HCD-Orbitrap-Testing'').

Raw data produced by the mass spectrometers were transcoded into either mzML or mz5 format using the msConvert tool of the ProteoWizard library [23] for further processing.

Peptide identification and results filtering

MS/MS spectra were identified using MyriMatch database search software. A complete list of MyriMatch search parameters are presented in Table S1. MyriMatch was configured to derive semitryptic peptides from the sequence database while looking for the following variable modifications: carbamidomethylation of cysteine (+57.0125 Da), oxidization of methionine (+15.996 Da), and formation of pyro-glutamic acid from N- terminal glutamines (−17.0265 Da). When modified, the software used the Basophile model for augmenting the theoretical MS/MS predictions for +3 and +4 precursors.

MyriMatch matched peaks between experimental and predicted MS/MS. Resulting PSMs were scored with three different systems: MVH, HGT and RST. The MVH system segregates experimental peaks into three intensity classes and measures the point probability of matching a given combination of peaks by random chance using a multivariate hypergeometric distribution [17]. The HGT system employs a hypergeometric distribution to measure the *P* value of obtaining more than the observed number of peak matches between the predicted and experimental MS/MS by random chance [18]. The RST system ranks experimental MS/MS peaks by increasing order of intensity, computes the intensity rank sum of peak matches, and estimates the *P* value of obtaining a better rank sum by random chance via a normal distribution [18]. MyriMatch was configured to sort the spectrum matches using either the MVH point probability or a *P* value derived from combining HGT and RST scores via Fisher's method. The software produces peptide identifications in standard pepXML formatted files. IDPicker [24] filtered peptide identifications from all searches at a q-value [25] of 2% using either MVH score or an optimized combination of HGT and RST scores.

Pattern of charge segregation events for highly charged peptides

Basophile was trained to predict fragment charge segregation for highly charged precursors. Three different models were trained using high-quality peptide identifications derived from ''NIST-CID'', ''Yeast-Multi-Enzyme-CID'' and ''HCD-Orbitrap-Training'' datasets (Table 1). Evidence of observed fragment ions for a PSM can be grouped in terms of charge segregation. Peptide bonds close to the N-terminus produce longer y ions than b ions; similarly, y ions near the N-terminus are likely to contain more basic residues than b ions. These two factors imply that y ions near the N-terminus compete more strongly for the protons that ionized the intact peptide. Conversely, when fragmentation occurs near the C-terminus, the b ions are longer and contain more basic residues. We separated the possible outcomes from charge segregation into regions of unambiguous and ambiguous charge segregations. For example, a +3 precursor can produce four unambiguous charge segregation outcomes: a triply-charged y ion (y + 3), a doubly-charged y ion and singly-charged b ion (b + 1; y + 2), a singly-charged y ion and doubly-charged b ion (b + 2; y + 1), and a triply-charged b ion (b + 3). For some peptide bonds, both outcomes may result; for example, a peptide bond may produce both singly and doubly-charged b and y ions. For +3s, three ambiguous regions fall between the four unambiguous outcome regions. Because these outcomes are not all equally spaced for peptides, we opted to emphasize only the most common charge segregation outcomes in Basophile, as discussed in subsection ''Constitution of charge segregation events.''

Ordinal regression-training of Basophile

Peptides from raw MS/MS data (Yeast-Multi-Enzyme-CID and HCD-Orbitrap-Training) were identified with MyriMatch software configured to use MVH score as primary sort order for matches. IDPicker filtered the resulting peptide identifications at a stringent 2% *q*-value. PSMs were grouped by precursor charge state and peptide sequence (including modifications). We selected the highest scoring MS/MS from each group for training.

Ordinal regression is a classification algorithm that deals with data with multiple outcomes, which models the probability of observing a positive outcome by using a sigmoid function: $p = h(x) = \frac{1}{1+e^{-\beta x}}$, where βx denotes the product of vectorized coefficients and factors. This is equivalent to $\text{logit} = \log(\frac{p}{1-p}) = \beta x$. This logit function then gives cutoff values to discriminate neighboring ordinal outcomes. Given a peptide cleavage site, Basophile computes the logit value (logarithmic odds) of observing a charge segregation event using an ordinal logistic regression function: $\log(\frac{p}{1-p}) = \beta_1 R_N + \beta_2 H_N + \beta_3 K_N + \beta_4 L_N + \beta_5 R_C + \beta_6 H_C + \beta_7 K_C + \beta_8 L_C$, where R_N, H_N, and K_N are number of Arg, His and Lys residues in N-terminal fragment; R_C, H_C, and K_C are number of Arg, His, and Lys residues in C-terminal fragment; L_N and L_C are number of other residues at N- and C-terminals, respectively.

Two training tables (one each for +3 and +4 precursors) were generated from the above PSMs of each dataset by custom software. Each row of the table corresponds to a peptide bond in a PSM. The row summarizes the counts of residues (R_N, H_N, K_N, R_C, H_C, K_C, L_N and L_C) for each peptide bond as well as the set of fragment ions observed in the MS/MS. The table generator removed noise peaks from the spectra using a 95% total ion current (TIC) threshold filter [17]. Having located the set of fragment ions for a given bond from the MS/MS spectrum, the software maps the fragment evidence to an ordinal label to describe the charge segregation outcome region. For example, if y ions from a bond of a triply-charged peptide were observed in both singly and doubly-charged form, the software would map this bond to a charge ambiguity region where both termini were capable of attracting two of the three protons. Table S2 presents a complete list of charge segregation events and evidence of observed fragment ions monitored for +3 and +4 precursors, while Table S3 presents a sample training table generated from triply-charged PSMs.

We employed ordinal logistic regression to process each training table and derive an ordinal logit function for predicting fragment charge states from the fragment basicity. A 5-fold cross-validation strategy was used to avoid over-fitting of the

function to the data. The regression provided weights for the basicity calculation function and decision table to predict which segregation region best models a given peptide bond. Table S4 presents the detailed ordinal functions of the trained models and the corresponding decision values. We implemented these advanced fragment prediction models for +3 and +4 precursors in MyriMatch alongside the Naive model. For comparison, we implemented a simple basic residue count (equal weights) based fragment charge state predictor in MyriMatch (BRCount model). This model allows the fragments to take on any charge below that of their precursor and less than or equal to the number of basic residues in that fragment. Searches with BRCount model followed the same pipeline. MyriMatch can be instructed at run time to apply a particular model for the database search.

Testing the efficacy of Basophile

High resolution precursor and fragments in the "HCD-Orbitrap-Testing" dataset were utilized to measure the efficacy of Basophile in reducing the number of fragments predicted for peptides. The MS/MS of +3 and +4 precursors were identified with the MyriMatch database search engine configured to use the Naive model for MS/MS prediction and MVH for results ranking. IDPicker filtered the resulting peptide identifications with MVH score to a stringent 2% *q*-value. Custom software in the C# programing language inspected each PSM, independently recapitulated the fragment predictions using Naive and Basophile-NIST models, matched the predicted fragments to experimental peaks and assessed the number of fragment hits and misses by each fragment charge state.

Authors' contributions

DLT supervised the project. DW developed the methodology, trained and tested Basophile. SD provided data sources for CID-NIST datasets, and provided the new scoring system, HGT and RST, in MyriMatch. Both authors contributed to project design. MCC provided experimental technical support. JDH provided manuscript review. The other authors provided datasets for training and testing. The first draft of the manuscript was drafted by DW, and SD rewrote several sections and contributed additional text. DLT completed final revisions. All authors read and approved the final manuscript.

Competing interest

The authors declared no conflict of interest.

Acknowledgements

This work was supported by the National Library of Medicine training grant (Grant No. 5T15LM007450-10). The authors gratefully acknowledge public datasets provided by Mengqiu Dong (NIBS) and Deniz Baycin (JHU).

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.11.004.

References

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
[2] Washburn MP, Wolters D, Yates 3rd JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001;19:242–7.
[3] Eng J, McCormack A, Yates III J. 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.
[4] Kapp EA, Schütz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, et al. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal Chem 2003;75:6251–64.
[5] Schütz F, Kapp EA, Simpson RJ, Speed TP. Deriving statistical models for predicting peptide tandem MS product ion intensities. Biochem Soc Trans 2003;31:1479–83.
[6] Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 2004;22:214–9.
[7] Arnold RJ, Jayasankar N, Aggarwal D, Tang H, Radivojac P. A machine learning approach to predicting peptide fragmentation spectra. Pac Symp Biocomput 2006;11:219–30.
[8] Frank AM. Predicting intensity ranks of peptide fragment ions. J Proteome Res 2009;8:2226–40.
[9] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 2004;76:3908–22.
[10] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem 2005;77:6364–73.
[11] Sun S, Meyer-Arendt K, Eichelberger B, Brown R, Yen CY, Old WM, et al. Improved validation of peptide MS/MS assignments using spectral intensity prediction. Mol Cell Proteomics 2007;6:1–17.
[12] Li S, Arnold RJ, Tang H, Radivojac P. On the accuracy and limits of peptide fragmentation spectrum prediction. Anal Chem 2011;83:790–6.
[13] Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev 2005;24:508–48.
[14] Jones JL, Dongre AR, Somogyi A, Wysocki VH. Sequence dependence of peptide fragmentation efficiency curves determined by electrospray ionization/surface-induced dissociation mass spectrometry. J Am Chem Soc 1994;116:8368–9.
[15] Dongré AR, Jones JL, Somogyi A, Wysocki VH. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. J Am Chem Soc 1996;118:8365–74.
[16] Wysocki VH, Tsaprailis GT, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. J Mass Spectrom 2000;35:1399–406.
[17] Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 2007;6:654–61.
[18] Dasari S, Chambers MC, Martinez MA, Carpenter KL, Ham AJ, Vega-Montoto LJ, et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. J Proteome Res 2012;11:1686–95.
[19] Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. TagRecon: high-throughput mutation identifi-

cation through sequence tagging. J Proteome Res 2010;9: 1716–26.
[20] Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography–tandem mass spectrometry. J Proteome Res 2010;9:761–76.
[21] Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, et al. PNovo: de novo peptide sequencing and identification using HCD spectra. J Proteome Res 2010;9:2713–24.
[22] Baycin-Hizal D, Tabb DL, Chaerkady R, Chen L, Lewis NE, Nagarajan H, et al. Proteomic analysis of Chinese hamster ovary cells. J Proteome Res 2012;11:5265–76.
[23] Kessner D, Chambers MC, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24:2534–6.
[24] Holman JD, Ma ZQ, Tabb DL. Identifying proteomic LC–MS/MS data sets with Bumbershoot and IDPicker. Curr Protoc Bioinformatics 2012; Chapter 13:Unit13.17.
[25] Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 2008;7:40–4.
[26] Paizs B, Suhai S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. Rapid Commun Mass Spectrom 2002;16:1699–702.
[27] Paizs B, Suhai S. Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage. J Am Soc Mass Spectrom 2004;15:103–13.
[28] Li W, Ji L, Goya J, Tan G, Wysocki VH. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. J Proteome Res 2011;10:1593–602.
[29] Prakash A, Tomazela DM, Frewen B, Maclean B, Merrihew G, Peterman S, et al. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J Proteome Res 2009;8:2733–9.