BASOPHILE: ACCURATE FRAGMENT CHARGE STATE PREDICTION IMPROVES PEPTIDE IDENTIFICATION RATES

By

DONG WANG

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

For the degree of

MASTER OF SCIENCE

In

Biomedical Informatics

December, 2012

Nashville, Tennessee

Approved:

Professor David L. Tabb

Professor Daniel C. Liebler

Professor Bing Zhang

CONTRIBUTION STATEMENT

The manuscript based on this master thesis has been submitted for publication in a peer-viewed journal with co-first authors of Dong Wang and Dr. Surendra Dasari, a former postdoc at Tabb's lab who is now working at Mayo Clinic as a Research Associate. Dr. Surendra Dasari provided significant help on this project. Dr. Dasari provided the CID-NIST data sources that were previously published in a peer-reviewed journal (see reference 19). He provided the new scoring system, HGT and RST, in MyriMatch, which is the next generation scoring system that may replace MVH. The whole Basophile fragmentation model was tested on the HGT and RST system. He spent significant time revising the manuscript, and some of its text also appears in this thesis. He was very helpful in providing technical support throughout the project.

ACKNOWLEDGMENT

I am grateful to the members of my Thesis Committee, Dr. David L. Tabb, Dr. Daniel C. Liebler, and Dr. Bing Zhang, who have provided me extensive professional guidance in scientific research. I would especially like to thank Dr. David L. Tabb, the chairman of my committee. This project would not have been possible without his support. I am indebted to Dr. Surendra Dasari, who devoted his time and efforts to writing source code for the new scoring system in MyriMatch, revising and reviewing the manuscript to make it for submission to peer-viewed journals after moving to Mayo Clinic. I am thankful to Matt C. Chambers for experimental technical support. I also gratefully acknowledge public datasets provided by Mengqiu Dong (National Institute of Biological Sciences), Deniz Baycin (Johns Hopkins University), and Daniel C. Liebler (Vanderbilt University).

This work was supported by NLM (National Library of Medicine) training grant 5T15LM007450-10. I am grateful to Dr. Cynthia S. Gadd, the Director of Educational Programs, and Rischelle Jenkins, the training program coordinator at Department of Biomedical Informatics. They've done incredibly well in creating an environment in which every trainee feels proud. I am also thankful to all of those in the department with whom I have had the pleasure to work during this project.

Nobody has been more important to me than the members of my family. I would like to thank my parents, my wife and my beloved daughter who is now a third grader. I love you all.

TABLE OF CONTENTS

Chapters Page

I. Introduction 1

II. Materials and Methods 5

Data sets 5

Peptide identification pipeline 10

Pattern of charge segregation events for highly charged peptides 12

Ordinal regression 13

Training of Basophile 14

Testing the prediction efficacy of Basophile 17

III. Results and Discussion 18

Naïve model is problematic for highly charged peptides 18

Constitution of charge segregation events 20

Comparison of Basophile models 24

Basophile reduces fragment peak list size 27

Effect of the small, but more accurate, peak lists on PSM scoring systems 29

Comparison of peptides under Naïve or Basophile 32

Comparison of Basophile and ppBasicity models 33

IV.Conclusion 36

References 37

LIST OF TABLES

Table 1. Data sets used in this study....5

Table 2. MyriMatch search configurations....10

Table 3. Charge segregation events for +3 and +4 peptides....16

Table 4. Sample training table including two peptides from NIST-CID data set....16

Table 5. Coefficients of predictor variables by ordinal regression....25

Table 6. Comparison of Naïve / Basophile exclusive peptide identifications in nine raw files of Yeast-CPTAC-CID (LTQ) data set. 33

LIST OF FIGURES

Figure 1. The flow chart of Basophile 11

Figure 2. Threshold perspective of ordinal regression..13

Figure 3. Fragment ion basicity and peptide charge segregation 14

Figure 4. Fragment-peak-matches for a typical CID +3 PSM of peptide "TLLEAIDAIEQPSRPTDKPLR" 19

Figure 5. MS/MS of highly charged precursors suffers from low identification rates 20

Figure 6. Precursor charge segregation events observed for basophile training peptides 23

Figure 7. Comparison of Basophile models and Naïve model 26

Figure 8. Basophile improves peak prediction accuracy 28

Figure 9. Comparison of Basophile and Naïve models on +3 and +4 peptides identified..............31

Figure 10. Overlap of +3 peptide identifications from Naïve model and Basophile model 32

Figure 11. Comparison of Basophile and ppBasicity model on +3 and +4 peptides identified 35

CHAPTER I

Introduction

Liquid chromatography based tandem mass spectrometry (LC-MS/MS) has become dominant in large-scale identification of proteins in complex biological samples [1-2]. In tandem MS (MS/MS), particular ions (precursors) are selectively passed through the first mass analyzer to generate intact peptide ions in the gas phase by protonation. Then the mass-selected ions pass through a reaction region where they are activated to fall apart to produce fragment (product). The *m/z* values of the dissociation products are then recorded by the second mass analyzer. The resulting “MS/MS” spectrum consists only of product ions from the selected precursor [3]. The most common way to excite the precursor ion is energetic collisions with a nonreactive gas, such as helium, and is referred to as collision-induced (activated) dissociation (CID or CAD). Alternative peptide fragmentation methods such as higher-energy collision dissociation (HCD) have consistently achieved significance [4].

Tandem MS relies highly on database search algorithms to identify peptides from tandem mass spectra where they enumerate peptides from the particular protein sequence database, predict their fragment ions, and match them to the experimental MS/MS spectra. It is widely accepted that the accuracy of the fragment prediction model plays an important role for database search algorithms [5]. The predicted theoretical spectrum must be sufficiently similar to the observed experimental spectrum in order for the identification to succeed. However, it is often difficult to make such predictions accurately due to the complex nature of peptide fragmentation. The most common model (Naïve model), introduced with Sequest [6], assumes that each peptide bond breaks with equal probability and each resulting fragment takes on all charges below that of the precursor ion. While this identification approach works well for most peptides, several peptides exhibit fragment ions that differ greatly from this ideal model, yielding low or insignificant scores, thus preventing automated positive identification [7]. Naïve model over-predicts the set of fragments expected for each peptide, especially for precursor peptides carrying more than two protons. Because data-dependent methods isolate ions of a particular peptide prior to fragmentation, this over-prediction is tolerable. In data-independent sets, however, tandem mass spectra are crowded with the fragments of many peptides, leading to a heightened potential for false-positive matching. Secondly, for highly charged precursors, multiple fragments may be predicted at the same *m/z*, double-counting any "hits". Lastly, as precursor charge increases, the rate of successful identification falls. Identifying peptides in data-independent sets will benefit substantially from fragmentation models that generate the set of ions most likely to be observed for each sequence.

There are more advanced and complicated statistical fragmentation models that were introduced recently, which typically dealt with peak intensities or intensity ranks. Kapp *et al* [8] and Schutz *et al* [9] produced linear regression models for predicting fragment ion intensities. A kinetic model was described by Zhang [10, 11] to produce realistic MS/MS of a peptide sequence based on the classical theory of reaction kinetics and the mobile proton model of peptide fragmentation. Machine learning approaches were used by Elias *et al* [12] and Arnold *et al* [13] with a probabilistic decision tree to model the probability of observing the fragment ion intensity, conditioned on a number of different peptide and fragment attributes. In similar fashion, Frank *et al* [14] predicted the intensity ranks of observable peptide fragments. Machine learning approaches were found to be generally more accurate than kinetic models in predicting fragmentation spectra, and both models are significantly more accurate than the ad-hoc models [15]. However, the intensity-based prediction models can become less accurate and more computationally intensive with large peptides and higher charge states. For example, the machine learning models by Frank to predict peak ranks worked well for singly and doubly charged peptides, but produced lower performance in triply-charged peptides primarily because the dynamics of the fragmentation pathways in triply-charged peptides are more difficult to predict given these peptides are longer and contain more basic amino acids [14]. It would be even worse for CID and HCD MS/MS with charges +4 and +5 or higher. Secondly, all these models are computationally too intensive for on-the-fly use in database search algorithms that process millions of candidate sequences. For example, Elias modeled the probability of observing fragment ion intensity conditioned on 63 different peptide and fragment attributes. Kinetic model [10, 11] included 236 parameters for doubly charged peptides which were thought to be important. The complexity undermines the feasibility and transferability among different search engines. In database search, the software designer must always be mindful of impacts on running time. For example, although the ByOnic database algorithms implemented a machine learning fragmentation model, it has to use heuristic-based rules for predicting fragment ion ranks to reduce computational complexity [14]. Thirdly, the intensity-based models only works when the scoring function of database search algorithms include intensity as part of it. Popular search engines like Mascot do not include a scoring system that deals with theoretical fragment intensities, and thus would not benefit from these advanced models. Altogether, it is necessary to create simple, fast, effective, and transferable fragment prediction systems that can be routinely brought to bear in common database searching algorithms.

The observed fragmentation pattern depends on various parameters including the amino acid composition and size of the peptide, excitation method, the charge state of the ion, etc. [3]. According to the “mobile proton” model [16-18], the proton(s) added to a peptide upon excitation will migrate to various protonation sites prior to fragmentation provided they are not sequestered by a basic amino acid side chain, hence amino acid composition (the absence or presence and type of a basic residue) plays deterministic roles on fragmentation efficiency. In this study, we created a new fragmentation model, Basophile, to accurately predict the charges of fragments based on the number of basic residues and the size of fragment for highly charged peptides. Complementary to CID model, we have alternative fragmentation strategies (HCD) to potentially improve identification of long, highly-charged peptides, and peptides containing multiple basic residues, since longer peptides containing one or more internal basic residues are poorly fragmented by CID [4]. Basophile was trained and tested with large collections of peptide-spectrum-matches (PSMs) aggregated from a variety of CID and HCD data sources, and has been implemented in MyriMatch software [19].For comparison, we have also implemented Protein Prospector’s prediction model (ppBasicity for short in this manuscript) [20] in MyriMatch. This model allows fragment to take on any charge state below that of precursor and up to the number of basic residues of that fragment. MyriMatch can be instructed at run time to apply a particular model (Naïve, Basophile or ppBasicity) for the database search. In contrast with more complicated fragmentation models, Basophile is fast, effective, and easily brought to bear in database search algorithms.

CHAPTER II

Materials and Methods

Data Sets

We gathered a diverse collection of peptide fragmentation spectra (MS/MS) for training and testing the Basophile model. Table 1 summarizes the data sets used in this study. Detailed description of data protocols are listed below.

Table 1. Data sets 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*	*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

*: these data were used for training Basophile-Yeast and testing other Basophile models.

1. NIST-CID. We downloaded the human ion trap library (on 11/29/2010) from the National Institutes of Standards and Technology (NIST) website http://peptide.nist.gov. This library contains representative CID-MS/MS spectra for 190,539 distinct peptides collected from human samples [21]. A majority (68%) of the candidates in the library are tryptic peptides. NIST-CID has a total of 165,499 distinct +2 peptides, 85,018 distinct +3 peptides, and 30,475 distinct +4 peptides.

2. Yeast-Multi-Enzyme-CID (Vanderbilt University). Proteins from yeast whole cell lysates were mixed with 0.1 ml 100mM Ambic and then 0.1ml TFE. Samples were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide (IAA). Protein mixture was apportioned into four aliquots and digested with one of the following enzymes: trypsin, chymotrypsin, lys-c, or proteinase-K (the individual data set was then named as Yeast-Multi-trypsin, Yeast-Multi-chymo, Yeast-Multi-lysC and Yeast-Multi-proK respectively). Then the digest was desalted with C18 SPE catridge and peptides were eluted with AcCN/H2O 80/20. The digest were then re-dissolved in 0.1 % Formic acid, and diluted to 200 ng/ul. 3ul of each digest was loaded into nanoLC/MS (ORBI) with 3 hr gradient LC profile. A total of 664,698 CID-MS/MS spectra were collected from all aliquots.

3. HCD-Orbitrap-Training. A diverse collection of HCD MS/MS spectra was assembled by combining shotgun proteomics data from five different samples: *M.musculus* brain tissue, *C. elegans* cells, *E. coli* cells and *C. griseus* cells.

1) *E.Coli* cells were analyzed at the Vanderbilt University's Mass Spectrometry Research Center (Nashville, TN). Lyophilized *E. coli* cells (1mg dry weight, Sigma EC11303) were lysed by addition of 200uL 500mM Tris (pH 7.5) with 50% trifluoroethanol (TFE). The *E. coli* lysate was reduced with 10mM tris(2-carboxythyl)phosphine (TCEP), and Cys residues were carbamidomethylated with 25mM iodoacetamide. The lysate was then diluted 5-fold with 100mM Tris and digested with 10ug trypsin (proteomics-grade, Sigma) at 37℃ overnight. Following digestion, peptides were desalted by solid-phase extraction (Sep-pak light C18 cartridge, Waters). First peptides were acidified by 2-fold dilution in 0.1% TFA, the peptide solution was loaded via syringe onto a preconditioned C18 cartridge, the cartridge was washed with 0.1% TFA, and peptides were

eluted with 60% acetonitrile/0.1% TFA. Three sequential 500uL elutions were performed, and eluates were dried by speed-vac concentration. *E. coli* peptides were reconstituted in 500uL 0.1% formic acid, generating a 2ug/uL solution. An aliquot of the desalted and concentrated digest was then diluted 1:15 in 0.1% formic acid, and this diluted solution was used for analysis via LC-coupled tandem mass spectrometry (LC-MS/MS). For five replicate experiments, 3uL (0.4ug) of *E. coli* digest was injected onto a capillary reverse phase analytical column (360µm O.D. x 100µm I.D.) using an Eksigent NanoLC Ultra HPLC and autosampler. The analytical column was packed with 20cm of C18 reverse phase material (Jupiter, 3 µm beads, 300Å, Phenomenox), directly into a laser-pulled emitter tip. Peptides were gradient-eluted at a flow rate of 500nL/min, and the mobile phase solvents consisted of 0.1% formic acid, 99.9% water (solvent A) and 0.1% formic acid, 99.9% acetonitrile (solvent B). The mobile phase gradient consisted of the following: 0-15 min, 2% B (during sample loading segment); 15-60 min, 2-40% B; 60-68 min, 40-90% B; 68-72 min, 90% B; 72-75 min, 90-2% B; 75-85 min, 2% B. Gradient-eluted peptides were mass analyzed on an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific), equipped with a nanoelectrospray ionization source. The instrument was operated using a data-dependent method with dynamic exclusion enabled. Full scan (m/z 400-2000) spectra were acquired with the Orbitrap (resolution 60,000), and the top 8 most abundant ions in each MS scan were selected for fragmentation via higher-energy collision induced dissociation (HCD). The precursor ions isolated for the 8 HCD MS/MS spectra per duty cycle were selected in order of least to most abundant, and a minimum signal of 4x103 was required to trigger MS/MS. An AGC target of 4x104, a maximum ion injection time of 500msec, an isolation width of 2 m/z, and 30% normalized collision energy were used to generate HCD

MS/MS spectra. Dynamic exclusion settings included a repeat count of 1, the exclusion list size was set to 500, the exclusion duration time was 60sec, and an exclusion mass width of 10ppm relative to the reference mass was applied. Singly-protonated precursor ions or those with unassigned charge states were rejected for MS/MS analysis, and monoisotopic precursor selection was enabled. A total of 16,492 HCD MS/MS spectra were collected from all aliquots.

2) *C.elegans* and *M.musculus* samples were analyzed at the National Institute of Biological Sciences (Beijing, China) [22]. *C.griseus* cells were analyzed at Johns Hopkins University (Richmond, WA) [23]. 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 211,788 HCD MS/MS spectra were collected from *M.musculus*, 105150 from *C.elegans*, and 855,745 from *C.griseus*.

4. Testing of Basophile and others

1) Yeast-CPTAC-CID (Vanderbilt University). Yeast whole cell lysates were previously analyzed at Vanderbilt University as part of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) initiative [24, 25]. In short, proteins were reduced with DTT, alkylated with IAA, and digested with trypsin. Peptide mixtures were subjected to either an LTQ (Yeast-CPTAC-CID(LTQ)) or an LTQ-Orbitrap (Yeast-CPTAC-CID(ORBI)) mass spectrometer (Thermo-Fisher, Waltham, MA). A total of 262,568 and 42,478 CID-MS/MS were collected from LTQ and LTQ-Orbitrap analyses, respectively.

2) Dicty-LTQ (Vanderbilt University): Dictyostelium cells (Ax3, racC-, and nlp/slp-cells) were cultured axenically in HL5 medium supplemented with 60 units of

penicillin and 60 mg of streptomycin per ml. Cells competent to chemotaxis toward cAMP (aggregation-competent cells) were obtained by pulsing cells in suspension (5 X 106 cells/ml) for 5 hrs with 30 nM cAMP. Cells were then lysed by passing through a 5 μm pore sized filter. This filtrate was then spun down at 30,000 g for 45 minutes at 4ºC to generate the membrane and soluble fraction. Membrane pellets were dissolved in 0.5% n-Dodecyl-β-D-Maltopyranoside. Membrane proteins were precipitated by adding 1/4 Vol 100% TCA and washed two times with 100% acetone. Then the samples were reduced, alkylated, and analyzed on an LTQ-XL mass spectrometer. A total of 169,021 CID MS/MS spectra were collected from all aliquots.

3) HCD-Orbitrap-Testing (Pacific Northwest National Laboratory). Shewanella oneidensis MR-1 samples were cultured in-house then digested with trypsin and analyzed by LC-FTICR using a fully automated, custom-built, four-column capillary LC system coupled online using an in-house manufactured ESI interface to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The capillary columns were made by slurry packing 3 μm Jupiter C18 bonded particles (Phenomenex, Torrence, CA) into a 35-cm long, 75- μm i.d. fused-silica capillary (Polymicro Technologies, Phoenix, AZ). Mobile phase A consisted of 0.1% formic acid in water and mobile phase B consisted of 0.1% formic acid in acetonitrile. Aliquots of each peptide sample were injected onto the reversed-phase column for LC-MS analysis. Mobile phase A was maintained at 100% for 20 min after which the composition was changed to 80% B over a 100 minute gradient. High mass accuracy spectra were collected via an orbitrap analyzer, and the six most intense peaks in the previous MS spectrum were then selected for high-resolution HCD MS/MS analysis. A total of 1,189,175 HCD MS/MS spectra were collected from all aliquots.

Peptide Identification pipeline

Raw data produced by the mass spectrometers were transcoded into either mzML or mz5 format using the msConvert tool of the ProteoWizard library [26]. Database search engine, MyriMatch (v2.1.119) was used in this study. MyriMatch was configured to derive semi-tryptic peptides from the sequence database with 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). The detailed settings are listed in table 2.

Table 2. MyriMatch search configurations.

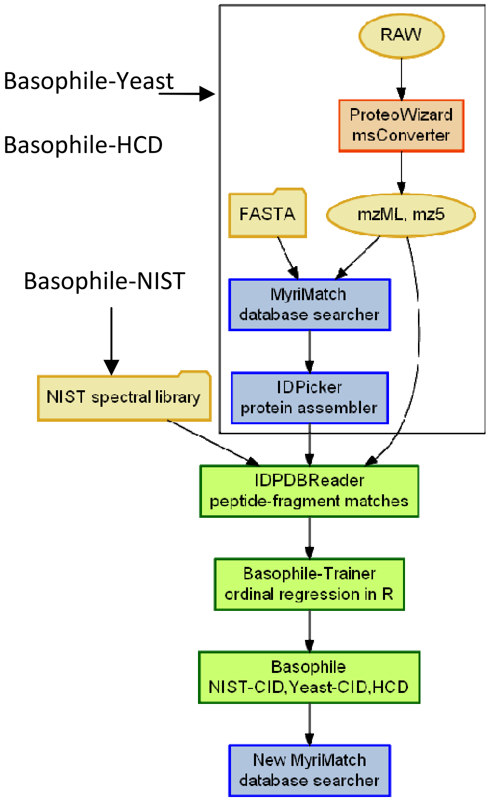
```
FragmentationAutoRule =  true
PrecursorMzToleranceRule = "auto"
MonoPrecursorMzTolerance = "10 ppm"
AvgPrecursorMzTolerance = "1.25 mz"
FragmentMzTolerance = "0.5 mz" *
PredictionModel            = "naive"#
SpectrumListFilters = "peakPicking true 2-;chargeStatePredictor false 4 2 0.9"
MonoisotopeAdjustmentSet = "[-1,2]"
TicCutoffPercentage = 0.95
MaxPeakCount = 150
CleavageRules =  "trypsin" $
MaxMissedCleavages =  2
MinTerminiCleavages =  1
DynamicMods = "M ^ 15.994915 (Q * -17.026549"
MaxDynamicMods = 3
StaticMods = "C 57.021464"
MaxResultRank = 2
MinPeptideLength =  5
UseSmartPlusThreeModel = false
NumChargeStates = 4
```

*: set 0.5 *mz* for LTQ / ORBI searches, and 10 ppm for HCD searches
#: qualified models are “naïve”, “basophile”, and “ppbasicity”.
$: charge to appropriate digestion rules

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. 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. The RST system ranks experimental MS/MS peaks by decreasing 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. MyriMatch was configured to sort the PSMs using either the MVH point probability or a p-value derived from combining HGT and RST scores via Fisher's Method. The software produced peptide identifications in standard pepXML formatted files. The IDPicker software (v3.0.433) [27] was used to filter peptide identifications at a q-value [28] of 2% using either MVH score or an optimized combination of HGT and RST scores. The results were written into idpDB file, which is a SQL database file. Figure 1 summarizes the flow of Basophile.

Figure 1. The flow chart of Basophile.

ProteoWizard is a software tool set which includes msConverter to convert the raw file into appropriate data format for database search. MyriMatch is a database search engine. IDPicker is a parsimonious protein assembler. IDBDBReader is a customized C# application for database file (idpDB) query, peptide analysis, and peptide fragment ions retrieval. Basophile-Trainer is R package by the author which includes 5-fold cross-validation and ordinal regression. Basophile-Yeast and Basophile-HCD started with raw data source, whereas Basophile-NIST was directly from human ion trap library.


Basophile-Yeast
Basophile-HCD
Basophile-NIST
RAW
ProteoWizard msConverter
FASTA
mzML, mz5
MyriMatch database searcher
NIST spectral library
IDPicker protein assembler
IDPDBReader peptide-fragment matches
Basophile-Trainer ordinal regression in R
Basophile NIST-CID,Yeast-CID,HCD
New MyriMatch database searcher

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” data sets (Table 1) and named Basophile-NIST, Basophile-Yeast, Basophile-HCD accordingly. 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 likely 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, behavior that bridges to adjacent 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 below in “Constitution of Charge Segregation Events.” By this rule, Basophile would only allow fragments to retain charges less than that of precursor charges.

Ordinal Regression

Ordinal regression is used to build models, generate predictions, and evaluate the importance of various predictor variables in cases where the dependent variable is ordinal in nature. The simplest ordinal data are those with two categories of outcome, Yes or No (for example), which can be analyzed by a binary logistic regression. We could imagine drawing a random number from a logistic distribution. If the number is above a threshold, the corresponding decision is a Yes, if it's less than the threshold, it's a No. In ordinal regression, we have more than 2 categories, and just like logistic regression, we then have multiple thresholds to distinguish one category from another. For example, in **Figure 2** right panel, A, B, and C are ordered categories. We have two thresholds for two logistic regressions: one for “A or B”, and one for “B or C". Mathematically, this reduces to a set of logistic regressions with different intercepts.

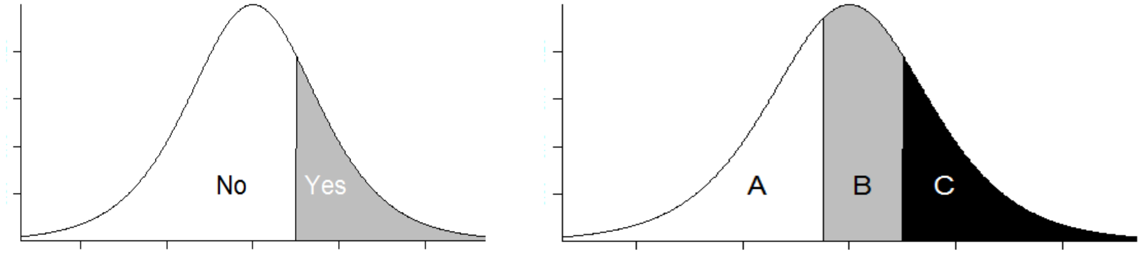


Figure 2. Threshold perspective of ordinal regression. The left panel is a logistic regression which deals with two categorical outcomes; the right panel is ordinal regression which decides among more than two categorical outcomes.

Basophile based its prediction on the basicity of fragment ions on either side to compete for precursor charges. If we go through peptide bonds from N-terminus toward C-terminus, b ions are getting longer and likely to contain more basic residues, y ions are getting shorter and likely to contain less basic residues, such that b ions are gaining ability to obtain charges, and y ions are losing ability to obtain charges. Under this assumption, charge segregation events with higher peptide bond numbers are likely to favor a higher charged b series. These charge segregation events are considered ordered in terms of peptide bond and thus ordinal regression applies. The relationship of fragment ion basicity, peptide charge segregation events, and the ordinal logit is exemplified in **Figure 3**. Ordinal logit is an increasing function that reflects N- or C-terminal ion basicity and determines the charge segregation event for each peptide bond. Charge segregation event “+2 C-term”, “Ambiguous”, and “+2 N-term” are ordered variables.

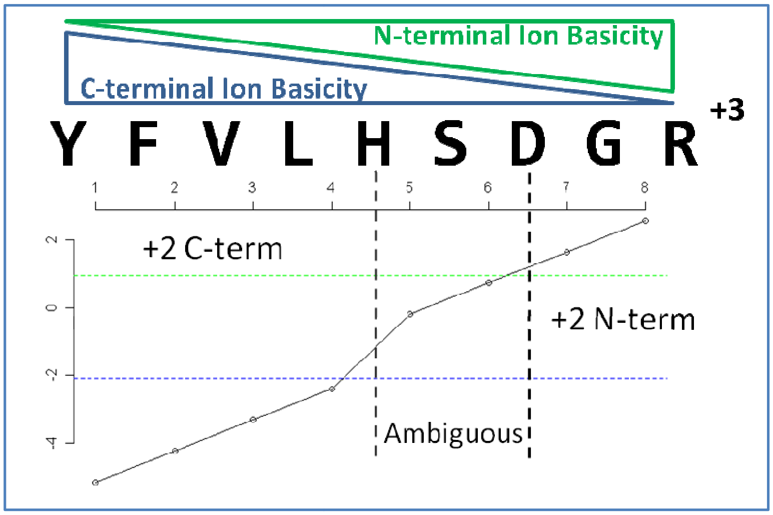


Figure 3. Fragment Ion Basicity and Peptide Charge Segregation. As the peptide bond increases, the N-term Ion Basicity increases and C-term Ion Basicity decreases. Down below the sequence is the ordinal logit calculated from regression function, where bonds are divided into three regions by cutoff values (green and blue dotted lines): +2 C-term (b+1, y+2), +2 N-term (b+2, y+1) and Ambiguous that takes both.

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 2% q-value. PSMs were grouped by precursor charge state, peptide sequence and modifications. We selected the highest scoring MS/MS from each group for training.

Given a peptide bond, Basophile computes the log scaled odds (p/1-p) of observing a charge segregation event using an ordinal logistic regression function $Log(p/1-p)=\beta_1R_N+ \beta_2H_N+ \beta_3K_N+ \beta_4L_N+ \beta_5R_C+ \beta_6H_C+ \beta_7K_C+ \beta_8L_C$, where R_N, H_N, and K_N are number of Arginine (Arg), Hisidine (His) and Lysine (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 residuals at N- and C- terminus, respectively.

Two training tables (one each for +3 and +4 precursors) were generated from the above PSMs of each data set by custom software (IDPDBReader). 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 after removing noise peaks from the spectra using a 95% Total Ion Current (TIC) threshold filter [19]. 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 3** presents a complete list of charge segregation events and evidence of observed fragment ions monitored for +3 and +4 precursors. **Table 4** presents a sample training table generated from triply charged PSMs.

Table 3. Charge segregation events for +3 and +4 peptides.

peptide group	segregation events	evidence of fragment ions
+3 peptides	b+1;y+2	(b+1), (y+2), (b+1, y+2)
	b+1,b+2;y+1,y+2	(b+1, y+1), (b+2, y+2), (b+1, b+2, y+1), (b+1, b+2, y+2), (b+1, y+1, y+2), (b+2, y+1, y+2), (b+1, b+2, y+1, y+2)
	b+2;y+1	(b+2), (y+1), (b+2, y+1)
+4 peptides	b+1;y+3	(b+1), (y+3), (b+1, y+3)
	b+1,b+2;y+2,y+3	(b+1, y+2), (b+2,y+3), (b+1, b+2, y+2), (b+1, b+2, y+3), (b+1, y+2, y+3), (b+2, y+2, y+3), (b+1, b+2, y+2,y+3)
	b+2;y+2	(b+2), (y+2), (b+2, y+2)
	b+2,b+3;y+1,y+2	(b+2, y+1), (b+3, y+2), (b+2, b+3,y+1), (b+2, b+3,y+2), (b+2, y+1,y+2), (b+3, y+1,y+2), (b+2, b+3/y+1,y+2)
	b+3;y+1	(b+3), (y+1), (b+3, y+1)

Table 4. Sample training table including two peptides from NIST-CID data set.

data	peptide	Bond	R_N	H_N	K_N	L_N	R_C	H_C	K_C	L_C	fragment events*	ordinal label
NIST	ITEHMLSLTR	1	0	0	0	1	1	1	0	7	0010	1
	ITEHMLSLTR	2	0	0	0	2	1	1	0	6	0110	1
	ITEHMLSLTR	3	0	0	0	3	1	1	0	5	1010	2
	ITEHMLSLTR	4	0	1	0	3	1	0	0	5	1000	2
	ITEHMLSLTR	5	0	1	0	4	1	0	0	4	1100	2
	ITEHMLSLTR	6	0	1	0	5	1	0	0	3	1010	2
	ITEHMLSLTR	7	0	1	0	6	1	0	0	2	1000	3
	~~ITEHMLSLTR~~#	~~8~~	~~0~~	~~1~~	~~0~~	~~7~~	~~1~~	~~0~~	~~0~~	~~1~~	~~0000~~	
	ITEHMLSLTR	9	0	1	0	8	1	0	0	0	1000	3
	KLALVVEGR	1	0	0	1	0	1	0	0	7	0100	1
	KLALVVEGR	2	0	0	1	1	1	0	0	6	0100	1
	~~KLALVVEGR~~#	~~3~~	~~0~~	~~0~~	~~1~~	~~2~~	~~1~~	~~0~~	~~0~~	~~5~~	~~0000~~	
	KLALVVEGR	4	0	0	1	3	1	0	0	4	1100	2
	KLALVVEGR	5	0	0	1	4	1	0	0	3	1110	2
	KLALVVEGR	6	0	0	1	5	1	0	0	2	1100	2
	KLALVVEGR	7	0	0	1	6	1	0	0	1	1010	2
	KLALVVEGR	8	0	0	1	7	1	0	0	0	1000	3

*: Four digits in this column represent existence (1) or non-existence (0) for a fragment type in the order of "y+1, b+1, y+2, b+2". For example, "1010" means y+1 and y+2 fragments were observed.

#: This peptide bond was deleted from training because no fragments were observed.

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 five-fold cross-validation strategy was used to avoid over-fitting of the function to the data, and customized coding was imbedded in Basophile-Trainer (a customized R package, see **Figure 1**). The regression provided weights for the basicity calculation function and decision table to predict which segregation region best models a given peptide bond. We then implemented these ordinal functions for +3 and +4 precursors in MyriMatch alongside the Naïve model.

Testing the Prediction Efficacy of Basophile

High resolution precursor and fragments in “HCD-Orbitrap-Testing” data set were utilized to measure the efficacy of Basophile charge segregation predictions. The MS/MS of +3 were identified with MyriMatch database search engine configured to use Naïve model for prediction and MVH for results ranking. IDPicker filtered the resulting peptide identifications at a stringent 2% q-value. Another program inspected each PSM, independently recapitulated the fragment predictions using Naïve and Basophile models, matched the predicted fragments to experimental peaks, and assessed the number of fragment hits and misses by each fragment type and charge state.

CHPATER III

Results and Discussion

Algorithms for peptide identification rely upon simplistic models of fragmentation to determine which fragments should be observed for a given sequence. Naïve model, the conventional fragmentation algorithm integrated in popular search engines, assumes a uniform breakage and charge segregation. The purpose of Basophile is to develop a simple but far-reaching model of fragmentation for charge segregation in support of database search. This project employs counts of Arg, His, and Lys residues to either side of a peptide bond in order to determine the charges in which fragment ions may be expected from either terminus.

Naïve model is problematic for highly charged peptides

The Naïve model has a predilection to over-predict fragments expected for a peptide, especially if its precursor carries more than two protons. A CID +3 PSM fragment table under Naïve model is described in **Figure 4**, wherein 43% of predicted fragments were not observed. On average, 57% of predicted fragments Yeast-CPTAC-CID (ORBI) PSMs never matched. Over-prediction rates are worse for HCD-Orbitrap-Testing PSMs, with 74% of predicted fragments missing from the corresponding MS/MS scan. Over-prediction increases the probability of peak matching by random chance because MS/MS of highly charged peptides are often crowded with peaks. False matches, in turn reduce the discrimination of correct matches from incorrect matches. Also, multiple predicted fragments may fall into single *m/z* bin, making the search engine double count fragment matches. Panel B in **Figure 4** reveals 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.

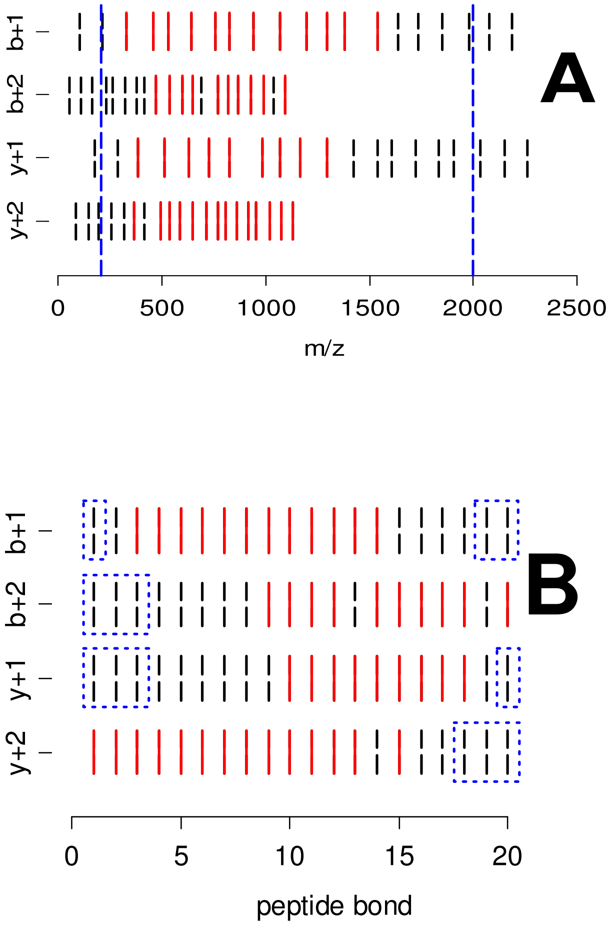


Figure 4. Fragment-peak-matches for a typical CID +3 PSM of peptide "TLLEAIDAIEQPSRPTDKPLR". All the short vertical lines (including long dashes and solid line) represent a predicted fragment ion in *m/z* (A) or per peptide bond (B) under Naïve model. Solid red lines indicate observed ones. The long blue dash lines on panel A indicate scan ranges of the MS/MS spectrum; the rectangles in doted blue lines on Panel B indicate that those ions are out of scan range.

Identification rates are correspondingly lower for highly charged peptides (**Figure 5**). 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 good matches from random ones. The use of fragmentation models that produce excessive fragment predictions, however, worsens matching further.

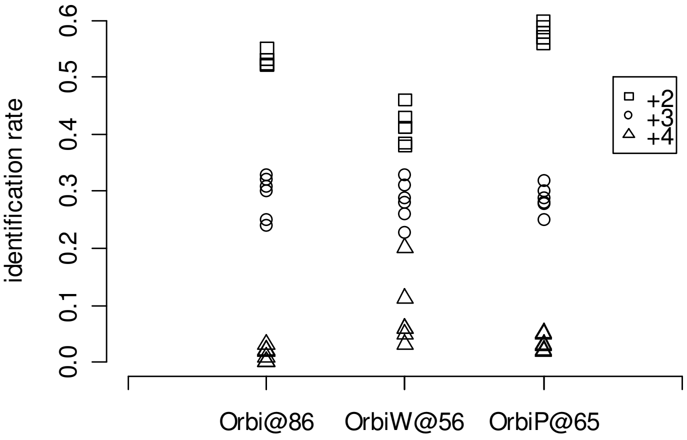


Figure 5. MS/MS of Highly Charged Precursors Suffers from Low Identification Rates. MyriMatch identified peptides from the Yeast-CPTAC-CID (LTQ-Orbi) data set, which featured six technical replicates for each of three instruments. The Naïve model was employed to predict fragments for matching. IDPicker 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.

Constitution of Charge Segregation Events

The Naïve 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 [29,30]. 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 6** summarizes the NIST-CID, Yeast-Multi-Enzyme-CID and HCD-Orbitrap-Training data sets. Doubly-charged precursors fragment in a manner similar to how the Naïve 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. For a typical doubly charged tryptic peptides, many of the protonation sites are accessible in a narrow energy range, such that the distribution of fragment charges are more close to evenly distributed, producing singly charged b ion and y ion with likely equal probability in mass spectrum [3]. However, for highly charged cases like triply charged peptides, more likely one or more of the protonation sites is favored than others leading to sequestration of the added protons, thus some fragment types with particular charge states appear more in mass spectrum. Basophile training was limited to models of the three most common patterns for +3 (exemplified in **Figure 3**) and the five most common outcomes for +4 peptides.

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,y+2;y+2,b+1“ (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 below. 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 data set, and a similar adjustment was applied.

HCD Adjustment Algorithm*:* For triply charged peptides, define ordinal label “b+1/y+2”, “b+1,y+2/y+2,b+1“ as “1”, “2”, and “3” respectively. This algorithm tries to trim the ordinal labels in a “non-decreasing order” heuristically by assigning “y+1 only” either label “2” or “3” through the following steps.

1. List the ordinal categories by peptide bond except “y+1 only”.

2. If the “y+1” only is in the middle of two solid ordinal categories, fill the blank based on the flanking ones. For example, if the flanking ones are “1” (upstream category) and “2” (downstream category), fill “2”. If the flanking labels suggest there is no way of judging this blank (for example, “2” and “1”), delete this row.

3. Fill the blanks with “3” where the upstream category is “3” and there is no downstream one. For example, if the category list shows “11123- - - -“, we finish as “111233333”.

4. With the partially finished bonds, calculate conditional probability p(label=2 | y+1 only) and p(label=3 | y+1 only) by using Bayesian theory. Triply charged scenario gives 0.5 for both. Fill the remaining blanks where they appear at the end of the peptide. For example, if the category list shows “11122- - - -“, we finish as “111222233”.

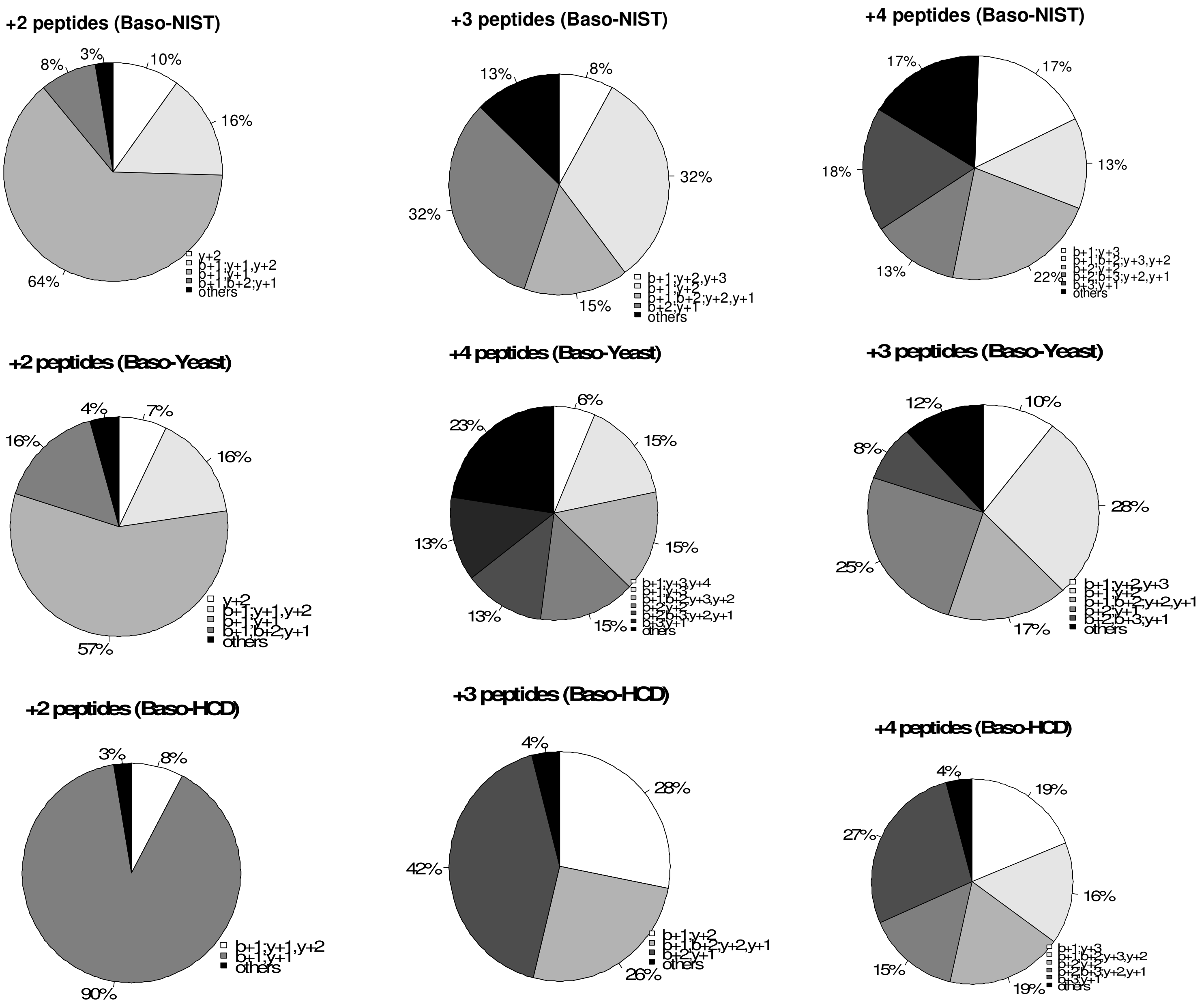


Figure 6. Precursor Charge Segregation Events Observed for Basophile training Peptides. PSMs were segregated by charge state. Charge states of the observed N- and C-terminal fragments were assessed for all peptide bonds. Frequencies of precursor charge segregation events are summarized here. Label "others" include all ordinal categories which are less than 5% and any other fragment pattern that fail to fit any category.

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 data sets differ in the relative distribution of basic residues and also the dissociation method employed to acquire their MS/MS. Basophile-NIST model represents the most peptide sequences, but they are almost universally from trypsin digestion. Basophile-Yeast features varied digestions, but it has far less training power because all data stem from the relatively simple yeast proteome; The Basophile-HCD model differs from the other two models in different fragmentation means and far higher fragment mass accuracy, protecting against false positive matches. 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 ≤ 0.01. However, SEs for +4 Basophile-Yeast and Basophile-HCD models were larger than corresponding Basophile-NIST model, reflecting Basophile-NIST' use of much larger spectral library for training. We chose the Basophile-NIST model as the preferred variant because of this reason. However, it is important to note that the values of coefficients derived from all three training sets followed the same order (**Table 5**). 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.

Table 5. Coefficients of predictor variables by ordinal regression. R_N, H_N, K_N and L_N denote N-terminal Arg, His, Lys, and other residuals; R_C, H_C, K_C and L_C denote C-terminal counterparts; CutOff 1-4 denote threshold value between ordinal labels.

+3 peptides												
model	R_N	H_N	K_N	L_N	R_C	H_C	K_C	L_C	CutOff1	CutOff2		
NIST	1.42	1.31	1.13	0.42	-1.68	-0.90	-1.17	-0.50	-2.23	0.78		
Yeast	1.11	0.97	0.79	0.39	-1.09	-0.87	-0.88	-0.41	-1.78	1.56		
HCD	1.09	0.97	0.75	0.33	-1.53	-0.81	-1.08	-0.42	-3.68	2.03		
+4 peptides												
model	R_N	H_N	K_N	L_N	R_C	H_C	K_C	L_C	CutOff1	CutOff2	CutOff3	CutOff4
NIST	0.79	0.80	0.73	0.30	-0.82	-0.54	-0.62	-0.30	-4.26	-1.94	2.00	4.28
Yeast	0.62	0.56	0.47	0.28	-0.69	-0.55	-0.62	-0.29	-4.25	-1.09	0.71	3.7
HCD	0.77	0.77	0.59	0.30	-1.24	-0.77	-0.89	-0.36	-5.92	-2.96	0.09	4.38

We performed 5-fold cross-validation tests and evaluated the performance of each of the classifiers (trained from training set) on the subset of the data set (testing set) to get a robust estimate of the prediction error rates in our models. In consideration that unambiguous category is subset of ambiguous one, we determined that it would be correct classification if the function predicts ambiguous category for an unambiguous category. For example, if Basophile predicts "b+1,b+2;y+1,y+2" for the peptide bond which produced evidence of observed ions supporting "b+1;y+2", then it is a correct prediction. By this criteria, 5-fold cross validation on NIST-CID data sets gave 7.59%, 7.63%, 7.73%, 7.65%, and 7.69% error rates in +3 identifications, and 15.67%, 15.83%, 15.85%, 15.51%, and 15.64% error rates in +4 identifications. By comparison, HCD-Orbitrap-training data set gave 8.56%, 8.67%, 8.65%, 8.60%, 8.58% error rates in +3 identifications, and 15.68%, 15.71%, 15.72%, 16.23%, 15.70% error rates in +4 identifications; Yeast-Multi-Enzyme-CID data set gave 9.88%, 9.73%, 10.05%, 9.87%, 10.02% error rates in +3 identifications, and 18.31%, 18.69%, 17.36%, 17.68%, 18.27% error rates in +4 identifications. It is obvious that Basophile-NIST had best prediction on the testing subsets for +3 identifications, followed by Basophile-HCD, and then Basophile-Yeast. Basophile-NIST and Basophile-HCD had very comparable error rates on testing subsets, and both are significantly lower than Basophile-Yeast. As such, Basophile-NIST is potentially the best model out of the three. Coefficients and cutoff values out of cross validation showed a very small variance. For example, NIST-CID showed variance of all coefficients and cutoff values exclusively less than 0.0001 for +3 identifications.

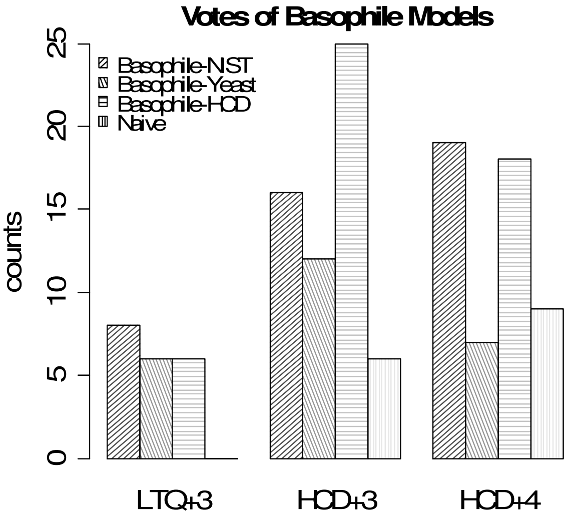


Figure 7. Comparison of Basophile models and Naïve model. For each LC-MS/MS experiment, the prediction model that produced the most identifications was given a "vote." Though the HCD-trained Basophile performed well in HCD data, Basophile-NIST performed well across the samples. The Naive model was competitive only in HCD data, reflecting that false positive matching is a smaller detriment in such data.

We compared the three Basophile models to Naïve model for peptide identification. To accomplish this, all trained models were implemented in the MyriMatch database search engine alongside the Naïve prediction model. Searches for each of the four prediction models were run separately on two LTQ data sets (Yeast-CPTAC-CID (LTQ) and Dicty-LTQ), and one HCD data set (HCD-Orbitrap-Testing) with the standard MVH scorer. **Figure 7** shows the number of files from the test data sets 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 Naïve model. **Figure 8** shows the number of fragments predicted and matched by the Naïve and Basophile-NIST models, grouped by fragment charge state. Compared to the Naïve model, Basophile-NIST reduced the number of fragment predictions by an average of 42% with only slight reductions in numbers 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 data set was rich in tryptic peptides that do not produce large numbers of b+2 fragments; a data set that enriches peptides with N-terminal basic residues might have matched more of these ions.

In contrast to the SQID model [31], Basophile produces a Boolean output, stating 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.

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 [32]. 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.

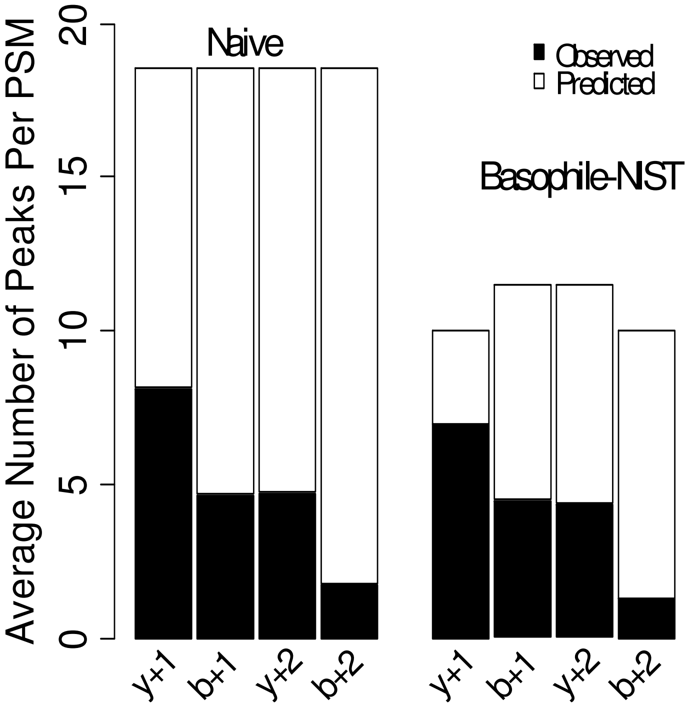


Figure 8. Basophile Improves Peak Prediction Accuracy. Basophile reduces the number of fragments predicted for peptide sequences. This reduction has a minimal impact on the number of matched ions for identified peptides, however. For +3 tryptic peptides, the number of matched b+2 fragments lags behind other classes of fragments.

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 Naïve 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) data set with MVH as the primary score. Searches were performed on 25 cluster nodes, each with two processor cores. In the ten LTQ files, searches using Naïve model took 42 minutes on average, while searches using Basophile took 30 minutes. Overprediction of fragments for peptides can contribute to the time required to search data sets.

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 9** compares the number of +3 and +4 peptides identified in four testing data sets when MyriMatch was employing the Basophile-NIST and Naïve models for the search. For LTQ-CID data sets, Basophile-NIST consistently improved the +3 peptide identification over Naive models (p-value < 0.01). However, the Basophile-NIST model failed to improve the peptide identifications when analyzing HCD-Orbitrap spectra. It appeared that the high-resolution fragment masses of HCD MS/MS neutralized any advantage gained from accurate fragment prediction, abolishing 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-trpsin data sets. All spectra were searched using the above mentioned protocol. Basophile-NIST did not significantly outperform Naïve on +4 MS/MS in both data sets (p-value > 0.05).

Both MVH and HGT+RST scorer benefited from Basophile in LTQ data set 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 reveal that fragment prediction models have a strong relationship with the PSM scoring systems that they support. 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 give less benefit.

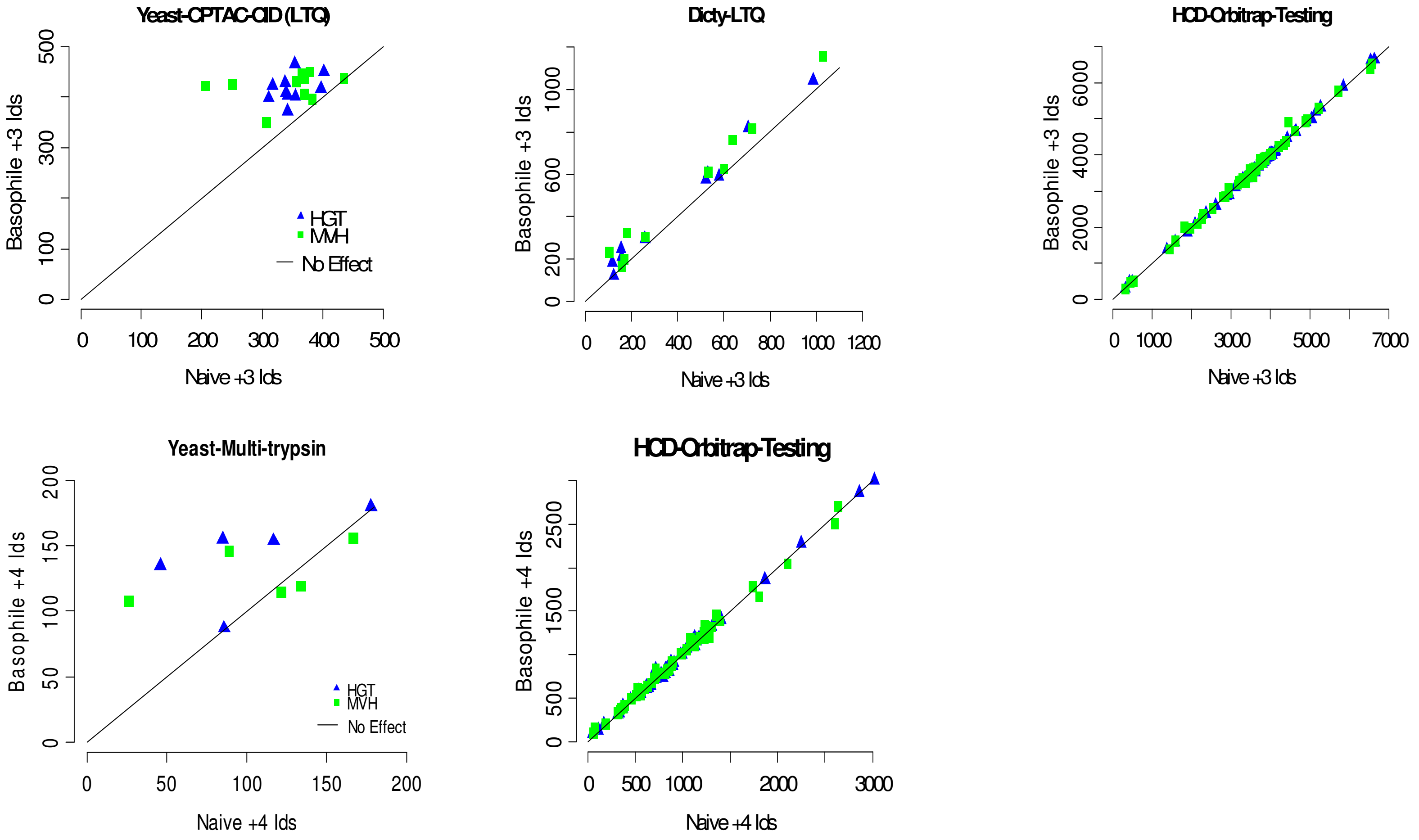


Figure 9. Comparison of Basophile and Naïve models on +3 and +4 peptides identified. MyriMatch employed Basophile-NIST and Naïve models for the search. Reduced but more accurate peak list benefits both scorers by improved peptide identifications in low resolution data, but not in high resolution ones.

Comparison of Peptides Under Naïve or Basophile

In order to explore how the peptides identified by MyriMatch under Basophile predictions compare with those from MyriMatch with naïve fragment predictions, we tested on the Yeast-CPTAC-CID (LTQ) data set. Replicates within this data set showed very similar patterns of +3 peptide overlap, and summed up in **Figure 10**. A large fraction (2897 summed distinct peptide) were found in intersection, while only a small fraction of peptide identified exclusively by either model (254 identifications for Naïve and 820 identifications for Basophile). The sections of the Venn diagram were analyzed in term of peptide length and MVH value. In one raw file, peptides in the overlap section had an average MVH score of 46.96 under Naïve model and 52.68 under Basophile model. Peptides that were exclusive to either Naïve or Basophile showed much lower MVH values, averaging 37.05 and 39.01, respectively. However, the significant difference between these two groups is that Naïve tended to identify shorter peptides (average length was 9.9) than Basophile (average length was 16.1). By comparison, the average length of peptides in the overlap section was 14.6. The overall peptide MVH score and length comparison in the other nine files are summarized in **Table 6**. As such, Basophile was most useful in improving recovery of longer peptide sequences.

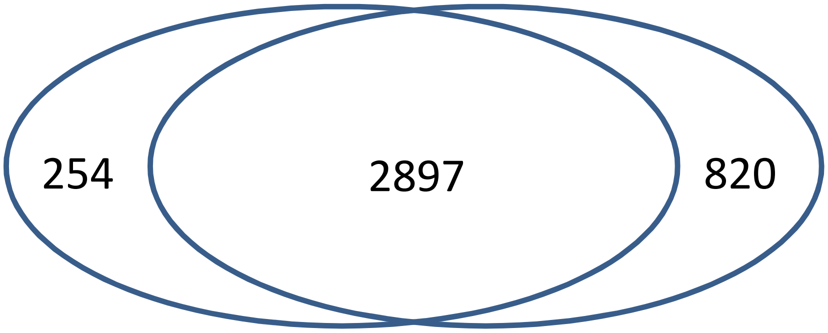


Figure 10. Overlap of +3 peptide identifications from Naïve model and Basophile model. Yeast-CPTAC-CID (LTQ) data (10 raw files) was searched using Naïve and Basophile model with MVH as sort scorer. A majority of peptide identifications are in the intersection part. Each raw file was analyzed separately, and the final sums are represented in the Venn diagram.

Table 6. Comparison of Naïve / Basophile exclusive peptide identifications in nine raw files of Yeast-CPTAC-CID (LTQ) data set. "Model exclusive peptides" means peptide identifications that were identified only by that specific model.

file	Naïve mvh	Naïve length	Basophile mvh	Basophile length
1	39.63	10.8	44.15	14.9
2	40.27	11.3	44.97	15.7
3	40.37	12.6	45.06	14.0
4	41.92	10.5	43.81	14.3
5	37.52	11.8	43.36	14.8
6	39.96	10.3	43.00	15.5
7	38.15	11.1	41.98	14.4
8	39.17	11.0	43.84	15.0
9	43.16	12.5	43.87	14.5

Comparison of +4 peptide identifications under two models was explored in HCD-Orbitrap-Testing data set since LTQ data set did not give +4 peptides. Basophile gave peptide identifications with an average of 29.0 in length while Naïve gave 22.2 instead. Again, Basophile tended to identify longer peptide sequences. It makes sense since higher charged peptides tends to have longer peptide sequences than lower charged ones.

Comparison of Basophile and ppBasicity models

The prediction model introduced by Protein Prospector (ppBasicity) predicts fragment charge based on the count of basic residues contained in a fragment. Both Basophile and ppBasicity count on basic residues for prediction, but Basophile is more realistic because the weights attached to each basic residue comes from training, and Basophile also takes fragment length into consideration.

Basophile-NIST model performed better than ppBasicity model when searching +3 precursors (**Figure 11**). Basophile-NIST increased the +3 identification rates by 27% (p-value < 0.001) and 36% (p-value < 0.01) compared to that of ppBasicity when using Yeast-CPTAC-CID (LTQ) and Dicty-LTQ data sets, respectively. However, Basophile-

NIST did not out-perform ppBasicity when using +3 precursors from HCD-Orbitrap-Testing data set and +4 precursors from all data sets (p-value>0.05).

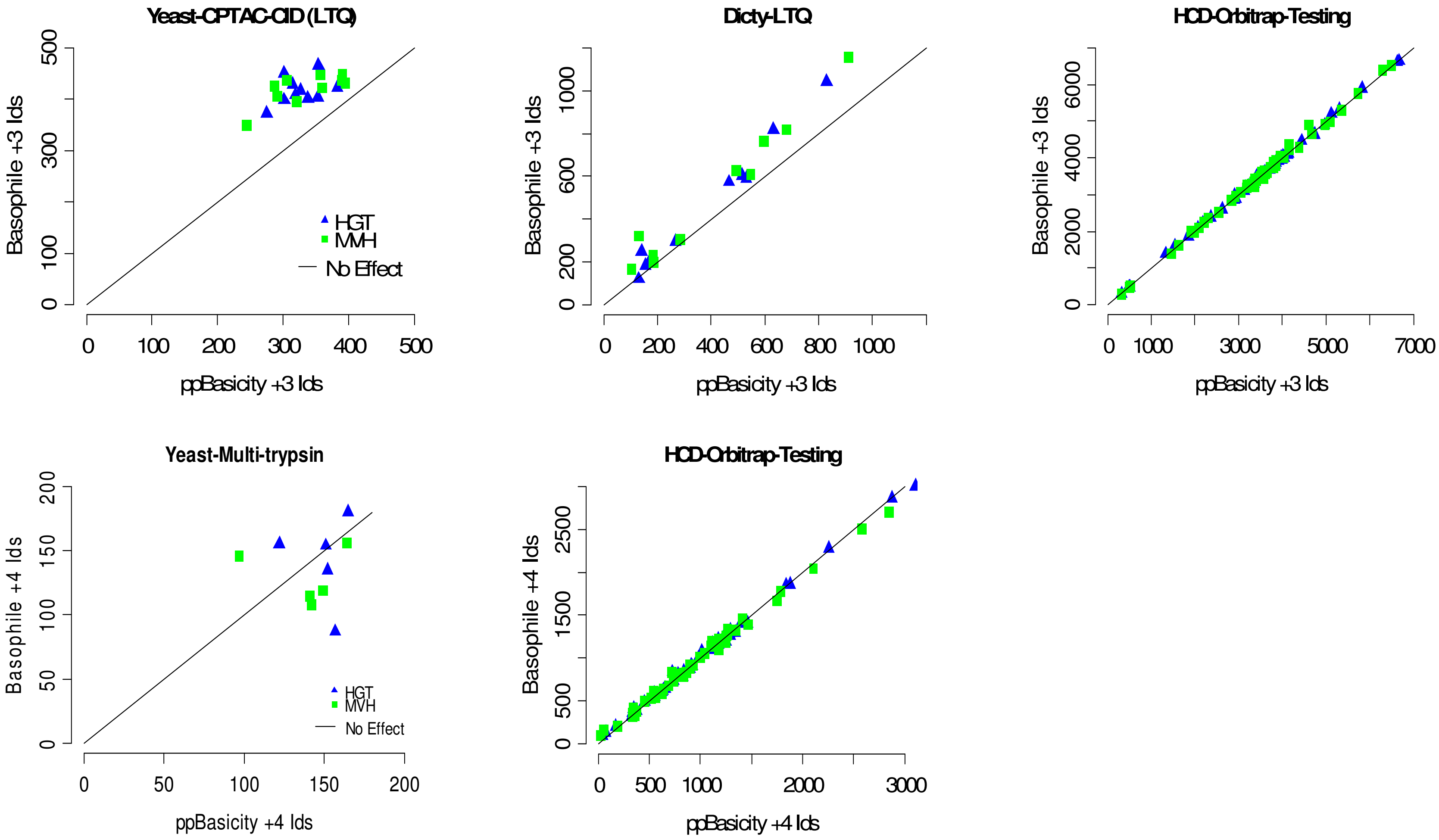


Figure 11. Comparison of Basophile and ppBasicity model on +3 and +4 peptides identified. MyriMatch employed Basophile-NIST and ppBasicity model for the search. Basophile-NIST outperformed ppBasicity by improved peptide identifications on low resolution data on +3 identifications, but failed otherwise.

CHAPTER IV

Conclusion

Basophile was designed to rapidly predict peptide fragmentation spectra (*m/z* values) from sequences that are being matched to MS/MS of +3 and +4 precursors. The model improves the specificity of predictions by reducing the number of unnecessary fragments that are routinely predicted for high charge state precursors. The reduction of fragments not only saves 25% of Naïve search time, but also potentially benefits SRM experiments by reducing the set of fragments screened in further iterations. 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 Naïve model. Basophile identifications features longer +3 and +4 peptides than Naïve model, which appear more frequently in higher changed peptides digested by trypsin. Basophile noticeably outperforms Protein Prospector's prediction model consistently in +3 identifications. Basophile also achieves simplicity by solving the prediction problem with an ordinal regression equation that can be easily incorporated into existing database search software for shotgun proteomic identification.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003, 422:198–207.

2. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001, 19(3):242-247.

3. Paizs B., Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectro. Rev.* 2005, 24:508-548.

4. Guthals A, Bandeira N. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Mol. Cell Proteomics*. 2012 May 17. [Epub ahead of print]

5. Frank AM. Predicting intensity ranks of peptide fragment ions. *J. Proteome. Res*. 2009, 8(5): 2226–2240.

11. Sun S, Meyer-Arendt K, Eichelberger B, Brown R, Yen CY, Old WM, Pierce K, Cios KJ, Ahn NG, Resing KA. Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol. Cell. Proteomics*. 2007, 6(1):1-17.

6. Eng J, McCormack A, Yates J., III. An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*. 1994, 5:976–989.

[7]. Brancia FL, Butt A, Beynon RJ, Hubbard SJ, Gaskell SJ, Oliver SG. A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis*. 2001, 22(3):552-559.

8. Kapp EA, Schütz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem*. 2003, 75(22):6251-6264.

9. Schütz F, Kapp EA, Simpson RJ, Speed TP. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans*. 2003, 31(6):1479-1483.

10. Zhang Z. Prediction of low-energy collision-Induced dissociation spectra of peptides. *Anal. Chem.* 2004, 76:3908–3922.

11. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem*. 2005, 77:6364–6373.

12. 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(2):214-219.

13. Arnold RJ, Jayasankar N, Aggarwal D, Tang H, Radivojac P. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput*. 2006:219-230.

[14]. Frank AM. Predicting intensity ranks of peptide fragment ions. *J Proteome Res*. 2009, 8(5):2226-2240.

15. Li S, Arnold RJ, Tang H, Radivojac P. On the accuracy and limits of peptide fragmentation spectrum prediction. *Anal Chem*. 2011, 83(3):790-6.

16. 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, 116:8368–8369.

17. Dongre´ 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, 118:8365–8374.

18. Wysocki VH, Tsaprailis GT, Smith LL, Breci LA. Mobile and Localized Protons: A framework for understanding peptide dissociation. *J. Mass Spectrom.* 2000, 35:1399–1406.

19. Tabb DL, Fernando CG, and Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* 2007, 6:654– 661

20. Clauser KR, Baker PR, Burlingame AL. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 1999, 71(14): 2871-2882.

21. Dasari S, Chambers MC, Martinez MA, Carpenter KL, Ham AJ, Vega-Montoto LJ, Tabb DL. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.* 2012, 11(3):1686-1695.

22. Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, Fu Y, Yuan ZF, Wang HP, He SM, Dong MQ. pNovo: de novo peptide sequencing and identification using HCD spectra. *J Proteome Res*. 2010, 9(5):2713-2724.

23. Baycin D, *et al*. Proteomic analysis of chinese hamster ovary (CHO) cells. Revision submitted to *J. Proteome Res.*

24. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. *J. Proteome Res.* 2010, 9(4):1716-1726.

25. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res*. 2010, 9(2):761-776.

26. Kessner D, Chambers MC, Burke R, Agus D, Mallick P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics.* 2008, 24(21):2534-2536.

27. 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.

28. 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(1):40-44.

29. Paizs B, Suhai S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun Mass Spectrom.* 2002, 16(17):1699-1702.

30. 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(1):103-113.

31. 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(4):1593-1602.

32. Prakash A, Tomazela DM, Frewen B, Maclean B, Merrihew G, Peterman S, Maccoss MJ. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *J. Proteome Res.* 2009, 8(6):2733-2739.