analytical chemistry

ARTICLE

pubs.acs.org/ac

On the Accuracy and Limits of Peptide Fragmentation Spectrum Prediction

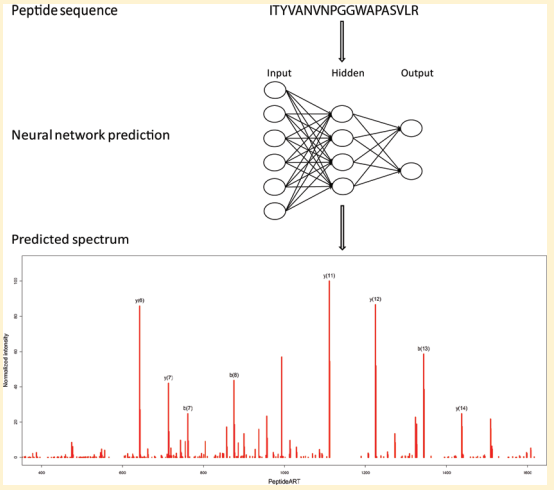
Sujun Li,[†] Randy J. Arnold,[‡] Haixu Tang,[†] and Predrag Radivojac*[,†]

[†]School of Informatics and Computing, Indiana University, Bloomington, Indiana 47408, United States

[‡]Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States

Supporting Information

ABSTRACT: We estimated the reproducibility of tandem mass spectra for the widely used collision-induced dissociation (CID) of peptide ions. Using the Pearson correlation coefficient as a measure of spectral similarity, we found that the within-experiment reproducibility of fragment ion intensities is very high (about 0.85). However, across different experiments and instrument types/setups, the correlation decreases by more than 15% (to about 0.70). We further investigated the accuracy of current predictors of peptide fragmentation spectra and found that they are more accurate than the ad-hoc models generally used by search engines (e.g., SEQUEST) and, surprisingly, approaching the empirical upper limit set by the average across-experiment spectral reproducibility (especially for charge +1 and charge +2 precursor ions). These results provide evidence that, in terms of accuracy of modeling, predicted peptide fragmentation spectra provide a viable alternative to spectral libraries for peptide identification, with a higher coverage of peptides and lower storage requirements. Furthermore, using five data sets of proteome digests by two different proteases, we find that PeptideART (a data-driven machine learning approach) is generally more accurate than MassAnalyzer (an approach based on a kinetic model for peptide fragmentation) in predicting fragmentation spectra but that both models are significantly more accurate than the ad-hoc models.



Tandem mass spectrum interpretation has been challenging from the early days of shotgun proteomics.[1] Original tools such as SEQUEST[2,3] and MASCOT,[4] which adopted a database search strategy that matches experimental tandem mass (MS/MS) spectra to the theoretical spectra of peptides in a protein database, are still widely used. However, even with the best tools available, a large fraction of MS/MS spectra are not identified.[5]

To increase the fraction of identified spectra, the recent development of database search tools has largely focused on two strategies. The first strategy attempts to incorporate additional experimental information into the peptide identification, e.g., to compare the reversed-phase retention time associated with the MS/MS spectra with the predicted retention time of the peptides,[6] to use accurate mass and time analysis in spectral matching,[7] to generate the consensus spectrum for multiple preclustered MS/MS spectra of the same peptide for database searching,[8] and to combine results from multiple MS/MS search engines.[9] The second strategy attempts to improve the scoring scheme for the spectral matching, e.g., to assess not only the number of matched peaks but also their intensities[10,11] or to design matching scores based on the amino acid-specific biases in peptide fragmentation.[12−15] A recent review by Barton and Whittaker provides an excellent discussion of the second group of algorithms as well as the physicochemical factors that are known to affect peptide fragmentation.[16]

For the peptides that fragment well, the database search or, in particular, the peptide-spectrum matching (PSM) problem becomes straightforward if the MS/MS spectra are available for all peptide sequences in the database, since it has been observed that the spectra were reproducible and distinct from one peptide to another. As a result, a new approach to peptide identification, based on the experience with small molecule identification,[17,18] was proposed. It matches experimental MS/MS spectra to the previously identified peptide spectra stored in peptide libraries.[19−23] It was shown that the peptide library approach can identify more spectra than the conventional database searching methods.[21,22] However, the estimates of the amount of increase in sensitivity and the number of identified peptides are still preliminary. In addition, the relative importance of searching smaller databases vs the use of peak intensities has not been quantified. In any case, the spectral library approach is practical only when the spectra have been characterized for all peptides in the sample (or at least all detectable[24] or proteotypic[25] peptides) and can be applied for the well-studied samples (e.g., blood samples) or relatively simple model organisms (e.g., yeast). As a result, hybrid approaches and

Received: August 28, 2010
Accepted: December 7, 2010
Published: December 22, 2010

workflows that combine conventional database searches with spectral library searches have emerged.[26,27]

The spectral library approach can be eliminated if the relative intensities, not only the occurrences of the fragment ions in the experimental spectrum, can be accurately predicted from a peptide sequence alone. Indeed, several computational methods have been developed using either physicochemical models of peptide fragmentation[28,29] or machine learning.[30−33] From limited benchmarking tests, these predictors, as well as those predicting the order of peak intensities,[34] have shown good accuracy and can potentially be used to assist peptide identification.

Peptide fragmentation is inherently stochastic. Combined with other random events such as fluctuations of the ionization source and ion detection, it can result in differences between fragment spectra of the same peptide even in the same experiment. When different instruments, experimental setups, or PSM algorithms are applied, the spectra matched to the same peptide can be significantly different. For example, Venable and Yates studied the variance of PSM scores and found that the distribution of scores depends not only on peptide sequence but also on its quantity.[35] Thus, methods relying on grouping of experimental spectra have been shown to improve peptide identifications.[8,36,37] Similarly, spectrum averaging over different experimental setups plays an important role in building spectral libraries.[22] As a result, it is necessary to further understand and quantify the variability of experimental fragmentation spectra corresponding to the same peptide. In light of the advent of more sophisticated algorithms for predicting fragmentation spectra, this can also help in determining the usefulness of such computational tools because their accuracy cannot be larger than the experimental reproducibility of the fragmentation spectra.

In this paper, we report a systematic assessment of the reproducibility of peptide fragmentation spectra as well as the accuracy of the current peptide MS/MS spectrum predictors for the most commonly used collision-induced dissociation (CID) instruments. We find that an average correlation between two MS/MS spectra repeatedly identified as the same peptide in the same experiment is very high over all precursor ion charge states. However, across different experiments, instrument types, or experimental setups, we find that this correlation decreases by 15% or even more (see Results) but is still substantially higher than the correlation of the ad-hoc models used in peptide search engines. We also computed the correlation coefficients between the experimental and predicted spectra for two predictors: MassAnalyzer, which uses a kinetic model of peptide fragmentation,[28,29] and PeptideART, which adopts a data-driven approach.[31] Both computational tools achieve considerable performance improvement over the ad-hoc models, although correlation coefficients are still somewhat lower, depending on the precursor ion charge state, than the across-experiment spectral reproducibility. Overall, this work supports the use of spectral libraries to most accurately model peptide fragmentation spectra. It also provides evidence that computational models such as MassAnalyzer and PeptideART are viable alternatives to spectral libraries in terms of accuracy but offer several advantages with respect to proteome coverage and storage requirements.

RESULTS

Reproducibility of CID-MS/MS Spectra of Identical Peptides. In the first experiment, we estimated the empirical upper limit for the reproducibility of experimental MS/MS spectra within and across proteomics experiments. For the within-experiment analysis, we used a subset of identified peptides with spectral counts greater than 1 from HUMAN, MOUSE, and YEAST data sets (see Materials and Methods). For each identified unique peptide, we calculated the average Pearson correlation coefficient over all pairs of spectra. Finally, the average correlation coefficient over all unique peptides is reported.

In Figure 1A−C, the distributions of correlation coefficients corresponding to the within-experiment replicated MS/MS spectra are shown. Only unique peptides were considered; thus, all pairs of experimental spectra that were identified as the same peptide were averaged and counted as one. The bars in Figure 1A−C represent the distribution of correlation coefficients for replicates of tandem mass spectra for +1, +2, and +3 precursor ions in three data sets. Overall, the average correlation coefficients for +1, +2, and +3 peptides were estimated to be 0.868 (precursor charge state +1; 825 unique peptides), 0.821 (+2; 8513), and 0.826 (+3; 2264) in data set HUMAN, 0.882 (+1; 56), 0.808 (+2; 2500), and 0.816 (+3; 595) in data set MOUSE, and 0.925 (+1; 27), 0.809 (+2; 439), and 0.762 (+3; 84) in data set YEAST. These results are also shown in Table 1 (column reproducibility[W]).

For the cross-experiment analysis, we used identical peptides identified across HUMAN and MOUSE data sets, as well as SHEWANELLA and DEINOCOCCUS data sets (Figure 1D,E; Table 1). Here, the average correlation coefficients were estimated to be 0.404 (precursor charge state +1; 5 unique peptide pairs), 0.658 (+2; 306), and 0.693 (+3; 50) across HUMAN and MOUSE data sets, whereas the average correlation across SHEWANELLA and DEINOCOCCUS data sets was estimated to be 0.692 (+1; 29), 0.713 (+2; 488), and 0.674 (+3; 57). Although the number of identical peptides across the pairs of data sets was smaller, these results show a significant decrease in spectrum reproducibility of 15% or more compared to a within-experiment spectral reproducibility ($P < 10^{-3}$; Wilcoxon test).

With respect to the specific fragment ion types, we observed that the neutral loss ions (e.g., $b-H_2O$, $y^{++}-NH_3$) are generally less reproducible than the regular fragment ions (e.g., b, y^{++}). Detailed per-ion results are shown in Tables S1−S2 (Supporting Information).

Prediction Accuracy of Computational Models. Prediction accuracy was estimated for two predictors of peptide fragmentation spectra, MassAnalyzer[28,29] and PeptideART.[31] In addition, we estimated the performance of three ad-hoc predictors, referred to as baseline 1, baseline 2, baseline 3; see Materials and Methods. Compared to its original version, PaptideART was retrained using similar features as in its original version but using multi-output neural networks in order to account for the dependencies between fragment ions. Each output corresponds to a specific type of fragment ion (27 types, compared to 11 in the earlier work[31]). PeptideART was trained in two modes: (i) on a specific data set, and (ii) on a set of unique peptides over HUMAN, MOUSE, SHEWANELLA, and YEAST data sets. In each situation, the model was evaluated using 5-fold cross validation; thus, no peptide was used both for training and testing in the same iteration. Peptides present in more than one data set were removed prior to training.

The correlation coefficients over the entire set of spectrum pairs are shown in Table 1. Somewhat surprisingly, the results indicate that the current predictors of peptide fragmentation spectra are within reach of the across-experiment spectral reproducibility (Table 1). In addition, both MassAnalyzer and PeptideART present significant improvements to any of the ad-hoc methods. Interestingly, data set-specific PeptideART was either less accurate or only marginally more accurate than the model trained

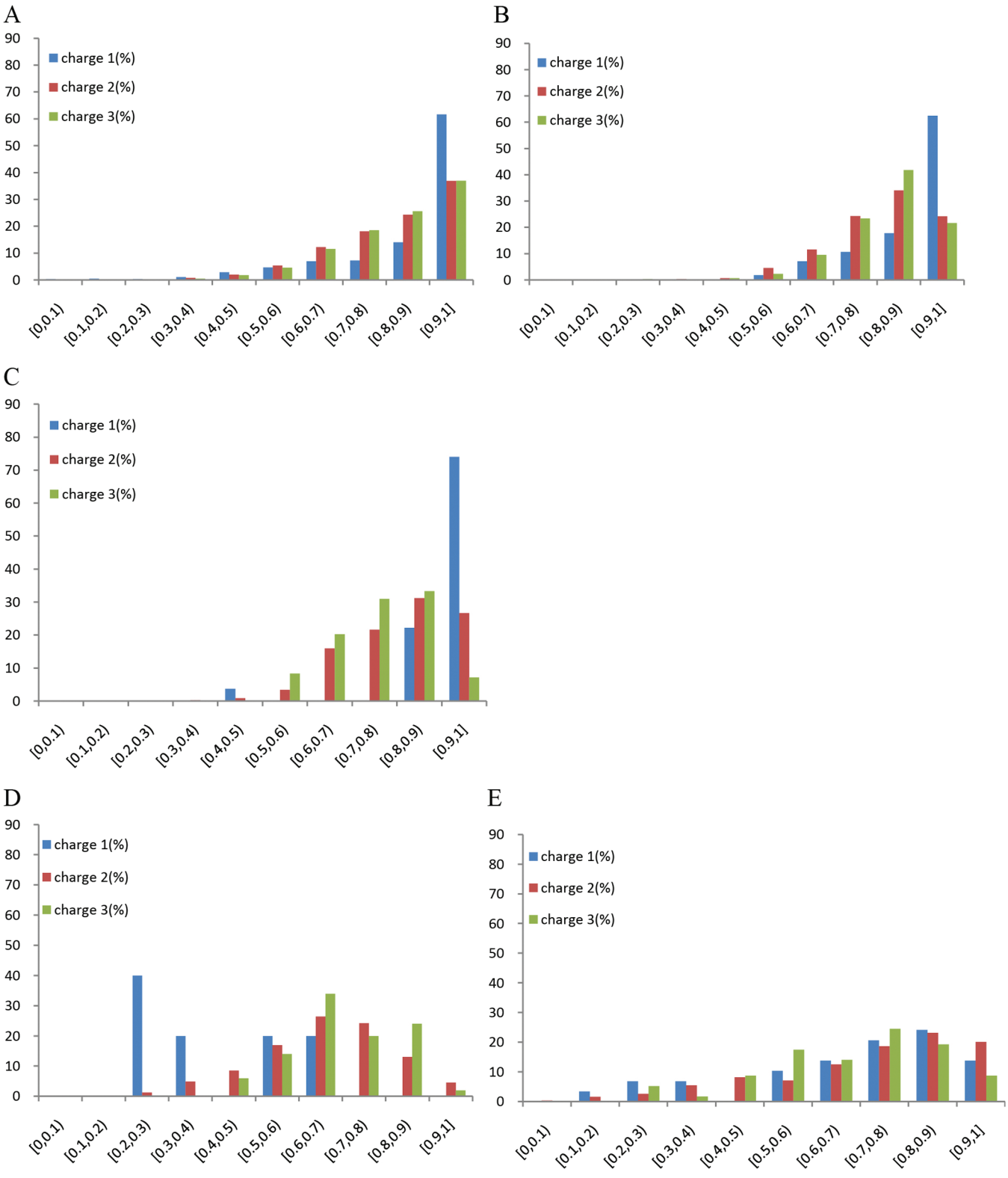


Figure 1. Histograms of spectral reproducibility over peptides identified multiple times in the same experiment (A: HUMAN; B: MOUSE; C: YEAST), as well as the histograms of reproducibility for the unique peptides identified in two different experiments (D: HUMAN VS MOUSE; E: SHEWANELLA VS DEINOCOCCUS).

over all data sets, and this accuracy was significantly lower than the within-experiment spectral reproducibility. This indicates that data-driven models were not able to capture idiosyncrasies of each particular experiment, even if trained for this purpose. Rather, they seem to have learned data set independent rules of peptide fragmentation. Several examples of predicted spectra for MassAnalyzer and PeptideART are shown in Figure S1 (Supporting Information). In addition, the Receiver Operating Characteristic (ROC) curve-based comparisons between models are provided in Tables S3 and S4 (Supporting Information).

dx.doi.org/10.1021/ac102272r |*Anal. Chem.* 2011, 83, 790–796

Table 1. Spectral Similarity (± Standard Deviation) between Experimental and Predicted Peptide Fragmentation Spectra on Four Different Data Sets[a]

charge	MA	ART^{HUMAN}	ART	MA + ART	HUMAN baseline 1	baseline 2	baseline 3	reproducibility^{W}	reproducibility^{A}
+1	0.380 ± 0.244	0.554 ± 0.213	**0.481 ± 0.213**	0.552 ± 0.210	0.179 ± 0.081	0.217 ± 0.104	0.444 ± 0.210	0.868 ± 0.178	0.404 ± 0.161
+2	0.602 ± 0.217	0.654 ± 0.182	**0.644 ± 0.187**	0.658 ± 0.169	0.226 ± 0.077	0.323 ± 0.112	0.444 ± 0.149	0.821 ± 0.147	0.658 ± 0.148
+3	0.509 ± 0.222	0.561 ± 0.164	**0.553 ± 0.163**	0.574 ± 0.152	0.197 ± 0.072	0.209 ± 0.099	0.329 ± 0.101	0.826 ± 0.152	0.693 ± 0.120

charge	MA	ART^{MOUSE}	ART	MA + ART	MOUSE baseline 1	baseline 2	baseline 3	reproducibility^{W}	reproducibility^{A}
+1	0.663 ± 0.212	0.662 ± 0.156	0.676 ± 0.157	0.698 ± 0.161	0.271 ± 0.099	0.374 ± 0.121	0.504 ± 0.180	0.882 ± 0.106	0.404 ± 0.161
+2	0.582 ± 0.222	0.596 ± 0.199	0580 ± 0.195	0.625 ± 0.199	0.198 ± 0.077	0.358 ± 0.131	0.411 ± 0.159	0.808 ± 0.116	0.658 ± 0.148
+3	0.515 ± 0.153	0.516 ± 0.184	0.487 ± 0.163	0.565 ± 0.187	0.169 ± 0.065	0.192 ± 0.093	0.293 ± 0.104	0.816 ± 0.105	0.693 ± 0.120

charge	MA	ART^{SHEWANELLA}	ART	MA + ART	SHEWANELLA baseline 1	baseline 2	baseline 3	reproducibility^{W}	reproducibility^{A}
+1	0.565 ± 0.222	0.627 ± 0.149	**0.627 ± 0.148**	0.634 ± 0.147	0.258 ± 0.102	0.314 ± 0.122	0.440 ± 0.153	N/A	0.692 ± 0.227
+2	0.663 ± 0.187	0.662 ± 0.154	0.666 ± 0.153	0.671 ± 0.140	0.216 ± 0.066	0.327 ± 0.092	0.450 ± 0.132	N/A	0.713 ± 0.210
+3	0.517 ± 0.179	0.571 ± 0.128	**0.571 ± 0.131**	0.591 ± 0.114	0.177 ± 0.047	0.194 ± 0.073	0.330 ± 0.073	N/A	0.674 ± 0.183

charge	MA	ART^{YEAST}	ART	MA + ART	YEAST (Glu-C) baseline 1	baseline 2	baseline 3	reproducibility^{W}	reproducibility^{A}
+1	0.637 ± 0.167	0.638 ± 0.171	0.638 ± 0.171	0.647 ± 0.109	0.254 ± 0.069	0.376 ± 0.098	0.431 ± 0.138	0.925 ± 0.101	N/A
+2	0.500 ± 0.203	0.601 ± 0.175	**0.601 ± 0.175**	0.597 ± 0.161	0.202 ± 0.071	0.274 ± 0.123	0.318 ± 0.128	0.809 ± 0.120	N/A
+3	0.407 ± 0.239	0.423 ± 0.161	0.423 ± 0.161	0.461 ± 0.161	0.150 ± 0.056	0.117 ± 0.060	0.242 ± 0.096	0.762 ± 0.108	N/A

[a] Spectral similarity was measured using the Pearson correlation coefficient. MA = MassAnalyzer, ART = PeptideART, ART$^{(D)}$ = PeptideART trained only on data set D, MA + ART = predictor constructed as an average of MassAnalyzer and PeptideART, baseline 1−3 = three baseline methods referred to in the Materials and Methods section, Reproducibility^{W} = reproducibility within the same sample, Reproducibility^{A} = reproducibility across different experiments (HUMAN vs. MOUSE; SHEWANELLA vs. DEINOCOCCUS). The values in bold indicate that the differences between PeptideART and MassAnalyzer are statistically significant with $P < 0.004$ (Wilcoxon test; Bonferroni-corrected value of 0.05).

Influence of Training Data on PeptideART. The accuracy of PeptideART was also estimated as a function of size of the training data. For a given data set size n, n tryptic peptides were selected uniformly randomly as training and evaluated on the remaining peptides from the combined HUMAN, MOUSE, and SHEWANELLA data sets. To obtain more stable estimates, this strategy was repeated 10 times for each data set size and the accuracy was averaged.

In Figure 2, the Pearson correlation coefficient between predicted and experimental spectra with different number of peptides (n) chosen as training data is shown. In addition to the standard correlation coefficient described in Materials and Methods (white boxes), here, we also estimated the correlation coefficient on the 27 ion types only between annotated ions of experimental and predicted spectra (shaded boxes). The results show that PeptideART is reasonably accurate when trained on as few as 100−200 peptides. This accuracy steadily increases with progressively larger data sets and plateaus at about 1000 peptides.

Running Time of PeptideART. The running time of PeptideART was estimated on a set of 10 000 randomly selected tryptic peptides from human. With about 0.04 s per peptide, creating a library for the entire human genome (~0.5 million tryptic peptides) would take roughly 5 h using a 2.66−3 GHz CPU and a single-threaded process.

DISCUSSION

In this study, we aimed to estimate the reproducibility of low-energy CID-MS/MS fragmentation spectra, within and across different samples and platforms, as well as to evaluate the predictors currently available in the public domain. We found that the reproducibility of peptide fragmentation spectra from the same experiment is consistently very high (Pearson correlation coefficient around 0.85) and is consistent for each protease type. On the other hand, reproducibility across different experiments that use similar ion trap instruments was significantly lower, although still high (around 0.70). This high reproducibility of mass spectra supports peptide identification approaches that utilize spectral libraries[19,21,22,38,39] over the generic strategies of modeling peptide fragmentation spectra.

We also evaluated two predictors of peptide fragment spectra, MassAnalyzer and PeptideART (with PeptideART retrained for this purpose). We found that their prediction accuracy is generally good but dependent on the charge state of the precursor peptide. The best prediction performance was achieved for singly and doubly charged precursor ions, followed by the triply charged precursors. This may be expected, since higher charge state spectra have more possible product ions, including multiply charged ones that may be formed from fragmentation events. Importantly, we estimated that the accuracy of the predicted spectra is relatively similar, with few exceptions, to the spectrum reproducibility across experiments. This strongly suggests that, in terms of accuracy of peptide identification, fragment spectrum predictors are good alternatives to spectral libraries, even with relatively small training data. We note that we used Pearson correlation coefficient as a primary measure of spectral similarity, but similar results were obtained when we applied a square root operation on raw peak intensities (Table S5, Supporting Information). Application of the

dx.doi.org/10.1021/ac102272r |Anal. Chem. 2011, 83, 790–796

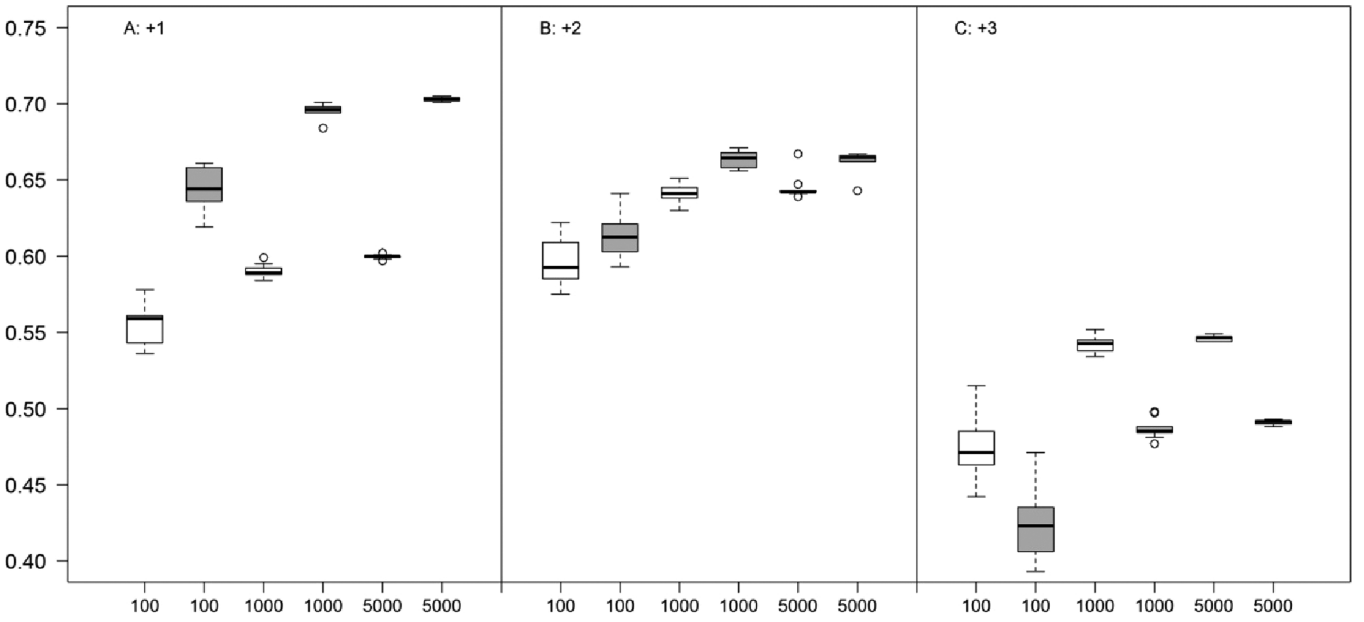


Figure 2. Box plots showing the influence of data set size on PeptideART model for three charges of precursor ions (A: +1; B: +2; C: +3). White boxes represent correlation coefficient as described in Materials and Methods. Shaded boxes represent correlation coefficients over 27 fragment ion types only.

square root function was previously shown to be a good preprocessing step for PSM algorithms.[21,26]

Computational models also offer several advantages over spectral libraries. Once trained, they require significantly less storage space than libraries of annotated spectra (e.g., there are >0.5 million human peptides only in the Swiss-Prot database;[40] if stored, their spectra would require more space than almost any trained machine learning model). In the context of database search, computational models provide theoretical spectra of complete proteomes (with decoy) and may impact the number and confidence of identified proteins and potentially even the estimation of false discovery rates. Finally, computational models can be trained for platforms where spectral libraries have low coverage. For example, even for the commonly used platforms such as CID, only 15% of human tryptic peptides (based on Swiss-Prot) have currently been stored in the NIST library of peptide fragmentation spectra (detailed data not shown).

Compared to MassAnalyzer that was developed based on the current understanding of peptide fragmentation pathways, PeptideART exploited large data sets of annotated spectra to achieve generally higher accuracy. Not only is this useful for accurate peptide identification but also it suggests that the chemistry of peptide fragmentation is difficult to model and not fully understood.

MATERIALS AND METHODS

Data Sets. Five data sets were used in this study. Mouse liver samples in data set MOUSE were digested with trypsin and analyzed by 2D-LC-MS/MS using a ThermoFinnigan LTQ linear ion trap instrument. MASCOT was adopted to search against the IPI mouse v3.71 forward database combined with the reverse database. Peptides with MASCOT scores higher than 40 were selected: 18 107 peptide-spectrum matches (PSMs) were retained, of which 20 PSMs were identified from the reverse database (false discovery rate FDR = 0.22%, peptide level FDR = 0.64%). The final data set contained 67 unique peptides with charge +1, 3218 peptides with charge +2, and 883 peptides with charge +3 (Table 2).

Table 2. Number of Unique Identified Peptides in Each of the Data Sets

charge	HUMAN	MOUSE	SHEWANELLA	DINOCOCCUS	YEAST
+1	1259	67	6010	31	39
+2	11234	3218	11155	796	707
+3	3323	883	1941	183	208

The second data set, referred to as SHEWANELLA, originally included a total of 28 311 identified spectra (7175 +1, 17 647 +2, and 3489 +3) from the *Shewanella oneidensis* and *Deinococcus radiodurans* proteomes[13] collected using HPLC with LCQ ion trap instruments. The peptides were identified using SEQUEST. In order to ensure a high quality of identified spectra (FDR not provided in the original paper), we applied new cutoffs to the set of peptide-spectrum matches (Xcorr = 2.0 for +1, Xcorr = 3.0 for +2, Xcorr = 4.0 for +3 peptides). The new data set contained 6010 +1, 11 155 +2, and 1941 +3 peptides.

Data set HUMAN comes from a human cell line.[41] The MS/MS proteomics analyses were carried out on an extract of the erythroleukemia cell line K562 grown in suspension. After trypsin digestion, a multistage gradient delivered by an Agilent 1100 Series HPLC (Agilent Technologies, Santa Clara, CA) was used to elute peptides into the electrospray ionization source of an LCQ ion trap mass spectrometer (ThermoElectron, San Jose, CA). In this work, we used InsPecT[42] to search against a database (IPI human v3.57 forward and reversed databases combined) which resulted in 84 471 PSMs with *P*-values below 0.01. Among them, 63 PSMs were from the reversed database (FDR = 0.15%, peptide level FDR = 0.73%). The final data set included a total of 1259 +1 peptides, 11 234 +2 peptides, and 3323 +3 peptides.

Data set DEINOCOCCUS, was created from 20 replicate analyses of the *D. radiodurans* proteome in our previous work.[43] The *D. radiodurans* samples were digested using trypsin, and the peptides were separated using nano-LC. The eluting peptides were electrosprayed into a ThermoFinigan LCQ Deca XP ion-trap mass spectrometer. The peptides were identified using MASCOT that

dx.doi.org/10.1021/ac102272r |*Anal. Chem.* 2011, 83, 790–796

searched forward and reverse *D. radiodurans* databases (FDR = 0.05%, peptide level FDR = 0.65%). This data set was used only to compare spectra of peptides identified both in DEINOCOCCUS and in SHEWANELLA in order to estimate reproducibility of fragment spectra across different experiments.

The last data set, YEAST, was constructed from the *Saccharomyces cerevisiae* mutant strain samples. The samples were digested using Glu-C to produce peptides terminated by aspartic or glutamic acid residues. The peptides in the digested sample were separated using a MudPIT experiment.[44] The released peptides were electrosprayed into a ThermoFinnigan LTQ mass spectrometer. The PSMs were generated using SEQUEST followed by PeptideProphet[45] with a probability cutoff of 0.95. The resulting number of unique peptides consisted of 39 +1 peptides, 707 +2 peptides, and 208 +3 peptides.

Computational Approaches to Predicting CID-MS/MS Spectra of Peptides. Two previously published methods were used to compare experimental and predicted spectra: (i) MassAnalyzer, an algorithm, introduced by Zhang,[28,29] which explicitly models the understood model of peptide fragmentation with parameter optimization based on the training CID spectra; (ii) PeptideART, a neural network-based model designed to predict the probability that a particular fragment ion will be observed.[31] PeptideART uses the outputted probabilities as estimates of the fragment ion intensities. For the purposes of this study, we retrained PeptideART using ensembles of 30 multioutput feed-forward neural networks, whereas the original version combined ensembles of single output networks. Thus, the retrained model better accounts for the dependencies between fragment ions. Additionally, we reduced the overall number of features (for speed), increased the number of predicted fragment ion types, and accommodated for the isotopic peaks using the method by Zhang.[28]

Features used to train PeptideART can be categorized in the following five groups: (i) peptide length and mass for the whole peptide as well as left and right fragment ions for a specific cleavage site; (ii) amino acid compositions for both fragments given the position of the cleavage site; (iii) physicochemical properties (basicity, helicity, hydrophobicity, pI) for both fragments;[31] (iv) distances from the termini to the nearest residues P, H, K, and R in both fragments; and (v) N-terminal and C-terminal amino acid for both fragments. The total number of features is 158. We considered the following 27 fragment ions: precursor, precursor$-H_2O$, precursor$-NH_3$, b, b$-H_2O$, b$-NH_3$, b$-H_2O-NH_3$, b$+H_2O$, a, a$-H_2O$, a$-NH_3$, y, y$-H_2O$, y$-NH_3$, y$-H_2O-NH_3$, b^{++}, $b^{++}-H_2O$, $b^{++}+H_2O$, $b^{++}-NH_3$, a^{++}, $a^{++}-H_2O$, $a^{++}-NH_3$, y^{++}, $y^{++}-H_2O$, $y^{++}-NH_3$, y^{+++}, b^{+++}. The doubly charged fragment ions were used for +2 and +3 precursor ions, while the triply charged fragment ions were used only for the +3 precursor ions.

In addition to MassAnalyzer and PeptideART, we also used three ad-hoc models. Baseline 1 model is the simplest scheme in which every possible fragment ion is assigned intensity of 1. Baseline 2 model outputs intensity of 1 for b- and y-ions, intensity of 0.5 for a-ions, intensity of 0.5 for ions with single neutral loss (e.g., b$-H_2O$ or y$-NH_3$), intensity of 0.25 for double neutral loss ions (e.g., b$-H_2O-NH_3$), and intensity of 0.25 for doubly charged fragment ions b^{++} and y^{++}. Finally, baseline 3 model outputs the prior probabilities of occurrence for each ion type (see Table S6, Supporting Information, for details), thus outputting different values depending on the fragment ions under consideration. For example, in the MOUSE data set, the b ions with intensity ≥1% of the total intensity were observed in 18.1% of cases, thus, in the theoretical spectrum, every b ion was assigned intensity of 0.181. For evaluation purposes, we used a publicly available version of MassAnalyzer, while PeptideART predictor was retrained using the data sets above and evaluated using 5-fold cross-validation on a set of unique peptides across different data sets.

Measuring Similarity of Fragment Spectra. Two performance measures were used to assess the reproducibility of experimental spectra and the quality of predicting experimental spectra: (i) the Pearson correlation coefficient and (ii) the area under the ROC curve (AUC). In the case of reproducibility estimation, for each confidently identified peptide, the spectra matched to this peptide were selected, with all identifications being above the score threshold (based on false discovery rate, Xcorr, or PeptideProphet probability value). The reproducibility was estimated by averaging the correlation coefficients over all spectrum pairs for a particular peptide and then further averaged over all unique peptides. Peptides identified based on a single matched spectrum, regardless of the score, were omitted. Therefore, data set SHEWANELLA was not used for within-experiment reproducibility analysis because the authors included only the highest scoring spectrum for each peptide (i.e., spectral count for each peptide in SHEWANELLA was 1). In the case of assessing the quality of prediction of fragment ions, we selected the highest scoring experimental spectrum for each peptide and then compared it with the predicted spectrum.

Correlation Coefficient. Given two spectra, S_a and S_b, each spectrum was binned using 1200 bins in the m/z range from 200 to 2000 (the size of the bin was selected to correspond to the tolerance used to match fragment ions, ±0.8). The highest peak in each bin was selected to represent the bin; thus, each spectrum was encoded into a 1200-dimensional vector. The Pearson correlation coefficient was calculated between such pairs of 1200-dimensional vectors.

Area under the ROC Curve. AUC was computed by assuming that spectrum S_a was the "correct" spectrum and spectrum S_b was its "prediction". Each fragment ion in S_a whose intensity was ≥1% of the total intensity of the spectrum was considered to be positive, while all other fragment ions were considered to be negative. A sliding threshold t, ranging from 0 to the maximum intensity (all spectra were normalized to 0−1 interval), was then applied to spectrum S_b to calculate sensitivity (sn, true positive rate) and specificity (sp, true negative rate). AUC was obtained as an area under the observed curve with (1 − sp) as the x-axis and sn as the y-axis, over the entire set of n pairs. While this approach gives more weight to longer peptides, it is a more stable estimate than an average of pairwise AUCs, given that a relatively small number of fragment ions comprise most of the total intensity of the spectrum.

■ ASSOCIATED CONTENT

Supporting Information. Additional information as noted in the text: (1) examples of predicted spectra for MassAnalyzer and PeptideART; (2) ion reproducibility analysis of experimental spectra; (3) ROC tables for PeptideART and MassAnalyzer; (4) correlation coefficients for PeptideART and MassAnalyzer when square root was applied to peak intensities; (5) prior probability tables in different data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: predrag@indiana.edu.

■ ACKNOWLEDGMENT

We thank Yong Fuga Li for assembling peptide identifications for the DEINOCOCCUS data set. We also thank Mark Goebl and Ross Cocklin, Quanhu Sheng, Vicki Wysocki, and Katheryn Resing for providing us with the YEAST, MOUSE, SHEWANELLA, and HUMAN data sets, respectively. This work was supported by the National Institutes of Health Grant R01 RR024236-01A1 and National Cancer Institute Grant U24 CA126480-01. Finally, we thank the reviewers for their comments that improved the quality of this paper. Availability: PeptideART is freely available at www.informatics.indiana.edu/predrag.

■ REFERENCES

(1) Dongre, A. R.; Eng, J. K.; Yates, J. R., 3rd *Trends Biotechnol* **1997,** *15* (10), 418–425.
(2) Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 976–989.
(3) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995,** *67* (8), 1426–1436.
(4) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999,** *20* (18), 3551–3567.
(5) Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D. *Methods* **2005,** *35* (3), 223–236.
(6) Spicer, V.; Yamchuk, A.; Cortens, J.; Sousa, S.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2007,** *79* (22), 8762–8768.
(7) May, D.; Fitzgibbon, M.; Liu, Y.; Holzman, T.; Eng, J.; Kemp, C. J.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *J. Proteome Res.* **2007,** *6* (7), 2685–2694.
(8) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2008,** *7* (1), 113–122.
(9) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. *J. Proteome Res.* **2008,** *7* (1), 245–253.
(10) Sadygov, R.; Wohlschlegel, J.; Park, S. K.; Xu, T.; Yates, J. R., 3rd *Anal. Chem.* **2006,** *78* (1), 89–95.
(11) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. *J. Proteome Res.* **2007,** *6* (2), 654–661.
(12) Huang, Y.; Triscari, J. M.; Pasa-Tolic, L.; Anderson, G. A.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *J. Am. Chem. Soc.* **2004,** *126* (10), 3034–3035.
(13) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *Anal. Chem.* **2005,** *77*, 5800–5813.
(14) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D; Yates, J. R., 3rd *Anal. Chem.* **2003,** *75* (5), 1155–1163.
(15) Savitski, M. M.; Kjeldsen, F.; Nielsen, M. L.; Zubarev, R. A. *J. Proteome Res.* **2007,** *6* (7), 2669–2673.
(16) Barton, S. J.; Whittaker, J. C. *Mass Spectrom. Rev.* **2009,** *28* (1), 177–187.
(17) Hertz, H.; Hites, R. A.; Biemann, K *Anal. Chem.* **1971,** *43* (6), 681–691.
(18) Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. *J. Am. Soc. Mass Spectrom.* **1999,** *10* (4), 287–299.
(19) Yates, J. R., 3rd; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998,** *70* (17), 3557–3565.
(20) Craig, R.; Cortens, J. P.; Beavis, R. C. *Rapid Commun. Mass Spectrom.* **2005,** *19* (13), 1844–1850.
(21) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. *Anal. Chem.* **2006,** *78* (16), 5678–5684.
(22) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007,** *7* (5), 655–667.
(23) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. *Nat. Methods* **2008,** *5* (10), 873–875.
(24) Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P. *Bioinformatics* **2006,** *22* (14), e481–e488.
(25) Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. *Nat. Rev. Mol. Cell Biol.* **2005,** *6* (7), 577–583.
(26) Bern, M. *Proceedings of the 8th Annual International Conference on Computational Systems Bioinformatics*, Life Sciences Society: Woodside, CA, 2009; pp 117−126.
(27) Ahrne, E.; Masselot, A.; Binz, P. A.; Muller, M.; Lisacek, F. *Proteomics* **2009,** *9* (6), 1731–1736.
(28) Zhang, Z. *Anal. Chem.* **2004,** *76* (14), 3908–3922.
(29) Zhang, Z. *Anal. Chem.* **2005,** *77* (19), 6364–6373.
(30) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat. Biotechnol.* **2004,** *22* (2), 214–219.
(31) Arnold, R. J.; Jayasankar, N.; Aggarwal, D.; Tang, H.; Radivojac, P. *Pac. Symp. Biocomput.* **2006,** 219–230.
(32) Barton, S. J.; Richardson, S.; Perkins, D. N.; Bellahn, I.; Bryant, T. N.; Whittaker, J. C. *Anal. Chem.* **2007,** *79* (15), 5601–5607.
(33) Klammer, A. A.; Reynolds, S. M.; Bilmes, J. A.; MacCoss, M. J.; Noble, W. S. *Bioinformatics* **2008,** *24* (13), i348–356.
(34) Frank, A. M. *J. Proteome Res.* **2009,** *8* (5), 2226–2240.
(35) Venable, J. D.; Yates, J. R., 3rd *Anal. Chem.* **2004,** *76* (10), 2928–2937.
(36) Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., 3rd *Anal. Chem.* **2003,** *75* (10), 2470–2477.
(37) Tabb, D. L.; Thompson, M. R.; Khalsa-Moyers, G.; VerBerkmoes, N. C.; McDonald, W. H. *J. Am. Soc. Mass Spectrom.* **2005,** *16* (8), 1250–1261.
(38) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. *J. Proteome Res.* **2006,** *5* (8), 1843–1849.
(39) Liu, J.; Bell, A. W.; Bergeron, J. J.; Yanofsky, C. M.; Carrillo, B.; Beaudrie, C. E.; Kearney, R. E. *Proteome Sci.* **2007,** *5*, 3.
(40) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. *Nucleic Acids Res.* **2005,** *33* (Database Issue), D154–159.
(41) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. *Anal. Chem.* **2004,** *76* (13), 3556–3568.
(42) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. *Anal. Chem.* **2005,** *77* (14), 4626–4639.
(43) Li, Y. F.; Arnold, R. J.; Tang, H.; Radivojac, P. *J. Proteome Res.* **2010,** *9* (12), 6288–6297.
(44) Radivojac, P.; Vacic, V.; Haynes, C.; Cocklin, R. R.; Mohan, A.; Heyen, J. W.; Goebl, M. G.; Iakoucheva, L. M. *Proteins* **2010,** *78* (2), 365–380.
(45) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002,** *74* (20), 5383–5392.