analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 8705−8711

pubs.acs.org/ac

AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility

Zhiqiang Gao,[†,§,∥] Cheng Chang,[‡,∥] Jinghan Yang,[†,§] Yunping Zhu,[*,‡,⊥] and Yan Fu[*,†,§]

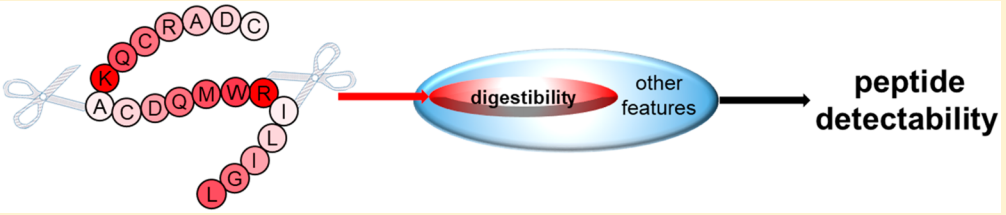
[†]National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[‡]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China

[§]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[⊥]Anhui Medical University, Hefei 230032, China

Supporting Information



ABSTRACT: The selection of proteotypic peptides, that is, detectable unique representatives of proteins of interest, is a key step in targeted proteomics. To date, much effort has been made to understand the mechanisms underlying peptide detection in liquid chromatography−tandem mass spectrometry (LC-MS/MS) based shotgun proteomics and to predict proteotypic peptides in the absence of experimental LC-MS/MS data. However, the prediction accuracy of existing tools is still unsatisfactory. We find that one crucial reason is their neglect of the significant influence of protein proteolytic digestion on peptide detectability in shotgun proteomics. Here, we present an Advanced Proteotypic Peptide Predictor (AP3), which explicitly takes peptide digestibility into account for the prediction of proteotypic peptides. Specifically, peptide digestibility is first predicted for each peptide and then incorporated as a feature into the peptide detectability prediction model. Our results demonstrated that peptide digestibility is the most important feature for the accurate prediction of proteotypic peptides in our model. Compared with the existing available algorithms, AP3 showed 10.3−34.7% higher prediction accuracy. On a targeted proteomics data set, AP3 accurately predicted the proteotypic peptides for proteins of interest, showing great potential for assisting the design of targeted proteomics experiments.

In shotgun proteomics, proteins are first digested by some enzyme into peptide mixtures, which are then analyzed by liquid chromatography−tandem mass spectrometry (LC-MS/MS). In recent years, MS-based targeted proteomics, such as multiple reaction monitoring (MRM) experiments, are capable of the sensitive identification and quantification of proteins of interest and have become a promising powerful tool for biological or clinical studies, such as the verification of candidate biomarkers.[1,2] The first key step in developing an MRM assay is the selection of proteotypic peptides, that is, the peptides that are unique representatives of their corresponding proteins and are detectable in proteomic experiments.[3] There are two major approaches to proteotypic peptide selection, that is, the experimental approach and the computational approach. The former selects previously identified peptides as proteotypic peptides, while the latter predicts proteotypic peptides using empirical or machine learning models. Although the experimental approach has been successfully applied in MRM assays, it has some limitations. For example, not all target proteins have experimental evidence, especially the proteins identified by literature mining. Thus, researchers are paying more attention to the computational approach. However, the mechanisms underlying peptide detection are still unclear, which hinders the development of accurate proteotypic peptide prediction algorithms.

To date, much effort has been devoted to understanding the mechanisms of peptide detection and predicting proteotypic peptides. In early research, empirical score functions were constructed based on the physiochemical properties of peptides, such as hydrophobicity, peptide length, and isoelectric point.[4,5] In recent years, several machine learning-based algorithms[6−12] have been developed to predict proteotypic or high-responding peptides. In these studies, designing the features characterizing the peptides is one of the key steps. The likelihood of observing a peptide in a shotgun proteomics experiment is governed by many factors, such as

Received: June 1, 2019
Accepted: June 6, 2019
Published: June 6, 2019

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

Table 1. Summary of the Four Public Data Sets Used for Training and Testing AP3

data set name	usage	instrument	proteins identified	peptides identified	proteins after filtration	peptides after filtration
Yeast	training	Orbitrap Fusion	3959	43088	1556	29172
E. coli	testing	LTQ Orbitrap Elite	2898	38035	1077	23674
Mouse	testing	Q Exactive Plus	5304	58184	2120	39095
Human	testing	LTQ Orbitrap	6422	80490	2606	55289

the physicochemical properties of the peptide, the LC-MS/MS behavior of the peptide, and the abundance of its corresponding protein.[13,14] Tang et al.[13] first proposed the concept of peptide detectability, which is defined as the probability that a peptide would be observed in a standard sample analyzed by a standard proteomics routine. They also invented a machine learning model that used 175 features derived from the peptide sequence to predict the peptide detectability. Later, Sander et al.,[9] Mallick et al.,[8] and Eyers et al.[11] developed their models using 596, 1010, and 1186 features, respectively. These features mainly include AAindex-derived features[15] and sequence-derived features. Recently, Muntel et al.[16] considered the protein abundance as an additional feature and obtained improved performance. However, the protein abundance information is generally unavailable in the absence of experimental LC-MS/MS data.

A common limitation of the above algorithms is that they only considered the features of peptides themselves but did not consider the influence of protein proteolytic digestion on peptide detectability. As we know, a typical shotgun proteomics experiment can be divided into two successive processes, protein proteolytic digestion and peptide detection by LC-MS/MS, and the peptide detectability is the probability of observing a peptide in the whole shotgun proteomics experiment, not only in the peptide detection process, by LC-MS/MS. It is important to know which peptides and what proportions of them could be digested out and subjected to subsequent LC-MS/MS analysis. Previous studies have demonstrated that the process of protein digestion by the commonly used enzyme trypsin is always incomplete.[17] The cleavage probability of a tryptic site is mainly determined by the amino acids surrounding the site and affected by other factors such as the local conformation and tertiary structure of its protein as well as experimental conditions of digestion.[17] Several algorithms have been demonstrated to be successful in predicting the cleavage probabilities of tryptic sites exclusively from the adjacent amino acids.[18,19] Note that half of the amino acids surrounding a cleavage site are outside the peptide sequence and therefore are not characterized by previous peptide detectability prediction algorithms. For more accurate prediction of peptide detectability, the peptide digestibility must be explicitly taken into account.

Here, we present a novel algorithm named AP3 (short for Advanced Proteotypic Peptide Predictor) that, for the first time, integrates the protein digestion and peptide LC-MS/MS detection processes together to predict proteotypic peptides. Specifically, it first predicts the peptide digestibility and then incorporates it as a feature into the peptide detectability prediction model. Our results showed that peptide digestibility was the most important feature for predicting proteotypic peptides, increasing the 10-fold cross-validation AUC (area under the receiver operating characteristic (ROC) curve) from 0.8891 to 0.9269. Furthermore, we demonstrated that the peptide detectability prediction model trained on a yeast data set retained good predictive power over three independent comprehensive data sets from *E. coli* (AUC 0.9405), mouse (AUC 0.9313), and human (AUC 0.9213) samples. Compared with existing tools, including PeptideSieve,[8] CONSeQuence,[11] ESP Predictor,[6] and PPA,[16] AP3 exhibited 10.3−34.7% higher accuracy in AUC. At last, we showed that AP3 can effectively predict proteotypic peptides for targeted proteomics in the absence of experimental LC-MS/MS data.

EXPERIMENTAL SECTION

Data Sets. A public large-scale yeast data set[20] was used to train the AP3 model. To test the generalization performance of AP3, we used three independent public data sets from other organisms, that is, *E. coli*,[21] mouse,[22] and human.[23] In all of these data sets, the proteolytic digestion was performed by incubation with trypsin overnight. For the human data set, the lymph node and salivary gland data in their original publication were used. We reanalyzed the raw MS files as follows.

Mass spectra in each data set were searched against the corresponding organism sequences from the UniProt database using the Andromeda search engine[24] in the MaxQuant software[25] (version 1.6.0.1). Carbamidomethylation on cysteine was set as a fixed modification. Methionine oxidation and protein N-terminal acetylation were set as variable modifications. Protein sequences were theoretically digested using fully tryptic cleavage constraints, and up to two missed cleavage sites were allowed. The precursor mass tolerance was set to 20 ppm for the first search and 4.5 ppm for the main search. The mass tolerance for fragment ions was set to 0.05 Da for Q Exactive Plus and 0.5 Da for other instruments. False discovery rates at the protein and peptide levels were both set to 1%.

Filtration of Identified Proteins. To construct more confident training and test data sets for AP3, the identified proteins were further filtered by their spectral counts (SCs) and sequence coverages. The SC of a protein was calculated as the sum of SCs of its identified peptides. It is reasonable to assume that a protein is confidently identified if it has a large SC and high sequence coverage. Therefore, we sorted the proteins by both their SCs and sequence coverages and kept those proteins that appeared in the top 50% of both ranks. Table 1 provides a summary of the four data sets.

Workflow of AP3 Algorithm. As Figure 1 illustrates, the development workflow of AP3 has two main components: a peptide digestibility predictor and a peptide detectability predictor. To avoid overfitting and confounding the peptide digestibility and detectability in the training stage, the identified filtered proteins are randomly divided into two halves. The first half is used to train the peptide digestibility predictor and the other half is used to train the peptide detectability predictor. A random forest classifier is first trained to predict the cleavage probabilities of tryptic sites. Then, the digestibility of each peptide is calculated from the cleavage probabilities of its tryptic sites and integrated into the peptide detectability predictor as one of the features that characterize the peptide. Next, other 587 physicochemical features of peptides are calculated and the feature selection is performed.

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

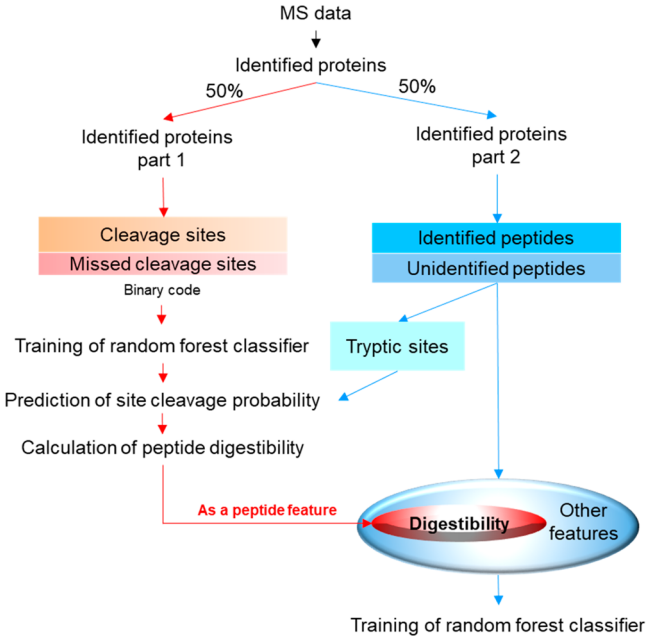


Figure 1. Development workflow of AP3. The identified proteins are divided into two halves to completely separate the training sets of peptide digestibility predictor and peptide detectability predictor. The first half is used to train a random forest classifier for predicting cleavage probabilities of tryptic sites. The other half is used to train another random forest classifier for predicting peptide detectability. The predicted digestibility is integrated as a feature into the peptide detectability predictor.

Finally, another random forest classifier is trained for predicting the peptide detectability.

Peptide Digestibility Predictor. *Constructing the Training Set.* Identified peptides were first mapped to their corresponding protein sequences. Then, the cleavage information on potential tryptic sites, for example, the arginine (R) and lysine (K) residues, in the first half of identified proteins was collected, including the SCs of the peptides observed on the N-terminal (SC_N) and C-terminal (SC_C) of the tryptic site and the SCs of the peptides containing this tryptic site as a missed cleavage site (SC_M). Then, tryptic sites were labeled as positive sites if (1) SC_N was at least 1, (2) SC_C was at least 1 and (3) SC_M was zero, and as negative sites if both SC_N and SC_C were zero and SC_M was at least 2. Previous studies have demonstrated that the cleavage probability of a tryptic site is influenced mainly by the adjacent amino acids.[18,19,26] Therefore, for each tryptic site, a 9-mer consisting of the tryptic site and four adjacent residues on each side was extracted. If the tryptic site was located on the N or C terminus of a protein, resulting in insufficient amino acids to form a 9-mer, the character Z was added to make up a 9-mer. Each character in the 9-mer, except for the tryptic site (arginine or lysine for trypsin enzyme), was converted into a 21-dimensional binary vector that indicated whether one of the 20 amino acids or the character Z appears. In the binary vector, if one amino acid appeared, the corresponding position was set to 1, and other positions were set to 0. Thus, each 9-mer was converted into a 168-dimensional binary vector that retained both the position and the residue-specific information.

Random Forest Classifier for Cleavage Probability Prediction. We used the random forest algorithm to predict the cleavage probabilities of tryptic sites. A random forest is a nonlinear ensemble classifier consisting of a collection of independent unpruned trees.[27] Its randomness is reflected in two aspects:[27] random selection of a training subset for each tree by bootstrap and random selection of the feature for the best split at each node. To implement the random forest algorithm, we used the TreeBagger toolbox in MATLAB, which outputs a probability for each site to measure the probability of being cleaved. The number of trees was set to 200. The number of randomly selected features for each node was set to the square root of the number of all features.

Peptide Digestibility Calculation. The peptide digestibility, which is defined as the probability of the peptide being produced by the protein digestion process, is calculated from the predicted cleavage probabilities of tryptic sites using the following formula:

$$\text{peptide digestibility} = e_N{}^* e_C{}^* \prod_{i=1}^{n} (1 - e_i)$$

where e_N and e_C are the predicted cleavage probabilities of the N- and C-terminal tryptic sites of the peptide, respectively, e_i is the predicted cleavage probability of the i-th missed cleavage site in the peptide, and n is the total number of missed cleavage sites in the peptide.

Peptide Detectability Predictor. *Constructing the Training Set.* For each protein in the second half of the training data set, in silico digestion was performed with up to two missed cleavage sites and the length range of the digested peptides was set as the same as that of those identified peptides. The peptides with more than one SC were taken as positive peptides, and the unidentified digested peptides were taken as negative peptides. To date, the mechanisms underlying peptide detectability are still not clear, so we collected as many computable features of peptides as possible. The first was the peptide digestibility which was calculated for each peptide using the method described above. To our knowledge, the new feature peptide digestibility has never been used before. Other features are as follows: peptide length, number of missed cleavage sites, peptide molecular weight, and frequencies of 20 amino acids were calculated from the peptide sequence. A total of 544 amino acid-related physicochemical features from AAindex[15] were used. For each of the 544 AAindex features, the numerical values of the constituent amino acids in a peptide were averaged to produce a single value. In addition, 20 other features collected from previous studies[7,11,13,28,29] about peptide detectability were added. Finally, a total of 588 features were used to characterize each peptide (Table S1).

Random Forest Classifier for Peptide Detectability Prediction. AP3 employed another random forest classifier to predict the peptide detectability. The random forest toolbox and parameter settings were the same as used for predicting cleavage probability. In construction of the training set, the number of identified peptides was usually far less than the number of unidentified digested peptides. To overcome this imbalance problem, we adopted the down-sampling technique[30] to the negative peptides (i.e., the majority class). In brief, the same number of negative peptides as that of positive peptides were randomly selected from the whole set of negative peptides to form a balanced training set. Next, since the scales of the selected features were different, z-score normalization[31] was employed for each feature to obtain a zero mean and unity variance. To reduce the redundancy of features

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

and increase the computation efficiency, feature selection was performed using the minimum redundancy maximum relevance (mRMR) method[32] (Supporting Information, Note 1).

After peptide detectability prediction, the peptides for each protein of interest were sorted by their predicted detectabilities in descending order and the top peptides were selected as the proteotypic peptides of this protein.

RESULTS AND DISCUSSION

We trained the AP3 algorithm using the Yeast data set, tested its generalization ability on the *E. coli*, Mouse and Human data sets, and compared it with existing tools. In addition, we evaluated the contribution of the feature peptide digestibility to the performance of AP3.

Performance of Tryptic-Site Cleavage Probability Predictor. There were 3959 proteins with 43088 peptides identified in the Yeast data set. After filtering by the sequence coverages and SCs of the proteins, 1556 proteins with 29171 peptides remained. These proteins were randomly divided into two halves. The first half consisting of 778 proteins were used to train the cleavage probability predictor. Following the construction rules of the cleavage probability training set described in the "Experimental Section", 3867 positive tryptic sites and 2370 negative tryptic sites were obtained. The trained cleavage probability predictor had a 10-fold cross-validation AUC of 0.9708 (Figure 2), and the average AUC for the three test data sets was 0.9684 (Figure S3). These results demonstrated the accuracy of the cleavage probability predictor.

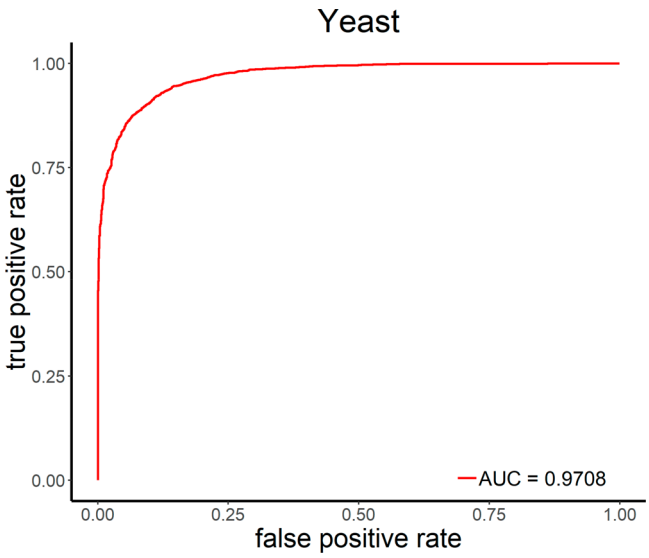


Figure 2. Ten-fold cross-validation ROC curve demonstrating the performance of cleavage probability predictor on the Yeast data set.

Performance of Peptide Detectability Predictor. In the second half of the Yeast data set used for training the peptide detectability predictor, there were 12850 identified peptides with SC > 1, which were taken as positive peptides, and the same number of negative peptides were randomly selected from the unidentified digested peptides with SC > 1. For each peptide, 588 features were calculated, including the peptide digestibility. Then, we selected 29 features in AP3 algorithm using the mRMR method (Supporting Information, Note 1).

The peptide detectability prediction model trained with the 29 selected features obtained a 10-fold cross-validation AUC of 0.9269 on the training set from the Yeast data set (Figure 3A).

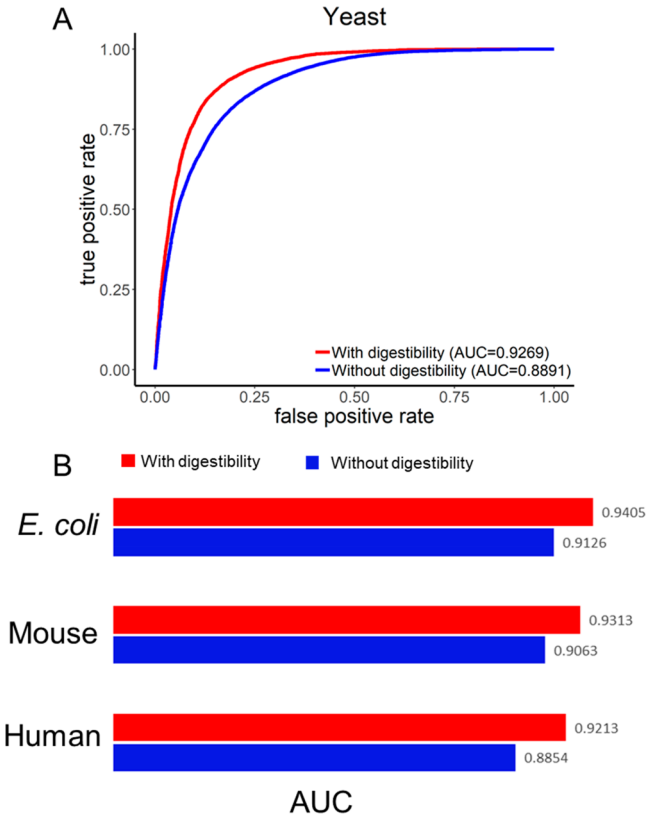


Figure 3. Performance comparisons of the AP3 algorithm with/without the peptide digestibility feature included in the peptide detectability model. (A) The 10-fold cross-validation ROC curves on the training data set (Yeast). (B) The AUCs on the three test data sets.

The AUCs for the three test data sets (*E. coli*, Mouse, and Human) were 0.9405, 0.9313, and 0.9213, respectively (Figure 3B). These results illustrated that our model retained good predictive power on independent data sets of other organisms. To demonstrate the generalization ability of AP3, we also tested the trained model on the three additive published data sets[33−35] with different LC columns and elution conditions (Table S3). The result demonstrated that LC columns and elution condition do not affect the good performance of AP3 (Figure S4). During the analysis, we trained the peptide detectability model 100 times independently and in each time used the down-sampling technique[30] to randomly select the same number of negative peptides as positive peptides. The results showed that our model has a good stability (Figure S5).

Notably, our proposed feature, peptide digestibility, showed an excellent ability in predicting the peptide detectability. First, peptide digestibility was the top one among the selected features and this conclusion had also been validated on the three test data sets (Table S2). Second, we compared the accuracy and generalization ability of the peptide detectability predictors with/without the peptide digestibility included in the selected features on the four data sets. Note that the model without the peptide digestibility feature was trained using all

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

the proteins after filtration, rather than half of these proteins. By including peptide digestibility, the 10-fold cross-validation AUC increased from 0.8891 to 0.9269 on the training data set (Figure 3A), and the AUCs increased from 0.9126 to 0.9405, from 0.9063 to 0.9313 and from 0.8854 to 0.9213 on the three test data sets (*E. coli*, Mouse, Human), respectively (Figure 3B). This conclusion based on the selected features was also confirmed by comparing the accuracy and generalization ability of the peptide detectability predictors with/without the peptide digestibility feature included in the whole set of features (without feature selection) on the four data sets (Figure S6). At last, we investigated the relationship between the peptide digestibility and the number of missed cleavage sites and found that they were indeed correlated with each other, but the former was much more powerful than the latter in predicting the peptide detectability (Supporting Information, Note 2 and Figure S7). By using the peptide digestibility alone, the random forest classifier exhibited a 10-fold cross-validation AUC of as high as 0.8266 (Figure S7B). All these results demonstrated that peptide digestibility is the most important feature for predicting proteotypic peptides.

Comparison with Existing Tools. To evaluate the performance of our algorithm, we compared AP3 with existing available tools, including PeptideSieve,[8] CONSeQuence,[11] ESP Predictor,[6] and PPA.[16] For all these tools, we used the available trained models provided by the authors of their publications. Note that PeptideSieve, CONSeQuence, and ESP Predictor were all trained on Yeast data sets, while PPA was trained on a human data set. PeptideSieve took protein sequences as the input in the FASTA format. The maximum number of missed cleavage sites was set to 2, and the maximum peptide mass was set to 6000 Da. We tested all types of PeptideSieve (PAGE_ESI, PAGE_MALDI, MUDPIT_ESI, MUDPIT_ICAT). CONSeQuence was run online (http://king.smith.man.ac.uk/CONSeQuence/), with the number of missed cleavage sites set to 2 and the prediction type set to ANN only. The ESP Predictor was also run online (https://genepattern.broadinstitute.org/) using the default parameters. The Perl script PPA.pl was downloaded from the Web site http://software.steenlab.org/rc4/PPA.php. The ROCs of different tools on the three test data sets are shown in Figure 4. The results showed that AP3 outperformed other tools in terms of true positive rate at arbitrary given false positive rate. On all the three test data sets, AP3 exhibited superior performance to other four tools (PeptideSieve, ConSequence, ESP Predictor, and PPA), increasing the AUC by 10.3%, 16.3%, 34.7%, and 25.4% on average on the three test data sets, respectively. Further, among these tools, only PPA provided retraining solution. Thus, we also retrained PPA on our Yeast data set and compared its performance with AP3 on the three test data sets. The result still supported that AP3 is significantly more accurate than PPA (Figure S8).

MRM Assay Validation. One direct application of a peptide detectability predictor is selecting proteotypic peptides for targeted proteomics assays. We applied AP3 and other four available tools to an MRM data set published by Fusaro et al.[6] Briefly, this data set consists of 14 proteins, each of which has several experimentally validated proteotypic peptides. We first predicted the peptide detectabilities of all in silico digested peptides of these proteins and then measured the performances of different algorithms in terms of the protein sensitivity, which is defined as the percentage of the proteins which have at least one validated proteotypic peptides included in the top five peptides with the highest predicted detectabilities for each protein.[6] The protein sensitivity of AP3 was 100% (14/14), while the protein sensitivities of PeptideSieve, CONSeQuence, ESP Predictor, and PPA were all 93% (13/14; Table S4). The

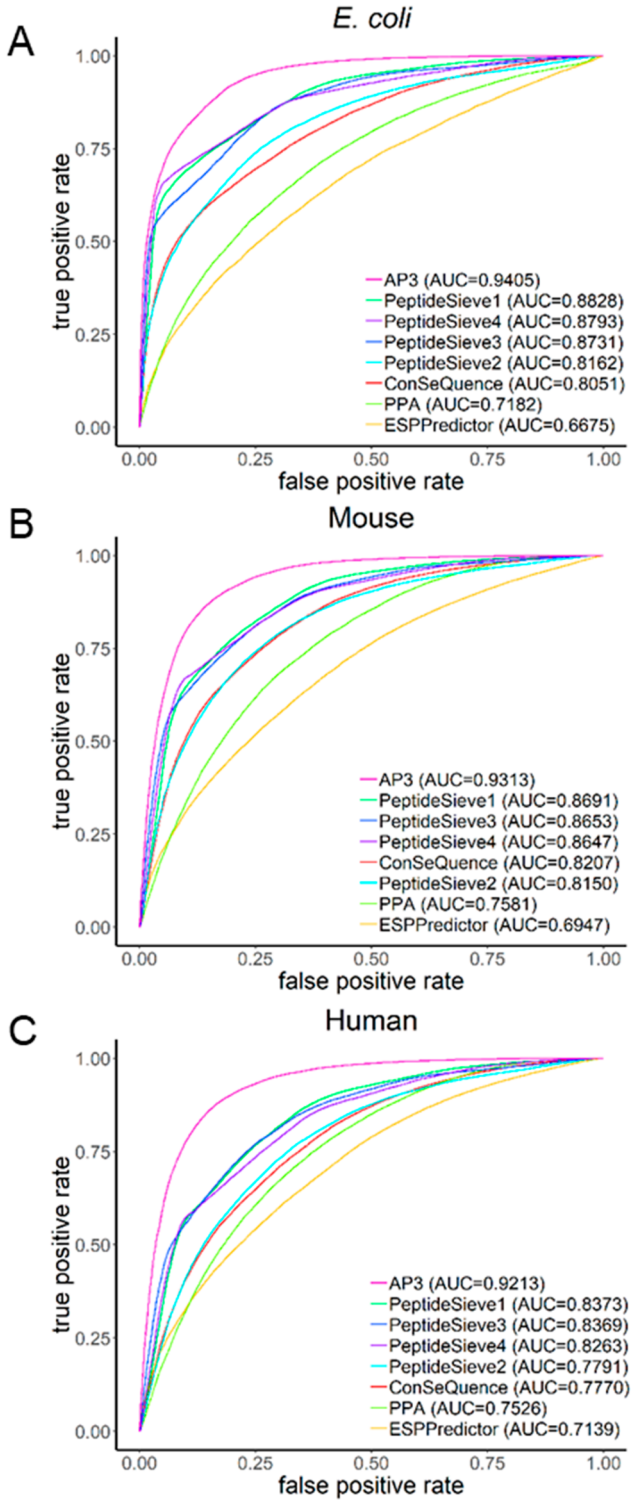


Figure 4. Performance comparisons between AP3 and other tools on the three test data sets (A) *E. coli*, (B) Mouse, and (C) Human, respectively. Abbreviations: PeptideSieve1, PeptideSieve_ICAT_ESI; PeptideSieve2, PeptideSieve_MUDPIT_ESI; PeptideSieve3, PeptideSieve_PAGE_ESI; and PeptideSieve4, PeptideSieve_PAGE_MALDI.

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

results indicated that AP3 was capable of accurately selecting proteotypic peptides for MRM-MS assays.

Software Availability. The AP3 software and the user guide document are freely available at http://fugroup.amss.ac.cn/software/AP3/AP3.html. For ease of use, we provide a graphical user interface for AP3.

CONCLUSIONS

In this study, we presented an algorithm, named AP3, for predicting peptide detectability. For the first time, we incorporated peptide digestibility into the peptide detectability prediction model as a novel and powerful feature. We demonstrated that peptide digestibility can greatly increase the accuracy of the peptide detectability prediction. AP3 enables the selection of candidate proteotypic peptides for the proteins of interest in the absence of high-quality MS-based experimental evidence, especially for the proteins identified by methods other than proteomics, such as genomic experiments or literature mining. This study may also have a significant effect on improving protein quantification, designing targeted proteomics assays, and discovering biological biomarkers for early diagnosis and therapy.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b02520.

1. Feature selection. 2. Relationship between peptide digestibility and the number of missed cleavage sites. The incremental feature selection curve was plotted by 10-fold cross-validation, as the top 50 features selected by mRMR were added successively to the random forest classifier (Figure S1); Validation of feature selection (Figure S2); ROC curves demonstrating the performance of cleavage probability model applied to the three test data sets (Figure S3); The performance of AP3 on three published data sets with different LC columns and elution conditions (Figure S4); Demonstration of the stability of peptide detectability model (Figure S5); Performance comparisons of the AP3 algorithm with/without the peptide digestibility feature included in all the 588 features (Figure S6); Comparative analysis of peptide digestibility and the number of missed cleavage sites (Figure S7); Performance comparison between AP3 and PPA on the three independent published data sets (A) *E. coli*, (B) Mouse, and (C) Human, respectively (Figure S8); Summary of the three additive published data sets with different LC columns and elution conditions (Table S3) (PDF)

The 588 features used to characterize the peptides in AP3 (Table S1) (XLSX)

The lists of selected features in four data sets by mRMR (Table S2) (XLSX)

Validation result on MRM assay data set (Table S4) (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: yfu@amss.ac.cn.

*E-mail: zhuyunping@gmail.com.

ORCID

Cheng Chang: 0000-0002-0361-2438

Yunping Zhu: 0000-0002-7320-7411

Yan Fu: 0000-0001-6896-5931

Author Contributions

[‖]These authors contributed equally to this work (Z.G. and C.C.). Y.F., C.C., and Y.Z. designed the algorithms and experiments. Z.G., C.C., and Y.J. implemented the algorithms and performed the data analysis. Z.G., C.C., and Y.F. wrote the manuscript. All authors edited and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Basic Research Program of China (2017YFA0505002 and 2015CB554406), the International S&T Cooperation Program of China (2014DFB30010), the Strategic Priority Research Program of CAS (XDB13040600), the National Natural Science Foundation of China (21605159), and the NCMIS CAS.

REFERENCES

(1) Parker, C. E.; Borchers, C. H. *Mol. Oncol.* **2014**, *8* (4), 840−858.

(2) Rifai, N.; Gillette, M. A.; Carr, S. A. *Nat. Biotechnol.* **2006**, *24* (8), 971−983.

(3) Demeure, K.; Duriez, E.; Domon, B.; Niclou, S. P. *Front. Genet.* **2014**, *5*, 305.

(4) Le Bihan, T.; Robinson, M. D.; Stewart, I. I.; Figeys, D. *J. Proteome Res.* **2004**, *3* (6), 1138−1148.

(5) Ethier, M.; Figeys, D. *J. Proteome Res.* **2005**, *4* (6), 2201−2206.

(6) Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A. *Nat. Biotechnol.* **2009**, *27* (2), 190−198.

(7) Webb-Robertson, B. J.; Cannon, W. R.; Oehmen, C. S.; Shah, A. R.; Gurumoorthi, V.; Lipton, M. S.; Waters, K. M. *Bioinformatics* **2008**, *24* (13), 1503−1509.

(8) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; et al. *Nat. Biotechnol.* **2007**, *25* (1), 125−131.

(9) Sanders, W. S.; Bridges, S. M.; McCarthy, F. M.; Nanduri, B.; Burgess, S. C. *BMC Bioinf.* **2007**, *8* (Suppl 7), S23.

(10) Wedge, D. C.; Kell, D. B.; Gaskell, S. J.; Lau, K. W.; Hubbard, S. J.; Eyers, C. *Gecco 2007 Genet. Evol. Comput. Conf.* **2007**, *12*, 2219−2225.

(11) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. *Mol. Cell. Proteomics* **2011**, *10* (11), M110.003384−M110.003384.

(12) Qeli, E.; Omasits, U.; Goetze, S.; Stekhoven, D. J.; Frey, J. E.; Basler, K.; Wollscheid, B.; Brunner, E.; Ahrens, C. H. *J. Proteomics* **2014**, *108*, 269−283.

(13) 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.

(14) Jarnuczak, A. F.; Lee, D. C. H.; Lawless, C.; Holman, S. W.; Eyers, C. E.; Hubbard, S. J. *J. Proteome Res.* **2016**, *15* (9), 2945−2959.

(15) Kawashima, S.; Ogata, H.; Kanehisa, M. *Nucleic Acids Res.* **1999**, *27* (1), 368−369.

(16) Muntel, J.; Boswell, S. A.; Tang, S.; Ahmed, S.; Wapinski, I.; Foley, G.; Steen, H.; Springer, M. *Mol. Cell. Proteomics* **2015**, *14* (2), 430−440.

(17) Siepen, J. A.; Keevil, E.-J.; Knight, D.; Hubbard, S. J. *J. Proteome Res.* **2007**, *6* (1), 399−408.

(18) Lawless, C.; Hubbard, S. J. *OMICS* **2012**, *16* (9), 449−456.

(19) Fannes, T.; Vandermarliere, E.; Schietgat, L.; Degroeve, S.; Martens, L.; Ramon, J. *J. Proteome Res.* **2013**, *12* (5), 2253−2259.

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711

(20) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol. Cell. Proteomics* **2014**, *13* (1), 339−347.
(21) Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrné, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M. *Nat. Biotechnol.* **2016**, *34* (1), 104−110.
(22) Malmström, E.; Kilsgård, O.; Hauri, S.; Smeds, E.; Herwald, H.; Malmström, L.; Malmström, J. *Nat. Commun.* **2016**, *7*, 10261.
(23) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. *Nature* **2014**, *509* (7502), 582−587.
(24) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. *J. Proteome Res.* **2011**, *10* (4), 1794−1805.
(25) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.
(26) Hubbard, S. J. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1998**, *1382* (2), 191−206.
(27) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.
(28) Braisted, J. C.; Kuntumalla, S.; Vogel, C.; Marcotte, E. M.; Rodrigues, A. R.; Wang, R.; Huang, S.-T.; Ferlanti, E. S.; Saeed, A. I.; Fleischmann, R. D.; et al. *BMC Bioinf.* **2008**, *9*, 529.
(29) Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. *Proteins: Struct., Funct., Genet.* **2003**, *52* (4), 573−584.
(30) Chen, C.; Liaw, A.; Breiman, L. *Using Random Forest to Learn Imbalanced Data*; Univ. of California, Berkeley, 2004; Vol. *110*, pp 1−12.
(31) Jain, A.; Nandakumar, K.; Ross, A. *Pattern Recognit.* **2005**, *38* (12), 2270−2285.
(32) Ding, C.; Peng, H. *J. Bioinf. Comput. Biol.* **2005**, *3* (2), 185−205.
(33) Ding, C.; Jiang, J.; Wei, J.; Liu, W.; Zhang, W.; Liu, M.; Fu, T.; Lu, T.; Song, L.; Ying, W.; et al. *Mol. Cell. Proteomics* **2013**, *12* (8), 2370−2380.
(34) Ge, S.; Xia, X.; Ding, C.; Zhen, B.; Zhou, Q.; Feng, J.; Yuan, J.; Chen, R.; Li, Y.; Ge, Z.; et al. *Nat. Commun.* **2018**, *9* (1), 1−16.
(35) Tyanova, S.; Albrechtsen, R.; Kronqvist, P.; Cox, J.; Mann, M.; Geiger, T. *Nat. Commun.* **2016**, *7*, 10259.

DOI: 10.1021/acs.analchem.9b02520
Anal. Chem. 2019, 91, 8705−8711