Integrated modeling of peptide digestion and detection for the prediction of proteotypic peptides in targeted proteomics

Zhiqiang Gao[1, 3||], Cheng Chang[2||], Yunping Zhu[2]*, Yan Fu[1, 3]*

1. NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
2. State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.
3. School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

|| Contributed equally to this work

* To whom correspondence should be addressed:

Yan Fu, Email: yfu@amss.ac.cn

Yunping Zhu, Email: zhuyunping@gmail.com

ABSTRACT

The selection of proteotypic peptides, i.e., detectable unique representatives of proteins of interest, is a key step in targeted shotgun proteomics. To date, much effort has been made to understand the mechanisms underlying peptide detection via mass spectrometry (MS) and to predict proteotypic peptides in the absence of empirical MS data. However, the performance of existing tools is still unsatisfactory. We find that one crucial reason is their neglect of the close relationship of protein proteolytic digestion to peptide detection in shotgun proteomics, although the former has been demonstrated to be predictable as well. To overcome this gap, we present here the first algorithm that considers the peptide digestion process for proteotypic peptide prediction. Specifically, our algorithm (named AP3) models the peptide digestion and detection processes in two successive steps and incorporates the predicted peptide digestion probability into the peptide detectability prediction model. Data analysis demonstrated that peptide digestion probability is the most important feature for the accurate prediction of proteotypic peptides in our model. Compared with the best available algorithms, AP3 showed 15.3%-17.5% higher accuracy. Finally, AP3 accurately predicted the proteotypic peptides for a targeted proteomics assay, showing great potential for assisting the design of targeted proteomics experiments.

Targeted proteomic assays, 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, i.e., peptides that are unique representatives of their corresponding proteins and possess good mass spectrometry (MS) detectability[3]. There are two major approaches for proteotypic peptide selection, i.e., the empirical approach and the computational approach. The former selects previously identified peptides as proteotypic peptides, while the latter predicts proteotypic peptides from the physiochemical properties of peptides. Although the empirical approach has been successfully applied in MRM-MS assays, it has some limitations. For example, not all target proteins have experimental evidence, especially proteins identified by literature mining. Moreover, there is randomness in peptide detection, i.e., some peptides identified in one experiment may not be identified in the next experiment. Thus, researchers are paying increased attention to the computational approach. However, the mechanisms underlying peptide detection are still unclear, which hinders the development of accurate algorithms to predict proteotypic peptides.

To date, much effort has been devoted to understanding the mechanisms underlying peptide detection. In early research, Le et al.[4] and Ethier et al.[5] proposed empirical score functions based on hydrophobicity, peptide length and isoelectric point. Recently, several algorithms[6,7,8,9,10,11,12] have been developed to predict proteotypic peptides or high-responding peptides based on supervised machine learning. In such algorithms, designing the features that describe peptides is a key problem. Many factors govern the likelihood of observing a peptide in a proteomics experiment, such as the physicochemical properties of the peptide, the abundance of the associated protein, and

the identification procedure[13,14]. Tang et al.[13] proposed the concept of peptide detectability, which was 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 algorithm based on 175 features derived from the peptide sequence to predict peptide detectability. Later, Sander et al.[9], Mallick et al.[8] and Eyers et al.[11] developed algorithms using 596, 1010, and 1186 features, respectively. These features included mainly AAindex[15]-derived features and other sequence-derived features. Recently, Muntel et al.[16] considered protein abundance as an additional feature and obtained improved performance. However, information on protein abundance is generally unavailable in the absence of experimental MS data.

A key limitation of the above algorithms is that they do not make (full) use of the protein proteotypic digestion information. As we know, a typical bottom-up proteomics experiment can be divided into two continuous processes: protein proteolytic digestion and peptide detection. An easily overlooked fact is that the products of protein proteolytic digestion are uncertain. That is, we do not know exactly which peptides and what proportions of them will be produced by digestion and undergo subsequent MS detection. Therefore, the accurate prediction of peptide detectability requires considering the process of protein proteolytic digestion. Previous studies have demonstrated that the commonly used enzyme trypsin exclusively cleaves the C-terminal to arginine or lysine[17], and this process is always incomplete[18]. In addition to the local conformation, tertiary structure and experimental condition, the cleavage probability of a tryptic site is mainly influenced by the amino acids surrounding the site[18]. Several algorithms have been proposed to predict the cleavage probability of tryptic sites from the adjacent amino acids[19,20].

Here, we propose a new machine learning algorithm named AP3 (short for Advanced Proteotypic Peptide Predictor) that, for the first time, integrates the peptide digestion and detection processes together to predict proteotypic peptides. Specifically, it incorporates the peptide digestion probability predicted in the first stage into the peptide detectability prediction model in the second stage. Training on a yeast dataset showed that peptide digestion probability was the most important feature for predicting proteotypic peptides, increasing the 10-fold cross-validation accuracy (area under the ROC curve, AUC) from 0.889 to 0.943. Furthermore, we demonstrated that the AP3 model trained on the yeast dataset retained predictive power over three independent comprehensive datasets, including *E. coli* (AUC 0.9383), mouse (AUC 0.9311) and human (AUC 0.9252), and exhibited superior performance (15.3%-17.5% higher in AUCs) to that of existing tools (PeptideSieve[8], CONSeQuence[11], ESP Predictor[6], and PPA[16]). Lastly, we showed that AP3 can be used to effectively select proteotypic peptides for MRM-MS assay development in the absence of experimental MS data.

Method

As Fig. 1 illustrates, AP3 has two major components: a digestion probability predictor and a peptide detectability predictor. A cleavage model is trained first to predict the cleavage probabilities of tryptic sites. Then, the peptide digestion probability is calculated and integrated into the peptide detectability predictor as a feature that characterizes peptides. Finally, feature selection is performed, and the peptide detectability prediction model is trained using the selected features.

Peptide digestion probability predictor

Constructing the training set. Identified peptides are mapped to their corresponding protein sequences. The cleavage information of tryptic sites in the identified protein sequences is collected, including the spectral count (SC) of the

peptides observed on the left of the tryptic site (L), the SC of the peptides observed on the right of the tryptic site (R) and the SC of the peptides in which this tryptic site is a missed cleavage (O). Then, tryptic sites are classified as positive sites if (1) L is at least 1, (2) R is at least 1 and (3) O is zero and as negative sites if both L and R are zero and O is at least 2.

To train a machine learning model, tryptic sites in the training set should be represented by numerical vectors. Previous studies have demonstrated that the cleavage probability of a tryptic site is influenced mainly by the amino acids adjacent to the tryptic site[21,20]. Therefore, for each positive or negative site, a 9-mer is taken consisting of the tryptic site and four residues on both sides. If the tryptic site is located on the N or C terminus of a protein, resulting in insufficient amino acids to form a 9-mer, the character Z is added to make up the 9-mer. Each character in the 9-mer, except for the tryptic site (arginine or lysine), is converted into a 21-dimensional binary vector that indicates whether one of the 20 amino acids or the character Z appears. For one binary vector, if one amino acid appears, the corresponding position is 1, and other positions are set to 0. Thus, each 9-mer is converted into one 168-dimensional binary vector that retains both the position and residue-specific information. The positive sites are labeled with 1, and the negative sites are labeled with 0.

Random forest classifier. A random forest is a nonlinear ensemble classifier consisting of a collection of independent unpruned trees[22]. The randomness of this algorithm is reflected in two aspects: randomly selecting a training subset for each tree by bootstrap and randomly selecting features for the best split at each node. The trained forest predicts the classification problem by letting all trees vote for the most popular class. The fraction of trees that vote high can be used as a probabilistic output, for example, the cleavage probability in our problem. Generally, as the number of trees

increases, the generalization error of the forest would decrease and converge gradually to a limit, but the running time would also increase rapidly[22]. Based on our test, using 200 trees achieves a good trade-off between accuracy and efficiency for our problem (data not shown). The number of randomly selected features for each node is set as its default value, that is, the square root of the number of all features. The sensitivity and specificity of the model are measured as the area under the receiver operating characteristic curve (AUC). The random forest model is implemented in MATLAB R2014a.

Calculating peptide digestion probability. For the remaining proteins, we perform an in silico digestion with a dataset-dependent peptide length range and up to 2 missed cleavage sites and then predict the cleavage probabilities of all tryptic sites associated with the digested peptides using the trained cleavage probability model. The dataset-dependent peptide length range means that the longest/shortest peptide length allowed is set to the longest/shortest length of peptides in the dataset, respectively. The digestion probability of a peptide, which is defined as the probability of the peptide being produced by the protein digestion process, is calculated according to the following formula:

$$e_{pep} = e_l * e_r * \sum_{i=0}^{n} (1 - e_i)$$

where e_{pep} is the digestion probability of this peptide, e_l and e_r are the cleavage probabilities of the left and right tryptic sites of the peptide sequence, respectively, e_i is the cleavage probability of the missed cleavage site i in the peptide sequence, and n is the number of missed cleavage sites in the peptide sequence.

Peptide detectability predictor

Constructing the training set. In the data preprocessing phase, we filtered proteins according to SC and sequence coverage to ensure high confidence in the remaining proteins. All the digested peptides of the remaining proteins theoretically have the chance of being observed. Identified peptides with SC larger than 1 are taken as positive peptides, and unobserved digested peptides are taken as negative ones.

To date, the mechanism underlying peptide detection is still not clear, so we collect as many related computer features as possible. Finally, a diverse set of 588 features are used to characterize each peptide (Supplementary Table 1). The peptide length, number of missed cleavage sites, peptide molecular weight and frequencies of 20 amino acids are first calculated according to the peptide sequence. Then, 544 amino acid-related physicochemical features from AAindex[15] are considered. For each of the 544 AAindex features, the numerical values of constituent amino acids in a peptide are averaged to produce a single value. Then, 20 additional features collected from previous studies[7,11,13,23,24] about peptide detection are added. Finally, we find that the peptide digestion probability can distinguish proteotypic peptides from unobserved digested peptides (KL distance = 3.09). The KL distance is a measure of the “distance” between two distributions. A larger KL distance indicates that the feature distribution of positive peptides is strongly different from the feature distribution of negative peptides. Thus, we add the peptide digestion probability to the feature set of the peptide detectability model for the first time.

Feature selection. To characterize the behavior of peptides in MS, 588 physicochemical features for each peptide are calculated. However, some features may be highly correlated with others. For example, peptide length and molecular weight have similar distributions in the positive set and the negative set (Supplementary Fig.

1). Therefore, feature selection is performed by using the minimum redundancy maximum relevance (mRMR) approach[25]. mRMR can sort features by considering the relevance to the dependent variable and the redundancy with higher-ranked features. The top 50 features are selected as candidate features. To select the minimum set of features with sufficient predictive performance, the top 50 candidate features are added into the peptide detectability model one by one in order. A 10-fold cross-validation strategy is adopted to evaluate the performance of the trained model as the number of added features increases.

Random forest classifier. The training set of peptide detectability models is usually an imbalanced set because the number of identified peptides is far less than the number of unobserved digested peptides in most cases. However, as the random forest algorithm cannot address this situation automatically, we employ the down sampling technique to handle imbalanced data sets[26]. In brief, the number of training samples for each class is set to the size of the minority class, and samples within the majority class are selected randomly together with the minority class to form a balanced training set. Because the scales of the selected features are different, z-score normalization[27] is employed for each feature to obtain a zero mean and unity variance. The random forest classifier is trained on the training set with selected features. The number of trees is set to 200, and the number of randomly selected features at each node takes the default value, that is, the square root of the number of all features. A 10-fold cross-validation AUC is calculated to evaluate the performance of our model. Then, the trained random forest model can be used to predict the detectability of digested peptides. Finally, we sort the peptides for each protein of interest in descending order by the peptide detectability and select the top peptides as the proteotypic peptides of this protein.

Software availability

AP3 is written in MATLAB and C++. The input of AP3 requires only the protein sequences in FASTA format, and the output file contains the detectability of all digested peptides of these proteins. The executable file and associated document are available at http://fugroup.amss.ac.cn/software/AP3/AP3.html.

Results

Datasets

In recent years, yeast has become the preferred model organism for studying proteomics[28]. Here, a public large-scale yeast dataset (Herbert, 2014)[21] was used as training data for our algorithm AP3. Descriptions of the sample preparations and experimental protocols are described in detail in (Herbert, 2014)[29]. To validate the generalization performance of our algorithm, we also used three publicly available datasets from other organisms: *E. coli* (Schmidt, 2016)[30], mouse (Malmstrom, 2016)[31] and human (Wilhelm, 2014)[32]. For the human dataset, the data of the lymph node and salivary gland were used. The raw files from the four public datasets were downloaded and reanalyzed. Table 1 provides a summary of the four public datasets used in our study.

MS data preprocessing

For the raw files produced by liquid chromatography coupled with tandem MS, peaks are searched against the corresponding organism sequences in the UniProt database using the Andromeda search engine included in the software MaxQuant[33] (version 1.6.0.1). Carbamidomethylation on cysteine is set as a fixed modification. Oxidation on methionine and protein N-terminal acetylation are set as variable modifications. Peptides are searched using fully tryptic cleavage constraints, and up to

two missed cleavage sites are allowed. The precursor mass tolerance is set to 20 ppm for the first search (for the identification of the maximum number of peptides for mass and retention time calibration) and 4.5 ppm for the main search (for the refinement of the identifications). The mass tolerance for fragment ions is set to 0.5 Da. False discovery rates at the protein and peptide levels are both set to 1%.

To construct a more precise training set, identified proteins are filtered to ensure that the remaining proteins are highly confident. We define the SC of a protein as the sum of its related peptide SCs. It is reasonable to assume that the larger the SC or sequence coverage of a protein, the higher is its confident. Therefore, proteins are sorted twice according to their SCs and sequence coverages respectively. Proteins that appear in the top 50% of both ranks remain for further analysis.

Performance of peptide digestion probability predictor

There were 3959 proteins with 43088 peptides identified in the yeast dataset. After filtering by the sequence coverages and SCs of proteins, 1556 proteins with 29171 peptides remained. Following the construction rules of the cleavage probability training set described in the Methods section, 7778 positive tryptic sites and 4854 negative tryptic sites were obtained. As shown in Supplementary Fig. 2, the trained cleavage probability model had a 10-fold cross-validation AUC of 0.975, and the average AUC for the three test datasets was 0.975. These results indicated that the cleavage probability predictor could accurately predict the cleavage probabilities of tryptic sites. The digestion probabilities of peptides were calculated based on the predicted cleavage probabilities of tryptic sites.

Feature selection for peptide detectability prediction

According to the down sampling technique, 25363 identified peptides with at least 2 SCs were taken as positive peptides, and 25363 peptides that were randomly selected

from the unobserved digested peptides were taken as negative peptides. For each peptide, 588 features were calculated, including the peptide digestion probability. Using the feature selection algorithm mRMR, 31 features were ultimately selected with the maximum AUC (Fig. 3 and Table 2). We grouped the 31 selected features into six categories: digestion, hydrophobicity, structural, charge, energy and other. There was broad agreement that the hydrophobicity, structure, charge and energy terms had a large impact on the peptide detection[6,7,8,9,11]. Hydrophobicity was represented by the hydrophobic residues and hydrophobicity coefficient in reversed-phase high-performance liquid chromatography (RP-HPLC). Nine of 31 selected features were related to secondary or tertiary structures, suggesting that peptide structure also influenced peptide detection. Charge played an important role in peptide detection, concurring with (Mallick, 2007), which demonstrated that peptide ionization, fragmentation, and detection were intimately linked to charge[8]. The selected feature "Activation Gibbs energy of unfolding" was consistent with a previous study, which claimed that Gibbs free-energy transfer between amino acids led to an increased response in peptides with nonpolar regions[34].

Notably, our proposed feature peptide digestion probability had good performance in the feature selection process. To illustrate the importance of the feature peptide digestion probability, we performed an in-depth analysis of this feature. First, the peptide digestion probability was the top feature among the selected features, and this conclusion had also been validated on the three test datasets (Supplementary Table 2). Second, we compared the generalization ability of peptide detection algorithms before and after the feature was included in the selected features on the three test datasets. By incorporating peptide digestion probability, the 10-fold cross-validation AUC increased 6.0% on the training dataset, and the AUCs increased 2.8%, 2.7% and 4.5% on the three

test datasets (*E. coli*, mouse, human), respectively (Fig. 4A, 4B). These results indicated that incorporating peptide digestion probability could significantly increase the accuracy and generalization ability of our model. Moreover, we calculated the detectability score for each proteotypic and undetected digested peptide for the human dataset with/without the feature peptide digestion probability. The different distributions between the peptide digestion probabilities of the proteotypic and undetected peptides and the larger scores for the proteotypic peptides indicated that our model was able to classify the proteotypic and undetected peptides (Fig. 4C). Importantly, 94.1% of proteotypic peptides received scores above 0.5, which indicated that the majority of proteotypic peptides could be predicted correctly. The above conclusions based on the selected features were also confirmed by comparing the generalization ability of peptide detection algorithms before and after the peptide digestion probability was included in all features on three test datasets (Supplementary Fig. 3). Third, using only one feature peptide digestion probability, the random forest classifier exhibited a 10-fold cross-validation AUC of 0.855 (Supplementary Fig. 4). The above analyses illustrated that peptide digestion probability was the most important feature for predicting proteotypic peptides.

Narrowing down the 588 features to the 31 selected features greatly decreased the computational complexity of our model and provided novel insights into the mechanism underlying peptide detection. To validate the feature selection performance, we compared the prediction performances between using all 588 features and using the 31 selected features. The results indicated that the simplified model using the 31 selected features had similar accuracy to that of the full model using all 588 features and showed better generalization ability than the full model (Fig. 5).

Relationship between peptide digestion probability and the number of missed cleavage sites

Integrating peptide digestion probability into the feature set of the peptide detectability model was inspired by the close relationship between protein proteolytic digestion and peptide detection and the fact that peptide digestion probability directly reflects the protein proteolytic digestion. However, other features may also have a relationship with trypsin digestion, such as the number of missed cleavage sites in the peptide sequence. To demonstrate that peptide digestion probability is a better representative of trypsin digestion involving peptide detection than the number of missed cleavage sites, we performed a comparative analysis of the two features. First, the peptide digestion probability distributions of peptides with different numbers of missed cleavage sites showed that the more missed cleavage sites the peptides had, the smaller the peptide digestion probability was (Supplementary Fig. 5). This result indicated that these two features were indeed negatively correlated with each other. Second, the KL distance was calculated to measure the distinguishing ability of peptide digestion probability and missed cleavage number for proteotypic peptides and unobserved digested peptides. The results showed that the KL distance of peptide digestion probability (3.09) was larger than that of the number of missed cleavage sites (1.66). Third, each of the two features was combined with the other 586 features to form two sets of 587 features. The two feature sets resulted in 10-fold cross-validation AUCs of 0.942 (the feature set including peptide digestion probability) and 0.891 (the feature set including missed cleavage site number), respectively. Finally, we also generated three models using the two features separately and together. As shown in Supplementary Fig. 4, the 10-fold cross-validation AUC of the model using both features was 0.865, the 10-fold cross-validation AUC of the model using only peptide

digestion probability was 0.855, and the 10-fold cross-validation AUC of the model using only the number of missed cleavage sites was 0.808. All the above results demonstrated that peptide digestion probability was a more powerful feature for peptide detectability prediction than the number of missed cleavage sites.

Performance evaluation of AP3

The peptide detectability prediction model was trained on the yeast dataset, and a 10-fold cross-validation strategy was applied to evaluate the performance of this model. AP3 obtained a 10-fold cross-validation AUC of 0.9428 on the yeast training dataset (Fig. 4A). Three independent public datasets (*E. coli*, mouse and human) were used to measure the generalization performance. The AUCs for the three independent test datasets (*E. coli*, mouse and human) were 0.9383, 0.9311, and 0.9252, respectively (Fig. 5). This result illustrated that AP3 retained predictive power on datasets in other organisms.

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]. PeptideSieve, CONSeQuence and ESP Predictor were all trained on a yeast dataset, while PPA was trained on a human dataset. 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 internal 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 website http://software.steenlab.org/rc4/PPA.php. We provided a file containing peptide

sequences as PPA.pl and obtained an output file that contained peptide detectability for every peptide.

The ROCs of different peptide detectability tools on the three test datasets are shown in Fig. 6. The results showed that AP3 outperformed the other tools with respect to the true positive rate at any given false positive rate. Compared with the four available tools, AP3 exhibited superior performance: average increases of 17.6%, 15.3% and 17.6% in AUC on three test datasets, *E. coli*, mouse and human, respectively.

MRM assay validation

One direct application of a peptide detectability algorithm is selecting proteotypic peptides for targeted proteomics assays. We applied AP3 and other peptide detectability tools to an MRM dataset published by Fusaro et al.[6]. Briefly, this dataset consists of 14 proteins, each of which has several experimentally validated MRM peptides. We first predicted the peptide detectability of all digested peptides of these 14 proteins, and then the protein sensitivity[6], which was the percentage of proteins with one or more peptides predicted by the predictor to be among the five highest responding peptides, was used to measure the performance. The protein sensitivity of AP3 was 100% (14/14), while the protein sensitivities of ESP Predictor, PPA, and PeptideSieve were all 93% (13/14) (Supplementary Table 3). The results indicated that AP3 was capable of accurately selecting proteotypic peptides for MRM-MS assays.

Discussion

In this study, we proposed an algorithm for predicting peptide detectability from a proteomics dataset. Considering protein proteolytic digestion when predicting peptide detectability is one of the major improvements provided by our algorithm. For the first time, we integrate the peptide digestion probability into the feature set of the peptide detectability model as a novel and significant feature. We demonstrate that

incorporating peptide digestion probability can significantly increase the performance of the peptide detectability model. Another advantage of AP3 is its generalization ability. Several previous studies[8,9] have noted that the peptide detectability model should be specific to experimental and instrumental conditions because a classifier constructed on one dataset does not perform well on another dataset. Our study here showed that this specialization may not be necessary. AP3 retained predictive power and had improved performance over existing tools on three independent test datasets (*E. coli*, mouse and human) using the trained model on the yeast dataset. Perhaps for certain specific conditions, the peptide detectability model should be retrained to be more adapted to their respective conditions. The third advantage of AP3 is its ease of use. AP3 needs only the sequences of proteins of interest as input, which facilitates its usage. For example, it enables the selection of candidate proteotypic peptides for proteins of interest in the absence of high-quality MS-based experimental evidence, especially for proteins identified by methods other than proteomics, such as genomic experiments or literature mining.

In summary, AP3 is a robust algorithm for the prediction of proteotypic peptides for a given protein based entirely on the peptide sequence and its neighboring regions in its parent protein. At the same time, we propose and demonstrate that peptide digestion probability is the most important feature for peptide detectability prediction. This study may have a significant effect on improving protein quantification, designing targeted proteomics assays, and developing biological biomarkers for early diagnosis and therapy.

References

1. Parker, C. E. & Borchers, C. H. Mass spectrometry based biomarker discovery, verification, and validation—quality assurance and control of protein biomarker assays. *Mol. Oncol.* **8,** 840–858 (2014).

2. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24,** 971 (2006).

3. Demeure, K., Duriez, E., Domon, B. & Niclou, S. P. PeptideManager: a peptide selection tool for targeted proteomic studies involving mixed samples from different species. *Front. Genet.* **5,** 305 (2014).

4. Le Bihan, T., Robinson, M. D., Stewart, I. I. & Figeys, D. Definition and Characterization of a “Trypsinosome” from Specific Peptide Characteristics by Nano-HPLC−MS/MS and in Silico Analysis of Complex Protein Mixtures. *J. Proteome Res.* **3,** 1138–1148 (2004).

5. Ethier, M. & Figeys, D. Strategy to Design Improved Proteomic Experiments Based on Statistical Analyses of the Chemical Properties of Identified Peptides. *J. Proteome Res.* **4,** 2201–2206 (2005).

6. Fusaro, V. A., Mani, D. R., Mesirov, J. P. & Carr, S. A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotech* **27,** 190–198 (2009).

7. Webb-Robertson, B. J. *et al.* A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics* **24,** 1503–1509 (2008).

8. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25,** 125–31 (2007).

9. Sanders, W. S., Bridges, S. M., McCarthy, F. M., Nanduri, B. & Burgess, S. C. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **8 Suppl 7,** S23 (2007).

10. Wedge, D. C. *et al.* Peptide Detectability following ESI Mass Spectrometry: Prediction using Genetic Programming. *Gecco 2007 Genet. Evol. Comput. Conf. Vol 1 2* 2219–2225 (2007).

11. Eyers, C. E. *et al.* CONSeQuence: Prediction of Reference Peptides for Absolute Quantitative Proteomics Using Consensus Machine Learning Approaches. *Mol. Cell. Proteomics* **10,** M110.003384-M110.003384 (2011).

12. Qeli, E. *et al.* Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *J. Proteomics* **108,** 269–283 (2014).

13. Tang, H. *et al.* A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22,** e481-8 (2006).

14. Jarnuczak, A. F. *et al.* Analysis of Intrinsic Peptide Detectability via Integrated Label-Free and SRM-Based Absolute Quantitative Proteomics. *J. Proteome Res.* **15,** 2945–2959 (2016).

15. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **27,** 368–369 (1999).

16. Muntel, J. *et al.* Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). *Mol Cell Proteomics* **14,** 430–440 (2015).

17. Olsen, J. V, Ong, S.-E. & Mann, M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Mol. Cell. Proteomics* **3,** 608–614 (2004).

18. Siepen, J. A., Keevil, E.-J., Knight, D. & Hubbard, S. J. Prediction of Missed Cleavage Sites in Tryptic Peptides Aids Protein Identification in Proteomics. *J. Proteome Res.* **6,** 399–408 (2007).

19. Lawless, C. & Hubbard, S. J. Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *OMICS* **16,** 449–56 (2012).

20. Fannes, T. *et al.* Predicting Tryptic Cleavage from Proteomics Data Using Decision Tree Ensembles. *J. Proteome Res.* **12,** 2253–2259 (2013).

21. Hubbard, S. J. The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1382,** 191–206 (1998).

22. Breiman, L. Random Forests. *Mach. Learn.* **45,** (2001).

23. Braisted, J. C. *et al.* The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* **9,** 529 (2008).

24. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins Struct. Funct. Genet.* **52,** 573–584 (2003).

25. Ding, C. & Peng, H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. in *Proceedings of the IEEE Computer Society Conference on Bioinformatics* 523-- (IEEE Computer Society, 2003).

26. Chen, C. & Liaw, A. Using random forest to learn imbalanced data. *Discovery* (2004).

27. Jain, A., Nandakumar, K. & Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognit.* **38,** 2270–2285 (2005).

28. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).

29. Hebert, A. S. *et al.* The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **13,** 339–347 (2014).

30. Schmidt, A. *et al.* The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* **34,** 104–110 (2016).

31. Malmström, E. *et al.* Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat. Commun.* **7,** 10261 (2016).

32. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

33. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

34. Cech, N. B. & Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **72,** 2717–2723 (2000).

Acknowledgments

This work was supported by the Strategic Priority Research Program of CAS (XDB13040600), the National Basic Research Program of China (2017YFA0505002 and 2015CB554406), the National Natural Science Foundation of China (21605159 and 21475150), the Innovation Program (16CXZ027), the International S&T Cooperation Program of China (2014DFB30010), and the NCMIS CAS.

Author contributions

Y.F and Y.Z. designed the algorithms and experiments. Z.G. and C.C. implemented the algorithms and performed the data analysis. Z.G. and C.C. wrote the initial manuscript. All authors edited and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figures

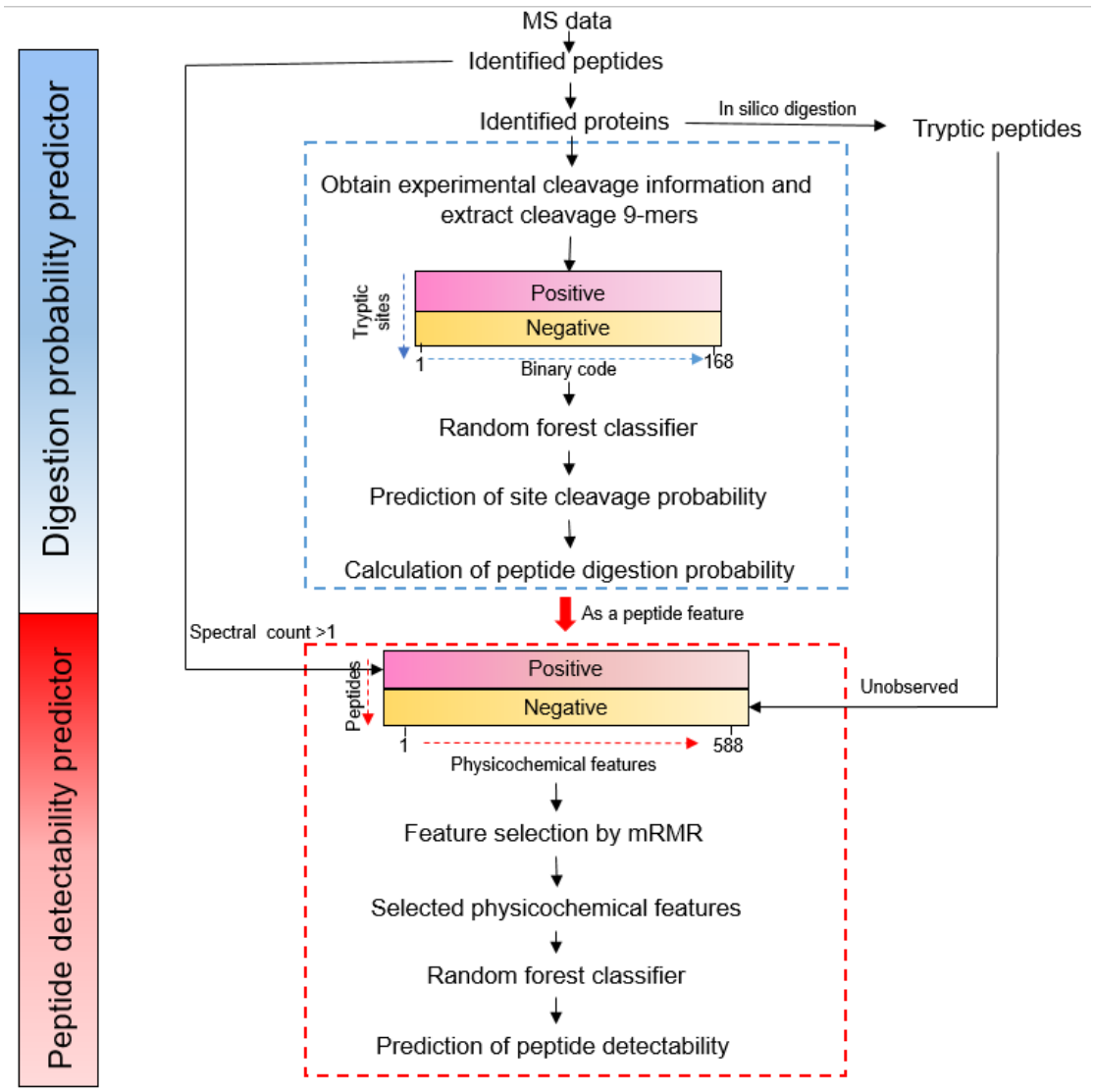


Figure 1. Overview of the development of the AP3 algorithm. AP3 has two major components: a digestion probability predictor (blue rectangle) and a peptide detectability predictor (red rectangle). Peptides identified in LC-MS/MS analysis are mapped to their corresponding protein sequences, and a training set is constructed from the cleavage information of the tryptic sites. Then, the digestion probability prediction model is trained by a random forest model to predict the cleavage probabilities of tryptic

sites and calculate the digestion probability of training peptides for the peptide detectability prediction model. Identified peptides with SC larger than 1 are taken as positive samples, and unobserved digested peptides are taken as negative samples. A 588-dimensional vector of physicochemical features is calculated for each training sample. Then, the peptide detectability prediction model is trained using the physicochemical features selected by mRMR and used to predict the detectability of peptides of proteins of interest given only their protein sequences.

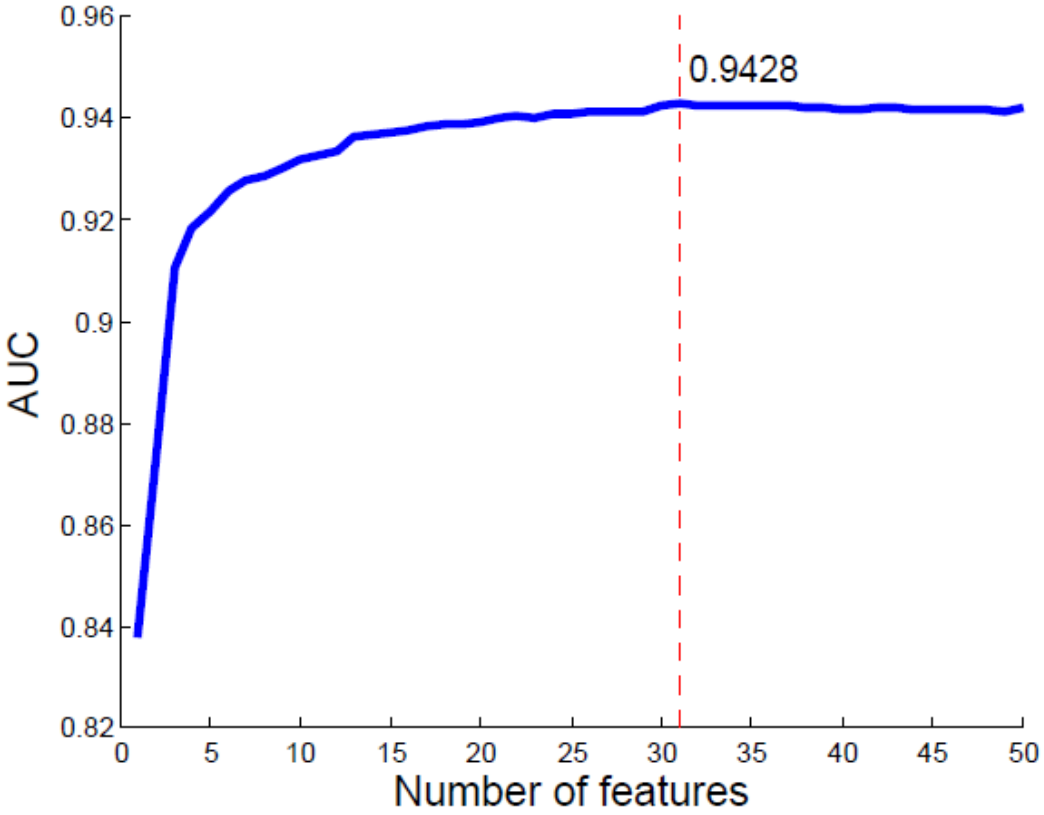


Figure 2. The incremental feature selection (IFS) curve was plotted by 10-fold cross-validation as the top 50 features selected by mRMR were added successively to the random forest model. Ultimately, we selected 31 features with a maximum AUC of 0.9428.

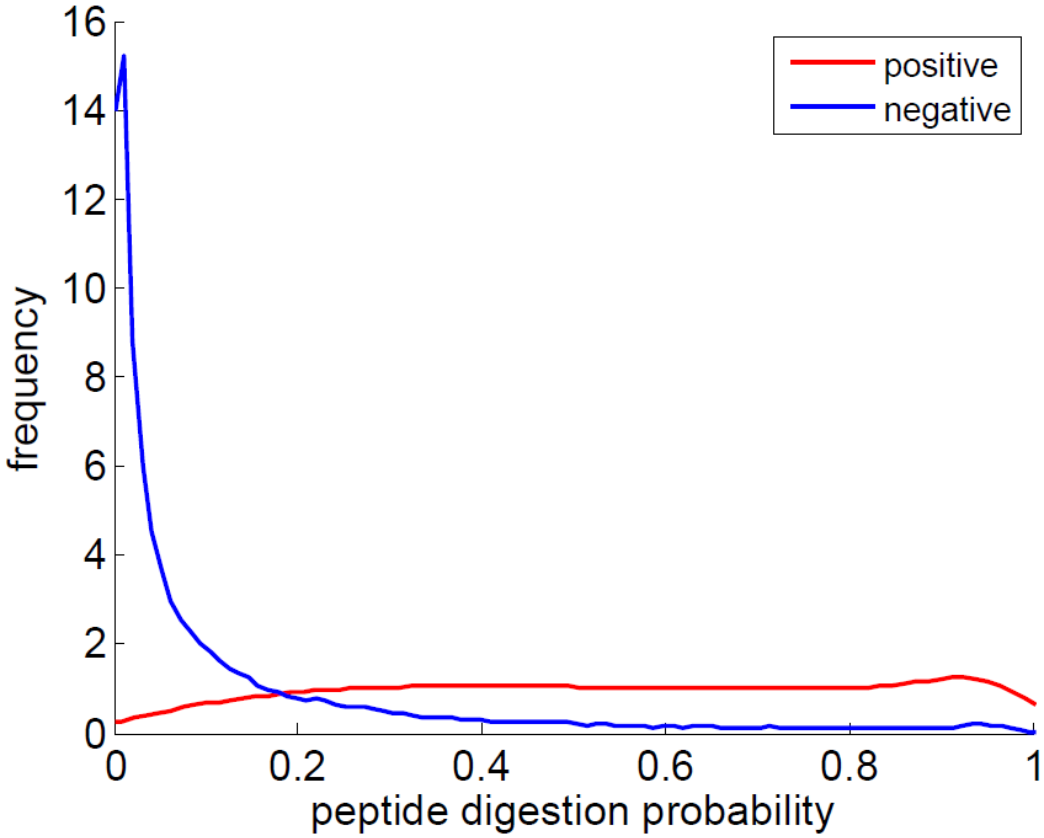
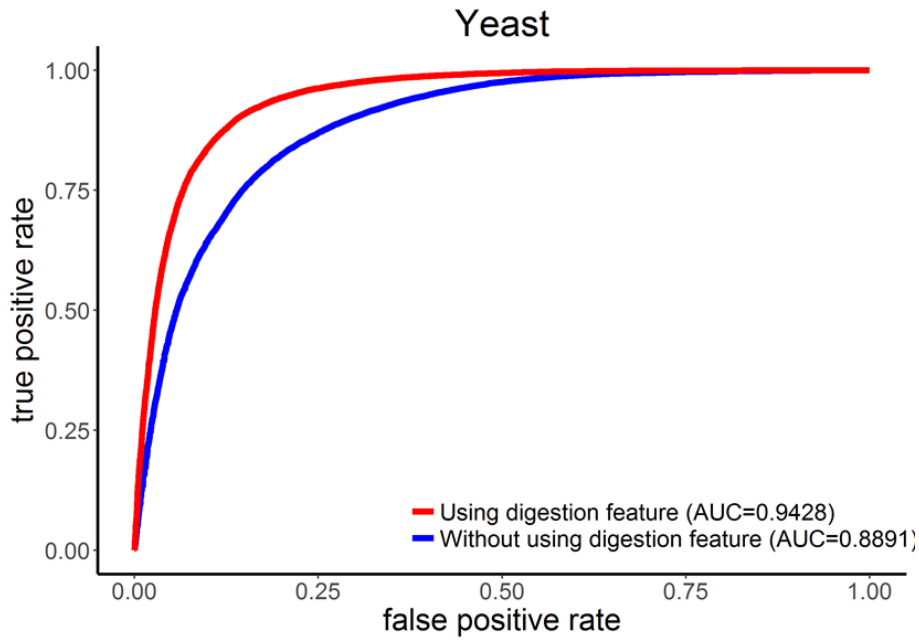
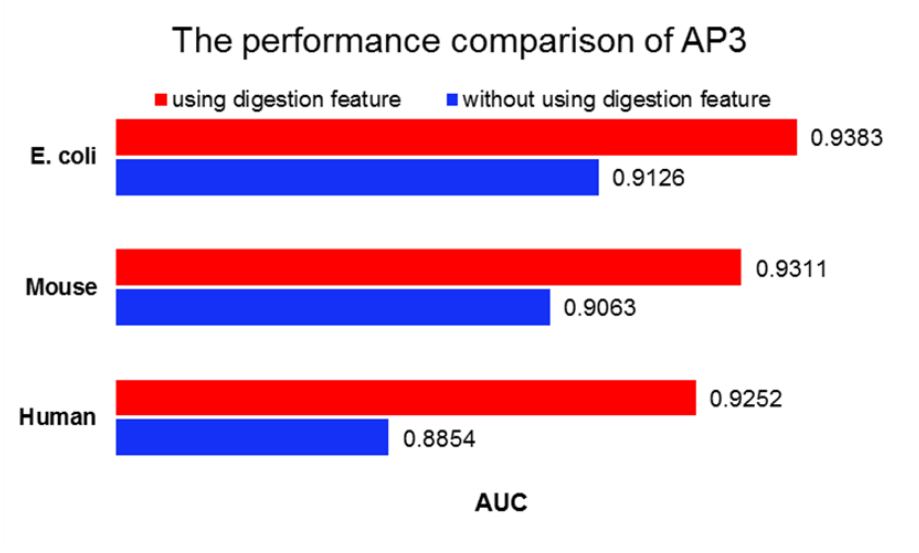


Figure 3. The distribution of the feature “peptide digestion probability” in the training set of the peptide detectability prediction model. The red and blue lines represent the distribution of the feature “peptide digestion probability” of the positive and the negative peptides, respectively.

A


Yeast
true positive rate
false positive rate
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Using digestion feature (AUC=0.9428)
Without using digestion feature (AUC=0.8891)

B


The performance comparison of AP3
using digestion feature
without using digestion feature
E. coli
0.9383
0.9126
Mouse
0.9311
0.9063
Human
0.9252
0.8854
AUC

C

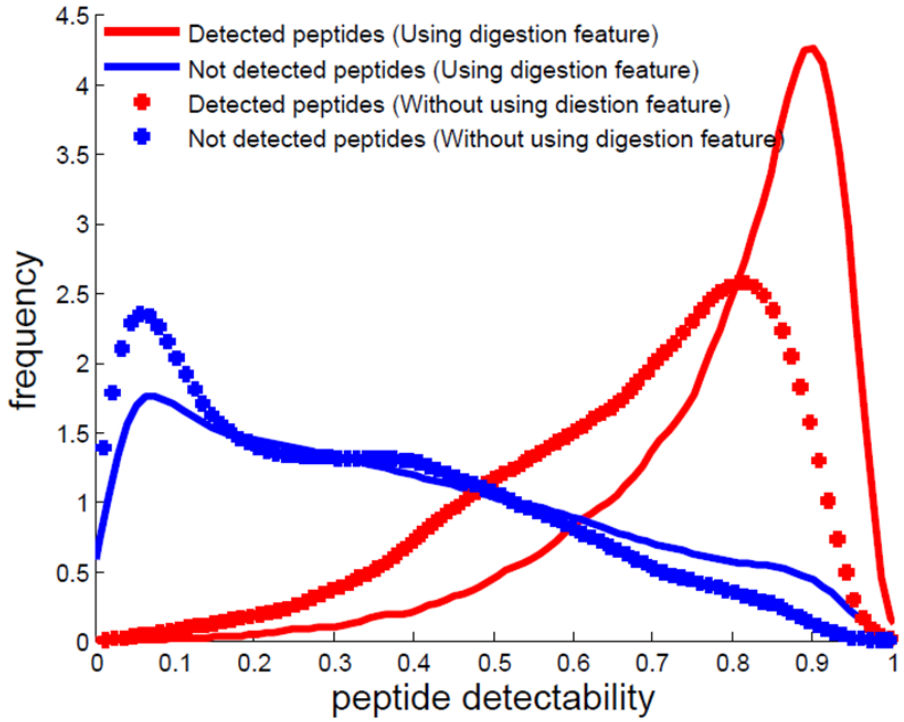
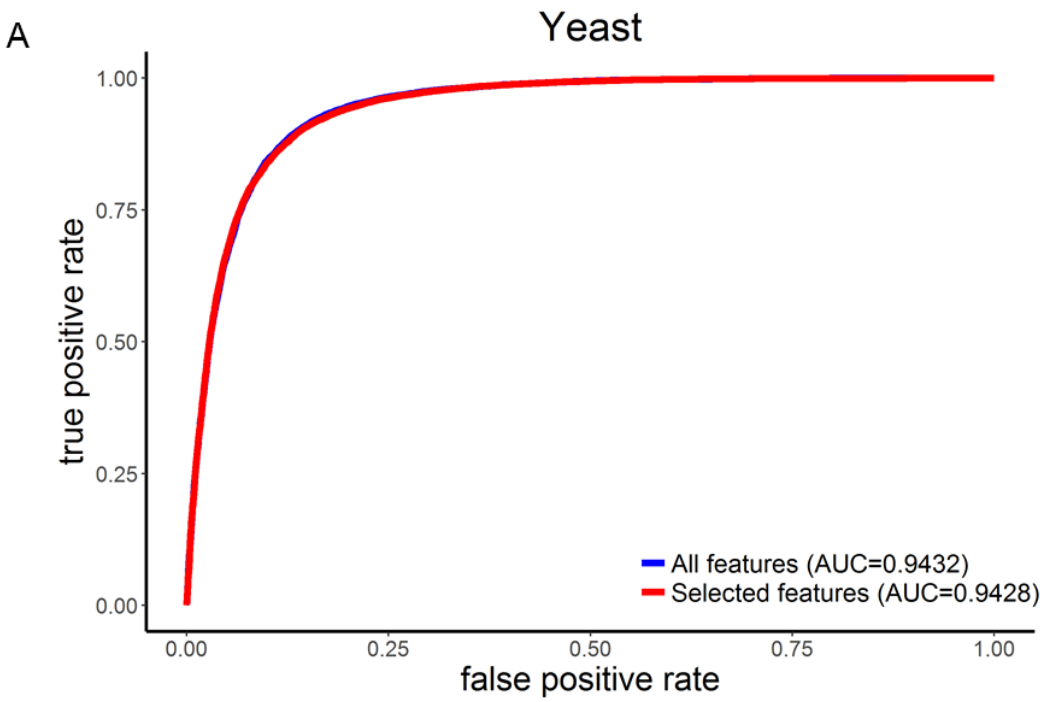

Detected peptides (Using digestion feature)
Not detected peptides (Using digestion feature)
Detected peptides (Without using diestion feature)
Not detected peptides (Without using digestion feature)
frequency
peptide detectability
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4. Comparison of the performance of peptide detectability predictors using (red) and not using (blue) the feature peptide digestion probability together with 30 other selected features. (A) The 10-fold cross-validation ROC curves were plotted for the peptide detectability model on the training set for two cases: using the digestion feature (peptide digestion probability) and without using this feature together with the 30 other selected features. (B) The AUCs for the peptide detectability model were calculated on three test datasets for two cases: using the digestion feature (peptide digestion probability) and without using this feature together with 30 other selected features. (C) Distributions of predicted peptide detectability for detected (red) versus undetected (blue) peptides of identified proteins for the human dataset for two cases: using the digestion feature (peptide digestion probability) and without using this feature together with 30 other selected features.



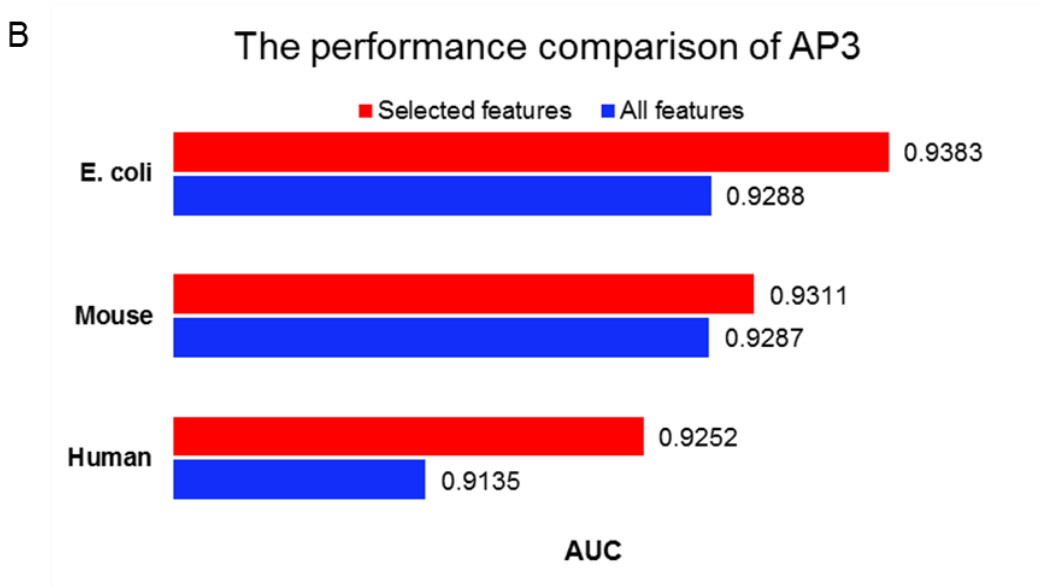


Figure 5. Validation of feature selection. (A) Comparison of the generalization abilities of all features-based model and selected features-based model on the training dataset. The red and blue lines represent the 10-fold cross-validation ROCs obtained using the 31 selected features and all 588 features, respectively. (B) Comparison of the generalization abilities of all features-based model and selected features-based model on three independent test datasets. The AUC values are given in the figure.

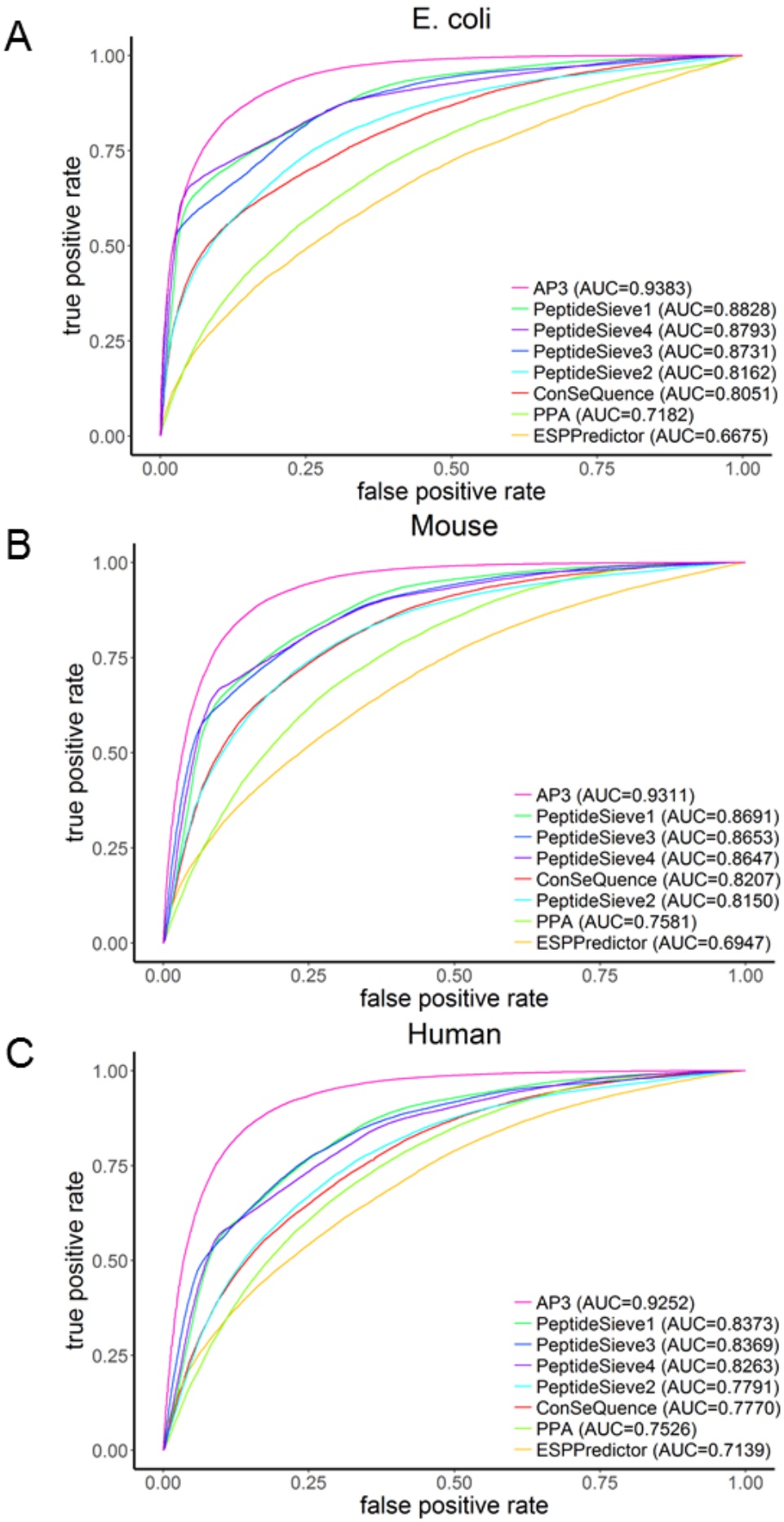

A
E. coli
true positive rate
false positive rate
0.00
0.25
0.50
0.75
1.00
AP3 (AUC=0.9383)
PeptideSieve1 (AUC=0.8828)
PeptideSieve4 (AUC=0.8793)
PeptideSieve3 (AUC=0.8731)
PeptideSieve2 (AUC=0.8162)
ConSeQuence (AUC=0.8051)
PPA (AUC=0.7182)
ESPPredictor (AUC=0.6675)
B
Mouse
AP3 (AUC=0.9311)
PeptideSieve1 (AUC=0.8691)
PeptideSieve3 (AUC=0.8653)
PeptideSieve4 (AUC=0.8647)
ConSeQuence (AUC=0.8207)
PeptideSieve2 (AUC=0.8150)
PPA (AUC=0.7581)
ESPPredictor (AUC=0.6947)
C
Human
AP3 (AUC=0.9252)
PeptideSieve1 (AUC=0.8373)
PeptideSieve3 (AUC=0.8369)
PeptideSieve4 (AUC=0.8263)
PeptideSieve2 (AUC=0.7791)
ConSeQuence (AUC=0.7770)
PPA (AUC=0.7526)
ESPPredictor (AUC=0.7139)

Figure 6. Performance comparison between AP3 and other tools on three independent test datasets. The AUCs are given in the figure legend. Abbreviations: PeptideSieve1: PeptideSieve_ICAT_ESI; PeptideSieve2: PeptideSieve_MUDPIT_ESI; PeptideSieve3: PeptideSieve_PAGE_ESI; and PeptideSieve4: PeptideSieve_PAGE_MALDI.

Tables

Table 1. Description of the four large-scale public datasets used for the development and evaluation of AP3

Dataset	yeast	human	*E. coli*	mouse
Reference	ref 1	ref 2	ref 3	ref 4
Species	*Saccharomyces cerevisiae*	*Homo sapiens*	*Escherichia coli*	*Mus musculus*
Instrument	Orbitrap Fusion	LTQ Orbitrap	LTQ Orbitrap Elite	Q Exactive Plus
Digestion	Trypsin	Trypsin	Trypsin	Trypsin
Protein groups observed	6422	3959	2898	5304
Peptides observed	80490	43088	38035	58184

References 1-4:

1. Hebert, A. S. *et al.* The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **13,** 339–347 (2014).
2. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587 (2014).
3. Schmidt, A. et al. The quantitative and condition-dependent Escherichia coli proteome. Nat. Biotechnol. 34, 104–110 (2016).
4. Malmström, E. et al. Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. Nat. Commun. 7, 10261 (2016).

Table 2. Selected features of the yeast dataset obtained using the mRMR method

Rank	Peptide feature
1	Peptide digestion probability
2	Nonpolar hydrophobic residues
3	The number of missed cleavage sites
4	Optimized propensity to form a reverse turn (Oobatake et al., 1985)
5	Normalized relative frequency of helix end (Isogai et al., 1980)
6	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H_2O
7	Basic residues
8	Cysteine
9	Weights for alpha-helix at the window position of 0 (Qian-Sejnowski, 1988)
10	Activation Gibbs energy of unfolding, pH 7.0 (Yutani et al., 1987)
11	Methionine
12	Information measurement extended without H-bond (Robson-Suzuki, 1976)
13	Net charge at pH 7
14	Tryptophan

15	Phenyalanine
16	Glutamine
17	pK-N (Fasman, 1976)
18	Serine
19	Tyrosine
20	Histidine
21	VL2C disorder (Vucetic et al., 2003)
22	Charged polar hydrophilic residues
23	Electron-ion interaction potential (Veljkovic et al., 1985)
24	Positive charge (Fauchere et al., 1988)
25	STERIMOL minimum width of the side chain (Fauchere et al., 1988)
26	Free energy change from epsilon (i) to epsilon (ex) (Wertz-Scheraga, 1978)
27	Threonine
28	Vihinen flexibility
29	Linker propensity from long dataset (linker length is greater than 14)
30	Molecular weight of peptide
31	AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa, 1992)

All the selected features are classified as digestion, hydrophobicity, structural, charge, energy and other.