An adaptive classification model for peptide identification

Xijun Liang1, Zhonghang Xia2*, Ling Jian1, Xinnan Niu3, Andrew Link3

From Fourth IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2014)
Miami Beach, FL, USA. 2-4 June 2014

Abstract

Background: Peptide sequence assignment is the central task in protein identification with MS/MS-based strategies. Although a number of post-database search algorithms for filtering target peptide spectrum matches (PSMs) have been developed, the discrepancy among the output PSMs is usually significant, remaining a few disputable PSMs. Current studies show that a number of target PSMs which are close to decoy PSMs can hardly be separated from those decoys by only using the discrimination function.

Results: In this paper, we assign each target PSM a weight showing its possibility of being correct. We employ a SVM-based learning model to search the optimal weight for each target PSM and develop a new score system, CRanker, to rank all target PSMs. Due to the large PSM datasets generated in routine database searches, we use the Cholesky factorization technique for storing a kernel matrix to reduce the memory requirement.

Conclusions: Compared with PeptideProphet and Percolator, CRanker has identified more PSMs under similar false discover rates over different datasets. CRanker has shown consistent performance on different test sets, validated the reasonability the proposed model.

Background

As the protein plays central roles in the interaction processes, identification and quantification of proteins in a variety of samples becomes a fundamental task in proteomics [1]. In the commonly used protein identification process, mass spectrometry (MS)-based strategies coupled with sequence database searching routinely generate a large number of peptide spectrum matches (PSMs), however, only a fraction of PSMs with high confidence scores are selected as true PSMs by using statistical and machine learning algorithms [2].

For peptide identification, a number of commercial and non-commercial database search tools [3-6] have been developed to rank the PSMs based on scoring functions and report top-scored ones as target PSMs. In the early stage, empirical filters [7,8] were described to validate the target PSMs, in which all above the defined thresholds are accepted as correct and those below the thresholds are assumed to be incorrect. However, the criteria for empirical filters may not be easily defined as scoring metrics used in database search tools, the quality of the mass spectrometry data, and the type of mass spectrometer used in the LC/MS/MS experiments vary.

Recently, machine learning approaches were introduced for improving the accuracy of discrimination between correct and incorrect PSMs based on PSM data models. A widely used algorithm, PeptideProphet [9], employs an unsupervised learning approach to identify correct and incorrect PSMs. In PeptideProphet, posterior probabilities of the PSMs are computed by using the expectation maximization (EM) method based on the assumption that these PSM data are drawn from a mixture distribution of correct and incorrect PSMs. Semi-supervised learning approaches exploit decoy data and use them as references for better estimation of discriminant scores. In [10], the PeptideProphet algorithm was
extended to incorporate decoy PSMs into a mixture probabilistic model at the estimation step of the EM with a semi-supervised learning framework. The restrictive parametric assumptions were removed by using the variable component mixture model and the semi-parametric mixture model. Percolator [11] is another advanced post-database searching method based on semi-supervised learning. The goal of Percolator is to increase the number of correct PSMs reported under the minimal FDR or q-value. Starting with a small set of trusted correct PSMs and a set of incorrect PSMs from searching a decoy database, Percolator iteratively adjusts the learning model to fit the dataset by ranking high-confidence PSMs higher than decoy peptide matches. The peptide identification can also be solved by a supervised learning approach which first trains a classifier with labels of PSMs already known and then uses the classifier to assign labels to those unknown PSMs [12]. In [13], a fully supervised SVM method is proposed to improve the performance of Percolator. Different with other supervised learning methods using decoy databases, De-Noise [14] labels all target PSMs as “correct”, but those low-scoring ones are treated as noises. The performance of a post-database search algorithm is usually evaluated by computing FDRs based on searching a target-decoy protein database [15-19].

De-Noise has shown its efficiency on eliminating incorrect target PSMs or noisy PSMs based on weights of the protease attributes. However, parameter selection is a big challenge in De-Noise. Based on the fuzzy SVM learning model, FC-Ranker [20] needs much fewer parameters and less input from the user than De-Noise does. FC-Ranker incorporates sample clustering procedure into the SVM classifier to estimate confidence on good target PSMs. Different with the traditional SVM model, in which the weight of training error is equally contributed by each data sample, FC-Ranker uses a fuzzy classification model to estimate the possibility of each target PSM being correct. The final score of each sample is determined by the combination of the value of discriminant function and fuzzy silhouette index. However, FC-Ranker does not provide an efficient method for calculating the weight of each PSM.

Similar to [20], we cast peptide identification as a binary classification problem in which “good” PSMs are labeled as “+1” and “bad” PSMs are labeled as “-1”. In this paper, to overcome the weight problem of FC-Ranker, we deal with the weight of training error as a variable, and employ the primal SVM technique [21] to re-formulate the classification problem as the CRanker classification model. In order to handle large PSM datasets, we use the Cholesky factorization technique to improve memory utilization in model training. A new scoring policy is proposed to rank all PSMs, and users can select those top-scored PSMs according to FDRs. The CRanker method has been validated on a number of PSM datasets generated from the SEQUEST database search tool. Compared with benchmark post-database search algorithms PeptideProphet and Percolator, CRanker has identified more “good” PSMs at the same false discovery rates (FDRs).

Methods

Peptide identification and classification problem
In sequence database searching, a large number of PSMs are routinely generated, however, only a fraction of them are correct. The task of peptide identification is to choose those correct ones from database search outputs. We formulate it as a binary classification problem, in which “good” PSMs are assigned to class “correct” or “+1” and “bad” PSMs to class “incorrect” or “-1”. Different with typical classification problems, the target PSMs are not trustworthy, i.e., ‘+1’ labels (corresponding to target PSMs) are not reliable. To overcome this problem, FC-Ranker introduces weight \( \theta_i \in [0,1] \) to indicate the reliability of \( i \)-th PSM, where 1 represents the highest confidence level and 0 the lowest confidence level. In fact, the learning model should rely more on reliable PSMs than untrustworthy ones.

Formally, the classification problem for peptide identification is described as follows. Given a set of \( l \) PSMs, denoted by \( \Omega = \{ (x_i, y_i) \}_{i=1}^l \subseteq \mathbb{R}^d \times \{-1, 1\} \) (Let \( \Omega = \{ (x_i, y_i) \}_{i=1}^l \subseteq \mathbb{R}^d \times \{-1, 1\} \) be a set of \( l \) PSMs), where \( x_i \in \mathbb{R}^d \) represents its \( i \)-th PSM record with \( q \) attributes, and \( y_i = 1 \) or \( -1 \) is the corresponding label indicating a target or decoy PSM. Let

\[
\Omega_+ = \{ y_i = 1 \}, \quad \Omega_- = \{ y_i = -1 \}.
\]

SVM-based classifiers have shown its advantages in peptide identification [14,20]. A typical SVM finds a discriminant function \( \Psi \) by solving

\[
\min_{\Psi} \sum_{i=1}^l \theta_i \text{Loss}(x_i, y_i) + c_1 ||\Psi|| \tag{1}
\]

where \( c_1 > 0 \) is a constant, \( \text{Loss}(x_i, y_i) \) is the loss function of \( (x_i, y_i) \), and \( ||\Psi|| \) is the norm of \( \Psi \) for regularization. In FC-Ranker, \( \theta_i \), \( i = 1, \ldots, l \) are treated as parameters and it is a challenge to determine their values.

In [20], Problem (1) is solved by the linear-programming SVM model as follows

\[
\min_{a,b,\xi,r} -r + c \sum_{i \in \Omega} \theta_i \xi_i
\]

s.t.

\[
\begin{align*}
\sum_{i=1}^l a_i y_i & \geq r - \xi_i, \quad i \in \Omega, \\
r & \geq 0, \\
-1 & \leq a_i \leq 1, \xi_i \geq 0, \quad i \in \Omega,
\end{align*}
\]

where \( \alpha \in \mathbb{R}^l, b \in \mathbb{R}^l, \xi = [\xi_1, \ldots, \xi_l] \in \mathbb{R}^l, \) and \( r \in \mathbb{R}^l \). Note that in this model, \( \theta_i \) is a parameter, and it is not trivial to choose a good one.
Cranker method

Cranker classification model

In this section, we cope with weight $\theta_i$ as a variable and reformulate Problem (1) as Cranker classification model. A new score scheme is developed for identifying correct PSMs based on Cranker solution. Note that all ‘-1’ labels (decoy PSMs) are reliable, and hence, $\theta_i = 1$, $i \in \Omega_-$. Moreover, we consider constraint $\sum_{i \in \Omega_-} \theta_i \geq \bar{\theta}$, where $\bar{\theta} > 0$ is a constant, to identify as many good PSMs as possible. Hence, we solve the following optimization problem:

$$\min_{\psi, \theta} \sum_{i=1}^{l} \theta_i \text{Loss}(\Psi(x_i), y_i) + c_1 \|\Psi\|$$

s.t. $\theta_i = 1$, $i \in \Omega_-$,
$$0 \leq \theta_i \leq 1, i \in \Omega_-,$$
$$\sum_{i \in \Omega_-} \theta_i \geq \bar{\theta},$$

where $c_1 > 0$ is a constant.

Technically, we move $\sum_{i \in \Omega_-} \theta_i \geq \bar{\theta}$ to the objective function, and reformulate model (3) as

$$\min_{\psi, \theta} \sum_{i=1}^{l} \theta_i \text{Loss}(\Psi(x_i), y_i) + c_1 \|\Psi\| - c_2 \sum_{i=1}^{l} \theta_i$$

s.t. $\theta_i = 1$, $i \in \Omega_-$,
$$0 \leq \theta_i \leq 1, i \in \Omega_-,$$

where $c_2 > 0$ is a constant.

By using the primal SVM technique [21], we formulate the Cranker classification model as

$$\min_{\psi, \theta} \beta^T k \phi + c_1 \sum_{i=1}^{l} \theta_i \cdot \left( \max \left( 0, -y_i k(x_i, \cdot) \right) \right)^p - c_2 \sum_{i=1}^{l} \theta_i$$

s.t. $\theta_i = 1$, $i \in \Omega_-$,
$$0 \leq \theta_i \leq 1, i \in \Omega_-,$$

where $K = (K_{ij})_{l \times l}$, $K_i = k(x_i, x_j)$, $k(\cdot, \cdot)$ is a given kernel, $K_i$ denotes the $i$-th column of $K$. The solution of model (5)

defines a discriminant function $\Psi(x) = \sum_{i=1}^{l} \beta_i k(x_i, x)$

Choose parameters $c_1$ and $c_2$

Parameters $c_1$ and $c_2$ play a critical role in determining the value of discrimination function $\Psi(x_i)$. We aim at $\Psi(x_i) \geq 0$ if $x_i$ is a correct target PSM and $\Psi(x_i) < 0$ otherwise. Notice that $y_i \geq 0$ for target PSMs, and $y_i < 0$ for decoys. We have $y_i \Psi(x_i) \geq 0$ for both correct target PSMs and decoys. Particularly, for $x_i$ with weight $\theta_i$, it contributes degree of confidence $-c_2 \theta_i$ to the value of the objective function in problem (5). Meanwhile, $x_i$ generates an empirical loss $c_1 \theta_i \eta_i$ where $\eta_i = \text{Loss}(y_i, \Psi(x_i)) = \max \left\{ 0, 1 - y_i k(x_i, \cdot) \right\}^p, p \geq 1$. In order to guarantee that the objective function of problem (5) decreases a certain amount, we enforce the loss $\theta_i (c_1 \eta_i - c_2) \leq 0$, which holds if and only if $0 \leq \eta_i \leq \frac{c_2}{c_1}$. It implies

$$y_i \Psi(x_i) = y_i k(x_i, \cdot) \beta \geq 1 - \left( \frac{c_2}{c_1} \right)^{1/p}.$$

Hence, if parameters $c_1$ and $c_2$ satisfy

$$\frac{c_2}{c_1} \leq 1,$$

we have $1 - \left( \frac{c_2}{c_1} \right)^{1/p} \geq 0$, and then $y_i \Psi(x_i) \geq 0$.

Moreover, if we choose parameters $c_1$ and $c_2$ such that $c_2 > 1$, then there exists a degeneration risk that $\beta = 0$ and $\theta_i = 1$ for all $i \in \Omega_+$ (i.e., all target PSMs are identified as correct), in which case we have objective function value $l(c_1 - c_2) < 0$.

Therefore, we always select parameters $c_2 \leq c_1$ in Cranker.

Cholesky factorization for large datasets

For large PSM datasets, the kernel matrix $K \in R^{l \times l}$ usually is not sparse, and thus, it is a big challenge to load whole $K$ in memory once. Usually, the number of sample features is much less than the number of samples, and kernel function $k$ provides a convenient and cheap transformation. We aim to design a low-rank approximation of large kernel matrix $K$ by Cholesky factorization, and request pairwise similarities between PSMs sequentially. Specifically,

$$K \approx L L^T$$

where $L \in R^{l \times r}, L_{ij} = 0$ if $i < j$, $L_{1,1} \geq L_{2,2} \geq \ldots \geq L_{r,r}$ are the square roots of the first largest $r$ eigenvalues of $K$. The details can be referred to [22].

Calculate the scores of PSMs

Based on Cranker discriminant function $\Psi(\cdot)$, we assign PSM $(x_i, y_i)$ a score

$$\text{score}(i) = \frac{2}{\pi} \arctan \left( \frac{\Psi(x_i)}{c_1} \right).$$

A large score value indicates the PSM is more likely to be correct. The PSMs are ordered according to their scores, and a certain number of PSMs are output to satisfy a pre-selected FDR.

Results and discussion

We evaluated the performance of Cranker by comparing it with PeptideProphet and Percolator based on PSMs generated from the SEQUEST search engine. The Cranker algorithm was implemented with Matlab version R2010b running on a PC with Intel Core i5-2400 CPU 3.10 GHz × 4 and 8 Gb RAM.

Experimental setup

Shotgun proteomics using multidimensional liquid chromatography coupled with tandem mass spectrometry were performed on all biological samples, including universal proteomics standard set (UPS1), the $S. cerevisiae$
Gcn4 affinity-purified complex (Yeast), \textit{S. cerevisiae} transcription complexes using the Tal08 minichromosome (Tal08) and Human Peripheral Blood Mononuclear Cells (PBMC). The RAW files generated from the different LC/MS/MS experiments were converted to mzXML format with the program ReadW. The MS/MS spectra were extracted from the mzXML file using the program MzXML2Search and all data was processed using the SEQUEST software. For PeptideProphet, we used the Trans Proteomic Pipeline V.4.0.2 (TPP), and the search outputs were converted to pep.XML format files using the TPP suite. For Percolator, we converted the SEQUEST outputs to a merged file in SQT format [23]. The UPS1 dataset, developed by Sigma-Aldrich company, contains 48 purified human proteins digested with trypsin. The SEQUEST search results include 17,335 PSMs, among which 8974 PSMs match target peptides and 8361 PSMs match decoy peptides. The Yeast dataset contains 6652 proteins and SEQUEST outputs 14,892 PSMs, among which 6703 PSMs match target peptides and 8189 PSMs match decoy peptides. For Tal08 complexes, the tryptic peptides were analyzed on an LTQ-Orbitrap XL (ThermoFisher) mass spectrometer using monoisotopic precursor selection (MiPS). It contains 69,560 PSMs, among which 42,222 PSMs match target peptides and 27,338 PSMs match decoy peptides. PBMCs were analyzed with both LTQ-Orbitrap XL and LTQ-Orbitrap Velos. A 6-step MuDPIT experiments was performed on a LTQ-Orbitrap XL using either MiPS (orbit-mips) or MiPS-off (orbit-nomips). The orbit-mips dataset contains 103,679 PSMs, including 68,334 targets and 35,345 decoys, and the orbit-nomips dataset contains 117,751 PSMs, including 76,395 targets and 41,356 decoys. For the LTQ-Orbitrap Velos experiments, 11-Step MuDPIT experiments were performed similar to Orbitrap XL experiments with either MiPS (velos-mips) or MiPS-off (velos-nomips). The velos-mips dataset contains 301,879 PSMs, including 208,765 targets and 93,114 decoys, and the velos-nomips dataset contains 447,350 PSMs, including 307,549 targets and 139,801 decoys. Samples are digested with trypsin. There are three types of tryptic peptides: full-digested, half-digested and none-digested. The detailed PSMs are summarized in Table 1.

Each dataset was divided into a training set and a test set according to 50/50 ratio. For large-sized datasets, such as Tal08 and PBMCs, we randomly select 20,000 samples from the training set for model training. This procedure was repeated \( n \) times, and let \( \Psi_1(x), i = 1, \ldots, n \) be the discriminant function for the \( i \)-th time.

Then, discriminant function

\[
\Psi(x) = \frac{1}{n} \sum_{i=1}^{n} \Psi_i(x)
\]

was employed in all experiments. We set as \( n = 6 \) in this paper. The PSM is represented by a vector of nine attributes: xcorr, deltacn, sprank, ions, hit mass, enzN, enzC, numProt, deltacnR. The first five attributes inherit from SEQUEST and the last four attributes are defined as

- \( \text{enzN} \): A boolean variable indicating whether the peptide is preceded by a tryptic site;
- \( \text{enzC} \): A boolean variable indicating whether the peptide has a tryptic C-terminus;
- \( \text{numProt} \): The number that the corresponding protein matches other PSMs;
- \( \text{deltacnR} \): deltacn/xcorr.

Weight 1.0 was assigned for xcorr and deltacn, and 0.5 for all others. In CRanker learning model, we set parameter \( c_1 \) and \( c_2 \) as 1.0, \( p \) as 2 and choose the Gaussian kernel with kernel argument \( \sigma = 1.0 \).

### Results

Table 2 shows that the total numbers of PSMs identified by CRanker, PeptideProphet, and Percolator over all datasets (training and test) at \( FDR = 0.05 \). As we can see, CRanker can identify more PSMs the other two algorithms.

Table 3 shows the performance of CRanker on test dataset. The last column of Table 3 indicates the ratios of PSMs identified on test set and whole dataset. As the

| Table 1. Statistics of datasets |
|--------------------------------|
| **Target** | **Decoy** |
| **Total** | **Full** | **Half** | **None** | **Total** | **Full** | **Half** | **None** |
| UPS1 | 17335 | 8974 | 645 | 2013 | 6316 | 8361 | 236 | 2588 | 5537 |
| Yeast | 14892 | 6703 | 1453 | 1210 | 4040 | 8189 | 106 | 1465 | 6618 |
| Tal08 | 69560 | 42222 | 14983 | 6809 | 20520 | 27338 | 419 | 5877 | 21042 |
| orbit-mips | 103679 | 68334 | 26760 | 15647 | 35915 | 62446 | 93114 | 24682 | 65912 |
| orbit-nomips | 117751 | 76395 | 28561 | 17490 | 62446 | 93114 | 24682 | 65912 |
| velos-mips | 301879 | 208765 | 110404 | 35915 | 62446 | 93114 | 24682 | 65912 |
| velos-nomips | 447350 | 307549 | 134117 | 77052 | 96380 | 139801 | 3414 | 34985 | 101402 |

Full, Half, None: number of PSMs with full-digested peptides, half-digested PSMs and none-digested PSMs, resp.
training data is randomly chosen, 50% ratio is an ideal scenario. On four PBMC datasets, the ratios are very close to 50%, indicating that CRanker classifier learned from training data works for the whole dataset. CRanker has shown very close learning performance on all datasets except UPS1. CRanker slightly overfitted on the test dataset of UPS1 (43.26%) as the training dataset is relatively small.

We have also looked at overlapping PSMs among PeptideProphet, Percolator and CRanker. Figure 1 shows the overlap of the identified target PSMs by the three methods on ups1, yeast, tal08 and 4 PBMC datasets. On all the datasets, the target PSMs output by CRanker have large overlap with PeptideProphet and Percolator. The details are list in Table 4. On ups1, PeptideProphet has 497 (87.8%) target PSMs shared by CRanker; Percolator has 390 (89.0%) target PSMs shared by CRanker. On all the other 6 datasets, these percentages exceed 90%. The results indicate that the majority of PSMs validated by PeptideProphet and Percolator were also validated by CRanker.

We finally compared the performance of CRanker, PeptideProphet, and Percolator by receiver operating characteristic (ROC). Due to the space limit, we included only ROCs on orbit-nomips (Figure 2) and velos-nomips (Figure 3) datasets. As we can see, CRanker reaches highest true positive rates (TPRs) throughout all false positive rates (FPRs) levels among the three algorithms in both figures.

Stability of CRanker
As training data points are randomly chosen from training datasets, the performance of CRanker classifier may vary slightly. We counted the outputs of CRanker in 20 runs on orbit-mips and velos-mips datasets.

Let $P_i$ and $\#P_i$ be the set of PSMs and the number of PSMs identified by CRanker at $i$-th run, $i = 1, \ldots, m$. We compared the similarity of $P_i$ and $P_j$, $i \neq j$, $i, j = 1, \ldots, m$ by

$$s_{ij} = \frac{1}{2} \left( \frac{\#(P_i \cap P_j)}{\#P_i} + \frac{\#(P_i \cap P_j)}{\#P_j} \right).$$

Then the stability of CRanker on a dataset is defined as the mean of all pairwise similarities over $m$ runs:

$$S = \frac{1}{m} \sum_{i, j=1, i \neq j}^{m} s_{ij}.$$

Table 5 and Table 6 show the numbers of PSMs identified by CRanker in 20 runs on orbit-mips and velos-mips, respectively. The stability of CRanker is $S = 99.17\%$ on orbit-mips and $S = 99.53\%$ on velos-mips.

Table 2. Target PSMs output by PeptideProphet, Percolator, and CRanker

| Data      | Method   | Total | TP   | FP  |
|-----------|----------|-------|------|-----|
| ups1      | PepProphet | 382   | 566  | 16  |
|           | Percolator | 450   | 438  | 12  |
|           | CRanker   | 601   | 585  | 16  |
| yeast     | PepProphet | 1481  | 1443 | 38  |
|           | Percolator | 1429  | 1394 | 35  |
|           | CRanker   | 1491  | 1455 | 36  |
| tal08     | PepProphet | 16025 | 15638 | 387 |
|           | Percolator | 14725 | 14371 | 354 |
|           | CRanker   | 16806 | 16390 | 416 |
| orbit-mips| PepProphet | 34035 | 32233 | 802 |
|           | Percolator | 34118 | 32270 | 848 |
|           | CRanker   | 35003 | 34123 | 880 |
| orbit-nomips | PepProphet | 36542 | 35673 | 869 |
|           | Percolator | 36962 | 36096 | 866 |
|           | CRanker   | 37337 | 36416 | 921 |
| velos-mips| PepProphet | 123908 | 120961 | 2947 |
|           | Percolator | 125701 | 122568 | 3133 |
|           | CRanker   | 125783 | 122624 | 3159 |
| velos-nomips | PepProphet | 180182 | 175789 | 4393 |
|           | Percolator | 178082 | 173719 | 4363 |
|           | CRanker   | 183492 | 178900 | 4592 |

PepProphet, Percolator, and CRanker denote the PSMs identified by PeptideProphet, Percolator, and CRanker, respectively.
Conclusion
We have proposed a new scoring system CRanker for peptide identification, in which the confidence on each PSM is taken into account in the model training process. CRanker employs the primal SVM technique and copes with the weight of each PSM as a variable. We use the Cholesky factorization technique to improve memory utilization in model training for large datasets.
PSM datasets. The performance of CRanker has been compared with benchmark algorithms PeptideProphet and Percolator over a variety of PSM datasets. The experimental studies show CRanker outperforms the other two by identifying more targets at the same FDRs.
### Table 5. Number of PSMs identified in 20 runs (orbit-mips)

|   | TP     | FP     |   | TP     | FP     |
|---|--------|--------|---|--------|--------|
| 1 | 33756  | 882    | 2 | 33772  | 866    |
| 3 | 33723  | 854    | 4 | 33756  | 882    |
| 5 | 33756  | 882    | 7 | 33673  | 844    |
| 9 | 33689  | 836    | 10| 33747  | 891    |

Average elapsed time for each run: 1848.9 s.

### Table 6. Number of PSMs identified in 20 runs (velos-mips)

|   | TP     | FP     |   | TP     | FP     |
|---|--------|--------|---|--------|--------|
| 1 | 122272 | 3158   | 2 | 122252 | 3178   |
| 3 | 122421 | 3189   | 5 | 122287 | 3163   |
| 7 | 122009 | 3068   | 9 | 122001 | 3076   |
| 10| 122284 | 3146   | 20| 122034 | 3043   |

Average elapsed time for each run: 2854.5 s.

### Abbreviations

PSM: peptide spectrum match; SVM: support vector machine; ROC: receiver operating characteristic; FDR: false discovery rate.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

XL and ZX designed the CRanker classification model and wrote the manuscript. LJ and XL designed the parameter selection and experiments. XL and ZX designed the CRanker classification model and wrote the

### Acknowledgements

The proteomics data was generated with support from NIH grant GM064779 and Vanderbilt University School of Medicine IDEA's Program grant 1-04-066-9530. AJL and XN were supported by NIH grant GM064779. LJ was partially supported by the Natural Science Foundation of China under Grant 61473011 and the Natural Science Foundation of Shandong Province under Grant ZR2014AP004.

### Declarations

Publication of this article was funded by the Natural Science Foundation of Shandong Province under Grant ZR2014AP004. This article has been published as part of BMC Genomics Volume 16 Supplement 11, 2015: Selected articles from the Fourth IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCBABS 2014): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S11.

### Authors' details

1. College of Science, China University of Petroleum, 66 Changjiang West Road, 266580 Qingdao, China. 2. Department of Computer Science, Western Kentucky University, 1006 College Heights Blvd, 42101 Bowling Green, USA. 3. Dept. of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, 37232 Nashville, USA.

### References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003, 422(6928):198-207.
2. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of Proteome Research 2007, 7(11):2092-2123.
3. Eng JK, McCormack AL, Yates Jr. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 1994, 5(11):976-989.
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20(18):3551-3567.
5. Geer LY, Markey SP, Kowalak JA, Wagner L, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res 2004, 3(5):958-64.
6. Craig R, Beavis RC. Tandem: matching proteins with tandem mass spectra. Bioinformatics 2004, 20(9):1466-1467.
7. Link A, Eng JI, Schieltz DM, Carmack E, Mize DJ, Morris DR, Garvik BM, Yates Jr. Direct analysis of protein complexes using mass spectrometry. Nature Biotechnology 1999, 17(7):676-682.
8. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology 2001, 19(3):242-7.
9. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. Analytical chemistry 2002, 74(20):5383-5392.
10. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. Journal of proteome research 2007, 7(1):254-265.
11. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods 2007, 4(11):923-925.
12. Anderson D, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide ms/ms spectra and sequest scores. Journal of proteome research 2003, 2(2):137-146.
13. Spivak M, Weston J, Bottou L, Kall L, Noble WS. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. Journal of proteome research 2009, 8(7):3737-3745.
14. Jian L, Niu X, Xia Z, Samir P, Sumanasekera C, Mu Z, Jennings JL, Hoek RL, Allos T, Howard LM, Edwards KM, Weil PA, Link AJ. A novel algorithm for validating peptide identification from a shotgun proteomics search engine. J Proteome Res 2013, 12(3):1108-1119.
15. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature methods 2007, 4(3):207-214.
16. Lam H, Deutsch EW, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. Journal of proteome research 2010, 9(11):605-610.
17. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. Journal of Proteome Research 2008, 7(1):47-50.
18. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. Journal of Proteome Research 2008, 7(1):29-34.
19. Higgi RE, Kniemad BM, Bonner Freeman A, Gelbert LM, Patil ST, Hale JE. Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. Journal of Proteome Research 2007, 6(5):1758-1767.
20. Liang X, Xia Z, Niu X, Link A, Pang L, Wu FX, Zhang H. Peptide identification based on fuzzy classification and clustering. Proteome Science 2013, 11(1):1-9, doi:10.1186/1477-5956-11-51.
21. Chapelle O. Training a support vector machine in the primal. Neural Comput 2007, 19(5):1155-1178, doi:10.1162/neco.2007.19.5.1155.
22. Fine S, Scheinberg K. Efficient svm training using low-rank kernel representations. J Mach Learn Res 2002, 2:243-264.
23. Bill N: SQF File Format [https://noble.gs.washington.edu/proj/crux/sqt-format.html].

doi:10.1186/1471-2164-16-S11-S1

Cite this article as: Liang et al. An adaptive classification model for peptide identification. BMC Genomics 2015 16(Suppl 11):S1.