Wu *et al. BMC Genomics* 2018, **19**(Suppl 6):567
https://doi.org/10.1186/s12864-018-4923-3

BMC Genomics

RESEARCH Open Access

A new estimation of protein-level false discovery rate

Guanying Wu[1†], Xiang Wan[2†] and Baohua Xu[1*]

From 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017)
Honolulu, Hawaii, USA. 30 May - 2 June 2017

Abstract

Background: In mass spectrometry-based proteomics, protein identification is an essential task. Evaluating the statistical significance of the protein identification result is critical to the success of proteomics studies. Controlling the false discovery rate (FDR) is the most common method for assuring the overall quality of the set of identifications. Existing FDR estimation methods either rely on specific assumptions or rely on the two-stage calculation process of first estimating the error rates at the peptide-level, and then combining them somehow at the protein-level. We propose to estimate the FDR in a non-parametric way with less assumptions and to avoid the two-stage calculation process.

Results: We propose a new protein-level FDR estimation framework. The framework contains two major components: the Permutation+BH (Benjamini–Hochberg) FDR estimation method and the logistic regression-based null inference method. In Permutation+BH, the null distribution of a sample is generated by searching data against a large number of permuted random protein database and therefore does not rely on specific assumptions. Then, *p*-values of proteins are calculated from the null distribution and the BH procedure is applied to the *p*-values to achieve the relationship of the FDR and the number of protein identifications. The Permutation+BH method generates the null distribution by the permutation method, which is inefficient for online identification. The logistic regression model is proposed to infer the null distribution of a new sample based on existing null distributions obtained from the Permutation+BH method.

Conclusions: In our experiment based on three public available datasets, our Permutation+BH method achieves consistently better performance than MAYU, which is chosen as the benchmark FDR calculation method for this study. The null distribution inference result shows that the logistic regression model achieves a reasonable result both in the shape of the null distribution and the corresponding FDR estimation result.

Keywords: FDR, Proteomics, Permutation, Null distribution

Background

In shotgun proteomics, the identification of proteins is a two-stage process: peptide identification and protein inference [1]. In peptide identification, experimental MS/MS spectra are searched against a sequence database to obtain a set of peptide-spectrum matches (PSMs) [2–4]. In protein inference, individual PSMs are assembled to infer the identity of proteins present in the sample [5–7].

Inferred proteins are the most biologically relevant outcome of a shotgun experiment. Therefore, the ability of accurately inferring proteins and directly assessing such inference results is critical to the success of proteomics studies. To date, many effective protein inference algorithms have been developed such as ProteinProphet, ComByne and MSBayesPro. However, the problem of accurate assessment of statistical significance of protein identifications remains an open question [8, 9]. Past research efforts towards this direction can be classified into *p*-value based approaches and false discovery rate (FDR) approaches:

*Correspondence: orthodontist_wu@163.com
†Guanying Wu and Xiang Wan contributed equally to this work.
[1]The Dental Center of China-Japan Friendship Hospital, Beijing, China
Full list of author information is available at the end of the article

- p-value based approaches provide a single protein-level p-value for each reported protein.
- FDR approaches apply a single threshold to all proteins identified from the data rather than generate individual significance values for each protein.

Both p-value based approaches and FDR approaches aim at controlling the quality of identified proteins, though they consider this problem from different perspectives. Unfortunately, available methods still deserve certain drawbacks, as summarized below:

1. **Reliance on specific assumptions.** Most methods depend on particular assumptions regarding the model or the distribution of false positive matches. For instance, the p-value based PROT_PROBE approach [10] assumes that protein identification by a collection of spectra follows a binomial model. Similarly, the representative of FDR approaches, MAYU [11] is based on the assumption that false positive PSMs are equally likely to map to either the target or decoy database and the number of false positive protein identifications is assumed to be hypergeometrically distributed.
2. **Reliance on the two-stage calculation process.** Generally, the protein-level confidence measure is obtained by combining peptide-level p-values (e.g., [8]). Such process may propagate errors at the peptide-level to the protein-level in a non-trivial manner [12].

Based on above observations, we propose a new framework for the protein-level FDR estimation that can avoid above-mentioned shortcomings. In this framework, we are permuting protein sequences and performing searching against these fake sequences on a dataset to get the corresponding null distribution at the protein-level before p-value and FDR calculation. Therefore, our calculation does not rely on the two-stage calculation process. In addition, we do not need to make any assumption on the distribution of protein identification scores since the permutation procedure is non-parametric. More importantly, once the null/permutation distribution is available, we can calculate p-values and the FDR without searching a decoy database. Experimental results on several real proteomics datasets show that our framework is effective in p-value and FDR calculation and outperforms MAYU consistently.

Although this framework is very appealing, the time required to perform the permutation procedure renders it infeasible to generate the null in an on-line manner before we have a fast permutation algorithm. To alleviate this problem, we suggest to do the permutation in an off-line manner and then store the null distributions for future use. When null distributions built on existing samples are not applicable in analyzing new-coming data with different features, we propose to use logistic regression to infer the null distribution from existing null distributions.

The rest of paper is organized as follows: "Methods" section illustrates the details of our methods. "Results and discussion" section presents the experiment results. "Discussion" section concludes the paper.

Methods

Overview of the methods

Proteins that are present in an experimental sample are true positives; others are false positives. Each protein is associated with a score measuring its confidence. The higher the score, more confident we are that the protein is in the sample. If we treat the protein score as a test statistic, the distribution formed by scores of false positives is the null distribution. Given N proteins, we can determine the p-value of each protein from the null distribution. For a certain FDR α, we can determine how many proteins are accepted based on their p-values according to the Benjamini-Hochberg (BH) method:

Algorithm 1 The Benjamini—Hochberg (BH) procedure [13]

1. Suppose $p_1, p_2, ..., p_N$ are ordered p-values associated with N proteins.
2. Fix the FDR α and let $p_1 \leq p_2 \leq ... \leq p_N$.
3. Define
$$L = \max\left\{c : p_c < \alpha \frac{c}{N}\right\}. \quad (1)$$
4. Accept all proteins l for which $p_l \leq p_L$.

Given a subset of proteins obtained by setting a protein score threshold, we can also determine the FDR according to the BH procedure.

In the method, the p-value calculation method and the BH procedure are well-established statistical routines. The major source of errors in estimating the FDR may come from the null distribution estimation. Generally, the more data we have, the more accurate the null distribution. Thus, we estimate the null distribution by using a permutation method, which can generate plenty of data for robust data analysis. However, the permutation method is inefficient. This method becomes computational expensive when handing large datasets. This motivates us to develop a method that can infer the null distribution of a new dataset from null distributions of known datasets. We can store the previous estimated null distributions and conduct the protein-level FDR estimation in an off-line mode. In this way, both accuracy and efficiency can be achieved. Details are provided in the following sections.

The permutation method

We employ the target-decoy technique to determine null distributions. In the permutation step, each sequence in the original protein database is randomly shuffled [14]. The shuffled proteins are appended into the original protein database to form a concatenated database. Then, proteins are identified from the concatenated target-decoy database. Protein identifications mapping to decoy sequences are false positives, whose scores are used to form the null distribution. When the sample size is small, we may not have enough false positives to form a reliable null distribution. Thus, the shuffling step and the protein inference step are repeated multiple times (e.g. 20 repeats). In each iteration, decoy protein scores are stored.

Suppose we obtain M decoy proteins in the above step. Let $\mathcal{Z} = \{z_1, z_2, ..., z_l, ..., z_M\}$ be the set of decoy protein scores. We partition the range of z_l values into K bins of equal length:

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k. \tag{2}$$

Here, $\mathcal{Z}_k$ contains protein scores belonging to the k-th bin. Define y_k as the count in the k-th bin:

$$y_k = \#\{z_s \in \mathcal{Z}_k\}. \tag{3}$$

and let x_k be the center point of $\mathcal{Z}_k$. Note that $\sum_{i=1}^{K} y_k = M$. Then, the set of points $\mathcal{H} = \{(x_1, y_1/M), ..., (x_K, y_K/M)\}$ describes the probability density function of the null distribution.

When a protein belongs to $\mathcal{Z}_{\hat{k}}, \hat{k} = 1, 2, ...K$. Then, its p-value can be approximated as:

$$p_{l, z_l \in \mathcal{Z}_{\hat{k}}} = \frac{\sum_{k=\hat{k}}^{K} y_k}{M}. \tag{4}$$

Given N proteins, we can estimate their p-values. The determination of the relationship of the FDR and the number of proteins is straightforward by applying the BH procedure mentioned in the previous section.

In the permutation method, we need to shuffle and identify proteins multiple times. Thus, the notable limitation of the permutation method is its low efficiency. Null distributions of different samples can be stored in the protein database for future use in an off-line mode.

A general null distribution inference model

The protein identification result can be affected by various reasons such as the tandem MS peak count and the sample complexity. Null distributions built on existing samples may not be applicable in analyzing data with different tandem MS peak counts and a different sample complexity. Determining the null distribution of new data is time consuming by applying the permutation method. Thus, we design a way to infer the null distribution from existing null distributions in the case that high efficiency is desired.

A raw data can be described by many features. For instance, tandem MS peak counts and tandem MS spectral quality measured by the mean noise level. Suppose we have I existing samples and each sample can be described by J features. Denote features of the i-th sample as $(r_{i,1}, r_{i,2}, ..., r_{i,J})$ and let $\mathcal{H}_i(x)$ be the null probability density function associated with the sample. The feature of a new sample is denoted as $(r_{0,1}, r_{0,2}, ..., r_{0,J})$ and our objective is to infer its null density function $\mathcal{H}_0(x)$.

For a protein score belonging to the k-th bin $\tilde{x} \in \mathcal{Z}_k$, we collect the following information from existing samples: Then, the relationship of the probability $\Pr_{i,k}(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_k)$ and J features can be described by the following logistic regression model:

$$\log\left(\frac{\Pr_{i,k}}{1 - \Pr_{i,k}}\right) = \beta_{k,0} + \sum_{j=1}^{J} \beta_{k,j} r_{i,j}, i = 1, 2, ..., I. \tag{5}$$

After fitting the logistic regression model, we estimate the the probability $\Pr_0(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_k)$ of the new sample as:

$$\Pr_0(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_k) = \frac{1}{1 + e^{-\left(\beta_{k,0} + \sum_{j=1}^{J} \beta_{k,j} r_{0,j}\right)}}. \tag{6}$$

For bins $\mathcal{Z}_1, \mathcal{Z}_2$,... and $\mathcal{Z}_{K-1}$, we collect information as shown in Table 1, conduct logistic regression by model (5) and obtain the fitting coefficients $\beta_{k,j} (k = 1, 2, ...K - 1; j = 0, 1, 2, ..., J)$ as shown in Table 2. It is unnecessary to perform logistic regression on the last bin $\mathcal{Z}_K$ because $\Pr_i(x \le \tilde{x} | \in \mathcal{Z}_K) = 1, i = 0, 1, 2, ..., I$. We use a coefficient table to store the information:

Then, the density function $\mathcal{H}_0(x)$ can be approximated as:

$$\mathcal{H}_0(x \in \mathcal{Z}_k) = \Pr_0(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_k) - \Pr_0(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_{k-1}). \tag{7}$$

When $k = 1$, $\mathcal{H}_0(x \in \mathcal{Z}_k)$ becomes ill-posed because $\mathcal{Z}_0$ is undefined. In this case, let $\Pr_0(x \le \tilde{x} | \tilde{x} \in \mathcal{Z}_0) = 0$. By using the coefficient table and equation (7), we obtain the null density function $\mathcal{H}_0$.

Table 1 The probability $\Pr_{i,k}(x \le \tilde{x} | \tilde{x} \in \boldsymbol{\mathcal{Z}}_k)$ can be calculated from the null probability density function $\boldsymbol{\mathcal{H}}_i(x)$

Sample	feature 1	...	feature j	...	feature J	$\Pr_i(x \le \tilde{x} | \tilde{x} \in \boldsymbol{\mathcal{Z}}_k)$
Sample 1	$r_{1,1}$	...	$r_{1,j}$	...	$r_{1,J}$	$P_{1,k}$
...	...	...	...	...	...	...
Sample i	$r_{i,1}$	...	$r_{i,j}$	...	$r_{i,J}$	$P_{i,k}$
...	...	...	...	...	...	...
Sample I	$r_{I,1}$	...	$r_{I,j}$	...	$r_{I,J}$	$P_{I,k}$

Table 2 The coefficient table for null distribution inference

	Intercept	feature 1	feature 2	...	feature J
1-th bin	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	...	$\beta_{1,J}$
...	...	...	...	...	...
k-th bin	$\beta_{k,0}$	$\beta_{k,1}$	$\beta_{k,2}$	...	$\beta_{k,J}$
...	...	...	...	...	...
$(K-1)$-th bin	$\beta_{K-1,0}$	$\beta_{K-1,1}$	$\beta_{K-1,2}$	...	$\beta_{K-1,J}$

The feature protein database

We can use a feature table to organize our data. A feature database is shown in Fig. 1. The feature database contains a protein database, which is used to perform protein inference. The null distributions in the feature database are obtained by the permutation method based on existing samples. For each existing sample, its features are extracted and stored in the feature table. The logistic regression coefficients are obtained by fitting the model (5) on the existing samples. The off-line strategy means that: To obtain the null distribution of a new sample, we neither need to apply the permutation method nor have to perform logistic regression fitting.

When a new raw data is input, protein inference and feature extraction are performed. We can use existing protein inference algorithms such as ProteinProphet to identify proteins from the protein database in the feature database. Then, we can compare the new sample with samples in the feature database based on their features. We can measure the similarity of two samples by calculating the correlation of their feature vectors. Similar samples are often encountered when we analyze replicate samples. If the similarity between the new sample and an existing sample i is high (e.g. the correlation of features is above 0.9), we use $\mathcal{H}_i(x)$ as the null distribution to calculate the protein-level FDR. If we cannot find any similar sample in the feature database, we can plug the coefficients in the coefficient table into function (6) and use Eq. (7) to infer a new null distribution for the new sample.

The permutation method takes lots of time. When the number of bins in the null distribution and the number of features are large, the logistic regression fitting may also take a great amount of time. The off-line information (i.e. the feature table and the coefficient table) is used to achieve a new null distribution without the permutation step and the logistic regression fitting step. Thus, it makes the protein-level FDR estimation efficient.

Our current implementation of the framework

When applying the proposed protein-level FDR estimation framework, two key points are: features and similarity measurement. A sample can be described by features. When a novel sample is similar to an existing sample by comparing their features, the null distribution of

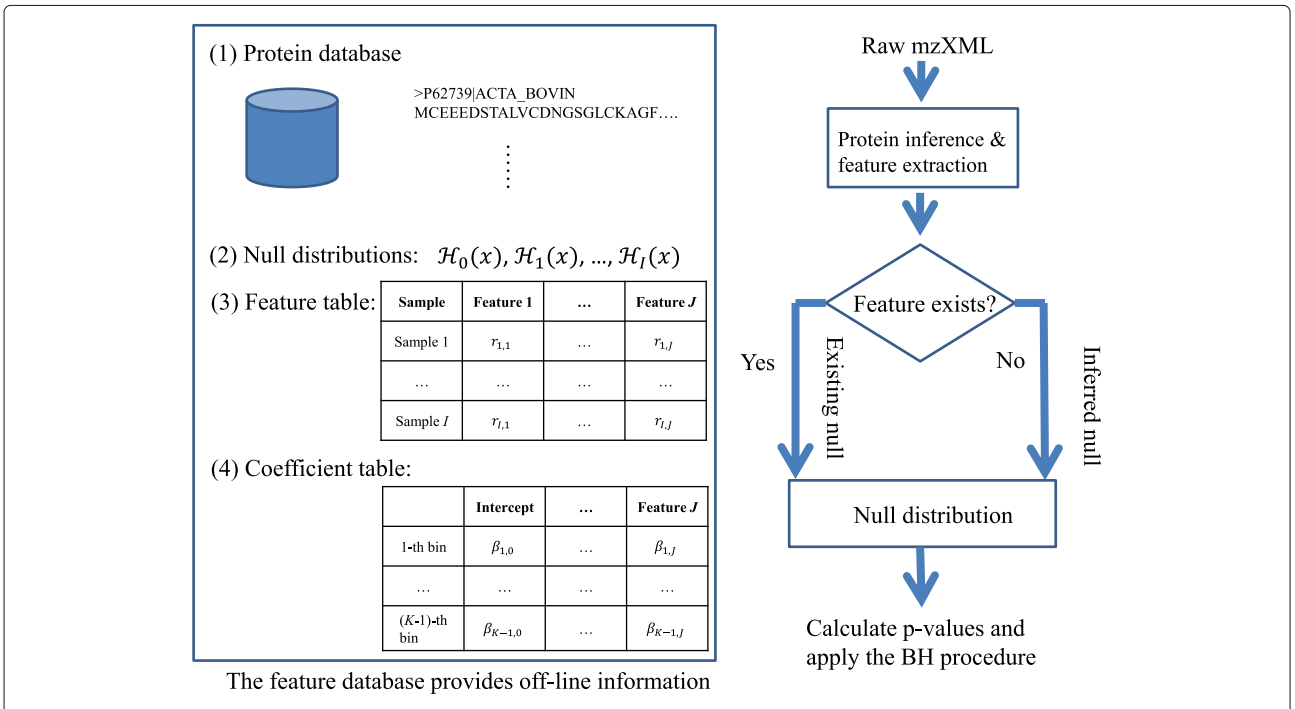

Fig. 1 A feature database contains null distributions, a feature table and a coefficient table. When features of a new sample exist in the feature table, the corresponding null distribution is chosen. Otherwise, a null distribution is inferred with information in the coefficient table. Then the null distribution is used to perform the protein-level FDR estimation

the existing sample will be used. Otherwise, a new null distribution is inferred by applying the logistic model based on sample features.

The error propagation from the peptide-level to the protein-level is non-trivial. Features of raw data such as tandem MS peak counts are faraway from the final protein inference result. Thus, these kinds of features may not have a clear connection with protein scores. In our current implementation, we determine to directly select features from protein scores.

First, we partition the range of protein scores of sample i into 10 bins of equal length. The probabilities of protein scores falling in 10 bins are denoted as $P_{i,1}$, $P_{i,2}$, ..., $P_{i,10}$. Then, we choose sample j as a reference sample. The similarity of protein identification results of sample i and sample j is measured by their Kullback-Leibler (KL) divergence:

$$D_{i,j} = \sum_{k=1}^{10} \mathrm{Pr}_{i,k} \log \frac{\mathrm{Pr}_{i,k}}{\mathrm{Pr}_{j,k}}. \quad (8)$$

The smaller the value of $D_{i,j}$, the more similar sample i and sample j. We choose the KL divergence from each sample to the reference sample as a feature, which is used to infer the null distribution and measure the sample similarity.

Overview of the experiments

The whole framework consists of two parts: The first part employs the permutation method and the BH procedure to estimate the FDR (Permutation+BH); the second part provides a logistic regression model to infer the null distribution of a new sample based on existing null distributions. We first conduct the experiment to verify Permutation+BH in FDR estimation. The performance of our method is compared to MAYU based on three datasets with groundtruth. Then, we conduct another experiment to illustrate the performance of our null distribution inference method. In the last part of our experiments, we discuss the reference dataset issue in our current implementation of our framework.

The whole framework is implemented in Ruby (v1.9.2p290). Target-decoy concatenated databases are generated from UniProtKB/Swiss-Prot (Release 2011_01) by appending shuffled protein sequences into the original protein database. Peptides are identified by X!Tandem (v2010.10.01.1) [4]. Then, ProteinProphet (Embedded in the Trans-Proteomic Pipeline v4.5 RAPTURE rev 0, Build 201109211427) is employed to perform peptide probability calculation and protein inference, respectively [5, 15].

In our experiments, we use six public available datasets: ISB, ABRF, Yeast, Yeast_Train, Human and Human_Test. The ISB dataset was achieved from a 18 standard protein mixture [16]. The sample was analyzed on a Waters/Micromass Q-TOF using an electrospray source. The ABRF sPRG2006 dataset contains 49 standard proteins. The Yeast and the Yeast_Train dataset were obtained by analyzing cell lysate on both LCQ and ORBI mass spectrometers from wild-type yeast grown in rich medium [17, 18]. The dataset contains a protein reference set which is used as the groundtruth. The Human dataset was obtained from human HEK293T cell lines and analyzed on the ORBI mass spectrometer. The Human_Test dataset was obtained by analyzing human serum samples with Thermo LTQ-FT. In our experiments, ".RAW" files are converted to ".mzXML" files by TPP. The addresses to access these datasets are shown in Table 3:

Results and discussion

FDR estimation

In this experiment, the first three datasets are used: ISB, ABRF and Yeast. For the ISB dataset, the 18 standard proteins together with 15 contaminants are marked as the groundtruth [16]. For the ABRF dataset, the 49 standard proteins and 78 contaminants are used as the groundtruth. Readers can refer to the supplementary document for more information [19]. For the Yeast dataset, all proteins in the protein reference set are treated as true proteins.

The permutation method includes two steps to obtain a null distribution: a shuffling step and a protein inference step. In the shuffling step, each protein sequence is shuffled and appended into the original protein database. In the protein inference step, proteins are identified from the target-decoy concatenated database with TPP. The shuffling step and the protein inference step repeat for 20

Table 3 Names and URLs of data files

Dataset name	File name	URL
ISB	QT20051230_S_18mix_04.mzXML	http://regis-web.systemsbiology.net/PublicDatasets/
ABRF	Lane/060121Yrasprg051025ct5.RAW	https://proteomecommons.org/dataset.jsp?i=71610
Yeast	YPD_ORBI/061220.zl.mudpit0.1.1/raw/000.RAW	http://aug.csres.utexas.edu/msnet/
Yeast_Train	YPD_ORBI/070119-zl-mudpit07-1/raw/000.RAW	http://aug.csres.utexas.edu/msnet/
Human	YPD_LCQ/060b.RAW	http://aug.csres.utexas.edu/msnet/
Human_Test	PAe000281_mzXML_200909301914/B06-7017_c.mzXML	http://www.peptideatlas.org/repository/

times. The protein mapping to a decoy sequence is considered to be a false positive. The protein probabilities of all false positives from the 20 protein identification results are collected. Then, the histogram of the protein probabilities is built and used as the null probability density function. In general, the smaller the bin size, the more detail the null distribution contains and more data points are desired to build the null distribution. In our experiments, the length of each bin is empirically chosen to be 0.003.

After the null distribution has been built, the p-value of each protein is calculated according to Eq. (4). Next, p-values of all proteins are sorted in the ascending order. Then, the BH procedure is conducted on p-values to obtain the relationship of the number of proteins and the FDR.

We apply our method and MAYU to the three protein datasets to estimate the FDR. The true FDR is calculated as the ratio of the number of proteins belonging to the groundtruth set and the total number of protein identifications. The performances of different methods are validated by comparing the absolution difference between the estimation and the groundtruth. The results based on three datasets are shown in Fig. 2.

According to our experimental results, our method and MAYU are comparable in performance on the ISB dataset. For the ABRF dataset, our method is better than MAYU on average. Our method is dominantly better than MAYU on the Yeast dataset.

Null distribution inference

In this dataset, the first five datasets listed in Table 3 are used to fit the logistic regression model (5) and the Human_Test dataset is used to validate the null distribution inference result.

In this experiment, we choose the ISB dataset as a reference dataset and calculate the KL divergence from other samples to the ISB dataset by using Eq. (8). The feature table for six datasets is shown in Table 4.

The null distributions of six samples are obtained by the permutation method. The first five null distributions are used to infer the null distribution of Human_Test. The last null distribution will be used as a reference.

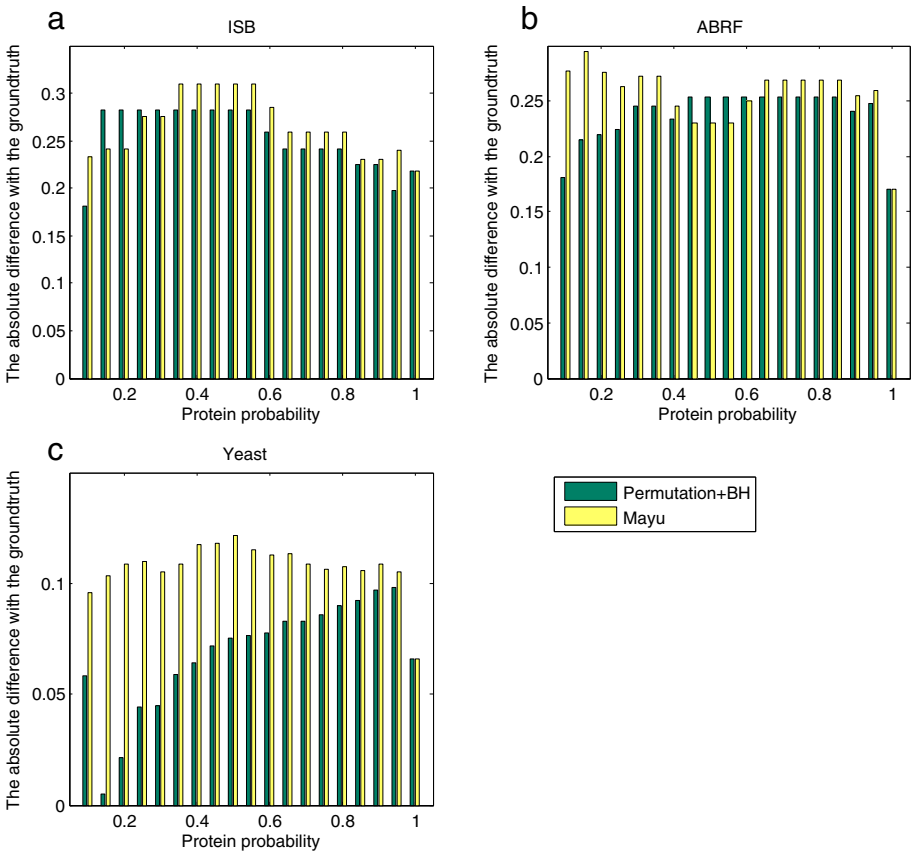

Fig. 2 a, **b** and **c** show the absolute differences between FDRs estimated and true FDRs. The smaller the difference, the better the performance. In (**a**), our method is comparable with MAYU. In (**b**), our method is better than MAYU on average. In (**c**), our method is dominantly better than MAYU

Table 4 The feature table for five datasets and one test set

Sample	KL divergence
ISB	0.000
ABRF	0.0301
Yeast	0.3697
Human	0.2206
Yeast_Train	0.2781
Human_Test	0.4159

Let the first five samples be sample 1,2,3,4 and 5, respectively. Human_Test is denoted as sample 0. The KL divergence from sample i to the reference sample is denoted as $r_{i,1}$. Since the bin length 0.003, the number of bins is $K = 334$. The probability $\Pr_i(x \leq \tilde{x} | \tilde{x} \in \mathcal{Z}_k)$ is calculated from the null distribution of sample i for bin k. The logistic regression coefficients for bin k are obtained through the following model:

$$\min_{\beta_{k,0},\beta_{k,1}} \sum_{i=1}^{5} \mathcal{L}\left(\log\left(\frac{\Pr_{i,k}}{1-\Pr_{i,k}}\right) - \beta_{k,0} - \beta_{k,1} r_{i,1}\right). \quad (9)$$

Here, $\beta_{k,0}$ is the intercept coefficient; $\beta_{k,1}$ is the sample KL divergence coefficient, respectively; $\mathcal{L}$ is a loss function measuring the error in estimation. In our experiment, we choose $\mathcal{L} = ||\cdot||_2$. In our Ruby program, we implement a R (v2.13.1) interface from which users can call any robust loss function such as the Huber loss. We partially show the coefficient table in Table 5.

The feature table and the coefficient table are stored in the feature database. When analyzing the new sample (i.e. Human_Test), we just plug coefficients in the coefficient table in Eq. (6) and use Eq. (7) to get the null probability density function of Human_Test. The result is shown in Fig. 3.

Figure 3a shows the probability density functions of the inferred null and the null obtained by the permutation method. The peak height of the inferred null is overestimated compared with that of the permutation null. In Fig. 3b, Permutation+BH and InferredNull+BH estimate the FDR by applying the BH procedure to the null distribution generated by the permutation method and the inferred null distribution, respectively. According to the result shown in Fig. 3b, the performance of InferredNull+BH is closer to Permutation+BH than that of MAYU. The correlation of the inferred null distribution and the permutation generated null distribution is 0.9052.

Table 5 Coefficient table. Note that we only need to conduct logistic regression on the first $K - 1$ bins

Bins	Intercept coefficients	KL divergence coefficients
1-th bin	-6.0539	-18.6072
...	...	...
333-th bin	5.1731	3.3703

The reference dataset

In our current implementation of our framework, we need to reference dataset. The reference dataset is chosen to calculate the KL divergence of each sample to the reference dataset. Then, the KL divergence is used as a feature in both similarity measurement and null distribution inference. In the previous experiment, we take the ISB dataset as the reference dataset. The KL divergence is a non-symmetric measure of the difference between two probabilities. Thus, a different reference dataset may lead to a different null distribution inference result. In the following experiment, we take the ISB dataset, the ABRF dataset and the Yeast dataset as the reference dataset, respectively. The result is shown in Fig. 4. The correlations of the inferred null distribution and the permutation generated null distribution when using the ISB dataset, the ABRF dataset and the Yeast dataset as a reference dataset are 0.9052, 0.6779 and 0.3317, respectively. According to the experimental result, the best performance in null distribution estimation is achieved when the ISB dataset, which contains 18 proteins, is taken as the reference dataset. The performance is worst when the Yeast dataset containing hundreds of proteins is used as the reference dataset.

Readers may be interested in the results of inferring the null distributions of other samples (other than Human_Test) by using the ISB dataset as the reference dataset. In the following experiment, we conduct two extra experiments to inferring the null distributions of the ABRF dataset and the Yeast dataset. In each of the two experiments, we treat either the ABRF dataset or the Yeast dataset as test data and using remaining datasets as training data. The result is shown in Fig. 5.

The difference between the inferred null distribution and the empirical null distribution obtained by the permutation method may be caused by the following reasons:

- The data used in fitting the logistic regression model may be neither typical nor enough. Data representative to different conditions are desired to obtain a robust regression model. When information of some typical conditions are missing, it is may be hard to make it up by using mathematical models.
- In our experiment, we only consider one feature in our logistic model. The single feature may not explain all kinds of variation in the protein inference process. A more accurate model can be achieved by using more features and plenty of data in building the feature database.

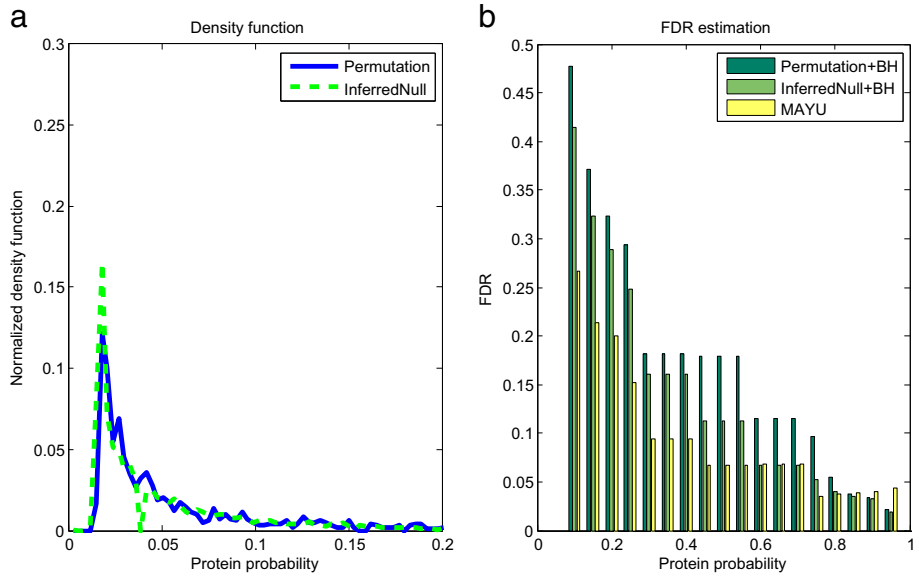

Fig. 3 a shows the probability density functions of the inferred null distribution and the null distribution obtained by the permutation method. The correlation between the inferred null distribution and the generated null distribution using permutation is 0.9052. **b** shows the FDR estimation results of Permutation+BH, InferredNull+BH and MAYU

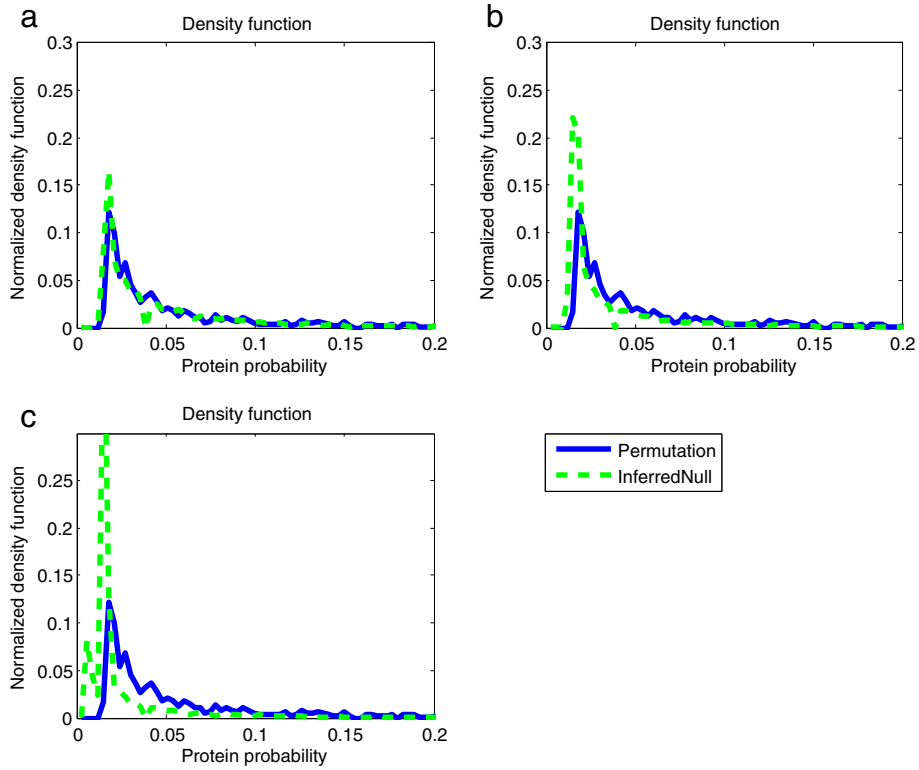

Fig. 4 a, **b** and **c** show null distribution inference results by using the ISB dataset, the ABRF dataset and the Yeast dataset as a reference dataset, respectively. The correlations between the inferred null distribution and the permutation generated null distribution in (**a**), (**b**) and (**c**) are 0.9052, 0.6779 and 0.3317, respectively

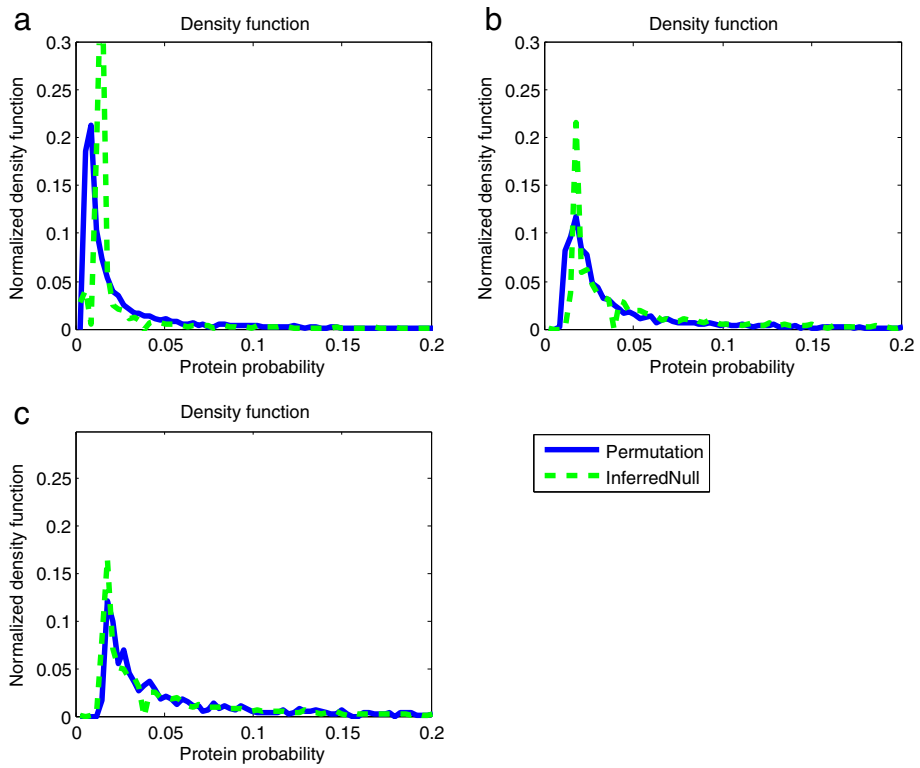

Fig. 5 In (**a**), the ABRF dataset is used as the test data and other datasets are the training data; in (**b**), the Yeast dataset is used as the test data and the other datasets are training data; in (**c**), the Human_Test dataset is used as test data and other datasets are the training data. The correlations between the inferred null distribution and the permutation generated null distribution in (**a**), (**b**) and (**c**) are 0.4259, 0.8116 and 0.9052, respectively

- Even if we obtain an ideal logistical model with perfect coefficients, it is not guaranteed that the new data is not an outlier. It is often the case that the new data locates at a point that has a certain distance to the ideal model.

The biggest advantage of the null inference method is its efficiency. Once the coefficient table is obtained beforehand, the inference of the null distribution is just the calculation of deterministic functions (6) and (7). The whole process of the FDR estimation just takes a few seconds. This should benefit the large-scale data analysis.

Discussion

In the Permutation+BH method, we use the permutation method to generate the null distribution and apply the BH procedure to estimate the FDR. The method does not relies on specific assumptions and works directly at the protein-level. Thus, the problems related to improper assumption and error propagation are avoided. The flexibility of our method also implies that it can be used with any search method or protein inference method of the user's choice. According to our experimental results based on three datasets, our method performs better than MAYU. We believe that this is partly due to a more accurate estimation of the null distribution through the increased sampling by our permutation method, than in a typical 1:1 target-decoy approach as used in MAYU.

In the Permutation+BH method, the efficiency is low because we need to shuffle protein sequences and conduct protein inference multiple times. We propose an off-line strategy to handle this issue. In the off-line strategy, a feature protein database is built beforehand with null distributions obtained from existing samples, a feature table and a coefficient table. When a new sample cannot find a match in the feature table, a new null distribution is inferred by directly plugging the coefficients in the coefficient table into the logistic model. The logistic regression model provides an efficient way to infer the null distribution. Our model is currently trained only on a few samples and including only 1 feature, limiting its accuracy. We will seek to improve this model by adding many more features and training the model on more datasets, as part of our future work.

Conclusions

In this paper, we propose a protein-level FDR estimation framework. The framework includes two major components: the Permutation+BH FDR estimation method and

the logistic regression-based null distribution inference method. The Permutation+BH method first applies the permutation to generate the null distribution and then uses the BH procedure to estimate the FDR. However, this method is inefficient for online identification. Therefore, we propose the logistic regression-based null distribution inference method to handle this issue. In our experiment based on three public available datasets, our Permutation+BH method achieves consistently better performance than MAYU, which is chosen as the benchmark FDR calculation method for this study. The null distribution inference result shows that the logistic regression model achieves a reasonable result both in the shape of the null distribution and the corresponding FDR estimation result.

Acknowledgements
The abridged abstract of this work was previously published in the Proceedings of the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017), Lecture Notes in Computer Science: Bioinformatics Research and Applications [20].

Funding
This work was supported in part by International S&T Cooperation Program of China (ISTCP) under Grant No. 2014DFA31520 and by Shenzhen Fundamental Research Fund under Grant No. QTD2015033114415450.

Availability of data and materials
All the data discussed in this study are publicly available and the sources for downloading these data are listed in Table 3.

About this supplement
This article has been published as part of *BMC Genomics* Volume 19 Supplement 6, 2018: Selected articles from the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-6.

Authors' contributions
GYW, XW and BHX developed the model, conducted the simulation studies and real data analysis, and wrote the draft of the manuscript. BHX revised the manuscript. GYW and XW conducted the real data analysis and finalized the manuscript. BHX provided the guidance on methodology and finalized the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not-applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]The Dental Center of China-Japan Friendship Hospital, Beijing, China.
[2]ShenZhen Research Institute of Big Data, ShenZhen, China.

Published: 13 August 2018

References

1. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods. 2007;4:787–97.
2. Eng J, McCormack AL, Yates III JR. An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976–89.
3. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.
4. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7.
5. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75:4646–58.
6. Bern M, Goldberg D. Improved ranking functions for protein and modification-site identifications. J Comput Biol. 2008;15(7):705–19.
7. Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H. A Bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol. 2009;16(8):1183–93.
8. Spirin V, Shpunt A, Seebacher J, Gentzel M, Shevchenko A, Gygi S, Sunyaev S. Assigning spectrum-specific p-values to protein identifications by mass spectrometry. Bioinformatics. 2011;27:1128–34.
9. Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol. 2006;24:333–8.
10. Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem. 2004;76:1664–71.
11. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics. 2009;8:2405–17.
12. Gupta N, Bandeira N, Keich U, Pevzner PA. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom. 2011;22(7):1111–20.
13. Friedman J, Tibshirani R, Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009.
14. Reidegeld KA, Eisenacher M, Kohl M, Chamrad D, Körting G, Blüggel M, Meyer HE, Stephan C. An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. Proteomics. 2008;8:1129–37.
15. 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. Anal Chem. 2002;74:5383–92.
16. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng J, Aebersold R, Martin DB. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. J Proteome Res. 2008;7:96–103.
17. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2006;25:117–24.
18. Ramakrishnan SR, Vogel C, Kwon T, Penalva LO, Marcotte EM, Miranker DP. Mining gene functional networks to improve mass-spectrometry-based protein identification. Bioinformatics. 2009;25:2955–61.
19. Gerster S, Qeli E, Ahrens CH, Bühlmann P. Protein and gene model inference based on statistical modeling in k-partite graphs. Proc Natl Acad Sci. 2010;107:12101–6.
20. In: Cai Z, Daescu O, Li M, editors. Proceedings of the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017), Honolulu, Hawaii, May 30 - June 2, 2017. New York City: Springer; 2017.