Chapter 7

False Discovery Rate Estimation in Proteomics

Suruchi Aggarwal and Amit Kumar Yadav

Abstract

With the advancement in proteomics separation techniques and improvements in mass analyzers, the data generated in a mass-spectrometry based proteomics experiment is rising exponentially. Such voluminous datasets necessitate automated computational tools for high-throughput data analysis and appropriate statistical control. The data is searched using one or more of the several popular database search algorithms. The matches assigned by these tools can have false positives and statistical validation of these false matches is necessary before making any biological interpretations. Without such procedures, the biological inferences do not hold true and may be outright misleading. There is a considerable overlap between true and false positives. To control the false positives amongst a set of accepted matches, there is a need for some statistical estimate that can reflect the amount of false positives present in the data processed. False discovery rate (FDR) is the metric for global confidence assessment of a large-scale proteomics dataset. This chapter covers the basics of FDR, its application in proteomics, and methods to estimate FDR.

Key words False discovery rate, Posterior error probability, Target-decoy, Peptide spectrum matches, Statistical validation, Shotgun proteomics

1 Introduction

In any large-scale high-throughput study, including genomics and proteomics, a large number of statistical hypothesis are tested, usually independently, for significance [1]. Each hypothesis tested (a gene, a transcript, a peptide, etc.) yields a *p*-value, *e*-value, or a score that depicts the quantitative measure of that hypothesis being correct. In proteomics, shotgun proteomics data is searched using a database search algorithm that provides such confidence metrics after searching spectra against peptides in a given FASTA database.

While such metrics can reflect the "*goodness of fit*" of an experimental spectrum to the assigned peptide and related chances of error in its identification, it cannot reflect the associated error in the whole dataset. For example, selecting peptide spectral matches (PSMs) or hits with $p\text{-value} \leq 0.05$ means that each PSM has a 5 % or less chance of incorrectly being assigned as a significant match. This, however, does not mean that all PSMs passing this threshold

Klaus Jung (ed.), *Statistical Analysis in Proteomics*, Methods in Molecular Biology, vol. 1362,
DOI 10.1007/978-1-4939-3106-4_7,

will have a collective error of 5 %. Any PSM, even with a low estimated error of 5 %, could turn out to be wrong. It could occur by (highly unfortunate and rare) chance that all PSMs (each with 5 % estimated error) turn out to be wrong.

One cannot estimate the percentage of false hits in the accepted PSMs by using *p*-value as a metric because this is a *single spectrum-specific* significance measure. To assess global false hits or error rate, one needs to understand population-level false estimation metrics. This is a classic case of what is called as the *multiple testing problem*; that is, when multiple independent statistical hypothesis tests are conducted, single hypothesis significance measures (like *p*-value) are neither sufficient nor amenable to extrapolation to calculate population error rate. By random chance alone, there will be many hits which may turn out to be false [2] in a collection of hits, each with $p\text{-value} \leq 0.05$ with a final error rate of more than 5 % globally.

1.1 False Discovery Rate

Adjusting for multiple comparisons can be achieved by applying Bonferroni correction which readjusts significance threshold ($\alpha = 0.05$) to control the false positives. For n spectra, the population-level significance threshold becomes $0.05/n$, to adjust for the error rate of 5 % globally. This method is very stringent and false positives are extremely low when n is large. But this occurs at the cost of false negatives. To avoid false positives, this method excludes many true positives with good scores and *p*-values. False discovery rate (FDR) is a measure of the incorrect PSMs among all accepted PSMs [2–4]. Proposed by Benjamini and Hochberg [5] as an alternate to the Bonferroni correction, it is defined as the rate of false positives in accepted hits. FDR is a less stringent metric for global confidence assessment. In the context of proteomics, it is a global estimate of the false positives present in the results obtained by a database search algorithm. There are many different strategies to estimate FDR like the *nonparametric* simple target-decoy (TD) database searches [4, 6] and *parametric* or *semi-parametric* mixture modeling approaches used in the Trans-proteomics pipeline (TPP) [7–10].

For estimating the FDR, a null model is required for which a decoy database search is carried out in proteomics. In the TD search strategy, the database search is carried out on the true (target) as well as null (decoy) database. A decoy database is constructed by shuffling, randomizing or by simply reversing the target database. It is the simplest approach to calculate FDR and requires no distributional assumptions, i.e., *nonparametric* in nature. The basic assumption made for TD approach is that the number of false PSMs in decoy search will be equal to the number of false PSMs in target search above a given threshold score. The database search for this approach can be performed together (concatenated) or separately with the decoy database. A *concatenated* search can be

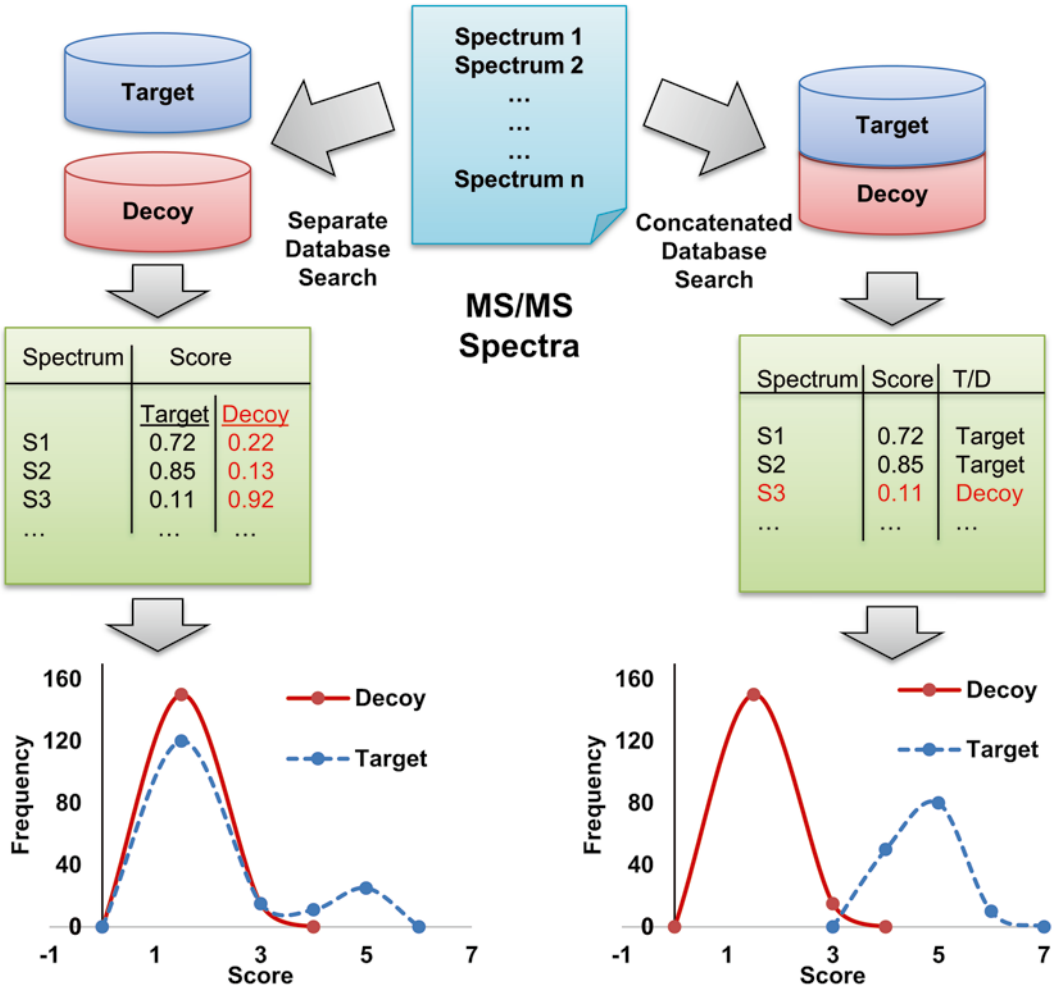

Fig. 1 There are two database search strategies—separate or combined database search. In separate search target and decoy databases are searched separately and FDR is estimated using Kall's method (*see* text). Each spectrum has one target and one decoy best score. In combined/concatenated approach, one unified target-decoy database is searched in which both TD peptides compete with each other. Each spectrum has one best score, either from target or decoy but not both. This also changes the score distributions

conducted by combining the target and decoy databases together (both target and decoy) as proposed by Elias and Gygi [6]. In a *separate* database search [4], as the name implies, both target and decoy databases are searched separately and scored using a search algorithm. The number of false positives divided by the total hits allows for easy calculation of FDR. Overview of this process is shown in Fig. 1.

1.2 q-*Value*

While FDR is a global measure of population error rate, this communicates nothing about confidence of individual PSMs. A PSM-specific metric is needed which conveys the confidence measure of a

particular PSM after FDR correction has been applied. Thus, *q-value* was introduced by Storey and Tibshirani [1], which is defined as the minimum FDR cutoff at which a particular PSM can be accepted. It is the property of a single PSM rather than a set of PSMs. The *q*-value of a PSM provides a direct measure of significance for a particular PSM with respect to the complete dataset and the risk accrued to the total accepted matches if that hit is deemed significant. For example, if a PSM with *q*-value 0.07 seems biologically important, we will need to lower the FDR threshold to 7 % in the dataset to accept this PSM as significant. A sorted list of hits by *q*-value becomes a monotonous function of search score/*p-value* and is thus easily interpretable [2]. A dataset can be revisited to select a biologically important hit without the need to recalculate the FDR.

1.3 Posterior Error Probability

Posterior error probability (PEP) is the probability of a PSM to be incorrect. Borrowing from the example given by Kall et al. [11], a PEP of 1 % would signify that there is 99 % chance for the PSM to be correct. It is also referred to as local FDR, as unlike FDR it measures the error rate associated with a single PSM. From Fig. 2, it can be seen that FDR represents the ratio of area under incorrect (decoy) region for any given score threshold *x* to the area under the total region for the same threshold. PEP is ratio of the infinitesimally small areas (virtually the height) of incorrect to total hits

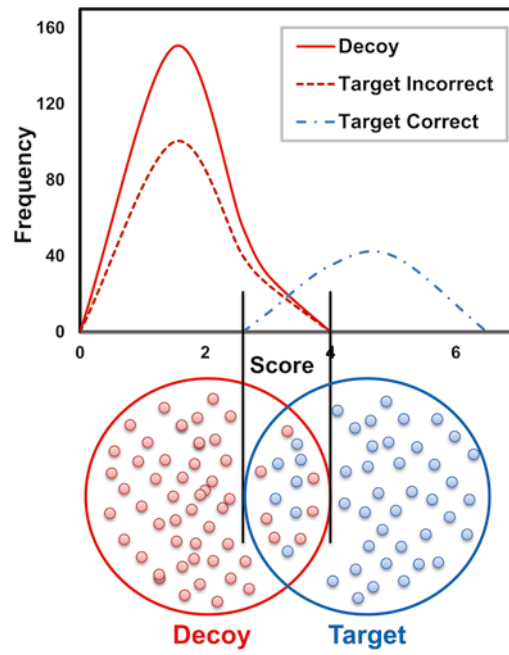

Fig. 2 There is considerable overlap between the correct and incorrect hits shown here as score frequency distribution and the corresponding set-based visualization. FDR is applied to control the proportion of false hits from getting accepted

in a local score region of x. Both are complementary to each other but have different meanings. While the q-value conveys the risk (error introduced) in the whole dataset if we accept the PSM at hand, the PEP on the other hand informs us whether the PSM is likely to be correct or not.

FDR can be calculated from PEP by integrating (summing up) all the PEPs. PEPs can be accurately calculated by using machine learning to learn the model parameters from labeled (correct and incorrect) training data. For any given score x, the PEP can be predicted from the model parameters. This strategy is used in PeptideProphet [8] and ProteinProphet [9].

2 Materials

Any FDR strategy would require a large-scale shotgun proteomics experiment data. This data should be searched using a database search algorithm. A normal desktop computer with enough memory according to dataset size should be useful. Perl programming language and ProteoStats [12] should be configured on the computer.

3 Methods

To calculate FDR, the spectra are searched against a target and a decoy database to obtain the top-ranked PSMs. This can be searched either separately or in a combined database search, each with different assumptions on target decoy competition and false-positive estimation. This section briefly explains how to use ProteoStats library for FDR estimation. ProteoStats requires the data to be searched using separate TD approach as it can perform the TD competition after the search as suggested by Fitzgibbon et al. [13]. More details can be found in ProteoStats documentation, supplementary material of [12], and blog post on ProteoStats [14]. Based on how TD matches are defined (in terms of TD competition) and how false positives are defined, there are five methods for FDR calculation. Mainly separate and concatenated FDR are two main methods based on the mode of search. Several variants of these formulations have been proposed which may provide better results [15, 16]. Though the community majorly uses Kall's or Elias and Gygi's formulae due to ease of calculation, there is no consensus on which formula is better or appropriate. Currently, the proteomics community agrees on any method as long as it is clearly defined. We also propose that users can try all methods in ProteoStats for a specific workflow for standardizing the protocol according to search engine, decoy strategy, data quality, etc. In our hands, we have found FDR with percentage of incorrect target (PIT) correction and refined methods to work better than others.

3.1 FDR Calculation Using ProteoStats

TD searches are completed separately and results in the form of target and decoy top hits provided as input to ProteoStats. When the searches are conducted separately, all different FDR methods can be applied *a posteriori*, but if a concatenated search is used, only concatenated FDR method can be applied as the correspondence between TD top hits is lost. ProteoStats removes the peptides identical in decoy and target considering isoleucine and leucine as identical. The resulting TD sets are sorted separately on the basis of scores/*e*-values/*p*-values from best to worst and depending on the search strategy chosen the FDR, *q*-value, and receiver operating curve (ROC) are calculated [12]. There are other supplementary modules for chart generation and comparing results (*see* **Note 1**).

For the calculation of FDR, there are different methods/formulae available:

1. Separate/simple FDR (FDR_S)
   This method by Kall et al. [4] assumes that the number of decoys passing the threshold (D) represents the number of false positives in the target PSMs (T) above the same threshold (*see* **Note 2**):

$$FDR_S = \frac{D}{T} \tag{1}$$

2. Concatenated FDR (FDR_C)
   This method by Elias and Gygi [17] assumes that for any number of decoys (D) passing a given threshold, there are equal number of false hits in target PSMs (T) above that threshold. Adding up the false hits in decoy and target, the number of false positives is therefore double of the decoy count above threshold. In this search, TD competition results in either target or a decoy best hit for any given spectrum, the reference population in which FDR is calculated is not the same as other methods:

$$FDR_C = \frac{2 \times D}{(T + D)} \tag{2}$$

3. FDR with PIT correction (FDR_{PIT})
   Kall's formula for simple FDR did not consider the incorrect target PSMs during simple FDR calculation which tilts the balance of random matches in decoy's favor due to higher decoy population. To correct for this effect, it was suggested to calculate the PIT, which is used as a factor to accurately determine the FDR. Note that the name PIT is a misnomer as it is a fraction and not a percentage. It is similar to the notion of fraction of true negatives, Π_0, as defined in genomics [1]:

$$FDR_{PIT} = PIT \times \frac{D}{T} \tag{3}$$

4. Refined separate FDR (FDR_{RS})
   In this formulation, Navarro et al. [15] propose that FDR should be calculated in the correct reference population (only targets). They argued that the estimated false positives should not be doubled blindly by observing decoy hits above threshold directly. The decoy PSMs above threshold should not be considered false positives if they do not score more than the corresponding target PSM. This causes inflated false-positive estimation and leads to overestimation of FDR. The hits could be above threshold only in target (target only, TO) or only in decoy (decoy only, DO). When both are above threshold, either target could be better (target better, TB) or the decoy (decoy better, DB). The FDR in the correct reference population is calculated by estimating the correct false positives and dividing by corrected total population [15]:

$$\mathrm{FDR}_{\mathrm{RS}} = \frac{(2 \times \mathrm{DB} + \mathrm{DO})}{(\mathrm{TB} + \mathrm{TO} + \mathrm{DB})} \tag{4}$$

5. Refined concatenated method (FDR_{RC})
   In this formulation, Cerqueira et al. [16] argued that since decoy hits are by definition false, they can be disregarded in FDR estimation and thus the FDR_C formula is changed to yield the following formula:

$$\mathrm{FDR}_{\mathrm{RC}} = \frac{D}{(T - D)} \tag{5}$$

3.2 Peptide to Proteins

FDR calculated at PSM level is not the same as protein-level FDR. Although the goal of a shotgun proteomics experiment is to assess protein-level significance, the hypotheses tested from shotgun proteomics data are spectra. Due to variable abundances of different proteins and the non-random distribution of peptides across these proteins, one-to-one correspondence between peptide and protein FDRs does not exist. Due to this incongruity, calculating protein FDRs is complicated.

Since the decoy database is made out of the target database with equal number of proteins and same protein length distribution, we can simply use the same algorithm (*see* **Note 3**) to estimate the number of false positives owing to the identification of the decoy proteins. This step requires a robust protein score which can help in estimation of protein FDR. The FDR for protein estimation is calculated as the ratio of the expected number of false-positive protein identifications (those that have a hit to the decoy database proteins) to that of the total number of protein identifications mapping to the target database at any threshold protein score. For protein FDR, MAYU software can be used which performs protein identification-level FDR on the basis of peptide identifications [18].

4 Notes

1. Calculating FDR using ProteoStats [12] software is accurate and reliable. It is written in Perl and it can be integrated in any kind of pipeline easily. In terms of file format inputs, it provides great flexibility as it can read proteomics output file format from different database searches like Mascot [19], MassWiz [20], OMSSA [21], X!Tandem [22], MyriMatch [23], and Comet [24]. These results may be processed using a rescorer like FlexiFDR [25] for MassWiz, Percolator [26–28] for Mascot, and OScore for Sequest [29] to improve identification results. Tab-delimited files can also be used for FDR to support other algorithms. It provides with CSV/Excel-based output files which can easily be interpreted and used for further analysis in R. It also contains plotting functionalities for visual analysis of the results obtained. Comparison of results and Venn charts can also be generated.
2. A general algorithm is described for FDR estimation using Kall's method. The pictorial representation is shown in Fig. 3. Please note that this is a generic algorithm for simple FDR calculation. This code is provided within ProteoStats for all FDR formulae, so it need not be manually calculated.
   (a) Sort target results on score/*p*-value/*e*-value from best to worst hit.
   (b) Sort decoy results on score/*p*-value/*e*-value from best to worst hit.

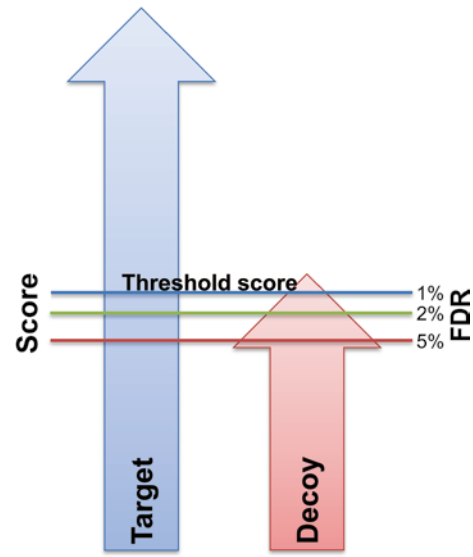

Fig. 3 Target and decoy scores are sorted and iteratively different thresholds can then be applied until a desirable level of FDR is achieved. FDR is the ratio of decoy/target hits at any particular threshold

(c) For every target score as threshold, count D and T, and the number of decoys and targets above threshold.

(d) Calculate FDR using Eq. 1. This algorithm explains simple FDR using Kall's method. Other methods differ in definition of false-positive counts, so should be accordingly calculated.

(e) This FDR is also the q-value for the PSM serving as score threshold, and other PSMs with same score.

(f) A tabulated list of number of targets and corresponding q-value can be used to create an ROC plot.

3. After FDR is calculated, the PSMs and peptides are used for protein inference. Similar TD approaches can be used for estimating protein-level FDR estimates although the correspondence between peptide and protein FDR is not same due to non-random distribution of peptides across proteins due to varying abundances. MAYU [18] is a tool for protein FDR calculation though it does not infer proteins. IDPicker [30] is another tool which integrates the process but can only perform FDR for concatenated search. ProteoStats will be updated in near future to support protein-level FDR calculation.

Acknowledgement

S.A. is supported by SRF grant and A.K.Y. is supported by Innovative Young Biotechnologist Award (IYBA) grant and DDRC-SFC grant from Department of Biotechnology (DBT), India. Authors acknowledge Manu Kandpal for proofreading the manuscript.

References

1. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440–9445
2. Choi H, Nesvizhskii AI (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res 7:47–50
3. Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Proteomics 73:2092–2123
4. Kall L, Storey JD, MacCoss MJ et al (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7:29–34
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300
6. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214
7. Choi H, Ghosh D, Nesvizhskii AI (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res 7:286–292
8. Keller A, Nesvizhskii AI, Kolker E et al (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392
9. Nesvizhskii AI, Keller A, Kolker E et al (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658

10. Tabb DL (2008) What's driving false discovery rates? J Proteome Res 7:45–46
11. Kall L, Storey JD, MacCoss MJ et al (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 7:40–44
12. Yadav AK, Kadimi PK, Kumar D et al (2013) ProteoStats—a library for estimating false discovery rates in proteomics pipelines. Bioinformatics 29:2799–2800
13. Fitzgibbon M, Li Q, McIntosh M (2008) Modes of inference for evaluating the confidence of peptide identifications. J Proteome Res 7:35–39
14. Yadav AK, Perez-Riverol Y (2014) ProteoStats: computing false discovery rates in proteomics. BioCode's notes, computational proteomics & bioinformatics. http://computationalproteomic.blogspot.com/2014/08/proteostats-computing-false-discovery.html
15. Navarro P, Vazquez J (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. J Proteome Res 8:1792–1796
16. Cerqueira FR, Graber A, Schwikowski B et al (2010) MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification. J Proteome Res 9:2265–2277
17. Elias JE, Gygi SP (2010) Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol 604:55–71
18. Reiter L, Claassen M, Schrimpf SP et al (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 8:2405–2417
19. Perkins DN, Pappin DJ, Creasy DM et al (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3567
20. Yadav AK, Kumar D, Dash D (2011) MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J Proteome Res 10:2154–2160
21. Geer LY, Markey SP, Kowalak JA et al (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964
22. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467
23. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6:654–661
24. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13:22–24
25. Yadav AK, Kumar D, Dash D (2012) Learning from decoys to improve the sensitivity and specificity of proteomics database search results. PLoS One 7, e50651
26. Brosch M, Yu L, Hubbard T et al (2009) Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res 8:3176–3181
27. Spivak M, Weston J, Bottou L et al (2009) Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. J Proteome Res 8:3737–3745
28. Wright JC, Collins MO, Yu L et al (2012) Enhanced peptide identification by electron transfer dissociation using an improved mascot percolator. Mol Cell Proteomics 11:478–491
29. Shao C, Sun W, Li F et al (2009) Oscore: a combined score to reduce false negative rates for peptide identification in tandem mass spectrometry analysis. J Mass Spectrom 44:25–31
30. Ma ZQ, Dasari S, Chambers MC et al (2009) IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 8:3872–3881