Published in final edited form as:

Proteomics. 2009 January ; 9(1): 194–204. doi:10.1002/pmic.200800330.

On the estimation of false positives in peptide identifications using decoy search strategy

Changyu Shen^{1,§,*}, Quanhu Sheng^{2,3,§}, Jie Dai³, Yixue Li³, Rong Zeng³, and Haixu Tang²

¹Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202

²School of Informatics, Indiana University, Bloomington, IN47408

³Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Abstract

False positive control/estimate in peptide identifications by mass spectrometry is of critical importance for reliable inference at the protein level and downstream bioinformatics analysis. Approaches based on search against decoy databases have become popular for its conceptual simplicity and easy implementation. Although various decoy search strategies have been proposed, few studies have investigated their difference in performance. With datasets collected on a mixture of model proteins, we demonstrate that a single search against the target database coupled with its reversed version offers a good balance between performance and simplicity. In particular, both the accuracy of the estimate of the number of false positives and sensitivity is at least comparable to other procedures examined in this study. It is also shown that scrambling while preserving frequency of amino acid words can potentially improve the accuracy of false positive estimate, though more studies are needed to investigate the optimal scrambling procedure for specific condition and the variation of the estimate across repeated scrambling.

Keywords

Decoy databases; False positive; Mass Spectrometry; Peptides; Sensitivity

1 Introduction

Currently, a common approach for high-throughput protein identification is the “bottom-up” or “shotgun” strategy that integrates liquid chromatography (LC), tandem mass spectrometry (MS/MS) and subsequent computer-aided data analysis, i.e. database search. In a typical LC-MS/MS experiment, proteins in a sample are first digested into peptides using some protease, separated in LC system and then subject to tandem mass spectrometry analysis. During the analysis, each peptide is first ionized and its mass-to-charge ratio (m/z) is measured by a mass spectrometer (MS scan). The most abundant peptide ions are then selected for a second mass analysis, in which the peptide ions are fragmented and the resulting fragment ions are measured again by a mass spectrometer (MS/MS scan) to generate their m/z signatures. Finally, the signature is compared with the theoretical

*To whom correspondence should be addressed: Changyu Shen, Ph.D., Division of Biostatistics, Indiana University School of Medicine, 410 West 10th Street, Suite 3000, Indianapolis, IN 46202, Phone: 317-274-1641, Fax: 317-274-2678, chashen@iupui.edu.

§These authors contributed equally to this work.

The authors declare no conflict of interest.

Supporting information is available.

signature of each candidate peptide in a *Target Database* using a scoring scheme that measures their similarity and the one with the best score is assigned as the identified peptide. The identified peptides are then assembled to identify proteins.

As powerful as the shotgun approach is, peptide identifications are subject to uncertainty. Particularly, variation and noises can be introduced to the experimental spectra through various experimental stages and there is uncertainty in prediction model used to generate the theoretical spectra. Both factors contribute to false identifications--experimental spectra are assigned to incorrect peptides. This is a critical analytical issue since the quality and validity of the identification of proteins and downstream analysis of these proteins heavily depend on the accuracy of the peptide identifications. A simple solution to the problem is to use the score as the selection criterion to determine the peptides that are mostly likely to be true by applying some thresholds. At the fundamental level, there are two issues in peptide identifications: a) to reduce the uncertainty in peptide identifications; and b) to measure the uncertainty.

A number of approaches have been proposed in the literature to tackle a) [1-10]. As for b), many adopted the p -value or E -value [4,6-8,11], or false discovery rate (FDR) as measures of statistical significance [12-18]. These concepts originated from the traditional framework of statistical-hypothesis testing, where an alternative hypothesis (H_1) is tested against a null hypothesis (H_0). In the context of peptide identification using MS/MS, the null hypothesis H_0 is that the score of a peptide-spectrum-match (PSM) is either random or driven by homology of the identified peptide to the true peptide [19]; and the alternative hypothesis H_1 is that the score reflects a draw from the pool of the scores of correct matches. The goal is to control/estimate false positives based on the score distributions under H_0 or/and H_1 .

A common feature of the p -value, E -value or FDR based approaches is the requirement of the estimate of the score distribution under H_0 (also called the *null distribution*). The process that yields the false assignment is likely to be complicated [19], which poses a great challenge for such an estimation task. Recently, the decoy search strategy has become popular [20], in which the target database is altered to represent a pool of false sequences, called the *decoy database*, and the spectra are searched against the decoy database to generate a score distribution that is used to simulate the null distribution. The method is simple in concept and easy to implement with reasonable accuracy [20]. Nevertheless, two issues are still not adequately addressed.

1. Should the target and decoy databases be concatenated for a single search (the *Target-Decoy approach*), or should the search be conducted separately for the two databases (the *Target|Decoy approach*)? Although the two strategies appear to be different in the implementation of the search, they actually are the same operation described in different ways [21]. The real distinction lies in how the results are interpreted. For the Target-Decoy approach, the identification of peptides and estimate of the null distribution are integrated: if the best score achieved in the target database for a spectrum beats that of the decoy database, then the corresponding peptide in the target database is considered as a candidate of true positive; otherwise, the best score of the decoy database is included for the simulation of the null distribution. The rationale of this approach is that sequences in the decoy database should receive the same competition from the true sequences as do the false sequences in the target database, whereas the downside is that the true sequences receive competition from more false sequences, which could compromise sensitivity. On the other hand, for the Target|Decoy approach, the identification of peptides and estimation of the null distribution is carried out separately. The search against the target database will identify candidates of true positives and the search against the decoy allows the simulation of the null

distribution by including the best score for each spectrum. The rationale of this approach is that it will not compromise sensitivity, whereas the downside is that the false sequences in the decoy database do not receive competition from the true sequences, which could lead to over-estimate of false positives [20].

2. What is the best way to alter the database to balance simplicity in implementation and performance? The most popular method for constructing a decoy database is to simply reverse every single sequence in the target database. Nonetheless, there are many other ways to alter the target database and their difference is not fully studied.

In this article, we describe a comparative study based on a mixture of model proteins to seek some answers/insights regarding the two issues above. Specifically, we examined various decoy-based approaches for their estimate of the number of false positives and sensitivity, which are the most practically relevant measures. The estimate of the number of false positives at various thresholds is also a reflection of the quality of the estimation of the null distribution. We show that target-decoy strategy with the decoy being the reversed version of the target database offers a good balance between performance and simplicity. Scrambling sequences while preserving frequency of amino acid words could potentially improve the accuracy of the estimate of the number of false positives (as measured by the distance index) under certain occasions. Further studies are needed to investigate the optimal scrambling procedure for specific condition and the variation of the estimate across repeated scrambling.

2 Materials and Methods

2.1 Tryptic digestion of standard proteins

Nineteen proteins (description of the 19 proteins can be found in the Supporting Information) were mixed with equal amount. About 1 mg proteins were dissolved in 500 μ L solution containing 50 mM ammonium bicarbonate and 6 M Guanidine HCl. 10 μ L of 1 M DTT was then added in the mixture. The mixture was incubated at 37 °C for 150 min, and 80 μ L of 1 M IAA was added and incubated for an additional 40 min at room temperature in darkness. The protein mixture was ultra filtrated using a 3 kDa ultra filtration membrane into 50 mM ammonium bicarbonate buffer with pH 8.3, and incubated with trypsin (30:1) at 37°C overnight. Undigested proteins and trypsin were removed by ultra filtration using a 10 kDa ultra filtration membrane. The tryptic peptide mixtures were collected and lyophilized for further analysis.

2.2 1D-LC-MS/MS analysis

A Surveyor liquid chromatography system (Thermo Finnigan, San Jose, CA, USA) equipped with a C18 trap column (RP, 320 μ m $\times$ 20 mm, Column Technology Inc., CA, USA) and an analytical C18 column (RP, 150 μ m $\times$ 100 mm, Column Technology Inc., CA, USA) was used. The HPLC solvents used were 0.1% formic acid (v/v) aqueous (A) and 0.1% formic acid (v/v) acetonitrile (B). About 30 μ g peptide mixture was first loaded into the trap column with a 3 μ L/min flow rate after the split, and then the reversed-phase gradient was from 2 to 40% mobile phase B in 90 min at 150 μ L/min flow rate before the split and 2 μ L/min after the split. A linear ion trap/Orbitrap (LTQ-Orbitrap) hybrid mass spectrometer (ThermoFinnigan, San Jose, CA, USA) equipped with an ESI microspray source was used to for MS/MS experiment with ion transfer capillary of 275 °C and ESI voltage of 3.2 kv. Normalized collision energy was 35.0. The resolving power of the LTQ-Orbitrap mass analyzer was set at 60,000 for the precursor ion scans ($m/\Delta m$ 50% at m/z 400).

We carried out mass analysis under three different modes. In all three modes, one full MS scan was acquired in the Orbitrap, parallel to three MS/MS scans in the LTQ linear ion trap on the three most intense ions from the full MS spectrum. In the first mode, one full MS and three MS/MS were executed. The second mode was the same as the first one except that m/z (445.120025) was used as an internal lock mass to calibrate ions. These two modes used the following Dynamic Exclusion™ settings: repeat count, 2; repeat duration, 30 seconds; exclusion duration, 90 seconds. The third mode was the same as the second mode except that the repeat count is set to 5 in the dynamic exclusion settings. In this study, we considered the data generated from the first mode (13553 peak lists), which will be referred to as “*The Small Dataset*”; and the data generated from all three modes (39913 peak lists), which will be referred to as “*The Large Dataset*”.

2.3 Database Construction

We consider two ways of creating the decoy databases: inverting or scrambling sequences. For the method of scrambling, we also consider the preservation of the frequency of amino acid words of length k ($k=1, 2$ or 3). To facilitate the explanation of various databases used in the search, we create some notations (Table 1). Since O, R, and S are much larger than T in this study, databases composed of T and O, R or S will be referred to as n -fold database, where n is the number of O, R or S (e.g. T-R-S is a 2-fold database). T-O will be referred to as the target database instead of a 1-fold database.

2.4 Database Search

Database search was performed using SEQUEST [2] and MASCOT [6]. The parameters used in database search are as follows: peptide mass tolerance, 0.8 Da; internal miss cleavage allowed, 1; static modification on cysteine (57.0215 Da), 1. PeptideProphet [17] was used to calculate a probability score based on SEQUEST parameters. PeptideProphet probability scores and MASCOT E -values, two popular statistical measures, were then used to select positive hits for SEQUEST and MASCOT search, respectively. We applied various thresholds (PeptideProphet: 0.7-0.99 incremental by 0.01; MASCOT E -values: 0.01-0.1 incremental by 0.003) for the selection of positive peptides.

The number of false positives was estimated at three levels: 1) spectra; 2) unique peptide ions and 3) unique peptide sequences. The difference lies in the way the counting unit is defined. For example, suppose there are 1000 spectra with matching score above some threshold, which corresponds to 800 unique peptide ions and 600 unique peptide sequences; and it turns out that 200 spectra are falsely assigned, which leads to 150 false peptide ions and 120 false peptide sequences. Then the number of false positives at levels 1), 2) and 3) is 200, 150, and 120, respectively.

3 Results

We will show the results at the level of unique peptide sequences in this section. Results at other levels are rather similar and can be found in the Supporting Information. Throughout this section, we will call the peptide with the maximum score (with respect to some spectrum) a *hit*. For simplicity, the actual number of true and false positives will be abbreviated by TP and FP, respectively. Similarly, the actual false discovery rate will be abbreviated by FDR. The estimate of the actual number of false positives will be called “estimated FP”.

3.1 SEQUEST and PeptideProphet

3.1.1 Accuracy of the estimate of the number of false positives—We created a series of 1-fold and 2-fold databases to examine the performance of the reversed and

scrambled databases in terms of the accuracy of the estimated FP. For 1-fold databases, T is combined with one other database (O, R or S) and the hits not in T that pass the specified score threshold represent the false positives. The purpose here is to examine how the number of hits in R or S mirrors that in O when facing the competition from T. For 2-fold databases, T-O (or G) is combined with either R or one of the S databases. The purpose here is to examine how well the number of hits in R or S mimics that in O when the decoy actually is competing with O in addition to T. It should be noted that the 1-fold and 2-fold search are not exactly the same as the Target/Decoy and Target-Decoy approach, which will be studied in a later section.

The results from search against 1-fold and 2-fold databases are shown in Fig 1, where the number of hits in R or S (estimated FP) under different thresholds is plotted against the number of hits in O (FP). It can be seen that the estimated FP based on different decoys for the large dataset is more consistent than those for the small dataset, particularly for 1-fold database search. If the scale of the number of false positives is taken into account, there is less variation for the large dataset than the small dataset (i.e. smaller coefficient of variation). For 1-fold database search, S(1,1)/S(1,2) seems to be a preferred decoy for the small dataset since it is closer to the truth or at least conservative, where R tends to be over-optimistic. For the large dataset, both R and S consistently under-estimate FP and R is the least liberal approach. For 2-fold database search, the magnitude of variation among R and S do not seem to differ much between the small and large datasets, except the search against T-O-S(2,1), which deviates from others to substantial extent for the small dataset. Overall, the estimated FP is less biased than that based on 1-fold search.

The primary reason that we created two copies of decoy databases for each scrambling method is that we want to evaluate the difference between the replicates to get some idea on the magnitude of variation. In other words, not only do we want to estimate FP, but also we want to assess the uncertainty of the estimate when the creation of the decoy has a random nature. Note our evaluation of variation is different from that in [20], where the variation is induced by randomized assignment instead of repeated scrambling. Intuitively, one would expect that replicates of the same type of scrambling are more similar to one another than to replicates from other type of scrambling. We can see some evidence along this intuition, e.g., S(1,1) is more similar to S(1,2) than to S(2,1) in the upper-left graph of Fig 1, though the magnitude is not clearly sensible without a quantitative measure.

We created a metric to quantify the “distance” among curves shown in Fig 1. Specifically, suppose x_i is the number of hits from database A when the i th threshold is applied, and y_i is the corresponding number for database B. Then the *distance index*, denoted by DI , is calculated as

$$DI(A, B) = \sqrt{\frac{\sum_{i=1}^n [2(x_i - y_i)/(x_i + y_i)]^2}{n}}, \quad (1)$$

where n is the number of thresholds under consideration. Essentially, DI can be thought of as the square root of the average adjusted squared distance. Here, “adjusted” refers to the fact that the squared distance $(x_i - y_i)^2$ is adjusted by the scale at the i th threshold $[x_i + y_i]/2]^2$. The term $2(x_i - y_i)/(x_i + y_i)$ is also known as the “symmetrized percent change [22]. If x represents the actual number of false positives at different thresholds and y represents the corresponding estimates, then DI is the square root of the average adjusted squared error. Obviously, small value of DI indicates high level of similarity between x and y .

In Table 2, we show the *DI* values among number of hits in O, R or S for search against 1-fold databases. It can be seen that scrambling that preserves single amino acid frequency (*S*(1,)) is the best approach for the small dataset and R is preferred for the large dataset. Overall, *S*(1,) out-performs others. For the three scrambling approaches, the distance between replicates seems less than the distance between scrambling methods, except *S*(3,). Moreover, the distance between replicates tends to increase as the length of the preserved amino acid words increases. This is likely to be caused by the fact that there is an increasing chance of generating sequences with extreme scores in the decoy by preserving the frequency of long amino acid words. Specifically, by preserving frequency of long amino acid words, the chance of generating a sequence similar to some true sequence increases, leading to high-score hits. On the other hand, it also limits the diversity of new sequences in the decoy, which results in increased chance of low-score hits. The combination of the two factors effectively inflates the two tails of the score distributions and therefore more variation. Considering both performance and the complexity of the procedure, R and *S*(1,) are better choices than other scrambling approaches. For 2-fold approach, we show in Table 3 the *DI* values between O and the corresponding R/S. It can be seen that R is in general better than others.

In summary, 1-fold and 2-fold database search approaches seem to perform similarly well in terms of the accuracy of the estimated FP, though the 1-fold approach is over-confident for the large dataset. As simple as R is, it is at least comparable with scrambling methods. Although the two replicates of each scrambling method do not allow formal statistical inference, the calculation of *DI* suggests that there might be some difference among the three scrambling approaches, particularly the between-replicate variation.

3.1.2 Sensitivity—Although an accurate estimated FP is desirable, we certainly do not want the accuracy to be achieved at substantial cost of sensitivity. This concern arises because the 2-fold search exposes the true sequences to not only the competition from the sequences in O, but also that from R or S, which could undermine sensitivity. Nevertheless, it should be noted that the sequences in O will also face competition from the decoy, which could reduce the number of false positives as well (hits in decoy are treated as true negatives). Therefore, the gain/loss in sensitivity (at a fixed false positive level) by combining some decoy with the target database will depend on the magnitude of the two competing effects.

We studied the sensitivity at various situations where decoy databases of different sizes are concatenated to the target database. First, we examined the Receiver Operating Characteristic (ROC) curves, which are shown in Fig 2. In the first row of Fig 2, the actual numbers of true positives (TP) are plotted against the actual numbers of false positives (FP) for search against databases of various sizes. It can be seen that increasing the size of decoy databases actually result in a better ROC curve in a broad range of the FP, particularly for the large dataset. Therefore, although concatenation of the decoy with the target database could lead to overall fewer hits in T for any score threshold, the same effect on O is such that the ROC is better than without the concatenation. The consequence is that one can adjust the threshold in a way such that the concatenated database has a higher TP at fixed FP level (as compared with search against the target only). Moreover, as for the SEQUEST search coupled with PeptideProphet probability score, adding decoy does not compromise the maximum TP that can be achieved (first row of Fig 2). Note that the ROC curves of 2-fold and 3-fold databases do not differ much, suggesting a plateau of the competition effect. The second row in Fig 2 shows the ROC curves for the search against T-O and various 2-fold databases. It can be seen that various decoys perform quite similarly to each other for the large dataset. There is some difference for the small dataset, particularly the low FP end, where T-O-*S*(1,1) and T-O-*S*(1,2) seem better.

We also studied the relationship of sensitivity and actual false discovery rate (FDR), which is not exactly the same as the ROC curve. Since $FDR = FP / (TP + FP)$, sensitivity versus FDR curve essentially describes the relationship between TP and $FP / (TP + FP)$. The third row of Fig 2 shows that there is no obvious evidence that adding decoy to target in general will lead to lower sensitivity at fixed FDR level for the small dataset, though several of the 2-fold databases do demonstrate slightly lower sensitivity as compared with T-O. On the other hand, for the large dataset, concatenation improves sensitivity.

In summary, the ROC and the sensitivity versus FDR curves together suggest that concatenating decoy with target could improve, or at least not significantly compromise, the sensitivity for SEQUEST search coupled with PeptideProphet probability score.

3.2 MASCOT

3.2.1 Accuracy of the estimate of the number of false positives—Fig 3 is the MASCOT counterpart of Fig 1. R and most scrambling approaches (except S(3,2)) perform reasonably well in the 1-fold scenario and we do not observe the over-confidence for the large dataset as revealed in Fig 1. For 2-fold database search, R is better than most scrambling approaches for the Small Dataset and comparable to most scrambling approaches for the Large Dataset (see Tables A.2 and A.3 in Supporting Information for DI values). Note that T-O-S(3,2) deviates from others substantially at opposite directions as revealed in the upper- and lower-left graphs in Fig 3. A similar deviation of the former is also observed in Fig 1 (lower-left), where S(2,1) seriously underestimates the FP. Thus, there is potential instability associated with scrambling methods that preserve amino acids of length greater than one. Calculation of *DI* values again suggests that there is some evidence that the distance between replicates of the same scrambling approach tends to increase as the length of the preserved amino acid words increases, though the evidence is not very strong. It also seems that the various scrambling approaches might not differ substantially from one another on average (see Supporting Information, Tables A.2 and A.3).

In summary, reversed decoy is a good choice for the estimation of FP based on 1-fold search. For 2-fold search, some scrambled databases seem to be able to provide better estimated FP than R (i.e. S(1,)), but the performance is quite variable.

3.2.2 Sensitivity—Fig 4 is the MASCOT counterpart of Fig 2. It can be seen that concatenation of target and decoy databases leads to some improvement in the ROC curves at the low FP end, and 2-fold and 3-fold databases do not differ significantly (first row). In addition, a feature that is not observed in Fig 2 is that the maximum TP achieved by searching against T-O is higher than search against the concatenated databases (at the cost of more false positives). This is clear evidence that inclusion of more false sequences could compete off true sequences. The second row of Fig 4 demonstrates that there is some variation among the different 2-fold databases at high FP level, particularly for the Large Dataset. The third row shows that adding decoy to the target database does compromise the sensitivity versus FDR curves, which is related to the fact that the maximum TP achieved in the concatenated database is compromised. This is because the search against T-O can generate high TP and FP such that the FDR is the same as that of the search against the 2-fold databases that generates low TP and FP. This leads to a higher TP at fixed FDR for T-O.

In summary, 2-fold search improves the ROC curve at the low FP level, but compromises the sensitivity versus FDR curve.

3.3 Realistic Situations

The 1-fold and 2-fold database searches are not exactly the same as the TargetDecoy and Target-Decoy approaches in practice. Unlike the 1-fold and 2-fold searches where the decoy only includes the altered false sequences in the target database, we do not know what the true sequences are in reality. Therefore, the realistic decoy in practice will include not only the altered false sequences but also the altered true sequences. In the context of the example described in this article, this will not make much difference between 2-fold search and Target-Decoy approach since the number of true sequences is much smaller than that of the false sequences ($T \ll O$). Nevertheless, there is a fundamental difference between 1-fold and TargetDecoy approaches: unlike the search against T-R or T-S, the search against RG or SG in TargetDecoy approach does not expose the altered sequences to the competition from the true sequences. A potential issue is whether or not the lack of competition from the true sequences will lead to more hits in RG or SG than those in O that competes with T, which could lead to an over-estimate of the FP.

Fig 5 shows the estimated FP using the TargetDecoy and Target-Decoy strategy based on MASCOT *E*-values. It can be seen that the TargetDecoy approach indeed tends to over-estimate the FP. On the other hand, the Target-Decoy results do not seem to differ very much from the 2-fold results shown in Fig 3 (second row). Similarly, the ROC curves of the Target-Decoy show similar pattern as the 2-fold search (not shown).

4 Discussion

False positive control/estimate is critically related to the quality of high-throughput peptide/protein identification using tandem mass spectrometry, which has profound impact on downstream bioinformatic analysis of these peptides/proteins. We studied various decoy-based strategies for the estimation of FP using data collected on a mixture of model proteins. We focused on two issues: (a) how the decoy should be assembled and (b) how the search should be executed. To compare different methods, we considered two criteria: (i) the accuracy of the estimated FP and (ii) the sensitivity (at a fixed FP or FDR level). In terms of (a), reversing sequences offers a simple and reasonably accurate way to re-construct characteristics of the false sequences in the target database that are relevant to the score distribution. It is comparable by criteria (i) and (ii) to most scrambling strategies studied in this article for both PeptideProphet probability scores and MASCOT *E*-values. As for (b), our study suggests that search against a concatenated database composed of the target and decoy databases has more advantages than search against the target and decoy separately. In terms of criterion (i), the study of PeptideProphet probability scores indicates that the former approach has less bias in the estimated FP (Fig 1), and the TargetDecoy approach based on MASCOT *E* value tends to over-estimate the FP due to the lack of competition from the true sequences (Fig 5). In terms of criterion (ii), same preference is evident by better ROC curve for both PeptideProphet score and MASCOT *E*-value, and better TP versus FDR curve for PeptideProphet score. In summary, a Target-Decoy strategy with decoy being the reversed version of the target database is a practically simple and reasonably accurate strategy for the estimation of FP, which could also improve sensitivity.

Results from 1-fold/2-fold and SEQUEST/MASCOT search do not offer a very clear view on how we should preserve local sequence nature by scrambling to mimic the false sequences in the target database. It seems scrambling could offer better performance than reversing under some scenarios. Nevertheless, it is not clear when and why it will perform superiorly. In addition, the scrambling demonstrates quite some variation across replicates, particularly with the 2-fold search or Target-Decoy strategy. Although the search against 1- and 2-fold databases serves as a theoretical investigation due to its impracticality, it generates some interesting findings deserving further studies. For instance, the scrambling

method that preserves the frequency of amino acids (S(1,)) seems more stable than the other two scrambling-based approaches. A study with more replicates might offer more precise insight into the variation pattern.

It is still a challenging task to understand the mechanistic nature of how the decoy strategy works, given the complexity and volume of the sequences in a database. For any spectrum, it is possible that some of the candidate peptides (i.e. peptides that fall into the relevant mass interval) are generated by two *in silico* cleavages at the C-terminus of the same amino acid (i.e. lysine or arginine for trypsin). These candidates will be preserved by reversing because the reversed version has the same mass. If an original candidate is generated by cleavages at the C-terminus of different amino acids, then the reversed version will not be a candidate because of the mass difference. However, for the same reason, some new candidates could be generated. Therefore, at the level of number of candidates, reversed database could offer a reasonable match to the target database, though the goodness of match could differ from spectrum to spectrum. This is likely to be part of the reason why decoy generated by reversing sequences provides a reasonably good estimated FP. The mechanism of scrambling is much more complicated. One observation we made in our study is that preserving long amino acid words tends to result in less stable estimate (Table 2). One possible reason is that preserving long amino acids has higher chance of generating peptides with extreme scores, as compared with preserving shorter amino acid words. On one side, it has higher chance of generating peptides very similar to the true peptide, leading to high scores. On the other hand, it limits the number of combinations and could be more likely to generate low scores. The inflated “tails” in both directions in the score distribution can cause higher variation in number of hits passing certain threshold.

In general, the performance of various decoy-based strategies is likely to depend on the nature of the target database (i.e. nature of the sequences, size, and the number of true sequences). Our study is based on the data collected on a mixture of model proteins. The conclusions made here might not be generalized to practical applications where complex biological samples are analyzed with distinct sequence nature and variable database sizes. In particular, the much higher scale of the number of true sequences in practice could have substantial impact on the performance of the various decoy-based strategies. Due to the lack of a complex data with golden standard, it is still a challenging task to rigorously evaluate the performance of various decoy approaches for complex samples. Nevertheless, our results on the estimation of FP are consistent with some findings based on complex samples [20]. The ongoing efforts on spectrum library construction [23-25] offer an opportunity to tackle the difficulty. Potentially, a large number of spectra with known precursor ions can be thought of as the spectra from a hypothetical sample, which can be used for similar studies as described in this article. This is likely to generate some insight on the performance of various decoy-based strategies in estimating the number of false positives for more realistic scenarios.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Changyu Shen's effort in this work is supported by Department of Defense grant BC030400, Indiana Alzheimer's disease center (IADC), and Indiana University Melvin and Bren Simon Cancer Center. Dr. Haixu Tang acknowledges the support from NIH/NCRR grant 5P41RR018942 and NSF grant DBI-0642897.

References

1. Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*. 2001; 17 1:S13–21. [PubMed: 11472988]
2. Eng JK, McCormack AL, Yates JR 3rd. 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; 5:976–989.
3. Feng J, Naiman DQ, Cooper B. Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Anal Chem*. 2007; 79:3901–3911. [PubMed: 17441689]
4. Geer LY, Markey SP, Kowalak JA, Wagner L, et al. Open mass spectrometry search algorithm. *J Proteome Res*. 2004; 3:958–964. [PubMed: 15473683]
5. Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem*. 2003; 75:435–444. [PubMed: 12585468]
6. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20:3551–3567. [PubMed: 10612281]
7. Sadygov RG, Yates JR 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem*. 2003; 75:3792–3798. [PubMed: 14572045]
8. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*. 2007; 6:654–661. [PubMed: 17269722]
9. Xue X, Wu S, Wang Z, Zhu Y, He F. Protein probabilities in shotgun proteomics: evaluating different estimation methods using a semi-random sampling model. *Proteomics*. 2006; 6:6134–6145. [PubMed: 17133366]
10. Zhang N, Aebersold R, Schwikowski B. ProbiD: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*. 2002; 2:1406–1412. [PubMed: 12422357]
11. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem*. 2003; 75:768–774. [PubMed: 12622365]
12. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*. 1995; 57:289–300.
13. Efron B, Tibshirani R, Storey JD, Tusher VG. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*. 2001; 96:1151–1160.
14. Muller, P.; Parmigiani, G.; Rice, K. ISBA 8th Meeting on Bayesian Statistics; Alicante, Spain. 2006.
15. Shen C, Wang Z, Shankar G, Zhang X, Li L. A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics*. 2008; 24:202–208. [PubMed: 18024968]
16. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*. 2002; 64:479–498.
17. 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; 74:5383–5392. [PubMed: 12403597]
18. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–4658. [PubMed: 14632076]
19. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res*. 2008; 7:47–50. [PubMed: 18067251]
20. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007; 4:207–214. [PubMed: 17327847]
21. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*. 2008; 7:40–44. [PubMed: 18052118]

22. Berry DA, Ayers GD. Symmetrized Percent Change for Treatment Comparisons. *The American Statistician*. 2006; 60:27–31.
23. Craig R, Cortens JC, Fenyo D, Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res*. 2006; 5:1843–1849. [PubMed: 16889405]
24. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem*. 2006; 78:5678–5684. [PubMed: 16906711]
25. Lam H, Deutsch EW, Eddes JS, Eng JK, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 2007; 7:655–667. [PubMed: 17295354]
26. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics (Oxford, England)*. 2001; 17:282–283.
27. Coward E. Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics (Oxford, England)*. 1999; 15:1058–1059.

Abbreviations

DI	distance index
FDR	false discovery rate
FP	actual number of false positives
TP	actual number of true positives
ROC	receiver operating characteristic
PSM	peptide-spectrum-match

Fig 1.

Estimates of the number of false positives at the peptide level based on SEQUEST search and PeptidProphet probability scores at different thresholds (0.7 to 0.99 by 0.01): 1-fold and 2-fold databases. Estimated FP: number of hits in R or S, FP: number of hits in O.

Fig 2.

Sensitivity at the peptide level based on SEQUEST search and PeptideProphet probability scores at different thresholds (0.7 to 0.99 by 0.01). **First row:** TP versus FP for all types of databases. 2-fold: T-O-R, T-O-S(1,1), T-O-S(1,2), T-O-S(2,1), T-O-S(2,2), T-O-S(3,1), T-O-S(3,2); 3-fold: T-O-R-S(1,1), T-O-R-S(1,2), T-O-R-S(2,1), T-O-R-S(2,2), T-O-R-S(3,1), T-O-R-S(3,2). TP: number of hits in T, FP: number of hits in O, FDR: $FP/(TP+FP)$. **Second row:** TP versus FP for 2-fold databases (T-O shown as a reference). **Third row:** TP versus FDR for 2-fold databases.

Fig 3.

Estimates of the number of false positives at the peptide level based on MASCOT *E*-values at different thresholds (0.01 to 0.1 by 0.003): 1-fold and 2-fold databases. Estimated FP: number of hits in R or S, FP: number of hits in O.

Fig 4.

Sensitivity at the peptide level based on MASCOT *E*-values at different thresholds (0.01 to 0.1 by 0.003). Graph contents are the same as Fig 2.

Fig 5.

Estimates of the number of false positives at the peptide level based on MASCOT *E*-values at different thresholds (0.01 to 0.1 by 0.003): Target|Decoy and Target-Decoy search. Estimated FP: number of hits in RG or SG, FP: number of hits in O.

Table 1
Notations of various sequence databases

Label	Description
T	True sequences: the collection of the 19 standard protein sequences in the sample. T in practice represents the pool of true sequences in the target database.
O	False sequences: non-redundant IPI sequences of Arabidopsis database (version 3.17, downloaded from EBI). Software cd-hit[26] was used to remove the redundant sequences that have more than 0.7 sequence identity with true sequences or other longer sequences in original IPI database. O in practice represents the false sequences that are mingled with the true sequences.
G	Target database: the union of T and O. G in practice is the database searched against.
R	Reversed false sequences: the reversed version of O.
S	Scrambled false sequences: the scrambled version of O. $S(i,j)$ refers to the j th copy of the scrambling that keeps the frequency of the words composed of i amino acids. Software shufflet [27] was used to scramble the sequences.
RG	Reversed target database: reversed version of G.
SG	Scrambled target database: the scrambled version of G. $SG(i,j)$ refers to the j th copy of the scrambling that keeps the frequency of the words composed of i amino acids.
When database are concatenated, the resulting database will be labeled by the participating databases connected by hyphens. For example, T-R-S refers to the union of databases T, R and S.	

Table 2

Distance index in terms of number of hits in O, R or S at the unique peptide sequence level for search against various 1-fold databases: SEQUEST and PeptideProphet results. Hits in scrambled databases are averaged (across the two replicates) when calculating the distance with other search. Within scrambling category comparison is based on the distance between the two replicates (e.g. the distance between T-S(1,1) and T-S(1,2) is 0.36 for The Small Dataset).

	The Small Dataset					The Large Dataset				
	T-O	T-R	T-S(1,)	T-S(2,)	T-S(3,)	T-O	T-R	T-S(1,)	T-S(2,)	T-S(3,)
T-O	0	0.83	0.33	0.98	0.55	0	0.38	0.42	0.79	0.66
T-R	0.83	0	0.96	0.28	0.67	0.38	0	0.20	0.48	0.36
T-S(1,)	0.33	0.96	0.36	1.10	0.60	0.42	0.20	0.22	0.45	0.29
T-S(2,)	0.98	0.28	1.10	0.75	0.78	0.79	0.48	0.45	0.18	0.22
T-S(3,)	0.55	0.67	0.60	0.78	1.08	0.66	0.36	0.29	0.22	0.36

Table 3

Distance index of number of hits in O and that in R or S at the unique peptide sequence level for search against various 2-fold databases: SEQUEST and PeptideProphet results.

Database	T-O-R	T-O-S(1,1)	T-O-S(1,2)	T-O-S(2,1)	T-O-S(2,2)	T-O-S(3,1)	T-O-S(3,2)
Comparison	O vs. R	O vs. S(1,1)	O vs. S(1,2)	O vs. S(2,1)	O vs. S(2,2)	O vs. S(3,1)	O vs. S(3,2)
The Small Dataset	0.31	0.76	0.84	1.26	0.60	0.29	0.41
The Large Dataset	0.43	0.87	0.83	0.68	0.25	0.81	0.77