Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches

Zheng Zhang¹*, Meghan Burke¹, Yuri A. Mirokhin¹, Dmitrii V. Tchekhovskoi¹, Sanford P. Markey¹, Wen Yu², Raghothama Chaerkady³, Sonja Hess³, Stephen E. Stein¹*

¹Mass Spectrometry Data Center, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899 USA

²Research Bioinformatics, MedImmune LLC, One MedImmune Way, Gaithersburg, Maryland 20878 USA

³Antibody Discovery and Protein Engineering, Protein Sciences, MedImmune LLC, One MedImmune Way, Gaithersburg, Maryland 20878 USA

Correspondence to:

*E-mail: <u>zheng.zhang@nist.gov</u> Tel: +1 301 975 5828. <u>steve.stein@nist.gov</u>. Tel: +1 301 975 2505.

KEYWORDS: *peptide mass spectral library, target-decoy approach, PeptideProphet algorithm, false discovery rate*

ABBREVIATIONS: FDR, false discovery rate; **SLS**, Spectral library searching; **SDS**, sequence database searching; **LC-MS/MS**, liquid chromatography-tandem mass spectrometry; **PSM**, peptide-spectrum match; **NIST**, National Institute of Standards and Technology; **NCBI**, National Center for Biotechnology Information; **NIH/NCI**, National Institutes of Health/National Cancer Institute; **CPTAC**, Clinical Proteomic Tumor Analysis Consortium; **HCD**, higher-energy collisional dissociation; **iTRAQ**, isobaric tags for relative and absolute quantitation; **IDs**, identifications; **SUI**, sequence uniqueness index.

ABSTRACT

Spectral library searching (SLS) is an attractive alternative to sequence database searching (SDS) for peptide identification due to its speed, sensitivity, and ability to include any selected mass spectra. While decoy methods for SLS have been developed for low mass accuracy peptide spectral libraries, it is not clear that they are optimal or directly applicable to high mass accuracy spectra. Therefore, in this paper we report the development and validation of methods for high mass accuracy decoy libraries. Two types of decoy libraries were found suitable for this purpose. The first, referred to as Reverse, constructs spectra by reversing a library's peptide sequences except the C-terminal residue. The second, termed Random, randomly replaces all non-Cterminal residues and either retains the original C-terminal residue or replaces it based on the amino-acid frequency of the library's C-terminus. In both cases the m/z values of fragment ions are shifted accordingly. Determination of FDR is performed in a manner equivalent to SDS, concatenating a library with its decoy prior to a search. The utility of Reverse and Random libraries for target-decoy SLS in estimating false positives and FDRs was demonstrated using spectra derived from a recently published synthetic human proteome project.¹ For data sets from two large-scale label-free and iTRAO experiments, these decoy building methods yielded highly similar score thresholds and spectral identifications at 1% FDR. The results were also found to be equivalent to those of using the decoy-free PeptideProphet algorithm. Using these new methods for FDR estimation, MSPepSearch, which is freely available search software, led to 18% more identifications at 1% FDR and 23% more at 0.1% FDR when compared with other widely-used SDS engines coupled to post-processing approaches such as Percolator. An application of these methods for FDR estimation for the recently reported 'hybrid' library search¹⁶ method is also made. The application of decoy methods for high mass accuracy SLS permits the merging of these results with those of SDS, thereby increasing the assignment of more peptides, leading to deeper proteome coverage.

INTRODUCTION

A critical step in all bottom-up proteomic studies is the accurate identification of as many peptides as possible in the biological material under investigation. This has primarily been done through sequence database searching (SDS) methods that match input spectra with predicted peptide ion spectra contained in a protein sequence database.^{2,3,4} High-throughput liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis has become routine and the sharing of high-quality peptide spectral data from large-scale proteomic studies is now a common practice. This has enabled the development of comprehensive mass spectral libraries for use in peptide identification by spectral library searching (SLS). In this approach, query spectra are matched to a collection of library spectra initially identified by either SDS^{2,3,4} or from synthetic peptides.¹ SLS is an attractive alternative to SDS for peptide identification due to its speed, sensitivity, and ability to include any selected mass spectra.^{5,6,7,8} A general view is that SDS and SLS results can reinforce each other leading to an increase in numbers of confidently identified peptides and proteins. When extensive libraries are unavailable, SDS remains the principal choice in discovery proteomics. A main concern in using SLS is the "incompleteness" of libraries. However, with the publication of two extensive studies mapping the human proteome and the recent release of peptide spectral data expected to lead to a comprehensive synthetic human proteome,^{1,9} the time is ripe for SLS to become routinely used alongside SDS in

discovery proteomics for samples of human origin. Preliminary tests show, not surprisingly, that improved results are obtained if SDS and SLS results are merged. One reason for the relative rarity of such practice is the uncertainty in merging results of SLS with those of SDS. This stems primarily from the lack of widely accepted and practiced decoy methods for computing false discovery rates (FDRs) in SLS analyses. The principal objective of this paper is to provide such a method.

In SDS, estimating FDR (i.e., the estimated fraction of wrong answers in the result) with targetdecoy methods has been extensively investigated,^{10,11} and is currently the accepted standard practice. In a target-decoy SDS analysis, experimental spectra are searched against both a normal protein sequence database (i.e., the target) and a concatenated decoy version of the database. The number of decoy hits above a chosen threshold score is used to estimate the number of false positives and the FDR. The most widely employed methods for generating decoy databases are reversing (a simple reversal of presumed amino acid sequences)¹⁰ and randomization (randomly generating protein sequences according to the occurrence frequencies of amino acids in a target database).¹¹

Recently efforts have also been made in adopting the target-decoy approach to SLS. Instead of decoy sequences, decoy spectral libraries are constructed for FDR estimation. Several decoy library methods have been developed and tested with low-resolution peptide spectral libraries. The first such method was reported by Lam et al.¹³ in which a shuffle-and-reposition strategy was used to construct decoy spectra. A refined version, DeLiberator,²¹ was implemented by giving special treatment to unannotated peaks. Another decoy generating method, precursor swapping, is also available.¹² These decoy methods have been utilized in various SLS applications including mining protein modifications.²² However, decoy library methods have not been reported for high mass accuracy peptide spectra, which is the focus of this paper. Because of the much higher degree of specificity of these spectra, it is not clear to us how well the low mass accuracy spectra methods will perform. In fact, as described later, one method that performed well for low mass accuracy spectra.

In the present study, we used the general concepts of reversing and randomization from SDS, and developed algorithms to create reverse and random peptide spectral libraries for high mass accuracy SLS. We tested the performance of these decoy libraries in estimating FDRs with data sets in which whether or not an individual spectrum was in the library was known and with data sets from large-scale high-resolution LC-MS/MS experiments. In addition, we compared the results with those obtained from the decoy-free PeptideProphet algorithm,^{14,15,20} and with those from target-decoy SDS. By applying similar decoy strategies, the methods described in this study bridge the gap in target-decoy searches between SDS and SLS, enabling integration of SDS and SLS results in a common FDR estimation framework. The integration of results from both will benefit the general proteomics community, especially for those interested in exploiting the unique strengths of both SDS and SLS and combining their results for greater coverage.

METHODS

Decoy library building: Two types of decoy libraries, Reverse and Random, were developed and tested for high mass accuracy peptide SLS using NIST-developed libraries and a library searching program MSPepSearch (peptide.nist.gov¹). Briefly, a decoy library is constructed from a target peptide spectral library spectrum-by-spectrum. For each library spectrum, the process begins by building a decoy peptide whose sequence is generated either randomly or by reversing the target spectrum's peptide sequence. The target spectrum is then translated into a decoy spectrum by shifting its fragment ions to computed m/z values expected for the decoy peptide (referred to as reposition in Lam et al.).¹³ Input msp-formatted libraries are pre-annotated using methods described in ref. 8, which is taken advantage of by our decoy methods. The algorithms were implemented in 'R'. The code is available from authors upon request. General descriptions of the Reverse and Random algorithms follow below. For both, no alterations to fragment ion intensities, the number of amino acids, or precursor charge state are made.

Reverse: For each spectrum in a target library, its associated peptide sequence is reversed to form a decoy peptide. The C-terminal residue, mostly K or R for a spectral library of tryptic peptides, is kept in its original position and only the order of the non-C-terminal amino acids is inverted. For modifications, those that are specific to an amino acid (e.g., oxidation of methionine, carbamidomethylation of cysteine, isobaric tag on lysine) move along with the amino acid to its new position, and those associated with the N-terminus (e.g., isobaric tagging for quantification) remain and are attached to the new N-terminal end after reversing. Exclusion of palindromes (peptides whose forward and reversed sequences are the same) is optional, and they are relatively infrequent and found to have no noticeable effect on results. Once the decoy sequence and modifications are determined, m/z values of fragment ions (e.g., a, b, and y ions for higher-energy collisional dissociation (HCD) fragmentation, and their 13C isotopes and neutral losses of ammonia and/or water) in the target spectrum are shifted to the corresponding m/z values predicted for the decoy sequence. Any mass errors (the difference between experimental and theoretical masses) of fragment ions are retained in the shifted m/z values. Fragment ions derived from losses from the parent ion are not shifted.

Random: In this method, all non-C-terminal amino acids are randomly replaced with amino acids with the constraint that the distribution of amino acids in the decoy library matches the distribution in the target library. To preserve positions of protease-specific residues, the decoy peptide's C-terminus is either taken directly from that of the target peptide (FixC) or randomly selected based on the target library's C-terminal residue distribution (RanC). Modifications are also randomly assigned to the decoy peptide as per their occurrences in the target library. For RanC, we first calculate the frequencies of library peptides' C-terminal amino acids and their modifications, and decoy C-terminal amino acid is generated from this frequency distribution. If the target peptide contains an N-terminus-specific modification (e.g., in libraries of isobaric-tagged spectra), this same modification is added to the N-terminus of the decoy peptide. In addition, steps are taken to ensure that (1) a newly generated decoy sequence does not match any peptide in the target library or any previously generated decoy sequences, maintaining the uniqueness of decoy peptides, and (2) if a peptide ion has multiple spectra (i.e., peptide-spectrum matches (PSMs)) in a target library, the same decoy sequence is reused to generate all the corresponding decoy spectra, keeping the same number of unique amino acid sequences in the

¹ Please go to <u>http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload</u> to download libraries used in this work (Table 1).

target and decoy libraries. When the decoy peptide sequence and its modification(s) are established, m/z values of fragment ions in the target spectrum are replaced by decoy peptide values in the same manner as for the Reverse method. Parent-related fragment ions are also shifted by the difference in precursor mass between the original and decoy peptides.

Spectrum Library Searches: Freely available MSPepSearch software (<u>peptide.nist.gov</u>) was used to perform SLS using a precursor ion tolerance of 20 ppm and fragment ion tolerance of 50 ppm for most searches. For fragment ion tolerance, we tested 50 ppm and 20 ppm, and found no significant difference, with 50 ppm giving slightly more IDs in some cases. For the hybrid spectral library searches¹⁶, the fragment ion tolerance was 40 ppm. Label-free and iTRAQ 4-plex labeled human tryptic peptide spectral libraries are available on-line (<u>peptide.nist.gov</u>).¹⁷ Spectra in these libraries originated from HCD fragmentation, where the best quality spectrum (e.g., highest MSGF+ score) was used for each identified peptide ion at each reported fragmentation energy.⁷ Both libraries contain more than 1 million HCD spectra and cover approximately one-third of all identifiable tryptic peptides in the human proteome.¹⁷ The NIST format libraries were built using the freely available program Lib2NIST (<u>peptide.nist.gov</u>).

Data Sources: Testing data sets of label-free and iTRAQ 4-plex labeled spectra were obtained from PRIDE¹⁸, NIST and the NIH/NCI CPTAC program (proteomics.cancer.gov). Table 1 summarizes data sets and libraries (peptide.nist.gov).¹⁷

Determining the Threshold Scores at Fixed FDR levels: A target-decoy strategy was employed to establish score thresholds for achieving false discovery rates (FDRs) at desired levels. Determination of FDR of SLS is performed in a manner equivalent to SDS, concatenating a library with its decoy prior to a search. For example, at 1% FDR, a score cutoff was selected so that above that cutoff, the numbers of PSMs from searching a decoy library was approximately 1% (FDR was estimated by 2xDecoy/(Decoy+Target)) of the numbers of PSMs from searching a target library. The target-decoy methods were then compared with the decoy-free PeptideProphet algorithm.²⁰ Correct and incorrect MSPepSearch scores were modeled by the distributions suggested in ref. 20, to obtain score cutoffs.

Sequence Database Searches: To further compare SLS with SDS, we performed SDS with Sequest³ coupled to Percolator¹⁹ in Proteome Discoverer (2.1). Testing data sets and mass tolerance settings were the same as for the library search for a fair comparison. For label-free spectra, oxidation on M was set as a variable modification and CAM on C as fixed modification. Semi-tryptic peptides were allowed. The search engine rank was 1 and FDR level was usually set at 1% or otherwise mentioned. RefSeq NCBI Aug. 2014 human fasta file was downloaded (55 926 sequences).

RESULTS AND DISCUSSION

As described above, we developed methods for constructing two types (Reverse and Random) of decoy peptide spectral libraries. The methods are conceptually equivalent to the reversing and randomization of records in SDS, but are implemented differently because of the spectrum-centric nature of SLS versus the sequence-centric SDS. Examples of decoy spectra created by the

Reverse and Random methods as well as the corresponding target spectra are displayed in Fig. 1A and Fig. 1B, respectively.

Unlike SDS, where methods for generating decoy sequence databases are generally accepted, no widely practiced decoy methods have been developed for high-resolution SLS. Two reported decoy library methods include peptide Precursor-Swap¹² and shuffle-and-reposition¹³ developed with low-resolution peptide spectra. For the shuffle-and-reposition method, its limitations include a possible failure to produce distinct decoy sequences for target peptides that are of low complexity (e.g., sequences consisting of the repeat of only one or a few amino acids) or relatively short in length. These will be further discussed later.

When creating decoy peptides, our option for the C-terminus is to either copy it directly from target peptides or select randomly based on its distribution in a target library. Since spectral libraries are usually composed of peptides resulting from digestion by specific enzymes (mostly trypsin), our approach preserves the enzyme-specific property (e.g., tryptic) of the target peptides.

Characterizing Decoy Spectral Libraries

Just as a decoy used in hunting should appear indistinguishable from a real prey, ideally a decoy library should possess characteristics equivalent to those of a target library – it should be realistic and wrong. These characteristics include distributions of peptide length and amino acids, precursor mass, and fragment peak abundances, as well as modifications and the number of distinct peptide sequences. In other words, decoy and target libraries should resemble each other statistically except that their sequences and spectra are different.

Since the basic idea behind the Reverse method is a simple reversal of a peptide's amino acid sequence, a decoy peptide spectral library constructed by this method will have identical amino acid occurrence frequencies, peptide lengths, precursor mass distributions, numbers of distinct peptide sequences, and modifications. A potential concern is that a decoy library may share some palindromic peptide sequences with a target library. However, the number of peptide palindromes encountered is very small. For example, in the human iTRAQ 4-plex labeled library, our largest collection with 390 009 distinct peptides, there were only 65 palindromic sequences (e.g. SLDLDISK). No measurable effect was found when these were deleted from the decoy library (data not shown). I and L are treated separately as any other amino acids in calculating their frequencies and in generating decoy amino acids according to their frequency distribution. The only times I and L are considered the same are when determining the distinctness of decoy sequences and whether a sequence is palindromic. A more significant concern is that peptides from proteins containing significant non-random distributions of amino acids will retain this non-random nature even after shuffling. To examine this idea, we defined a sequence uniqueness index (SUI) as the number of unique amino acids in a peptide divided by its length (the total number of amino acids). An SUI close to 0 suggests low complexity while an SUI close to 1 suggests few of the amino acids are repeated. Supplementary table 1 shows that a substantial fraction of peptides with low SUI values was found in the target and Reverse libraries, but significantly less in the Random library. It suggests that the Random method avoids

the problem of non-random distributions of amino acids inherent in the Reverse method, and therefore suppresses this variety of "homology".

Another way of comparing results of different methods is to find relative numbers of matches from the target versus decoy libraries at different ranks for target-decoy concatenated library searches. Results are shown in Fig. 2A for the top 10 hits using a large-scale label-free proteomic data set.²³ We found that for the Reverse, Random, and Precursor-Swap methods, the average percent of decoy IDs were 48.4%, 45.3%, and 40.3% respectively from ranks 2 to 10 (rank 1 is heavily influenced by the fraction of search spectra found in the library). Also shown are results from Elias et al.¹⁰ for Sequest (an SDS method) using reversed protein sequences as the decoy and using a Jurkat cell digest as the test data. Our Reverse method closely matched their results. The lower value for the Random versus the Reverse method presumably reflects the elimination of sequence "homologies" (discussed in the next section), which, in effect, serves to lower scores for some identifications. The 99.4% precursor mass overlap shown in Fig. 3 and supplementary table 3 for 20 ppm tolerance suggests that this is not caused by a mismatch in precursor mass distributions. It was also found that decoy ratios for the Random method between 10 and 40 ppm were virtually the same. In any case, the actual effect on computed score thresholds and number of identifications at a fixed FDR level between Reverse and Random methods is rather small (see later). The low values (especially for ranks 1 to 3) for the Precursor-Swap method¹² appear to originate from a mismatch of shifted precursor ion masses and un-shifted product ion masses that did not significantly affect results for low mass accuracy spectra.

As detailed in the Methods section, a decoy peptide spectral library constructed by the Random method will have the same peptide length distribution and the same number of distinct peptide sequences (i.e. search space). A separate step ensures that there are no shared peptide sequences between the target and decoy libraries. Supplementary table 2 shows an example of amino-acid occurrence frequencies in non-C-terminal residues and in C-terminal residues of a target/decoy library pair together with the normalized frequencies of modifications associated with an amino acid. The amino-acid and modification frequencies were virtually identical. In addition, all three pairwise target/ranC comparisons for both non-C-terminal and C-terminal residues gave virtually the same amino-acid distributions (t-test's p = 1 and correlation coefficient r > 0.9999).

For the Random method to be reliable, derived decoy libraries should: (1) have precursor mass distributions similar to target libraries and (2) perform similarly to each other. To test point (1) we compared the precursor mass distributions of three random decoy libraries with that of their target label-free library. Fig. 3 shows a representative example of histograms for precursor masses in a 500 m/z to 500.5 m/z range with an m/z tolerance of 20 ppm. Results of an overall examination given in supplementary table 3 show that \geq 51.4% of target precursor masses (rounded to 4 decimal digits) had exact matches in each of the decoy libraries (this generally means that the chemical formulas of the decoy and library peptides match), and \geq 96.0%, \geq 98.4%, and \geq 99.4% of target precursor masses were found in each decoy libraries had effectively identical distributions of precursor masses to the target library within the mass accuracy of a typical proteomic analysis. Furthermore, to test point (2) we compared the performance of these three random decoy libraries and found they performed equivalently in target-decoy SLS, which will be addressed later.

Testing the Reverse and Random Methods in Target-Decoy SLS

The proposed methods in target-decoy SLS were tested using a selected subset of the HCD spectra for over 330 000 synthetic tryptic peptides made publicly available through PRIDE.¹ We downloaded their first pool raw files and performed a target-decoy SDS with MSGF+. After passing our criteria for inclusion in a library and excluding short (<10 residues) and long (>25 residues) sequences, 34 340 spectra (each representing a distinct peptide sequence; if a peptide ion had multiple PSMs, only its highest scoring spectrum was selected) were extracted and used as a test data set. These spectra were derived from different HCD energies (25, 30, and 35), covering a wide range of real-world experimental conditions. This makes them especially suitable for testing with our library, whose spectra were collected from different sources and conditions (labs, instruments, and energies). A subset (19 343) of these spectra (whose corresponding peptide sequences were found to be present in our human label-free spectral library) served as "in library" spectra (all expected to match). In Fig. 4A, we used these spectra as inputs to search against the original (target) library and its Reverse and Random decoys separately. The MSPepSearch score distribution from the target search was well separated from two decoy searches, as expected. The other 14 997 spectra whose sequences had no corresponding entries in our library served as "not in library" spectra (none expected to match). In Fig. 4B, results show that using these spectra as query spectra to search against the original (target) library and its Reverse and Random decoys separately, MSPepSearch score distributions from two decoy searches were highly similar to that from the target search, as well as from searches using 'in library' spectra searched against decoy libraries.

While the Reverse and Random methods gave similar results (Fig. 4), as we further examined high scoring false hits, we found in Reverse method a high proportion of them were peptides with relatively few different amino acids – an extreme case was QQQQLQQQQR that gave an SUI value as low as 0.27. We calculated the peptide SUI values of inputs ("in library" or "not in library" spectra as query spectra) for matches scored above 250 (Table 2), and noticed that target/Reverse searches had significantly more matches with SUI values of 0.1 to 0.5 than Random search had after normalization. In addition, this finding was reliable for the three different versions of Random decoys (RanC1, 2, and 3). The result suggests that a low-complexity peptide spectrum has a higher chance of finding a homologous match in a library. This increased "homology matching" could partially explain the slightly higher score cutoff at a fixed FDR using a Reverse decoy than a Random one (also supported by Fig. 5 and 6, which will be discussed later).

We also show, in Fig. 4A and 4B, score distributions for the Precursor-Swap method using synthetic peptide spectra that are present and absent, respectively, in the target library. Most notable is its significant lower scores, especially in the higher-score region. This is consistent with the findings shown in Fig. 2.

We also determined decoy/target ratios broken down by rank for the 14 997 "not in library" synthetic peptide spectra (Fig. 2B). For ranks 1 to 10, ratios for the Reverse and Random methods varied from 46% to 48% and 43% to 45% respectively. The former closely matched published values by Lam et al.,¹³ obtained from the shuffle-and-reposition method based on low

resolution SLS. On the other hand, the Precursor-Swap method, which worked well for low mass accuracy spectra,¹² gave poor results for this high mass accuracy test. Decoy ratios for the Random method were uniformly about 3% lower than for the Reverse method, consistent with experimental proteomic data results (Fig. 2A).

Label-Free and iTRAQ 4-plex labeled Spectrum Searching

To further examine Reverse and Random strategies for library identifications of high mass accuracy peptide spectra, we performed target-decoy searches using the NIST-developed comprehensive label-free peptide spectral library and iTRAQ 4-plex labeled peptide spectral library (<u>peptide.nist.gov</u>)¹⁷ with high-resolution Orbitrap HCD data acquired in large-scale studies. The results at 1% FDR cutoff were then compared with those obtained by the decoy-free PeptideProphet algorithm initially designed for SDS and later adapted for SLS.^{14,15}

We first used 605 113 label-free spectra from PRIDE¹⁸ (Table 1) as query spectra to search against the NIST-developed label-free spectral library (1 127 970 spectra, 320 824 distinct peptide sequences) and its decoys, comparing the thresholds and the number of IDs at 1% FDR with that of the decoy-free PeptideProphet algorithm. Fig. 5 shows that the numbers of IDs obtained from different methods at 1% FDR were very similar: (A) 331 373 IDs by PeptideProphet algorithm, (B) 328 512 IDs by Reverse method (0.86% less than A), (C) 334 940 IDs by Random method, FixC (1.08% greater than A), (D) 334 951 IDs by Random method, RanC (1.08% greater than A). We further compared three different versions of the Random libraries (RanC1, 2 and 3) and obtained nearly identical IDs for each: (RanC1) 334 951 vs. (RanC2) 335 005 vs. (RanC3) 335 026.

Since isobaric labeling alters the distribution of peak intensities by increasing the contribution from b-ions, presumably due to their stabilization by the tag, we separately examined FDR results for this class of spectra. We tested a set of iTRAQ 4-plex labeled query spectra (994 133) against the NIST human iTRAQ 4-plex labeled spectral library (1 201 632 spectra, 390 009 distinct peptide sequences) and its decoys. Results are shown in Fig. 6. Again at 1% FDR, the numbers of IDs obtained from different methods were quite similar: (A) 304 468 IDs by PeptideProphet algorithm, (B) 306 227 IDs by Reverse method (0.58% greater than A), (C) 306 368 IDs by Random method, FixC (0.62% greater than A), (D) 308 908 IDs by Random method, RanC (1.46% greater than A).

In summary, using both labeled and unlabeled data sets, we found that Reverse and Random decoy libraries were suitable for FDR estimation using the target-decoy approach, and results were equivalent to those obtained by the decoy-free PeptideProphet algorithm. As discussed earlier for the synthetic peptide search results, cutoffs using the Random approach were slightly lower than that of using Reverse approach so that the number of IDs were slightly higher (1% to 2% more IDs), presumably due to the removal of peptides with highly non-statistical amino acid occurrences.

While it is customary to create decoy libraries of the same size as the target library, there is no inherent need for this, since the number of hits from a decoy library should scale linearly with its number of peptides. In fact, if libraries contain too few entries, FDR results could be quite unreliable. To examine the issue, we determined effects of library size on FDR estimation, for

library subsets diminished in size by factors of two from 1/2 to 1/64 relative to the original labelfree library based on the number of peptides. We tested these new libraries for target-decoy SLS using the Random decoy method in triplicates with a test data set of 77 953 label-free spectra as input (Table 1). At 1% FDR, the results were highly reproducible for all library sizes tested (the variation in number of IDs was less than 3%). We conclude that for the number of search spectra, decoy libraries are significantly larger than required to produce reproducible results. We also note that Random libraries have an advantage over Reverse libraries in that there is no limit of their potential size. Each Random version will contain different sequences. In principle, making these libraries large enough, they could be even employed to estimate the false positive potential of an individual peptide spectrum.

Comparing Spectral Library and Sequence Database Searching

The proposed method for FDR measurement for high mass accuracy SLS enables the direct integration of SLS and SDS results and provides a basis for comparing differences in results. To make this comparison, a set of 77 953 label-free query spectra was used (Table 1) for SLS using MSPepSearch, and SDS using Sequest coupled to Percolator,¹⁹ which is a widely used postprocessing approach. At 1% FDR, MSPepSearch identified 35 095 PSMs and Sequest found 32 696 PSMs or 7% fewer. Some 22% (7357 PSMs) of the latter IDs were not found by MSPepSearch. Of these, 90% were not represented in the library. If these were added, SLS would have made approximately 28% more IDs than SDS. We also compared the performance at both the 1% and 0.1% FDR level and found that at the higher level of confidence the relative performance of SLS improved further. To illustrate, Fig. 7 shows a scatter plot of library vs. Sequest scores for 31 464 common peptide ions identified by both search programs as rank 1 hits without any cutoffs. While Sequest scores were evenly distributed across the whole range, library scores were more concentrated in the upper score region. At 1% FDR, SDS identified only 139 (0.4% of 31 464) additional peptide ions, while SLS identified 5587 (18% of 31 464) more peptide ions than SDS. A decrease in the FDR level from 1% to 0.1% reduced the number of PSMs for SLS significantly less than for SDS (964 vs. 2229, respectively), showing that SLS on average identifies peptides more confidently than SDS. The percentage of SLS-identifiedonly further increased to 23% at 0.1% FDR level. The situation is analogous to a scenario where the same amount of sample is injected into two mass spectrometers but a higher signal is observed in the instrument with a higher sensitivity. This observation demonstrates the additional advantage of SLS in validation studies where both high sensitivity and low error rate are required. Given the known variations in performance with search settings, spectra, and versions, we note that these results should not be taken as a true measure of the relative performance of SLS and SDS methods. We simply report our findings for one particular case to show how decoy libraries enable such comparisons. They do show that SLS can significantly increase the confidence of identification due, no doubt, to its smaller search space and use of product ion intensities leading to increased selectivity.

Hybrid Mass Spectral Library Search with Decoy Spectral Libraries

The present methods are applicable to the recently reported 'Hybrid' spectral search method for finding unexpected modifications.¹⁶ Similar to ultra-wide precursor tolerance search in SDS, hybrid SLS is an alternative search method allows the matching of both ions containing and not containing a modification present in only one of the peptides being compared. This search

method is available in NIST MSPepSearch and NIST MS Search (peptide.nist.gov). It can identify spectra for which the peptide ion is not in the spectral library but for which a peptide that differs by a single modification is in the library. Consequently, multiple high-scoring matches can potentially be made for a single query spectrum – a basic difference between this search and the traditional SLS. A representative fraction from the PRIDE test data set (Table 1) containing 43 284 spectra was chosen for a hybrid SLS against the label-free target spectral library and either the corresponding Random (Fix C) or Reverse decoy spectral library. The resulting score thresholds corresponding to 1% FDR were 600 and 570 for the Reverse and Random decoy libraries, respectively. The lower score threshold observed for the Random decoy library corresponds to a gain of 11% of target library identifications compared to the Reverse decoy library, significantly greater than found for conventional searches, suggesting the hybrid search may be more influenced by sequence 'homologies' than the conventional search. The higher threshold scores presumably result from the fact that hybrid search permits both direct and loss peaks to match, resulting in more matching peaks and therefore higher scores.

Conclusions

The methods reported here for determining FDR values for high mass accuracy mass spectral library searching enable results of library searches to be directly integrated with those of more widely used sequence searching methods. The availability of a reliable means for estimating these values should encourage the use of SLS, which offers advantages in speed and sensitivity. The reliability of these methods was tested using synthetic peptide spectra available from a recent publication.¹ These tests show that the Reverse and Random decoy library construction methods are both effective for FDR estimation, and results are consistent with the entirely unrelated method of Peptide-Prophet.¹⁴ We also showed that the Reverse method yields slightly higher score thresholds than the Random method presumably due to fewer homology matches in the latter. However, differences in these methods are rather small, leading to threshold scores at 1% FDR for the Reverse method to be approximately 50 higher than for the Random method, which we find causes a slight reduction, typically 2%, in reported IDs. We also demonstrated that decoy libraries are effective in computing FDRs in the recently reported hybrid SLS method ¹⁶ for identifying unexpected modifications. We point out the recently reported spectral libraries of synthesized¹ peptides, which were used for testing, will also provide a boost in library coverage. However, at its current level of completeness, it is shown that SLS can yield more peptide identifications than SDS, especially at the highest levels of confidence. Furthermore, SLS performance would be expected to be further enhanced by the application of postprocessing methods, such as Percolator,¹⁹ which is commonly used to significantly increase SDS identification rates.

SUPPORTING INFORMATION:

Table S1. Comparison of the Sequence Uniqueness Index (SUI) in Label-free Target or Reverse Decoy Libraries Relative to Random Library

Table S2. Amino Acid Distribution in non-C-terminal Residues and in C-terminal Residues of a Target Library (Human Label Free) and a Decoy Library (by Random Method)

Table S3. Comparison of the Precursor Mass Distributions at Four Tolerances of Three Random Decoy Libraries with Their Target Label-free Library

ACKNOWLEDGMENTS

We acknowledge support from the NIH/NCI CPTAC program through an Interagency Agreement, ACO15005, with NIST.

NIST COMMERCIAL DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

(1) Zolg D. P.; Wilhelm M.; Schnatbaum K.; Zerweck J.; Knaute T.; Delanghe B.; Bailey D. J.; Gessulat S.; Ehrlich H.; Weininger M.; Yu P.; Schlegl J.; Kramer K.; Schmidt T.; Kusebauch U.; Deutsch E. W.; Aebersold R.; Moritz R. L.; Wenschuh H.; Moehring T.; Aiche S.; Huhmer A.; Reimer U.; Kuster B. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, 14, 259-262.

(2) Eng J. K.; McCormack, A. L.; Yates J. R. 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) Perkins D. N.; Pappin D. J.; Creasy D. M.; Cottrell J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, 20, 3551-3567.

(4) Hernandez P.; Muller M.; Appel R.D. Automated protein identification by tandem mass spectrometry: issues and strategies *Mass Spectrom Rev.* **2006**, 25, 235-254.

(5) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **2008**, *5*, 873-875.

(6) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **2011**, 11, 1075-1085.

(7) Shao, W.; Lam, H. Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass Spectrom Rev.* **2016**, 36, 634-648.

(8) Zhang, Z; Yang, X; Mirokhin, Y.A.; Tchekhovskoi, D.V.; Ji, W.; Markey, S.P.; Roth, J.; Neta, P.; Hizal, D.B.; Bowen, M.A.; Stein, S.E. Interconversion of Peptide Mass Spectral Libraries Derivatized with iTRAQ or TMT Labels. *J Proteome Res.* **2016**,15, 3180-3187.

(9) Marx, H.; Lemeer, S.; Schliep, J.E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A.J.; Kuster, B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol.* **2013**, 31, 557-564.

(10) Elias, J.E.; Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, 4, 207-214.

(11) Wang, G.; Wu, W.W.; Zhang, Z.; Masilamani, S.; Shen, R.F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* **2009**, 81, 146-159.

(12) Cheng, C.Y.; Tsai, C.F.; Chen, Y.J.; Sung, T.Y.; Hsu, W.L. Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications. *J. Proteome Res.* **2013**, 12, 2305-2310.

(13) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome Res.* **2010**, 9, 605-610.

(14) Keller, A.; Nesvizhskii, A. I.; 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.

(15) Shao, W.; Zhu, K.; Lam, H. Refining similarity scoring to enable decoy-free validation in spectral library searching. *Proteomics*. **2013**, 13, 3273-3283.

(16) Burke, M.C.; Mirokhin, Y.A.; Tchekhovskoi, D.V.; Markey, S.P.; Heidbrink Thompson, J.; Larkin, C.; Stein, S.E. The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. *J Proteome Res.* **2017**, 16, 1924-1935.

(17) The NIST Libraries of Peptide Tandem Mass Spectra. http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload.

(18) Sinha, A.; Ignatchenko, V.; Ignatchenko, A.; Mejia-Guerrero, S.; Kislinger, T. In-depth proteomic analyses of ovarian cancer cell line exosomes reveal differential enrichment of functional categories compared to the NCI 60 proteome. *Biochem Biophys Res Commun.* **2014**, 445, 694-701.

(19) Kall L.; Canterbury J.D.; Weston J.; Noble W.S.; MacCoss M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. **2007**, 4, 923-925.

(20) Ma, K.; Vitek, O.; Nesvizhskii, A.I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*. **2012**, 13, S1.

(21) Ahrné, E.; Ohta, Y.; Nikitin, F.; Scherl, A.; Lisacek, F.; Müller, M. An improved method for the construction of decoy peptide MS/MS spectra suitable for the accurate estimation of false discovery rates. *Proteomics.* **2011**, 11, 4085-4095.

(22) Horlacher, O.; Lisacek, F.; Müller, M. Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries. *J. Proteome Res.* **2016**, 15, 721–731.

(23) Heffner, K.M.; Hizal, D.B.; Yerganian, G.S.; Kumar, A.; Can, Ö.; O'Meally, R.; Cole, R.; Chaerkady, R.; Wu, H.; Bowen, M.A.; Betenbaugh, M.J. Lessons from the Hamster: Cricetulus griseus Tissue and CHO Cell Line Proteome Comparison. *J Proteome Res.* **2017**, 16, 3672-3687.

Туре	Tag	Instrument	Spectra	Source
Exp ^a	Label-	Q Exactive	605 113	PRIDE
	free			(PXD000695)
Exp ^a	Label-	Fusion	77 953	NIST
	free	Lumos		
Exp ^a	iTRAQ	Q Exactive	994 133	CPTAC
	4-plex			
Lib ^b	Label-	Q Exactive,	1 127 970	peptide.nist.gov
	free	LTQ		(Sep 23, 2016)
		Orbitrap		
		Velos		
Lib ^b	iTRAQ	Q Exactive,	1 201 632	peptide.nist.gov
	4-plex	LTQ		(Nov 26, 2014)
	-	Orbitrap		(····································
		Velos		

Table 1. Description of Human Test Data Sets and Libraries (All derived from digestion with either Trypsin alone or with Lys-C and fragmented by beam-type collision-cell)

^aExperimental data set. ^bSpectral library.

	Query Not in Library					Query in Library	
SUI	Target	Rev	RanC1	RanC2	RanC3	Rev	RanC
0.1-0.2	0	0	0	0	0	1	0
0.2-0.3	0	0	0	0	0	1	0
0.3-0.4	3	1	0	1	0	7	1
0.4-0.5	23	14	6	7	6	18	5
0.5-0.6	112	61	46	43	40	111	64
0.6-0.7	175	97	63	59	80	181	111
0.7-0.8	120	86	65	71	61	115	85
0.8-0.9	68	41	27	26	29	62	41
0.9-1	16	15	7	7	8	8	14
Total	517	315	214	214	224	504	320

 Table 2. Numbers of False Positives Above Score 250 at Various SUI Levels for Reverse and Random Methods in Target-Decoy SLS^a with Synthetic Peptide Search Spectra

^aEither 14 997 synthetic peptide spectra not in library or 19 343 synthetic peptide spectra in library as query spectra for searching against the target/Reverse/RanC library. RanC1, 2, and 3 were three different versions of Random decoy libraries.

FIGURE LEGENDS

Figure 1. Examples of Decoy Spectra Created by the Reverse and Random Methods. (A) The Reversed Method with FixC. A spectrum of the peptide ion AYTDELVELHR/2+ (upper panel, in red) has been reversed to create a decoy spectrum, HLEVLEDTYAR/2+ (lower panel, in blue). (B) The Random Method with RanC. A spectrum of peptide ion AYVLLGDDSFLER/2+ (upper panel, in red) has had both its non-C and C terminal residues randomly replaced to create a decoy spectrum of MFHAAHYTLNDIK/2+ (lower panel, in blue). The m/z of fragment ions are shifted accordingly with intensity unchanged. For clarity, only major peaks are labeled. Sequence and charge state are also labeled.

Figure 2. Decoy Fraction vs. Rank. (A) The decoy fraction of top 10 identifications of 110 374 HCD spectra from a proteomic data set.²³ 110 374 query spectra searched against concatenated target-decoy human HCD libraries (Reverse in yellow circle; Random in orange circle; Precursor-Swap in gray circle, and linked by solid lines). * Also shown are adapted results of Elias et al.¹⁰ using an SDS engine (Sequest) for a sequence reversal method (blue circle linked by dashed line). (B) The fraction of decoy hits for rank up to 10 of 14 997 not-in-library synthetic HCD peptide spectra¹. 14 997 query spectra searched against concatenated target-decoy human HCD libraries (Reverse in yellow circle; Random in orange circle; Precursor-Swap in gray circle, for clarity ranks are linked by solid lines). ** Also shown are adapted results of Lam et al.¹³ using a shuffle-and-reposition method based on low-resolution libraries and test data (blue circle linked by dashed line).

Figure 3. Example of histograms for precursor mass overlapping. Precursor distribution histogram in target label-free library (Target: orange square, linked by orange line) in the 500 m/z to 500.5 m/z range, with a mass tolerance of 20 ppm, vs. those of three decoy libraries generated by Random method (ranC1: gray triangle; ranC2: yellow cross; ranC3: blue circle, linked by blue line).

Figure 4. Testing Target, Reverse, Random, and Precursor-Swap Libraries with Synthetic Peptide Spectra. Library search score distribution histograms (Target in blue; Reverse {rev} in yellow; Random {ranC} in orange; Precursor-Swap {swap} in gray) generated either with (A) in-library-spectra: 19 343 synthetic peptide spectra or (B) not-in-library spectra: 14 997 synthetic peptide spectra as query spectra. MSPepSearch scores are shown on the x-axis (bin=10) and frequency counts are shown on the y-axis.

Figure 5. FDR vs. Score for four methods for label-free spectra. (A) Decoy-free PeptideProphet algorithm and three decoy methods: (B) Reverse; (C) Random FixC; (D) Random RanC. 605 113 label-free HCD query spectra were used to search against the NIST label-free HCD spectral library (1 127 970 spectra, 320 824 distinct peptide sequences, <u>peptide.nist.gov</u>)¹⁷ and their decoys. 1% FDR thresholds are marked by an arrow. MSPepSearch scores are shown on the x-axis and FDR values are shown on the y-axis in a log scale. Threshold scores and numbers of identifications (in parentheses) at the 1% FDR level are shown on each plot.

Figure 6. FDR vs. Score for four methods for iTRAQ 4-plex labeled spectra. (A) Decoy-free PeptideProphet algorithm and three decoy methods: (B) Reverse; (C) Random FixC; (D)

Random RanC. 994 133 iTRAQ 4-plex labeled HCD query spectra were used to search against the NIST iTRAQ 4-plex labeled HCD spectral library (1 201 632 spectra, 390 009 distinct peptide sequences, <u>peptide.nist.gov</u>)¹⁷ and their decoys. 1% FDR thresholds are marked by an arrow. MSPepSearch scores are shown on the x-axis and FDR values are shown on the y-axis in a log scale. Threshold scores and numbers of identifications (in parentheses) at the 1% FDR level are shown on each plot.

Figure 7. A comparison of spectral library search (NIST MSPepSearch) with sequence database search (Sequest) scores. The library search and Sequest search were done with the same set of 77 953 label-free query spectra. Only peptide ions contained in the library (31 464 peptide ions) were considered. Each blue dot represents a peptide ion. The x-axis represents Sequest scores and the y-axis represents library scores. 1% FDR (black solid line) and 0.1% FDR (dark red dash line) cutoff scores of SLS and SDS were determined individually. Total 1: # of PSMs identified by SLS at 1% FDR; Total 2: # of PSMs identified by SLS at 0.1% FDR; Total 3: # of PSMs identified by SDS at 1% FDR; Total 4: # of PSMs identified by SDS at 0.1% FDR. Total 1 – Total 2 = 964; Total 3 - Total 4 = 2229.



Figure 1. Examples of Decoy Spectra Created by the Reverse and Random Methods



(B) Random



m/z

Figure 2. Decoy Fraction vs. Rank

(A) With Test HCD Spectra



110 374 Test HCD Spectra

(B) With "Not in Library" Synthetic HCD Spectra





Figure 3. Example of histograms for precursor mass overlapping

Figure 4. Testing Target, Reverse, Random, and Precursor-Swap Libraries with Synthetic Peptide Spectra



(A) With "In Library" Spectra

(B) With "Not in Library" Spectra





Figure 5. Label-free Spectra Searched against Label-free Spectral Library



Figure 6. iTRAQ 4-plex Labeled Spectra Searched against iTRAQ 4-plex Spectral Library



Figure 7. Comparison of Spectral Library Search and Sequence Database Search

for TOC only



Table S1: Comparison of the Sequence Uniqueness Index (SUI) in Label-free Target or Reverse Decoy Libraries Relative to Random Library

SUI	target or rev (#)	target or rev (%)	ranC (#)	ranC (%)
0.1-0.2	309	0.03	0	0.00
0.2-0.3	2,648	0.23	82	0.01
0.3-0.4	18,081	1.60	6,209	0.55
0.4-0.5	69,948	6.20	48,505	4.30
0.5-0.6	199,466	17.68	182,337	16.17
0.6-0.7	286,281	25.38	296,658	26.30
0.7-0.8	262,299	23.25	279,767	24.80
0.8-0.9	196,527	17.42	215,072	19.07
0.9-1	92,411	8.19	99,340	8.81
total	1,127,970	100	1,127,970	100

AA*	Non-C-Terminal				C-Terminal			
	#_Target	%_Target	#_Decoy	%_Decoy	#_Target	%_Target	#_Decoy	%_Decoy
Α	1,220,319	8.03	1,215,639	8.00	701	0.06	732	0.06
С	3213	0.02	3240	0.02	2	0.00	1	0.00
C (CAM)	206,436	1.36	206,543	1.36	523	0.05	473	0.04
D	1,010,987	6.65	1,009,284	6.64	1070	0.09	1065	0.09
E	1,455,284	9.58	1,458,374	9.60	984	0.09	967	0.09
F	593,252	3.90	591,183	3.89	2020	0.18	1939	0.17
G	1,072,914	7.06	1,070,270	7.04	1167	0.10	1193	0.11
Н	417,912	2.75	415,223	2.73	6571	0.58	6382	0.57
I	788,279	5.19	789,448	5.19	786	0.07	896	0.08
К	281,458	1.85	282,586	1.86	587,092	52.05	585,270	51.89
L	1,595,704	10.50	1,596,433	10.51	1808	0.16	1626	0.14
М	210,485	1.39	211,891	1.39	424	0.04	453	0.04
M (0)	137,378	0.90	138,505	0.91	333	0.03	334	0.03
Ν	601,281	3.96	605,427	3.98	1516	0.13	1483	0.13
Р	1,015,065	6.68	1,014,626	6.68	314	0.03	275	0.02
Q	792,739	5.22	790,466	5.20	826	0.07	886	0.08
R	137,673	0.91	137,329	0.90	517,338	45.86	519,544	46.06
S	1,097,697	7.22	1,100,658	7.24	769	0.07	671	0.06
т	861,850	5.67	861,894	5.67	486	0.04	515	0.05
V	1,094,413	7.20	1,093,668	7.20	1474	0.13	1417	0.13
W	137,872	0.91	137,245	0.90	126	0.01	176	0.02
Υ	464,541	3.06	466,820	3.07	1640	0.15	1672	0.15
Total	15,196,752	100	15,196,752	100	1,127,970	100	1,127,970	100

Table S2: Amino Acid Distribution in non-C-terminal Residues and in C-terminal Residues of a Target Library (Human Label Free) and a Decoy Library (by Random Method)

*modifications were labeled as: C (CAM): Carbamidomethyl of C; M (O): Oxidation of M.

Table S3: Comparison of the Precursor Mass Distributions at Four Tolerances of Three Random Decoy Libraries with Their Target Label-free Library

<i>m/z</i> Tolerance	Target	ranC1	ranC2	ranC3
0 ppm	1,127,970	580,267 (51.4%)	579,543 (51.4%)	579,790 (51.4%)
5 ppm	1,127,970	1,082,937 (96.0%)	1,083,440 (96.1%)	1,082,918 (96.0%)
10 ppm	1,127,970	1,109,521 (98.4%)	1,109,543 (98.4%)	1,109,599 (98.4%)
20 ppm	1,127,970	1,121,455 (99.4%)	1,121,598 (99.4%)	1,121,486 (99.4%)