Spectrum-to-Spectrum Searching Using a Proteome-wide Spectral Library*[S]

Chia-Yu Yen‡, Stephane Houel‡§, Natalie G. Ahn‡§, and William M. Old‡¶

The unambiguous assignment of tandem mass spectra (MS/MS) to peptide sequences remains a key unsolved problem in proteomics. Spectral library search strategies have emerged as a promising alternative for peptide identification, in which MS/MS spectra are directly compared against a reference library of confidently assigned spectra. Two problems relate to library size. First, reference spectral libraries are limited to rediscovery of previously identified peptides and are not applicable to new peptides, because of their incomplete coverage of the human proteome. Second, problems arise when searching a spectral library the size of the entire human proteome. We observed that traditional dot product scoring methods do not scale well with spectral library size, showing reduction in sensitivity when library size is increased. We show that this problem can be addressed by optimizing scoring metrics for spectrum-to-spectrum searches with large spectral libraries. MS/MS spectra for the 1.3 million predicted tryptic peptides in the human proteome are simulated using a kinetic fragmentation model (MassAnalyzer version2.1) to create a proteome-wide simulated spectral library. Searches of the simulated library increase MS/MS assignments by 24% compared with Mascot, when using probabilistic and rank based scoring methods. The proteome-wide coverage of the simulated library leads to 11% increase in unique peptide assignments, compared with parallel searches of a reference spectral library. Further improvement is attained when reference spectra and simulated spectra are combined into a hybrid spectral library, yielding 52% increased MS/MS assignments compared with Mascot searches. Our study demonstrates the advantages of using probabilistic and rank based scores to improve performance of spectrum-to-spectrum search strategies. ***Molecular & Cellular Proteomics 10: 10.1074/mcp.M111.007666, 1–15, 2011.***

High-resolution hybrid mass spectrometers and improved methods for sample preparation and chromatography have enabled routine quantitative profiling of thousands of proteins in a single sample. Typically, profiling of complex biological samples is performed by "bottom up" proteomics, where proteins are proteolyzed and peptides separated by one or more dimensions of chromatography before mass spectrometry analysis. Peptide ions are isolated and dissociated in the gas phase to yield tandem mass spectra (MS/MS)[1], which are interpreted by algorithms to identify the fragmented peptides present in the sample. "Spectrum-to-sequence" or sequence-based approaches are commonly employed for MS/MS identification, for example using database search algorithms that match MS/MS spectra to sequences in a protein database, by first generating theoretical fragmentation spectra for different peptide sequences, and then scoring overlap between model and experimental spectra. In general, the models for generating theoretical MS/MS spectra use simple fragmentation rules that consider all backbone fragmentation events as equally likely. This ignores well-known residue-specific effects on backbone cleavage, which contribute to the variable fragment intensities characteristic of different peptide sequences (1). Consequently, the scoring functions used by many search algorithms show poor discrimination in separating valid peptide sequences from incorrect or false positive assignments. Previous studies showed that discrimination of sequence-based approaches is improved by using a kinetic fragmentation model to evaluate the chemical plausibility of MS/MS assignments (2–4). "Spectrum-to-spectrum" or spectral library searching, is an alternative to sequenced based approaches, where experimental MS/MS are directly matched against a library of previously identified reference spectra, assembled from MS/MS assigned to peptides with high confidence (5, 6). Spectral library searching has two advantages over sequence-based approaches. First, because the number of spectra in reference libraries normally employed is small compared with the number of database peptide sequences, search times are significantly reduced. Second, true peptide MS/MS assignments are more easily distinguished from false positive assignments with the added fragment intensity information. Most spectral library search

From the ‡Department of Chemistry and Biochemistry, §Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309

Author's Choice—Final version full access.

Received January 6, 2011, and in revised form, April 26, 2011

Published, MCP Papers in Press, April 30, 2011, DOI 10.1074/mcp.M111.007666

[1] The abbreviations used are: MS/MS, tandem MS or fragmentation spectrum; FDR, false discovery rate determined by the *q*-value; SS, simulated spectrum/spectra; DP, Dot product; RDP, ranked DP; SIM, similarity score; RSIM, ranked SIM; SHP, hypergeometric probability score by single-candidate consideration; MHP, hypergeometric probability score by multi-candidate consideration; UPS1, the universal protein sample data set generated by the Sigma UPS1 sample.

algorithms use a dot product score to measure the similarity in fragment intensity patterns between the candidate MS/MS and library spectra (5, 7). This is because dot product scores have shown better performance compared with other scores for searching small molecule spectral libraries (8). In contrast, scores used in spectrum-to-sequence search tools, such as Mascot (9), OMSSA (10), and MyriMatch (11), are often based on probabilistic functions that match fragment ion masses, largely ignoring fragment intensity information in observed MS/MS. By matching peak intensities, spectrum-to-spectrum methods fully exploit the intensity patterns unique to different peptides, including those of noncanonical fragment ions that are not predicted by simple fragmentation models often used in sequence-based searching. As a result, spectral library searching can show increased score discrimination and sensitivity over sequence-based methods (7).

The compact sizes of reference libraries, although providing advantages of search speed and discrimination, are far from comprehensive in their coverage over human proteins. In many human tissues and cancers, much of the proteome and its associated complement of modifications remain undiscovered; consequently, they are incompletely represented in reference libraries. For example, the "Human IT Library" from the National Institute of Standards and Technology (NIST) covers only 21% of amino acids in the human proteome (12). Accordingly, spectrum-to-spectrum search methods are limited to rediscovery of previously identified sequences, such as in targeted proteomics applications. For this reason, database search algorithms remain the primary peptide identification tool employed for new protein discovery.

We recently reported a method for addressing the limited coverage of reference libraries, which uses a kinetic gas phase peptide fragmentation model (3) to create a library of simulated MS/MS spectra for all predicted peptides in the human proteome (13). In this way, a "proteome-wide library" can be searched using spectrum-to-spectrum search software in the same manner used for smaller reference libraries, maintaining the advantages of direct intensity comparisons, but extending the search to larger numbers of peptides typically covered only by database search algorithms. However, we observed lower performances of spectrum-to-spectrum searching against proteome-wide libraries compared with conventional sequence-based tools such as Mascot.

Here we present a new strategy for searching proteome-wide spectral libraries, comprised of kinetically simulated MS/MS spectra, and incorporated into an efficient search application, Spec2spec. We evaluate the contributions of increased search space, proteome coverage and the quality or accuracy of spectral intensity predictions on discrimination performance in spectral library searching. We show that current limitations in spectral library search tools include the scoring functions, which are not optimized for proteome-wide libraries, because of larger search spaces representing more than 20-fold the number of peptides contained in reference libraries, and the quality of simulated spectra used to increase proteome coverage. Thus, although the high proteome coverage in simulated proteome-wide libraries increases the number of unique peptide identifications compared with reference libraries, the increased library size degrades performance of dot product and similarity scoring. To address this limitation, we present new scoring metrics, including probabilistic scores based on a hypergeometric model of random peak matching in library spectra, and dot product scores based on peak intensity rankings. The new scores enable proteome-wide library searching with more discriminatory power, outperforming sequence-based searching with Mascot. Furthermore, we evaluate the use of target-decoy search methods for estimating false discovery rates (FDR) in spectrum-to-spectrum searching. We identify score-dependent biases which lead to underestimated FDR with smaller reference libraries, compared with proteome-wide simulated libraries. These findings demonstrate the potential for replacing traditional spectrum-to-sequence searching with spectrum-to-spectrum searching against proteome-wide simulated libraries in discovery proteomics.

EXPERIMENTAL PROCEDURES

Data Collection—Liquid chromatography (LC)-LC-MS/MS was performed using a LTQ-Orbitrap mass spectrometer (Thermo Scientific) interfaced with a nanoAcquity ultra performance liquid chromatography (UPLC) (Waters, Milford, MA), operated in two-domensional fractionation mode. Peptide mixtures (5 μl, 0.2–20 μg) were first separated on a Xbridge BEH C18 column (5 cm $\times$ 300 μm i.d, 5 μm bead diameter with 150 Å pore size, Waters) using a step gradient of 2% for each fraction from 97% buffer A (20 mM ammonium formate, pH 10) to 21% buffer B (100% acetonitrile). Steps were loaded onto a trap column (Waters C18 Symmetry, 20 mm $\times$ 180 μm i.d., 5 μm bead), washed and placed in line with a second dimension BEH C18 reversed-phase column (25 cm $\times$ 75 μm i.d., 1.7 μm bead, 100 Å pore size, Waters) before elution with a linear gradient from 95% buffer A (0.1% formic acid) to 40% buffer B (0.1% formic acid, 80% CH_3CN) in 120 min at a flow rate of 300 nL/min.

MS/MS were collected on the 10 most intense precursor ions, enabling monoisotopic precursor and charge selection settings, and excluding ions with unassigned charge state. Dynamic exclusion settings were: 30 s repeat duration, 180 s exclusion duration, 20 ppm exclusion width, and repeat count of 1. The maximum injection time for Orbitrap parent scans was 500 ms, allowing one microscan and AGC of 1×10^6. The maximal injection time for the LTQ MS/MS was 250 ms, with one microscan and automatic gain control (AGC) of 1×10^4. The normalized collision energy was 35%, with activation Q = 0.25 for 30 ms, and isolation width 2.0 Da.

Data Sets—The Sigma universal protein standard (UPS1, Sigma Aldrich) containing 48 purified human recombinant proteins present in equimolar ratios (14) was used as the defined protein mixture. Proteins were reduced with dithiothreitol and alkylated with iodoacetamide before overnight digestion with modified trypsin (Promega) at a 1:20 (w/w) trypsin to protein ratio. One picomole of this mixture was analyzed by two-dimensional-UPLC-MS/MS. Raw files were then extracted with extract_msn.exe (distributed with Bioworks 3.2), using the parameters -M1.4 -B85 -T4500 -S5 -G1 -I35 -C0.

For the purpose of evaluating bias and score distributions of target and decoy library searches, an *E. coli* consensus library (ver. 2009_05_21) was downloaded from the NIST website and converted

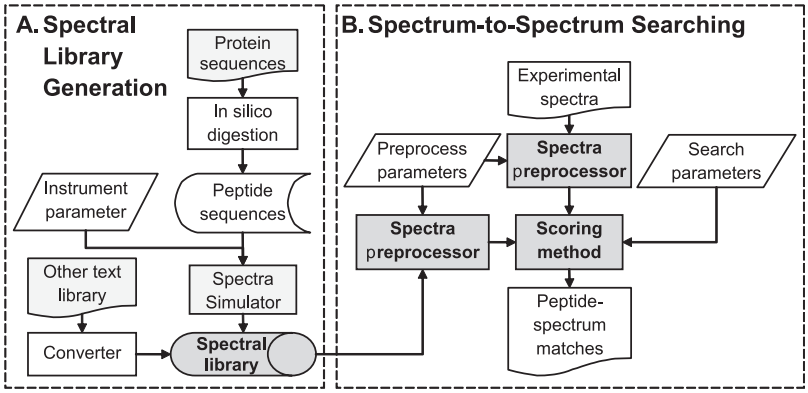

FIG. 1. **Spectrum-to-spectrum searching workflow.** *A*, Spectral library generation. Our in-house application takes a FASTA database and generates peptide sequences by *in-silico* digestion. These sequences are then simulated by MassAnalyzer, and the results are stored in an in-house text format. The approach allows generation of a large library covering most of the human proteome and a decoy library. Alternatively, the source of a library can be the text formatted NIST or X!Hunter library. An application converts these into our in-house text format. *B*, Spectrum-to-spectrum searching. The abstract view of a spectrum-to-spectrum searching application, which takes 3 major inputs: the experimental spectra in MGF format, a spectral library, and a set of search parameters. These are all taken into account in the scoring process.

TABLE I
Spectral libraries used in this study

	Unfiltered NIST	NIST[a]	TargetSS[b]	DecoySS[c]
Version	2/4/2009	–	IPI v3.27	IPI v3.27
Total spectra	249,896	154,612	4,050,732	3,963,009
UPS1 spectra[d]	12,041	4420	15,060	–
Unique sequences	155,218	110,960	1,343,602	1,314,334
UPS1 sequences[e]	6555	2442	4999	–
% UPS1 coverage[f]	59.5%	49.7%	70.7%	–

[a] Spectra are limited to those peptide ions represented in the TargetSS library.

[b] MassAnalyzer (v. 2.1) simulated spectra for peptides derived from in silico trypsin digest of the IPI human protein database version 3.27, with masses between 900 to 4,500 Da and peptide length at least nine amino acids. Charge states up to +3 are simulated.

[c] Simulated spectra for peptides derived from in silico trypsin digest of a reversed version of the IPI 3.27 sequence database. The same criteria as (b) are used to filter peptide sequences.

[d] Spectra corresponding to peptides from one of the 48 Sigma UPS1 standard proteins + known contaminants (22).

[e] Unique sequences matching to proteins in the UPS1 mixture, as in (d).

[f] The coverage is calculated by counting the number of covered amino acids in the whole protein database.

into MGF format (12). Spectra were filtered by the following criteria: not common to IPI human protein database version 3.27, cysteines modified with carbamidomethyl, charge state up to three, and peptides with nine or more amino acids. After filtering, 36,368 spectra remained.

Spec2spec: Software for Searching Proteome-wide Spectral Libraries—Spectrum-to-spectrum search applications typically consist of three main components: a spectral preprocessor, which includes ion filtering and intensity scaling, a spectral library and the scoring method (Fig. 1). Current software, such as X!Hunter (15), BiblioSpec (16), or SpectraST (7), do not allow optimization of each of these components independently. For example, X!Hunter can efficiently search large libraries, but at the expense of reduced ion representation in library MS/MS spectra. To address this need, we designed a cross-platform spectral library search application, Spec2spec, written in Java with a flexible object-oriented architecture to allow independent optimization of each component. In this architecture, spectral filters and scoring methods are predefined as abstract classes, which simplify the development and testing of new filters and scoring methods. To enable efficient searches of large simulated libraries, we prefiltered and partitioned the libraries by *m/z* and charge, and searched the partitions in multiple threads. This sacrificed the flexibility to customize filtering methods, but significantly reduced the loading time to an average of 1 min per library partition (13). The search times for Spec2spec were on the same order as those for sequence algorithm searching; searches of the UPS1 database required 26 min. on average whereas Mascot required 21 min. The overall workflow for spectral library generation and spectrum-to-spectrum searching is shown in Fig. 1.

Construction and Filtering of Spectral Libraries—A human simulated proteome-wide library ("TargetSS") was constructed as described previously (13) (Fig. 1*A*, Table I). Peptide sequences were generated by *in silico* tryptic digest of the IPI human protein database version 3.27 (17), including peptides with up to two missed cleavages, parent masses between 900–4500 Da and nine or more amino acids. Peptides corresponding to unlikely missed cleavage products were removed (18). A dynamic-link library version of MassAnalyzer (version 2.1) was used to simulate spectra in batch mode for those peptides with up to three charge states, using the parameters: instrument LTQ, collision energy 35%, activation time 30 ms, isolation window 2 Da, and resolution 800 at 400 *m/z*. A simulated decoy library ("DecoySS") was generated following the same methods, except that protein sequences were reversed before *in silico* proteolysis (Table I). An in-house application was then used to gather forward and reversed

simulated MS/MS into spectral libraries using a custom text format. In addition, an in-house application was written to convert between NIST-MSP, X!Hunter and our own text formats.

The NIST ion trap human reference library, build Feb 4, 2009 (12), was downloaded and filtered to remove spectra of nontryptic peptides, peptides less than nine amino acids, and charge state greater than three (Table I). To normalize comparisons between the NIST and other libraries, spectra corresponding to modified peptides, excepting carbamidomethylated cysteine containing peptides, were removed. Two spectra from this version of the NIST reference library corresponded to "standard protein peptides" (proteins in the Sigma UPS1 sample), but were found to be misannotated. Therefore, identifications assigned to these two spectra were labeled as true hits (Supplementary Methods).

To test the effects of search space, proteome coverage, and spectral quality, three more libraries were constructed by concatenating or merging libraries described above in different combinations (illustrated in supplemental Fig. S6). A "NIST+DecoySS" library was constructed by concatenating NIST and DecoySS libraries. A "SSNIST+DecoySS" library was constructed by simulating MS/MS for peptides in the NIST reference library, and concatenating these spectra with the DecoySS library. A "Hybrid" library was generated by merging TargetSS and NIST libraries, and replacing TargetSS spectra with corresponding reference spectra from the NIST reference library. Therefore, the Hybrid library is exactly the same size as the TargetSS library.

Scoring Methods

Dot Product Scoring Metrics—The dot product (DP) score treats each spectrum as a vector of the ordered peak intensities and measures the cosine of the angle between the spectra (8). Ions in two spectra are aligned and matched with a specified fragment ion tolerance. When multiple candidate ions are within the tolerance range, the peak with the highest value of the observed intensity divided by the difference between observed and predicted *m/z* is chosen for matching.

$$DP = \frac{\sum I_{obs} \times I_{sim}}{\sqrt{\sum I_{obs}^2 \times \sum I_{sim}^2}} \quad \text{(Eq. 1)}$$

In Equation 1, I_{obs} and I_{sim} are the intensities of observed and simulated spectra, respectively. The similarity score (SIM) is related to DP but places greater weight on lower intensity ions (3):

$$SIM = \frac{\sum \sqrt{I_{obs} \times I_{sim}}}{\sqrt{\sum I_{obs} \times \sum I_{sim}}} \quad \text{(Eq. 2)}$$

Calculating DP for two spectra with square root transformed intensities is mathematically equivalent to SIM. The square-root transformation used by SIM has been shown to provide higher discrimination in reference library searches (5, 8).

Ranked DP and SIM—We developed new scores based on DP and SIM, which use peak intensity ranks in place of actual intensities. In these equations, rank one is assigned to the peak with least intensity whereas the highest rank is assigned to the peak with most intensity. When a peak in the first spectrum does not have a corresponding peak with matched m/z in the second spectrum, it is matched to a peak with rank zero in the second spectrum. The resulting ranked DP and SIM scores are:

$$RDP = \frac{\sum R_{obs} \times R_{sim}}{\sqrt{\sum R_{obs}^2 \times \sum R_{sim}^2}} \quad \text{(Eq. 3)}$$

$$RSIM = \frac{\sum \sqrt{R_{obs} \times R_{sim}}}{\sqrt{\sum R_{obs} \times \sum R_{sim}}} \quad \text{(Eq. 4)}$$

where R_{obs} and R_{sim} are the intensity-based ranks of fragment ions in the observed and simulated spectra, respectively.

Hypergeometric Probability Scores—We also developed probabilistic scores using a hypergeometric distribution to model the frequency of random matching of fragment ions between experimental and library spectra. In spectrum-to-sequence searching, a hypergeometric probability distribution closely approximates the frequency of randomly matching MS/MS fragments to those predicted from a sequence database (19), and scoring functions based on this model have shown higher performance than other probabilistic methods in database searching (19, 20). Probabilistic scores typically consider only the *m/z* for fragment ion matches and ignore peak intensity. Therefore, we developed a scoring function where peaks from the library and experimental spectra are prefiltered by intensity, before matching and probability calculations.

The hypergeometric probability score by multi-candidate consideration (MHP) uses a hypergeometric distribution to model the frequency of random matches between fragment ions in an experimental spectrum and the set of all fragment ions found in library spectra within a certain precursor mass tolerance:

$$MHP = -\ln\left[\frac{\binom{K}{K_1}\binom{N-K}{N_1-K_1}}{\binom{N}{N_1}}\right] \quad \text{(Eq. 5)}$$

The terms in parentheses are binomial coefficients. N represents the number of all fragment ions from library spectra with precursor masses that fall within tolerance of the precursor mass of the experimental spectrum, *i.e.* from all candidate library spectra. K represents the number of N peaks that match ions in the experimental spectrum within tolerance. N_1 is the number of fragment ions in a candidate library spectrum, and K_1 is the number of N_1 peaks that match ions in the experimental MS/MS. Natural logarithms of the binomial coefficients are used to simplify the calculation of the final score (11).

MHP is adapted from a hypergeometric score described by Sadygov *et al.* (19), which was used to model random matching to predicted fragment ions in a sequence database, rather than a spectral library. By considering random matches to the global background of all candidate fragment ions in a spectral library, MHP should correct for mass and size dependent biases that arise with other scores, such as Sequest's XCorr (21). Consistently, the hypergeometric score described for spectrum-to-sequence searching was shown to be largely independent of peptide charge state and thus peptide mass (19).

The SHP score considers matches between experimental and candidate library spectra, without considering background matches within the library.

$$SHP = -\ln\left[\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}\right] \quad \text{(Eq. 6)}$$

The experimental spectrum is first divided into 1-*m/z* bins. In this equation, N represents the total number of these bins between the lowest m/z peak and the highest *m/z* peak, m represents the number of ions in the experimental spectrum, k represents the number of ions in the experimental spectrum which match the library spectrum, and n represents the number of ions in the library spectrum. The hypergeometric probability score by single-candidate consideration (SHP) was adapted from a hypergeometric score described by Tabb

et al. (11), except that SHP uses a univariate, rather than a multivariate, hypergeometric distribution and library spectra are used in place of predicted fragment *m/z* ladders from a protein sequence database.

Performance Assessment—To evaluate and compare search discrimination using different scores and libraries, we used a sample of known composition (Sigma UPS1) containing 48 purified and 103 contaminating proteins (22). MS/MS assignments to peptides from known proteins were assumed true, while assignments to other proteins were assumed false. This allowed the FDR for searches to be calculated as (# of accepted false assignments) ÷ (# of all accepted assignments). We refer to this method of FDR calculation as the "protein standard FDR."

FDR can alternatively be estimated using a target-decoy library search, where a library of decoy spectra are generated by simulations based on a kinetic fragmentation model (13, 23). In concatenated library searches, the target library was concatenated with a decoy version of the target library. Decoy assignments were considered false and the FDR calculated as 2 × (# of accepted decoy library assignments) ÷ (# of all accepted assignments) (24). In separated searches, the target and decoy libraries were searched independently, with FDR = (# of accepted decoy library assignments) ÷ (# of accepted target library assignments). False discovery rates shown in receiver operating characteristic (ROC) curves and tables were calculated as *q*-values to avoid complications when multiple score thresholds yielded the same FDR, especially within the low FDR range (25, 26).

Filtering and Search Criteria—Ions in the experimental and library spectra were filtered before searching, using the following procedure. First, ions representing neutral loss events within the range of −50 to +5 *m/z* around the parent ion were removed. Second, each spectrum was divided into windows 100 m/z wide and the six most intense peaks from each window were selected (all other peaks were removed). The parent mass tolerance for searches was ±1.2 Da and the fragment ion tolerance was ±0.5 *m/z*.

RESULTS

Spectrum-to-spectrum search algorithms evaluate MS/MS assignments using scoring functions, whose discriminatory power is measured by the ability to distinguish true from false identifications. In a previous study, we showed that although the dot product scores (DP and SIM) yielded good discrimination when used for searching against libraries of previously observed spectra, their performance degraded when the search space was expanded by 10-fold to include 1.3 million tryptic peptides in the human proteome (13), simulating MS/MS spectra using a kinetic fragmentation model (3). The poor performance of DP and SIM scores motivated the development of metrics with higher discrimination for searching simulated proteome-wide libraries. To gain insight into the factors important for discrimination, and to provide a baseline against which to compare the performances of scores developed in this study, we first evaluated the performance of DP and SIM when used to search a smaller library comprised of observed reference spectra.

Dot Product Scores Show Poor Discrimination—DP and SIM were evaluated in searches of a NIST human reference library (Table I), which contains consensus reference MS/MS covering 17% of amino acids in the human proteome. MS/MS spectra collected by LC-MS/MS on proteins of known composition (Sigma UPS1 standard) were searched against the NIST reference library, and performance was evaluated using ROC plots using the protein standard FDR calculation (Fig. 2). Also shown are ROC plots for the spectrum-to-sequence

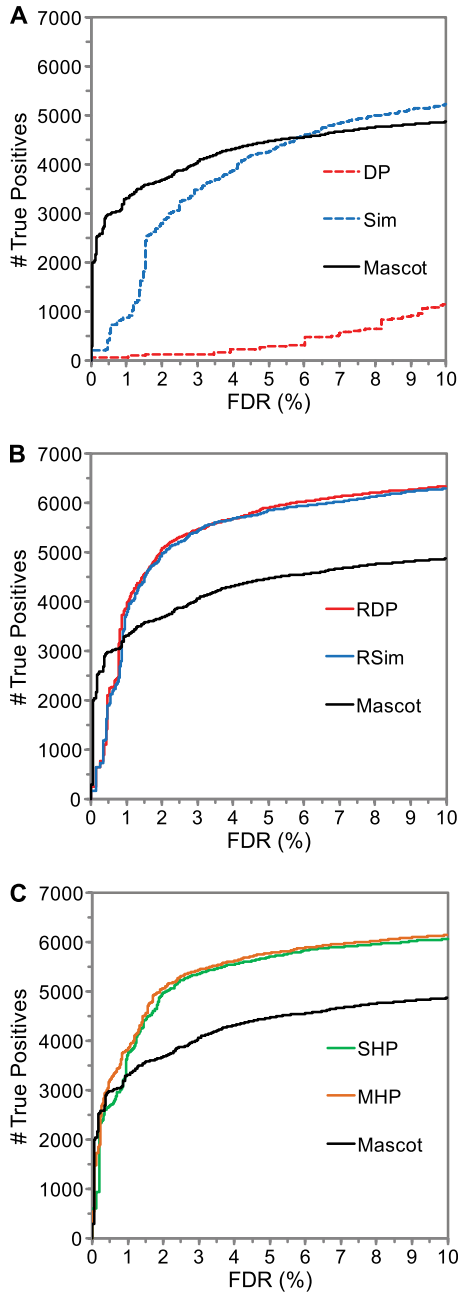

FIG. 2. **New scoring methods show improved discrimination over Mascot and dot product scores.** The UPS1 standard protein dataset was searched against the NIST reference spectral library to evaluate search discrimination for (*A*) DP and SIM, (*B*) RDP and RSIM, and (*C*) SHP and MHP. Performance of the Mascot ions score (black line) is shown using a peptide sequence database corresponding to those peptides represented in the NIST reference library. The correct hits are MS/MS spectra whose assignments match the peptides for protein in the Sigma UPS1 protein mixture (see Experimental Procedures).

search algorithm, Mascot, searched against a database with number of peptides equivalent to those in the NIST reference library. The dot product scores yielded poor discrimination, with DP showing ~fivefold lower sensitivity than SIM over a wide range of FDR, and Mascot identifying more true positives (TPs) than SIM at FDR < 1% (Fig. 2A). SIM and DP are closely related, because SIM is mathematically equivalent to DP calculated with peak intensities that have been square root transformed. The square root transformation used in SIM places greater weight on lower intensity peaks, allowing potentially informative backbone fragment ions to be included (5, 8, 27). This can be important for MS/MS spectra dominated by a few intense fragment ions. For example, peptides with strong N-terminal proline cleavages often generate other backbone fragment ions, which have low intensities but are important for assignments.

We sampled a number of high-scoring false assignments from a search of the NIST reference library using DP, and found numerous cases where false assignments were dominated by a few very intense ions in the spectra (supplemental Fig. S1). In each case, the DP score was elevated primarily by a small number of matches to high intensity fragments. We hypothesized that by placing more weight on matched peaks with lower intensity, SIM more effectively penalizes this class of false positive assignments, and that the dramatic difference in sensitivity between DP and SIM was because of their different scaling of peak intensity measurements. Based on these observations, we developed four new scoring metrics, which emphasize matching of lower intensity fragment ions and thus increase discrimination in proteome-wide library searches.

Rank-based Scores Improve Discrimination in Reference Library Searching—An alternative method to increase the relative weights for lower intensity peaks is to ignore intensity measurements and instead score based on intensity rankings of fragment ions. We thus modified DP and SIM to replace intensities with ranks assigned after sorting peaks in each spectrum by increasing intensity, resulting in calculations for ranked DP (RDP) and ranked SIM (RSIM) (*Experimental Procedures*). These rank based scores significantly improved sensitivity when searching against the NIST reference library, compared with DP and SIM (Figs. 2*A*, 2*B*). The performances of RDP and RSIM were comparable, and both outperformed Mascot with ~35% higher sensitivity at FDR = 2%.

Discrimination increases when search score distributions for true and false assignments are more completely separated. Consequently, increased discrimination occurs when (1) scores for false assignments are suppressed, and/or (2) scores for true assignments are increased. We compared true *versus* false score distributions using DP and RDP to determine whether the increase in discrimination with RDP was due to suppression of false assignments or enhancement of true assignments. Although scores for both true and false assignments decreased with RDP compared with DP, RDP scores for false assignments were suppressed to a greater degree than scores for true assignments (data not shown). Thus, RDP increases discrimination primarily by penalizing false matches.

Interestingly, at low FDR (< 0.8%), the Mascot ions score yielded higher sensitivity than either RDP or RSIM (Fig. 2*B*). To investigate this further, we manually examined spectra for the 10 highest scoring false positive matches from the RDP search, which account for approximately half of the false positives below 0.8% FDR. In 9 of the 10 cases, the experimental and library spectra showed a high degree of similarity, *i.e.* the spectra in each matched pair likely corresponded to the same peptide and the spectral match was valid. We hypothesized that these cases were labeled as false positives because of: (1) their correspondence to unknown protein contaminants in the Sigma UPS1 protein mixture, and/or (2) peptide sequence annotations for some NIST reference library spectra were incorrect (discussed in supplementary Methods). Thus, the lower sensitivity for RDP and RSIM at low FDR may be artifactual, because true spectral matches were counted as false. Nevertheless, the effect was complicated by the small number of cases with high scores. Above FDR = 1%, error estimates were more precise, and the rank-based scores showed significantly increased sensitivity over sequence-based Mascot ion scores.

Probabilistic Scores Perform as well as Rank-based Metrics for Reference Library Searching—Probability based scores are widely used in sequence-based searching (9, 10), but most algorithms score peaks matched by *m/z* without considering intensities. Spectral library searching, on the other hand, evaluates spectral matches primarily by intensity, placing less emphasis on peak matching in the *m/z* dimension. We extended probability based scoring to spectral library searching, in a way that evaluates matches in both *m/z* and intensity dimensions, potentially improving score discrimination. Two probability-based scores were developed, SHP and MHP, which used a hypergeometric probability distribution to model the random chance of matching observed to library fragment ions. Although peak intensities are not explicitly used in the probability calculation, they are used to select peaks for matching and scoring from both the experimental and candidate library spectrum. In this way, peaks of higher intensity are more likely to be selected from library spectra for matching and scoring against peaks selected from experimental spectra. We determined empirically the optimal number of peaks to select within each 100 Da window. For DP and SIM, the standard protein data set was searched against TargetSS with parent tolerance 1.2 Da, fragment tolerance 0.5 Da and selecting 3 to 25 peaks per 100 Da window in both experimental and library spectra, and the number of correct assignments at 5% FDR was compared (data not shown). We found that selecting the six most intense ions per 100 Da window gave the best discrimination. Peaks selected from the two spectra were

matched based on an *m/z* tolerance, and the numbers of matching and nonmatching peaks were used to calculate SHP and MHP (*Experimental Procedures*). SHP considers only the experimental MS/MS and the library spectrum being scored, and is thus library independent. In contrast, MHP incorporates a term for background matching of candidates in the library spectra, and is thus library dependent.

Performances of SHP and MHP were evaluated by searching the UPS1 data set against the NIST reference library, compared with Mascot searches of an equivalent search space (Fig. 2*C*). Both scores showed higher sensitivity than Mascot over a broad range of FDR (Fig. 2*C*). Moreover, MHP showed slight but consistently higher sensitivity than SHP, particularly at low FDR. In contrast, the two ranked scores showed greater discrimination than the probability scores above 1% FDR, which may reflect greater weighting of peak intensities by RDP and RSIM. Overall, each new score resulted in significantly higher discrimination compared with the dot product scores, DP and SIM, under all conditions, as well as improvement over Mascot above 1% FDR (Figs. 2*B*, 2*C*). Consistent with trends for RDP and RSIM, SHP and MHP histograms showed increased separation between true and false assignments compared with DP and SIM, primarily by lowering scores for false assignments (data not shown). These results demonstrated that rank- and probability-based scores provide a significant advantage over traditional dot product metrics for searching small reference libraries such as NIST, as well as significant improvements over the sequence-based probabilistic scoring algorithm used by Mascot.

Proteome-wide Library Searching with New Scores Improves Discrimination Over Mascot—We next tested the performance of rank- and probability-based scores in searching a simulated spectral library covering the human proteome. This addresses a limitation of spectrum-to-spectrum search methods, in which the size of the reference libraries restricts peptide assignments because of their low coverage of peptides in human proteins. We hypothesized that the increased coverage over human proteins in a proteome-wide library would increase the number of unique peptide identifications, and that the new scoring metrics would have increased discriminatory power over DP and SIM.

MassAnalyzer is based on an empirical kinetic model of gas-phase peptide fragmentation during collision-induced dissociation (CID) in quadrupole ion trap mass spectrometers, and predicts MS/MS spectra with reasonable accuracy for doubly and triply charged peptides up to 5000 Da (3, 4). This application (version 2.1) was used to generate a library of simulated MS/MS spectra corresponding to tryptic peptides in the human proteome filtered for mass, charge state, and sequence as described under "Experimental Procedures". We constructed a "TargetSS" library, covering >99% of proteins and 79% of amino acids in the International Protein Index human protein database (Table I). MS/MS spectra from the UPS1 dataset were searched against the TargetSS library,

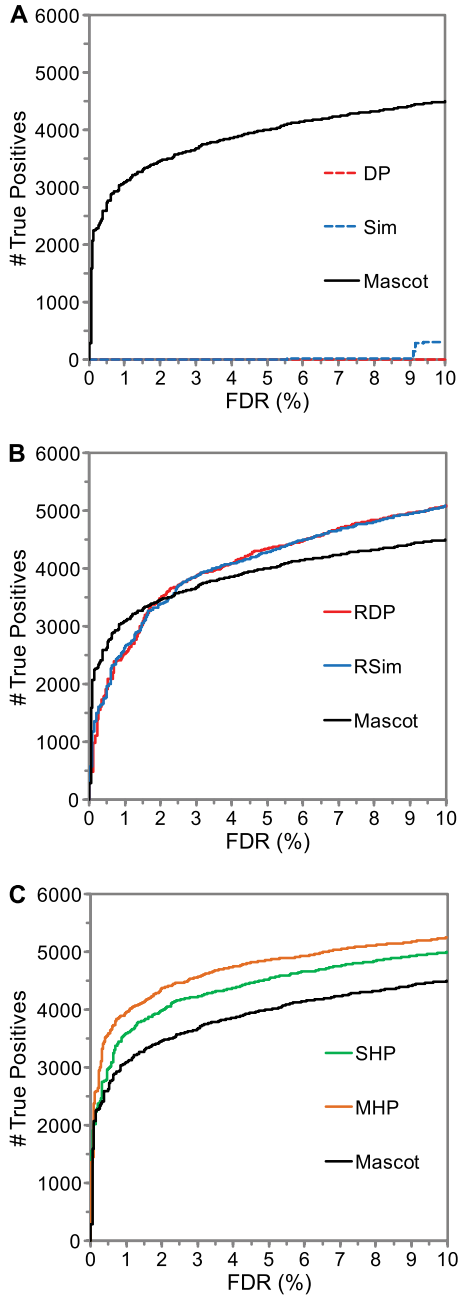

FIG. 3. **Proteome-wide simulated library searched using different scores.** (*A*) DP and SIM, (*B*) RDP and RSIM, and (*C*) SHP and MHP. The UPS1 standard protein dataset was searched with seven scoring methods against the TargetSS library, consisting of simulated spectra based on *in silico* digest of the human IPI 3.27 database. Performance is also shown for the Mascot ions score (black line) search against an equivalent peptide sequence database.

and the performances for each score as well as a Mascot search of the equivalent peptide database were compared by ROC analysis (Fig. 3).

As seen in NIST reference library searches, DP and SIM performed poorly in searches of the TargetSS library; below 10% FDR, DP identified only seven true positives and SIM

TABLE II

The number of identified MS/MS and unique sequences comparing four different spectral libraries using different scoring metrics. The Sigma UPS1 dataset was searched against each library using RDP, RSIM, SHP, and MHP scores. Thresholds correspond to 3% FDR, where FDR is calculated as the proportion of accepted assignments that fail to match known proteins in the UPS1 sample. Mascot was searched using peptide databases equivalent to each library, using the ions score for FDR threshold determination

Library	Correct MS/MS and unique sequence assignments at 3% FDR[a]				
	RSIM	RDP	SHP	MHP	Mascot
NIST[b]	5434 (651)	5457 (651)	5363 (647)	5451 (657)	4055 (617)
TargetSS[c]	3859 (678)	3867 (675)	4223 (715)	4557 (732)	3661 (656)
NIST+DecoySS[d]	4494 (611)	4628 (618)	4415 (614)	4626 (623)	3111 (545)
SSNIST+DecoySS[e]	2970 (534)	3033 (538)	3306 (561)	3425 (552)	3111 (545)
Hybrid[f]	5281 (743)	5387 (744)	5247 (753)	5583 (769)	3661 (656)

[a] The number of accepted MS/MS assignments in a search of the Sigma UPS1 standard, using a score threshold corresponding to 3% FDR. The number of unique sequences identified is indicated in parentheses.

[b] Filtered NIST human LTQ reference library, build Feb 4, 2009 (see Table I).

[c] TargetSS library (see Table I).

[d] Filtered NIST reference library concatenated with DecoySS library (see Table I).

[e] Spectra were simulated for peptides in NIST reference library (SSNIST) and concatenated with DecoySS library.

[f] The Hybrid library is the union of the NIST reference library and the TargetSS library, where simulated spectra in TargetSS library are replaced with NIST reference spectra only for peptide ions represented in both libraries. Library size is equal to that of TargetSS library.

identified fewer than 500 true positives (Fig. 3*A*). RDP and RSIM improved significantly over DP and SIM, and yielded higher sensitivity than Mascot above 2% FDR (Fig. 3*B*). SHP and MHP yielded the greatest discrimination in TargetSS searches, outperforming Mascot and the other scoring methods consistently over a wide range of FDR values. MHP showed 8% greater sensitivity over SHP (3% FDR, Fig. 3*C*), suggesting an advantage in considering library dependent variations for fragment ion matching. The increased discrimination observed with SHP and MHP over Mascot reflects the advantage of using intensity information contained within the simulated spectra. Indeed, when simulated spectra in the TargetSS library were manipulated to remove the relative intensity information, sensitivity decreased by fivefold with MHP and 1.6-fold with SHP at 3% FDR (supplemental Fig.2), both reduced below the sensitivity of Mascot in Fig. 3. Thus, the relative intensity information in simulated spectra significantly increased discrimination, and was manifested best using the probability scores, SHP and MHP.

Proteome-wide Library Searching Identifies More Unique Sequences but Fewer MS/MS—Searches of the TargetSS library yielded lower numbers of MS/MS assignments, compared with NIST reference library searches, using every score (compare Figs. 2 and 3). However, a different trend emerged when unique peptides were compared, where every score yielded more unique peptides in TargetSS library searches over NIST library searches. This is seen in Table II, where MHP assigned 16% fewer MS/MS (4557 *versus* 5451) but 11% more unique sequences (732 *versus* 657) in TargetSS *versus* NIST library searches. Similarly, RSIM assigned 29% fewer MS/MS (3859 *versus* 5434), but 4% more unique sequences (678 *versus* 651) in TargetSS *versus* NIST reference library searches. We hypothesized that the reason more unique peptides were identified despite lower sensitivity in proteome-wide TargetSS library searches was the higher coverage of tryptic peptides, allowing MS/MS assignments to peptides not represented in the NIST reference library. Indeed, 93% of the unique sequences identified only by the TargetSS library searches (3% FDR) were absent in the NIST reference library (supplemental Fig. S4). Thus, the increased coverage of the simulated library enables matching to peptide sequences not present in the NIST reference library, resulting in increased numbers of unique peptide identifications.

We next examined why TargetSS library searches were less sensitive than NIST library searches, with respect to numbers of MS/MS assignments. RDP and RSIM showed a more dramatic reduction in sensitivity for TargetSS *versus* NIST library searches, compared with SHP and MHP. The greater weight on peak intensities with RDP and RSIM suggests that the accuracy or quality of relative intensities in the simulated spectra might underlie decreased performance. The fragmentation of certain peptides might be modeled poorly by MassAnalyzer because of missing chemical mechanisms and oversimplifying assumptions, resulting in inaccurate relative intensities in simulated MS/MS spectra (28). Another important difference is the increased search space of the TargetSS library, which contains 26-fold more spectra than the NIST reference library. A larger search space might increase opportunities for false positive matches by random chance, requiring higher score thresholds and thus fewer accepted assignments. Similarly, sequence-based scoring algorithms show reduced discrimination when the search space is expanded (29). Thus, important differences between the NIST and TargetSS libraries that may affect discrimination include proteome coverage, search space size, and quality of simulated spectra (*i.e.* the accuracy of peak intensity simulations by MassAnalyzer). We next examined the contribution of each these parameters on score discrimination.

Searching Larger Spectral Libraries Reduces Discrimination—To assess the effect of increased search space on score

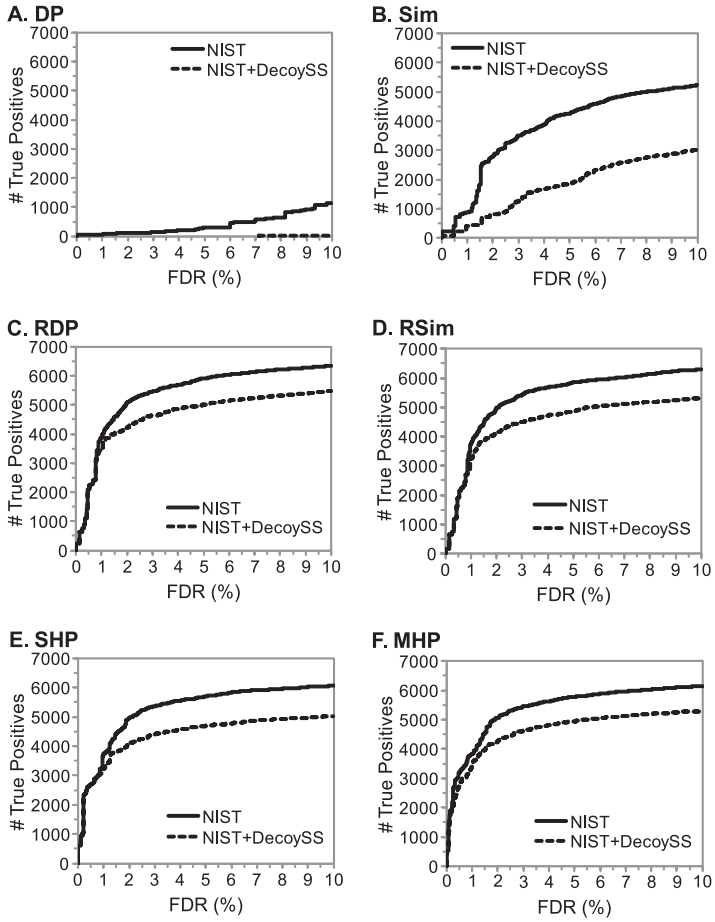

FIG. 4. **Rank- and probability-based scores are more resistant to effects of increased search space.** ROC plots for searches of the UPS1 dataset against the NIST human reference library and the same library concatenated with a human DecoySS library (see Table I). Performance for (*A*) DP, (*B*) SIM, (*C*) RDP, (*D*) RSIM, (*E*) SHP, and (*F*) MHP are shown. Increasing the search space with decoy spectra lowers discrimination performance, with DP and SIM showing the greatest effect.

discrimination in spectral library searching, we artificially expanded the size of the NIST reference library by 26-fold. MassAnalyzer was used to generate 3.96 million "decoy spectra" from reversed protein sequences in the human database (*Experimental Procedures*; Table I). These were concatenated with the NIST reference library to form a "NIST+DecoySS" library. In this way, the NIST reference library could be compared with a library with increased search space, while maintaining the characteristics of the NIST spectra. The UPS1 dataset was searched against the NIST and NIST+DecoySS libraries, and discrimination was compared using ROC analysis (Fig. 4). The protein standard FDR calculation was used to estimate specificity, where assignments to peptides from known proteins in the UPS1 sample were considered true and all other assignments, including those to decoy spectra, were considered false.

DP showed the worst performance, with sensitivity that precluded evaluation below 10% FDR, and SIM showed the largest performance decrease as a result of the increased search space introduced by decoy spectra (Figs. 4*A*, 4*B*). The new scoring methods showed less pronounced reductions in sensitivity with increased search space; MHP showed the smallest reduction (15.1%) and SHP showed the largest reduction (17.7%).

Further examination of RDP score distributions for true and false assignments from NIST *versus* NIST+DecoySS searches suggested two effects contributing to reduced discrimination with the larger library. First, the number of matches to false candidates made by random chance might be expected to increase with the larger spectral library. Second, MS/MS spectra might be correctly assigned in the NIST search, but assigned to higher scoring false spectra in the NIST+DecoySS search (29, 30); this effect, termed "distraction," both reduce the number of correct assignments while increasing the incorrect assignments. Both effects could raise score thresholds needed to maintain low FDR, thus reducing sensitivity. Random chance assignments were evident from the score distribution for false assignments from the NIST+DecoySS library search, which was dramatically shifted toward higher scores relative to false assignments for the NIST library search (supplemental Fig. S3). However, we found only three true assignments in the NIST search that were "distracted" to favor false assignments in the NIST+DecoySS search (above 3% FDR). Thus, the reduced sensitivity was mainly caused by increased false assignments with higher scores made by random chance, rather than the depletion of true assignments by distraction.

However, because the search space was artificially expanded with simulated spectra and not observed reference spectra, we cannot rule out that the "quality" of the simulated spectra comprising the appended decoy library contributed to the reduced sensitivity. A 26-fold expansion in search space using observed spectra would require nearly 4 million validated MS/MS spectra, which is clearly well beyond the size of any current reference library. To accommodate a smaller pool of reference spectra, while still allowing for a significant expansion in search space, we created a small target library by appending the NIST UPS1 reference library, containing 3,542 consensus reference spectra assembled from analyses of the UPS1 sample, and the NIST Mouse library, containing 156,075 spectra from *Mus musculus* samples. This target library was appended with either reference spectra from the NIST *Drosophila melanogaster* library, or corresponding simulated spectra. The resulting libraries allowed us to test the effect of a 70% increase in search space using observed and simulated decoy spectra (Suppl. Table I). At 5% FDR, sensitivity for SIM decreased 19% from 5,220 to 4,220 when the search space was expanded with observed spectra. When simulated spectra were used to increase the search space, we observed a 12% decrease in sensitivity. Thus, the increased search space resulted in reduced discrimination with either real or simulated spectra, with simulated decoy spectra showing a less pronounced effect.

Effects of Spectral Quality of Simulated Spectra on Search Discrimination—To examine the contribution of simulated spectral quality to performance, we constructed a library where the NIST reference spectra in the NIST+DecoySS library were replaced with spectra simulated by MassAnalyzer, corresponding to the same peptides ("SSNIST+DecoySS" library). The UPS1 data set was searched against both libraries, and the numbers of true assignments at 3% FDR were compared (Table II). Sensitivity was significantly lower for the SSNIST+DecoySS library, using any score. The greatest reduction in sensitivity was observed using RDP, which assigned 35% fewer MS/MS spectra in the SSNIST+DecoySS search compared with the NIST+DecoySS search (3033 *versus* 4628). Corresponding reductions in sensitivity were 34%, 25% and 26% respectively, for RSIM, SHP, and MHP.

ROC analyses for RDP and MHP searches revealed reduced sensitivity with SSNIST+Decoy compared with NIST+Decoy searches over a range of FDR values (Figs. 5*A*, 5*B*). The rank-based scores showed larger reductions in sensitivity than probability-based scores, when reference spectra were replaced by their simulated counterparts (Figs. 5*A*, 5*B*, compare solid and dashed green curves). Because RDP and RSIM place more emphasis on peak intensity differences than SHP and MHP, they might be expected to show more sensitivity to inaccurate predictions of relative intensities in simulated spectra. The same trend was seen when comparing searches using the TargetSS *versus* NIST libraries, where rank scores showed larger reductions in sensitivity for TargetSS searches than SHP and MHP. Score distributions for searches against NIST+DecoySS and SSNIST+DecoySS libraries were also compared with interrogate effects on scoring of true *versus* false assignments. Indeed, the distributions of true MS/MS assignments in the SSNIST+DecoySS search were shifted toward lower scores compared with the corresponding distribution for the NIST+DecoySS search (Figs. 5*C*, 5*E* and Figs. 5*D*, 5*F*), indicating that the experimental MS/MS spectra score lower against simulated spectra and are therefore not as well matched, compared with high-confidence reference spectra.

Increased Proteome Coverage Improves Discrimination—We next assessed the effect of increased proteome coverage on discrimination in spectral library searching. The SSNIST+DecoySS library, which concatenates the simulated NIST library and the DecoySS library, includes only 8.3% of peptide sequences contained within the simulated proteome-wide library, TargetSS. Because the two libraries have approximately equal numbers of spectra, the effect of coverage on discrimination can be measured while holding search space constant. The UPS1 dataset was searched against each library, and sensitivity was compared at 3% FDR (Table II, SSNIST+DecoySS *versus* TargetSS). Using each score, the sensitivity increased in searches against the TargetSS library compared with SSNIST+DecoySS, with respect to both MS/MS assignments and unique peptides. MHP showed the greatest increase in MS/MS assignments (from 3425 to 4557; +33%), similar to the increased numbers of unique peptides (from 552 to 732; +32.6%). In each case, the increased MS/MS and unique peptide assignments in TargetSS searches were the result of matches to peptides not present in the SSNIST+DecoySS library, indicating that the sensitivity gains were a direct result of increased proteome coverage. Discrimination was also compared by ROC analysis for RDP and MHP searches of the two libraries (Figs. 5*A*, 5*B*, TargetSS *versus* SSNIST+Decoy). MHP showed a greater increase in discrimination compared with RDP (Figs. 5*A*, 5*B*), a difference likely due to greater sensitivity of the rank-based scores to imperfect spectral simulations. Mascot searches of equivalent peptide databases showed smaller increases in MS/MS assignments (18%) or unique peptides (20%) than any of the spectral library searches, revealing that sequence-based search methods gain less from increased proteome coverage than simulated spectral library methods.

Hybrid Library Searching: Combining Reference and Simulated Spectra—The findings above showed that although the number of peptide identifications are increased using a proteome-wide library, the simulated spectra used to generate this library nevertheless match experimental spectra less well than spectra from reference libraries. We hypothesized that proteome-wide library searching might be optimized by combining simulated and reference spectra together in one library ("Hybrid library"), replacing 154,612 simulated spectra in the TargetSS library with their counterpart spectra from the NIST

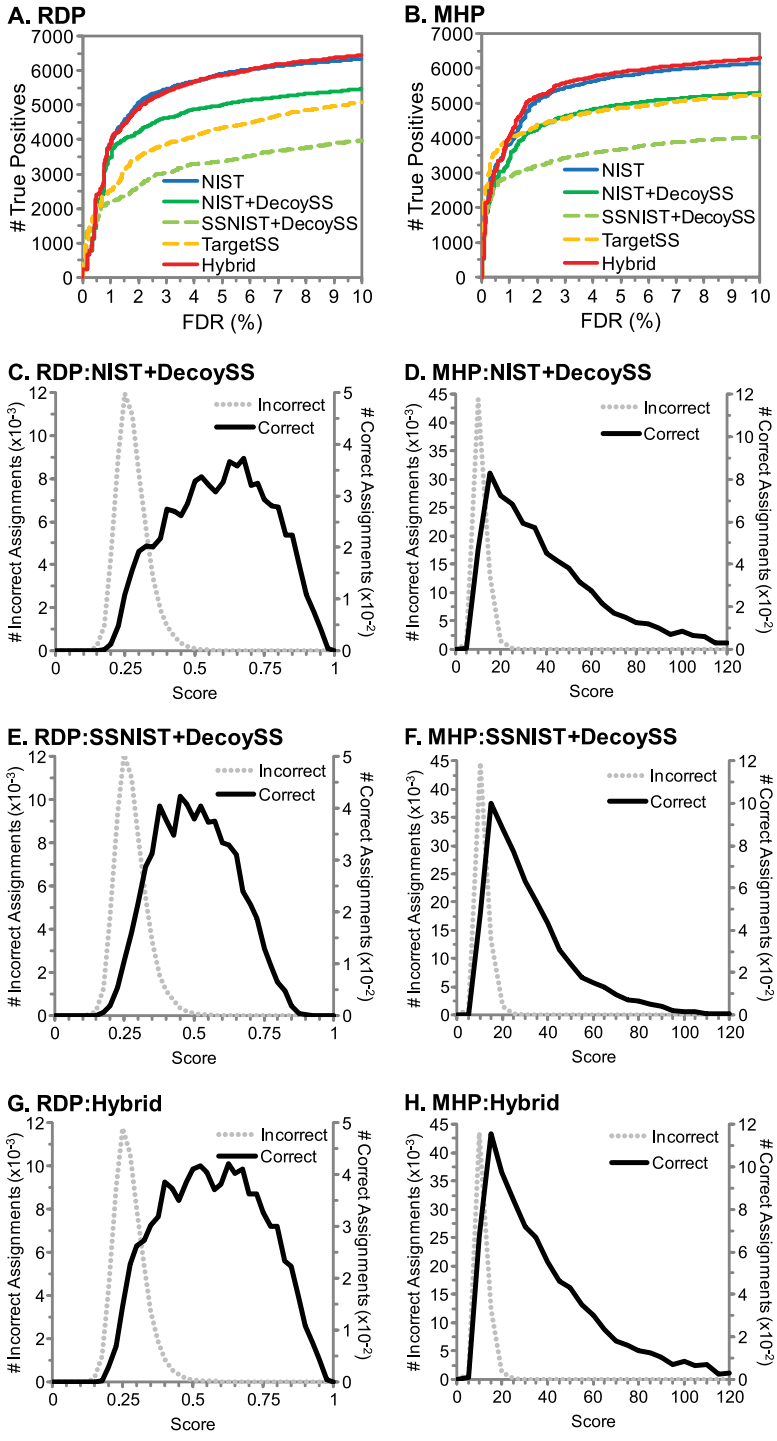

FIG. 5. **The effect of proteome coverage, spectral quality, and search space on discrimination performance.** Search performances for five libraries were compared with assess the contribution of proteome coverage, differences in spectral quality between reference and simulated spectra, and search space. Only the results of (*A*) RDP and (*B*) MHP are shown. Panels *C*, *E*, and *G* show RDP score distributions and panels *D*, *F*, and *H* show MHP score distributions for searches of NIST+DecoySS (*C*, *D*), SSNIST+DecoySS (*E*, *F*), and Hybrid (*G*, *H*) spectral libraries. The SSNIST+DecoySS library was generated by replacing reference spectra in the NIST+DecoySS with simulated counterparts to assess to effect of simulated *versus* reference intensities on discrimination.

reference library. In this way, we might gain the advantage of high coverage provided by simulated spectra, while maintaining the benefits of high quality observed spectra available from reference libraries.

Spectra from the UPS1 data set were searched against a Hybrid library and compared with searches against the TargetSS library, measuring sensitivity at 3% FDR (Table II), with ROC analyses shown for RDP and MHP (Figs. 5*A*, 5*B*). The Hybrid library searches showed high performance over a wide FDR range, comparable or better than the NIST reference library (Figs. 5*A*, 5*B*). Although the numbers of MS/MS assigned to the Hybrid library were comparable to NIST searches, the Hybrid library consistently identified more unique sequences using every score (Table II, Hybrid *versus* NIST). For example, using RDP, 1.3% fewer MS/MS spectra were assigned in searches of Hybrid *versus* NIST libraries

(5387 *versus* 5457 at 3% FDR), but 14% more unique sequences (744 *versus* 651) were identified. Using MHP, Hybrid searches increased both the number of MS/MS assignments (2.4%, from 5451 to 5583) and the number of unique sequences (17% from 657 to 769). Searches performed using every score resulted in increased amino acid coverage of identified proteins with searches of the Hybrid library compared with the NIST reference library (supplemental Fig. S5). Overall, Hybrid library searching with MHP identified the most MS/MS spectra and unique sequences compared with any other score and library combination, illustrating the higher performance when combining simulated and reference spectra in a single high coverage library. Thus, despite the advantages of the small search space of the NIST reference library, the Hybrid library consistently identified more unique sequences, because of increased proteome coverage afforded by simulated spectra.

The performance of the Hybrid library also exceeded that of the simulated library. This was seen when comparing searches against Hybrid *versus* TargetSS libraries using RDP and MHP (Figs. 5*A*, 5*B*), and measuring sensitivity at 3% FDR for all scores (Table II). Searches of the Hybrid library yielded significantly more MS/MS assignments and unique peptides, compared with the TargetSS library (Table II). For example, using RDP, the sensitivity of MS/MS assignments using Hybrid searches increased by 39% (from 3867 to 5387 at 3% FDR), and unique peptide identifications increased by 10% (from 675 to 744). MHP yielded the highest sensitivity improvement in Hybrid over TargetSS searches, which increased by 23% MS/MS assignments (from 4557 to 5583), and by 5% unique peptides (from 732 to 769). Rank-based scores showed greater increases in sensitivity (RDP: +28%, RSIM: +27%) than probability-based scores (SHP: +20%, MHP: +18%).

Interestingly, the differences in sensitivity between Hybrid and TargetSS library searches were comparable to the differences between NIST+DecoySS and SSNIST+DecoySS searches. This was seen by comparing the differences in ROC curves between Hybrid *versus* TargetSS searches (Figs. 5*A*, 5*B*, red *versus* yellow dashed curves), to the differences between NIST+DecoySS *versus* SSNIST+DecoySS searches (Figs. 5*A*, 5*B*, solid green *versus* dashed green curves). Both comparisons showed higher performances of libraries containing spectra of higher quality or simulation accuracy. Thus, replacing simulated spectra with reference spectra improved discrimination to a similar extent in two different library backgrounds, one containing simulated target spectra and the other containing simulated decoy spectra.

In summary, by systematically examining the effects of search space size, proteome coverage, and spectral quality, we conclude that using simulated spectra to construct proteome-wide libraries has effect of reducing sensitivity by expanding the search space, while increasing sensitivity by providing increased coverage. The simulated spectra perform less well than reference library spectra because of poorer spectral quality, but constructing a Hybrid library, which substitutes high-confidence spectra in place of simulated spectra, fully compensates for penalties incurred by the large search space and imperfections in the kinetic simulations while conferring the advantage of proteome-wide coverage. The Hybrid library outperforms a widely used sequence-based algorithm, while allowing FDR statistics to be estimated from parallel searches of decoy spectral libraries.

Estimating FDR for Spectrum-to-Spectrum Searches Using Target-Decoy Strategies—A major advantage of using simulated spectral libraries is the ability to estimate false discovery rates, which is essential when analyzing complex biological samples of unknown composition. The target-decoy search strategy is widely used in sequence-based approaches, where searches against a database of decoy sequences are used to estimate the proportion of false assignments among all assignments accepted above a given score threshold. Commonly, the decoy database contains a set of reversed protein sequences generated from the target database. One of the assumptions in estimating the FDR by this method is that the probabilities of random matches to target and decoy sequences are equal (23, 24). If instead there are biases for or against the decoy sequences, such biases must be incorporated into the FDR (31, 32).

Previously, we described the use of the kinetic fragmentation model to generate decoy spectral libraries, enabling the application of target-decoy methodology to spectral library searching (13). The idea was further explored by Lam *et al.* (23) who reported a significant bias against matches to decoy spectra generated by MassAnalyzer. Therefore, to assess the degree of bias against the DecoySS library, MS/MS spectra from an NIST spectral library of *E. coli* peptides were searched against the human TargetSS library concatenated with the DecoySS library. Because assignments to both TargetSS and DecoySS spectra are necessarily false, biases for or against decoy spectra would be reflected by any deviation from unity in the ratio of target:decoy matches. Each score was used to assess target-decoy bias in searches of the NIST, TargetSS and Hybrid libraries, each concatenated with corresponding decoy spectra generated using MassAnalyzer (Fig. 6).

The frequency of random matches to decoy spectra ranged between 48.5–49.3% for all scores using the concatenated TargetSS+DecoySS library, indicating a small but systematic bias against decoy spectra (Fig. 6*A*, supplemental Fig. S7). This bias was greatest using the NIST target-decoy library, ranging between 44.8–50.5% (Fig. 6*B*), and ranging between 47.1–48.9% for the Hybrid target-decoy library (Fig. 6*C*). Notably, the bias against simulated decoy spectra reported here for the NIST target-decoy library was much smaller than reported previously using a simulated library 3.2-fold smaller than ours (23). The results suggest that larger concatenated decoy libraries show lower bias against decoy spectra. With-

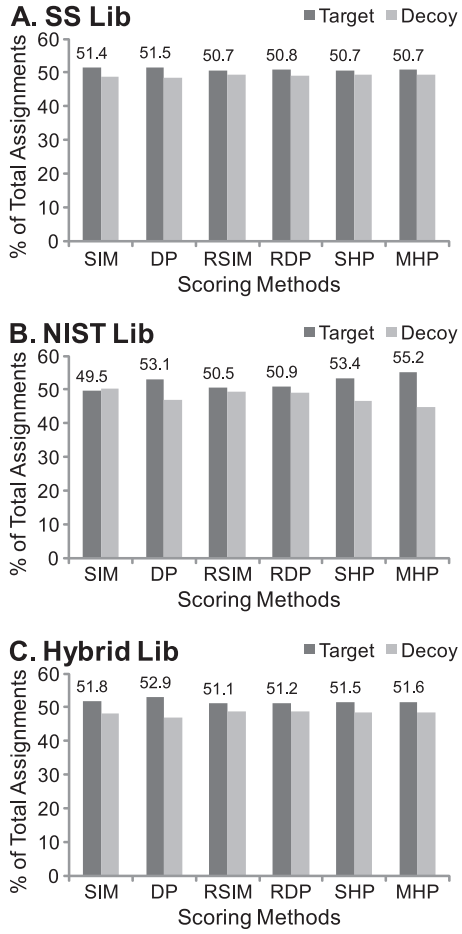

FIG. 6. **Large simulated libraries show reduced bias against decoy spectra compared with the NIST reference library.** Spectra from the NIST *E. coli* reference library were searched against concatenated versions of (*A*) NIST, (*B*) TargetSS, and (*C*) Hybrid libraries, using MassAnalyzer to simulate MS/MS for reversed library peptides. The fraction of total assignments matched to target and decoy spectra is shown for each score.

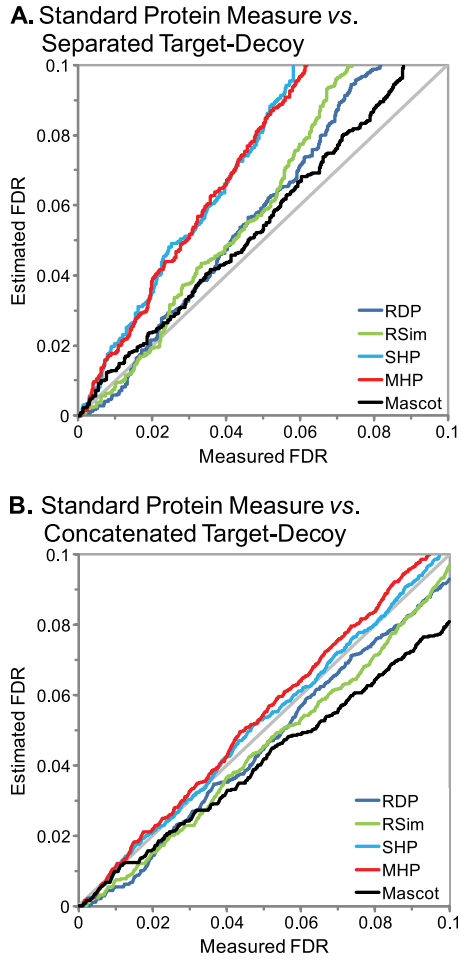

FIG. 7. **Separated and concatenated target-decoy methods show score-specific biases in FDR estimates.** For each scoring method, the UPS1 data set was searched against either separate TargetSS and DecoySS libraries, or a concatenated Target-DecoySS library, comparing measured FDR to estimated FDR. FDR levels were estimated by measurements from (*A*) separated target-decoy searches, or (*B*) the concatenated target-decoy searches, and compared with the measured FDR at varying score thresholds (Supplementary Procedures).

out large numbers of candidates considered for each MS/MS match, potential biases may be amplified using smaller libraries, leading to underestimates of false positives that may not be accounted for in calculations of FDR.

Separated versus Concatenated Target-Decoy Strategies—Finally, we compared performances of separated *versus* concatenated target-decoy searches, where MS/MS were either searched against a target and decoy databases in two separate parallel searches, or searched against a single concatenated target-decoy database. In sequence-based searching, the separated approach tends to be more conservative, overestimating the FDR (25). Conversely, the concatenated approach may underestimate the FDR if biases for or against decoy matches are not accounted for in the FDR calculation (31). We examined whether the two target-decoy strategies showed score-specific biases in spectral library searches, by searching the UPS1 dataset against the TargetSS and DecoySS libraries separately, or against the two libraries concatenated together. For the separated search, the number of false matches (FP) was reported as the number of accepted decoy matches, and for the concatenated search, we reported 2 × FP (24). The estimates of FDR derived from target-decoy measurements were compared with the measured FDR, calculated by the protein standard FDR method, where true and false matches were taken as the number of matches to standard and nonstandard peptides, respectively (from the target library) (Fig. 7).

The results showed that FDR estimated using separated target-decoy searches was overly conservative using all scores (RDP, RSIM, SHP, MHP), and overestimated the measured FDR (Fig. 7*A*). The Mascot ions score showed the closest concordance to the measured FDR, whereas SHP and MHP were the most conservative, resulting in 1.5- to 2-fold

overestimation of the number of false matches. Conversely, the concatenated method was less stringent than the separated approach but showed lowest systematic bias, where FDR was underestimated by Mascot, RDP, and RSIM, and slightly overestimated by SHP and MHP. These trends for spectral library target-decoy searches were consistent with those observed by others using sequence-based target-decoy methods, where the separated method (*e.g.* for SEQUEST XCorr) showed a threefold higher estimates of FDR compared with the concatenated method, and likely overestimated the true FDR (33). This is because in separated searches, all spectra are scored for both decoy and target libraries, whereas in concatenated searches only the highest scored match to target or decoy candidates is reported (competition). High scoring decoy matches normally considered in the separated method do not contribute to FP estimates in the concatenated method because of this score competition. Thus, the separated method uses a larger set of false matches to estimate FPs, and as a result may lead to overestimates of FDR (25), consistent with the trend observed in Fig. 7*A*. Overall, the results indicate that the target-decoy strategy, whether separated or concatenated, must be used with caution when comparing scores and search engines in unknown samples, where score-specific biases may confound estimated FDR levels.

DISCUSSION

In this study, we demonstrate the use of large proteome-wide spectral libraries for MS/MS peptide identification using rank- and probability-based scoring methods. The new scoring metrics are more resistant to library size expansion compared with dot product-based methods, and outperformed Mascot for both small and large sized libraries. We demonstrate that high coverage libraries can be successfully generated by simulations using a kinetic fragmentation model, and when searched with our new scoring methods, result in increased peptide identifications and amino acid coverage over the identified proteins. The best search discrimination at the level of unique peptide sequences was observed using a hybrid library, which combines observed reference MS/MS with spectra generated by kinetic simulation. In this way, we gain the advantage of high accuracy in reference spectral intensities, while retaining the comprehensive proteome coverage afforded by simulated spectra. The increase in search discrimination attained when simulated spectra were replaced with observed spectra indicates a systematic difference in the simulated spectra, perhaps because the gas-phase fragmentation of certain peptides was imperfectly modeled by MassAnalyzer. Work ongoing in our lab to develop an improved kinetic model indicates that certain classes of peptides with unusual fragmentation chemistries indeed are poorly modeled, because of dissociation mechanisms not accounted for in the original model.

Interestingly, while searches against a small reference library identified larger numbers of MS/MS compared with searches of the TargetSS library, these assignments represented a smaller number of unique sequences. We attribute this effect to the higher protein coverage of the larger proteome-wide libraries. We expect that as reference libraries grow in size, the performance of hybrid libraries will show a corresponding increase in performance due to the presence of larger proportions of high quality observed MS/MS. Similarly, improvements in gas-phase peptide fragmentation models (4) will enable more accurate prediction of MS/MS intensities and translate to increased discriminatory power of Spec2spec library search methods.

One important question is whether spectra generated with different instruments and activation methods can be identified by searching ion-trap reference libraries and libraries simulated with MassAnalyzer's ion-trap fragmentation model. Previous studies have demonstrated that ion-trap libraries can be used to identify triple quadrupole and Qtof MS/MS, but at the expense of lower sensitivity due to minor differences in the fragmentation and spectral characteristics (34, 35). Thus, when searching non-ion trap data against simulated spectra, using instrument specific kinetic models for library generation would likely result in increased performance. While the current version of MassAnalyzer is capable of simulating collision-cell fragmentation spectra (Qtof), the underlying model has not been published.

Target-decoy search strategies are used for estimating statistical significance of MS/MS assignments, and have seen recent use in evaluating false discovery rates in spectral library search algorithms (13, 23). Concatenated target-decoy searches generally show a bias against simulated decoy spectra, with smaller reference libraries showing higher degree of bias than the larger simulated libraries. Furthermore, tests with the UPS1 standard protein mixture indicated that the separated target-decoy method systematically overestimates FDR and FP for all scores, with Mascot showing the lowest bias. Conversely, the concatenated method systematically underestimated the FDR using Mascot, RDP, and RSIM, with the hypergeometric scores showing the least amount of bias. The optimal target-decoy strategy for application to large scale unknown samples is likely score dependent, and may require correction of the underlying bias to enable accurate comparisons of different scores and search algorithms.

In summary, we have developed and adapted new scoring methods to a spectrum-to-spectrum search strategy optimized for large simulated libraries with high proteome coverage. In addition, these scores are applicable to the smaller reference libraries, with discrimination performance surpassing the sequence-based search algorithm Mascot. Moreover, our hybrid library approach shows the highest discrimination performance for all scores. The general approach demonstrated in this study is a significant step toward the use of the spectrum-to-spectrum searching as a primary protein identification tool in proteomic workflows.

Acknowledgments—We thank Shaojun Sun and Karen Meyer-Arendt for insightful discussions, and Paul Rudnick for help with identifying misannotated reference library spectra.

* This work was supported by the W.M. Keck Foundation and National Cancer Institute grants R01 CA126240 and R01 CA125291, part of NCI Clinical Proteomic Technologies for Cancer (http://proteomics.cancer.gov) initiative.

[S] This article contains supplemental Methods, Figs S1 to S7, and Table S1.

¶ To whom correspondence should be addressed: Department of Chemistry and Biochemistry, University of Colorado, Boulder Colorado 80309. Phone: 303-492-5519; Fax: 303-492-2439; E-mail: William.Old@colorado.edu.

REFERENCES

1. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass. Spectrom.* **35,** 1399–1406
2. Sun, S., Meyer-Arendt, K., Eichelberger, B., Brown, R., Yen, C. Y., Old, W. M., Pierce, K., Cios, K. J., Ahn, N. G., and Resing, K. A. (2007) Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol. Cell Proteomics*. **6,** 1–17
3. Zhang, Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76,** 3908–3922
4. Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **77,** 6364–6373
5. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., and MacCoss, M. J. (2006) Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.* **78,** 5678–5684
6. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., and Aebersold, R. (2008) Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods*. **5,** 873–875
7. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667
8. Stein, S. E., and Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.* **5,** 859–866
9. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
10. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome. Res*. **3,** 958–964
11. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome. Res.* **6,** 654–661
12. Stein, S. E., and Rudnick, P. A. (2009) NIST Peptide Mass Spectral Libraries. Human Peptide Mass Spectral Reference Data, H. sapiens, ion trap, Official Build Date: Feb. 4, 2009. National Institute of Standards and Technology, Gaithersburg, MD, 20899, http://peptide.nist.gov.
13. Yen, C. Y., Meyer-Arendt, K., Eichelberger, B., Sun, S., Houel, S., Old, W. M., Knight, R., Ahn, N. G., Hunter, L. E., and Resing, K. A. (2009) A simulated MS/MS library for spectrum-to-spectrum searching in large-scale identification of proteins. *Mol. Cell Proteomics*. **8,** 857–869
14. Andrews, P. C., Arnott, D. P., Gawinowicz, M. A., Kowalak, J. A., Lane, W. S., Lilley, K. S., Martin, L. T., and Stein, S. (2006) ABRF-sPRG 2006 study: a proteomics standard. Poster available at http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf. In *ABRF* 2006 Long Beach, CA
15. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467
16. Frewen, B., and MacCoss, M. J. (2007) Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinformatics*. Chapter 13, Unit 13 17
17. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
18. Yen, C. Y., Russell, S., Mendoza, A. M., Meyer-Arendt, K., Sun, S., Cios, K. J., Ahn, N. G., and Resing, K. A. (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.* **78,** 1071–1084
19. Sadygov, R. G., and Yates, J. R., 3rd (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75,** 3792–3798
20. Sadygov, R. G., Good, D. M., Swaney, D. L., and Coon, J. J. (2009) A new probabilistic database search algorithm for ETD spectra. *J. Proteome. Res.* **8,** 3198–3205
21. MacCoss, M. J., Wu, C. C., and Yates, J. R., 3rd (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74,** 5593–5599
22. Lane, W. S., Nesvizhskii, A. I., Searle, B., Tabb, D. L., Kowalak, J. A., and Seymour, S. L. (2007) Bioinformatic Evaluation of Datasets Derived from the ABRF sPRG Proteomics Standard. Poster available at http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearchGroup/Studies/sPRG-BIC2007poster.pdf. In *ABRF* 2007 Tampa, FL
23. Lam, H., Deutsch, E. W., and Aebersold, R. (2010) Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome. Res.* **9,** 605–610
24. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*. **4,** 207–214
25. Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome. Res.* **7,** 29–34
26. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 9440–9445
27. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R., 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.* **75,** 2470–2477
28. Zhang, Z., and Bordas-Nagy, J. (2006) Peptide conformation in gas phase probed by collision-induced dissociation and its correlation to conformation in condensed phases. *J. Am. Soc. Mass. Spectrom.* **17,** 786–794
29. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76,** 3556–3568
30. Ning, K., Fermin, D., and Nesvizhskii, A. I. (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10,** 2712–2718
31. Fitzgibbon, M., Li, Q., and McIntosh, M. (2008) Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome. Res.* **7,** 35–39
32. Choi, H., and Nesvizhskii, A. I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome. Res.* **7,** 47–50
33. Wang, G., Wu, W. W., Zhang, Z., Masilamani, S., and Shen, R. F. (2009) Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* **81,** 146–159
34. Yates, J. R., 3rd, Morgan, S. F., Gatlin, C. L., Griffin, P. R., and Eng, J. K. (1998) Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* **70,** 3557–3565
35. Zhang, X., Li, Y., Shao, W., and Lam, H. (2011) Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **11,** 1075–1085