RESEARCH ARTICLE

Development and validation of a spectral library searching method for peptide identification from MS/MS

Henry Lam[1], Eric W. Deutsch[1], James S. Eddes[1], Jimmy K. Eng[1, 2], Nichole King[1], Stephen E. Stein[3] and Ruedi Aebersold[1, 4]

[1] Institute for Systems Biology, Seattle, WA, USA
[2] Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[3] National Institute of Standards and Technology, Gaithersburg, MD, USA
[4] Institute of Molecular Systems Biology, ETH Zurich, Switzerland and Faculty of Science, University of Zurich, Switzerland

Received: August 14, 2006
Revised: October 23, 2006
Accepted: December 23, 2006

A notable inefficiency of shotgun proteomics experiments is the repeated rediscovery of the same identifiable peptides by sequence database searching methods, which often are time-consuming and error-prone. A more precise and efficient method, in which previously observed and identified peptide MS/MS spectra are catalogued and condensed into searchable spectral libraries to allow new identifications by spectral matching, is seen as a promising alternative. To that end, an open-source, functionally complete, high-throughput and readily extensible MS/MS spectral searching tool, SpectraST, was developed. A high-quality spectral library was constructed by combining the high-confidence identifications of millions of spectra taken from various data repositories and searched using four sequence search engines. The resulting library consists of over 30 000 spectra for *Saccharomyces cerevisiae*. Using this library, SpectraST vastly outperforms the sequence search engine SEQUEST in terms of speed and the ability to discriminate good and bad hits. A unique advantage of SpectraST is its full integration into the popular Trans Proteomic Pipeline suite of software, which facilitates user adoption and provides important functionalities such as peptide and protein probability assignment, quantification, and data visualization. This method of spectral library searching is especially suited for targeted proteomics applications, offering superior performance to traditional sequence searching.

Keywords:
Peptide identification by MS/MS / Spectral library / Spectral similarity / Targeted proteomics

1 Introduction

Proteomics, which aims at the systematic identification and quantification of all proteins in a biological system, has become a popular and effective analytical approach in life science research [1–4]. Over the past decade, thanks to the convergence of several important technological advances, MS-based proteomics technologies have progressed at a rapid pace and gained wide acceptance and usage. In one of the most frequently practiced proteomics workflows, commonly known as shotgun proteomics, a protein sample of interest is first digested with a proteolytic enzyme (trypsin being the most common) to yield peptides that are amenable to LC-MS/MS analysis. The peptides in the resulting mixture are chromatographically resolved, ionized by techniques such as ESI or MALDI before being analyzed by a mass spectrometer. A fraction of the peptide ions are selectively isolated by the mass spectrometer and subjected to CID to yield characteristic MS/MS spectra that allow the determi-

Correspondence: Professor Ruedi Aebersold, Institute of Molecular Systems Biology, HPT E 78, Wolfgang-Pauli-Str. 16, 8093 Zurich, Switzerland
E-mail: aebersold@imsb.biol.ethz.ch
Fax: +41-206-299-6573

Abbreviations: NIST, National Institute of Standards and Technology; **Th,** Thomson, unit of mass-to-charge ratio (Da/e); **TPP,** trans proteomic pipeline

nation of the peptide sequence. Finally, the peptide ions subjected to CID are identified by computational tools that seek to match the observed MS/MS spectra to the predicted ones corresponding to putative peptide sequences generated by *in silico* digestion of protein sequences in public databases [5, 6].

This last step of sequence database searching, that is, the inference of the peptide sequence from the MS/MS spectra of fragmented peptide ions, is a challenging, error-prone, and computationally expensive exercise, and has been a subject of intense research since the early days of proteomics [7–10]. Several popular computational tools developed for that purpose have emerged over the years, each employing different algorithms and heuristics to achieve an acceptable balance of sensitivity and accuracy [11–16]. However, despite the tremendous improvement in computer hardware and software over the past decade, this step often remains the bottleneck of many shotgun proteomics experiments. A great deal of computational resources are often required for this process, limiting the application of this powerful analytical technique to only those research groups that can afford the costly computational infrastructure and the personnel to maintain it. With the advent of more powerful mass spectrometers that are able to churn out MS/MS spectra at much faster rates, as well as the growing interest in doing more sophisticated experiments that require a larger body of data, this problem is expected to get worse in time, unless an alternative approach to sequence searching can be developed to ease the computational load of a proteomics facility.

Long before the time of proteomics, analytical chemists learned how to infer the structure of an unknown chemical species from its characteristic fragmentation pattern, typically generated by electron impact, in the mass spectrometer. While this process occasionally involves manual or computer-assisted assignment of peaks in the mass spectra to possible fragments, a general and commonly practiced method of identification is by library searching [17–19]. In this approach, a spectral library is meticulously compiled from a large collection of experimentally observed mass spectra of known compounds. An unknown spectrum can then be identified by comparing it to all the candidates in the spectral library to determine the best match with the highest spectral similarity. In this paper, we report the development and validation of SpectraST, a functionally complete, high-throughput software tool to perform spectral library searching on peptide CID spectra against carefully compiled spectral libraries for peptides. We will demonstrate that this method has the potential to gradually replace traditional sequence searching methods in proteomics experiments, thereby alleviating some of limitations discussed above.

This approach has been proposed and explored previously, although the main difficulty at the time was the availability of a reliable and comprehensive spectral library for peptides [20]. With the explosion of proteomics data in recent years, the time is ripe to revisit the idea, with some preliminary demonstration of success being reported in two recent publications [21, 22]. The availability of online data repositories developed in recent years [23–27], as well as emerging unified standards for representing shotgun proteomics data, such as mzXML [28] and mzData (http://psidev.sourceforge.net/ms/#mzdata), have made it possible to collect and catalog an adequate set of peptide CID spectra to create a high-quality spectral library. For the spectral library to be comprehensive (containing sufficient entries to cover a high proportion of the observed proteome) and accurate (containing high-quality and truly characteristic MS/MS spectra that can be confidently mapped to peptides), it is important to gather raw MS/MS spectra from a wide variety of sources, to identify them as accurately as possible, to filter out the inevitable false positives, and to process the spectra to reduce noise and other experimental artifacts. Such an endeavor is undertaken by the National Institute of Standards and Technology (NIST), in collaboration with the PeptideAtlas project of our group [24, 29]. The resulting high-quality spectral libraries were then used for spectral searching in this study.

The approach of spectral library searching represents a significant shift of paradigm for proteomics research. While traditional sequence searching methods are suitable for the discovery of novel segments of the proteome (as long as the peptide sequences are in the sequence database), they are less efficient for in-depth investigation of a subproteome that has already been mapped out. In traditional sequence searching, an unknown spectrum is sent for search against the entire sequence database, without any input from previous knowledge of the observed proteome [30]. Moreover, the expected fragmentation pattern for each candidate peptide is generated anew every time, while the great wealth of spectral information obtained in previous experiments remains unused. Spectral library searching, on the other hand, is a natural way to incorporate previous knowledge about observed peptides and their respective fragmentation patterns into a new search. Therefore, it is ideal for targeted proteomics, in which the experimental goals do not call for the discovery of previously unobserved peptides, but rather for the systematic and repeated investigation of a certain predefined set of peptides or proteins of interest, for instance, in a time-series experiment or in a comparative study across many samples [30].

To facilitate user adoption and to capitalize on the wealth of other useful computational tools available, we have developed SpectraST in concert with our open-source and widely used Trans Proteomic Pipeline (TPP) suite of software [31], which allows users to perform the entire proteomic workflow in a standardized, user-friendly manner. Functionalities available to SpectraST users include the importation of raw mass spectrometry data files from various instruments and vendors, automatic validation and probability assignment at the peptide and protein levels, quantification, and large

dataset management and visualization, among others. As such, SpectraST offers the unique advantage of usability over other spectral searching options.

The SpectraST program was validated with four different datasets consisting over 30 000 MS/MS spectra and over 4 200 confidently identified peptides, using the yeast (*S. cerevisiae*) spectral library developed at NIST. We conducted a thorough comparison of its performance to that of SEQUEST [11] one of the most popular sequence search engines for peptide identification from MS/MS spectra, and demonstrated the superior performance of SpectraST over SEQUEST, in terms of speed, sensitivity, and false discovery rates.

2 Materials and methods

2.1 Creation of consensus spectral libraries

A library of "consensus spectra" derived from the CID of 34 426 peptide ions was created from 43 LC-MS/MS datasets of tryptic digests of baker's yeast (*S. cerevisiae*), taken mostly from public data repositories [23-27], and released as part of the freely available NIST Library of Peptide Ion Fragmentation Spectra (The NIST Library of Peptide Ion Fragmentation Spectra, June 2006 Version. Available for download at http://www.peptideatlas/org/SpectralLibraryDownload.php.).

These datasets were acquired in shotgun proteomics experiments in many laboratories exclusively on ion trap instruments. Details about the methods used to compile the library can be found in the documentation for the library, and will be further described in a future publication. Briefly, four popular sequence search engines (SEQUEST [11], MASCOT [12], X!Tandem [13], and OMSSA [14], were used for preliminary identifications of MS/MS spectra. Multiple spectra passing a strict confidence threshold that were identified as the same peptide ions (*i.e.* replicates) were combined to create a characteristic "consensus spectrum." Peak intensities were averaged across the replicates and only peaks that occur in a majority of the replicates were included in the consensus spectrum. Various quality filters were employed to exclude incorrectly identified or highly impure spectra. Some important factors considered included the noisiness of the spectra, the fraction of abundance from unassigned peaks, similarity to a theoretical spectrum derived from known fragmentation rules, the sequence characteristics of the peptide (termini, number of missed cleavages, *etc.*), and the presence of multiple charge states, modifications, and/or sub-sequences of the same sequence. Various useful pieces of information about a consensus library spectrum, including the sources of the replicates, detailed sequence search results, and various quality metrics, were provided in the library file and can be easily examined. The best-scoring replicate spectrum of each peptide ion represented in the library was also included in the current release, but was ignored in this study.

2.2 SpectraST software implementation

SpectraST was written in C++ and compiled on a LINUX platform, although a Windows-compatible version will be made available together with the TPP software suite. It runs in two modes: a *Create* mode and a *Search* mode. In the *Create* mode, a text file containing the library spectra and their associated peptide identifications is parsed to create a precursor m/z-indexed library for fast searching. In the *Search* mode, unknown query spectra (in mzXML [23], format) with known precursor m/z values are searched against this indexed library for the best matching spectrum of similar precursor m/z values. The design of SpectraST is highly modular, such that various functionalities, including the import of the library files, the indexing, the loading of the query spectra, the similarity scoring of spectra pairs, and the output of the results, are readily extensible. Many behaviors of the program are also controllable *via* user-defined options. Written with the enormous data-generating capacity of modern mass spectrometer in mind, the program operates efficiently on large data files and library files, and uses efficient spectral processing and comparison algorithms. To allow for maximum usability, SpectraST is designed to run on modern personal computers with modest specifications. In addition, the program is also integrated seamlessly with the TPP software tools so that the user can take advantage of the upstream and downstream processing capabilities of TPP.

2.3 Spectral processing and similarity scoring in SpectraST

SpectraST is designed to be highly customizable, and all behaviors described below can be further optimized by the users to suit their needs. By default, SpectraST preprocesses both the query and library spectra in the following manner. Spectra having fewer than six peaks overall, or having negligible signal anywhere above the m/z value of 500 Thomson (Th) (likely to be non-peptide impurity) are discarded. The remaining spectra are noise-reduced by removing all peaks below the intensity threshold of 2.0 (arbitrary unit) and re-scaled by taking the square root of all raw peak intensities to de-emphasize dominant peaks. Unassigned peaks present in the library spectra are penalized by multiplying their intensities by a factor of 0.2. Next, all scaled peak intensities are placed into 1 Th-wide bins. To "spread out" each of the peak to enable the matching of corresponding, but slightly m/z-shifted peaks, a fraction of the peak intensity is also assigned to the neighboring bins. The bins are then normalized by the magnitude of the spectral vector, as follows in Eq (1):

$$\hat{I}_j = \frac{I_j}{\sqrt{\sum_j I_j^2}} \quad (1)$$

where I_j and $\hat{I}_j$ are the raw intensity and the normalized intensity of the j^{th} bin, respectively. A simple spectral dot product function is used to determine spectral similarity.

Previous work in the MS analysis of small molecules has demonstrated that the spectral dot product is the most effective among several similarity scoring functions [18], The spectral dot product (D) is calculated as follows in Eq (2):

$$D = \sum_{\mathrm{j}} \hat{I}_{\mathrm{library,j}} \hat{I}_{\mathrm{query,j}} \tag{2}$$

where $\hat{I}_{\mathrm{library,j}}$ and $\hat{I}_{\mathrm{query,j}}$ are the normalized intensity of the j^{th} bin of the library spectrum, and that of the matching bin (of the same m/z value) of the query spectrum, respectively. Since the binned intensities are normalized, the value of D is always between 0 and 1, with the latter value indicating identical spectra. In addition to the dot product, two other metrics are used to aid in the discrimination of good and bad matches. The first of these, the normalized difference between the dot product of the top hit (D_1) and the runner-up (D_2), is calculated as follows in Eq. (3):

$$\Delta D = \frac{D_1 - D_2}{D_1} \tag{3}$$

A large ΔD implies that the top hit clearly stands out from other candidates, and hence is more likely to be correct. The last metric used, called the dot bias (DB), measures how much of the dot product is dominated by a few peaks, and is given by Eq. (4):

$$DB = \frac{\sqrt{\sum_{\mathrm{j}} \hat{I}^2_{\mathrm{library,j}} \hat{I}^2_{\mathrm{query,j}}}}{D} \tag{4}$$

DB attains a value of 1 for the case when the dot product is due to a single matching peak, and a value of $1 \Big/ \sqrt{(number\ of\ bins)} \approx 0$ when the contribution to the dot product is evenly distributed among all bins. DB values that are too large or too small often signify doubtful identifications that have inflated dot products due to the matching of a few dominant peaks, or to the matching of many small peaks that are likely noise.

The candidates are ranked by a discriminant scoring function, F, which is calculated as follows in Eq. (5):

$$F = 0.6D + 0.4\Delta D - b \tag{5}$$

where b, a penalty assessed for unfavorable dot bias values, is given by Eq. (6):

$$b = \begin{cases} 0.12 & \text{if } DB < 0.1 \\ 0.12 & \text{if } 0.35 < DB \leq 0.4 \\ 0.18 & \text{if } 0.4 < DB \leq 0.45 \\ 0.24 & \text{if } DB > 0.45 \\ 0 & \textit{for all other values of } DB \end{cases} \tag{6}$$

The various parameters of spectral processing and similarity scoring, as well as the discriminant scoring function, were chosen in a trial-and-error manner to obtain satisfactory performance for one test dataset (Dataset I, see Table 1), and the same parameters and discriminant scoring function were applied without change to the others (Datasets II, III and IV). Note that all of these parameters are easily adjustable within the open-source and modular framework of SpectraST, and the user is encouraged and empowered to optimize them for their own needs. The "default" parameters described above, though not necessarily optimal for all possible situations, should be generally adequate for a wide variety of applications.

2.4 Preparation of the test datasets

The four test datasets are listed in Table 1. Dataset I was a published dataset available in the public repository Peptide Atlas (http://www.peptideatlas.org/, dataset name "pxproteome"), and the materials and methods pertaining to its creation can be found in ref. [32]. The other three datasets (II, III and IV) were obtained from shotgun proteomics experiments of whole yeast lysate. All reagents were purchased from Sigma (St. Louis, MO) unless otherwise stated. *Saccharomyces cerevisiae* (strain BY4741) was grown and lysed

Table 1. The four *S. cerevisiae* test datasets used to validate SpectraST

No.	Instrument	Sample[a)]	MS/MS Spectra[b)]	Modification on cysteines	Included in NIST library construction
I	LCQ[c)]	Yeast peroxisomes	23354	Cleavable ICAT[d)]	Yes
II	LCQ[c)]	Whole yeast extract	2670	CAM[e)]	No
III	LTQ[f)]	Whole yeast extract	3240	CAM[e)]	No
IV	Q-TOF[g)]	Whole yeast extract	1935	CAM[e)]	No

a) Sample preparation is given in detail in Section 2.4
b) Numbers of spectra passing SpectraST's quality filter and actually searched
c) ThermoFinnigan LCQ Deca
d) Acid-cleavable isotope coded affinity tag
e) Carbamidomethyl, introduced by reacting with iodoacetamine.
f) ThermoFinnigan LTQ
g) Micromass Q-TOF Ultima

as described in ref. [33]. The whole cell lysate was then reacted with iodoacetamide and digested with trypsin (Promega, Madison, WI) according to the procedure described in ref. [34]. Next, the tryptic digest was separated by strong cation exchange chromatography as described in ref. [35] using a polysulfoethyl-A 2.1 mm × 200 mm column (PolyLC, Columbia, MD). The fractions used in Dataset II and III were collected from 24 to 26 min, and the whole cell lysate digest was used without fractionation for Dataset IV. The samples were then loaded onto a 2 cm long Magic C18 (5 μm 100Å) pre-column connected to a 10 cm long Magic C18 (5 μm 200Å) analytical column (Michrom Bioresources, Auburn, CA), and finally to the mass spectrometer, in an in-house built RP-ESI device [36]. A wash of 5 min with a solution of 5% ACN and 0.1% formic acid was followed by a 60 min linear gradient from 15% to 35% ACN. Both the wash and the gradient were delivered using an Agilent 1100 series pump, and the gradient was followed by a cleaning and equilibration step before the next run. The mass spectrometers used were: ThermoFinnigan LCQ Deca (Thermo Corporation, Waltham, MA) for Datasets II, ThermoFinnigan LTQ (Thermo Corporation) for Dataset III, and Micromass Q-TOF Ultima (Waters Corporation, Milford, MA) for Dataset IV. MS/MS was performed in an automatic fashion, with MS/MS spectra acquired on the three most abundant peaks not on the exclusion list per MS scan. Raw data files were converted to mzXML files in centroid mode using converters available in the TPP package.

2.5 Searching test yeast datasets by SpectraST and SEQUEST

The four test datasets (Table 1) were searched by SpectraST against the NIST yeast consensus spectral library with a ± 3.0 Th precursor m/z value window. The same datasets were searched by SEQUEST (version 27) against *S. cerevisiae* protein sequences in the NCI non-redundant protein sequence database (ftp://ftp.ncifcrf.gov/pub/nonredun/, version on date 9/20/2005), appended with common contaminants (trypsin, human keratins). The following search parameters were used: ± 3.0 Da precursor mass window, isotopically averaged mass, at least one tryptic terminus, at most one missed tryptic cleavage, variable methionine oxidation, and acid-cleavable ICAT modification on cysteines for the ICAT dataset (Dataset I).

2.6 Comparing the performance of SpectraST and SEQUEST

The top peptide hits returned by both SpectraST and SEQUEST were subjected to automated validation and probability assignment by PeptideProphet [16]. The discriminant function for SpectraST is given in Eq. (5), and the optimized discriminant function as described in ref. [16] was used without change for SEQUEST. A conservative probability cutoff of 0.99, above which a peptide hit was considered positive, was used. The number of positive hits, the PeptideProphet-predicted sensitivity and false discovery rate, as well as the actual peptide identification for each spectrum, were compared between SEQUEST and SpectraST. Manual validation was employed for identifications made by SpectraST but missed by SEQUEST to determine their correctness. Factors such as the presence of a majority of the common fragment ions, the global similarity to the library spectrum, the overall noisiness, the number of unexplained high-abundance peaks, and the presence of continuous ion series were used as criteria for manual validation.

3 Results

3.1 Speed

On a machine with an Intel P4 3.4 GHz single CPU and 2 GB of RAM, SpectraST finished the library creation step of over 30 000 yeast consensus spectra in less than 2 min. Against a library of this size, SpectraST achieved a typical search speed of 0.005 s per query spectrum for a dataset with over 23 000 spectra. In contrast, a SEQUEST search of the same dataset requires on average 6.4 s per query spectrum (calculated from the time elapsed reported in the SEQUEST .out files), to complete on the same machine. Thus, SpectraST is able to complete the search about three orders-of-magnitude faster than SEQUEST against this library. We expect the per-query search time to be largely proportional to the library size.

3.2 Sensitivity and error analysis by PeptideProphet

PeptideProphet analysis was performed on the four test datasets (Table 1) after searching them by both SpectraST and SEQUEST. The receiver operator characteristics curves were plotted for all four datasets for both search engines (Fig. 1). The results clearly show that SpectraST is able to obtain higher sensitivity at all desired false discovery rates, for all four datasets. This is the direct consequence of the fact that SpectraST is more capable of distinguishing between a good and a bad match than SEQUEST. The same conclusion can be drawn from looking at the histogram of the discriminant scores of the top hits and noting the better score separation between presumed positive and negative distributions for SpectraST than for SEQUEST (Fig. 2).

3.3 Comparison of the peptide identifications

The peptide identifications made by SpectraST and SEQUEST were compared for each query spectrum, and the results were presented in Table 2. As shown, a majority (72–86%) of SEQUEST positive hits ($P_{seq} \geq 0.99$) are identified to

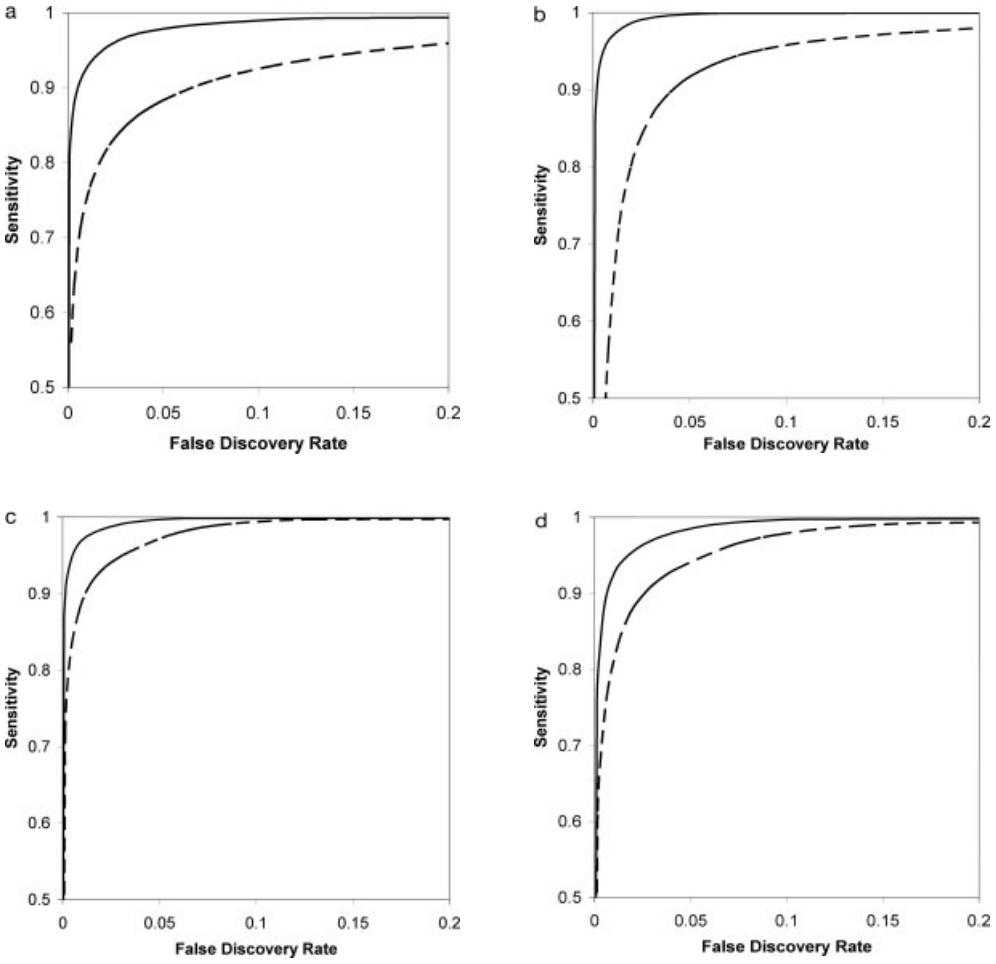


Figure 1. Receiver operator characteristic curves for both SEQUEST and SpectraST searches, for all four test datasets. Sensitivities and false discovery rates are calculated by PeptideProphet. Solid line: SpectraST; Dotted line: SEQUEST. (a) Dataset I, (b) Dataset II, (c) Dataset III, (d) Dataset IV

be the same peptide by SpectraST, most of them as positive hits ($P_{spec} \geq 0.99$). Among the remaining SEQUEST positive hits not identified by SpectraST, 97% are not present in the NIST yeast consensus spectral library to begin with, and therefore should not be expected to be found by SpectraST. Thus, SpectraST is almost always able to recover highly-confident identifications made by SEQUEST, provided that the identification is in the library.

On the other hand, SpectraST assigned considerably more positive hits ($P_{spec} \geq 0.99$) to Datasets I and II than SEQUEST, but fewer positive hits to Datasets III and IV. Among the SpectraST positive hits, 84–96% matched the corresponding SEQUEST identifications, most of them matching $P_{seq} \geq 0.99$ identifications. In this study, all positive hits that are identical to their respective SEQUEST top-hit of the same query spectrum were presumed correct, and the remaining ones were manually validated. It was found that all but a few of these extra hits are correct, testifying to the accuracy of SpectraST. Upon closer inspection, many of them actually correspond to the non-top hits returned by SEQUEST, but are buried because only the top hit is considered for validation. Because SEQUEST does not use peptide sequence information (*e.g.* number of tryptic termini, number of missed tryptic cleavage, and whether or not a cysteine is present (for the ICAT dataset)) to rank its hits, it is sometimes the case that the lower hit matching the SpectraST identification has better sequence characteristic than the top hit and therefore is more likely to be correct. Some others are missed by SEQUEST because they fall outside of the precursor mass window employed by SEQUEST. More often, the query spectrum is often noisy or consisting of dominant peaks that may have confused the sequence search engine, which does not take into full account the peak intensities.

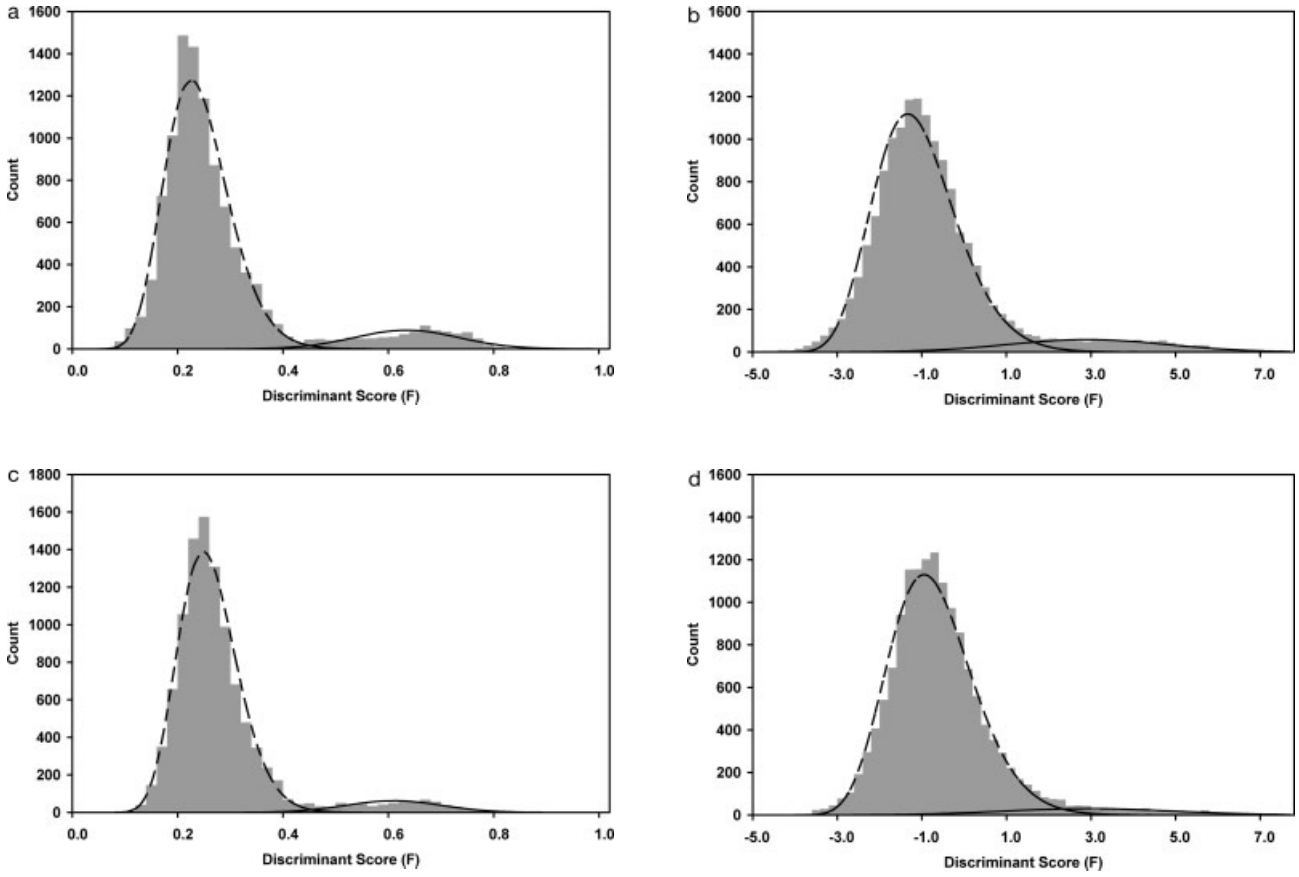


Figure 2. The histograms of top-hit discriminant scores of +2 and +3 peptide identifications, from both SEQUEST and SpectraST searches, for Dataset I, illustrating the superior ability of SpectraST to distinguish between good and bad hits. Shaded area: the observed histogram, Solid line: Positive distribution (modeled as a Gaussian distribution) fitted by PeptideProphet, Dotted line: Negative distribution (modeled as a Gamma distribution) fitted by PeptideProphet.
(a) SpectraST +2 peptide identifications, (b) SEQUEST +2 peptide identifications, (c) SpectraST +3 peptide identifications, (d) SEQUEST +3 peptide identifications

4 Discussion

4.1 Advantages of spectral searching

The results presented here suggest that the spectral searching approach to peptide identification has several important advantages over the traditional sequence search approach. First, the search space for spectral library searching is much smaller compared to that of sequence searching. Proteomics researchers have long observed that some peptides are detected all the time in shotgun proteomics experiments while some never are, for a variety of reasons. In traditional sequence search approaches, however, all putative peptides resulting from the *in silico* digestion of the sequence database are being considered as possible candidates, while in reality, most of these peptides will never be observed experimentally. As a result, a vast majority of the computational time is wasted on fruitlessly matching these peptides to each of the query spectrum. Recently developed methods address this deficiency by limiting the search space to so-called proteotypic peptides, which are the ones found empirically to have a high chance of being observed [30, 37]. In spectral searching, similarly, the library only consists of spectra of peptides that have been observed numerous times in the past, and therefore represents a much reduced search space. Table 3 shows the huge discrepancy between the search space of spectral search and traditional sequence search, especially when less stringent search parameters are used.

To further reduce the search space, spectral searching also does not require the knowledge of the charge of the precursor ion undergoing collision-induced dissociation. Each library spectrum is associated with both a peptide sequence and a precursor charge; spectra of different precursor charges are often included for the same peptide sequence. Hence, a spectral search engine like SpectraST simply needs to match the precursor m/z values of the unknown and library spectra. In contrast, when the pre

Table 2. Comparison of the positive hits found by SEQUEST and SpectraST for the four test datasets. Positive hits are defined as hits having PeptideProphet assigned probabilities (P_{seq} for SEQUEST, P_{spec} for SpectraST) of over 0.99

SEQUEST positive hits ($P_{seq} \geq 0.99$)		**Dataset I (LCQ)**		**Dataset II (LCQ)**		**Dataset III (LTQ)**		**Dataset IV (Q-TOF)**	
		No. spectra	%	No. spectra	%	No. spectra	%	No. spectra	%
Identical to	$P_{spec} \geq 0.99$	910	74	357	81	916	65	405	61
SpectraST top hit	$P_{spec} < 0.99$	83	7	23	5	99	7	112	17
Not identical to	Not in spectral library	236	19	55	13	387	27	139	21
SpectraST top hit	In spectral library	1	0.1	4	1	12	1	7	1
TOTAL		**1230**	100	**439**	100	**1414**	100	**663**	100
SpectraST positive hits ($P_{spec} \geq 0.99$)		No. spectra	%	No. spectra	%	No. spectra	%	No. spectra	%
Identical to	$P_{seq} \geq 0.99$	910	59	357	36	916	81	405	66
SEQUEST top hit	$P_{seq} < 0.99$	378	25	494	50	170	15	163	26
Not identical to	Correct[a)]	252	16	131	13	42	4	38	6
SEQUEST top hit	Incorrect[a]	5	0.3	3	0.3	0	0	12	2
TOTAL		**1545**	100	**985**	100	**1128**	100	**618**	100

a) By manual validation as described in Section 2.7

Table 3. The vast difference between the search space of spectral search and sequence search. Both the entire search space over all *m/z*, and a typical search space encountered in a single query with an observed precursor *m/z* of 800 Th and a precursor *m/z* window of ± 3 Th, are shown

Criterion		**Search space**	
Sequence characteristics	Precursor m/z range	Sequence Search[a)]	Spectral Search[b)]
Tryptic, with no missed internal cleavage	300–2000	3.5×10^5	1.1×10^4
	797–803	7.2×10^2	1.0×10^2
At least semi-tryptic, with at most 1 missed internal cleavage	300–2000	1.5×10^7	1.7×10^4
	797–803	7.3×10^4	1.6×10^2

a) The number of +2/+3 spectra of unmodified peptide ions meeting criteria in NIST's yeast consensus library

b) The number of +2/+3 unmodified peptide ions with sequences meeting criteria in the yeast sequence database against which SEQUEST search is performed in this study.

cursor charge cannot be experimentally determined (*e.g.* by using high mass accuracy instruments), sequence search methods often have to "guess" the precursor charge by examining the query spectrum, or in cases where such "guessing" is deemed too unreliable, perform multiple searches on the same spectrum, each time assuming a different precursor charge. Often, spectra that appear to be multiply charged are searched twice (assuming precursor charges of +2 and +3, respectively), essentially almost doubling the search space and search time. In addition, having to determine which of the search result is to be believed adds further complexity to the validation process.

Similarly, spectra of peptide ions with PTMs can be explicitly included in the library, and therefore can be recovered by spectral searching in the same way as unmodified peptide ions. Currently, the NIST libraries only contain peptides with oxidized methionine, N-terminal acetylation, and carbamidomethyl- and ICAT-modified (both non-cleavable and cleavable) cysteine, but it is a simple matter to incorporate peptides with other types of interesting modifications into the library, provided there is available data to do so. In sequence searching, PTMs are added to all applicable putative peptide sequences on the fly. While this approach offers the user great flexibility, the search space is substantially increased, sometimes making the search prohibitively costly. On the other hand, in spectral searching, only peptides that are found previously to be modified are included in the library and considered in the search, thus dramatically reducing the search space. The actual modification site of each modified peptide is also already encoded in the library entry, obviating the need to consider all the possible permutations of modified sites.

As seen in Table 2, the reduction in search space does not impair the ability of SpectraST to recover a great majority of the SEQUEST-identifiable hits, including semi-tryptic and modified peptides. Using the spectral searching approach, the user is able to obtain comparable sets of positive identifications with much better discriminating power at a fraction of the speed, as compared to a sequence searching approach. In fact, the reduction in search space is directly responsible for most of the time gain; the search space is reduced by about a factor of 500 (Table 3) while the per-query search time is reduced by about a factor of 1000.

Another advantage of spectral searching is its more precise similarity scoring. The global similarity of the matching spectra is considered, and the intensities of all peaks are

accounted for naturally. Hence, peaks that are experimentally observed to be big will also carry a proportionally bigger weight in the overall similarity score, and *vice versa*. In traditional sequence search methods, because the expected fragment-ion peak intensities are not known and not easily predicted, they are often simply ignored or given an ad-hoc value based on the ion type. This fails to take into account that the peak intensities are strong functions of the peptide sequence and the charge state, as has been reported in numerous publications [38–40]. Moreover, spectral searching needs no assumption on what fragment ions one should expect to see in the MS/MS spectra; it simply gathers similarity information from all peaks, assigned or not, that are consistently present in many experimentally observed spectra. On the contrary, traditional sequence search methods can only assume the presence of fragment ions corresponding to well described and commonly observed fragmentation mechanisms. Therefore, a good match found by SpectraST is expected to clearly stand out from random matches, so much so that it is often obvious to the naked eye that a good match is made. The same cannot be said of traditional sequence search methods, in which the theoretical, predicted spectra based on peptide sequence often do not globally resemble the experimental spectra that they are supposed to match. A few examples of matched spectra by SpectraST are shown in Fig. 3 to illustrate the improved precision and simplicity of spectral matching.

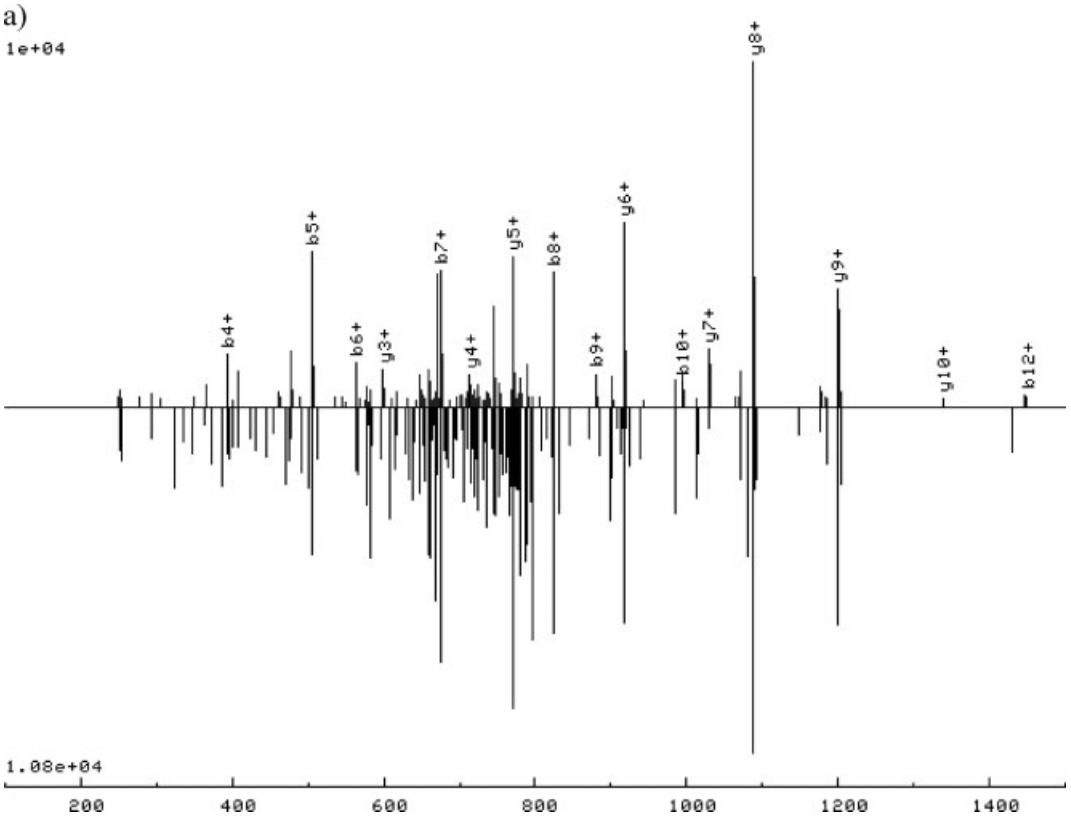
It has also been demonstrated in this study that spectral searching can sometimes yield more positive hits for the same dataset than a routine sequence search with typical search parameters, as observed for Datasets I and II (Table 2). This can be ascribed to a number of factors. First, because of the more precise similarity scoring, spectral searching is better able to discriminate between good and bad hits, resulting in higher probabilities being assigned to correct hits. Thus, a higher number of correct hits will be reported at the same confidence threshold. The more precise similarity scoring also allows spectral library searching to sometimes spot decent matches for some noisy or contaminated spectra, for which a sequence search engine might be confused (see Fig. 3). Second, because of the improved speed, one can afford to cast a wider net to search for possible hits. For instance, in this study, SpectraST searches all test datasets with a $\pm$ 3 Th precursor m/z window, which amounts to a $\pm$ 3 Da window for +1 peptide ions, a $\pm$ 6 Da window for +2 peptide ions, and a$\pm$ 9 Da window for +3 peptide ions. SEQUEST, on the other hand, searches a $\pm$ 3 Da window for all charges. (It is worthy to note that a constant m/z window is more appropriate than a constant mass window since the mass spectrometer directly measures m/z values.) Hence SpectraST is able to recover a few more positive hits that fall outside the SEQUEST precursor mass tolerance. Third, and perhaps most importantly, is that the library is constructed from the search results of multiple sequence search engines. Studies have shown that different sequence search engines have different strengths and weaknesses, and often detect largely overlapping, but not identical sets of positive hits [30, 44]. Since the library includes the high-confidence identifications by different sequence search engines, searching against such a library amounts to searching with multiple sequence search engines and combining the results, which is an approach championed by some to enhance the sensitivity and specificity of sequence searching methods [42, 43]. In other words, spectral searching implicitly allows one to take advantage the strengths of several sequence search engines, without the added time and effort, to increase the number of positive hits.

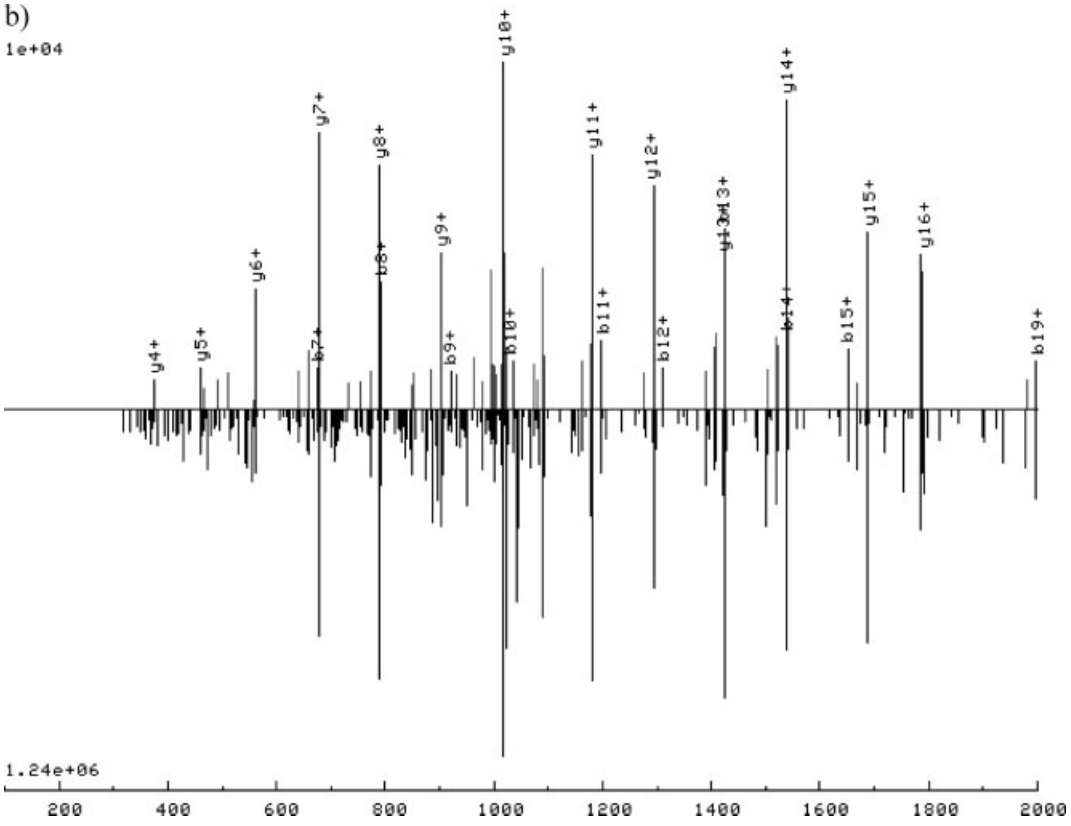
Intuitively, the precision of spectral searching allows the use of a simpler similarity scoring function. The spectral dot product is much easier to calculate than the cross correlation function that SEQUEST uses, and is shown in this study to be perfectly adequate to allow discrimination of good and bad hits. The binning of peaks into 1-Th window simplifies the spectra and further speeds up the calculation. SpectraST also does not need to sort peaks, match peaks one-by-one, or deconvolute peaks from different ion series, procedures used in some sequence search methods that are potentially computationally intensive. The simplicity of the scoring methods also contributes to the speed improvement of SpectraST over SEQUEST.

4.2 Library quality

In spectral searching, false positives can be traced to one of two causes: the scoring function assigning a decent score to a poor spectral match (often due to noisiness and/or dominant matching peaks); or a misidentified or low-quality library spectrum. The former is inevitable in any scoring method, and its extent can be estimated by conventional methods of statistical analysis such as PeptideProphet. The latter is unique to spectral library searching, and warrants careful examination. Unlike in traditional sequence searching, in which the search score is directly linked to the correspondence between the query spectrum and the peptide sequence, spectral searching assumes that the library spectrum must be identified correctly in the first place, such that one can correlate spectral similarity to the likelihood of a correct match. In other words, a library spectrum that carries an incorrect identification may lead SpectraST to misidentify a query spectrum even though the similarity score is extremely high. By the same token, library spectra that are noisy or contaminated may also create false positives by scoring very well against noisy and contaminated query spectra. This implies that the conventional means of estimating false discovery rates, which assume an inverse relationship between the search score and the false discovery rate, can be potentially misleading if the library is not of the highest possible quality.

As shown in Table 2, the number of high-confidence but incorrect identifications (as determined by manual validation) made by SpectraST is remarkably low, and none are caused by a misidentified or low-quality library spectrum.





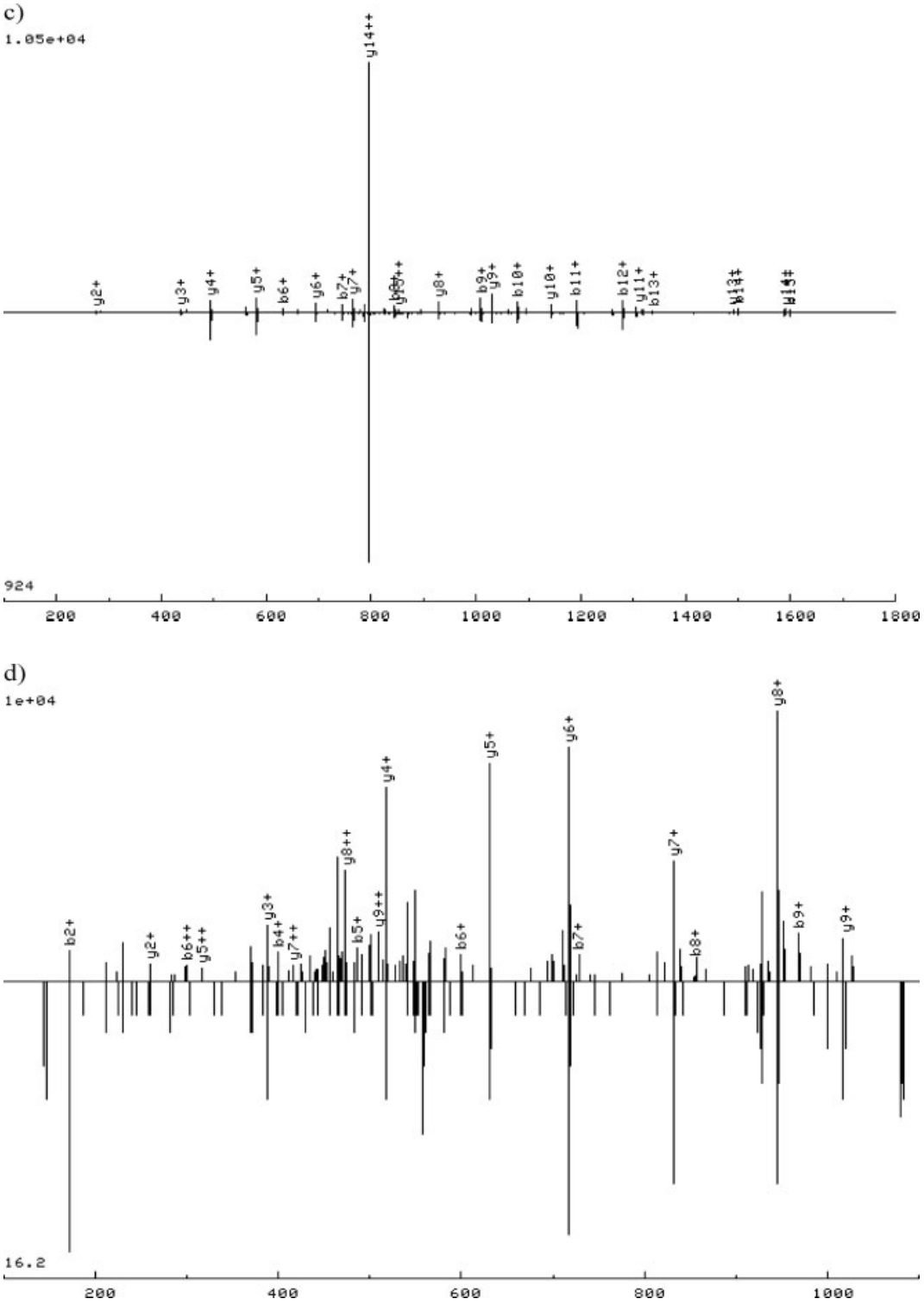


Figure 3. Example spectra matched by SpectraST. The matched library spectrum, with peak assignments, (top) is shown head-to-tail with the query spectrum (bottom), using a visualization tool available in the TPP. Both are pre-filtered, pre-scaled raw spectra. SEQUEST failed to find a positive identification for any of the four MS/MS spectra shown, despite that the peptide is in the sequence database and within the precursor mass tolerance employed. SpectraST found the following identifications: a) VGVHLGLFGNC[339]IK (+2) from Dataset I; b) AVGSLTFDENYNLLDTSGVAK (+2) from Dataset II; c) AIPGEYITYALSGYVR (+2) from Dataset III; d) VANNSLEQLK (+2) from Dataset IV. The noisiness of (a) and (b), and the dominant peak of (c) might have confused SEQUEST, whereas for (d), SEQUEST picked a semitryptic peptide over the more likely fully tryptic identification made by SpectraST.

Searching against earlier, preliminary versions of the NIST spectral library, however, resulted in a much higher number of incorrectly identified positive hits, with a majority of them mapping to problematic library spectra that have since been removed from the library by various quality filters. These problematic library spectra either resulted from false positive identification by the sequence search engines, were heavily contaminated with impurities or coeluting species, or were simply very noisy. Therefore, in order for spectral searching to be a truly reliable alternative to sequence searching, we advocate a more conservative and deliberate approach in building spectral libraries that requires strict quality control and careful validation, than those taken in refs. [21, 22].

4.3 Library coverage

The proteome coverage of the NIST spectral libraries varies from species to species, and should approximately mirror that of the data repositories from which the experimental spectra are collected. For yeast, it is estimated that about 75% of the named yeast protein coding genes are covered in the Yeast PeptideAtlas, which represents the highest degree of proteome coverage for any eukaryotic organism to date [44]. From a discovery point of view, the library coverage is of interest, because one would of course not expect spectral searching to detect peptides outside of the library coverage. Currently, the libraries have plenty of room for improvement. For instance, only peptides with oxidized methionine, N-terminal acetylation, and carbomidomethyl- and ICAT-modified (both non-cleavable and cleavable) cysteine are represented in the libraries, and are thus identifiable by spectral searching. Peptides that are seen only in more advanced instruments, such as in linear ion traps, are also lacking. However, it can certainly be expected that the libraries will gradually become more complete as more and more identified spectra, including those from peptides with other interesting modifications, are being incorporated. Optimistically, with the rapid accumulation of data, we will soon reach the point where additional shotgun proteomics experiment will add very few new peptides to the list of observable peptides. When that happens, library searching should be considered a viable option to replace sequence searching for most applications. On the other hand, most shotgun proteomics experiments today are more focused in scope, and a general high-coverage spectral library may not be as useful as one that is constructed specially for the subproteome of interest. The methodology and computational tool described in this paper should apply equally well to these specialized libraries.

4.4 Application to spectra from different instruments

The results reported here also suggest that the method is generally applicable across different instruments, although further work has to be done to determine if it is more advantageous to create separate spectral libraries for different instruments. At this stage, the spectral library used in this study is constructed exclusively from data generated on 3-D ion traps, simply because that is what is available. As shown in Fig. 1 and Table 2, the same library can also be used to search datasets acquired on linear ion traps and TOF instruments, with comparable performance in terms of discriminating good and bad hits. It should be noted that while the MS/MS spectra of the same peptide ion generated in different types of ion traps are largely similar, there are some significant differences between the spectra from ion traps and those from TOF mass detectors [45]. However, we observed that SpectraST is largely unaffected by these differences, as its performance on the Q-TOF dataset (Dataset IV) relative to that of SEQUEST was not severely affected, except for a slightly higher incidence of incorrect identifications. The same enhancement in discriminating power relative to SEQUEST, as seen in Figs. 1 and 2, shows that SpectraST's similarity scoring remains highly effective despite the minor differences in CID fragmentation patterns between the two types of instruments. Of greater concern, however, was the fact that SpectraST no longer outperforms SEQUEST in terms of the number of positive hits found (Table 2). Closer observation revealed that a great majority of the extra hits that were found by SEQUEST but missed by SpectraST were not present in the spectral library. These are probably lower-abundance peptides that are picked up only in the more advanced machines, which unfortunately, are not represented in the current spectral libraries. Hence, one can conclude that the decrease in the number of hits is largely attributable to library coverage issues, and not to a decrease in effectiveness of the searching algorithm. Therefore, once a sufficiently large body of data is generated from these advanced instruments and incorporated into the latest spectral libraries, we expect the performance of SpectraST to improve accordingly with respect to these advanced instruments.

In conclusion, the long-proposed possibility of using spectral reference libraries to identify peptides has been explored, and its effectiveness demonstrated in this paper. A functionally complete, high-throughput, and thoroughly validated search tool, SpectraST, is made readily accessible to the proteomics community as open-source software. Special care has been taken to integrate the software into the existing data processing pipeline to facilitate user adoption. Utilizing a carefully constructed reference spectral library, made possible by the rapidly growing amount of shotgun proteomics data generated and made public just in the past few years, the method was tested using four test datasets covering three different instruments, and found to offer comparable performance to SEQUEST at vastly improved speed. In particular, SpectraST consistently outperforms SEQUEST in terms of the ability to discriminate between good and bad hits, leading to improved sensitivity and false discovery rates.

Incomplete library coverage is found to be the key issue that needs to be addressed before the method can replace sequence searching methods when the discovery of new peptides or proteins (or their modified forms) remains a

major experimental goal. However, spectral searching is certainly applicable and more efficient for targeted proteomics applications in which the focus is not on discovery but rather on studying proteome segments that have been already discovered.

We would like to emphasize that the advances reported in this paper were made possible by a conscious effort by some in the proteomics community to gather and catalog peptide CID spectra and to make them readily accessible, as well as by the generosity of others who are willing to share their data. We therefore call on the proteomics community to continue to contribute their data, wherever possible, to these online data repositories, so that the spectral library can be further improved for all to use.

This project has been funded in part with federal funds the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179.

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 NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

5 References

[1] Patterson, S. D., Aebersold, R., *Nat. Genet.* 2003, *33*, 311–323.

[2] De Hoog, C. L., Mann, M., *Annu. Rev. Genomics Hum. Genet.* 2004, *5*, 267–293.

[3] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[4] Mann, M., Jensen, O. N., *Nat. Biotechnol.* 2003, *21*, 255–261.

[5] Aebersold, R., Goodlett, D. R., *Chem. Rev.* 2001, *101*, 269–295.

[6] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[7] Patterson, S. D. *Nat. Biotechnol.* 2003, *21*, 221–222.

[8] Sadygov, R. G., Eng, J., Durr, E., Saraf, A. *J. Proteome Res.* 2002, *1*, 211–215.

[9] Sadygov, R. G., Cociorva, D., Yates, J. R., III, *Nat. Methods* 2004, *1*, 195–202.

[10] MacCoss, M. J., *Curr. Opin. Chem. Biol.* 2005, *9*, 88–94.

[11] Eng, J. K., McCormack, A. L., Yates, J. R., III., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[12] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[13] Craig, R., Beavis, R. C., *Bioinformatics* 2004, *20*, 1466–1467.

[14] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., *J. Proteome Res.* 2004, *3*, 958–964.

[15] Zhang, N., Aebersold, R., Schwikowski, B., *Proteomics* 2002, *2*, 1406–1412.

[16] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[17] Domokos, L., Hennberg, D., Weimann, B., *Anal. Chim. Acta* 1984, *165*, 61–74.

[18] Stein, S. E., Scott, D. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 859–866.

[19] Ausloos, P., Clifton, C. L., Lias, S. G., Mikaya, A. I. *et al., J. Am. Soc. Mass Spectrom.* 1999, *10*, 287–299.

[20] Yates, J. R., III, Morgan, S. F., Gatlin, C. L., Griffin, P. R., Eng, J. K., *Anal. Chem.* 1998, *70*, 3557–3565.

[21] Craig, R., Cortens, J. P., Fenyo, D., Beavis, R. C., *J. Proteome Res.* 2006, *5*, 1843–1849.

[22] Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., MacCoss, M. J., *Anal. Chem.* 2006, *78*, 5678–5684.

[23] Pedrioli, P. G., Eng., J. K., Hubley, R., Vogelzang, M. *et al., Nat. Biotechnol.* 2004, *22*, 1459–1466.

[24] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. *et al., Genome Biol.* 2004, *6*, R9.

[25] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., *Nat. Biotechnol.* 2004, *22*, 471–472.

[26] Craig, R., Cortens, J. P., Beavis, R. C., *J. Proteome Res.* 2004, *3*, 1234–1242.

[27] Kristensen, D. B., Brond, J. C., Nielsen, P. A., Andersen, J. R. *et al., Mol. Cell. Proteomics* 2004, *3*, 1023–1038.

[28] Martens, L., Hermjakob, H., Jones, P., Taylor, C. *et al., Proteomics* 2005, *5*, 3537–3545.

[29] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii *et al., Nucleic Acids Res.* 2006, *34*, D655–D658.

[30] Kuster, B., Schirle, M., Mallick, P., Aebersold, R., *Nat. Rev. Mol. Cell Biol.* 2005, *6*, 577–583.

[31] Keller, A., Eng., J. K., Zhang, N., Li., X., Aebersold, R., *Mol. Syst. Biol.* 2005, *1*, 17.

[32] Marelli, M., Smith, J. J., Jung, S., Yi, E. *et al., J. Cell Biol.* 2004, *167*, 1099–112.

[33] Smolka, M., Zhou, H., Aebersold, R., *Mol. Cell. Proteomics* 2002, *1*, 19–29.

[34] Lopez-Ferrer, D., Capelo, J. L., Vazquez, J., *J. Proteome Res.* 2005, *4*, 1569–74.

[35] Martin, D. B., Gifford, D. R., Wright, M. E., Keller, A. *et al., Cancer Res.* 2004, *64*, 347–355.

[36] Yi, E. C., Lee, H., Aebersold, R., Goodlett, D. R. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2093–2098.

[37] Craig, R., Cortens, J. P., Beavis, R. C., *Rapid Commun. Mass Spectrom.* 2005, *19*, 1844–1850.

[38] Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S. *et al., Anal. Chem.* 2003, *75*, 6251–6264.

[39] Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., Gygi, S. P., *Nat. Biotechnol.* 2004, *22*, 214–219.

[40] Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L. *et al., Anal. Chem.* 2005, *77*, 5800–5813.

[41] Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P., *Nat. Methods* 2005, *2*, 667–675.

[42] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D. *et al., Anal. Chem.* 2004, *76*, 3556–3568.

[43] Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A. *et al., Proteomics* 2005, *5*, 3475–3490.

[44] King, N. L., Deutsch, E W.., Ranish, J. A., Nesvizhskii, A. I. *et al., Genome Biol.* 2006, *7*, R106.

[45] Steen, H., Mann, M., *Nat. Rev. Mol. Cell Biol.* 2004, *5*, 699–711.

