PEPTIDE IDENTIFICATION USING BOTH SPECTRUM LIBRARIES AND PROTEIN DATABASES

Marshall Bern
Palo Alto Research Center, 3333 Coyote Hill Rd.
Palo Alto, CA 94304, USA
Email: an_ bern@parc.com

The "standard" method for peptide identification by tandem mass spectrometry compares observed mass spectra to predicted mass spectra, computed from protein databases. Another approach, actively promoted in the last few years, compares observed mass spectra to previously observed, "library" spectra. In this paper we describe algorithms and software for combining the two methods in a way transparent to the user. The software applies the same dot-product scoring algorithm to theoretical and library spectra, so that results from the two types of searches are directly comparable. We show that combined database and library search outperforms either method alone.

1. INTRODUCTION

In the past 15 years, shotgun proteomics[1, 2] has emerged as the dominant paradigm for analysis of protein samples. In this method, a complex protein sample is digested with a protease such as trypsin into a still-more-complex peptide mixture, which is then separated by liquid chromatography (LC) and assayed by tandem mass spectrometry (MS/MS). The tandem mass spectra are most often identified by *database search*, that is, by comparison with predicted, "theoretical" spectra of peptides in a protein database. There are numerous search tools for this comparison; the most popular ones are Mascot[3], SEQUEST[4], and X!Tandem.[5] Tandem spectra of peptides, however, are not completely predictable, as the fragment ions and their relative intensities (peak heights) depend upon the instrument parameters and the peptide chemistry in some complicated and poorly understood way. The *spectrum-library* approach offers an alternative to database search; the idea here is to compare each unknown spectrum to previously observed well-identified spectra, rather than to theoretical spectra. This approach offers shorter search times, because the number of frequently observed peptides (tryptic or otherwise) from some organism will typically be at least 100 times smaller than the total number of peptides in the proteome. In the long run, once spectrum libraries offer sufficient coverage, the approach should also offer greater sensitivity, because an unknown spectrum should more closely match an observed spectrum than a theoretical spectrum of the correct peptide.

Here we propose a hybrid method that enables a graceful transition to the spectrum-library approach. The spectrum library can be built in-house in the course of biological studies, using one MS set-up, rather than relying on public-access spectrum libraries from a variety of set-ups. Database-search and spectrum-library scoring use exactly the same peaks (a-, b-, and y-ions, neutral losses, and so forth), and only the predicted intensities change, so that the system can compare the scores from the two approaches directly. The system automatically chooses the best match, and if the match exceeds a score threshold, it optionally records the spectrum for subsequent searches. Indeed our hybrid tool consists of two separate programs: ByOnic[6] for database search and a new program called LyBrary for library search. The hybrid approach does not offer the speed-up of the pure spectrum-library approach, but it offers greater sensitivity because it can identify a peptide the first time it is observed.

We tested our hybrid method using two biological samples (Jurkat cell lysate and mouse blood plasma) for which we had numerous technical replicates. We address the following questions: How much sensitivity improvement is possible with the spectrum-library approach? How much would database search improve with accurate intensity prediction? Can the spectrum-library approach improve the limit of detection (that is, the lowest concentration at which proteins are reliably detected)?

2. BACKGROUND

We start with an example peptide AGFAGDDAPR^{++} with fragmentation spectrum shown in Figure 1. The most important fragment ions correspond to prefixes and suffixes of the peptide sequence, and these ions are conventionally named a-, b-, and c-ions and x-, y-, and z-ions, respectively. The most common ions produced by CID fragmentation (collision-induced dissociation) are the b- and y-ions. The ion number indicates the number of residues, so that the b3 ion from the peptide AGFAGDDAPR^{++} is AGF^{+} and the y6 ion is GDDAPR^{+}. CID also produces some a-ions; the a4 ion is essentially the b4 ion with a loss of carbon monoxide (28 Daltons).

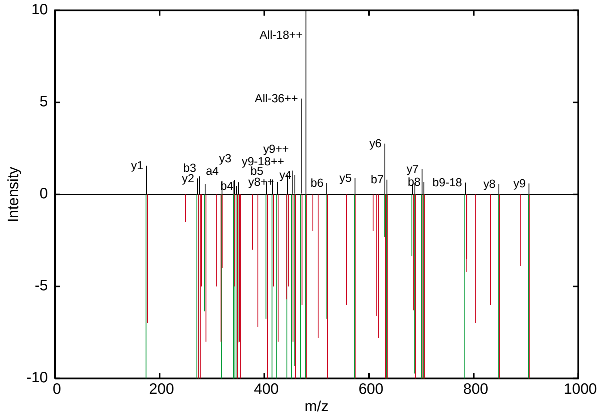


Fig. 1. A tandem-MS spectrum of AGFAGDDAPR^{++} from actin, acquired on a Thermo LTQ ion-trap instrument. Peaks growing up show relative intensity (percent of total ion current); 0.5 has been added to each peak for visibility. Peaks growing down show ByOnic's rank-based $I_{obs}(s)$ values and rank-based $I_{ref}(s)$ predictions. for the same spectrum. Rank "flattens" the peak intensities, taking them partway to presence/absence. Unobserved peaks like a3 at 248 and b9 at 802 Da have negative $I_{obs}(s)$ values (not shown).

The rationale for the spectrum library approach is the following: given the sequence AGFAGDDAPR, it is hard to predict that y6 and y7 would be the most intense y-ions, that b9−18 (water loss from b9) would be more intense than b9 (actually missing from Fig. 1), and that a4 would be the only prominent a-ion, yet these seem to be stable features of low-energy CID MS/MS spectra, repeated in other spectra of the same peptide as shown in Figure 2. Some other features, however, are apparently not so stable, for example, All−18^{++} and All−36^{++} (the full peptide with water losses) are prominent in Fig. 1 but not in Fig. 2. As shown in Fig. 3, different instrument types and different precursor ion charges give different intensity patterns. In this work we limit attention to a single type of instrument, which would be the likely scenario for in-house library search. In order to maximize our library coverage, we also limit attention to +2 precursors, the most important charge state for ion-trap instruments employing CID fragmentation.

The spectrum-library approach was proposed by Yates et al.[7] in 1998, but the first large-scale efforts are the Global Proteome Machine (GPM) by Beavis et al.[8, 10] and the PeptideAtlas project by Desiere, Aebersold, et al.[12], both of which now have 10^5 – 10^6 well-identified spectra. (An early effort by NIST focused mostly on small molecules.[13]) The initial annotations for GPM are made by X!Tandem and the library search software is called X!Hunter. Now that the GPM library includes some 400,000 "proteotypic" peptides, GPM also offers a mode called X!P3 in which the library search identifies the proteins[14], and then X!Tandem makes a broader search, including modifications of observed peptides and other peptides from the same proteins. The search tool associated with the PeptideAtlas project is SpectraST,[15] part of the Trans Proteomic Pipeline from the Institute for Systems Biology. Yet another project built library-search software called BiblioSpec,[16] and demonstrated cross-instrument identification using two types of ion trap, Thermo LCQ and LTQ.

All of these efforts use very simple scoring algorithms. For example, X!Hunter uses only the 20 largest peaks in a library spectrum.[10] BiblioSpec and SpectraST use more peaks, but they both round mass-over-charge (m/z) measurements to the closest integer, rather than using a settable mass tolerance. In all three cases, the peaks are not necessarily identified peaks; the software relies on averaging multiple spectra to remove noise and improve m/z measurements. (Spectrum clustering[11] also averages spectra, but it averages them prior to identification.) Because LyBrary uses only identified peaks, it records exact (theoretical) m/z values and discards unexplained "noise" peaks. Exact m/z values are of course a great advantage, especially in identifying high-resolution spectra (QTOF, Orbitrap, FTICR) using low-resolution library spectra (ion-trap). This

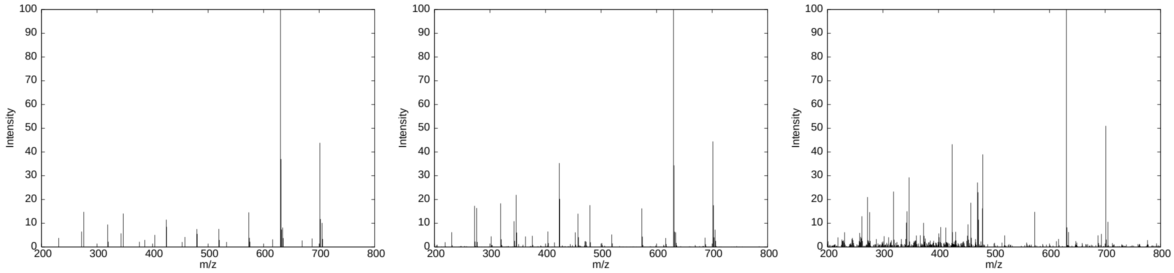


Fig. 2. More Thermo LTQ MS/MS spectra of AGFAGDDAPR^{++} from three different chromatographic runs and three different organisms (human, mouse, and *C. elegans*). In all cases the tallest peaks include y2 at 272, b3 at 276, a4 at 319, y3 at 343, b4 at 347, y8^{++} at 424, y5 at 573, y6 at 630 Da, and y7 at 701. All-18^{++} and All-36^{++} at 480 and 471 are most prominent in the rightmost spectrum.

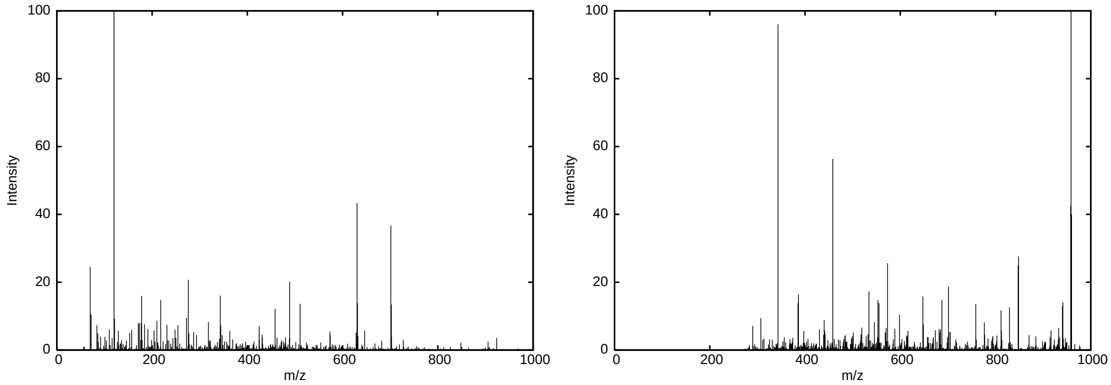


Fig. 3. Agilent QTOF spectrum of AGFAGDDAPR^{++} on the left, and Thermo LTQ spectrum of AGFAGDDAPR^{+} on the right. In the QTOF spectrum, prominent peaks include b3 at 276, b4 at 347, All^{++} (without water loss) at 489, y6 at 630, and y7 at 701. The peaks at 70 and 120 (the tallest) are immonium ions from proline and phenylalanine, rarely observed in ion-trap spectra. In the spectrum of the singly-charged peptide, the tallest peaks are y3 at 343, y4 at 458, y5 at 573, y8 at 848, and All-18^{+} at 958.

design choice, however, has a potential downside: maybe there are unusual—and hence unidentified—signal peaks, e.g., doubly charged b-ions, characteristic of the peptide, that LyBrary does not consider. We address this question in Section 5.

In previous studies, the library approach has always given much greater speed, but because of incomplete libraries has not always given greater sensitivity.[9] Craig et al.[10] report 1000-fold speed improvement and 50% better sensitivity on a sample containing only bovine serum albumin; in this case the spectrum library had essentially complete coverage. Lam et al.[15] report that on four yeast data sets, SpectraST always gave a better sensitivity/specificity tradeoff (ROC curve) than SEQUEST, but SEQUEST gave a larger number of high-probability matches on two of the data sets. Frewen et al.[16] report that with their most complete library (for *E. coli*), BiblioSpec could make 91% of the identifications made by SEQUEST. Library coverage of yeast and *E. coli* is much more complete than for organisms with larger proteomes,[15] so it is fair to say that coverage remains the major limitation of the spectrum-library approach. We designed LyBrary to fill this need: library search that works even without high coverage.

3. ALGORITHMS

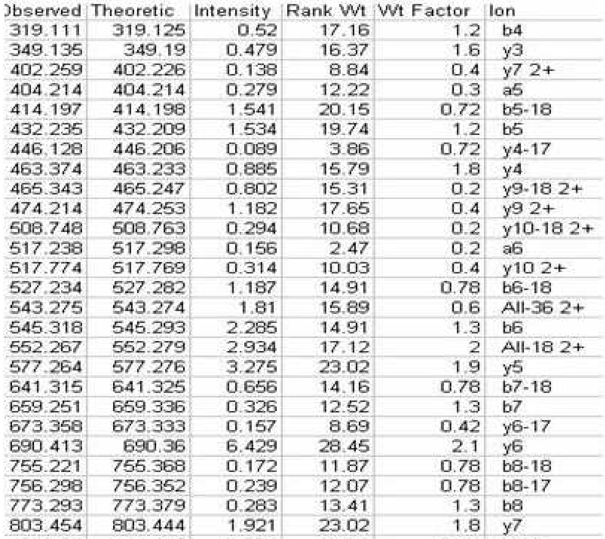
LyBrary is built on top of ByOnic,[6] a conventional database search program. Given an unknown spectrum, ByOnic (respectively, LyBrary) scores *candidate* peptides using a scoring function that essentially takes the dot product of two vectors: the observed spectrum and the theoretical (library) spectrum. Specifically, given an observed spectrum S

and a theoretical or library spectrum T, the score is

$$\text{Score}(S,T) = \sum_{\substack{s\in S \\ t\in T}} \text{Acc}(s,t) \cdot \text{I}_{\text{obs}}(s) \cdot \text{I}_{\text{ref}}(t), \quad (1)$$

where s is a peak in S and t a peak in T. A peak is a pair of numbers: an m/z measurement and an intensity, roughly proportional to ion count.

The most informative ions in MS/MS spectra of peptides are b- and y-ions, which correspond to prefixes and suffixes of the amino acid sequence. The mass of a b-ion is the sum of the residue masses along with 1.007 Dalton for the mass of a proton; a y-ion includes the residue masses along with a proton and water (18.011 Da). Other commonly observed ions are water and ammonia losses from b- and y-ions, denoted by −18 and −17 in Figure 4; doubly charged y-ions, such as "y9 2+"; and neutral losses from the precursor ion such as "All-36 2+" and "All-18 2+".



)bserved	Theoretic	Intensity	Rank Wt	Wt Factor	Ion
319.111	319.125	0.52	17.16	1.2	b4
349.135	349.19	0.479	16.37	1.6	y3
402.259	402.226	0.138	8.84	0.4	y7 2+
404.214	404.214	0.279	12.22	0.3	a5
414.197	414.198	1.541	20.15	0.72	b5-18
432.235	432.209	1.534	19.74	1.2	b5
446.128	446.206	0.089	3.86	0.72	y4-17
463.374	463.233	0.885	15.79	1.8	y4
465.343	465.247	0.802	15.31	0.2	y9-18 2+
474.214	474.253	1.182	17.65	0.4	y9 2+
508.748	508.763	0.294	10.68	0.2	y10-18 2+
517.238	517.298	0.156	2.47	0.2	a6
517.774	517.769	0.314	10.03	0.4	y10 2+
527.234	527.282	1.187	14.91	0.78	b6-18
543.275	543.274	1.81	15.89	0.6	All-36 2+
545.318	545.293	2.285	14.91	1.3	b6
552.267	552.279	2.934	17.12	2	All-18 2+
577.264	577.276	3.275	23.02	1.9	y5
641.315	641.325	0.656	14.16	0.78	b7-18
659.251	659.336	0.326	12.52	1.3	b7
673.358	673.333	0.157	8.69	0.42	y6-17
690.413	690.36	6.429	28.45	2.1	y6
755.221	755.368	0.172	11.87	0.78	b8-18
756.298	756.352	0.239	12.07	0.78	b8-17
773.293	773.379	0.283	13.41	1.3	b8
803.454	803.444	1.921	23.02	1.8	y7

Fig. 4. A screenshot of ByOnic's scoring report for the peptide SSGSLLNNAMK^{++}. The first two columns give observed and theoretical m/z values. The third column shows relative intensity (percent of total ion current); the fourth column shows rank-based intensity; and the fifth column, "Wt Factor" shows ByOnic's prediction of the rank-based intensity in arbitrary units. The score is the dot product of columns 4 and 5, with each summand in the dot product weighted according to the closeness of the match in the first two columns.

To be included in the sum for Equation (1), the m/z measurements for peaks s and t must match within a user-defined tolerance, typically 0.4 Daltons per charge (Thompsons) for an ion-trap instrument. Acc is a function that returns 1.0 if the m/z measurements match exactly and drops to 0.0 in a bell-shaped curve as the difference between the measurements increases to the user-defined tolerance.[6] ByOnic uses a rank-based function for weighting observed intensity:

$$\text{I}_{\text{obs}}(s) = A/(2 + \text{Rank}(s)) + B \cdot \text{RelI}(s) - C \quad (2)$$

After correcting for isotope peaks and varying instrument sensitivity across the m/z range, Rank(s) is 1 for the tallest peak in S, 2 for the second tallest, and so forth. RelI gives peak s's fraction of the total ion current in S, and A, B, C are empirically chosen positive constants, such that the Rank term dominates over the RelI term. C is included so that a predicted but unobserved peak makes a small negative contribution to Score(S,T); this helps level the comparison of candidates with different numbers of predicted peaks. See Figures 1 and 4.

ByOnic includes rule-based "expert system" code (following the lead of Zhang[17]) that attempts to predict $\text{I}_{\text{obs}}(s)$ based on the peptide sequence, charge state, instrument type, and so forth; this prediction is used as $\text{I}_{\text{ref}}(s)$. The expert system includes chemical knowledge such as the facts that y5 – y9 ions tend to be strong, that a2 and a4 are the most likely a-ions, and that cleavage is likely on the C-terminal side of proline (hence the strong y1 peak in Figure 1). The expert system is of course obviated by the library approach, and for LyBrary we tried various choices for $\text{I}_{\text{ref}}(s)$ as described below. ByOnic's rank-based $\text{I}_{\text{obs}}(s)$ and $\text{I}_{\text{ref}}(t)$ were chosen for robustness; with the library approach more aggressive intensity weighting is possible.

Scorers in other search tools differ in various ways. For example, Mascot uses 0 / 1 values for both $\text{I}_{\text{obs}}(s)$ and $\text{I}_{\text{ref}}(t)$, using the intensities only to decide where to cut off the observed peak list and which "peak series" (a-, b-, y-ions, etc.) to score. X!Tandem uses relative intensity for $\text{I}_{\text{obs}}(s)$ (normalized to the tallest peak rather than the total ion current) and unit intensities for $\text{I}_{\text{ref}}(t)$. SEQUEST uses relative intensity within an m/z band for $\text{I}_{\text{obs}}(s)$ and unit intensities for $\text{I}_{\text{ref}}(t)$. None of the major database search engines include an Acc term, but several de novo sequencers do.[18]

3.1. Library Spectrum Intensity Weights

LyBrary's scorer differs from ByOnic only in the weights $\mathrm{I}_{\mathrm{ref}}(s)$. We explored three possibilities. In all cases, we normalized the Euclidean length of the vector of $\mathrm{I}_{\mathrm{ref}}(s)$ values to agree with the length of the vector predicted by ByOnic's expert system. LyBrary scores are thus comparable to ByOnic scores, but we expect LyBrary scores for correct matches to be somewhat higher, because the peak intensities from library spectra should be more accurate than those from theoretical spectra, that is, the angle between the observed and library spectra, regarded as vectors, should be closer to zero. For a given peptide p, let $S(p)$ denote the library spectrum matching p with highest ByOnic score. We also tried using the average of all library spectra for p, but the results were almost identical.

(1) **Rank-based intensities.** In this option, LyBrary set $\mathrm{I}_{\mathrm{ref}}(s)$ equal to $\mathrm{I}_{\mathrm{obs}}(s)$ in $S(p)$.

(2) **Relative intensities.** LyBrary set $\mathrm{I}_{\mathrm{ref}}(s)$ to the relative intensity of s in $S(p)$.

(3) **Square root.** LyBrary set $\mathrm{I}_{\mathrm{ref}}(s)$ to the square root of the relative intensity of s in $S(p)$.

3.2. Software Architecture

For a top-scoring peptide-to-spectrum match (PSM), ByOnic writes out a scoring report such as the one shown in Figure 4. The report details each peak down to rank 20 – 200 (depending upon peptide mass and the number of peaks in the spectrum), its intensity, its predicted intensity (if any), and so forth. The report also includes predicted peaks not observed and large unexplained peaks (at least 0.5% of the total intensity in the MS/MS spectrum).

LyBrary really consists of two programs: Archive and LibScore. Archive parses and reformats ByOnic's scoring reports into a spectrum library. The library is organized by precursor mass, so that one file includes all peptides with precursor mass 1600 – 1700 Da, another includes all with precursor mass 1700 – 1800, and so forth. In order to score an unknown spectrum, LibScore opens the relevant library files and scores all the peptides with the right precursor mass. For example, if the unknown spectrum has precursor mass 1708.78, LibScore run without modifications (and precursor mass tolerance at most 8 Da) would open only the file with masses 1700 – 1800 Da. If oxidized methionine were enabled, then LibScore would also open the file with masses 1600 – 1700 Da in order to score peptides of (unmodified) mass $1708.78 - 15.995$ with one M[+16], peptides of mass $1708.78 - 31.990$ with two M[+16]'s, and so forth. LibScore, like ByOnic, sets reasonable limits on the numbers of modifications per peptide, at most two M[+16]'s, at most one sodiation, and so forth.

LibScore uses ByOnic's predicted peak intensities to normalize the previously observed peak intensities; these normalized intensities, an equal-length but different-direction vector, then substitute for ByOnic's predictions in the scoring subroutine. LibScore writes its output in the same plain-text format as ByOnic, so that subsequent programs like ComByne[19] (a peptide-to-protein integration tool) can use output from either program, or even concatenated files produced by any combination of searches from either tool. Our hybrid approach actually consists of a run of ByOnic and a run of LibScore (on a previously built library), a concatenation of the two outputs, and then a run of ComByne to produce the final report, which we generally read into an Excel spreadsheet. ComByne always picks the highest scoring PSM for each spectrum, regardless of the origin (ByOnic or LibScore) of the PSM.

LibScore actually has two ways to make modification identifications: it can use either a previous observation of the same peptide with (exactly) the same modifications, or a previous observation of the same peptide without modifications. (In all the experiments reported in Section 4, however, we used only the latter path.) LibScore cannot currently use an observation of DNSTM[+16]GYMMAK to identify DNSTM[+16]GYM[+16]MAK. When LibScore uses an unmodified library peptide to identify a modified unknown, it assumes that the peak intensity pattern is unchanged, only the masses are shifted. This assumption is reasonable for most low-mass modifications, and has been validated in Bandeira's work on spectral networks analysis,[20] but the assumption is not wholly true for M[+16], which sometimes loses 64 Da, and is quite untrue for phosphorylation (S[+80] and T[+80]), which has a prominent neutral loss of

98 Da. In the case that both modified and unmodified peptides are in the library, LibScore obliviously scores the candidate match both ways and retains only the higher score.

4. EXPERIMENTAL RESULTS

We used two data sets, described below. Due to lack of sufficient data, we did not attempt to use a spectrum library built for one type of instrument to make identifications on another type, nor did we attempt to identify the same peptide in different charge states.

(1) **Jurkat Cell Lysate.** Five LC-MS/MS runs on a Thermo LTQ Orbitrap, with Orbitrap single-MS and LTQ MS/MS. These runs are essentially technical replicates, differing only in details of the data acquisition (e.g., whether the top 5 or 10 peaks in the single-MS scan were selected for MS/MS).

(2) **Mouse Blood Plasma.** Six LC-MS/MS runs on a Thermo LTQ of MARS-depleted mouse blood plasma, spiked with low concentrations of 13 soluble human proteins. (MARS is "multiple affinity removal system" for removing serum albumin and 5 other abundant proteins in order to improve dynamic range.) The spiked proteins were at two different concentrations, 1 μg/ml and 10 μg/ml, with 3 technical replicates at each concentration.

4.1. Complete Coverage

We used the Jurkat sample to test how much sensitivity gain is possible with the spectrum library approach, assuming a best-case scenario in which the library has complete coverage of all the peptides in the sample. In this computational experiment, we first used ByOnic for a conventional database search using the IPI human protein database with about 49,000 protein sequences. We included reversed protein sequences as "decoys" in order to measure false positive and false discovery rates.[21] We ran searches with and without modifications enabled; the modifications considered were oxidation (M), deamidation (N and Q), pyro-glu (N-terminal E and Q), acetylation (C, S, K and N-terminus), disulfide bridge (since the sample had no cysteine treatment), carbamylation (K, R and N-terminus).

Orbitrap precursor masses were good to about ±7 ppm, so we judged 20 ppm to be a safe tolerance for precursor masses. High precursor mass accuracy alone is very informative and hence can mask differences between scoring algorithms, so we also ran searches with a "fictitious" precursor mass tolerance of 5 Da.

Table 1. Numbers of matches to 30 abundant peptides for database search (ByOnic) and spectrum library (LyBrary), using either 20 ppm or 5 Da precursor mass tolerance. We report average numbers over the 5 runs, for two different searches: a no-modification search, and a search with 13 modifications enabled. The first four lines of the table show that library search with rank-based weighting gives about 10% – 20% greater sensitivity than database search with rank-based weighting. The last three lines show that rank and square-root weighting beat relative intensity.

Search	# No mod	# Mods
ByOnic (5 Da)	133.6	142.0
ByOnic (20 ppm)	137.4	146.4
LyBrary (rank, 5 Da)	147.0	155.8
LyBrary (rank, 20 ppm)	157.6	172.6
LyBrary (sqrt root, 5 Da)	145.6	154.0
LyBrary (relative, 5 Da)	133.4	142.2

The spectrum library for run 1 included all high-scoring spectra from runs 2–4, even those that matched decoy peptides. This "aggressive" policy allowed us to generate the library automatically, without any manual curation. The library for run 1 included 6804 peptides, including 348 reversed peptides. We used a similar leave-one-out approach for all 5 runs. For both database search and library search, we counted the number of matches (of any score) to 30 abundant tryptic peptides found in all 5 runs, all of which are true (non-reversed) peptides. This simulates the case of complete coverage, because we only count peptides represented in all spectrum libraries. Matches to the top 30 peptides are presumed correct, because the top 30 peptides represent less than 0.5% of the spectrum library and less than 0.01% of the protein database. We use the number of matches to the top 30 peptides as a proxy for the overall number of correct matches ("sensitivity" or "recall"), because it is hard to validate matches to low-ranking peptides or proteins in complex natural samples.

Table 2. Numbers of matches to the top 100 proteins in run 1 of the Jurkat cell lysate for database search (ByOnic), spectrum library (LyBrary), and combined search. The spectrum library was built using runs 2 – 5.

Search	# Spectra	# Mod Spectra	# Unique	Coverage of Top 3 Proteins		
ByOnic	1567	216	1241	32.6%	44.7%	20.7%
LyBrary (rank-based)	1717	208	1206	24.6%	30.0%	16.5%
LyBrary (relative intensity)	1704	195	1184	25.6%	30.0%	17.4%
LyBrary (sqrt relative intensity)	1715	194	1199	24.6%	30.0%	16.5%
ByOnic + LyBrary (rank-based)	1846	255	1406	34.1%	43.8%	22.1%

As shown in Table 1, library search outperforms database search by a small amount. Rank-based and square-root of relative intensity outperformed raw relative intensity, which gave the top few peaks too much consideration. Top peaks such as "All-18 2+" (the entire peptide, doubly charged, with one water loss) do not discriminate between candidate peptides of the same precursor mass very effectively because most peptides can lose water. Frewen et al.[16] also found that using the square-root gave better results. Surprisingly, high precursor mass accuracy gave LyBrary a bigger boost than it gave ByOnic. We attribute this to the sizes of the library and the database. 20 ppm precursor accuracy typically limits the number of library possibilities to about 10 (100 if modifications are enabled) so that even very poor spectra with few fragment peaks can be identified, but the number of database possibilities is still on the order of 10^4 (or 10^5 with modifications).

The experiment described in this section also gives a partial answer to another of our questions. If we continue to develop ByOnic's intensity prediction, without making any other improvements, we can expect to achieve only moderate gains in sensitivity, at most about 20% more identifications at the spectrum level.

4.2. Incomplete Coverage

In this section, we drop the complete-coverage assumption and address a more realistic scenario in which the spectrum library is built from a small number of similar samples and hence has incomplete coverage. We again used a leave-one-out approach, with the spectrum library for run 1 built using runs 2 – 5, but now we consider not just the top 30 peptides, but all peptides in top proteins. Table 2 reports the number of spectra matched to peptides from the top 100 proteins in run 1 of the Jurkat cell lysate, for ByOnic, LyBrary, and a hybrid run, which ran both ByOnic and LyBrary as explained above. We used 20 ppm precursor mass tolerance and searched tryptic and semitrypic +2 peptides with the same modification list as above. The top-100 protein list was compiled by an initial run of ByOnic and ComByne, but a protein list compiled using LyBrary and ComByne is not much different.

As shown in the table, ByOnic alone matched fewer spectra to the top 100 proteins, but found more unique peptides and more modified peptides. The difference in performance was modest—LyBrary gave less than 10% more matched spectra. Runs 2 – 5 (not shown) give similar results, with LyBrary's edge varying from 9% to 12%. On all runs, the hybrid approach gave the best results on all measures, with an edge of 15% to 21% over ByOnic alone.

On this sample, it appears that a set of four runs of exactly the same material on exactly the same proteomics set-up gives decent but not complete coverage. LyBrary found almost as many unique peptides as database search, but its coverage of the top 3 proteins fell short of ByOnic's coverage. This is consistent with MS/MS studies[22] in which repeat runs consistently yield some new peptide and protein identifications. Table 3 gives our own study of this type. ByOnic can make these new identifications, but LyBrary alone cannot identify peptides not represented in the library. Conversely, LyBrary's identifications that were not found by ByOnic often matched poor spectra to already-identified abundant peptides at the beginning or end of their elution pulses, or matched poor spectra to modified versions of abundant peptides. The first type of extra identification is not especially useful, but the second type may be quite important if the modifications are biologically active. Because ByOnic and LyBrary have different strong points, the hybrid approach gives the best overall analysis of the data.

Table 3. Numbers of peptides and proteins found by combining ByOnic's database-search identifications from 6 repeat runs of the mouse blood plasma sample. The peptide number is the number of distinct peptides in the 3 most abundant proteins (Alpha-2-macroglobulin, complement C3, and murinoglobulin), with different modification states considered distinct. Coverages gives percent coverage for the top 3 proteins. The protein number is the number of proteins ranked above the 3rd highest reverse.

Runs	# Peptides	Coverages	# Prots
1	281	58% 61% 50%	139
1–2	353	61% 67% 54%	150
1–3	388	62% 69% 57%	154
1–4	450	67% 75% 60%	161
1–5	506	67% 77% 61%	167
1–6	535	68% 79% 61%	167

What about protein sensitivity? We used ComByne[19] to integrate peptide identifications into protein identifications. ByOnic finds 269 proteins at 1% FDR as measured by the number of reversed proteins, that is, ComByne's ranked list put the third highest reversed protein at rank 272. LyBrary alone finds fewer proteins—typically around 200—and the number is unstable and hard to estimate because the spectrum library contains only a few reversed peptides for gauging FDR. The best result was obtained by ByOnic plus LyBrary (rank-based) using a conservative spectrum library, which included only extremely high-scoring spectra and no spectra matching reversed peptides. This combination found 296 proteins at 1% FDR.

4.3. Improved Limit of Detection?

Sample 2 consists of mouse blood plasma, spiked with either low (1 μg/ml) or high (10 μg/ml) concentrations of 13 soluble human proteins. We built a spectrum library using 3 chromatographic runs with high concentrations, where the spiked proteins are fairly easy to detect, to test whether the spectrum library approach would help find the spiked proteins in the lower concentration samples, where many of the spiked proteins are missed by ByOnic (and every other database search engine we have tried). The question we would like to answer is whether low-abundance proteins are missed because they have no MS/MS spectra or because their MS/MS spectra are too poor to be identified.

Again we used an aggressive spectrum library including both forward and reversed proteins. We searched the spectra for tryptic and semitryptic peptides, assuming only +2 precursor charge, without any modifications enabled. For the mouse blood plasma sample, which was taken on a Thermo LTQ instrument (without Orbitrap), the charge cannot be reliably determined in advance, but previous searches on this sample showed that +2 precursors predominate and that less than 10% of the peptides carry modifications. Overall results were consistent with the earlier experiments. ByOnic alone matched 1516 spectra representing 894 unique peptides to the top 100 proteins. LyBrary alone matched 1582 peptides representing 614 unique peptides. The hybrid approach matched 1778 spectra representing 916 unique peptides. The top-ranked reversed protein had rank 120 for ByOnic, 110 for LyBrary, and 113 for ByOnic + LyBrary. All three approaches found the same 8 spiked proteins with about the same confidence, in the sense that for all three approaches 5 of the 8 proteins were reliably identified (above the top-ranking reverse protein) and 3 were in the gray zone (below the top-ranking reverse, but well above the noise zone, the point in the ranked list at which half of the protein identifications are reversed).

This result suggests that although the hybrid approach does indeed offer higher sensitivity at the spectrum level (17% more matches to the top proteins, consistent with the results on the Jurkat sample), it does not offer commensurate improvement at the protein level. We believe that low-abundance proteins, at least in single-LC runs of blood plasma, are most often missed because they have no MS/MS spectra at all. Blood plasma is very rich in peptides, so that the usual top-5 or top-10 approach to shotgun proteomics (picking only the 5 or 10 biggest single-MS peaks for MS/MS) will fail to acquire MS/MS spectra for many low-abundance proteins.

5. DISCUSSION

Library search potentially improves sensitivity in two ways. First, it provides a focused database containing only observable "proteotypic" peptides.[9] Even our aggressive library-building strategy gave libraries of less than 10,000 peptides, but the full IPI human protein database contains on the order of 10^6 tryptic peptides. Second, library search gives more accurate intensity predictions for fragment peaks, because the predictions are based on previous observations rather than general principles. Thus one might hope that library search would give very large sensitivity gains over database search. Unfortunately this does not seem to be the case. Craig et al.[10] reported 50% greater sensitivity in the complete-coverage case; and we report much less, only 10% – 20% improvement, in Table 1.

Why did our library search program fall short? We think that ByOnic provides a better baseline than X!Tandem and leaves less room for improvement. X!Tandem does not predict peak intensities, and does not score neutral losses nor doubly charged ions, which are often among the top 20 peaks. Even though ByOnic's expert system cannot predict relative intensity very accurately, it can predict rank-based intensity reasonably well. ByOnic's predictions of rank-based intensity have median correlation coefficient 0.559 with observed rank-based intensities. This number is the correlation of columns 4 and 5 (Rank Wt and Wt Factor) of the scoring reports (Figure 4), with the median taken over 920 unique peptides from the blood plasma sample. Observed and re-observed rank-based intensities—the analogous statistic for LyBrary—give median correlation coefficient 0.780. The correlation coefficient for observed and re-observed relative intensities is higher (around 0.9) because a few strong peaks dominate.

Finally, we return to the question of whether the library approach should use spectra containing all observed peaks as in GPM and PeptideAtlas or only identified peaks. In this work we used only identified peaks for compatibility with database search. After running our computational experiments, we tentatively conclude that this choice did not adversely affect the spectrum library approach. The number of large unexplained peaks (greater than 0.5% of total ion current) is not overwhelming. For example in the aggressive library for run 4, there are about 28,000 large unexplained peaks and 174,000 explained peaks (large and small) in 6606 library spectra. Predicted but unobserved peaks are common—129,000 in the same library.

What are the large unexplained peaks? Are they ion types not considered by ByOnic? For CID ion-trap spectra of +2 precursors, ByOnic considers the following peaks (and scores them if they fall into the right m/z range): b1 – b_{n-1}, where n is the length of the candidate peptide, y1 – y_{n-1}, along with single water and ammonia losses from these ions. ByOnic also scores doubly charged y-ions from y4 to y_{n-1}, the a-ions a2 – a8, and single and double neutral losses from the precursor ion. ByOnic does not score the following ions: internal fragments, a-ions larger than a8, doubly charged b-ions, double neutral losses (e.g., two waters or one water and one ammonia) from b- or y-ions, and triple neutral losses from the precursor ion. We have observed all of the ignored ions just named, but statistics on our training sets suggest that these ions are infrequent and not worth scoring. In fact, ByOnic deliberately "over-scores", including some infrequent peaks (b1, a3, water losses from small y-ions) just for completeness.

In manual inspection of a small number (10s) of library spectra, the single most common explanation for large unexplained peaks was unrecognized isotope peaks. ByOnic requires fairly tight tolerances on m/z and intensity in order to dismiss a peak as an isotope peak of another (explained) peak. If a peak does not fit within these tolerances, then it is considered unexplained. The second most common explanation was no explanation—expert inspection could find no ion from the identified peptide that would explain the peak.

ACKNOWLEDGMENTS

The author thanks Don Kirkpatrick of Genentech for the Jurkat spectra, Christopher Becker of PPD, Inc. (Menlo Park) for the spiked mouse plasma spectra, and the NIH for grant GM085718, which supported this work.

References

1. D.C. Liebler. *Introduction to Proteomics: Tools for the New Biology.* Humana Press, 2002.
2. G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology.* MCC Press, 2003.
3. D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20 (1999), 3551-3567.
4. J.K. Eng, A.L. McCormack, and J.R. Yates, III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5 (1994), 976–989.

5. R. Craig and R.C. Beavis. X!TANDEM: matching proteins with mass spectra. *Bioinformatics* 20 (2004), 1466–1467.
6. M. Bern, Y. Cai, and D. Goldberg. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* 79 (2007), 1393–1400.
7. J.R. Yates, III, S.F. Morgan, C.L. Gatlin, P.R. Griffin, J.K. Eng. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* 70 (1998), 3557–3565.
8. R. Craig, J.P. Cortens, and R.C. Beavis. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Research* 3 (2004), 1234–1242.
9. B. Domon and R. Aebersold. Challenges and opportunities in proteomics data analysis. *Mol. Cell. Proteomics* 5.10 (2006), 1921–1926.
10. R. Craig, J.P. Cortens, D. Fenyö, and R.C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Research* 5 (2006), 1843–1849.
11. D.L. Tabb, M.J. MacCoss, C.C. Wu, S.D. Anderson, and J.R. Yates, III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.* 75 (2003), 2470–2477.
12. F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S.N. Loevenich, and R. Aebersold. The PeptideAtlas project. *Nucleic Acids Research* 34 (2006), D655–D658.
13. P. Ausloos, C.L. Clifton, S.G. Lias, A.I. Lias, S.E. Stein, D.V. Tchekhovskoi, O.D. Sparkman, V. Zaikin, and D. Zhu. The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrometry* 10 (1999), 287–299.
14. R. Craig, J.P. Cortens, and R.C. Beavis. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* 19 (2005), 1844–1850.
15. H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7 (2007), 655–667.
16. B.E. Frewen, G.E. Merrihew, C.C. Wu, W.S. Noble, and M.J. MacCoss. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* 78 (2006), 5678–5684.
17. Z. Zhang. De novo peptide sequencing based on divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal. Chem.* 76 (2004), 6374–6383.
18. M. Bern and D. Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comp. Biology* 13 (2006), 364–378.
19. M. Bern and D. Goldberg. Improved ranking functions for protein and modification-site identifications. *RECOMB 2007*, T. Speed and H. Huang (eds.), LNBI 4453, Springer, 444–458. Also *J. Comp. Biology*, 15 (2008), 705–719.
20. N. Bandeira, D. Tsur, A. Frank, and P. Pevzner. Protein identification by spectral networks analysis. *Proc. Nat. Academy of Sciences, USA*, 104 (2007), 6140–6145.
21. J.M. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Research* 2 (2003), 43–50.
22. H. Liu, R.G. Sadygov, J.R. Yates, III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76 (2004), 4193–4201.