Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data

TOWARD MORE EFFICIENT IDENTIFICATION OF POST-TRANSLATIONAL MODIFICATIONS, SEQUENCE POLYMORPHISMS, AND NOVEL PEPTIDES

Alexey I. Nesvizhskii‡§¶, Franz F. Roos§‖, Jonas Grossmann§‖, Mathijs Vogelzang‡, James S. Eddes‡, Wilhelm Gruissem‖, Sacha Baginsky‖, and Ruedi Aebersold‡‖

In mass spectrometry-based proteomics, frequently hundreds of thousands of MS/MS spectra are collected in a single experiment. Of these, a relatively small fraction is confidently assigned to peptide sequences, whereas the majority of the spectra are not further analyzed. Spectra are not assigned to peptides for diverse reasons. These include deficiencies of the scoring schemes implemented in the database search tools, sequence variations (e.g. single nucleotide polymorphisms) or omissions in the database searched, post-translational or chemical modifications of the peptide analyzed, or the observation of sequences that are not anticipated from the genomic sequence (e.g. splice forms, somatic rearrangement, and processed proteins). To increase the amount of information that can be extracted from proteomic MS/MS datasets we developed a robust method that detects high quality spectra within the fraction of spectra unassigned by conventional sequence database searching and computes a quality score for each spectrum. We also demonstrate that iterative search strategies applied to such detected unassigned high quality spectra significantly increase the number of spectra that can be assigned from datasets and that biologically interesting new insights can be gained from existing data. Molecular & Cellular Proteomics 5:652–670, 2006.

Proteomics, the systematic identification and characterization of all proteins expressed in a cell, has become a key analytical approach in the life sciences (1). The dramatic progress of proteomic research over the last decade has been catalyzed by several, seemingly independent developments. First, the wealth of genomic sequence information generated by large scale sequencing projects and the development of computational gene prediction and annotation tools have produced sequence databases that are expected to contain most coding gene regions. These databases can be searched with proteomic data and constrain the proteomic search space (2). Second, technological improvements in mass spectrometry and peptide and protein separation techniques allow rapid and sensitive protein identification from minute amounts of complex biological samples (for reviews, see Refs. 1, 3, and 4). Third, the development of computational tools for the assignment of MS/MS spectra to peptide sequences and the statistical validation of these assignments support the consistent analysis of large datasets with no or minimal human intervention (5). Collectively these developments resulted in the emergence of shotgun proteomics, a strategy based on the combination of tandem mass spectrometry-based peptide sequencing and sequence database searching, which now routinely permits the identification of hundreds to thousands of proteins in a single experiment.

Shotgun proteomics creates significant computational challenges (5–8). Large numbers (on the order of $10^5$) of MS/MS spectra acquired in each experiment need to be computationally processed to identify peptides that produced them and to infer what proteins were present in the original sample. In most high throughput studies, peptide identification is performed by searching MS/MS spectra against protein sequence databases. A number of automated database search tools have been developed for that purpose, including commercial and open source programs (9–17). These programs correlate the experimental MS/MS spectra with theoretical fragmentation patterns of peptides obtained from a sequence database and use various scoring schemes to find the best matching peptide sequence. This high throughput protein identification process, however, is prone to false positives resulting from incorrect peptide assignments to MS/MS spectra by the database search tools (5, 18–21). The problem of false positives has received significant attention in recent years. As a result, statistical approaches and computational tools were developed for assigning confidence measures to peptide and protein identifications and for estimating the false identification rates. These tools reduce the need for time-consuming manual verification of peptide assignments.
and allow faster and more consistent analysis of large scale datasets (5).

Despite the progress in developing new database search tools and methods for statistical validation of peptide assignments to MS/MS spectra, the number of spectra that remain “unassigned” (i.e., the sequence of the peptide that produced the spectrum is not known) in any experiment is significant. In fact, of all MS/MS spectra acquired in a typical shotgun proteomic experiment only a relatively small fraction (e.g., typically less than a half in the case of the ion trap instruments) is assigned a peptide with high confidence. The use of higher quality instruments having better mass accuracy and resolution alleviates the problem but does not eliminate it. Reasons for such a high failure rate include the deficiencies of the scoring schemes used to quantify the degree of similarity between the experimental spectrum and those predicted for database peptides, ambiguities in the determination of the charge state of the peptide ions selected for fragmentation (in low mass accuracy instruments), the presence of spectra arising from non-peptidic contaminants, concurrent fragmentation of multiple different precursor ions, and the low quality of many spectra due to excessive noise or incomplete peptide fragmentation (5, 7, 22, 23). However, a significant fraction of high quality, peptide-derived spectra also remain unassigned, and the failure of the database search tools to interpret them correctly cannot be explained by the reasons mentioned above.

One source of unassigned high quality spectra are peptides containing a post-translational or chemical modification. To speed up the analysis, MS/MS database searching is often performed in a way that does not anticipate the presence of modifications in the peptides analyzed, and the modified peptides are therefore missed. Another source for such spectra are peptides whose sequence is not present in the searched protein sequence database, e.g., peptides corresponding to unanticipated alternative splice forms or sequence variants of known proteins (polymorphisms). Although identification of modified, mutated, or novel peptides can be of high significance, the high quality spectra that remain unassigned after the initial database search pass through the data in most high throughput experiments are not further analyzed. Finding such spectra would normally entail manually sifting through large amounts of low quality data, which is rarely practiced due to the ever increasing number of newly acquired datasets waiting to be processed and interpreted. Thus, there is a clear need to develop computational tools for automated spectrum quality assessment that could be used to detect unassigned high quality spectra and mark them for subsequent reanalysis. Such quality assessment tools can also be used for a different task, e.g., to filter out low quality spectra prior to database searching to reduce the computational time or to assist in the process of discriminating between correct and incorrect peptide assignments.

In this work, we present a dynamic quality scoring approach for finding high quality unassigned spectra in large shotgun proteomic datasets. The basic idea behind the method is that in the first database search pass through the data high confidence peptide assignments are generally based on high quality MS/MS spectra. Therefore, the notion of what constitutes a high quality spectrum can be learned from the analyzed data itself, i.e., without relying on a training dataset created using spectra from a different experiment (24). This way the statistical classifier is automatically developed for each dataset anew, ensuring the robustness of the method toward variations in the MS/MS spectrum properties caused by differences in acquisition methods or instrument to instrument variability. This distinguishes this method from other recently described approaches (25, 27–29) that can produce inaccurate results when applied to spectra that differ significantly from those used for training. The statistical spectrum quality classifier is computed using an extended set of spectrum features, including those designed to take into account the knowledge of the peptide fragmentation process.

The accuracy of the method is evaluated using a dataset of MS/MS spectra from a recent experiment on human lipid rafts (30). The results indicate that the spectrum quality classifier enables fast and automated detection of high quality spectra left unassigned during the initial computational pass through the data. We also demonstrate that by interrogating those unassigned high quality spectra more comprehensively using existing protein sequence databases and by searching against genomic databases, one can significantly increase the number of identified peptides, including peptides containing modifications and sequence polymorphisms. Furthermore reanalysis of high quality spectra can lead to the identification of novel peptides, e.g., peptides confirming computationally predicted alternative splice forms.

**EXPERIMENTAL PROCEDURES**

**Experimental Datasets**

Three experimental datasets of MS/MS spectra were used to evaluate and optimize the quality scoring method and to investigate the sources of peptides that produced those spectra. All spectra were acquired using ESI ion trap tandem mass spectrometers.

1) The *Haemophilus influenzae* dataset consisted of 15 LC-MS/MS runs and was used previously as a test dataset for the development of a statistical model for validation of peptide assignments to MS/MS spectra (19). Sample proteins were digested using trypsin, and the resulting peptide mixtures were separated using reverse phase and strong cation exchange (SCX) chromatography prior to MS/MS se-

---

1 The term “novel” peptide here refers to a peptide identified by searching an unannotated genomic database and whose sequence is not present in any major protein sequence database for the corresponding organism.

2 A. I. Nesvizhskii, unpublished. See also Ref. 18.

3 The abbreviations used are: SCX, strong cation exchange; IPI, International Protein Index; SQS, spectrum quality score; LDF, linear discriminant function; ROC, receiver operator characteristic; EST, expressed sequence tag; SNP, single nucleotide polymorphism; LIME1, Lck-interacting transmembrane adaptor 1 protein.
quencing. The dataset contained more than 16,000 multiply charged MS/MS spectra generated from the membrane fraction of the sample. The spectra were searched using SEQUEST as described previously (19), and peptide assignments to the spectra were processed using PeptideProphet. Of the 31,728 search results (multiply charged spectra were searched twice, assuming 2+ or 3+ charge state, because the exact charge could not be determined), 4229 MS/MS spectra were assigned a peptide with probability above 0.9 (3457 and 730 assignments to spectra of doubly charged and triply charged precursor ions, respectively). This dataset was used to optimize the method for computing spectrum quality scores and to test the accuracy of the method as a function of the dataset size.

2) The Arabidopsis dataset was acquired from four protein mixtures derived from cultured Arabidopsis thaliana cells. The crude extract was loaded on a one-dimensional SDS gel, and single bands were cut out. Each band was digested with trypsin, and peptides were sequenced by LC-MS/MS. In total, 3420 MS/MS spectra were submitted to a SEQUEST database search against a protein database of Arabidopsis and known contaminants allowing for semitryptic (tryptic at one terminus only) peptides with a mass tolerance of 3 Da and with specifying carboxyamidomethylation of cysteines and methionine oxidation as variable modifications. This search resulted in 6720 peptide assignments (counting 2+/-3+ duplicates). Of those, 924 were peptide assignments to spectra with a probability of being correct greater than 0.90 as computed by PeptideProphet. Additional database searches were performed using this dataset to test the ability of the method to recover high quality unassigned spectra.

3) The human lipid raft dataset was taken from a large scale quantitative proteomic experiment on lipid raft plasma membrane domains from human Jurkat T cells (30). The full dataset is available from the PeptideAtlas MS/MS data repository (31). Nine LC/MS/MS runs (flow-through sample, SCX fractions 31 through 39) were selected for the analysis in this work. The spectra, 12,864 in total, were first searched using SEQUEST against the human International Protein Index (IPI) database (32), version 2.35, allowing for semitryptic peptides with a mass tolerance of 3 Da and allowing for methionine oxidation as a variable modification. This resulted in 24,197 peptide assignments, counting 2+/-3+ duplicates (711 singly charged, 12,153 doubly charged, and 12,044 triply charged). Of those, 4171 were peptide assignments to spectra with a PeptideProphet probability of being correct equal or greater than 0.9 (4034 peptide assignments to 2+/-3+ spectra and 137 assignments to 1+ spectra). The results were further analyzed using ProteinProphet (19), resulting in 315 proteins having a ProteinProphet probability above 0.9 identified by 660 unique peptides. The number of proteins was determined by counting the number of entries in the minimal list of proteins sufficient to explain all observed peptides (19). The list of peptide assignments and the corresponding list of protein accession numbers are provided in Supplemental Table S1 (the results of the analysis of the entire dataset, including ICAT samples, can be found on the PeptideAtlas website). In calculating the number of assigned spectra, a spectrum was counted only if it was assigned a peptide corresponding to a protein with ProteinProphet probability greater than 0.9. To eliminate random matches to proteins correctly identified by other peptides, an additional peptide level constraint was applied: a spectrum was counted as assigned only if the peptide was identified (from this or another spectrum in the dataset) with PeptideProphet probability equal to or greater than 0.5.

To determine what types of peptides generated the high quality spectra unassigned in the initial search, a subset of the initially unassigned MS/MS spectra was reanalyzed using a number of additional searches. 1) “Large mass tolerance” search: SEQUEST, semitryptic, 5-Da mass tolerance (larger mass tolerance compared with the initial 3-Da search). 2) “4+/5+ charge state” search: SEQUEST search, semitryptic, 3-Da mass tolerance, assuming 4+ or 5+ charge state. 3) “Pyro-Glu” search: SEQUEST, semitryptic, 3-Da mass tolerance, allowing for conversion of N-terminal glutamic acid (loss of 17 Da) as a variable modification. Due to a limitation of SEQUEST, the search was performed allowing for modified residues to be located anywhere in the sequence; peptides that did not contain glutamine at the N terminus were filtered out at the data validation stage.

4) “Mascot” searches: Mascot, tryptic peptides only (note that the Mascot tryptic search allows for removal of the initiating Met), two missed cleavages or less, 3-Da mass tolerance, allowing for N-terminal acetylation (+42 Da) or carbamoylation (+43 Da) as a variable modification.

5) Miscellaneous searches: X! Tandem (13) search allowing for more than one type of modification per peptide; SEQUEST and Mascot searches allowing for modifications not specified in the previous searches (e.g., conversion of N-terminal glutamic acid residues to pyrogulutamic acid, phosphorylation, guanidination, etc.).

These searches were performed against the same version of the IPI database as the initial search (version 2.35). The results from different searches were processed using Interact and PeptideProphet and then manually scrutinized. Validated peptide sequences were imported in a relational structured query language database, and the total number of assigned spectra, the number of unique peptides, and the minimum number of protein identifications sufficient to explain all identified peptides were calculated. In addition, the spectra were searched again against a genomic database as described under “Results.”

Computational Method

Separation of Spectra into Peak List Subsets (Low and High Intensity Components)

Prior to computing spectrum features, two peak list subsets are extracted from each spectrum. The main reason for this step is a reduction of the number of noise peaks in the spectrum whose presence can lower the discriminating power of some spectral features. The two subsets of peaks are extracted for each spectrum in parallel using the following approaches.

1) A signal intensity cutoff is applied using a robust percentile-based approach, creating the “high intensity” peak subset. To account for the significant variability of the fragment ion intensities across the spectrum, the spectrum is divided into five equally sized m/z sections. Within each section, the peaks are sorted according to their intensity, and the intensity at a given percentile is used as the cutoff value (i.e., the signal intensity at the 50th percentile would be equal to the median intensity of the peaks in a section). All peaks with intensity above the cutoff are assigned to the high intensity peak subset. Depending on the instrument type and settings, MS/MS spectra may contain a large number of noise peaks. To make the percentile approach more robust, the number of peaks taken into account is restricted to the 400 most intense peaks per 1000-Da interval.

2) Spectrum deisotoping is performed based on a Poisson distribution model, thus creating the “deisotoped” peak subset. The peak heights of the isotope patterns are predicted by a Poisson distribution model with fragment masses, and the “heavy isotope excess” constant is estimated by dividing the sum of the average masses of all amino acids by the sum of the monoisotopic masses. The theoretical
distribution and the actual spectrum are compared, and the \( \chi^2 \) test statistic is calculated to measure the quality of the fit. All peaks with the \( \chi^2 \) value below 20 are assigned to the deisotoped peak list subset. This represents a relatively loose threshold appropriate for low mass resolution data where the observed isotope distributions of heavy ions tend to deviate substantially from the theoretical distributions.

Spectrum features such as sequence tags, complementary fragments, and neutral losses (described below) are calculated exclusively on the high intensity peak subset as defined by the percentile cutoff, whereas the general spectrum features are calculated using the full spectrum as well as the two peak list subsets (the high intensity peak subset and the deisotoped peak subset). The percentile cutoff was optimized separately for each spectrum feature as described under “Results.”

**Spectrum Features**

Several classes of spectrum features were introduced to measure the quality of MS/MS spectra. These features were designed to characterize the overall distribution of peaks in the spectrum (general spectrum features) as well as to detect certain patterns (sequence tags, complementary fragment ions, and neutral losses) expected to be present in high quality MS/MS spectra but absent in low quality spectra.

**General Spectrum Features**—Eight general descriptive features were calculated for each peak list subset.

(a) Number of peaks, square root-transformed.
(b) Arithmetic mean of the peak intensities, log-transformed.
(c) Standard deviation of the peak intensities, log-transformed.
(d) Smallest \( m/z \) range containing 95% of the total peak intensity.
(e) Smallest \( m/z \) range containing 50% of the total peak intensity.
(f) Total ion current per \( m/z \) (total ion current divided by feature d), log-transformed.
(g) Standard deviation of the consecutive \( m/z \) gaps between all peaks, log-transformed.
(h) Average number of neighbor peaks within a 2-Da interval around any peak.

The log or square root transformation of some of the spectrum features was applied to obtain a more symmetric shape of the distributions and to minimize the variance across the spectra and resulted in improved discriminating power of the composite spectrum quality classifier derived using the linear discriminant function approach (described under “Spectrum Quality Score”). For each spectrum, the entire peak list and the two subsets (the high intensity peak subset and the deisotoped peak subset) were extracted and characterized by the eight descriptive features, resulting in 24 general spectrum features. Clearly many of these spectrum features are redundant, and there is a particularly high degree of correlation between those spectrum features that are derived by applying the same descriptive feature to the three different peak subsets. Nevertheless having each spectrum feature computed in triplicate slightly improved the overall performance of the method.

**Sequence Tags**—A high quality peptide MS/MS spectrum should contain an extended series of peaks separated by the mass differences corresponding to one of the 20 amino acids. Thus, if a spectrum does not allow extraction of at least a short sequence tag (partial sequence) it is unlikely to be of high quality. The sequence tag-based spectrum features are calculated using the high intensity peak subset only. The following four spectrum features measuring the presence of amino acid mass differences in the spectrum are computed.

(a) The length of the longest sequence tag that can be extracted from the spectrum (24). The tags are extracted using a fast dynamic programming algorithm in a single pass through the spectrum and assuming that all peaks correspond to singly charged fragment ions from the same ion series (e.g. all \( y^- \) or \( b^- \) ions).
(b) The average (per peak) length of all extracted sequence tags, calculated by summing up the lengths of the tags involving each peak in the high intensity subset of the spectrum and dividing it by the number of peaks.
(c) The number of peak pairs corresponding to an amino acid mass difference (sequence tag of length one).
(d) A derived version of the above feature computed using the peak intensities as weighting factors to account for the increased likelihood of more intense peaks being true fragment ions. For every pair of peaks, the geometric mean of their intensities is calculated, and the spectrum feature is computed as a ratio of the sum of the geometric means of peak pairs corresponding to an amino acid mass divided by the sum of the geometric means of all peak pairs.

In the case of peptides containing modifications, the sequence tags are bound to be interrupted by mass differences not corresponding to any of the standard amino acid residue masses. Although the masses of typical modifications can be included in the list of amino acid masses, this would lead to an increased number of peak pairs occurring at random. The additional measures counting the isolated amino acid mass differences (features c and d) are thus useful in the case of modified peptides where extraction of long sequence tags using the standard set of 20 amino acids might not be possible.

**Complementary Fragment Ions and Neutral Losses**—The fragmentation of multiply charged peptide ions via collision-induced dissociation results in the charges being distributed among the resulting fragment ions. The doubly charged parent ions most often dissociate into two singly charged fragments, which leads to a high degree of symmetry in the spectrum. In contrast, the triply charged parent ions dissociate into two charged fragments, but the charges are unevenly distributed with one fragment carrying two charges. When a precursor ion with the \( m/z \) value \( m_z \) and charge \( z_s \) breaks down into two ions (in the case of ion trap mass spectrometers, primarily \( b \) ions and \( y \) ions), the overall mass and charge are preserved but split up between the two fragment ions, thus meeting the following criteria (33, 34).

\[
m_p \cdot z_p = m_z \cdot z_s + m_{z_p} \cdot z_y
\]

\[
z_p = z_s + z_y \quad (0 < z_p < z_s, \quad 0 < z_y < z_s)
\] (Eq. 1)

Although the fragmentation process is strongly dependent on the sequence of the peptide and the resulting fragmentation patterns observed in MS/MS spectra can be quite complex, a high quality spectrum should still contain a significant number of complementary peak pairs (fragment ions whose masses add up to the mass of the precursor ion). Three spectrum features are computed (using the high intensity peak subset only) in the following way.

(a) The number of complementary peak pairs satisfying the above conditions (Equation 1).
(b) The number of complementary peak pairs (feature a) divided by the background count (the number of complementary pairs expected by chance). The background count is estimated by averaging the numbers of peak pairs satisfying the same conditions (Equation 1) but using 10 incorrect \( m/z \) values of the precursor ion obtained by shifting the correct precursor ion mass by ±10, 20, 30, 40, and 50 Da.
(c) Similar to feature b but with the background count subtracted from the number of complementary peak pairs instead of taking the ratio.

When the charge state of the precursor peptide is not known (e.g. 6 The length of the sequence tag is implemented as an optional spectrum quality filter in SpectrumMill (www.chem.agilent.com).
multiply charged ions in the case of low mass resolution data), the number of complementary peaks is computed first assuming the $2^+$ charge state (in which case all observed fragment ions are expected to be singly charged) and then assuming the $3^+$ charge state (in which case both singly charged and doubly charged fragments are allowed). Then the largest of the two counts is selected and used as the number of complementary peak pairs feature (feature a), and the two derivative features (b and c) are computed as described. It should be noted that some of the spectrum features used here for quality assessment purposes, especially the number of complementary fragments and its derivatives, can be used to assist in the determination of the precursor ion charge state directly from the MS/MS spectrum (33, 34). In the case of high resolution mass spectrometers where the charge state can be accurately determined, the entire analysis can be performed separately for spectra with different precursor ion charge states.

Similarly neutral losses such as loss of ammonia (loss of 17 Da), water (18 Da), and carbon monoxide (28 Da) are often found in MS/MS spectra of peptide ions (see e.g. Ref. 35). The presence of neutral losses in the spectrum is useful for distinguishing peptide ion spectra from those of contaminants or background noise. They are detected in the same way as the complementary fragments described above, thus adding nine more spectrum features and bringing the total number to 40.

**Spectrum Quality Score**

Improved discrimination between assigned and unassigned spectra (and thus, to a large degree between high quality and low quality spectra) can be achieved by combining all separate spectrum features into a single discriminant score (the spectrum quality score (SQS)). The spectrum quality score is a weighted combination of the individual scores $s_i$, described above computed according to the discriminant function (Equation 2).

$$\text{SQS}(s_1, s_2, \ldots, s_n) = c_0 + \sum_{i=1}^{n} c_i s_i \quad \text{(Eq. 2)}$$

In the linear discriminant function (LDF) approach used in this work the constant $c_0$ and the weights $c_i$ are derived in such a way that the ratio of between-class variation to within-class variation is maximized under the assumption of multivariate normality. This assumption does not necessarily hold for all of the spectrum features, and nonlinear classification methods should in general allow better discrimination. Nevertheless the LDF analysis is a robust multivariate technique that classification methods should in general allow better discrimination.

The weighing factors, measuring the relative contribution of each spectrum feature to the overall quality score, are determined using the linear discriminant analysis. Discriminant function coefficients are determined automatically for each analyzed dataset anew (“dynamic” quality scoring). Finally the statistical classifier is applied to the spectra left unassigned in the initial search to find those that are of high quality. Such spectra are the most interesting candidates for subsequent reanalysis because they are likely to contain valuable biological information that is worthwhile to be explored further.

**Method Optimization and Testing**—The computational method was first tested and optimized using several datasets. The optimization of the spectrum features was performed using the *H. influenzae* dataset. Spectrum features such as...
sequence tags, complementary fragment ions, and neutral losses were found to be sensitive to the peak intensity cutoff. As the peak intensity cutoff increases, an increasingly higher fraction of the noise peaks are eliminated, thus reducing the number of peak pairs and sequence tags occurring by chance. This, in turn, improves the discriminating power of the spectrum features. This trend, however, holds up only to a certain point after which the performance starts degrading due to exclusion of true fragment ions from the high intensity peak subset. Fig. 2A illustrates this in more detail for one particular feature, the length of the longest sequence tag that can be extracted from the spectrum. The plot shows the frequency of extracting a tag (the longest tag for each spectrum) of a certain length among assigned and unassigned spectra in the *H. influenzae* dataset when using all peaks in the spectrum and after applying an 80% intensity cutoff (i.e., using only the top 20% of most intense peaks). For both assigned and unassigned spectra, the distributions are shifted toward shorter tag lengths when the sequence tags are extracted from the high intensity peak subsets compared with using all peaks. The average length of the longest tag extracted from the assigned or unassigned spectra was around 15 and 6 amino acids, respectively, when the tags were extracted using all peaks. These numbers dropped to ~7 and 2 at the 80% intensity threshold. To find the optimal intensity cutoff, the significance of each spectrum feature was measured using the non-parametric two group comparison test (Mann-Whitney z score). Fig. 2B shows the statistical significance of several spectrum features that are representative of each group (sequence tags, complementary fragments, and neutral losses) plotted as a function of the intensity cutoff. For each class of spectrum features, the discriminating power was evaluated, and the parameters were optimized. The plot demonstrates that in the case of sequence tag-based features, the optimal intensity cutoff (the cutoff resulting in max-

---

**Fig. 1.** Overview of the dynamic quality scoring approach for finding unassigned high quality spectra. The results of the initial database search are statistically evaluated using e.g. PeptideProphet. Computed probabilities are used to automatically assemble a training dataset of assigned (probability of being correct above 0.9) and unassigned (probability below 0.1) spectra. For each MS/MS spectrum, various numerical descriptors of the spectrum are calculated, and a statistical classifier is developed. The classifier is then applied to all unassigned spectra in the dataset to find those that are of high quality. Such spectra, if reanalyzed more comprehensively, can result in the identification of post-translational modifications, sequence polymorphisms, or novel proteins. DB, database; LDA, linear discriminant analysis.

---

8 Denoising of MS/MS spectra is also known to increase the sensitivity of the database search-based peptide identification process (38, 39).
A similar analysis was performed for other classes of spectrum features, and the intensity cutoffs of 70 and 50% for the complementary fragments and neutral losses, respectively, were chosen. In the case of neutral losses, the choice of cutoff was slightly arbitrary in the sense that a whole contiguous range of cutoffs yielded similar performance, reflecting the absence of a pronounced maximum in the curve on Fig. 3B. In addition, similar trends were investigated for the mass tolerance parameter, and the tolerance of 0.7 Da was selected for computing the sequence tag-based features, 0.9 Da was selected for the complementary fragments, and 0.5 Da was selected for neu-

![Fig. 2. Optimization of spectrum quality features. A, distribution of the length of the longest sequence tag among assigned (green) and unassigned (blue) spectra extracted at different peak intensity cutoffs, 80% (solid line) and 0% (i.e. using all peaks in the spectrum, dashes). Inset, mean length of the longest sequence tag extracted from the assigned (green) or unassigned (blue) spectra plotted as a function of the peak intensity cutoff. B, discriminating power (Mann-Whitney z score) of several representative spectrum features: the length of the longest sequence tag that can be extracted from the spectrum (blue, solid), the number of complementary fragments (magenta, dash-dot), and the number of peaks corresponding to the loss of water (green, dashes) computed as a function of the peak intensity cutoff.](image)

![Fig. 3. Discriminating power of the spectrum quality classifier. A, ROC curves showing the discriminating power of the spectrum quality classifier computed using all features (black curve, □) and all except general features (purple, ◊) as well as using features from each single feature group: general features (green, ●), complementary fragments (magenta, ▲), sequence tags (blue, ▼), and neutral losses (red, ◌). B, ROC curves showing the performance of the classifier as a function of the training dataset size: 2029 assigned, 6747 unassigned (black curve, □); 444 assigned, 1312 unassigned (green, ●); 111 assigned, 362 unassigned (magenta, ▲); and 40 assigned, 142 unassigned (blue, ▼).](image)
tal losses. In general, these parameters should be optimized separately for each type of mass spectrometer. Ideally the optimization process should be carried out using the same on the fly created training dataset used to determine the discriminant function coefficients in Equation 2, thus making the method even more robust. However, preliminary tests indicated that slight changes in the model parameters did not significantly affect the overall performance of the spectrum quality classifier, and the implementation of the dynamical optimization process was deferred to future work.

The overall performance of the composite spectrum quality classifier was tested on the *H. influenzae* dataset using receiver operator characteristic (ROC) plots. To generate ROC curves, the discriminant function was developed using 50% of the data and applied to the remaining 50%. Fig. 3A plots the spectrum quality score obtained by combining all spectrum features using the LDF approach as well as scores computed using spectrum features from a single category only (e.g. using general spectrum features only). The plot demonstrates that by combining multiple different spectrum quality features into a composite score one can achieve better discrimination between assigned and unassigned spectra than by using any class of spectrum features alone. Many spectrum feature groups have high discrimination power individually, but the significance of each group varies. For example, the spectrum features describing the presence of complementary fragment ions in the spectrum are the most discriminative of all classes of spectrum features in the region of low sensitivity (below 0.5). This is the region of interest when the task is to find a relatively small number of unassigned spectra of very high quality. At the same time, the spectrum features from the general category allow better discrimination in the region of high sensitivity, which is more relevant to the problem of filtering out low quality spectra prior to database searching without a significant loss in the number of peptide identifications.

Within each class of spectrum features, not all individual features contributed equally to the discrimination. Furthermore combining several redundant features did not result in a significantly improved discrimination overall. For example, using the length of the longest sequence tag extracted from the spectrum as the only sequence tag-based feature allowed almost as good overall discrimination as using all four features from that category combined. The effect of inclusion of redundant or marginally discriminating features on the accuracy of the classifier was investigated by performing a stepwise analysis (addition of one new feature at each step), and ROC curves were plotted to measure the performance of the classifier at each step (data not shown). Importantly addition of redundant features or features with low discriminating power did not negatively affect the performance of the classifier but resulted in a slight improvement. Furthermore the use of multiple related features made the method more robust toward unexpected variations in the spectrum properties.

Based on these considerations, all 40 spectrum features were kept in the analysis.

Another important consideration is the minimum size of the dataset needed to realize the advantages of dynamic training. This was investigated by applying the method to several *H. influenzae* datasets of decreasing size. The results of this analysis are shown in Fig. 3B. Overall the plot demonstrates that the method can be used even with datasets of relatively small size. As expected, the best results were achieved with large datasets consisting of more than several LC-MS/MS runs. However, even for a dataset consisting of a single LC-MS/MS run (with 111 and 362 assigned and unassigned spectra, respectively, in the dynamically created training dataset), the classification performance was satisfactory and only marginally worse than that observed for larger datasets. The accuracy of the classifier started degrading to a significant degree only when applied to very small datasets (with less than 50 spectra assigned in the initial search). Thus, in a typical shotgun analysis of complex protein mixtures, in most cases even a single LC-MS/MS run generated on current mass spectrometers should be sufficient to train the classifier because it is unlikely that it would contain fewer than several thousand spectra of which minimally a few hundred spectra would be identified confidently in the initial database search. In case of very small datasets, rather than training the classifier dynamically it would better to use a larger dataset generated in another (but similar) experiment. The larger size of the training dataset would then compensate for the loss of accuracy due to the differences in the spectrum properties between the dataset of interest and the training dataset.

**Evaluation of the Method**—The ability of the method to effectively recover unassigned high quality spectra was evaluated using the *Arabidopsis* dataset. The same dataset was used to investigate the robustness of the method toward the presence of ‘misclassified’ (high quality) spectra in the unassigned subset of the on the fly created training set. The tests were carried out in the following way. First the spectra were search against the full database and allowing for Met + 16 as a variable modification (normal search). The same spectra were than searched against a modified database, or allowing for no modifications, thus restricting the search space to prevent the identification of some spectra (called ‘masked’ spectra) that were successfully assigned in the normal search (see Fig. 4A). These flawed searches were performed in three different ways by making the following changes compared with the normal search.

1) Search allowing for no modifications, resulting in 94 masked spectra.
2) Search against a modified database created by removing sequences of four high abundance proteins, resulting in 122 masked spectra.
3) Search with the mass of serine statically increased by 5 Da, resulting in 402 masked spectra.
The spectrum quality classifier was developed using the results of the flawed searches. The performance of the method was then evaluated by plotting the distribution of quality scores for the masked spectra and comparing it with the distribution for the assigned spectra (see Fig. 4, B, C, and D). In the first two tests (Fig. 4, B and C), the distribution of quality scores for the masked spectra remained similar to the distribution observed for the assigned spectra. The performance was worse in the third case (the search with incorrect serine mass) where the number of high quality spectra in the unassigned subset of the dynamically created training dataset far exceeded the number of assigned spectra, a situation that is unlikely to occur in practice (Fig. 4D).

Overall the tests described above demonstrate that the approach should be able to detect unassigned high quality spectra based on their similarity to successfully identified spectra in any typical experiment. In those cases where it is expected that only a small fraction of all high quality spectra would be successfully assigned in the initial database search (e.g., proteomic studies using organisms with unsequenced or partially sequenced genome), a better approach would be to use a classifier trained on a different, more suitable dataset generated using the same type of mass spectrometer.

**Application of the Method to a Human Lipid Raft Dataset**—The tests described in the previous section demonstrate that the method is able to recover high quality spectra left unassigned due to intentionally introduced flaws in the database search parameters or in the sequence database itself. However, for practical applications, it is more important to estimate how many high quality spectra are left unassigned in a

---

**Fig. 4.** Evaluation of the ability of the method to detect unassigned high quality spectra. A, correlation between the probabilities of peptide assignments in the “normal” search (allowing for Met + 16 as a variable modification) and in the ‘flawed’ search (no modifications allowed). The performance of the classifier was evaluated by its ability to assign high quality scores to the masked spectra, which are located in the lower right corner of the plot. B, the distribution of spectrum quality scores among assigned (green, dash-dot), unassigned (blue, solid), and masked spectra (magenta, short dashes) in the first test: database search allowing for no modifications. C, same as above, second test: database search with the sequences of four high abundance proteins removed from the database. D, same as above, third test: database search with incorrect mass of serine residues.
typical analysis, to test whether the dynamic quality scoring method is able to recover them, and to investigate what types of peptides produced those spectra. Answering these questions is also important for designing more efficient computational strategies for peptide identification.

To address these questions, a dataset of ion trap MS/MS data, a subset of a large human lipid raft experiment (30), was chosen for more detailed analysis (see “Experimental Procedures”). The spectra were initially searched against the human IPI database (version 2.35) using SEQUEST with 3-Da precursor ion tolerance, allowing for tryptic and semitryptic peptides, and with variable Met + 16 modification. These search parameters are equivalent to those implemented in the PeptideAtlas high throughput data analysis pipeline (31) and are quite typical for data generated using ion trap mass instruments. In total, the initial database search resulted, after applying ProteinProphet (19), in 315 proteins having a ProteinProphet probability above 0.9 (the minimal number of proteins sufficient to explain all observed peptides; see Refs. 19 and 40) identified by 653 unique peptides9 from 4172 assigned spectra (see Table I). The list of peptide assignments and the corresponding list of protein accession numbers are provided in Supplemental Tables S1 and S2.

Analysis of the database search results confirmed that this

| Category | Spectra | Peptides | New peptides | New proteins |
|----------|---------|----------|--------------|--------------|
| Tryptic  | 3552    | 627      | 627          | 311          |
| Semitryptic | 81       | 21       | 21           | 4            |
| Met + 16 | 92       | 34       | 4            | 0            |
| 347 additional peptide assignments resulting from reanalysis of unassigned high quality spectra, SQS > 1 |
| $M_{\text{calc}} - M_{\text{exp}} > 3$ Da | 46 | 6 | 1 | 0 |
| $4+/5+$ charge state | 13 | 6 | 2 | 0 |
| Other search tools (Mascot, X! Tandem) | 42 | 34 | 2 | 1 |
| N-terminal pyro-Glu from Gln | 48 | 3 | 14 | 6 |
| N-terminal acetyl/carbamyl | 33 | 12 | 6 | 3 |
| Other modifications | 8 | 0 | 0 |
| EST search | 7 | 2 |

9 Here the term “unique peptide” is used to describe the identification of the same peptide sequence from one or more different MS/MS spectra.

The results of this initial database search were used to train the spectrum quality classifier, which was then applied to all multiply charged spectra. Singly charged MS/MS spectra were removed from further analysis due to their relatively small contribution. The distribution of quality scores for the entire dataset of multiply charged spectra as well as for the
assigned and unassigned subsets is shown in Fig. 5. The plot indicates that close to 80% of all unassigned MS/MS spectra are of relatively low quality; these spectra might not be worth additional efforts to reanalyze them. However, a significant number of unassigned spectra are actually of high quality, which manifests itself in the presence of an extended right tail in the quality score distribution for the unassigned spectra. Fitting the quality score distribution observed for the assigned spectra to match the extended right tail of the quality score distribution for the unassigned spectra indicates that it should be possible to extract at least 25% more peptide identifications in addition to those obtained in the initial database search (see Fig. 5).

To determine what types of peptides generated the high quality spectra unassigned in the initial search, 632 unassigned spectra having high computed spectrum quality score (SQS above 1), representing less than 5% of the original dataset, were selected for more comprehensive reanalysis. All searches were performed against the same version of the IPI database as the initial search (version 2.35). The overall database search strategy was to minimize the number of redundant searches while attempting to identify as many different peptides as possible (see “Experimental Procedures” for details). Of all the spectra submitted for reanalysis, 359 were assigned a peptide in one of the additional searches, and six were determined to be correctly assigned in the initial search (but with low PeptideProphet probability of less than 0.1). The list of the additional peptide assignments is given in Supplemental Table S3. Comparing with the 1695 spectra of comparable quality (SQS above 1) that were assigned in the initial search, this represents a 20% increase in the number of assigned spectra. As expected, the higher the quality score of the initially unassigned spectrum was, the more likely it was that the spectrum would eventually be interpreted with high confidence after completion of all additional database searches (see Fig. 6). For example, among 268 initially unassigned spectra having a quality score between 1 and 1.5, it was possible to identify only 129 spectra (48%), whereas it was possible to identify 22 of the 28 spectra having a quality score above 3 (79%). This is consistent with the expected rates of spectrum assignments given the distributions of quality scores plotted in Fig. 5.

The distribution of peptide assignments, obtained in the additional searches, among different categories is shown in Table I. The search with large mass tolerance added 64 assigned spectra (46 unique peptides) that were missed in the initial search due to inaccurate measurement of the precursor ion mass (mass error $|M_{\text{calc}} - M_{\text{exp}}|$ of more than 3 Da). In addition, 20 MS/MS spectra (13 unique peptides) were arising from peptide ions with +4 or +5 charge states. Furthermore 42 unmodified peptides were identified from 64 spectra with the help of two other database search tools, Mascot and X! Tandem, and by performing SEQUEST searches using a different set of search parameters than those used in the initial

---

**Figure 5.** Application of the method to the human lipid raft dataset. Distribution of quality scores plotted for all spectra from the dataset (black curve) and separately for the assigned (dash-dot) and unassigned (dashes) spectra based on the initial database search is shown. The distribution of scores for the unassigned spectra has an extended right tail indicating the presence of high quality spectra among them. Short dashed curve represents an estimated fraction of the unassigned spectra that are of high quality.

**Figure 6.** Results of reanalysis of 632 high quality (SQS > 1) spectra unassigned in the initial search (human lipid raft dataset). The number of initially unassigned spectra that were eventually interpreted after completion of all additional database searches (black bars) and the number of spectra that remain assigned (white bars) as a function of spectrum quality are shown. Inset, fraction of initially unassigned spectra that got identified in one of the additional searches as a function of spectrum quality.
search (i.e. trypsin-constrained search and not allowing for any modifications).

In addition, a significant number of peptides containing modifications were identified among which peptides with the N-terminal glutamine converted to pyroglutamic acid (116 spectra, 48 unique peptides) and N-terminal acetylation or carbamylilation (68 spectra, 33 unique peptides) were the most common (in addition to the peptides containing oxidized methionine identified in the initial search). Other types of detected modifications included sodiation, tryptophan oxida-

Fig. 7. A, MS/MS spectrum supporting the identification of a tryptic peptide, YPIEHGIVTNWDMEK, from Actin, cytoplasmic 1 protein (Swiss-Prot accession number P02570) containing methylated histidine at position 5 in the peptide sequence and oxidized methionine at position 14. B, MS/MS spectrum supporting the identification of a tryptic peptide, LQGSATAEAQVGHQTAR, from a computationally predicted splice form of the Lck-interacting transmembrane adaptor 1 protein obtained by searching the unassigned high quality spectra against the EST database.
Quality Assessment and Iterative Analysis of MS/MS Data

...eight unique peptides in total). Fig. 7A shows one such example, the identification of a tryptic peptide, YPIEH-GIVTNWDDMEK, from Actin, cytoplasmic 1 protein (Swiss-Prot accession number P02570) containing methylated histidine at position 5 in the peptide sequence (position 73 in the protein sequence) that is a known post-translational modification annotated in the Swiss-Prot database.

Although it is informative to know how many new spectra can be assigned by performing additional, more comprehensive database searches, it is more important to know whether additional assignments add any new information to the data obtained from the initial search. Although a significant number of unmodified peptides (101 unique peptides) were identified by performing the additional searches (large mass tolerance, 4+/5+ charge state, X! Tandem and Mascot searches), only nine of those peptides were not identified from other spectra by SEQUEST in the initial search, resulting in one new protein identification and one identification with increased confidence. Similarly, most modified peptides were found to be arising from a small number of the most abundant sample proteins that were already identified on the basis of tryptic peptides seen in the unmodified form in the initial search. In fact, new protein identification resulted from any peptides containing modifications other than conversion of N-terminal glutamine to pyroglutamic acid and N-terminal acetylation.

Of the 48 unique peptides (116 MS/MS spectra) identified containing N-terminal pyroglutamic acid, all except one were not seen previously in the unmodified form in this dataset. However, this modification is likely attributable to sample handling because all of the identified peptides containing this modification were tryptic and not located at the N terminus of the protein. It is known that pyroglutamic acid readily forms at the N terminus of peptides beginning with glutamine when peptides are kept in acidic solution for 15 min or more, which is the case in shotgun proteomic analysis. Nevertheless, identification of peptides containing this modification was informative: it resulted in 10 new protein identifications and four protein identifications with increased confidence (identified by unmodified peptides in the initial search but with ProteinProphet probability below 0.9). Eight of those proteins could not be identified with high confidence based on the results of the initial SEQUEST search even when the search was extended to include the entire human lipid raft dataset of Ref. 30. These proteins include several membrane proteins that are likely to be relevant to the analysis of lipid rafts, e.g. KIAA0143 protein (IP10009217, Swiss-Prot accession number Q14156). The accession numbers of those proteins and the sequences of their corresponding peptides can be found in Supplemental Table S4 with one MS/MS spectrum for each of the identified peptides shown in Supplemental Table S5.

Of the 33 unique peptides (68 MS/MS spectra) identified containing N-terminal acetylation or carbamylation, none were seen in the unmodified form in the selected dataset (SCX flow-through sample, fractions 31 through 39 of the entire human lipid raft dataset), resulting in three new protein identifications and three identifications with increased confidence. However, additional analysis indicated that 21 of those peptides were also present in the unmodified form in other SCX fractions of the original dataset (flow-through sample, SCX fractions 43 and higher). All of the peptides that were identified in both modified and unmodified form (considering all SCX fractions) were tryptic, were located in the middle of the protein sequence, and were arising mostly from high abundance proteins. This strongly suggests that in those peptides the observed modification was a carbamylation, an artifact that occurred after protein digestion. N-terminal carbamylation is a common chemical modification known to occur when protein samples are digested in a solution containing urea, which was the case with the human lipid raft dataset used in this work (30).

In contrast, of the 12 peptides that were identified in the modified form only (considering all SCX fractions), 10 were located at the N terminus of their corresponding proteins starting with the residue immediately following the initiating methionine. In those 10 peptides the modification was likely an acetylation of the protein N terminus that occurred prior to digestion. The sequences of those peptides started with either Ala (four peptides), Ser (four peptides), or Thr (two peptides). As reviewed in Ref. 41, these residues are among the most frequently acetylated residues with the frequency of acetylation of Ala and Ser close to 100% and 80% in the case of Thr. For comparison, only three of the 21 peptides identified in both modified and unmodified form (considering all SCX fractions) started with Ala, Ser, or Thr. The identification of N-acetylated peptides is important because they confirm the translation initiation start sites and assist in the determination of the mature form of the protein present in the sample (40, 42, 43). It is also worth noting that three of the N-acetylated peptides identified here correspond to three low molecular weight proteins that could not be identified even when the initial SEQUEST search (i.e. search without looking for N-acetylated peptides) was extended to the entire human lipid raft dataset, including two less characterized proteins containing predicted transmembrane domains: RER1 protein (IP100005728, Swiss-Prot accession number O15258) and HSPC009 protein (IP100022277, Swiss-Prot/TrEMBL accession number Q9Y2R0) (see Supplemental Tables S4 and S5).

All database searches described above were performed against the human IPI database. This database is often used in the analysis of shotgun proteomic data because it represents a good tradeoff between the completeness of protein...
sequences and the degree of sequence redundancy (40). In particular, the sequence- and identifier-based construction of the IPI database significantly reduces the need for manual filtering of redundant sequences while maintaining cross-references to all its source data. However, minor sequence variants, e.g., polymorphisms, are not represented in the IPI database. Furthermore, this database is likely to be not complete with respect to alternative splicing. Thus, it was interesting to investigate whether any of the high-quality spectra unassigned in the initial search against the human IPI database could be assigned by searching more complete sequence databases.

To identify novel peptides or peptides containing sequence polymorphisms, the unassigned spectra were searched against a compressed human EST sequence database created and maintained by N. Edwards. This database represents the minimum size sequence database containing 30-amino acid-long sequences from the human EST database six-frame translation (44). To reduce the number of entries with sequencing errors, or simply redundant entries, the database was constrained to contain only those sequences that appear in an open reading frame of at least 50 unambiguous amino acids and that appear at least twice in the database. Importantly, the compact nature of the database allowed an order of magnitude reduction in the database search time compared with the search time against the original EST database. More detailed information regarding the construction of this database can be found on the corresponding website.

The search against the EST database resulted in the identification of seven peptides from 12 MS/MS spectra whose sequences were not present in the then-current version of the IPI database used in the initial search (version 2.35). The sequences of two of those peptides are now present in the latest version of that database (version 3.01), a positive development indicating continuing improvement of the protein sequence databases by their developers and annotators. In addition, the EST database search resulted in the identification of several sequence variants of known proteins. For example, a tryptic peptide, LSLGVFFPHDLR, from endonuclease G-like 1 protein (Swiss-Prot accession number Q9Y2C4) was identified from three MS/MS spectra, revealing a single nucleotide polymorphism (SNP) annotated in the dbSNP database (46) maintained by the National Center for Biotechnology Information (NCBI) (Val/Gly at the third position in the peptide sequence, refSNP identification number rs1141223). In another example, a tryptic peptide, LQGASTESYKDPTNIQLSK, was identified from two MS/MS spectra corresponding to an alternative transcript of the protein Spectrin α chain, brain (Swiss-Prot accession number Q13813, isoform 2), annotated in Swiss-Prot but not present as a separate sequence entry in the human IPI database.

Importantly, the EST database search also resulted in the identification of two novel peptides (see Supplemental Tables S4 and S5). One of the peptides, tryptic peptide LQGASTESYKDPTNEAQQVGHQHAR, is particularly interesting and deserves special attention. This peptide is present in four EST sequences and was identified by SEQUEST (Xcorr score of 2.8) and additionally confirmed by Mascot (expectation value, $5 \times 10^{-5}$) from one high-quality MS/MS spectrum (Fig. 7B). This intron-exon-spanning peptide identifies a novel splice variant of the Lck-interacting transmembrane adaptor 1 protein (LIME1). At the moment, the NCBI Entrez Protein sequence database contains only the ‘standard’ form of this protein, accession number NP_060276, which was shown to be a raft-associated protein in several recent studies (47, 48).

However, the identified peptide is present in several predicted alternative splice forms of LIME1 according to the AceView gene models (November 2004 version). All these predicted forms also contain a transmembrane domain and a putative raft-targeting motif, suggesting that all of them can be raft-associated. To gain further insight, the entire dataset (all SCX fractions from both the flow-through and ICAT experiments and without quality filtering) was researched against a database of all predicted LIME1 splice forms created based on the AceView gene models. This resulted in the identification of another MS/MS spectrum produced by the same peptide but in a different charge state, thus further increasing the confidence in this identification. However, no peptides from the standard form (or any other predicted form) were identified, suggesting that different forms of this protein can be expressed in different cell types. Although comprehensive follow-up analysis of these and other interesting findings was beyond the scope of this work, these results clearly illustrate that new biological insight can be gained by reanalysis of existing datasets. They also demonstrate the potential of mass spectrometry-derived data as a new source of information useful for validation and annotation of the genome (31, 40, 43, 49–51).

**DISCUSSION**

The quality of MS/MS spectra in shotgun proteomic experiments depends on multiple factors, including the complexity of the sample, the type of mass spectrometer, and the fragmentation parameters used (see e.g., Ref. 52). Moreover, even run to run variations in the signal to noise in MS/MS spectra acquired on the same instrument can be significant. The dynamic quality scoring method developed in this work can automatically adapt to a significant degree, to variations caused by acquisition methods or instrument to instrument differences. It is well suited to the problem of finding high

---

11 www.umiacs.umd.edu/~nedwards/research/pepseqdb.html.
12 Edward, N. J. (2005) Faster, more sensitive peptide identification from tandem mass spectra by sequence database compression, poster presented at the 1st Annual Symposium, United States Human Proteome Organization, March 12–15, Boston, MA.
13 www.ncbi.nih.gov/IEB/Research/Acembly.
Quality Assessment and Iterative Analysis of MS/MS Data

quality spectra left unassigned after the initial search of all spectra (with no loss of potential identifications due to filtering prior to database searching).

The basic idea behind the dynamic quality assessment method is that having the information on what spectra got assigned a peptide with high or low confidence during the first pass through the data, the notion of what constitutes a bad or good quality spectrum can be learned from the analyzed data itself, i.e. without relying on a training dataset of spectra taken from a different experiment. The statistical classifier is thus automatically developed for each dataset anew, making the method more robust toward variations in the MS/MS data than the static classifiers. Importantly it does not require creating training datasets using manual expert ranking (29),\(^2\) which is time-consuming and complicates adoption of the method to different instrument types, or generation of MS/MS data using samples of purified proteins (27, 28) in which case database searching tools, including X! Tandem, Mascot, and SpectrumMill. Many of the peptides containing modifications in the human lipid raft dataset corresponded to a base comprised of proteins already identified in the initial, fast sodiation) and for semitryptic peptides using a subset database search approach implemented in several existing databases invariably leads to high false identification rates simply due to increased database search size (22, 56). At the same time, even the basic follow-up analysis, not to mention

At the same time, the subset database search approach would likely be less successful in identifying peptides containing certain types of modifications, e.g. peptides with N-terminal acetylation. The strategy described in this work enables efficient analysis without limiting the protein search space because more comprehensive reanalysis is carried out only on those spectra that are of high quality and remain unassigned after the initial search. Both strategies, spectrum quality filtering and subset database searching, can be combined for even better computational efficiency. Quality scoring can also be combined with spectrum clustering (54, 55) to achieve even higher reduction in the number of spectra that need to be analyzed in each experiment. However, the development of efficient iterative database search strategies will also require new statistical models for combining peptide identifications obtained on different iterations and for computing accurate probabilities of their corresponding proteins.

It was also observed, in the human lipid raft dataset used here and in other similar datasets, that in most cases where a peptide containing a polymorphism was identified with high confidence, one or more related proteins (often the main variant of that protein) were already identified in the initial search against the IPI database. Similar to the subset database approach for the identification of modified peptides discussed above, one can largely limit the search for peptides containing polymorphisms to searching only those translated EST sequences that are homologues to proteins already identified by one or more peptides in the initial search against more compact databases such as IPI. At the same time, it remains to be seen whether the same approach can be used effectively for the purpose of identifying novel splice forms by first identifying the known protein forms and then performing a more comprehensive search around the corresponding regions of the genomic sequence. For example, this strategy would fail to identify the peptide from a novel splice variant of the LIME1 protein discussed above. Because novel or less characterized gene products are naturally of lower abundance relative to other proteins in the sample or have low molecular weight, many such proteins are likely to be identified by a single peptide.

The computational strategy that implements quality assessment and filtering, as described in this work, is particularly useful for the identification of sequence polymorphisms and novel proteins. Searching large numbers of MS/MS spectra against genomic databases requires significant computational resources. Even more importantly, searching unfiltered (i.e. including low quality spectra) datasets against such databases invariably leads to high false identification rates simply due to increased database search size (22, 56). At the same time, even the basic follow-up analysis, not to mention
the biological validation, of each identification of a novel peptide to understand its nature and significance requires querying multiple databases, performing protein and DNA sequence alignment, etc. and is very time consuming. Thus, large genomic databases, including EST databases, should be searched using MS/MS spectra of sufficiently high quality to minimize the number of incorrect identifications and to reduce the efforts spent on the interpretation of those erroneous findings.

In addition, in the case of low mass accuracy mass spectrometry data, such as ion trap data used in this work, unambiguous peptide identification is often not possible due to inability to resolve among different sequence alternatives (e.g. Ile/Leu, Asp/Asn, Glu/Gln/Lys, etc.). For example, it was observed that many MS/MS spectra with inaccurately determined precursor ion mass (mass error of more than 3 Da, which prevented their identification in the initial search) matched with high scores to incorrect (but highly homologous) peptides in the EST search. In addition, the EST search resulted in multiple peptide assignments to spectra that appeared to be identifications of sequence polymorphisms when in fact the true peptides that gave rise to those spectra were high abundance peptides containing a chemical modification. For example, when the unassigned high quality spectra in the human raft dataset were searched against the EST database, peptide TTVIMDSGHVTHTVPIYEGYALPHAIR was assigned to multiple MS/MS spectra with high scores, suggesting a polymorphism at position 10 in the peptide sequence (Asp/His substitution). However, the normal variant of this peptide (with Asp instead of His) was identified from more 40 MS/MS spectra in the initial search, and thus by far the most plausible explanation for the observed 22-Da mass difference is sodiation of the aspartate residue (addition of 22 Da to Asp, which makes it indistinguishable from His) rather than a point mutation. The abundance of such examples observed in this work warrants a general warning that, before reporting an ‘identification’ of a polymorphism or a novel peptide based on low mass accuracy data, it is necessary to eliminate the spectra for which a more likely explanation can be found (e.g. chemical modification of a high abundance peptide).

Application of multiple database search engines to the same dataset is not a new concept and is known to result in improved performance (see e.g. Ref. 57). The concept of meta-analysis has been discussed in the context of MS/MS database searching as well (22, 58). In this work, the reanalysis of the unassigned high quality spectra using multiple database search tools allowed identification of additional peptides and protein, although the overall increase in the number of proteins was not significant. A larger benefit of combining the results of multiple search engines could be in the improved discrimination between correct and incorrect identification, leading to an increase in the number of protein identifications at a certain fixed error rate. Still because the application of multiple database search tools increases the overall database search time, in many applications it should be sufficient to apply additional database search tools only to the remaining unassigned high quality spectra after performing the initial search using a single database search tool.

More comprehensive interrogation of high quality unassigned spectra performed in this work was largely limited to database searching using tools that are currently available in most proteomics laboratories. However, a number of other approaches have been described recently that could be better suited for the identification of modified peptides or sequence polymorphisms. These methods include de novo sequencing (Refs. 59–63; for a review, see Ref. 64), sequence tag-based error-tolerant database searching (8, 65, 66), and other approaches (26, 67–70). These computational algorithms can be used in conjunction with the quality scoring method described here as well. The human lipid raft dataset and the results of its comprehensive analysis such as the list of peptide and protein identifications, reported in this work, can serve as a useful test dataset for the development and benchmarking of new and improved peptide identification strategies. For example, it will be interesting to apply those tools to the high quality spectra from the human lipid raft dataset that still remain unassigned after all the additional efforts (see Supplemental Table S6). Moreover the analysis performed here with regard to the extraction of sequence tags from MS/MS spectra should be useful for optimizing peptide identification methods involving extraction of sequence tags and their use as low pass database filters (65, 66), for spectrum library searching (45), or for any other applications involving the separation of true fragments from noise peaks in the spectra.

The main aim of this work was to develop a robust approach for finding high quality spectra left unassigned in the initial database search and to investigate what prevented their identifications. A different application of the quality scoring method is simply to filter out low quality MS/MS spectra prior to database searching to reduce the computational time spent on the initial search (25). This would be a particularly attractive option in the case of MS/MS datasets generated using mass spectrometers such as LTQ, which provide much higher data acquisition rate at the expense of producing high numbers of low quality spectra. To use it for that purpose, the spectrum quality classifier can first be trained using a dataset of already searched MS/MS spectra acquired under similar experimental and acquisition conditions and then applied to all newly acquired spectra prior to database searching (without retraining the classifier). In a large scale study, a simple approach would be to search all spectra acquired in a few LC-MS/MS runs, train the classifier, and then apply it to all remaining spectra prior to database searching.

The choice of the filtering threshold is important for all practical applications and requires further discussion. The quality score threshold selected in this work (SQS > 1) rep-
resented a reasonable trade-off between the sensitivity and specificity of the quality score-based filtering, i.e., elimination of more than 95% of all spectra at the expense of losing less than 25% of additional peptide identifications that potentially can be extracted from the remaining unassigned spectra. The distribution of quality scores, however, is dataset-dependent, reflecting, among other factors, the overall proportion of high quality spectra in the dataset. Furthermore depending on the goals of the analysis, it might be desirable to filter the data with different stringency. The selection of the filtering threshold can be facilitated by the conversion of the SQS into a probability that the spectrum belongs to the class of good quality spectra, $p_{\text{good}}$. This conversion can be performed by modeling the observed distributions of spectrum quality scores as a mixture of the two underlying distributions constructed separately for each class (good and bad quality spectra) and then using those distributions to convert the spectrum quality scores into class probabilities using Bayes’ formula. The probabilities computed this way are more transparent than somewhat arbitrary SQSs and allow easy estimation of the sensitivity and specificity rates resulting from filtering the data using any minimum $p_{\text{good}}$ value as a filtering threshold.

Some of the spectrum features, introduced in this work, can also be used as additional parameters helpful in the process of validating peptide assignments to MS/MS spectra. Computational tools such as PeptideProphet and ProteinProphet, used in this work, already utilize an extended set of parameters in addition to the database search scores. As a result, the probabilities computed by those tools allow good discrimination between correct and incorrect identifications. Nevertheless those tools should benefit from the inclusion of certain spectrum features in their statistical model.

The calculation of spectrum quality scores in this work was based on 40 spectrum features, many of which are redundant or only marginally discriminative. It is possible that the number of spectrum features can be reduced without a significant decrease in the accuracy of the composite spectrum quality classifier. This question should be investigated more thoroughly in the future work using a more diverse set of experimental data (different instrument types and experimental conditions). If some of the spectrum features are indeed unnecessary under all conditions, they can be easily removed from the analysis if desired. However, the presence of redundant features does not create a problem for the method, and the increase in the computational time is insignificant. Furthermore new features, not considered in this work but possibly developed by others, could be easily added.

CONCLUSIONS

We developed a computational method for finding high quality spectra left unidentified in large shotgun proteomic datasets under most common conditions. The method does not require training datasets created using time-consuming manual ranking of MS/MS spectra or datasets generated using mixtures of purified proteins that might not adequately reflect the complexity of real data. Instead the quality scoring classifier is developed dynamically using the results of the initial database search pass through the data. This ensures the robustness of the method toward variations in the properties of MS/MS spectra caused by the differences in acquisition methods or instrument to instrument variability.

The practical utility of the method was investigated using a dataset of spectra from a human lipid raft experiment. Comprehensive computational reanalysis of the high quality spectra left unassigned in the initial search resulted in a 20% increase in the number of identified high quality spectra, including identifications of numerous modified peptides. Furthermore searching unassigned high quality spectra against the EST database led to the identification of peptides containing sequence polymorphisms and novel peptides, including a peptide confirming a computationally predicted splice form. It is thus expected that the strategy explored in this work will play an increasingly important role in the analysis of shotgun proteomic data and will help to realize the long touted potential of shotgun proteomic data as a rich source of information useful for the validation and annotation of the genomes.

The computational tool QualScore for finding unassigned high quality spectra will be made available freely and as an open source at www.proteomecenter.org.

Acknowledgments—We thank P. Widmayer and R. Jacob for contributions at an early stage of this work; P. von Haller and J. Watts for the human lipid raft dataset; Ulrike Schmidt for the vacuole dataset; S. Lin, N. Edwards, J. Eng, B. Fischer, V. Horejsi, J. Watts, and B. Wollischef for valuable discussions; and N. King and Brendan MacLean for technical assistance. We also acknowledge anonymous reviewers of the manuscript for comments.

REFERENCES
1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207
2. Apweiler, R., Bairoch, A., and Wu, C. H. (2004) Protein sequence databases. Curr. Opin. Chem. Biol. 8, 76–80
3. Yates, J. R. (2004) Mass spectral analysis in proteomics. Annu. Rev. Biophys. Biomol. Struct. 33, 297–316
4. Ferguson, P. L., and Smith, R. D. (2003) Proteome analysis by mass spectrometry. Annu. Rev. Biophys. Biomol. Struct. 32, 399–424
5. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation of shotgun proteomic data as a rich source of information useful for the validation and annotation of the genomes.
dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discov. Today 9, 173–181

6. Patterson, S. D. (2003) Data analysis—the Achilles heel of proteomics. Nat. Biotechnol. 21, 221–222

7. Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) Informatics for protein identification by mass spectrometry. Methods 35, 223–236

8. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic informatics. Int. Rev. Neurobiol. 61, 129–157

9. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989

10. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. 66, 4390–4399

11. Perkins, D. N., Pappin, D. J. C., Beausoleil, D. M., and Cottrell, P. C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

12. Clausner, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. 71, 2871–2882

13. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARs, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. Proteomics 2, 36–47

14. Zhang, N., Aebersold, R., and Schiwkowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2, 1406–1412

15. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectral data. Proteomics 4, 1406–1412

16. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, J. K., Adern, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Jonscher, K. R., Pierce, K. G., Old, W., and Hunter, L. (2004) Proteomic solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. Proteomics 2, 36–47

17. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1470

18. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, J. K., Adern, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Jonscher, K. R., Pierce, K. G., Old, W., and Hunter, L. (2004) Proteomic solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. Proteomics 2, 36–47

19. Colinge, J., Maselot, A., Giron, M., Desangis, T., and Magnin, J. (2003) OMS: towards high-throughput tandem mass spectrometry data identification. Proteomics 3, 1454–1463

20. Craig, R., Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1470

21. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392

22. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 5336–5346

23. Baldwin, M. A. (2004) Protein identification by mass spectrometry: issues to be considered. Mol. Cell. Proteomics 3, 1–9

24. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. L., Clausner, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data. Mol. Cell. Proteomics 3, 531–533

25. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wilets, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal. Chem. 76, 3556–3568

26. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzhizhsky, K. F., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer. Anal. Chem. 77, 5258–5268

27. Polevoda, B., and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. J. Mol. Biol. 325, 595–622

28. Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandenkerckhove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat. Biotechnol. 21, 586–591

29. Mann, M., and Pandey, A. (2001) Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. Trends Biochem. Sci. 26, 54–60

30. Edwards, N., and Lippert, R. (2004) Sequence database compression for peptide identification from tandem mass spectra, in Proceedings of the 4th Workshop on Algorithms in Bioinformatics (WABI 2004), Bergen, Norway (September 14–17, 2004), p. 230–241

31. Stein, S., Kilpatrick, L., Mautner, M., Neta, P., and Roth, J. (2005) Building
and using reference libraries of peptide mass spectra, in 53rd American Society for Mass Spectrometry Conference on Mass Spectrometry and Allied Topics, San Antonio, Texas (June 5–9, 2005), Abstr. MOEAm1155, American Society for Mass Spectrometry, Santa Fe, NM

46. Kitts, A., and Sherry, S. (2003) The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation, in The NCBI Handbook, pp. 5.1–5.30, National Center for Biotechnology Information, Bethesda, MD

47. Brdickova, N., Brdicka, T., Angelisova, P., Horvath, O., Spicka, J., Hilgert, I., Paces, J., Simeoni, L., Klíček, S., Merten, C., Schraven, B., and Horčíš, V. (2003) LIIME: a new membrane raft-associated adaptor protein involved in CD4 and CD8 coreceptor signaling. J. Exp. Med. 198, 1453–1462

48. Hur, E. M., Son, M., Lee, O. H., Choi, Y. B., Park, C., Lee, H., and Yun, Y. (2003) LIIME, a novel transmembrane adaptor protein, associates with p56Lck and mediates T cell activation. J. Exp. Med. 198, 1463–1473

49. Taylor, J. A., and Johnson, R. S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal. Chem. 73, 2594–2596

50. Chen, T., Kao, M. Y., Tepel, M., Rush, J., and Church, G. M. (2001) A dynamic programming approach to de novo sequencing via tandem mass spectrometry. J. Comput. Biol. 8, 325–337

51. Croft, W. B. (2000) Combining approaches to information retrieval, in Advances in Information Retrieval (Croft, W. B., ed) pp. 1–36, Kluwer Academic Publishers, Dordrecht, Netherlands

52. Standing, K. G. (2003) Peptide and protein de novo sequencing by mass spectrometry. J. Chromatogr. B 768, 291–298

53. McCowan, S. J., Terrett, J., Brown, C. G., Adam, P. J., Aldridge, L., Allen, J. C., Amess, B., Andrews, K. A., Barnes, M., Barnwell, D. E., Berry, J., Bird, H., Boyd, R. S., Broughton, M. J., Brown, A., Bruce, J. A., Brusten, L. C. M., Draper, N. J., Elsmore, B. M., Freeman, C. D., Giles, D. M., Gong, H., Gormley, D., Griffiths, M. R., Hawkes, T. D. R., Haynes, P. S., Heesom, K. J., Herath, A., Hollis, K., Hudson, L. J., Inman, J., Jacobs, M., Jarman, D., Klířová, I., Klířův, J. J., Knuthia, S. K., Lane, K. E., Lees, M. L., Loader, J., Longmore, A., McEwan, M., Middleton, A., Moore, S., Murray, C., Murray, H. M., Myatt, C. P., Ng, S. S., O’Neill, A., Parekh, R. B., Patel, A., Patel, K. B., Patel, S., Patel, T. P., Philip, R. J., Platt, A. E., Poyser, H., Prendergast, C., Prime, S., Redpath, N., Reeves, M., Robinson, A. W., Rohlf, C., Rosenbaum, J. M., Schenker, M., Scrivener, E., Shipston, N., Siddiqi, S., Southan, C., Spencer, D. I. R., Stamps, A., Steffens, M. A., Stevenson, D., Sweetman, G. M. A., Taylor, S., Townsend, R., Ventom, A. M., Waller, M. N. H., Weresch, C., Williams, A. M., Woolfscroft, R. J., Yu, X., and Lyall, A. (2004) Annotation of the human genome by high-throughput sequence analysis of naturally occurring proteins. Curr. Proteomics 1, 41–49

54. Wenner, B. R., and Lynn, B. C. (2004) Factors that affect ion trap data-dependent MS/MS in proteomics. J. Am. Soc. Mass Spectrom. 15, 150–157

55. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. Mol. Cell. Proteomics 4, 1180–1188

56. Beer, I., Barnea, E., Ziv, T., and Admon, A. (2004) Improving large-scale proteomics by clustering of mass spectrometry data. Proteomics 4, 950–960

57. Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., Verberkmoes, N. C., and McDonald, W. H. (2005) MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. J. Am. Soc. Mass Spectrom. 16, 1250–1261

58. Colinge, J., Cusin, I., Reffas, S., Mahe, E., Niknejad, A., Rey, P. A., Mattou, H., Moniatte, M., and Bouguereler, L. (2005) Experiments in searching small proteins in unannotated large eukaryotic genomes. J. Proteome Res. 4, 161–174

59. Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. 6, 327–342

60. Tabb, D. L., Saraf, A., and Yates, J. R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem. 75, 6415–6421

61. Frank, A., Tanner, S., Bafna, V., and Pevzner, P. (2005) Peptide sequence tags for fast database search in mass-spectrometry. J. Proteome Res. 4, 1287–1295

62. Hernandez, P., Gras, R., Frey, J., and Appel, R. D. (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3, 870–878

63. Liska, A. J., and Shevchenko, A. (2003) Combining mass spectrometry with database interrogation strategies in proteomics. Trends Anal. Chem. 22, 91–99

64. Searle, B. C., Dasari, S., Turner, M., Reddy, A. P., Choi, D., Wilmart, P. A., McCormack, A. L., David, L. L., and Nagalla, S. R. (2004) High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal. Chem. 76, 2220–2230

65. Hansen, B. T., Davey, S. W., Ham, A. J., and Liebler, D. C. (2005) P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. J. Proteome Res. 4, 358–368