Published in final edited form as:

Anal Bioanal Chem. 2012 September ; 404(4): 1115–1125. doi:10.1007/s00216-012-6011-x.

Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines

Yao-Yi Chen

Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575, USA

Surendra Dasari

Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575, USA

Ze-Qiang Ma

Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575, USA

Lorenzo J. Vega-Montoto

Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575, USA

Ming Li

Division of Cancer Biostatistics, Vanderbilt University Medical School, Nashville, TN 37232-6848, USA

David L. Tabb

Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575, USA david.l.tabb@vanderbilt.edu

Abstract

Spectral counting has become a widely used approach for measuring and comparing protein abundance in label-free shotgun proteomics. However, when analyzing complex samples, the ambiguity of matching between peptides and proteins greatly affects the assessment of peptide and protein inventories, differentiation, and quantification. Meanwhile, the configuration of database searching algorithms that assign peptides to MS/MS spectra may produce different results in comparative proteomic analysis. Here, we present three strategies to improve comparative proteomics through spectral counting. We show that comparing spectral counts for peptide groups rather than for protein groups forestalls problems introduced by shared peptides. We demonstrate the advantage and flexibility of this new method in two datasets. We present four models to combine four popular search engines that lead to significant gains in spectral counting differentiation. Among these models, we demonstrate a powerful vote counting model that scales well for multiple search engines. We also show that semi-tryptic searching outperforms tryptic searching for comparative proteomics. Overall, these techniques considerably improve protein differentiation on the basis of spectral count tables.

Keywords

Label-free comparative proteomics; Spectral counting; Combining database search engines

Introduction

Shotgun proteomics based on tandem mass spectrometry has become a widespread method for analyzing complex biological mixtures. It begins by digesting protein mixtures and separating the resulting peptides by liquid chromatography. After peptide MS/MS spectra are acquired, they are matched to database peptide sequences by search engines such as Sequest [1], Mascot [2], X!Tandem [3], and MyriMatch. Proteins are assembled from these raw identifications by validation tools [4–8] that convert arbitrary search scores into statistical measures [9]. Proteins can then be filtered by customized criteria for further analysis. Because shotgun analyses can represent complex proteomes in considerable depth, a key question is how one can compare shotgun proteome inventories to reveal molecular characteristics of biologically distinct phenotypes to discover clinically important biomarkers. Improvement in protein differentiation broadly benefits the identification and validation of molecular markers that relate to various biological or medical outcomes, thus improving the current state of the art in biological research and clinical practice.

Prior work has demonstrated that a frequency-based analysis approach using the number of observed spectral counts for each protein provides a rough measure of protein levels in complex protein mixtures, especially for more abundant proteins [10–12]. In shotgun proteomics, however, a particular peptide may correspond to multiple potential protein sources. In systems where proteins of multiple species are present, such as xenograft models of cancer, shared peptides are very common, and so, a difference in one protein may masquerade as a difference in a second protein that shares peptides with the first. Several approaches have been proposed to help solve this problem [4, 13–15] by employing protein parsimony, discarding shared peptides, or adjusting spectral counts of shared peptides by unique peptides. These approaches have the disadvantage of losing information or attempting to apportion large numbers of spectra on the basis of relatively small sets of differentiating spectra. IDPicker, a protein assembly tool [4, 9], organizes peptides into groups when they match identical sets of proteins, and it similarly organizes proteins into groups when they match identical sets of peptides. This structure enables the development of methods to differentiate proteomes in units of “peptide groups” that do not overlap with each other.

Comparative proteomics spans two complementary goals. First, researchers may seek to *differentiate* the proteomes of two sample cohorts, seeking the proteins that appear in one sample to a significantly greater degree than in another. Second, researchers may seek to *quantify* the extent to which proteins change in magnitude between sample cohorts. This article considers the first of these goals, leaving quantification as a topic for experimental methodologies better designed for this purpose, such as selected reaction monitoring [16]. The evidence produced for a protein in “shotgun” experiments is the result of a high-throughput sampling process. As a result, which spectra are captured from a particular protein digest will vary among experiments [17]. Spectral counts attributed to a particular protein group may vary due to random sampling or due to differences in protein quantity. In general, one expects to collect both more spectra from individual peptides (potentially varying in charge state or modification state) as well as more peptides from a particular protein group as the concentration of that protein rises compared to the sample background. As a result, finding significant differences requires the ability to compare variation in replicates to variation between cohorts.

Spectral counting depends upon identification, and yet little evaluation of its dependence on search engines has appeared in the peer-reviewed literature. Search results differ from one search engine to the next, depending on both the type of mass spectrometer used and the configuration of the search. In biological samples, often the most interesting proteins are lowest in abundance, and meaningful changes in protein abundance may be small in magnitude. Detecting these differences may be visible by one search engine but not another because of differences in match scoring. Even if the search engine is held constant, the way in which the tool is configured may significantly impact the set of identifications produced. Deciding between a “fully tryptic” search and a “semi-tryptic” search would seem to primarily impact the amount of time required, but this decision has been shown to significantly alter the set of peptides identified from a mixture [9, 18]. The impact of trypsin specificity configurations on protein differentiation has not been considered in depth. Stepping beyond a single search scenario, researchers have demonstrated that collating results from multiple search engines improves sensitivity for inferring protein inventories [19–21], so long as false positives are kept under control. It would seem that the improved coverage available through multiple search engines would be a boost for differentiation, as well. How to leverage the increased information yield, however, has not yet been described.

In this manuscript, we characterize three strategies for improving comparative proteomics through spectral counting. We demonstrate that the problem of shared peptides can be resolved through comparison for peptide groups rather than proteins, giving examples of differences that would be confused by standard approaches. We examine the gains achieved for spectral counting when collating search results from a set of four high-performance peptide identifiers. We determine the impact of tryptic and semi-tryptic searching for spectral count tables to frame recommendations for best practices. Taken together, these techniques enable higher-quality differentiation on the basis of spectral count tables.

Materials and methods

Data sources

ABRF data—We used a dataset from the Association of Biomolecular Research Facilities (ABRF) iPRG 2009 study. In that study, two samples of *Escherichia coli* lysates (labeled “red” and “yellow”) were digested with trypsin then analyzed with LC-MS/MS on an LTQ-Orbitrap with five technical replicates for each sample. The red and yellow replicates were derived from the same *E. coli* lysate sample running on two halves of one gel with a single region excised from each half (The “green” and “blue” proteomic datasets). Proteins in the changing region of red and yellow cohorts were enriched in blue and green cohorts, respectively (for more information, see Fig. S1 in the Electronic supplementary material). A differential protein key list was built by comparing the differentially expressed proteins between the less complex blue and green cohorts with a significance level of 0.05. Eighty-five percent of the proteins in the key list corresponded with the mass regions excised from the gel. The proteins significantly expressed in the blue cohort that were also significantly expressed in the red cohort were considered as true positives. Similarly, the proteins significantly expressed in the green cohort that were also significantly expressed in the yellow cohort were considered as true positives.

CPTAC data—We used a dataset created by the Clinical Proteomic Technology Assessment for Cancer (CPTAC) program [17]. In the study, a yeast lysate was spiked with a mixture of 48 human proteins (Sigma-Aldrich UPS1) at several levels of concentrations. Each sample was analyzed with triplicates on seven independent instruments of four models (Thermo Fisher LTQ, LTQ-XL, LTQ-XL-Orbitrap, and LTQOrbitrap). Groups A, B, C, D, and E were yeast spiked with UPS-1 at 0.25, 0.74, 2.2, 6.7, and 20 fmol/μl, respectively.

Data were processed using a FASTA database combining the yeast and human proteomes. Search parameters are provided in the Electronic supplementary material 1.

HNSCC data—The Head and Neck Tissue Repository at Vanderbilt University [22] collected 20 head and neck squamous cell carcinomas (HNSCC). These cancerous samples can be compared to 20 normal tonsillectomy tissues, also from this repository. Peptides were separated by isoelectric focusing and liquid chromatography, followed by MS/MS analysis on a LTQ-Orbitrap.

ASW480 data—Adenomatous polyposis coli (APC) is a negative regulator of Wnt signaling. Mutation of APC occurs in up to 60 % of colorectal cancer tumors. Patrick Halvey at Vanderbilt University has examined the proteomes of two colon tumor cell lines: SW480APC (APC restored) and SW480Null (mutant APC). Biological samples from three independent cell cultures were injected in duplicate for a total of six replicate measurements for the SW480null and SW480APC cell lines. Peptides were separated by isoelectric focusing and analyzed by LC-MS/MS on an LTQ. A subset of proteins found to be differentially expressed by LC-MS/MS were validated by targeted proteomics (LC-MRM-MS) [23]. The results from label-free shotgun proteomics were confirmed by targeted proteomics in that study.

Database search pipeline

MS/MS scans were converted to mzML file format by the msConvert tool in the ProteoWizard [24] library to provide input files for TagRecon (TR) [25], MyriMatch (MM) [26], and X!Tandem (XT) search. These files were then converted to DTA format by ScanSifter [25, 27] to enable Sequest (SQ) search. All protein databases contained sequences in both forward and reverse orientations for estimation of protein and peptide identification error rates. For LTQ data, MM, TR, and XT applied a precursor tolerance of 1.25 m/z, while SQ applied a 2.5-Da mass tolerance. For Orbitrap data, MM and XT applied a precursor tolerance of 10 or 40 ppm, while TR applied 0.01 m/z tolerance and SQ applied a 0.1-Da mass tolerance. The search results were processed by IDPicker to yield a 5 or 2 % false discovery rate (FDR). Peptides passing these thresholds were considered as legitimate identifications. IDPicker assembled protein identifications from peptides using parsimony rules [4, 9].

Statistically significant differences in protein spectral counts between different groups were calculated using quasi-likelihood generalized linear modeling (GLM) by QuasiTel [22]. Proteins with p values less than 0.05 were considered as differential proteins. Differentially expressed proteins were mapped to genes and compared for enrichment of defined classes against a reference set of all identified proteins. Search configurations, dataset information, and identified peptides are shown in Table S1 and S4, and Electronic supplementary material 2.

Model for peptide group-based spectral count differentiation

IDPicker generates tables reporting the number of spectral counts for each peptide group (Fig. 1). We used Fisher's exact test instead of GLM to compute a p value for each peptide group because the GLM includes additional covariates in the comparisons which may diminish accuracy for peptide groups with low spectral counts. We also used Fisher's exact test to compute a p value for each protein group as a comparison method. We employed the Benjamini–Hochberg FDR method to correct p values for multiple hypothesis testing [28]. Statistical techniques for the peptide group-based analysis differed from those employed in the search algorithm combination and semi-trypsin evaluations. These latter examinations employed the standard QuasiTel GLM for differentiation.

Common data analysis practices in comparative proteomics reflect the belief that FDR (multiple hypothesis testing corrected p value) is good both as a qualitative and a quantitative indicator of the overall significance of the results. The use of FDR-based meta-analysis was previously demonstrated in ChIP-chip meta-analysis [29]. The corrected p values of peptide groups corresponding to the same protein group were combined using Stouffer's Z -reverse normal transform method [30] to estimate the significance level of changes at the protein group level.

The weighted Stouffer's inverse normal transform method we built, described in Eqs. (1) and (2), took peptide p value, sample size, and effect direction (4) into consideration to compute a protein p value. Optimal weights for the weighted Z method were given by the square root of the spectral counts of peptide groups divided by their occurrence in protein groups (3). By this strategy, unique peptide groups are assigned higher weights than the shared peptides.

$$S_{\text{pro}} = \sum_{s=1}^{N_s} w_s d_s \phi^{-1} \left(1 - \frac{P_{\text{pep}}(S)}{2} \right) \quad (1)$$

$$p_{\text{pro}} = 2 \left(1 - \phi(|S_{\text{pro}}|) \right) \quad (2)$$

$$w_s = \sqrt{\frac{\text{peptide} \frac{S_p C(s)}{\text{occurrence}(s)}}{\sum_{i=1}^{N_s} \text{peptide} \frac{S_p C(i)}{\text{occurrence}(i)}}} \quad (3)$$

$d_s = +1$ for increased; -1 for decreased; 0 for unchanged from one sample to another * ϕ and ϕ^{-1} denote the standard normal cumulative distribution function and its inverse.

Models for combining search engines

We present statistical models to combine search results from four search engines. Heterogeneity among search engines results from factors including spectral pre-processing, theoretical spectrum prediction, and match scoring algorithms. As a result, FDR-based meta-analysis was necessary to summarize results. In the first model, spectral counts from each search engine were added together prior to differentiation. The combined spectral counts were analyzed by QuasiTel and corrected by the FDR method to compute p values. In the second model, we computed FDR-corrected p values of protein spectral counts separately by search engine. These p values were then combined for each protein using Stouffer's Z -transform probability test [31]. In the third model, we ranked the proteins by FDR-corrected p values from individual search engines (from smallest to largest). The ranks were then added together to compute a super rank for each protein. In the “Stouffer p -combo model” and “ p -rank sum model,” proteins that were not identified by any included search engine were excluded in the comparison.

Vote counting is well described for use in microarrays and peptide identifications [32, 33]. Rhodes et al. used a comparative meta-profiling which assesses the overlap of gene expression differentiation from a diverse collection of microarray datasets. Several modifications enable its use for protein spectral count differentiation. Briefly, the spectral count data were analyzed by QuasiTel, and p values from individual search engines were FDR corrected. We then defined a significance threshold— α ($\alpha_{\text{DEFAULT}}=0.05$)—and the number of top proteins we wanted to select— N_{SELECT} . For these thresholds, we then ranked proteins by the number of search engines that find each significant; this positions each

search engine as a “voter.” Within each class of proteins with the same vote counts, we then ranked proteins by the minimum of their p value from the combining search engines (minimum p value, increasing). This process ranked potential protein differences, with the most substantial changes at the top.

Assessing FDR for vote counts and best p values followed a permutation strategy. First, we counted the proteins for each possible number of vote counts ($N_1, N_2 \dots N_S$). Permuting the p values per search engine among proteins generated a set of randomly produced differences. We counted these differences for each possible number of vote counts ($E_1, E_2 \dots E_S$). The minimum meta-false discovery rate (mFDRmin) can then be calculated by:

$$\text{mFDRmin} = \text{minimum} \left(\frac{[E_i + 1]}{[N_i]} \right) \text{ for } i=0 \text{ to } S$$

Then, we assess the validity of α with the following criteria: If $\text{mFDRmin} < \alpha$, these proteins were found to be differentially expressed at the threshold α . If not, we repeated the enumeration of votes with the value of α lowered by 20 % at each iteration until either a valid α is defined or the number of differential proteins detected in two or more search engines reaches 0. A valid α should not fall so far that the number of proteins with at least one vote was less than N_{SELECT} . Furthermore, to be strict in the significant level of the threshold, we should find the smallest (most significant), valid α setting by lowering α by 20 % and repeat the previous validity testings iteratively. The algorithm was implemented in R (Electronic supplementary material 3). This model is tuned for the best performance when voter turnout is large, i.e., more search engines are deployed for each dataset.

Results and discussion

Peptide group-based spectral count differentiation improves protein differentiation

Peptide group-based spectral count differentiation better evaluates the impact of unique and shared peptide groups on protein differentiation, thus effectively reducing false positives. This method is most effective in reducing false positives when working with proteomic samples of higher organisms where a lot of shared peptide groups exist. Therefore, we tested the technique in ASW480 and HNSCC human proteomic datasets. In the ASW480 dataset, 6,042 peptide groups were identified, mapping to 7,325 proteins in 5,215 protein groups. We compared the cell line with and without the APC vector with protein group-based and peptide group-based techniques after MyriMatch search and IDPicker filtering. Of the differentiating proteins discovered by peptide group analysis, 95 % were also discovered through protein group analysis. Correspondingly, 81 % of the differential proteins from protein-based differentiation were also identified by peptide group-based differentiation (Figs. S2 and S4). At first, this would seem to imply higher sensitivity to differences in protein group analysis, perhaps due to more aggressive p value correction in the more numerous peptide group comparisons. Only five proteins were identified exclusively by peptide group-based differentiation, while 21 proteins were differentiated by protein-based but not peptide-based techniques. We examined these 21 proteins with a critical eye. In the example of protein groups for desmin and vimentin, four peptide groups were shared between desmin and vimentin and six other protein groups (Table S2). The p values of desmin and vimentin from protein-based spectral count differentiation were 0.023 and <0.00001 , respectively, signifying that these two proteins were both differentially expressed. However, we found that the spectral count of the unique peptide group of desmin had not significantly changed (p value >0.05). Desmin and vimentin share four peptide groups that were also shared by two to four other protein groups, causing cross talk between these proteins and others that were legitimately changing. The spectral count of these peptide

groups greatly impacted the total spectral count of desmin. These data demonstrate that shared peptides can cause unchanging proteins to become false-positive differences.

The p value for desmin was 0.155111 when differentiation was performed at the level of peptide groups with combination via Stouffer's inverse normal method [30]. Separating peptides by protein association revealed that the expression level of desmin had not significantly changed. On the other hand, the change of the vimentin level remained significant (p value <0.01). In fact, the lack of change for desmin was reinforced by microarray (p value of 0.98747) [23]. On the other hand, the enrichment analysis of proteomic data revealed that targets of transcription repressor ZEB1 were measured at lower levels in the SW480 null cell line, implying elevated ZEB1 activity in this cell line. Others have shown that disruption of the ZEB1/SMARCA4 binding causes an increase in CDH1 expression and a decrease in vimentin [34]. We also compared the two methods between replicates of the APC or control group which we knew should not show any differential proteins. Peptide group-based differentiation reduced the false-positive differentiation by 20–41 % (Fig. S3). These facts have shown that peptide group-based differentiation is robust against false positives induced by shared peptides.

Peptide group-based spectral count differentiation is also more sensitive to changes in unique peptide groups. In the HNSCC dataset, 4,011 proteins were assembled to 2,569 protein groups, with 2,941 peptide groups mapping to them. One hundred differential proteins were identified by peptide group-based differentiation (Fig. S2). As a test, we evaluated the biomarker set resulting from a comparison using only the peptide groups that mapped to a single protein group; limiting the information to this set of peptides, however, reduced detection of differentiating proteins by 22 % (Fig. S2). Of the proteins found to be differences from the peptide group-based technique employing all peptides, 94 % were also found through the protein-based technique. Of the protein-based difference set, 82 % were also observed through peptide-group differentiation. A majority of protein changes found by peptide group-based differentiation shared peptides with other protein groups. Myosin 14 was among the differences found by peptide group-based but not protein group-based techniques. This non-muscle myosin, which appears to play a role in cytokinesis and cell shape, was matched to five peptide groups (Table S3). Protein-based spectral count differentiation could not provide enough evidence (p value = 0.258 > 0.05) to show that myosin 14 was differentially expressed in the cancer group versus the control group. However, when we look closely into each peptide group, we find that the peptide group that contains sequences specific to this form of myosin changes significantly in spectral counts, increasing from 39 to 110 (2.82-fold, p value = 0.000388 < 0.05). By peptide group-based spectral count differentiation, the difference is significant (p value = 0.001588 < 0.05). Previous studies have shown that overexpression of myosin 14 inhibits cell growth [35], which coincides with the heightened expression in normal samples. Without peptide group-based comparison, this difference would be masked by other myosin forms.

Generally, protein and peptide group-based differentiations are highly concordant with each other (Fig. S4). The correlation coefficient for the p values of ASW480 proteins was 0.947, while the HNSCC set yielded a 0.942 correlation. After finding the differential proteins by p values, the fold change of a protein can be estimated by averaging the fold change of its peptides. Because there are more peptide groups than protein groups for an assembly, multiple testing adjustment reduces the count of significant differences more strongly for peptide groups than for protein groups. For example, of the 20 proteins that were disagreements between the two differentiation techniques in the ASW480 dataset, three proteins (CD2 antigen cytoplasmic tail-binding protein 2, envoplakin, heat shock protein beta-1) are proteins with only one peptide group. As a result, the set of spectral counts compared in protein group and peptide group techniques is the same. Once multiple testing

correction has been applied, though, Envoplakin shifts to a 0.0446 p value from protein group evaluation or to an insignificant 0.0581 p value from peptide group evaluation. Whether this constitutes the removal of a false-positive difference or losing sensitivity for real differences cannot be resolved from the data on hand.

Combining multiple search engines improves protein differentiation

Protein differentiation is considerably affected by search algorithms. In the ABRF iPRG *E. coli* dataset, 1,275 proteins in total were identified by the four search engines, while only 662 proteins were shared between all four search engines. The ability to identify truly differentiated proteins also varied among different search engines. MM, TR, XT, and SQ each identified 228, 225, 226, and 207 truly differentiated proteins, respectively (Fig. S5). Most truly differential proteins (derived from identifications in the “blue” and “green” samples) reach agreement between two or more search engines with consistent fold change directions. These results highlighted the necessity of combining search engines to detect more correct differences and reduce false discoveries. We applied four distinct models (see “Materials and methods” section) to combine different search engines. These models have shown their unique advantages to achieve better protein differentiation. We ranked the proteins by p values from the “count sum model” and “Stouffer p -combo model” and by super rank of “ p -rank sum model” from smallest to largest, or by vote counts from the “vote counting model” from largest to smallest, and chose the top 250 proteins (approximately the length of the key list) for true-positive and false-positive analyses. As shown in Fig. 2, generally, combinations of search engines outperform individual search engines. For the pairing of SQ and TR, the “Stouffer p -combo model” increased AUC by 12.7 %, from 0.796 to 0.897, and identified 18 more true-positive proteins than TR by itself. Combining all four search engines by the “ p -rank sum model” identified 3–13 % more true positive proteins than for any individual search engine; this combination revealed that adding all possible search engines is not guaranteed to outperform a well-selected set of search engines, since the MM+TR+SQ combination was more effective. Of the search engine pairs, XT and SQ appeared least effective at complementing each other.

Combining all the four search engines with the “vote counting model” produced the best true-positive ratio and lowest false-positive ratio, with 177 true positives out of top 250 differences, while the best number of true positives of other models is only 167. The “vote counting model” identified 20.5, 22.1, 22.1, and 22.9 % more true-positive proteins than searching by MM, TR, XT, or SQ individually. Combining three search engines such as MM+TR+SQ is also effective.

Combinations of search engines by these models were also evaluated in the context of the CPTAC LTQ dataset. We used data from C and E cohorts (a ninefold difference of UPS-1 spike concentration). In total, 45 out of 48 UPS-1 proteins were identified by the four search engines. MM, TR, XT, and SQ identified 42, 42, 41, and 40 UPS-1 proteins, respectively (Fig. S6). We ranked the proteins by p values, super rank, or vote counts and analyzed the top 50 proteins with the four models. Numbers of true positives among the top 50 proteins were compared in Fig. 3. Again, combinations of search engines outperformed individual search engines for revealing protein differences. Combining TR and SQ with the “Stouffer p -combo model” generated 32 % more true positives than SQ individually. Combining all four search engines by the “count sum model” identified 17.9–50.0 % more true-positive proteins than individual search engines.

Again, combining all the four search engines with the “vote counting model” produced one of the best true-positive ratio and lowest false-positive ratio, with 33 true positives out of top 50 differences. The “vote counting model” identified 17.9, 43.5, 22.2, and 50.0 % more true-positive proteins than searching by MM, TR, XT, or SQ individually. The advantage is not

distinctive here because of the small number of proteins in the “answer key.” Combining only two search engines was helpful for one dataset but not the other; the voting model benefits from a larger pool of votes (Figs. 2 and 3). For example, MM+XT, TR+XT, and MM+TR only identified around 150 true-positive proteins. In the ABRF dataset, only the combinations that included Sequest gave the highest performance, though this algorithm working alone had yielded the lowest number of true differences.

The four models for combining search engines have different strengths and weaknesses. In simply adding spectral counts for a protein identified by multiple search engines, a single spectrum might be counted multiple times. Although the multiple counting increases the confidence of identification and spectral count differentiation, it will get extreme p values because of the correlation between search results. In the “Stouffer p -combo model,” combining p values among algorithms increases the sensitivity of the collective analysis but has risks of bias towards idiosyncratically significant p values of one search engine. In the “ p -rank sum model,” the super rank comprises a non-parametric assessment of the results from individual engines. Drawing conclusions about which of these techniques is best would over-generalize from the two sample sets evaluated in this study, though combination is clearly beneficial. The “vote counting model” was most powerful when combining more search engines. Overall, combining search engines improves protein differentiation by not only increasing the protein inventories, but also increasing the pool of information available to differentiate each protein. Each combination of search engines allows for better discrimination than any individual search engine.

Semi-tryptic search outperforms tryptic search in protein differentiation

A given search engine may yield a different performance depending on its configuration. Bioinformaticists have argued for years that semi-tryptic searching, which allows the identification of peptides that differ from canonical trypsin specificity on one terminus, improves the inventories possible from proteomics [36]. We tested this parameter for its impact on comparative proteomics. Table S4 reports the number of identified peptides by fully tryptic and semi-tryptic searches.

We first compared red/yellow cohorts in the iPRG *E. coli* dataset. All the other configurations and analysis were identical. Figure 4 shows the ROC curve of differentiated protein expression with semi-tryptic or fully tryptic searches. Semi-tryptic search achieved better sensitivity and specificity than fully tryptic search, with AUC increased by 6 % (from 83.77 to 88.56 %). Similarly, when comparing true-positive and false-positive proteins at the cut point of p value 0.05, semi-tryptic search greatly increases true-positive proteins by 7.07 % for the same number of false positives. The improvement reveals that semi-tryptic search achieves better sensitivity and specificity than fully tryptic search for a sample in which many proteins offer stark differences between cohorts.

We next analyzed the CPTAC dataset (where the spiked proteins differed by a factor of three between each pair of five levels) with fully tryptic and semi-tryptic searches. We compared the spectral counts of proteins in these cohorts in pairs (Table 1). We chose a sampling of the possible fold changes, preferring samples where spike concentrations were greater. Semi-tryptic search generally outperformed fully tryptic search in AUC. Especially in the D and E cohorts, where UPS1 proteins were most dominant, semi-tryptic search increased AUC by 5.5 % (from 86.76 to 91.50 %).

The top 50 (approximately the number of proteins in the gold standard) most differentiated proteins for each pairwise comparison were evaluated against the list of proteins known to change, and the numbers of true positives and false positives were computed (TP/FP). At different spike levels, semi-tryptic search detects more true-positive proteins along with

fewer or unchanged false-positive proteins. Especially in B vs C, which contained only small amounts of spiked proteins with a threefold concentration difference, semi-tryptic search identified 55 % more true positives than fully tryptic search. Generally, semi-tryptic search provides better sensitivity and specificity than fully tryptic search, especially when comparing groups with small spike-in protein concentration changes (D vs E, C vs D).

Why would adding semi-tryptic peptide improve protein differentiation? When an algorithm fails to identify a spectrum, a semi-tryptic search will typically assign a semi-tryptic peptide to the spectrum (because random semi-tryptic peptides outnumber fully tryptic peptides by more than an order of magnitude). Software that separates correctly identified spectra from incorrectly identified ones exploits this information to identify a larger set of peptides, even if no semi-tryptic peptides are present. The most abundant proteins in a mixture are, in turn, more likely to produce semi-tryptic peptides in addition to fully tryptic peptides. As the concentration of UPS-1 proteins increases from group A to group E, the percentage of semi-tryptic peptides from these UPS-1 proteins was 0 % in group A and group B. The percentage increased to 6.9–7.0 % in group C and group D, and reached the highest—10.6 %—in group E. The increased identification of semi-tryptics from dominant proteins increases the power of semi-tryptic search in protein differentiation and expands the dynamic range of differentiation.

Conclusions

Spectral count differentiation benefits from a peptide group-based evaluation strategy, new models for combining database search engines, and care in search configuration. Peptide group-based spectral count differentiation helps to resolve the protein inference problem, giving particular power when untangling complex protein–peptide clusters. It can be used as an alternative or complementary differentiation method when working with complex comparative proteomic samples where a lot of shared peptide groups exist. In systems where proteins of multiple species are present, such as xenograft models of cancer or other samples that contain proteins from multiple eukaryotes, the method has great potential in improving protein differentiation. Due to the influence of multiple testing adjustment, this method may lose power for proteins near the p value threshold.

Three of the four tested models for combining search engines for differentiation proved to be effective. The “count sum model” can be easily implemented for almost any workflow and delivers solid performance, though false positives may prove problematic. The “ p -rank sum model” may be more robust against idiosyncratic performance for individual search engines. These two models can be used when combining two or three search engines; in this examination, MM+TR+SQ yielded the best performance. With the increased ability of incorporating three or more search engines, the “vote counting model” is very robust against idiosyncratic results for individual search engines. Its steady, high performance in these datasets suggested great potential for fielding many search engines at once.

These models may be most useful in biomarker discovery, where some proteins of interest are at low abundance. The use of multiple engines can broaden the pool of information available to differentiate proteins present at small quantities. These models can also apply to samples with large genomes when low-resolution mass analyzers have measured precursor masses; these searches compare very large numbers of candidate sequences to every spectrum, thus losing discrimination. In the future, these models may be developed by recognizing the unique contribution of each search engine. The search engines that provide more confident IDs with better sensitivity and specificity, such as MM in the two datasets above, should be afforded more importance. In the “count sum model,” excluding the

overlapping peptide spectrum matching by different search engines can also be used to reduce the type I error.

In both datasets, semi-tryptic peptide search outperforms fully tryptic peptide search in protein differentiation studies in multiple aspects including higher discovery rate, better specificity, and better sensitivity. Semi-tryptic search is more sensitive to small protein concentration changes. Ignoring the contributions of semi-tryptic peptides would sacrifice discrimination for levels of abundant proteins. If endogenous proteases are present in a sample, semi-tryptic search is obviously the choice for better protein differentiation, but the improved inventories are feasible through this option even in samples dominated by fully tryptic peptides. In the future, more general conclusions can be drawn by in-depth analysis of trypsin specificity configurations by search engines other than MM.

In conclusion, these three strategies yield higher-quality differentiation based on spectral counting. These strategies are each generic enough to enable their incorporation in many bioinformatics pipelines. Since the spectral counting strategy was introduced in 2004, it has become a standby for many laboratories. These advances will enable its application to samples where proteins share peptides in complex relationships, discrimination of correct peptides requires multiple pipelines, and a wide dynamic range of proteins is interrogated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

D. L. Tabb and Y.-Y. Chen were supported by U01 CA152647 from the National Cancer Institute. S. Dasari, Z.-Q. Ma, and L. J. Vega-Montoto were supported by R01 CA126218 from the National Cancer Institute.

Abbreviations

FP	False positive
MM	Myrimatch
SQ	Sequest
TP	True positive
TR	TagRecon
XT	X!Tandem

References

1. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom.* 1994; 5:976–989.
2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999; 20(18):3551–3567. [PubMed: 10612281]
3. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinforma.* 2004; 20(9):1466–1467.
4. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res.* 2007; 6(9):3549–3557. [PubMed: 17676885]
5. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res.* 2008; 7(1):29–34. [PubMed: 18067246]

6. Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, Chen X, Markey SP, Kowalak JA. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res.* 2004; 3(5):1002–1008. [PubMed: 15473689]
7. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods.* 2007; 4(3):207–214. [PubMed: 17327847]
8. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J Proteome Res.* 2008; 7(1):286–292. [PubMed: 18078310]
9. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res.* 2009; 8(8):3872–3881. [PubMed: 19522537]
10. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem.* 2004; 76(14):4193–4201. [PubMed: 15253663]
11. Zybilov B, Coleman MK, Florens L, Washburn MP. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem.* 2005; 77(19):6218–6224. [PubMed: 16194081]
12. Fu X, Gharib SA, Green PS, Aitken ML, Frazer DA, Park DR, Vaisar T, Heinecke JW. Spectral index for assessment of differential protein expression in shotgun proteomics. *J Proteome Res.* 2008; 7(3):845–854. [PubMed: 18198819]
13. Fermin D, Basrur V, Yocum AK, Nesvizhskii AI. Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. *Proteome.* 2011; 11(7):1340–1345.
14. Jin S, Daly DS, Springer DL, Miller JH. The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. *J Proteome Res.* 2008; 7(1):164–169. [PubMed: 18001079]
15. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem.* 2003; 75(17):4646–4658. [PubMed: 14632076]
16. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol Cell Proteomics.* 2007; 6(12):2212–2229. [PubMed: 17939991]
17. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res.* 2010; 9(2):761–776. [PubMed: 19921851]
18. Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. *Mol Cell Proteomics.* 2007; 6(9):1589–1598. [PubMed: 17533221]
19. Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteome.* 2009; 9(5):1220–1229.
20. Searle BC, Turner M, Nesvizhskii AI. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res.* 2008; 7(1):245–253. [PubMed: 18173222]
21. Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. *J Proteome Res.* 2011; 10(7):2949–2958. [PubMed: 21488652]
22. Li M, Gray W, Zhang H, Chung CH, Billheimer D, Yarbrough WG, Liebler DC, Shyr Y, Slebos RJ. Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res.* 2010; 9(8):4295–4305. [PubMed: 20586475]
23. Halvey PJ, Zhang B, Coffey RJ, Liebler DC, Slebos RJ. Proteomic consequences of a single gene mutation in a colorectal cancer model. *J Proteome Res.* 2012; 11(2):1184–1195. [PubMed: 22103262]

24. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinforma*. 2008; 24(21):2534–2536.
25. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. *J Proteome Res*. 2010; 9(4):1716–1726. [PubMed: 20131910]
26. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*. 2007; 6(2):654–661. [PubMed: 17269722]
27. Ma ZQ, Tabb DL, Burden J, Chambers MC, Cox MB, Cantrell MJ, Ham AJ, Litton MD, Oretto MR, Schultz WC, Sobecki SM, Tsui TY, Wernke GR, Liebler DC. Supporting tool suite for production proteomics. *Bioinforma*. 2011; 27(22):3214–3215.
28. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001; 125(1–2):279–284. [PubMed: 11682119]
29. Pyne S, Futcher B, Skiena S. Meta-analysis based on control of false discovery rate: combining yeast ChIP-chip datasets. *Bioinforma*. 2006; 22(20):2516–2522.
30. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol*. 2005; 18(5):1368–1373. [PubMed: 16135132]
31. Stouffer, SA.; Suchman, EA.; DeVinney, LC.; Star, SA.; Williams, RM, Jr. *The American soldier: adjustment during army life*. Princeton University Press; Princeton: 1949.
32. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*. 2004; 101(25):9309–9314. [PubMed: 15184677]
33. Edwards N, Wu X, Tseng C-W. An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clin Proteomics*. 2009; 5(1):23–36.
34. Sanchez-Tillo E, Lazaro A, Torrent R, Cuatrecasas M, Vaquero EC, Castells A, Engel P, Postigo A. ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene*. 2010; 29(24):3490–3500. [PubMed: 20418909]
35. Wan D, Gong Y, Qin W, Zhang P, Li J, Wei L, Zhou X, Li H, Qiu X, Zhong F, He L, Yu J, Yao G, Jiang H, Qian L, Yu Y, Shu H, Chen X, Xu H, Guo M, Pan Z, Chen Y, Ge C, Yang S, Gu J. Large-scale cDNA transfection screening for genes related to cancer development and progression. *Proc Natl Acad Sci U S A*. 2004; 101(44):15724–15729. [PubMed: 15498874]
36. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002; 74(20):5383–5392. [PubMed: 12403597]

Fig. 1.

Protein assembly, protein groups, and peptide groups in IDPicker. In this diagram, three peptide groups are associated with two protein groups. IDPicker groups peptides, such as the two peptides in the *orange box*, to “peptide groups” when they match to exactly the same proteins—histone-binding protein RBBP7 and RBBP4. “Protein groups” are sets of proteins such as RBBP4 and isoform 3 of RBBP4 that are indiscernible on the basis of the observed peptides. Peptide groups that only associate with one protein group are called unique peptide groups (*green box*). Peptide groups that associate with more than one protein group are called shared peptide groups (*orange box*). Both the protein and peptide groups are shown in IDPicker reports

Fig. 2.

Number of true-positive proteins out of the top 250 of proteins of corresponding combination of search engines in ABRF *E. coli* dataset. In the “count sum model,” results from different search engines are combined by adding spectral counts together; in the “Stouffer *p*-combo model”, *p* values from different search engines are combined by Stouffer's method. In the “*p*-rank sum model,” proteins were ranked by *p* value from individual search engines (from smallest to largest). The ranks are added together to compute a super rank for each protein. In the “vote counting model,” proteins were ranked by the number of search engines deeming them significant along with the best individual search engine *p* value

Fig. 3. Number of true-positive proteins in top 50 differentiated proteins using different combinations of search engines in CPTAC Study6 dataset

Fig. 4. Comparison of fully tryptic and semi-tryptic searching in ABRF dataset. **A** ROC curve of differentiated protein expression using semi-tryptic and fully tryptic searches in the iPRG dataset. **B** True positives and false positives at p value of 0.05 discovered through semi-tryptic and fully tryptic searches

Table 1

Fully tryptic versus semi-tryptic search in a yeast sample with spiked human proteins

	Fully tryptic			Semi-tryptic		
	True positive/false positive	Average spectral count ratio	Area under curve	True positive/false positive	Average spectral count ratio	Area under curve
27-fold difference						
B versus E	30/20	10.74	0.9786	32/18	12.36	0.9827
A versus D	25/25	9.67	0.9939	27/23	12.62	0.9944
9-fold difference						
C versus E	29/21	4.84	0.9777	31/19	4.33	0.9779
B versus D	22/28	2.45	0.9799	25/25	6.76	0.9765
3-fold difference						
D versus E	21/29	1.95	0.8676	24/26	2.12	0.9150
C versus D	21/29	2.53	0.8290	24/26	2.48	0.8841
B versus C	9/41	3.25	0.8859	14/36	3.21	0.8942

The amount of UPS-1 (Sigma-Aldrich) proteins that spiked in A, B, C, D, and E are 0.24, 0.67, 2.7, 6.7, and 20 fmol/μg yeast, respectively. Geometric average is calculated by geometric mean of the ratio: group 1 and group 2 indicate the comparison pairs. For a spike in protein that is not found in the search results, their spectral counts are set as zero