BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 8 2011, pages 1128–1134
doi:10.1093/bioinformatics/btr089

Systems biology

Advance Access publication February 23, 2011

Assigning spectrum-specific *P*-values to protein identifications by mass spectrometry

Victor Spirin[1], Alexander Shpunt[1,2], Jan Seebacher[3], Marc Gentzel[4], Andrej Shevchenko[4], Steven Gygi[3] and Shamil Sunyaev[1,*]

[1]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, [2]Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02143, [3]Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA and [4]MPI of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Although many methods and statistical approaches have been developed for protein identification by mass spectrometry, the problem of accurate assessment of statistical significance of protein identifications remains an open question. The main issues are as follows: (i) statistical significance of inferring peptide from experimental mass spectra must be platform independent and spectrum specific and (ii) individual spectrum matches at the peptide level must be combined into a single statistical measure at the protein level.

Results: We present a method and software to assign statistical significance to protein identifications from search engines for mass spectrometric data. The approach is based on asymptotic theory of order statistics. The parameters of the asymptotic distributions of identification scores are estimated for each spectrum individually. The method relies on new unbiased estimators for parameters of extreme value distribution. The estimated parameters are used to assign a spectrum-specific *P*-value to each peptide-spectrum match. The protein-level confidence measure combines *P*-values of peptide-to-spectrum matches.

Conclusion: We extensively tested the method using triplicate mouse and yeast high-throughput proteomic experiments. The proposed statistical approach improves the sensitivity of protein identifications without compromising specificity. While the method was primarily designed to work with Mascot, it is platform-independent and is applicable to any search engine which outputs a single score for a peptide-spectrum match. We demonstrate this by testing the method in conjunction with X!Tandem.

Availability: The software is available for download at ftp://genetics.bwh.harvard.edu/SSPV/.

Contact: ssunyaev@rics.bwh.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 2, 2010; revised on January 28, 2011; accepted on February 15, 2011

*To whom correspondence should be addressed.

1 INTRODUCTION

Improvements in specificity, sensitivity and throughput of protein identifications progress at unprecedented pace. However, the scope of available mass spectrometric and informatics technologies and divergence of their physicochemical and algorithmic bases contributes to growing concern regarding the reproducibility, transparency and statistical reliability of the proteome analysis.

In bottom-up proteomics, the identification proceeds via matching peptide MS/MS spectra to a database of sequences. The identity of proteins present in the sample is then inferred from individual peptide matches. While many peptide identifications are seemingly undisputed within any statistical framework, a large fraction of hits fall into 'borderline' significance range. The ability to accurately assess the statistical significance of these identifications would increase specificity and sensitivity of mass spectrometry analyses. More importantly, at the protein level, a single estimate of statistical significance should combine evidence from all tandem mass spectra matched to the same protein. Arbitrary conventions (Carr *et al.*, 2004) that are employed for maintaining protein identification consistency will be rendered unnecessary by such estimate.

At the peptide level, current identification methods can be classified into false discovery rate (FDR) approaches, *P*-value-based approaches, and posterior probability-based approaches. FDRs are typically estimated by searches against a decoy database (e.g. a database of reversed and/or randomized protein sequences) (Elias and Gygi, 2007; Kall *et al.*, 2008; Park *et al.*, 2008). A threshold determining statistical significance is selected specifically for a given dataset and a given peptide-to-spectrum scoring metric. More sophisticated statistical approaches involve simulations and analytical approximations (Ramos-Fernandez *et al.*, 2008).

However, a single FDR threshold is applied to all spectra from the dataset. Therefore, it is not informative about statistical significance of individual matches and the estimate of the FDR threshold usually has a large variance.

Specificity and sensitivity of identifications can be increased by stratifying peptide-spectrum matches (PSMs) using properties of spectra (Anderson *et al.*, 2003; Kall *et al.*, 2007) or peptide length (Cox and Mann, 2008) in a statistical or machine learning framework.

A number of recently developed methods report *P*-values specific to individual spectra. These methods are primarily based

on fitting the distribution of scores with a parametric model. Klammer *et al*. (2009) fit the distribution of SEQUEST Xcorr scores with Weibull distribution. Searle *et al*. (2008) fit the distribution of Mascot scores with a double-exponential Gumbel distribution. Fast SEQUEST (Eng *et al*., 2008) employs a faster algorithm which applies the Xcorr to every candidate peptide from the sequence database being queried. Spectrum-specific *E*-values are computed by approximating the distribution of Xcorr scores with an exponential distribution. OMSSA (Geer *et al*., 2004) approximates the scores of every peptide-spectrum match with Poisson distribution. Kim *et al*. (2009) address the issue of spectrum specificity by calculating a generating function and infer the probability of a correct spectrum identification based on all matching peptides. RAId_DbS (Alves *et al*., 2007) uses a score in the form of a weighted sum of logarithmic intensities and applies an extension of the Central Limit Theorem to assign statistical significance to the matches.

However, the approach based on fitting specific parametric models cannot be generalized to other platforms. Also, the parametric model with parameters estimated using the overall distribution of scores may not be accurate at the extreme tail.

Further, a single protein level rather than peptide level *P*-value is needed for practical applications.

The issue of statistical significance of protein identifications is addressed in ProteinProphet (Nesvizhskii *et al*., 2003) (posterior error probabilities), ComByne (Bern and Goldberg, 2008), Qscore (Moore *et al*., 2002), PROVALT (Weatherly *et al*., 2005), MAYU (Reiter *et al*., 2009) (FDRs), PANORAMICS (Feng *et al*., 2007) (combining peptide probabilities derived from Mascot score into protein probabilities) and X!Tandem (Craig and Beavis, 2004) (*E*-values). However, some of these methods are not based on spectrum-specific identifications, while others are dependent on particular assumptions regarding the platform or distribution of false-positive matches.

ProteinProphet, PROVALT, ComByne, MAYU and Qscore provide dataset-specific, rather than spectrum-specific, identifications. PROVALT uses heuristic approach to assign FDR for protein identifications from Mascot peptide scores. MAYU utilizes a set PSM above a user-specified PSM FDR threshold. The number of false-positive protein identifications and protein FDR is then estimated under the assumption of uniform distribution of false-positive PSMs over the target database. ProteinProphet builds on PeptideProphet (Keller *et al*., 2002). The latter combines SEQUEST Xcorr score adjusted for peptide length, the difference between the top score and the rest, logarithm of rank of Sp score and mass accuracy, and then uses an empirical formula to obtain Bayesian posterior probability for peptide identifications. These peptide identifications are then combined into protein probabilities.

In this article, we developed and extensively tested a method that
(i) relies on the information content of individual spectra and is fully independent of dataset properties, spectrum-to-sequence correlation model, instrumentation platform, data acquisition and processing settings; and
(ii) combines evidence from peptide identifications to assess statistical significance of protein identifications.

We report a single *P*-value for protein identifications combining evidence from individual peptides. Statistical significance of individual peptide identifications is based on asymptotic distribution of order statistics and is specific for each tandem mass spectrum. This increases sensitivity and specificity of both peptide and protein identifications.

We demonstrate how the software implementation of the method works in conjunction with the popular database search algorithm Mascot. However, our method is applicable to any database search that assigns a single numeric score for each spectrum match. We demonstrate this with an application of the same approach to peptide and protein identification using X!Tandem platform. We describe our statistical approach and tests of the method on real proteomic samples. We start with the statistics of peptide identifications and then proceed to the protein level.

2 METHODS

2.1 Datasets

Samples from two organisms—mouse and yeast—were used for obtaining datasets of peptide mass spectra. The mouse datasets are a triplicate LC-MS/MS analysis from the in-gel digest of the 49–64 kDa molecular weight region of mouse brain lysate. The yeast datasets are a triplicate LC-MS/MS analysis of an in-gel trypsin digest of the 70–150 kDa molecular weight region of a *Saccharomyces cerevisiae* whole-cell lysate, as previously described by Bakalarski *et al*. (2007). LC-MS/MS analyses for both the yeast and the mouse samples were performed on a Thermo LTQ-FT mass spectrometer in a data-dependent mode, with the FT10 method, as previously described by Bakalarski *et al*.

2.2 Statistical background

Peptide identification in mass spectrometry, similarly to any other database search, relies on the highest scoring hit. This simplifies the statistical analysis because the shape of the distribution of highest scores does not depend on the distribution of all scores, which is generally unknown.

According to extreme value theory, the maximum of a sample of independent identically distributed random variables after proper renormalization converges to one of three possible distributions, depending on the general properties of the distribution within the sample. In the case when the distribution within the sample is unbounded and has a light tail, which is the case for the Mascot search engine, the maxima of the sample converge to a double-exponential distribution (Gumbel, 2004). Thus, the probability P that the database search with spectrum i will result in a false identification with score s is given by:

$$P = 1 - F(s) = 1 - \exp(-\alpha_i \exp(-(s-\mu_i)/\beta_i)) \quad (1)$$

where $F(s)$ is the cumulative extreme value distribution function. For convenience, here we isolated dependence on the relative database size in the parameter α_i. This relative database size is spectrum dependent, since it represents not only the ratio of the database sizes but also the ratio of the number of peptides with a given precursor mass. Parameters μ_i and β_i are specific to the spectrum and depend only weakly on the relative databases size in the asymptotic regime. We note that we make no assumption about parametric form of the distribution of all scores. The distribution parameters only refer to the extreme value asymptotic form general for a large class of distributions.

Therefore, to compute a *P*-value of a peptide identification for an individual spectrum, we have to estimate two parameters of the distribution μ_i and β_i corresponding to this spectrum in any decoy database, and adjust for the target database size using the scaling factor α_i. An extremely important consideration is that α_i scales linearly with the effective size of the database. The effective database size for a peptide with a given precursor mass is just the total number of peptides with the same (within tolerance) precursor mass in this database. This means that the decoy database used to estimate the distribution parameters can be of any size, just large enough that asymptotic extreme value distribution would hold. The parameter α_i for any given

spectrum i can then be estimated as a ratio of the number of peptides with the corresponding precursor mass in the target and decoy databases. Second, extreme value distribution theorem is not limited to the maximal value of the Mascot score, but can be extended to any number k of top hits that the Mascot search returns for each spectrum. The joint probability density function of the highest k scores in the database for spectrum i [Equation (2)] depends on the same two parameters μ_i and β_i (Arnold *et al.*, 1992):

$$f(s_1,\ldots,s_k)=\exp(-\alpha_i\exp(-(s_k-\mu_i)/\beta_i)) \times\prod_{j=1}^{k}\frac{alpha_i}{\beta_i}(\exp(-(s_j-\mu_i)/\beta_i) \quad (2)$$

Equation (2) provides a way to estimate parameters μ_i and β_i for each spectrum from a search in a decoy database from a set of highest-scoring database hits $s_1,\ldots s_k$. We obtained maximum likelihood estimators analytically (see Supplementary Material):

$$\hat{\beta}_{\mathrm{iML}} = \langle s\rangle - s_k \quad (3)$$

$$\hat{\mu}_{\mathrm{iML}} = s_k+\hat{\beta}_{\mathrm{iML}}\ln k \quad (4)$$

Estimation by maximum likelihood does not guarantee that the estimators are unbiased. However, the bias of the estimators given by Equation (4) can be computed analytically, and unbiased estimators can be constructed as follows.

We will first consider the standardized Gumbel distribution,

$$F(s)=\exp(-\exp(-s)); \quad f(s)=\exp(-\exp(-s)-s),$$

corresponding to $\mu=0$ and $\beta=1$. We will then write the explicit results for general μ and β.

Starting with PDF of the i-th order statistic (Arnold *et al.*, 1992),

$$f_{(i)}(s_i)=\frac{\exp(-is_i)}{(i-1)!}\exp(-\exp(-s_i)),$$

one can obtain moments by integration:

$$\mathrm{E}\{s_i\} = \frac{1}{(i-1)!}\int_{-\infty}^{\infty}x\exp(-\exp(-x)-ix)dx = \gamma-\sum_{j=1}^{i-1}\frac{1}{j},$$

where $\gamma=0.577216...$ is the Euler–Mascheroni constant.

$$\mathrm{E}\{s_i^2\} = \frac{1}{(i-1)!}\int_{-\infty}^{\infty}x^2\exp(-\exp(-x)-ix)dx = \frac{\pi^2}{6}+\left(\gamma-\sum_{j=1}^{i-1}\frac{1}{j}\right)^2-\sum_{j=1}^{i-1}\frac{1}{j^2}$$

Consequently,

$$\mathrm{Var}\{s_i\}=\frac{\pi^2}{6}-\sum_{j=1}^{i-1}\frac{1}{j^2}.$$

Joint pdf of $s_1,s_2,\ldots,s_k, s_1\geq s_2\geq\ldots\geq s_k$ is

$$f(s_1,s_2,\ldots,s_k)=\exp\left(-\exp(-s_k)-\sum_{i=1}^{k}s_i\right)$$

Properly marginalizing the above distribution, we get, for $i<j$, the joint PDF of i-th and j-th statistic:

$$f_{(i),(j)}(s_i,s_j)=\frac{\exp(-\exp(-s_j))}{(i-1)!} \times\sum_{m=0}^{j-i-1}\frac{(-1)^{j-i-1-m}}{m!(j-i-1-m)!}\exp(-s_j(m+1))\exp(-s_i(j-1-m))$$

Using the following two formulas:

$$\int_{s_j}^{\infty}s_i\exp(-k_is_i)ds_i=\frac{(1+s_j)\exp(-k_is_j)}{k_i^2},$$

$$\int_{-\infty}^{\infty}s_j\exp(-\exp(-s_j)-k_js_j)ds_j\int_{s_j}^{\infty}s_i\exp(-k_is_i)ds_i = \frac{(k_i+k_j-1)!}{k_i^2}\times \left(\gamma-\sum_{n=1}^{k_i+k_j-1}\frac{1}{n}+k_i\left[\left(\gamma-\sum_{n=1}^{k_i+k_j-1}\frac{1}{n}\right)^2+\frac{\pi^2}{6}-\sum_{n=1}^{k_i+k_j-1}\frac{1}{n^2}\right]\right)$$

one obtains

$$\mathrm{E}\{s_i\cdot s_j\}=\left(\gamma-\sum_{n=1}^{j-1}\frac{1}{n}\right)^2 +\frac{\pi^2}{6}-\sum_{n=1}^{j-1}\frac{1}{n^2}+\left(-\sum_{n=1}^{i-1}\frac{1}{n}+\sum_{n=1}^{j-1}\frac{1}{n}\right)\left(\gamma-\sum_{n=1}^{j-1}\frac{1}{n}\right)$$

Therefore,

$$\mathrm{Cov}\{s_i,s_j\}=\frac{\pi^2}{6}-\sum_{n=1}^{j-1}\frac{1}{n^2}.$$

We see that $\mathrm{Cov}\{s_i,s_j\}=\mathrm{Var}\{s_j\}$.

With the basic formulas at hand, we are now in a position to compute the moments of our estimates. Skipping elementary algebra, we get:

$$\mathrm{E}\{\hat{\beta}_{\mathrm{ML}}\}=\frac{k-1}{k}, \quad \mathrm{Var}\{\hat{\beta}_{\mathrm{ML}}\}=\frac{k-1}{k^2};$$

$$\mathrm{E}\{\hat{\mu}_{\mathrm{ML}}\}=\gamma-\sum_{n=1}^{k-1}\frac{1}{n}+\frac{k-1}{k}\ln k, \quad \mathrm{Var}\{\hat{\mu}_{\mathrm{ML}}\}=\sum_{n=1}^{k-1}\frac{1}{n^2}.$$

With the explicit expressions for bias, we can form unbiased estimates of the parameters β and μ:

$$\hat{\beta}_{iub} = \frac{1}{k-1}\sum_{i=1}^{k}s_i-\frac{k}{k-1}s_k \quad (5)$$

$$\hat{\mu}_{iub} = s_k+\hat{\beta}_{\mathrm{iML}}\ln k-\hat{\beta}_{iub}\left(\gamma-\sum_{n=1}^{k-1}\frac{1}{n}+\frac{k-1}{k}\ln k\right) \quad (6)$$

2.3 The algorithm

This suggests the following strategy for estimating spectrum-specific P-values. For every spectrum obtained in a proteomics experiment we run a search, i.e. Mascot or X!Tandem, against a decoy sequence database. The computational efficiency can be greatly increased with only a limited sacrifice of accuracy if these searches would be run against a small size decoy database using a very large precursor mass tolerance. We estimate parameters μ_i and β_i for each spectrum using a number of top hits reported by Mascot. Parameter α_i is then computed to adjust to the effective real database size and mass tolerance (see Supplementary Materials). This approach assumes weak dependence of μ_i and β_i on database size in the asymptotic regime as discussed above. Mascot search against the real protein sequence database is then run and P-values for all individual peptide identifications are computed according to Equation (1).

In practice, the currently available version of Mascot reports only 10 hits for each spectrum, which is insufficient for reliable estimation of the distribution parameters. To circumvent this problem, we used several independent searches against small decoy databases and obtained maximum

likelihood estimators numerically and unbiased estimators using analytically derived correction (Supplementary Materials).

Each decoy database contains 10 000 Bernoulli sequences. The protein length distribution for these sequences is approximated by UniProtKB/Swiss-Prot mouse proteome. The frequencies of all amino acids are simulated to be the same as observed amino acid frequencies in the mouse proteome.

One can, just as well, use 100 databases and estimate parameters for each spectrum using only one top hit in each database. Our simulations indicate that the accuracy of parameter estimation is the same for both methods. But using 10 top hits in each database and reducing the required number of decoy databases to 10 significantly speeds up parameter estimation.

On the basis of P-values corresponding to individual peptide identifications, we compute P-values for protein identifications.

As a standard practice among Mascot users, for each spectrum we consider only the best peptide hit, and discard all other peptides with lower matching score.

P-values of individual peptides matching the same protein are combined into cumulative protein P-value using Stouffer's method (Whitlock, 2005). For the peptide match with k-th lowest P-value among all peptides from the same protein, we compute the probability that the k-th P-value would be as low or even lower by pure chance, using Z-test. This test converts the P-values, p_i, from each of the k individual independent tests into standard normal deviates Z_i. The sum of these deviates, divided by the square root of the number of tests, k,

$$Z_s = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{7}$$

can be shown to follow the standard normal distribution and provide a test for the common null hypothesis. The minimal value of this probability over all k peptides (corrected for multiple testing using permutations) is the P-value for the protein identification. We further convert these protein level P-values to E-values corresponding to the size of the sequence database. Alternatively, one could combine P-values of all peptides matching a protein. However, a common situation arises when, for example, two peptides are true hits, while the rest are noise. If one combines P-values of all such hits, the contribution from noise may render the overall P-value insignificant. Choosing the minimal P-value corrected for multiple testing solves this problem.

An important issue is clustering proteins identified by the same or overlapping groups of peptides. Redundant protein database entries, as well as homologous proteins and splicing variants may be identified by the same set of PSM from the experimental sample. This is a complicating issue, since a peptide belonging to more than one distinct protein in the database makes determination of the true corresponding protein ambiguous. A comprehensive review of possible scenarios and approaches is given in Nesvizhskii and Aebersold (2005).

In our method, protein expectation values are computed based on unique evidence for the group to which this protein belongs. This is similar to the approach used in IDPicker (Zhang *et al.*, 2007) in that each of our clusters represents a minimal set of proteins that explains all the peptides in this cluster. However, we do not eliminate proteins that explain a subset of peptides in a peptide cluster. We report such proteins in a different cluster with its corresponding P-value.

2.4 Software

The software is written in Perl and runs on a Linux computer. A standalone version is available at ftp://genetics.bwh.harvard.edu/SSPV/ The software input is: an mgf file and fasta file with a protein sequence database. Separate scripts make a combined forward/reverse sequence database and 10 random Bernoulli databases. The combined forward/reverse database is created for FDR plots. To increase the precision of the FDR plots, we use a reversed target database rather than a Bernoulli database to make it as similar to the target database as possible. After this, a shell script runs the 10 Mascot searches in these Bernoulli databases, then the Mascot search in the combined database, then starts the perl script which determines the parameters for each spectrum, and then processes the combined database search output and puts the results into a user-specified directory. The output contains text files with the list of identified peptides and proteins and their P-values. Processing a typical input containing 5000–10 000 spectra against a typical-size sequence database, which includes search in 10 random databases and parameter estimation for every spectrum, usually takes <5 min.

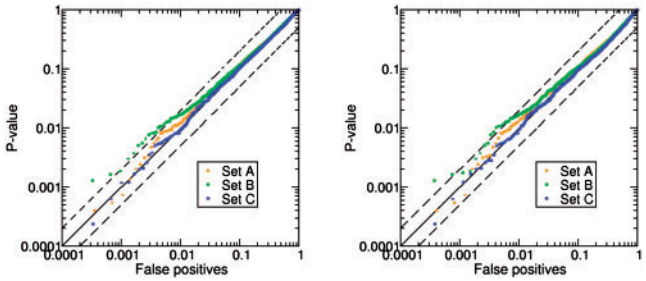

Fig. 1. Fraction of false positive peptide (left) and protein (right) identifications as a function of specified P-value threshold. Three sets of spectra obtained in triplicate LC-MS/MS data from the in-gel digest of the 49–64 kDa MW region of mouse brain lysate. Calculated P-values are accurate to within a factor of 2 (dashed lines).

3 RESULTS

We tested the method using two proteomic datasets obtained from in-gel trypsin digestion of (i) yeast and (ii) mouse brain proteins, analyzed with a Thermo LTQ-FT mass spectrometer (see Section 2). The datasets represent highly complex peptide mixtures from the two species. To confirm that P-values accurately estimate fraction of false positive identifications, we run searches in two sets of decoy databases. P-values corresponding to individual peptide or protein identifications determined in one decoy database set are plotted against the corresponding fractions of identifications in the other (qq-plot). By the definition of P-value, the fraction of false-positive identifications (estimated by the fraction of identifications in the decoy databases) at a given P-value should ideally be equal to this P-value. The qq-plots are shown in Figure 1 and in Supplementary Figures S1 and S2. As seen from the qq-plots, P-values provide practically acceptable approximation of the rate of false-positive identifications, i.e. predict fraction of decoy database hits below a specified P-value threshold. For example, the fraction of false-positive identifications at P-value threshold 0.01 in the three mouse datasets is in the range of 0.0064–0.0119, and in the yeast datasets of 0.0120–0.0126.

Next, P-values generally discriminate between searches in the forward (real) and reverse (decoy) databases (Fig. 2) better than peptide ion scores assigned by Mascot. At the 1% FDR threshold as determined by the search against the decoy database, selecting database hits by P-value rather than by score increases the number of identifications in the forward (real) sequence database. In addition, we tested the performance of spectrum-specific P-values using one other measure recommended by Matrix Science (www.matrixscience.com/pdf/2005WKSHP4.pdf). Figure 2B, shows the ROC curve when Mascot Ion score minus Identity score is used as a discriminant function, similar to Searle *et al.* (2008). We find that ion minus identity score performs approximately equally well to ion score in discriminating true

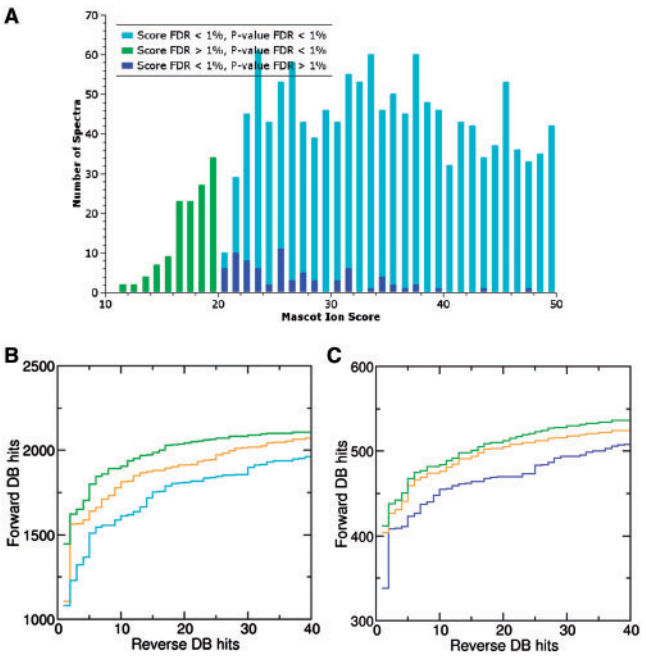

Fig. 2. (**A**) number of identified spectra with a given Mascot score for a specified FDR cutoff, as determined by identifications in forward and reverse protein databases. While high-scoring peptide matches are equally well identified by raw score and *P*-values (cyan bars), *P*-value approach has a significant advantage over Mascot-score approach for borderline-scoring peptides (green bars). Details are provided in Supplementary Tables S2–S5. (**B**) identified peptide hits. Three sets of spectra obtained in triplicate LC-MS/MS data from the in-gel digest of the 49–64 kDa MW region of mouse brain lysate. Set A is shown. Sets B and C are provided in Supplementary Figure S3. Orange curves: identifications by raw Mascot score; cyan: identifications by ion minus identity score; green: identifications corresponding to spectra-specific *P*-values. (**C**) Same set of spectra, identified protein hits. Blue: identifications by raw Mascot score. Forward and reverse proteins are identified by MASCOT score of the top peptide-spectrum match. Orange: forward versus reverse protein identifications by best *P*-value of PSM. Green: identifications by a single protein *P*-value that combines all PSMs from the same protein.

Table 1. Number of PSM for each of the mouse datasets

Mouse datasets, regardless of *P*-value				
		Forward		Reverse
	Mascot score	All hits	Confident hits	
Set 1	$10 < S < 20$	1229	426	666
	$20 < S < 30$	1845	1688	11
Set 2	$10 < S < 20$	1312	459	762
	$20 < S < 30$	1996	1813	9
Set 3	$10 < S < 20$	1298	486	697
	$20 < S < 30$	2113	1919	16
Mouse datasets, $P < 0.01$				
Set 1	$10 < S < 20$	170	154	5
	$20 < S < 30$	496	471	2
Set 2	$10 < S < 20$	151	140	3
	$20 < S < 30$	538	507	1
Set 3	$10 < S < 20$	130	122	6
	$20 < S < 30$	520	494	1

The search is performed in both forward and reverse sequence databases. The hits are binned by Mascot score into two ranges: $10 < S < 20$ and $20 < S < 30$. These ranges correspond to the borderline significance of peptide identification. All matches are included, regardless of *P*-value (Upper part). Only spectra with conservative $P < 0.01$ are included (Lower part). Confident hits: peptides supported by at least two other peptide identifications from the same protein.

Table 2. Discrimination of spectra by spectrum-specific *P*-value versus by raw Mascot ion score for the three mouse datasets

	$FDR_p < 1\%$ $FDR_I > 1\%$		$FDR_p < 1\%$ $FDR_I > 1\%$	
	Forward	Reverse	Forward	Reverse
Set 1	161	7	57	7
Set 2	142	7	67	7
Set 3	293	9	52	9

Number of forward database hits versus reverse database hits at 1% FDR level. Peptides with only *P*-values passing the FDR cutoff 1% versus peptides with only Mascot score passing the FDR cutoff 1%. FDR_p: *P*-value FDR, FDR_I: ion-score FDR.

positive versus false positive hits. Spectrum-specific *P*-values outperform either combination of ion and identity scores.

To test whether spectrum-specific *P*-values indeed allow peptide identifications with borderline ion scores, we analyzed the distribution of reverse database hits with respect to both ion score and *P*-values. As seen from Table 1, upper part, the fraction of reverse database hits for scores in the range 10–20 is approaching the fraction of forward database hits, indicating that in this score range any forward hit is as likely to be true as false. The fraction of reverse database hits for scores in the range 20–30 compared with forward identifications is much smaller. However, it is still of the order of 1%. Applying strict *P*-value threshold eliminates almost all reverse database hits (Table 1, lower part and Supplementary Tables S3 and S4) while retaining significant number of forward database hits. Most of these forward hits were supported by at least two confident peptide identifications from the same protein, confirming that these forward hits are likely correct.

Table 2 and Supplementary Table S5 is another demonstration that identification by spectrum-specific *P*-values significantly outperforms identification by Mascot ion score. We count the fraction of reverse database hits at 1% FDR level obtained using Mascot ion score but not by the 1% FDR threshold if sorted by *P*-value. Conversely, we count the fraction of reverse database hits at the 1% FDR level obtained by sorting database hits by *P*-value but not by Mascot ion score. Table 2 shows that the fraction of peptides passing only FDR cutoff is significantly higher than the fraction of peptides passing only Mascot score cutoff.

In all datasets, results of the search according to *P*-values leads to higher number of true positive hits (approximated by hits in the database of real protein sequences) per false-positive hit (approximated by hits in the database of reversed protein sequences). Selecting peptide identifications by *P*-value rather than score leads to increased sensitivity at any specificity threshold (Fig. 2B, Supplementary Figs S3 and S4).

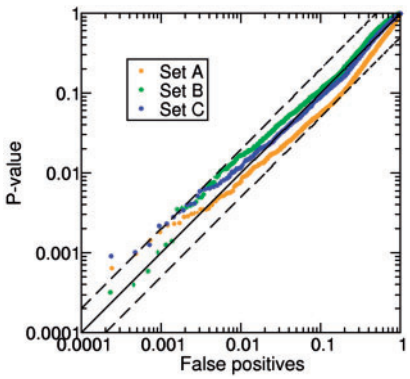

Fig. 3. Peptide qq-plots for the three mouse datasets; identifications with X!Tandem.

In addition to the performance comparison with raw Mascot scores, we tested the FDRs obtained with our method against those obtained using PeptideProphet. PeptideProphet source code TPP v4.4 was downloaded from http://proteinprophet.sourceforge.net/. For each of the three mouse and three yeast datasets, we ran xinteract to identify PSMs in combined target plus reversed database. Supplementary Figure S9 shows the comparison of FDRs for peptides identified with PeptideProphet and using spectrum-specific *P*-values. Spectrum-specific *P*-values outperform PeptideProphet by 5–10%, depending on the dataset.

At the protein level, the use of *P*-values combining evidence from multiple peptides increases sensitivity over a conventional method of protein identification based on the score of the best scoring peptide from the protein (Fig. 2C, Supplementary Figs S3–S4).

Importantly, the major increase in the number of true-positive identifications per false-positive identification (increase in sensitivity at a given specificity level) is observed at strict thresholds corresponding to low rate of false-positive identifications. This means that spectrum-specific *P*-values confidently validate medium-scoring true positives while efficiently suppressing high scoring false positives.

To exclude the possibility that these additional identifications in the forward database are false identifications, we repeated the analysis limiting it to confidently identified proteins, i.e. proteins with at least three matching peptides (Supplementary Fig. S5). Spectrum-specific *P*-values increase the number of these confidently true positive identifications at any FDR cutoff.

To demonstrate generality and platform independence of our approach, we applied spectrum-specific *P*-values to identifications made by X!Tandem platform (Supplementary Materials). We computed peptide *P*-values using parameters estimated from top hits according to X! Tandem hyperscore in searches against 50 decoy databases. These *P*-values are shown to provide accurate rate of false-positive identifications, as seen from qq-plots (Fig. 3, Supplementary Fig. S6). Discrimination between searches in forward and reverse databases by *P*-values versus X!Tandem *E*-values is shown in Supplementary Figures S7 and S8.

4 DISCUSSION

P-values for peptide and protein identifications based on the analysis of order statistics can be generally applied in conjunction with any peptide-to-spectrum matching software, regardless of the matching algorithm.

The specificity of spectrum *P*-values renders unnecessary empirical conventions to infer significance of peptide identifications. Matrix Science, for example, provides two ion score thresholds (www.matrixscience.com/pdf/2005WKSHP4.pdf). Ion identity threshold is a value representing a fixed probability of encountering a false match. However, this value is conservative and may not be achieved for some true matches, due to poor signal or fragmentation gaps. Therefore, Mascot also provides a second, less conservative, homology threshold. Our approach, by providing a single *P*-value for every spectrum, solves the problem of empirical conventions, based on these two thresholds, automatically.

Existing methods to assign statistical significance to peptide-to-spectrum matches either use a single threshold for the dataset as a whole (Elias and Gygi, 2007; Kall *et al*., 2008; Park *et al*., 2008; Ramos-Fernandez *et al*., 2008) or take into account spectrum specificity by stratifying spectra or peptides (Anderson *et al*., 2003; Cox and Mann, 2008; Kall *et al*., 2007). Other methods assign spectrum-specific measures of statistical significance by fitting parametric models to the distribution of peptide-to-spectrum matches (Craig and Beavis, 2004; Klammer *et al*., 2009; Searle *et al*., 2008). FAST SEQUEST (Eng *et al*., 2008) and X!Tandem (Craig and Beavis, 2004) assume the exponential decay of the score distribution. These methods use regression-based approaches to estimate spectrum-specific *E*-values. The generality of the assumption of the exponential tail is justified using the arguments presented in this work (convergence to the Gumbel distribution). The tail of the double exponential distribution is approximately exponential. However, the regression-based estimator employed by FAST SEQUEST and X!Tandem has two disadvantages addressed by our approach. First, it uses hundreds of points for the estimation. This may lead to inaccuracies if convergence to the Gumbel distribution is slow. Second, simulations of the true Gumbel distribution (Supplementary Fig. S10) show that the regression-based estimator is biased and inefficient resulting in *E*-values that are volatile and, on average, inflated.

Reliance on the asymptotic distribution of high order statistics allows avoiding any specific parametric models of the distribution of peptide-to-spectrum matches. Distribution of high order statistics, given the convergence to the asymptotic behavior, depends on the overall distribution of peptide-to-spectrum matches only through parameter values. Unbiased estimates of these parameters help approximating the far tail of the distribution and construct platform-independent spectrum-specific *P*-values.

5 CONCLUSION

We propose a method and software to assign statistical significance to peptide and protein identification by mass spectrometry. A single probabilistic confidence measure (*P*-value) is assigned to each identification. This *P*-value is spectrum-specific and independent of instrumentation or software scoring method. The platform independence is achieved because the method makes very few assumptions about the distribution of the PSM scores. It is based on the asymptotic extreme value distribution. The parameters of the extreme value distribution are estimated using newly obtained analytical unbiased estimates.

Therefore, our approach does not require introducing any empirical cutoffs or data stratifications. It also avoids fitting any parametric distribution to the discriminant function provided by a search engine. Such parametric models can be inaccurate at the important far tail of the distribution.

Protein level *P*-values are constructed by combining peptide level *P*-values using the Stouffer's method, so a single *P*-value is reported per protein identification. Spectrum specificity allows improving accuracy and FDR.

ACKNOWLEDGEMENT

We thank Dr Ivan Adzhubei for help with the software and Dr Ronald Beavis for help with installing X!Tandem.

Funding: National Institute of Health (Grants RL1 DE019022 and R01 GM070986).

Conflict of Interest: none declared.

REFERENCES

Anderson,D.C. *et al.* (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, **2**, 137–146.

Alves,G. *et al.* (2007) RAId_DbS: peptide identification using database searches with realistic statistics. *Biol. Direct*, **2**, 25.

Arnold,B.C. *et al*. (1992) *A First Course in Order Statistics*. Jihn Wiley, New York.

Bakalarski,C.E. *et al.* (2007) The effect of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics. *Anal. Bioanal. Chem.*, **389**, 1409–1419.

Bern,M. and Goldberg,D. (2008) Improved ranking functions for protein and modification-site identification. *J. Comput. Biol.*, **15**, 705–719.

Carr,S. *et al.* (2004) The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics*, **3**, 531–533.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechonol.*, **26**, 1367–1392.

Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Elias,G.E., and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Eng,J.K. *et al.* (2008) A Fast SEQUEST cross correlation algorithm. *J. Proteome Res.*, **7**, 4598–4602.

Feng,J. *et al.* (2007) Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Anal. Chem.*, **79**, 3901–3911.

Geer,L.Y. *et al.* (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.

Gumbel,E.J. (2004) *Statistics of Extremes*. Dover Publications, Inc.

Kall,L. *et al.* (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, **4**, 923–925.

Kall,L. *et al.* (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, **7**, 29–34.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Kim,S. *et al.* (2009) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, **7**, 3354–3363.

Klammer,A.A. *et al.* (2009) Statistical calibration of the SEQUEST Xcorr funtion. *J. Proteome Res.*, **8**, 2106–2113.

Moore,R.E. *et al.* (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass. Spectrom.*, **13**, 378–386.

Nesvizhskii,A.I. and Aebersold,R. (2005) Interpretation of shotgun proteomic data. The protein inference problem. *Mol. Cell. Proteomics*, **4**, 1419–1440.

Nesvizhskii,A.I. *et al.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4658.

Park,C.Y. *et al.* (2008) Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, **7**, 3022–3027.

Ramos-Fernandez,A. *et al.* (2008) Generalized method for probability-based peptide and protein identification from tandem mass spectrometry data and sequence database searching. *Mol. Cell. Proteomics*, **7**, 1748–1754.

Reiter,L. *et al.*(2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8**, 2405–2417.

Searle,B.C. *et al.* (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.*, **7**, 245–253.

Weatherly,D.B. *et al.* (2005) A euristic method for assigning a false-discovery rates for protein identifications from Mascot database search results. *Mol. Cell. Proteomics*, **4** , 762–772.

Whitlock,M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.

Zhang,B. *et al.* (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.*, **6**, 3549–3557.