Bioinformatics, 31(5), 2015, 699–706
doi: 10.1093/bioinformatics/btu717
Advance Access Publication Date: 31 October 2014
Original Paper

OXFORD

Structural bioinformatics

Mass spectrometry-based protein identification with accurate statistical significance assignment

Gelio Alves and Yi-Kuo Yu*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 24, 2014; revised on September 9, 2014; accepted on October 23, 2014

Abstract

Motivation: Assigning statistical significance accurately has become increasingly important as metadata of many types, often assembled in hierarchies, are constructed and combined for further biological analyses. Statistical inaccuracy of metadata at any level may propagate to downstream analyses, undermining the validity of scientific conclusions thus drawn. From the perspective of mass spectrometry-based proteomics, even though accurate statistics for peptide identification can now be achieved, accurate protein level statistics remain challenging.

Results: We have constructed a protein ID method that combines peptide evidences of a candidate protein based on a rigorous formula derived earlier; in this formula the database *P*-value of every peptide is weighted, prior to the final combination, according to the number of proteins it maps to. We have also shown that this protein ID method provides accurate protein level *E*-value, eliminating the need of using empirical post-processing methods for type-I error control. Using a known protein mixture, we find that this protein ID method, when combined with the Sorić formula, yields accurate values for the proportion of false discoveries. In terms of retrieval efficacy, the results from our method are comparable with other methods tested.

Availability and implementation: The source code, implemented in C++ on a linux system, is available for download at ftp://ftp.ncbi.nlm.nih.gov/pub/qmbp/qmbp_ms/RAId/RAId_Linux_64Bit.

Contact: yyu@ncbi.nlm.nih.gov

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

1 Introduction

Peptide identifications (ID) via mass spectrometry (MS) have become the central component in modern proteomics; this component, combined with additional analyses, routinely yields pragmatic **metadata**, including protein ID, protein quantification, protein structure and protein associations (Zhang *et al.*, 2013). These **metadata**, especially the associated statistical significance assignments, need to be as accurate as possible because they often form the building blocks for investigations at the systems biology level and influence the scientific conclusions drawn henceforth. In this article, we focus on protein ID, in particular on improving the accuracy of statistical significance assigned to proteins identified.

The need for robust developments towards accurate statistical significance assignments has been advocated (Huang *et al.*, 2012; Noble and MacCoss, 2012) despite the existence of many protein ID methods (Li and Radivojac, 2012; McHugh and Arthur, 2008; Serang and Noble, 2012). It has also been suggested (Spirin *et al.*, 2011) that the primary cause of unreliable significance assignment

for protein ID can be attributed to inaccurate significance assignment for peptide ID. Frequently used error-control/significance-assigning methods for peptide ID largely fall into two groups: proportion of false discovery (PFD), which is often incorrectly termed as false discovery rate (Benjamini and Hochberg, 1995), and spectrum-specific *P*-value/*E*-value (Alves *et al.*, 2007; Fenyo and Beavis, 2003; Park *et al.*, 2008). Methods belonging to the first group, controlling type-I error globally only, do not discriminate among identified peptides (Elias and Gygi, 2007). Methods belonging to the second group, capable of assigning per-spectrum per-peptide significance, can *properly* prioritize identified peptides when reported *P*-values/*E*-values are accurate; but the needed statistical accuracy is often unattainable due to improper heuristics or unjustifiable distribution assumptions (Alves *et al.*, 2007; Segal, 2008; Spirin *et al.*, 2011).

Given a tandem MS (MS/MS) spectrum and a quality score cutoff s_c, the *E*-value $E(s_c)$ should reflect the expected number of random peptides with scores the same as or better than s_c. (Similarly, the *P*-value $P(s_c)$ reflects the probability of finding a random peptide with quality score $s \geq s_c$.) In general, the *E*-value is obtained by multiplying the *P*-value by the total number of *qualified peptides* (whose masses fall in the range $[m_p - \delta, m_p + \delta]$ with m_p being the precursor ion's mass and δ the specified tolerance) in the database searched. Thus, besides providing the user with the numbers of false positives to anticipate, accurate *E*-value assignments enable ranking of candidate peptides across different spectra and experiments. In database searches in proteomics, the goal of accurate statistics can be approached in at least two ways. First, one may devise a scoring function whose resulting score distribution can be analytically characterized and thus used to infer the statistical significance (Alves *et al.*, 2007); if this is done correctly, the theoretical score distribution should fit well the bulk part of the normalized score histogram obtained from scoring all *qualified peptides* in the database of interest. Second, one may infer the spectrum-specific *P*-value via the normalized score histogram obtained from scoring all possible peptides (APP) (Alves and Yu, 2008); in this case, the database dependence appears only in the *E*-value, which is the *P*-value multiplied by the number of qualified peptides associated with the specified precursor ion mass and mass error tolerance. Either way yields database-specific *E*-values. Once a peptide *E*-value is obtained, one may transform it into the peptide database *P*-value (DPV) (Alves *et al.*, 2008b; Yu *et al.*, 2006), representing the likelihood of obtaining, in the database chosen, at least one peptide scoring equal to or better than the prescribed threshold. When combining *P*-values of peptides associated with a candidate protein, we use the peptides' DPVs.

Specifically, our proposed protein ID method combines peptide evidences of a candidate protein using a rigorous formula derived earlier (Alves and Yu, 2011); in this formula the DPV of every peptide is weighted, prior to the final combination, according to the number of proteins it maps to. Among the existing protein ID methods, the approach taken by Spirin *et al.* (2011) is closest to ours; both methods combine peptides' spectrum-specific *P*-values. There are, however, major differences between our method and that of Spirin *et al.* (2011). First, in our method, each candidate peptide of a query spectrum receives a DPV, allowing multiple matching peptides per spectrum. This is to accommodate entangled peptide co-fragmentation (Alves *et al.*, 2008a), observed more often when using lower resolution mass analyzers. For the method of Spirin *et al.* (2011), only the best peptide match per spectrum is considered and the peptide DPV thus represents the probability of having the *best* match score no worse than the prescribed threshold when searching a database. Since each random protein database only contributes one best match score, searching *many* random protein databases is required for the *P*-value assignment. Second, the candidate peptides' *P*-values are combined differently. Our method, down-weighting contributions of peptides mappable to multiple proteins, combines peptide DPVs directly using a rigorous formula (Alves and Yu, 2011); the method of Spirin *et al.* (2011) first transforms, for every candidate protein, the *P*-values of its associated peptides into *Z*-scores, combines them using Stouffer's formula (Whitlock, 2005), and then transforms the combined *Z*-score back to a final *P*-value with multiple hypotheses testing correction. Third, the cutoff conditions for peptides' *P*-values are different. Our method approximates DPVs (Alves *et al.*, 2008b; Yu *et al.*, 2006) by *E*-values, valid for small *E*-values, and retains all peptides whose *E*-values are less than one. That is, we have a global cutoff condition. For the method of Spirin *et al.* (2011), the peptide cutoff *P*-value varies by candidate protein: given a candidate protein, its corresponding peptides' *Z*-scores are first sorted in descending order; the *k*th *Z*-score is chosen as the cutoff provided that the maximum combined *Z*-score is reached while combining the top *k* *Z*-scores using the Stouffer's formula.

There exist many other protein ID methods, for example, ProFound (Zhang and Chait, 2000), ProteinProphet (Nesvizhskii *et al.*, 2003), DBParser (Yang *et al.*, 2004), EBP (Price *et al.*, 2007), PANORAMICS (Feng *et al.*, 2007), PROVALT (McHugh and Arthur, 2008), X!Tandem (Fenyo *et al.*, 2010), Scaffold (Searle, 2010) and npCI (Serang *et al.*, 2013), to name just a few. We refer the readers to recent review papers (Huang *et al.*, 2012; Serang and Noble, 2012) for details and more comprehensive listings of these methods. Although some of them do start with spectrum-specific peptide *P*-values, they often *assume* certain parametric forms for the peptide score distributions when searching a random database; other methods, however, only process outputs of specific peptide ID tools, limiting their uses to certain platforms. By discarding all but the *best* few peptide scores per spectrum per database search, the method of Spirin *et al.* (2011) does not rely on the accuracy of the full peptide score distribution from searching a random database and in principle can accept input from various peptide ID tools. Our method is free from the aforementioned problems for different reasons. Founded on a derived analytical formula, our method can be applied in general and will yield accurate protein *P*-values if the input peptide DPVs (or *E*-values) are accurate. When using peptide *E*-values reported by RAId_DbS, even though the parameters of the score distribution are determined via maximum-likelihood, the functional form of the score distribution is analytically derived (Alves *et al.*, 2007) rather than assumed. When the statistical significances are obtained from RAId_aPS (Alves *et al.*, 2010), for every scoring function implemented, the *P*-values are inferred by scoring APP instead of assuming that the score histogram follows a specific form; the peptide *E*-values are then obtained via multiplying the *P*-values by the respective numbers of qualified peptides.

The article is organized as follows. The mathematical underpinnings of our formalism will be described in Section 2. In Section 3, comparisons of our method with other approaches will be made; the accuracy of the reported protein *P*-value will be illustrated. Some technical but important issues will be addressed in Section 4. To keep the article focused, we relegate to supplementary information figures and tables that complement or corroborate the information contained in the main text.

2 Methods

2.1 Statistical protocols

Weighting the contribution of each peptide in protein ID is important. It helps mitigate the issue of peptide degeneracy, where an identified peptide is a subsequence of multiple database proteins. The optimal weighting scheme, however, can depend on the protein ID methodology employed. For the purpose of our study, namely, devising a method that yields accurate protein *P*-values, we opt for a simple weighting scheme: a peptide's weight is inversely proportional to the number of database proteins it maps to. Within a sample, when multiple spectral searches identify the same peptide but with different significance levels, only the most significant assignment of that peptide is retained for further analyses.

The foundation of our method is built upon a rigorous formula (Alves and Yu, 2011; Mathai, 1983) that enables weighted combination of *P*-values. When the weights are all identical, this formula reduces to Fisher's formula (Bahrucha-Reid, 1960; Fisher, 1932); when the weights are all different, this formula reduces to the formula of Good (1955). A detailed derivation and generalization to incorporating nearly identical weights can be found in Alves and Yu (2011), whose notation will be used to briefly summarize the content of the formula.

Let us assume that a given protein contains L identified peptides with *P*-values. Let us further group these L peptides, according to the number of database proteins a peptide maps to, into m groups with $1 \le m \le L$. Within each group k, the n_k peptide *P*-values are weighted equally; while peptide *P*-values in different groups are weighted differently.

The weighting enters our formalism through the following quantities of interest

$$\tau \equiv \prod_{k=1}^{m} \left[\prod_{j=1}^{n_k} p_{k;j} \right]^{w_k}, \tag{1}$$

$$Q \equiv \prod_{k=1}^{m} \left[\prod_{j=1}^{n_k} x_{k;j} \right]^{w_k}, \tag{2}$$

where each $p_{k;j}$ represents a reported peptide *P*-value, each $x_{k;j}$ represents a random variable drawn from an uniform, independent distribution over $[0, 1]$ and each w_k is a positive weight. The quantity of interest $\text{Prob}(Q \le \tau)$, representing the protein *P*-value, was obtained earlier (Alves and Yu, 2011) and is repeated below for clarity.

Let $F(\tau) \equiv \text{Prob}(Q \le \tau)$, one may show that

$$F(\tau) = \left[\prod_{l=1}^{m} r_l^{n_l} \right] \sum_{k=1}^{m} \sum_{\mathcal{G}(k)} \left\{ \frac{1}{r_k^{g_k+1}} H(-r_k \ln \tau, g_k) \times \left(\prod_{j=1, j\neq k}^{m} \frac{(n_j - 1 + g_j)!}{(n_j - 1)! g_j!} \frac{(-1)^{g_j}}{(r_j - r_k)^{n_j+g_j}} \right) \right\}, \tag{3}$$

where $r_k \equiv 1/w_k$ is the number of proteins a group-k peptide maps to, $\sum_{\mathcal{G}(k)}$ enumerates each set of nonnegative integers $\{g_1, g_2, \ldots, g_m\}$ that satisfies the k-dependent constraint $\sum_{i=1}^{m} g_i = n_k - 1$, and the function H is defined as

$$H(x, n) \equiv e^{-x} \sum_{k=0}^{n} \frac{x^k}{k!}. \tag{4}$$

See the Supplementary information for an example application of formula (3).

When searching a database with a prescribed peptide mass error tolerance δ, one often needs to score different numbers of database peptides for spectra with different precursor ion masses. That is, the number of tested hypotheses (database peptides in the mass range $[m_p - \delta, m_p + \delta]$) varies by the precursor ion mass m_p. The effect of varying number of multiple hypotheses tested can be properly accounted for by using the peptide DPVs (Alves *et al.*, 2008b; Yu *et al.*, 2006) for *P*-values ($p_{k;j}$) in equation (1); given a quality score cutoff S_c, the peptide DPV is defined as

$$P_{\text{db}}(S_c) = 1 - e^{-E(S_c)}, \tag{5}$$

where $E(S_c)$ represents the expected number of peptides having score $S \ge S_c$, and the DPV $P_{\text{db}}(S_c)$ represents the probability of seeing one or more peptides in a given random database with quality scores $S \ge S_c$. Another advantage of using DPV is that as a function of the quality score S, the *E*-value $E(S)$, determined by the search score histogram per spectrum and the number of qualified peptides (database-dependent), correctly takes into account both the spectrum-specificity and the database-specificity of scoring statistics.

Since the *E*-value specifies the expected number of random database peptides having scores equal to or better than the given cutoff, a peptide with *E*-value larger than one is more likely to be a false positive than a true positive. For this reason, when constructing the evidence peptide set for ID of a protein, we only include peptides with *E*-values <1. This implies that only peptide DPVs $< (e-1)/e$ are considered, leading to a combination of truncated *P*-values. Unfortunately, combining truncated *P*-values, even though doable, is far more complicated than using equation (3). However, two observations simplify the matter. First, it is evident from equation (5) that the DPV approaches the *E*-value when the *E*-value is small. Second, we note that confidently identified proteins must contain evidence peptides with high ID confidences (or small *E*-values). Therefore, for practical uses, we may approximate the DPV by its corresponding *E*-value. Because only *E*-values <1 are considered, the approximated DPVs (or simply the *E*-values) now encompass the full range between 0 and 1. Consequently, it is unnecessary to combine truncated *P*-values, and the simple formula (3) becomes applicable. The protein *E*-value is then obtained via multiplying the protein *P*-value by a Bonferroni correction factor; in this case, the Bonferroni factor is the number of protein clusters (described below) each having at least one evidence peptide with *E*-value <1.

We denote by a protein cluster a group of *entangled* proteins that share a substantial portion of evidence peptides. To avoid exaggerating the number of identified proteins, several existing methods (Huang *et al.*, 2012) report those entangled proteins as one. Adopting the same idea, we implemented this strategy via a transitive approach described below. One first sorts the identified proteins by the number of identified evidence peptides in descending order and using the rank of a protein in the sorted list as that protein's cluster index. Starting with the first protein as the reference protein, all other lower-ranking proteins sharing at least 95% of evidence peptides with the first protein will have their cluster indexes changed to that of the reference protein. One then moves the reference point (from the first) to the second protein, all other lower-ranking proteins sharing at least 95% of evidence peptides with the reference protein will have their cluster indexes changed to that of the reference protein. The reference point is then moved to the third protein and the process continues till the reference point moves through all proteins in the list. The most significant *P*-value within a cluster (containing one or more proteins) is regarded as the *P*-value associated with that cluster; the protein with the largest number of

evidence peptide is called the head of that cluster, the other proteins members of that cluster. An exception to the aforementioned clustering rule, however, is introduced to appropriately emphasize a protein's evidence peptides that are not shared by other proteins. We call evidence peptides of this kind *unique* peptides to a protein. When a protein has a unique evidence peptide with *E*-value $<10^{-4}$, our method does not allow this protein to be a member protein of any cluster.

2.2 MS/MS datasets

Sixty-three spectral datasets were categorized into four data groups. See Supplementary Tables S1–S4 for details. Protein mixtures giving rise to spectral datasets were reduced with iodoacetamide, resulting in the addition of the carbamidomethyl group (57.07 Da) to cystine residues. Each protein mixture was further digested with trypsin. Among these spectral datasets, there are also dataset-specific parameters such as the target database, the maximum number of missed cleavage sites allowed, the precursor-ion mass error tolerance and the product-ion mass error tolerance. The dataset-specific parameters are given in the figure caption to provide more information underlying the generation of the figures.

For brevity, we shall denote the MS/MS spectra obtained from a sample by SN followed by its sample index. For example, SN1 denotes the collection of MS/MS spectra acquired from mixture sample one. The first data group, SN1–SN15, contained MS/MS spectra from replicates of different dilutions of Sigma49, a protein standard mixture composed of 49 know human proteins. The second data group, SN16–SN26, was downloaded from the Pacific Northwest National Laboratory and contained spectra from 11 whole-cell-lysate samples of protein mixtures of *Escherichia coli K-12*. The third data group, SN27–SN30, consisted of spectra from four in-house whole-cell-lysate samples of protein mixtures of *E.coli K-12*. Downloaded from PeptideAtlas database, the fourth data group (SN31–SN63) was composed of spectra from SDS-PAGE protein fractionation extractions of human lung cells.

2.3 Protein databases and random databases

Because protein mixtures from *E.coli K-12* and *Homo sapiens* were analyzed using their corresponding MS/MS spectra, protein databases for both organisms were thus required. From UniProt http://www.uniprot.org/downloads, we downloaded 4303 non-redundant protein sequences of *E.coli K-12*. A non-redundant *H.sapiens* protein database, containing 31 236 protein sequences, was obtained from the NCBI site ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/.

When analyzing statistical significance, it is often required to have random (decoy) databases in addition to the organismal (target) databases. One common problem when using random databases is that for a given precursor ion mass the numbers of qualified peptides in the random database and in the organismal database may significantly differ. This causes an additional uncertainty in assessing statistical significance (Elias and Gygi, 2007; Wang *et al.*, 2009). We can avoid this problem by ensuring that the numbers of qualified peptides per spectrum are identical for both the random and the organismal databases: for each qualified peptide in the organismal database, we generate a corresponding random peptide by randomly shuffling its amino acids.

3 Results

The results will be described in the following order. First, we illustrate that our *E*-value assignments are accurate at both the peptide and the protein levels. We further show that using the formula proposed by Sorić (1989), our reported PFDs agree well with the target-decoy PFDs. Second, our protein *E*-value accuracy is compared with that of using the formulas in Spirin *et al.* (2011). By extending the formula of Sorić for the method of Spirin *et al.* (2011), we also evaluate the agreement between their reported PFDs and the target-decoy PFDs. Benchmarking with some of the existing protein ID methods will be described in the third part.

3.1 *E*-value accuracy

The input peptide DPVs for our protein ID method are obtained via Equation (5) using the *E*-values reported by RAId_DbS. For this reason, the input peptide DPVs (for protein ID) are synonymous with the reported peptide DPVs (from RAId_DbS). As mentioned earlier, the statistical accuracy of our protein ID method relies on the DPVs for the evidence peptides being accurate. We therefore start by comparing the input peptide DPV with its definition. In panel A of Figure 1, the abscissa records the peptide DPV, while the ordinate displays the *observed* DPV (i.e. fraction of spectra having at least one or more matching peptides with reported DPVs smaller than the specified threshold). The agreement between the observed DPV and the reported DPV indicates that the peptide DPVs used as input for our protein ID method are accurate.

To assess whether approximating peptide DPVs by their corresponding *E*-values for *E*-values <1 is reasonable or not, we plot in panel B of Figure 1 the observed peptide DPVs versus *E*-values. As expected, when *E*-values are close to 1, there is certain degree of disagreement; while for small *E*-values, the agreement is excellent. To assess the accuracy of the protein *P*-values reported by Equation (3), we compare them with the observed protein *P*-values. As described in Section 2, the reported proteins appear in clusters, each represented by a *head* protein and its *P*-value. The observed protein *P*-value is defined as the fraction of identified protein clusters (whose member proteins each containing at least one evidence peptide with *E*-value <1) that have reported *P*-values smaller than a

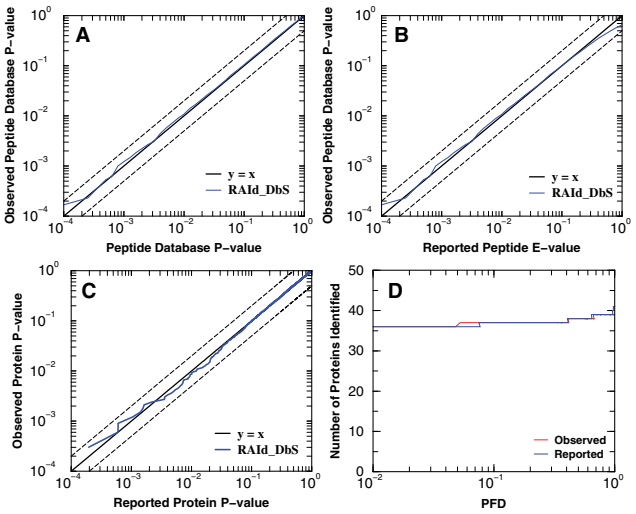

Fig. 1. Assessment of *E*-value accuracy. In panels **A–C**, the closer the displayed curves are to the $y = x$ line the better. In panel D, the closer the two displayed curves are to each other the better. See Section 3.1 for more details. For panels A–C, spectral dataset SN26 (*E.coli K-12* whole cell lysate) is used to search the *E.coli* database with mass accuracy ±0.033 Da for precursor and product ions. For panel D, spectral datasets SN13–SN15 (Sigma49 protein standard mixture) is used to search the *H.sapiens* database with precursor ion accuracy ±0.033 Da and product ion accuracy ±0.8 Da

given threshold. As shown in panel C of Figure 1, good agreement between the reported protein *P*-value and the observed protein *P*-value is obtained, indicating that our reported protein *P*-values are accurate. More protein *P*-value accuracy assessment examples can be found in Supplementary Figures S1–S3. With an accurate protein *P*-value, one can also obtain its corresponding protein *E*-value by multiplying it by the total number of protein clusters. In Supplementary Figure S4, we show that reported protein *E*-values obtained this way are accurate.

By having accurate protein *E*-values, one can avoid the uncertainty associated with using a decoy database (Gupta *et al.*, 2011) while estimating the proportion of false discoveries. In panel D of Figure 1, we plot two PFD curves: one is computed using the reported protein *E*-value to estimate the number of false ID (hence the PFD), while the other is computed using the observed PFD obtained from known target protein content in the sample (Sigma49). The excellent agreement between the observed PFD and the reported PFD indicates that one should be able to trust the PFD estimated from accurate reported protein *E*-values. More accuracy assessment examples of the reported PFD can be found in Supplementary Figure S5.

3.2 Comparison with an EVD-based method

Since the method of Spirin *et al.* (2011) is closest to ours, we also implemented their method and compute equivalent quantities for comparison. Following the Supplementary Material of Spirin *et al.* (2011), we have implemented 100 random databases each containing 10 000 random amino acid sequences. However, instead of generating sequences of uneven length, we opt for uniform length (each sequence is of length 350) and generate these random sequences using the background amino acid frequencies of Robinson and Robinson (1991). The EVD parameters are obtained by using only the best score per database search and by applying standard procedures described in Spirin *et al.*, (2011). The effect of database size difference, leading to rescaling of the α parameter, is done the same way as in Spirin *et al.* (2011).

A moment of reflection reveals that the best match *P*-value of Spirin *et al.* (2011) is in fact the DPV (Alves *et al.*, 2008b; Yu *et al.*, 2006). We therefore plot in panel A of Figure 2 the reported peptide DPVs against the observed peptide DPVs. The result indicates that the peptide DPV reported by Spirin *et al.* (2011) is quite accurate, with an uncertainty of a factor of 2 as reported by Spirin *et al.* (2011).

To have a fair assessment, the same procedure for clustering proteins is also applied to the proteins identified using protocols of Spirin *et al.* (2011). Database proteins that contain any of the best match peptides, one from each spectrum, form the effective protein set, within which each group of entangled proteins forms a cluster. The observed protein *P*-value is defined similarly: the fraction of identified protein clusters that have reported protein *P*-values less than the specified threshold. The reported protein *P*-value for the head protein of each cluster is obtained by applying the iterative procedure (involving uses of Stouffer's formula) described in Spirin *et al.* (2011). In panel B of Figure 2, the reported protein *P*-values are plotted against the observed protein *P*-values. The agreement between the reported protein *P*-values and the observed protein *P*-values is not as great as in the peptide case. The protein *E*-value is then obtained by multiplying the protein *P*-value by the total number of proteins in the effective protein set.

To construct a PFD curve, it is necessary to estimate the number of false ID at a given significance threshold. The number of false ID can be estimated either by using the reported protein *E*-values or the number of ID within the decoy databases. The latter is currently widely used mainly because accurate protein *E*-values (or *P*-values) are generally hard to attain. To investigate the agreement between the PFD curves obtained using decoy databases and using reasonably accurate protein *P*-values, we use spectra acquired from dataset SN26 and construct the PFD curves obtained using both approaches. The good agreement between our *E*-value-based PFD (Sorić, 1989) and the target-decoy-based PFD, displayed in panel C of Figure 2, is expected because, as shown in panel D of Figure 1, we have already found that the reported PFD and the observed PFD (computed by using a known protein mixture) are nearly identical. The disagreement between the *E*-value-based PFD and the target-decoy-based PFD using protocols of Spirin *et al.* (2011) seems to indicate that the moderate uncertainty in DPV can influence the accuracy of the overall PFD estimate in a substantial manner.

For RAId_DbS, the agreement between our *E*-value-based PFD and the target-decoy-based PFD is further tested using more spectral datasets (SN16–SN25), see Supplementary Figure S6. In addition to RAId score, RAId_aPS allows other scoring functions: XCorr, Hyperscore and Kscore. For completeness, we plot their corresponding protein *P*-value accuracy assessments in Supplementary Figures S7–S9; we also present the agreement tests between their *E*-value-based PFDs and the target-decoy-based PFDs in Supplementary Figures S10–S12.

3.3 Comparison with other methods

In terms of computation speed of our approach, like other methods, the most time-consuming component lies in peptide ID. Although the computation speeds of various peptide ID methods were surveyed by Diament and Noble (2011), RAId_DbS was not included. To complement the survey information, we provide the peptide/protein ID speed of our method below based on a single core usage

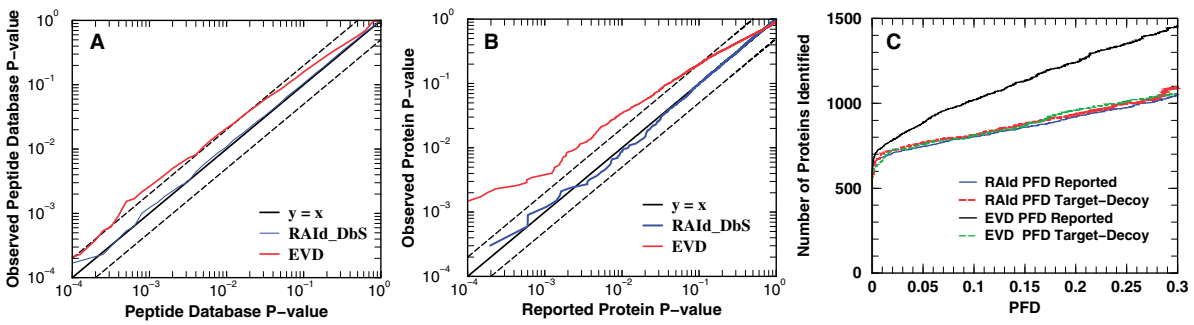

Fig. 2. Statistical accuracy comparison. Except that results from two methods are being displayed, panels **A–C** display similar information, respectively, to panels A, C and D of Figure 1. See Section 3.2 for more details

on an intel Xeon 2.8 GHz CPU (even though RAId_DbS is a multi-threading code capable of running on multiple cores in parallel). When searching an 18-MB protein database (allowing up to two missed cleavages) with trypsin as the enzyme, RAId_DbS analyzes approximately 10 MS/MS spectra/s and in approximately 0.5 s finishes protein ID by processing output of approximately 15 000 spectra.

The previous two subsections focus on the accuracy of type-I error control. Although it is possible to accurately control type-I error for some protein ID methods, this seems not the central focus of all protein ID methods. Many protein ID methods prefer to use the decoy database search results to pragmatically provide statistical significances for retrieval results from the target (organismal) database. When this approach is used, the retrieval results are displayed in terms of a parametric PFD plot: the parameter is some kind of significance score used to prioritize the ID, the abscissa shows the PFD and the ordinate displays the number of ID found in the target database. In general, a large number of target ID at a small PFD value indicates a good retrieval, provided that the number of decoy ID accurately reflects the number of false ID in the target database. However, one should note that the fulfilment of the aforementioned condition requires accurate type-I error control. Investigating and improving the statistical accuracy of type-I error control of existing protein ID methods is beyond the scope of this article and we believe that it is best done by developers of individual protein ID software.

To examine how our method compares with others under the pragmatic target-decoy approach, we analyze two large datasets from *E.coli* (SN27–SN30) and *H.sapiens* (SN31–SN63) using a variety of protein ID software along with a number of scoring functions. The list of software is given below (with both software version and scoring functions, if given, shown inside a pair of parentheses): RAId_DbS (v. Jan.12.2014; RAId), RAId_aPS (v. Jan.12.2014; XCorr, Kscore, Hyperscore), Mascot (v. 2.4.0, http://www.matrixscience.com/help.html), and X!Tandem (v. 2013.06.15; Hyperscore). The peptide ID software SEQUEST (Eng *et al.*, 1994) (v. 28) is only used in conjunction with other post-processing protein ID software. We list below the post-processing software used (with software version, peptide ID software and peptide scoring functions, if given, shown inside a pair of parentheses): iProphet (v. TPP 4.5; X!Tandem; Kscore), Proteome Discoverer (v. 1.3, http://www.thermofisher.com/en/home.html; SEQUEST, Mascot), and Scaffold Q+/Q+S (v. 4.0, http://www.proteomesoftware.com; SEQUEST, Mascot). The results are displayed in different panels of Figure 3. Before delving into the details of the results, we first provide the information relevant to the generation of the results.

In terms of peptide ID, RAId_DbS, RAId_aPS, Mascot, SEQUEST and X!Tandem used the same parameters: for *E.coli* whole cell lysate, SN27–SN30, the precursor ion mass error tolerance is ±0.033 Da, the product ion mass error tolerance is ±0.033 Da, and up to five missed cleavages are allowed; for *H.sapiens* lung cells, SN31–SN63, the precursor ion mass error tolerance is ±1.4 Da, the product ion mass error tolerance is ±0.4 Da, and up to two missed cleavages are allowed.

Both X!Tandem and Mascot have built-in protein ID capability, and the target-decoy approach was directly applied to estimate the protein level PFD. The peptide ID outputs from SEQUEST and Mascot were also further analyzed using Proteome Discoverer for protein ID and the target-decoy approach was applied to estimate PFD. For iProphet, we did not compute the PFD but downloaded the results for data group 4 from PeptideAtlas. Peptide ID in this case was done using X!Tandem (v. 2009.10.01; Kscore).

Whenever the decoy peptide search results are available, Scaffold computes the PFDs using the target-decoy approach; otherwise, it computes the PFDs using a probabilistic method. In Figure 3, three

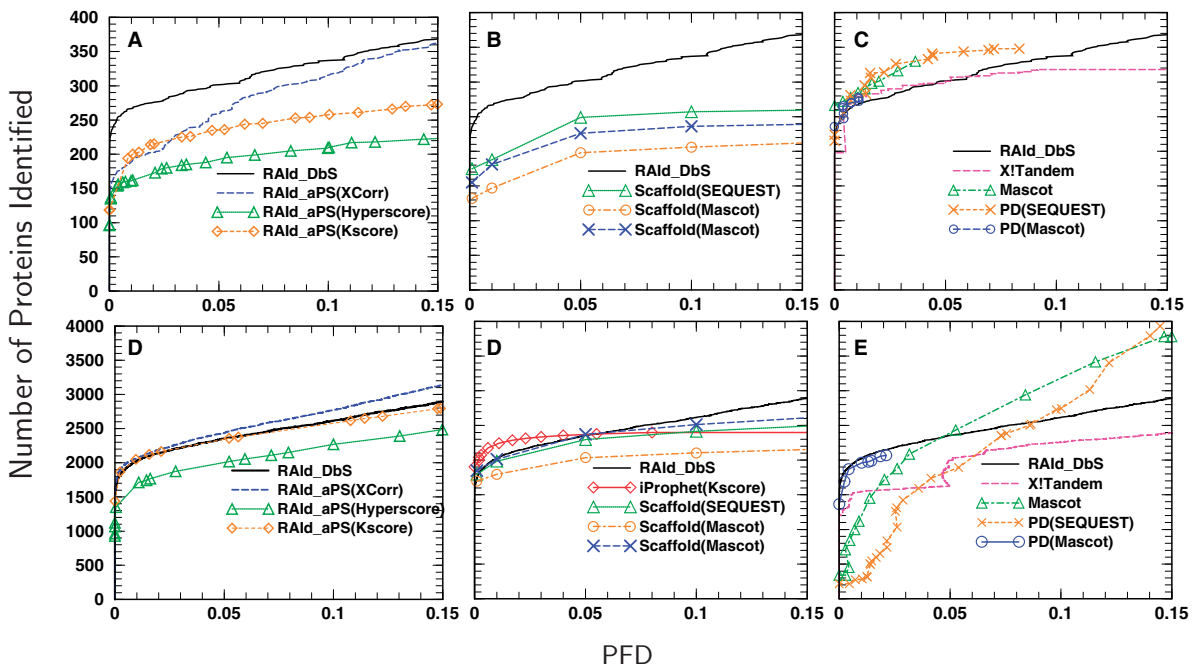

Fig. 3. Retrieval results of various methods based on their stated PFD values. Because we can only ensure the accuracy of type-I error of the proposed method, this figure only illustrates how our retrieval fares at the stated value when compared with other methods. See Section 3.3 for more details. To avoid clutter, results from using samples of *E.coli* whole cell lysyte, SN27–SN30, are displayed in panels **A–C**. Within each panel, the results from RAId_DbS are always shown as a reference curve. Similarly, results from using samples of *H.sapiens* lung cells, SN31–SN63, are also displayed in **D–F** panels. The iProphet (Shteynberg *et al.*, 2011) results in panel E were downloaded from PeptideAtlas instead of being computed

Scaffold PFD curves are displayed, two of which (shown in triangles and circles) are from target-decoy approaches. The protein PFDs under Scaffold were computed by fixing the peptide threshold at 20% PFD with a minimum of one evidence peptide per protein. We observed that changing the peptide threshold to lower values had a small effect on the number of proteins identified. We thus used the minimum of one peptide per protein to maintain consistency across all methods. For RAId_DbS and RAId_aPS, the PFD estimates do not require user-added target-decoy methods. RAId_DbS and RAId_aPS compute the PFDs using the Sorić formula (Sorić, 1989).

Examinations of different panels of Figure 3 indicate that the retrieval efficacy of the proposed method (shown in RAId_DbS and RAId_aPS PFD curves) is comparable with existing protein ID methods, even though only at the stated values. However, it should be noted that the proposed method does have a few advantages. First, it reports accurate protein *P*-values, providing accurate type-I error control. Second, the PFD curves obtained using this method show stability across different mass resolution requirement and datasets, while some methods seem to exhibit fluctuations of notable amplitudes.

4 Discussion

Our investigation indicates that it is possible to achieve faithful protein *P*-value assignment, hence accurate type-I error control, in protein ID. Since our approach is founded on a derived mathematical formula that requires accurate peptide *E*-values as input, it is evident that accurate protein *P*-values require accurate statistical significance at the peptide ID level.

The discrepancy between the computed protein *P*-value and the PFD results in our implementation of the method of Spirin *et al.* (2011) is interesting. Based on the results in Figure 2, the peptide *P*-values are reasonably accurate albeit exhibiting slightly larger fluctuations than the results from RAId_DbS. In addition to the possibility of accumulating uncertainty of peptides' *P*-values, the other possibility is that the iterative procedure to choose the combination yielding the most significant *Z*-score may skew the *P*-values toward the significant side. Investigation of the origin of the PFD and *P*-value discrepancy when using the method of Spirin *et al.* (2011), however, is beyond the scope of this study and might be most appropriately done by the authors of (Spirin *et al.*, 2011).

As explained earlier, we allow more than one candidate peptide per spectrum to accommodate cofragmentation of multiple precursor ions with proximate masses. (When a low-resolution mass analyzer is used, it can happen that more than one underlying peptide appears significant.) However, readers may ask why do we choose to use DPVs for lower-ranking peptides per spectrum instead of using ordered statistics. The reason is that in this context using ordered statistics beyond the first is not meaningful: the *n*th-ordered statistics assumes that for a given query spectrum the best $n-1$ scored peptides are spurious while the rank-*n* peptide is the underlying peptide whose fragmentation yields the query spectrum. This contradicts the general idea of using a scoring function: among candidate peptides of a query spectrum, the better a peptide scores the more likely it is the underlying peptide. On the other hand, when using the DPV for the rank-*n* peptide, we are essentially assuming that the top $n-1$ candidate peptides of the query spectrum are cofragmented underlying peptides and are not considered to be spurious.

The protein ID method proposed in this article illustrates the possibility of accurate type-I error control, providing a theoretically sound significance assignment method that is also pragmatically simpler than the target-decoy approach. This is particularly important since the number of identified proteins versus PFDs provides trustworthy retrieval results only if the reported PFDs truly reflects the proportion of false discoveries. Evidently, to achieve accurate type-I error control is a task best done by developers of individual software. Only when this is accomplished can a true retrieval comparison among different methods be done.

Since we did not focus on type-II error, there is definite room for improvement in terms of retrieval efficacy. We note that the information of negatives (segments of a candidate protein not covered by the protein's evidence peptides) is not used. We also believe that, in principle, scoring functions for peptide ID can also be improved to better separate true underlying peptides from false positives. Currently, we are using a flat peptide weight (by the number of proteins a peptide covers). It is perceivable that more sophisticated weighting may be useful in better separating true positive proteins from false positives. It is our plan to investigate these avenues of improvement in the near future.

Acknowledgements

We thank the administrative group of the National Institutes of Health Biowulf Clusters, where most computational tasks were carried out. We also thank the National Heart Lung and Blood Institute proteomics core for assistance.

Funding

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

Conflict of interest: none declared.

References

Alves,G. and Yu,Y.K. (2008) Statistical characterization of a 1D random potential problem—with applications in score statistics of MS-based peptide sequencing. *Physica A*, **387**, 6538–6544.

Alves,G. and Yu,Y.K. (2011) Combining independent, weighted P-values: achieving computational stability by a systematic expansion with controllable accuracy. *PLoS ONE,* **6**, e22647.

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

Alves,G. *et al.* (2008a) Detection of co-eluted peptides using database search methods. *Biol. Direct,* **3**, 27.

Alves,G. *et al.* (2008b) Enhancing peptide identification confidence by combining search methods. *J. Proteome Res.*, **7**, 3102–3113.

Alves,G. *et al.* (2010) RAId_aPS: MS/MS analysis with multiple scoring functions and spectrum-specific statistics. *PLoS One,* **5**, e15438.

Bahrucha-Reid,A. (1960) *Elements of the Theory of Markov Processes and Their Applications*. McGraw-Hill, New York.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological),* **57**, 289–300.

Diament,B.J. and Noble,W.S. (2011) Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.*, **10**, 3871–3879.

Elias,J.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.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.,* **5**, 976–989.

Feng,J. *et al.* (2007) Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics,* **23**, 2210–2217.

Fenyo,D. and Beavis,R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, **75**, 768–774.

Fenyo,D. *et al.* (2010) Mass spectrometric protein identification using the global proteome machine. *Methods Mol. Biol.*, **673**, 189–202.

Fisher,R.A. (1932) *Statistical Methods for Research Workers*. Vol. II. Oliver and Boyd, Edinburgh.

Good,I.J. (1955) On the weighted combination of significance tests. *J. R. Stat. Soc. Ser. B (Methodological),* **17**, 264–265.

Gupta,N. *et al.* (2011) Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.*, **22**, 1111–1120.

Huang,T. *et al.* (2012) Protein inference: a review. *Brief. Bioinform.*, **13**, 586–614.

Li,Y.F. and Radivojac,P. (2012) Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics.,* **13** (Suppl. 16), S4.

Mathai,A. (1983) On linear combinations of independent exponential variables. *Commun. Stat. Theory Methods,* **12**, 625–632.

McHugh,L. and Arthur,J.W. (2008) Computational methods for protein identification from mass spectrometry data. *PLoS Comput. Biol.*, **4**, e12.

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

Noble,W.S. and MacCoss,M.J. (2012) Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.*, **8**, e1002296.

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

Price,T.S. *et al.* (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Mol. Cell Proteomics,* **6**, 527–536.

Robinson,A.B. and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl Acad. Sci. USA,* **88**, 8880–8884.

Searle,B.C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics,* **10**, 1265–1269.

Segal,M.R. (2008) On *E*-values for tandem MS scoring schemes. *Bioinformatics,* **24**, 1652–1653.

Serang,O. and Noble,W. (2012) A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface,* **5**, 3–20.

Serang,O. *et al.* (2013) A non-parametric cutout index for robust evaluation of identified proteins. *Mol. Cell Proteomics,* **12**, 807–812.

Shteynberg,D. *et al.* (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell Proteomics,* **10**, M111.007690.

Sorić,B. (1989) Statistical "discoveries" and effect-size estimation. *J. Am. Stat. Assoc.,* **84**, 608–610.

Spirin,V. *et al.* (2011). Assigning spectrum-specific *P*-values to protein identifications by mass spectrometry. *Bioinformatics,* **27**, 1128–1134.

Wang,G. *et al.* (2009) Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.*, **81**, 146–159.

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.

Yang,X. *et al.* (2004) DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.*, **3**, 1002–1008.

Yu,Y.K. *et al.* (2006) Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res.*, **34**, 5966–5973.

Zhang,W. and Chait,B.T. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, **72**, 2482–2489.

Zhang,Y. *et al.* (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.*, **113**, 2343–2394.