Bioinformatics, 2016, 1–8
doi: 10.1093/bioinformatics/btw225
Advance Access Publication Date: 29 April 2016
Original Paper

OXFORD

Structural bioinformatics

Confidence assignment for mass spectrometry based peptide identifications via the extreme value distribution

Gelio Alves and Yi-Kuo Yu*

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

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on November 5, 2015; revised on March 24, 2016; accepted on April 16, 2016

Abstract

Motivation: There is a growing trend for biomedical researchers to extract evidence and draw conclusions from mass spectrometry based proteomics experiments, the cornerstone of which is peptide identification. Inaccurate assignments of peptide identification confidence thus may have far-reaching and adverse consequences. Although some peptide identification methods report accurate statistics, they have been limited to certain types of scoring function. The extreme value statistics based method, while more general in the scoring functions it allows, demands accurate parameter estimates and requires, at least in its original design, excessive computational resources. Improving the parameter estimate accuracy and reducing the computational cost for this method has two advantages: it provides another feasible route to accurate significance assessment, and it could provide reliable statistics for scoring functions yet to be developed.

Results: We have formulated and implemented an efficient algorithm for calculating the extreme value statistics for peptide identification applicable to various scoring functions, bypassing the need for searching large random databases.

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

Assigning accurate statistical significance to identified peptides is fundamentally important because these peptides form the basis for protein identifications, whose results are often used to infer biological processes and functions. Statistical errors at the peptide identification level inevitably propagate downstream, and may have far reaching consequences, undermining the biological conclusions drawn from high level analyses. Therefore, various approaches for bounding the occurrence of false identifications have been developed. One frequently employed method is to globally control the proportion of false discoveries (PFD) (Benjamini and Hochberg, 1995) per experimental dataset. The target-decoy approach (Elias and Gygi, 2007) is often used to estimate the PFD, with its numerator (the number of false positives in the target database) being approximated by the number of decoy hits and with its denominator being the number of peptides identified in the target database. However, as discussed by Gupta *et al.* (2011), this estimate can be inaccurate when a global grading standard to prioritize the peptide hits across spectra and across databases is lacking. Furthermore, as emphasized by Higdon *et al.* (2008), even with a correct grading standard present and under the same PFD cutoff, two candidate groups from two different experiments do not have identical statistical significances.

These aforementioned problems can be mitigated, however, by computing per spectrum *E*-values/*P*-values. Given a tandem mass

spectrometry (MS/MS) spectrum and a quality score cutoff S_c, the *E*-value $E(S_c)$ is defined to be 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 mass differences from the precursor ion are within the specified tolerance. There are multiple advantages to employing *E*-values/*P*-values for controlling false positives (type-I error): it permits peptide/protein prioritization across spectra and experiments (Alves *et al.*, 2007a); it allows the statistical significances from different analyses to be combined (Alves *et al.*, 2008a). Since the expected number of random matches per query can be defined without inferring experimental details, the *E*-value is an ideal choice for the universal grading standard for candidate peptides. This greatly assists peptide prioritization in the target-decoy approach. Furthermore, when candidate peptides are ranked by their *E*-values and a threshold *E*-value is chosen, one may estimate the number of false positives by multiplying the threshold *E*-value by the number of spectra analyzed. This estimated number of false positives divided by the total number of candidate peptides with *E*-values smaller than the threshold *E*-value is the PFD estimate of Sóric. Evidently, the realization of the statistical benefits described requires accurate *E*-values/*P*-values.

There exist several methods that compute *E*-values/*P*-values using spectrum-dependent information (Eng *et al.*, 2008; Fenyo and Beavis, 2003; Geer *et al.*, 2004; Klammer *et al.*, 2009). These methods use various parametric functions to fit the peptide score histogram *per spectrum* to estimate the *E*-values/*P*-values for the peptides identified. Even though these *assumed* parametric functions may fit the histogram, a per-spectrum *goodness-of-fit* (GOF) is not provided. Hence, there is no guarantee that such procedures can consistently yield accurate *E*-values/*P*-values (Alves *et al.*, 2007a; Segal, 2008). Nevertheless, there exist a few published methods that do not assume any parametric form for the score distribution and are able to compute accurate spectrum-specific significance consistently. One of these methods extends the central limit theorem (CLT) by analytically deriving a parametric distribution function, accounting for finite sample size and skewness, to fit the peptide score histogram per spectrum (Alves *et al.*, 2007b). Another method computes *E*-values/*P*-values by using a dynamic programming algorithm, also known as all possible peptides statistics (APPS), to generate the score histogram of all possible peptides whose masses are close enough to that of the precursor ion (Alves and Yu, 2008; Alves *et al.*, 2010; Kim *et al.*, 2008). Alternatively, similar to sequence alignment statistics (Yu and Hwa, 2001), one may use the extreme value distribution (EVD) for significance assignment. Indeed, Spirin *et al.* (2011) infer spectrum-specific statistics by employing the EVD with the parameters estimated from searching multiple random databases. More information about the CLT and the EVD is provided in Section 2.1.

Although all three aforementioned methods can provide accurate significance estimates (without assuming parametric fitting functions), each admits some limitations. Specifically, the CLT-based method is applicable only to a single choice of scoring function, the average of the sum of independent contributions; the APPS-based method requires each of its scoring functions be a sum of independent contributions; and the EVD-based method, whose parameter learning is in general challenging, can be applied only to scoring functions whose resulting score histograms fall in the domain of attraction of the EVD (Gumbel, 1958; Kotz and Nadarajah, 2000). Among these three methods, however, the EVD-based one offers perhaps the most flexibility: permitting scoring functions other than the sum of contributions, it may enable the development of more discriminative functions that improve the sensitivity and specificity of peptide identification.

To estimate the EVD parameters, Spirin *et al.* (2011) propose searching 100 (or 10) random protein databases and obtaining from each the best (or the best 10) score(s). Every random protein database of (Spirin *et al.*, 2011) is made of 10 000 random protein sequences with lengths distributed according to the mouse proteome. For each MS/MS spectrum, its corresponding set of best scores is used to find the EVD parameters maximizing the order statistics probability density function. This procedure, however, faces several challenges. First, for a given precursor-ion mass, the number of scored peptides may vary among the 100 (or 10) random protein databases. Second, even within the same random database, the number of scored peptides may also vary by precursor-ion masses. The set of best scoring peptides may thus be sampled from populations of varying size, affecting the accuracy of the estimated EVD parameters. Third, as the mass accuracy of the precursor ion increases, the small mass-error tolerance yields few database peptides to score, decreasing the accuracy of estimated EVD parameters because the large number of scored random peptides (NSRP) needed to accurately fit the high scoring tail of the EVD (Olsen *et al.*, 1999; Yu *et al.*, 2002) becomes unattainable. Finally, querying 100 or 10 random protein databases, as prescribed by Spirin *et al.* (2011), with a large set of MS/MS spectra can require a substantial computational cost, which may deter software developers from rigorously implementing EVD-based algorithms.

In this manuscript, we address the algorithmic challenges faced by Spirin *et al.* (2011) by completely eliminating the need for random protein databases for EVD parameter estimation. In our EVD implementation (including XCorr (Eng *et al.*, 1994), Hyperscore (Fenyo and Beavis, 2003), Kscore (MacLean *et al.*, 2006) and Rscore (Alves *et al.*, 2007b)) in RAId_DbS (Alves *et al.*, 2007b), random peptides used to estimate EVD parameters are generated on-the-fly during the program execution. This allows us to fix the NSRP = 100 000 regardless of the mass of the precursor ion, its mass error tolerance (δ), or the size of the protein database. Under this approach, the computational cost for generating high-scoring random peptides is much less than the original EVD implementation (Spirin *et al.*, 2011). However, our goal is not to design a *fast* significance assignment method, but a flexible and robust one that can be pragmatically implemented in most tools without adding much computational cost. For the convenience of the readers, we have summarized in Table 1 the acronyms used in this paper.

Table 1. The acronym table

APPS	All possible peptides statistics	DPV	Database P-value
NSRP	Number of scored random peptides	DG	Data group
PFD	Proportion of false discovery	MS/MS	Tandem mass spectrometry
δ	Precursor-ion mass error tolerance	GOF	Goodness-of-fit
ϵ	Product-ion mass grid spacing	EVD	Extreme value distribution
PTM	Post-translational modification	CLT	Central limit theorem
SAP	Single amino acid polymorphism		

2 Methods

2.1 Statistical significance assignment for peptides

Universality emerges when sampling a large number, n, of independently identically distributed random variables from some underlying distribution. The CLT states that if the underlying distribution has well defined first and second moments, the average of these n random variables follows the Gaussian distribution as $n \to \infty$. The theory of extreme, on the other hand, indicates that when the tail of the underlying distribution falls in the domain of attraction of an EVD, the maximum (or minimum) of these n random variables, as $n \to \infty$, distributes according to that EVD. Therefore, when employing the EVD theory, it is important to verify, analytically or numerically, that the underlying distribution in question belongs to the domain of attraction of the EVD (Gumbel, 1958; Kotz and Nadarajah, 2000). In this study, all the scoring functions considered have score distributions bounded by an exponential tail (theoretically or empirically obtained). One thus expects them to be in the domain of attraction of Gumbel-type EVD.

The Gumbel-type EVD can be written as

$$P[S \geq s; u, \lambda] = 1 - \exp[-\exp\{-\lambda(s-u)\}], \tag{1}$$

and its cumulant generating function is given by

$$g(t) \equiv \ln\left\{\int_{-\infty}^{\infty} e^{ts} \frac{\partial P[s; u, \lambda]}{\partial s}\, ds\right\} = ut + \ln[\Gamma(1 - t/\lambda)], \tag{2}$$

where $\Gamma(x)$ is the gamma function, u is the location parameter, $1/\lambda$ is the scale parameter and the P-value in Equation (1) represents the probability for the random score S to equal or exceed the sample maximum score s.

To estimate the u and λ parameters, we note that the mean ($\mu[s]$) and variance ($\sigma^2[s]$) correspond to the first two cumulants

$$\mu[s] = g'(t)|_{t=0} = u - \psi(1)/\lambda = u + \gamma/\lambda, \tag{3}$$

$$\sigma^2[s] = g''(t)|_{t=0} = \psi'(1)/\lambda^2 = \frac{\pi^2}{6\lambda^2}, \tag{4}$$

where $\gamma = 0.577215665\ldots$ is the Euler-Mascheroni constant, $\psi(x)$ is the digamma function and $\psi'(x)$ is the polygamma function.

2.2 EVD parameters and database *P*-value

For each MS/MS spectrum, to estimate the corresponding EVD parameters we use 100 high-scoring random peptides, each obtained from scoring a set of 1000 random peptides. These 100 000 peptides are generated in the following manner. For a given precursor-ion mass associated with an MS/MS spectrum, a set of peptides that are in the target database and are within δ of the precursor-ion mass is identified, and we refer to peptides in this set as *qualified peptides*. We then randomly select $N = 100$ qualified peptides (or $N =$ the total number of qualified peptides if there are less than 100 of them) to initiate the procedure described below.

First, for each of the N qualified peptides we generate $\lfloor 1000/N \rfloor$ random peptides by replacing parts of the qualified peptide with substitution tags that are less than 1000 Da and of unique amino acid composition. Without distinguishing leucine (L) from isoleucine (I), the use of 19 standard amino acids yields in total 3 128 177 unique substitution tags of lengths 2–14. In addition, the mass differences between the random peptides and the precursor ion are required to be smaller than δ. During this step, priority is given to unique amino acid compositions that yield random peptides with masses closest to the precursor ion. Along with the original N highest scoring peptides, this step yields in total $M = N(1 + \lfloor 1000/N \rfloor)$ peptides.

Second, for each of these M peptides, we randomly permute its amino acids, except for the terminal residue(s) targeted by the digestion enzyme(s), $\lceil 100\,000/M \rceil$ times to generate a total of $M \times \lceil 100\,000/M \rceil \geq 100\,000$ random peptides. Although we aim for $M \geq 1000$ random peptides with unique compositions prior to random permutations, sick cases may occur when δ takes too small a value. For example, the number of qualified peptides N can be zero for extremely small δ. For this case, one either declares no peptide match from the database or needs to enlarge δ. We always opt for the latter. In this case, we will increase δ until both of the following conditions are met: (a) at least one qualified peptide is found; (b) we obtain a large enough number M of random peptides with unique compositions such that the permutations of these M peptides can yield no less than 100 000 different random peptides.

Third, these $M \times \lceil 100\,000/M \rceil$ peptides are scored and randomly assigned to 100 different bins, each viewed as a random database and containing exactly 1000 random peptides. This step effectively mandates that the NSRP be 100 000 per spectrum. Fourth, the 100 best scores, one from each bin, are used to estimate $\mu[s]$ and $\sigma^2[s]$. These values are then substituted into Equations (3) and (4) to obtain

$$\lambda = \frac{\pi}{\sqrt{6\sigma^2[s]}}, \tag{5}$$

$$u = \mu[s] - \gamma/\lambda\,. \tag{6}$$

To better estimate the EVD parameters, we repeat the third and fourth steps described above 10 times, without rescoring the peptides and average the resulting EVD parameters.

Lastly, one has to account for the difference between the number of qualified random peptides, used to estimate the EVD parameters and the number of qualified target peptides. This difference can be compensated for (Yu and Hwa, 2001) by introducing an extra parameter k, which for our model is the ratio of the number of qualified target peptides to the number of qualified random peptides, i.e. $k =$ (number of qualified target peptides)/1000. The EVD, needed for computing P-values, in our study is thus given by

$$P[S \geq s; k, u, \lambda] = 1 - \exp[-k\exp\{-\lambda\,(s-u)\}], \tag{7}$$

which represents the probability of finding the best scoring peptide with a score larger than or equal to the threshold s.

One may view observing a very large score s as a rare event and model it by a Poisson process. Equation (7) then can be written as

$$P[S \geq s] = 1 - e^{-E(s)} \tag{8}$$

with the mean number of occurrences given by

$$E(s) = k\exp\{-\lambda(s-u)\}\,.$$

Here, the E-value $E(s)$ is the expected number of peptides having scores $\geq s$. The P-value in Equation (8) is then the probability of observing one or more database peptides with scores larger than or equal to the threshold s; so it is also referred to as the database P-value (DPV) (Alves *et al.*, 2008a; Yu *et al.*, 2006). Since both Equations (7) and (8) represent the same P-value, this establishes that the EVD P-value is the DPV (Alves and Yu, 2015) for the best scoring peptide per spectrum. It is also worth mentioning that when $E(s) \ll 1$ the DPV is well approximated by $E(s)$.

2.3 EVD model goodness-of-fit (GOF)

For any statistical model proposed, it is important to quantify how well the model fits the experimental data. For each MS/MS spectrum analyzed, one thus needs to assess how well the 100 max scores, used for estimating the EVD parameters, actually agree with the fitted EVD model. Inspired by the work of Kinnison (1989), we use as the GOF the correlation between the max scores $\{s_i\}_{i=1}^{100}$ and $-\ln[-\ln(RP(s_i))]$, where $RP(s_i) = 1 - R(s_i)/101$ is the rank percentile, with $R(s_i)$ being the rank of max score s_i. This correlation coefficient is computed for each of the 10 iterations used to estimate the EVD parameters, and its average is our measure for the GOF. Based on the table provided in (Kinnison, 1989), when 100 data points/scores are sampled from an EVD, one expects that more than 99% of the time the scores s_i and $-\ln[-\ln(RP(s_i))]$ have correlation greater than 0.92. Consequently, we select 0.92 as our correlation cutoff value: the EVD model is rejected if the GOF is less than 0.92; in this case, no candidate peptide is reported for the spectrum considered.

2.4 Scoring functions

In RAId_aPS (Alves *et al.*, 2010), we have implemented several scoring functions (and their associated filtering algorithms), including Kscore (a plug-in scoring function for X!Tandem (MacLean *et al.*, 2006)), Hyperscore (from X!Tandem (Fenyo and Beavis, 2003)), XCorr (from SEQUEST (Eng *et al.*, 1994)) and Rscore (from RAId_DbS (Alves *et al.*, 2007b)). This allows us to reuse them in RAId_DbS for this study. Note that in RAId programs a grid of mass points of spacing ϵ is used to speed up the scoring of product ions. More details about the mass grid and the four scoring functions are provided in the supplementary information. For each of these four scoring functions, we assess the accuracy of statistical significance assignment as well as retrieval efficacy at the peptide level under the proposed EVD method.

2.5 MS/MS data

High resolution MS/MS spectra, acquired in an LTQ Orbitrap instrument, with mass resolutions approximately 10 ppm for precursor ions and 100 ppm for product ions, from whole-cell-lysate samples for three strains of bacteria were downloaded from the Pacific Northwest National Laboratory website at http://omics.pnl.gov/. Assuming the mass of the charged fragments to be ≈1500 Da leads to an estimate of $\delta = 0.015$ Da for the precursor ion mass error tolerance and $\epsilon = 0.15$ Da for the product-ion mass grid spacing. The downloaded MS/MS datasets were sorted into three data groups (DGs), each from a separate bacterial strain. DG-1 contains 12 MS/MS datasets (175 569 spectra) from *Escherichia coli K-12*; DG-2, 9 MS/MS datasets (141 332 spectra) from *Mycobacterium tuberculosis H37Rv*; and DG-3, 8 MS/MS datasets (121 787 spectra) from *Salmonella typhimurium ATCC 14028*. Experimental details concerning sample preparations for DGs 1–3 can be found in (Mottaz-Brewer *et al.*, 2008; Schrimpe-Rutledge *et al.*, 2012). A summary of the datasets downloaded is provided in Supplementary Table S1.

2.6 Software parameters

To minimize the number of confounding factors, we analyze MS/MS data with the following fixed software/experimental parameters: cysteine residues are modified with iodoacetamide, resulting in the addition of the carbamidomethyl group (57.07 Da); trypsin is used to digest protein mixtures; the maximum number of missed cleavage sites allowed per peptide is 3; and only b- and y-series are used for scoring since they form the largest common set of *m/z* peaks used by all four scoring functions considered. Dataset-specific parameters can be found in the figure captions, each of which provides more information regarding the generation of the figure, such as protein sequence database (target/decoy/random), scoring function(s) and precursor-ion mass error tolerance (δ).

2.7 Protein databases

The three target databases contain non-redundant protein sequences respectively from *Escherichia coli* (4303 protein sequences), *Mycobacterium tuberculosis H37Rv* (11 081 protein sequences) and *Salmonella typhimurium ATCC 14028* (5482 protein sequences), which were downloaded from UniProt http://www.uniprot.org/uniprot/on July 15, 2013. For each MS/MS data group, the retrieval assessment requires, in addition to the target database, a decoy database that mimics the background of the target database. Although a different approach (Alves and Yu, 2015) was developed, we simply use the reversed protein sequences as the decoy. This is to facilitate comparison with other methods, since most of them also use reversed protein sequences as the decoy.

In addition to the decoy databases, a random protein database was employed to test the accuracy of the reported DPVs. With the Robinson-Robinson frequencies (Robinson and Robinson, 1991) as the amino acid occurrence probabilities, the random protein database contains 100 000 randomly generated proteins, each of 350 amino acids long.

3 Results

Although all four scoring functions mentioned in Section 2.4 are extensively tested for the accuracy of significance assignments and retrieval performance, we describe the results for only XCorr, termed RAId_DbS EVD(XCorr), in the main text. The results for the other three scoring functions are similar and are thus relegated to the supplementary information.

3.1 DPV accuracy and GOF

To investigate how the NSRP (i.e. the sample size) affects the accuracy of the reported DPVs, we test the method described in Section 2.2 with the following NSRP: 1000, 10 000, 100 000 and 1 000 000. Using these NSRP to estimate the EVD parameters via Equations 5 and 6, we examine how well the reported DPVs agree with the observed DPVs. Given a score threshold and the EVD parameters, the reported DPV is computed via Equation (7), while the observed DPV is the fraction of spectra having matching peptides with reported DPVs smaller than the threshold, i.e. the (reported) DPV specified. Agreement between the reported DPVs and the observed DPVs is measured by how well the curve traces the $y = x$ line for the entire range of DPVs, from 10^{-6} to 1. The curves in panel A of Figure 1 show that as the NSRP, used to estimate the EVD parameters, increases from 1000 to 100 000, there is a significant improvement in agreement between reported and observed DPVs. However, not much improvement is gained by raising NSRP from 100 000 to 1 000 000. Similar results are found for other scoring functions, see panel A's of Supplementary Figures S1–S3. This investigation also leads us to use NSRP = 100 000 for estimating the EVD parameters for all four scoring functions.

To assess the statistics' robustness against variation in the number of qualified database peptides, we compare the reported DPVs with the observed DPVs when artificially varying δ. Evidently, a large (small) δ yields a large (small) number of qualified database peptides for scoring. The curves in panel B of Figure 1 show that as

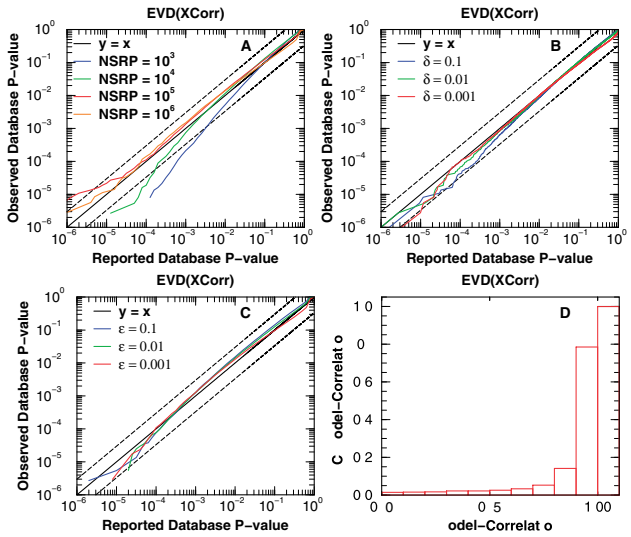

Fig. 1. DPV accuracy and assessment of the EVD model. In panels **A**–**C**, two dashes lines, $x = 3y$ and $x = y/3$, are plotted to show how close/off the measured curves are from the theoretical $y = x$ curve. All spectra in DG-1 (*E. coli*) were queried against the random database (Section 2.7). With δ=0.01 Da, panel (A) displays the observed DPVs versus the reported DPVs as NSRP varies from 10^3, 10^4, 10^5 to 10^6. With NSRP $=10^5$, panel (B) displays the accuracy of the reported DPV for different δs: 0.1 Da, 0.01 Da and 0.001 Da. With NSRP $=10^5$ and δ set to 0.01 Da, panel (C) displays the accuracy of the reported database DPVs under different internal mass spacings (ϵs): 0.1 Da, 0.01 Da and 0.001 Da. In panel (D) (with NSRP$=10^5$, $\delta = 0.01$ Da, and $\epsilon = 0.1$ Da), the cumulative frequency histogram of the model GOF is shown

δ varies, the agreement between reported and observed DPVs remains stable. Similar results are found for other scoring functions; see panel B's of Supplementary Figures S1–S3.

We also investigate the impact of varying ϵ, RAId's internal mass spacing (see Section 2.4). Evidently, varying ϵ changes the score obtained for each qualified peptide, influencing the EVD parameters estimated via Equations (5) and (6), and thus may yield different reported DPVs. However, if the statistics are correct, regardless of the superficial differences, the reported DPVs should still agree with observed DPVs. Panel C of Figure 1 as well as panel C's of Supplementary Figures S1–S3 show that for a fixed precursor-ion mass tolerance of 0.01 Da, strong agreement between the reported DPVs and the observed DPVs is found for the three different ϵ's: 0.1 Da, 0.01 Da and 0.001 Da.

An EVD model is useful for computing database *P*-values only if it describes well the distribution of the sampled maxima as mentioned in Section 2.3. Fortunately, this condition is met by the GOFs computed for the EVD models. Displayed in panel D of Figure 1 (and in panel D's of Supplementary Figs S1–S3) is the normalized cumulative frequency histogram for the model GOF. We find that more than 98% of the EVD models, with parameters obtained using Equations (5) and (6), have average correlation coefficients greater than 0.92. The strong agreement between the reported DPVs and the observed DPVs suggests that the score distributions for all four scoring functions fall in the domain of attraction of the EVD. It is possible that peptide properties such as peptide lengths and precursor-ion charges may impact the accuracy of EVD models, which can be measured by the GOFs. To systematically investigate this effect, we first group spectra according to their precursor ion charge states; within each group, we analyze the spectra using four scoring functions via EVD; spectra within the same group are further separated into subgroups according to the lengths of their best scoring peptides; the statistics of the GOFs and the number of spectra within each subgroups are documented. As shown in Supplementary Tables S2–S5, there is indeed a slight, albeit not significant, deterioration of the GOFs for short peptides when the charge states of the precursor ions are low.

3.2 Sorić PFD and the target-decoy PFD

In Figure 2, we examine, using three different DGs, the agreement between the PFD curves produced using the Sorić formula and using the target decoy approach. While these curves are constructed using RAId_DbS EVD(XCorr), similar results are obtained for the other scoring functions and the corresponding PFD curves can be found in the Supplementary Figures S4–S6. The Sorić PFD (Sorić, 1989) is given by

$$\mathrm{PFD}(\mathrm{DPV} \leq \mathrm{DPV}_c) \approx \frac{n_\sigma \times \mathrm{DPV}_c}{t(\mathrm{DPV} \leq \mathrm{DPV}_c)}, \qquad (9)$$

where DPV_c is the DPV cut-off, n_σ is the total number of MS/MS spectra from a given experiment, $t(\mathrm{DPV} \leq \mathrm{DPV}_c)$ counts the number of target database peptides (out of n_σ spectra) identified with $\mathrm{DPV} \leq \mathrm{DPV}_c$, and the total number of false positives with $\mathrm{DPV} \leq \mathrm{DPV}_c$ is approximated by $n_\sigma \times \mathrm{DPV}_c$. When the target-decoy strategy is used to estimate the PFD, the total number of false positives is estimated by the number of decoy peptides with $\mathrm{DPV} \leq \mathrm{DPV}_c$.

Figure 2 shows that the PFD curves computed using Sorić's formula and the target-decoy approach trace each other well for PFD values between 10^{-3} and 0.3; for PFD values less than 10^{-3}, the two curves slightly disagree. Similar results are obtained for other scoring functions and are shown in Supplementary Figures S4–S6. The deviation between the two curves for small PFD values (less than 10^{-3}) is expected. In this region, a small variation in the number of false positives can influence significantly the computed PFD value. The observed deviation at large PFD values (greater 0.3) is mainly due to the disagreement between the overall number of false positives estimated from the target and decoy databases. In this region, for a given MS/MS spectrum the best scoring qualified peptides from the target and decoy database can have very different DPVs, explaining the possible observed difference in the PFD.

3.3 Peptide retrieval comparison

In the previous section, when using the DPVs as the grading standard, the agreement between the target-decoy PFD and the Sorić PFD is illustrated. However, whether or not these PFD curves coincide with the *ideal* PFD curve (reflecting the ground truth) remains an open question. This issue is partially answered in an earlier publication (Alves and Yu, 2015), within which an ideal PFD curve is constructed by analyzing MS/MS spectra from a known protein mixture and agreement between the ideal PFD curve and the Sorić PFD curve was confirmed. However, given that a sample of known protein mixture is less complex than true biological samples, a more systematic and rigorous study (which is beyond the scope of the current paper) is needed in order to thoroughly examine the relationship between the ideal PFD and the target-decoy PFD. Nevertheless, with the target-decoy approach being well accepted in the community, we find it appropriate to compare retrieval results based on this approach.

In panels A–C of Figure 3, corresponding to DGs 1–3, we display the retrieval PFD curves of XCorr using the original SEQUEST [v. 28] and our EVD implementation. In Supplementary Figure S7,

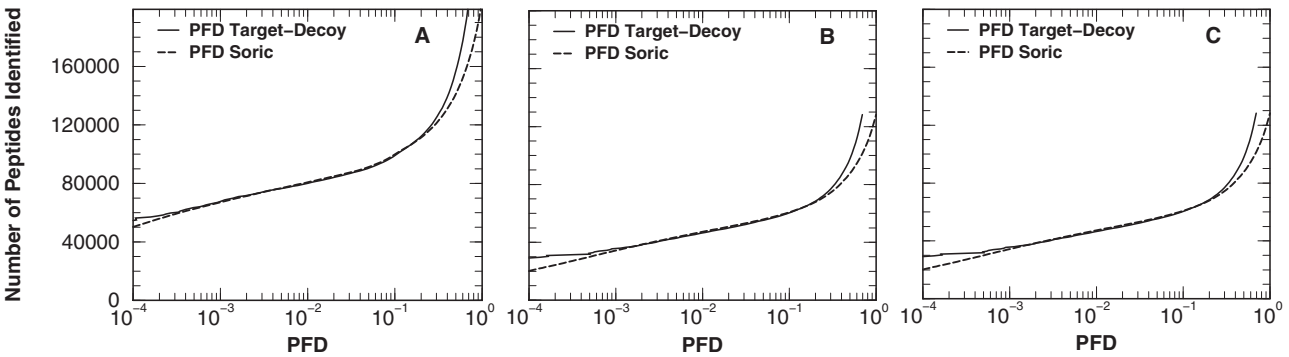

Fig. 2. Agreement between the Sorić's PFD and the target-decoy PFD when peptides are ranked by DPVs. The PFD curves through both approaches are displayed in panels **A**, **B** and **C**, respectively, for DG-1, DG-2 and DG-3; each DG is analyzed using the parameters mentioned in Section 2.6 and with the following additional parameters: $\delta = 0.015$ Da, $\epsilon = 0.15$ Da, target and decoy databases as described in Section 2.7

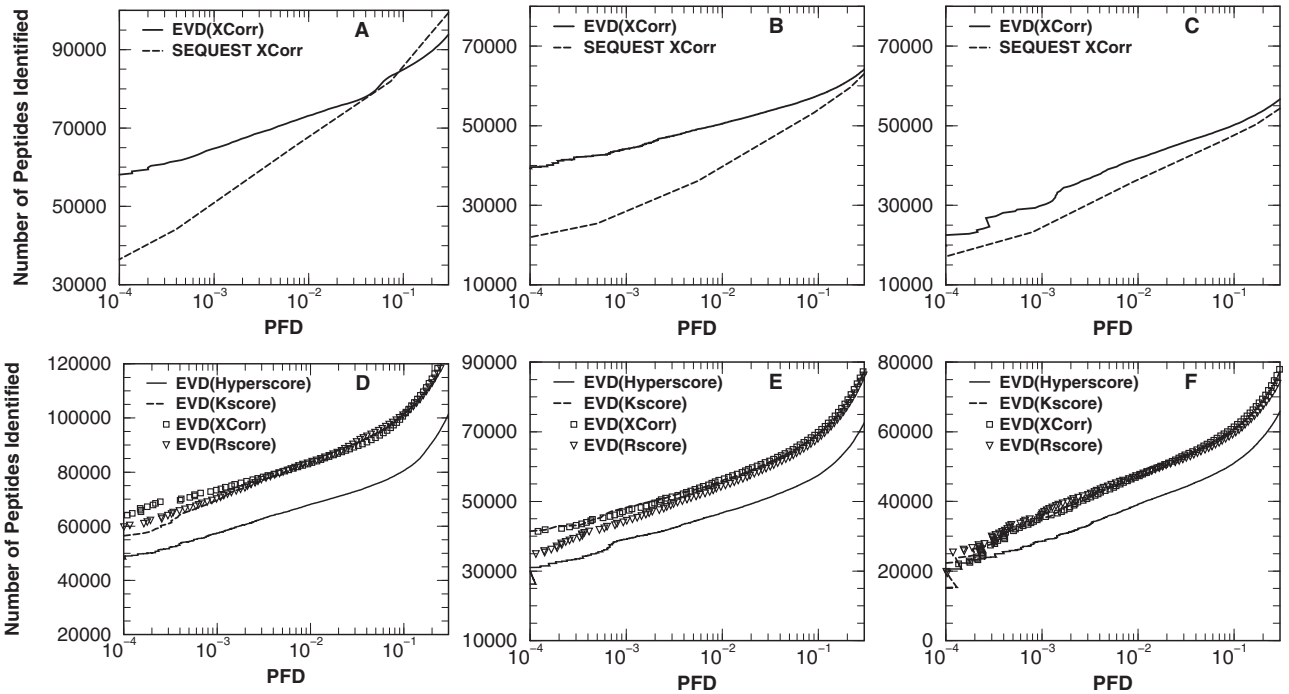

Fig. 3. Peptide retrieval comparison via target-decoy approach. DGs 1-3, analyzed using the same parameters as mentioned in the caption of Figure 2, yield the retrieval curves in panels **A**–**C** (and in panels **D**–**F**) respectively. Panels A–C display the retrieval PFD curves when peptides are ranked by the per spectrum EVD statistics and by the native SEQUEST program, both of which use XCorr as the scoring function. Panels D–F display, for various scoring functions, the retrieval PFD curves when peptides are ranked by the EVD statistics

we show similar comparison for Hyperscore using the original X!Tandem [v. 2013.06.15] and our EVD implementation as well as for Rscore using the CLT and EVD statistics through RAId_DbS. For Kscore, however, we are unable to obtain reasonable results through X!Tandem [v. 2013.06.15]. Hence we do not include this retrieval result for comparison. Also, when performing retrieval comparison between RAId_DbS EVD(XCorr) and SEQUEST XCorr (panels A-C) it is necessary to turn off in RAId_DbS the parameter 'isotopic error correction', which corrects for erroneous monoisotopic precursor-ion mass assignment, because SEQUEST does not have this option.

As shown in panels D–F (corresponding to DGs 1–3) of Figure 3, comparable retrieval results are observed for different scoring functions under the EVD strategy except for Hyperscore (whose superficially worse performance might be attributed to X!Tandem's aggressive filtering). The retrieval comparison between the EVD statistics (RAId_DbS) and the APPS strategy (RAId_aPS) is provided in the Supplementary Figure S8. It is evident that these two strategies perform comparably under various scoring functions. This is expected because both strategies have the same spectral filtering and scoring procedures, and both have accurate statistical significances assigned to identified peptides. One may notice that the PFD curves for RAId_DbS EVD(XCorr) in panels A, B and C of Figure 3 are lower than those in panels D–F. This is because the 'isotopic error correction' option mentioned earlier is turned off in order to fairly compare with the native SEQUEST program. We also observe from panels A–C of Figure 3 that, with the isotopic error correction turned off, RAId_DbS EVD(XCorr) still has a better retrieval than SEQUEST XCorr, albeit under the choice that only b- and y-series are used for scoring (see Section 2.6).

3.4 Perspective on computational speed

To gain the perspective on computational speed, we query a 35 MB database with 15 000 MS/MS spectra using one quad-core processor of 2.8 GHz under the same search parameters described in Section 2.6 along with $\delta = 0.01$ Da, $\epsilon = 0.1$ Da and Rscore as the scoring function. For CLT-based method, it takes 7 min; for EVD-based method, it takes 14 min; for APPS-based method, it takes 5 min. The EVD implementation, although is slower than both the CLT and the APPS-based implementations, offers flexibility in terms of scoring function choices without adding much computational cost.

4 Discussion

By definition, a decoy database contains no *true* peptide that produces any of the observed spectra. The primary function of a decoy database is to mimic the 'real' background one encounters when searching the target database. Therefore, if the decoy database is properly constructed, the distance between score distributions, resulting respectively from searching the target and the decoy databases, should be small. However, to firmly establish that a decoy database is a good choice requires perhaps more work than the data analysis itself. Of course, this work can be waived if accurate statistics can be obtained from the target database alone. For our EVD implementation, we have tested the effectiveness of error control using both random databases (for type-I) and retrieval PFD curves (for type-II). We find that the reported DPVs agree closely with the observed DPVs (see Fig. 1, Supplementary Figs S1–S3) and that the Sorić PFDs agree well with the target-decoy PFDs. The former shows that our assigned statistical significance coincides with its definition, while the latter indicates that the common choice of using reverse sequences as the decoy is reasonable.

In Section 2.3, the correlation between the set of maximum scores and the rank percentile (after double logarithm transformation) is introduced as the GOF of the EVD models, and a cutoff value of 0.92 is selected based on the work of Kinnison (1989). Evidently, raising the threshold to a higher value reduces the number of spectra for analyses, hence potentially reducing the number of peptides identified. That is, it increases the number of false negatives (leading to type-II error). On the other hand, lowering the threshold too much will include spectra for which the EVD models break down, hence potentially increasing the number of false identifications (leading to type-I error). Our investigation indicates that the threshold value 0.92 seems to yield a good balance between type-I and type-II errors.

The search options of RAId_DbS allow for the inclusion of post-translational modifications (PTMs) and/or single amino acid polymorphisms (SAPs). The inclusion of PTMs/SAPs or isotopic labeling enlarge the search space, hence generally reduces sensitivity as expected (Alves *et al.*, 2008b). However, the assigned statistical significances, when accurate, should naturally reflect this effect. For example, an identified database peptide with a low *P*-value will have a worse *E*-value when the search space is large as compared to the case of a smaller search space.

A question arises when one allows more than one candidate peptide per spectrum to accommodate cofragmentation of multiple precursor ions with proximate masses. The natural extension of the EVD for considering candidates other than just the best is the order statistics. However, it makes better sense to use DPVs for lower-ranking peptides per spectrum; see the discussion in (Alves and Yu, 2015). The DPV for each lower-ranking peptide per spectrum is simply obtained thorough applying the EVD parameters learned for the spectrum and that peptide's score in Equation (7).

Currently, in RAId_DbS's EVD implementation, for each spectrum, 100 000 random peptides are scored and randomly assigned to 100 bins to estimate the EVD parameters regardless of δs. In principle, there might exist better NSRP/bin combinations that can produce EVD parameters more rapidly or accurately. It is our plan to explore other NSRP/bin combinations in the near future.

Another interesting point to mention is the finite size effect associated with the EVD, an asymptotic distribution that occurs when the number of samples drawn from the population approaches infinity. Because we always draw only a finite number of peptides to score, finite size effects, such as those observed in sequence comparison statistics (Yu and Hwa, 2001), are indeed expected. However, because we are not referring the EVD parameters from a theory but are fitting them with a finite number of samples, the finite size effect is automatically taken into account. In addition, the k factor introduced in Equation (7) allows each spectrum's EVD parameters to be learned with the same NSRP while at the same time correcting for uneven sample sizes resulting from searching the target database.

As expected, we find that using the raw score to prioritize peptides yields worse retrieval than using the EVD statistics; see Supplementary Figure S9. In addition, we also find that when statistics are accurate, almost all scoring functions have highly similar retrieval performances, except for Hyperscore (whose superficially worse performance might be attributed to X!Tandem's aggressive filtering). The similar performances found may result from the fact that all scoring functions considered use the same *m/z* evidence (b- and y-series peaks). However, if this were the case, it indicates that no significant retrieval improvement can be expected from a new scoring function that simply tweaks or reweighs the contributions of commonly used evidence peaks; rather, a true advance in scoring is more likely to come from new scoring schemes that better capture the physical/chemical mechanism of fragmentations. This type of new scoring scheme may not be a simple sum of contributions. Nevertheless, if its score distribution falls in the domain of attraction of the EVD, the method elaborated in this manuscript can still be applied to provide accurate statistical significance assignments, hence bringing out the full strength of the new scheme.

Acknowledgements

We thank the administrative group of the Biowulf Clusters (of the National Institutes of Health), where all the computational tasks were carried out. We also thank Stephen Altschul for a critical reading of the manuscript.

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. (2015) Mass spectrometry-based protein identification with accurate statistical significance assignment. *Bioinformatics*, **31**, 699–706.

Alves,G. *et al.* (2007a) Calibrating E-values for MS2 database search methods. *Biol. Direct*, **2**, 26.

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

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

Alves,G. *et al.* (2008b) RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. *BMC Genomics*, **9**, 505.

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

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. p

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.

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

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.

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

Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York, USA.

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.

Higdon,R. *et al.* (2008) A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics*, **24**, 1225–1228.

Kim,S. *et al.* (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, **7**, 3354–3363.

Kinnison,R. (1989) Correlation coefficient goodness-of-fit test for the extreme-value distribution. *Am. Stat.*, **43**, 98–100.

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

Kotz,S. and Nadarajah,S. (2000) *Extreme Value Distributions*. Imperial College Press, London, UK.

MacLean,B. *et al.* (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics*, **22**, 2830–2832.

Mottaz-Brewer,H.M. *et al.* (2008) Optimization of proteomic sample preparation procedures for comprehensive protein characterization of pathogenic systems. *J. Biomol. Tech.*, **19**, 285–295.

Olsen,R. *et al.* (1999) Rapid assessment of extremal statistics for gapped local alignment. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 211–222.

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.

Schrimpe-Rutledge,A.C. *et al.* (2012) Comparative omics-driven genome annotation refinement: application across Yersiniae. *PLoS ONE*, **7**,

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

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.

Yu,Y.K. and Hwa,T. (2001) Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *J. Comput. Biol.*, **8**, 249–282.

Yu,Y.K. *et al.* (2002). Statistical significance and extremal ensemble of gapped local hybrid alignment. In: Lssig,M. and Valleriani,A. (eds.) *Biological Evolution and Statistical Physics, Volume 585 of Lecture Notes in Physics*. Springer, Berlin, Heidelberg, pp. 3–21.

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