

## RESEARCH ARTICLE

# Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines

Andrew R. Jones<sup>1</sup>, Jennifer A. Siepen<sup>2</sup>, Simon J. Hubbard<sup>2</sup> and Norman W. Paton<sup>3</sup>

<sup>1</sup> Department of Preclinical Veterinary Science, Faculty of Veterinary Science, University of Liverpool, Liverpool, UK

<sup>2</sup> Faculty of Life Sciences, University of Manchester, Manchester, UK

<sup>3</sup> School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, Manchester, UK

LC-MS experiments can generate large quantities of data, for which a variety of database search engines are available to make peptide and protein identifications. Decoy databases are becoming widely used to place statistical confidence in result sets, allowing the false discovery rate (FDR) to be estimated. Different search engines produce different identification sets so employing more than one search engine could result in an increased number of peptides (and proteins) being identified, if an appropriate mechanism for combining data can be defined. We have developed a search engine independent score, based on FDR, which allows peptide identifications from different search engines to be combined, called the *FDR Score*. The results demonstrate that the observed FDR is significantly different when analysing the set of identifications made by all three search engines, by each pair of search engines or by a single search engine. Our algorithm assigns identifications to groups according to the set of search engines that have made the identification, and re-assigns the score (*combined FDR Score*). The *combined FDR Score* can differentiate between correct and incorrect peptide identifications with high accuracy, allowing on average 35% more peptide identifications to be made at a fixed FDR than using a single search engine.

Received: May 30, 2008

Revised: July 28, 2008

Accepted: September 26, 2008

**Keywords:**

Decoy database / False discovery rate / MASCOT / OMSSA / X!Tandem

## 1 Introduction

High-throughput proteome analyses are now commonplace, allowing researchers to assess the proteins present in a sample and, by utilising new technologies, to quantify protein abundance on a large scale. The high-throughput methods

can generate large volumes of data for which manual verification of peptide and protein identification is not feasible, so automated methods are required for assigning significance. It is not yet clear how best to determine which peptide or protein identifications are correct, and how to optimise identification pipelines such that false discovery is kept sufficiently low, while maximising the number of proteins that can be identified correctly [1].

There are a number of software applications, both commercial and open-source, for identifying peptides from mass spectra [2–6]. Each application produces a set of non-standard, algorithm-dependent measures of the quality of peptide and protein identifications. Several search engines produce an expectation value (*e-value*), which relates to the likelihood of a peptide identification having been made incorrectly by chance. However, while *e-values* provide a good

---

**Correspondence:** Dr. Andrew R. Jones, Department of Preclinical Veterinary Science, Faculty of Veterinary Science, University of Liverpool, Liverpool, L69 7ZJ, UK

**E-mail:** andrew.jones@liv.ac.uk

**Fax:** +44-151-7944243

**Abbreviations:** **ABRF**, association of biomolecular research facilities; **AFS**, average FDR score; **e-value**, expectation value; **FDR**, false discovery rate

measure for ranking the quality of identifications within a single experiment, it has been demonstrated that  $e$ -values are not comparable between different packages [7]. Without a search engine independent measure, it is difficult to optimise identification pipelines, and researchers are likely to set stringent thresholds (often with a limited understanding of the underlying statistical model), to ensure that the rate of false positives is acceptably low. Algorithms have been developed that can estimate the probability that a peptide (or protein) identification is correct such as those implemented in the Trans-Proteomic Pipeline (TPP) [8, 9]. An alternative approach for validating identifications involves the use of a decoy database, for example comprising reversed or randomised protein sequences. The number of identifications made from the decoy database, compared with the total number from the target database, can be used to estimate the false discovery rate (FDR) for a given threshold [10].

It has been demonstrated that different software packages do not produce the same peptide identifications for large sets of spectra [7], particularly for peptides scoring close to the threshold for acceptance or rejection. This means that it should be possible to extract more identifications from a set of spectra by employing multiple search engines, if there is a framework for combining results. In this work, we have developed a search engine independent score assigned to each peptide-spectrum match based on observed FDRs. The score is a further extension of a  $q$ -value, which has recently been demonstrated for use in proteomics [11].  $q$ -values can be assigned to each peptide-spectrum match and they can be used to set thresholds that guarantee the estimated FDR is less than a given value. For instance, accepting all identifications with  $q$ -value  $<0.05$  results in FDR  $<0.05$  by definition. However,  $q$ -values have several properties that mean they have limited use for further calculations and cannot be used reliably to combine identification sets from different search engines (see Section 2 for algorithm used to calculate  $q$ -values and discussion of limitations). We have therefore adapted the calculation of  $q$ -values to create a metric called the *FDR Score*. The assignment of an *FDR Score* to each identification allows the identification sets produced by different search engines to be compared using a single metric and combined. We have integrated the results from MASCOT [6], and two open source applications OMSSA [4] and X!Tandem [3], and the process can be followed relatively simply by any laboratory using the search engines that are available to them.

Our results demonstrate that the FDR is far higher for peptides identified by only a single search engine. In contrast, if a peptide has been identified by all three search engines, very few false positives are observed. As such, we have developed an algorithm for calculating a second metric using a similar basis to the *FDR Score*, called a *combined FDR Score*, which re-assesses the rates of false discovery for identifications made by only one search engine, by each distinct pair of search engines, or by all three search engines after data have been combined. The *combined FDR Score* appears

to be a highly effective discriminator between a correct and incorrect peptide identification. For a fixed FDR of 1% FDR, on average gains of 35% total peptide identifications are possible over the best individual search engine.

## 2 Materials and methods

### 2.1 Datasets searched

Proteome datasets from the public data repository PeptideAtlas [12], shown in Table 1, were downloaded in mzXML format and converted to MASCOT generic format (.mgf) using the RAMP parser (<http://tools.proteomecenter.org/TPP.php>). The majority of the datasets were from yeast (*Saccharomyces cerevisiae*), selected to cover a range of contributing laboratories, experimental approaches and dataset sizes. The datasets were searched using OMSSA (version 2.1.1), MASCOT (version 2.0) and X!Tandem (version 07-07-01) using a parameter set matching the original search parameters as closely as possible: parent ion tolerance 2 Da, fragment ion tolerance 0.8 Da, default instrument setting and average mass setting. Datasets contained different types of variable and fixed modification including: isotope coded affinity tag (ICAT), carbamidomethyl of cysteine and oxidation of methionine. The system was also tested using searches of human and mouse data from PeptideAtlas, and validated by datasets released by Association of Biomolecular Research Facilities (ABRF) (<http://www.abrf.org>). ABRF have generated a standard protein set containing 49 known proteins, which allows the actual FDR to be calculated and compared to the estimates made from the decoy search approach. ABRF datasets were searched with parent ion tolerance 1.2 Da, fragment ion tolerance 0.6 Da and monoisotopic mass, to reflect more closely the parameters used by the laboratories that produced the data.

The following databases were used: Yeast SGD ORFs (<http://www.yeastgenome.org/>), IPI human and IPI mouse (<http://www.ebi.ac.uk/IPI/>). ABRF datasets were searched against a database constructed specifically by ABRF containing a combination of UniProt human, Swiss-Prot human and additional contaminant proteins expected to be in the sample. Decoy databases were created by reversing all the protein sequences, and adding the set of reversed sequences to the standard sequences in the same file, for each of the databases individually [10]. By searching the forward and reverse database simultaneously, standard and decoy sequences can compete equally to be the highest ranking identification for each spectrum, adequately representing how normal false positive identifications are made. There is still considerable discussion in the field about how best to construct decoy databases to model incorrect identifications, since it has been recognised that while reverse databases adequately model random false positives, they may not perfectly model false positives caused by closely homologous

**Table 1.** Datasets from PeptideAtlas re-searched in the analysis

Experiment accession	Num. spectra	Study authors	Species	Variable modifications	Fixed modifications
PAe000066	126 956	Omenn <i>et al.</i>	Human	Oxidation (M)	None
PAe000077	26 684	Raught <i>et al.</i>	Yeast	Oxidation (M), ICAT new (heavy)	ICAT new (light)
PAe000093	97 070	Flory <i>et al.</i>	Yeast	Oxidation (M), ICAT old (heavy)	ICAT old (light)
PAe000098	49 443	Omenn <i>et al.</i>	Human	Oxidation (M)	Carbamidomethyl (C)
PAe000138	117 110	Marelli <i>et al.</i>	Yeast	Oxidation (M), ICAT new (heavy)	ICAT new (light)
PAe000146	183 210	Raught <i>et al.</i>	Yeast	Oxidation (M), ICAT new (heavy)	ICAT new (light)
PAe000157	27 845	Marelli <i>et al.</i>	Yeast	Oxidation (M), ICAT new (heavy)	ICAT new (light)
PAe000158	133 608	Breci <i>et al.</i>	Yeast	Oxidation (M), ICAT new (heavy)	ICAT new (light)
PAe000160	61 625	Breci <i>et al.</i>	Yeast	Oxidation (M)	Carbamidomethyl (C)
PAe000162	169 028	Breci <i>et al.</i>	Yeast	Oxidation (M)	Carbamidomethyl (C)
PAe000165	183 838	Breci <i>et al.</i>	Yeast	Oxidation (M)	Carbamidomethyl (C)
PAe000166	52 632	Breci <i>et al.</i>	Yeast	Oxidation (M)	Carbamidomethyl (C)
PAe000167	156 016	Aebersold and Kregenow	Yeast	Oxidation (M)	None
PAe000292	42 919	Rong Wang	Mouse	Oxidation (M)	None

ICAT new corresponds with the newer form of the (227.13 Da · mass); ICAT old is the original form (442.2 Da · mass).

sequences [13, 14]. As consensus is reached in this area, new methods for creating decoy databases can be implemented in conjunction with our algorithm.

In the result set, only the top ranking peptide identification for each spectrum is included. We compared the result sets containing only the top ranking identification with result sets containing the top three ranking identifications, and discovered that there is no significant increase in the number of peptides identified for a fixed FDR (data not shown), but causes a significant increase in the computation time.

## 2.2 Calculating false discovery rate for peptide identifications

A method has been published by Elias and Gygi [10] for calculating FDR using decoy databases. Assuming a search is made against a database constructed from equal-sized target and decoy databases, the number of false positive peptide identifications is calculated for a given threshold by doubling the number of hits to the decoy database, following the logic that for every hit to a decoy sequence, there will be a 'silent' incorrect hit in the standard database (*FDR Method 1*). However, this measure of FDR can be misleading, since it does not reflect the FDR within the targets, which ultimately is the measure that researchers are interested in. It is trivial to remove the decoys from a set of peptide identifications (since they are flagged with a particular identifier). In the remaining set of targets, it can be assumed on average that the number of targets which are false positives is approximately equal to the number of decoy hits that have been removed, as discussed by Käll *et al.* (*FDR Method 2* [11]). This method creates a lower estimate of FDR than Method 1 as shown below. *FDR Method 2* is used in our algorithms.

### *FDR Method 1 (Elias and Gygi [10])*

False positives (FP) =  $2 \times$  decoy hits

True positives (TP) = All targets above threshold – FP

False discovery rate (FDR) = FP/FP + TP

Example for 1000 identifications above threshold, 20 decoy identifications.

FP = 40

TP = 960 (1000–40)

FDR = 40/1000 = 4%

### *FDR Method 2 (Käll *et al.* [11])*

All targets = Target hits only (above threshold)

False positives (FP) = Decoy hits

True positives (TP) = All targets above threshold – FP

False discovery rate (FDR) = FP/FP + TP

Example for 1000 identifications above threshold, 20 decoy identifications:

All targets = 980 (1000–20)

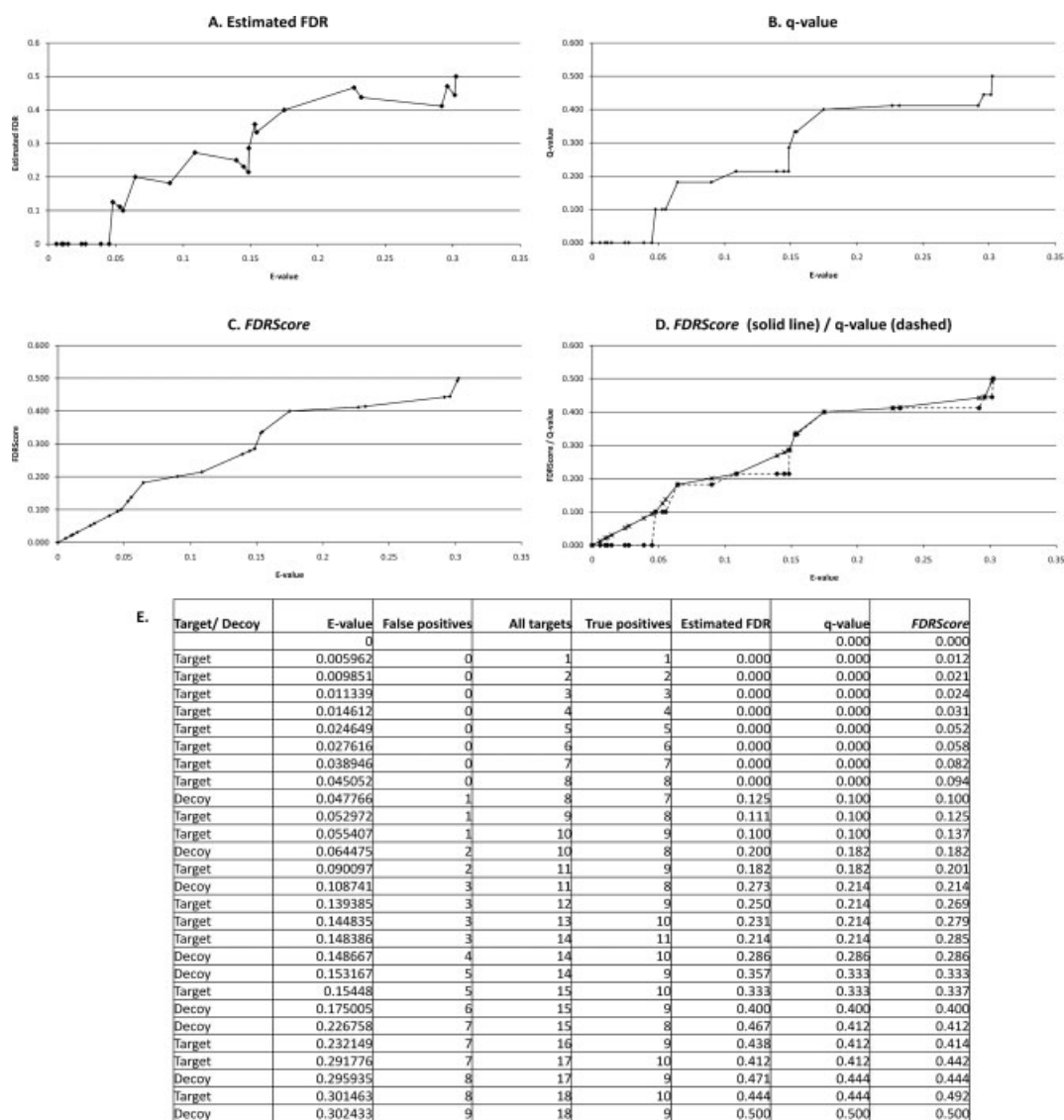
FP = 20

TP = 960 (980–20)

FDR = 2.04% (20/980)

## 2.3 Estimating FDR, *q*-values and *FDR Scores* for each identification

Three related measures are calculated for each peptide-spectrum match: the estimated FDR, the *q*-value and *FDR Score* (*Algorithm 1* and Fig. 1) as follows. In step 1, peptide-spectrum matches are ordered according to some measure of identification quality. In our implementation, *e*-values are used to rank identifications since MASCOT, OMSSA and X!Tandem all produce an *e*-value but other score types, for example MASCOT ion score, a SEQUEST XCorr or a more



**Figure 1.** A hypothetical dataset is shown for a series of peptide-spectrum matches ranked by increase in  $e$ -value. The four graphs display the relationship between the  $e$ -value and (A) FDR, (B)  $q$ -value, (C)  $FDR\ Score$  and (D)  $FDR\ Score$  overlaid on the  $q$ -value (to demonstrate their relationship and the method by which  $FDR\ Score$  is calculated). The table (E) shows the individual values used in the calculation of FDR,  $q$ -value and  $FDR\ Score$ .

complicated consensus score, could also be used. In step 2, for each score associated with a peptide-spectrum match, the cumulative FDR is estimated (from a decoy database search) that would result if that exact score was set as the threshold for acceptance or rejection of identifications. In step 3, a  $q$ -value is assigned to each match as the minimum FDR at which the identification could be made; in effect, the weakest threshold that could be set to include an identification without increasing the number of false positives. The  $q$ -values therefore follow a stepwise distribution, typically with each step-point caused by a decoy identification increasing the FDR (Fig. 1).

A  $q$ -value has a property that it can be used to set a threshold that guarantees the reported FDR is less than a given value

but  $q$ -values are less useful for further calculations for the following reasons.  $q$ -values follow a stepwise distribution where all target identifications with no intervening decoy identifications share the same  $q$ -value, so relative information about the quality of an identification on a local scale is lost. Furthermore, within a set of identifications that share the same  $q$ -value, the strongest identification will always be a decoy such that  $q$ -values are biased against decoys.

In step 4 of *Algorithm 1*, a new score called the  $FDR\ Score$ , is created by calculating the intercept and gradient of the line between the step points at which the  $q$ -value changes. The intercept and gradient are then used to convert the  $e$ -value to an estimate of FDR, for all identifications in between step

points (which share the same  $q$ -value). Across an entire identification set, the estimated FDR,  $q$ -value and  $FDR$  Score are roughly similar, but on a localised scale the  $FDR$  Score is more useful for further calculation. The  $FDR$  Score maintains the ordering of identification quality (which is lost in both estimated FDR and  $q$ -value), and each  $FDR$  Score assigned to a target identification is likely to be closer to the actual FDR

associated with a peptide-spectrum match than either the estimated FDR or the  $q$ -value. Finally, all target identifications scoring higher than the best decoy hit have an estimated FDR = 0 and  $q$ -value = 0, although their likelihood of being a false positive is not zero. The regression method used to calculate  $FDR$  Scores includes the origin (0,0) for the first calculation and hence no identification has an  $FDR$  Score = 0.

#### Algorithm 1

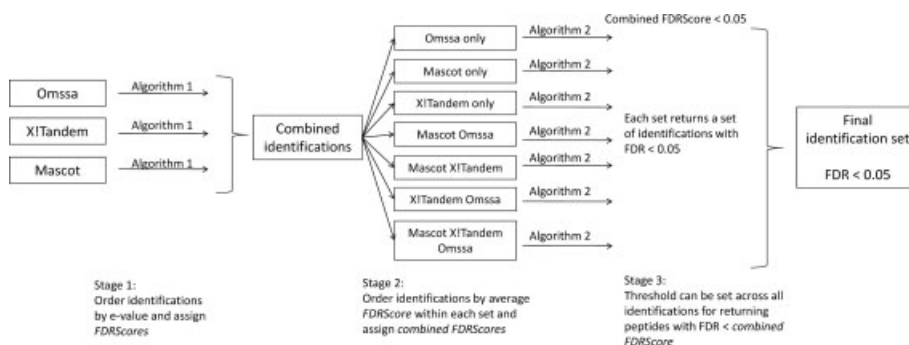
For each set of peptide identifications made by one search engine:

- 1) Order identifications according to  $e$ -value.
- 2) Traverse identifications from lowest  $e$ -value to highest:
  - a. calculate and store the estimated FDR ( $FDR_{est}$ ) for each identification according to  $FDR$  Method 2.
- 3) Traverse identifications from highest  $e$ -value to lowest, storing the lowest estimated FDR ( $FDR_{min}$ ) that has been observed so far.
  - a. For each identification, retrieve  $FDR_{est}$ :
    - i. If  $FDR_{est} > FDR_{min}$ ,  $q$ -value =  $FDR_{min}$ .
    - ii. Else,  $q$ -value =  $FDR_{est}$  and  $FDR_{min} = FDR_{est}$ .
- 4) Traverse the set of identifications (each member denoted  $i$ ) from lowest  $q$ -value to highest, identifying a set of step points, where the  $q$ -value of the identification changes ( $q$ -value $_i > q$ -value $_{i-1}$ ).
  - a. At a step-point, calculate the intercept  $c$  and gradient  $g$  between previous step-point $_{prev}$  ( $e$ -value $_{prev}$ ,  $q$ -value $_{prev}$ ) and current step-point $_{curr}$  ( $e$ -value $_{curr}$ ,  $q$ -value $_{curr}$ ).
  - b. For each identification with  $e$ -value,  $e_i$  between step-point $_{prev}$  and step-point $_{curr}$ :
    - i. Assign  $FDR$  Score =  $e_i * g + c$ .

## 2.4 Recalculating $FDR$ Scores to reflect search engine agreement

For each peptide-spectrum identification, the  $FDR$  Score approximates the discovery rate of false positives that would be observed if a particular threshold was used for an individual search engine. However, we observe that in the set of peptide-spectrum matches made by all three search engines, there is a far lower actual FDR than the general distribution. Indeed in certain datasets, decoy identifications are never observed in the set of identifications made by all three search engines, even if the individual scores from each search engine are weak. Conversely, within the set of peptide-spectrum identifications made by only a single search engine, the FDR is high, even for identifications with strong scores from the source search engine.

In order to quantify the effect of search engine agreement, all peptide-spectrum identifications are divided into seven sets according to which search engines have identified them (Fig. 2): (i) Tandem only, (ii) MASCOT only, (iii) OMSSA only, (iv) OMSSA and Tandem, (v) MASCOT and Tandem, (vi) OMSSA and MASCOT and (vii) MASCOT, OMSSA and Tandem. *Algorithm 1* is adapted to re-assign  $FDR$  Scores calculated within each of the seven distinct sets in *Algorithm 2*. Instead of identifications being ordered by  $e$ -value, they are ordered by the  $FDR$  Scores calculated in *Algorithm 1*. In sets 4–6, all peptide-spectrum matches have an  $FDR$  Score assigned from each of the two search engines, and in set 7 the peptide-spectrum matches have three  $FDR$  Scores, one assigned from each search engine. As such, an *average FDR Score* (AFS) is assigned to each identification (*Algorithm 2* step 1). For the AFS, the geometric mean



**Figure 2.** A flow chart of the stages in the calculation of  $FDR$  Scores for each individual search engine, and *combined FDR Scores* across search engines. The *combined FDR Score* can be used to set a single threshold to return identifications from each distinct set with estimated FDR approximately equal to the threshold.

is used (calculated as the  $n$ th root of the product of  $n$  numbers) as we found empirically that a geometric mean is a better differentiator between correct and incorrect identifications than an arithmetic mean, since an arithmetic mean can mask the contribution of low *FDR Scores* (data not shown).

*Algorithm 2* follows the same stages as *Algorithm 1* for calculating a  $q$ -value and a new *FDR Score*, called the *combined FDR Score* (to differentiate it from the value calculated in the first stage), for each peptide-spectrum match.

---

#### Algorithm 2

For each set of peptide identifications made by a particular combination of search engines:

- 1) Calculate AFS as:
    - a. AFS = The *FDR Score* from the search engine (se) making the identification (sets 1–3).
    - b. AFS = square root ( $FDR\ Score_{se1} * FDR\ Score_{se2}$ ) (sets 4–6).
    - c. AFS = cube root ( $FDR\ Score_{se1} * FDR\ Score_{se2} * FDR\ Score_{se3}$ ) (set 7).
  - 2) Order identifications according to AFS within that set.
  - 3) Traverse identifications from lowest AFS to highest:
    - a. Calculate the estimated FDR (*FDR<sub>est</sub>*) for each identification according to *FDR Method 2*.
  - 4) Traverse identifications from highest AFS to lowest, storing the lowest estimated FDR (*FDR<sub>min</sub>*) that has been observed so far:
    - a. For each identification, retrieve *FDR<sub>est</sub>*:
      - i. If  $FDR_{est} > FDR_{min}$ ,  $q\text{-value} = FDR_{min}$ .
      - ii. Else,  $q\text{-value} = FDR_{est}$  and  $FDR_{min} = FDR_{est}$ .
  - 5) Traverse the set of identifications (each member denoted  $i$ ) from lowest  $q$ -value to highest, identifying a set of step-points, where the  $q$ -value of the identification changes ( $q\text{-value}_i > q\text{-value}_{i-1}$ ).
    - a. At a step-point, calculate intercept  $c$  and gradient  $g$  between previous step-point<sub>prev</sub> ( $AFS_{prev}$ ,  $q\text{-value}_{prev}$ ) and current step-point<sub>curr</sub> ( $AFS_{curr}$ ,  $q\text{-value}_{curr}$ ).
    - b. For each identification with AFS  $a_i$  between step-point<sub>prev</sub> and step-point<sub>curr</sub>:
      - i. Assign *combined FDR Score* =  $a_i * g + c$ .
- 

*Algorithm 2* is performed independently for each of the seven sets of identifications and is demonstrated in Fig. 3 for an example experiment from PeptideAtlas. It can be observed that the FDR of peptide-spectrum matches made by only a single search engine (sets 1–3) for low AFSs is far higher than those made by a pair of search engines. For example, in set 2 (identifications made by MASCOT only) with *FDR Score* < 0.05 (high quality identifications according to MASCOT), the actual FDR is around 0.2, demonstrating that many of these identifications are false positives. In contrast, in the set of identifications where MASCOT and Tandem agree (set 5), for AFS < 0.05 the actual FDR is close to zero.

In the set of identifications made by all three search engines (set 7), the estimated FDR is low, and false positives are rarely observed. In certain datasets, decoy hits are not observed at any score threshold for identifications made by all three search engines. To correct for the size of the result set, an artificial decoy hit is added at the end of each data series, such that no identification has a *combined FDR Score* = 0.

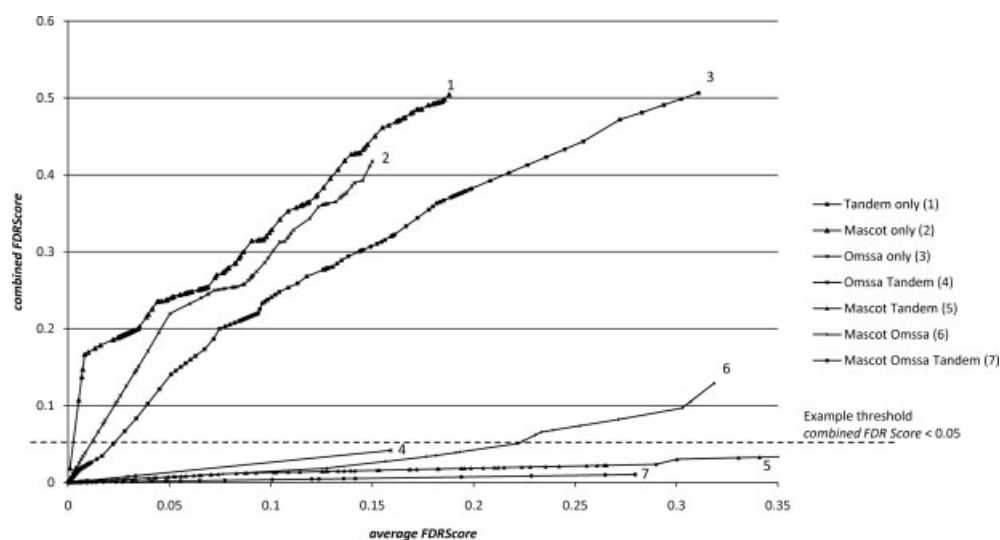
The *combined FDR Score* has the property that it can be used to set a threshold  $x$ , returning a set of identifications with  $FDR = x$  (in practice almost always <  $x$ ). The seven identification sets are distinct, thus a final set of peptide identifications is made by accepting all identifications (from

any set) with say *combined FDR Score* < 0.05. Each of the seven sets returns a certain number of identifications with no more than 5% FDR. In practice sets 1–3 (single search engine only) may return few, if any, identifications. This is demonstrated by the dashed line on Fig. 3. All identifications with a *combined FDR Score* below 0.05 would pass the threshold. The advantage of this approach is that it extracts the maximum number of peptide-spectrum matches that have the profile of being correct while excluding those that have the profile of being incorrect.

## 3 Results

### 3.1 Improvement in peptide discovery

The method outlined above was used to combine results and weight the contribution of different search engines. The number of peptide-spectrum matches made by each search engine with  $FDR < 0.01$  was calculated, and compared with the number of identifications using the *combined FDR Score* to set a threshold for data combined from the three search engines. Table 2 displays the percentage improvement using *combined FDR Scores* over the best individual search engine (defined as the search engine that returns the highest num-



**Figure 3.** Relationship between AFS and *combined FDR Score* from a single experiment within dataset PeptideAtlas PA162 for peptide-spectrum matches made by the seven sets (distinct combinations) of search engines. Each data point represents the calculated AFS *versus combined FDR Score* for a single identification (peptide-spectrum match).

**Table 2.** The maximum number of peptide-spectrum matches made by the combined scoring methods at *combined FDR Score* < 0.01 compared with the best performing single search engine

ID	Threshold (%)	Single search engine idents	Maximum search engine	Combined search idents	Percent improvement
PAe000066	1	5 405	OMSSA	5 794	7.2
PAe000077	1	2 636	OMSSA	3 016	14.4
PAe000093	1	5 924	OMSSA	6 914	16.7
PAe000098	1	184	MASCOT	271	47.3
PAe000138	1	1 915	OMSSA	2 628	37.2
PAe000146	1	18 710	OMSSA	20 776	11.0
PAe000157	1	648	OMSSA	866	33.6
PAe000158	1	11 503	OMSSA	17 091	48.6
PAe000160	1	7 687	Tandem	11 987	55.9
PAe000162	1	1 638	OMSSA	2 757	68.3
PAe000165	1	22 151	OMSSA	27 904	26.0
PAe000166	1	7 091	Tandem	9 007	27.0
PAe000167	1	1 194	Tandem	1 939	62.4
PAe000292	1	570	OMSSA	822	44.2
				Mean improvement	35.7

ber of identifications at a particular FDR threshold) in columns 2 and 3. On average the combined scoring method identifies 35% more peptides than the best individual search engine at FDR < 0.01. There is quite a large difference in the percentage gain between the lowest PA66 (7%) and the highest PA162 (68%). It is interesting to note that in the experiments where only modest gains in the number of peptide-spectrum matches are made (PA66, PA77, PA93 and PA146) that X!Tandem performs poorly, identifying only a fraction of the number of peptides than by OMSSA and MASCOT (Table 3). There is no obvious explanation for the performance of X!Tandem in these experiments since the

data files are being searched with parameter sets close to those specified in the PeptideAtlas repository, although it is possible that there is missing or incorrect information in the metadata that causes problems for X!Tandem. As such, data are effectively being combined from two search engines only. In other experiments, OMSSA, X!Tandem and MASCOT all perform similarly well, with either OMSSA or X!Tandem identifying the highest number of peptides at 1% FDR. Table 3 also displays the number of identifications made at 1 and 5% FDR for each pairwise combination of search engines. In most experiments, the combination of the top two search engines identifies almost as many peptides as using three

**Table 3.** The number of peptide-spectrum matches with *combined FDR Score* <0.01 and *combined FDR Score* <0.05 by MASCOT (M); OMSSA (O); X!Tandem (T); MASCOT and OMSSA (MO); OMSSA and X!Tandem (OT); MASCOT and X!Tandem (MT); and MASCOT, OMSSA and X!Tandem (MOT) across 14 different experiments

Experiment accession	Threshold	M	O	T	MO	OT	MT	MOT
PAe000066	0.01	1 883	5 405	178	5 865	5 499	1 993	5 794
PAe000066	0.05	3 387	6 254	335	6 197	6 276	3 475	6 232
PAe000077	0.01	1 256	2 636	204	3 009	2 738	1 346	3 016
PAe000077	0.05	1 687	3 166	391	3 357	3 256	1 788	3 426
PAe000093	0.01	2 520	5 924	1 096	6 900	6 310	3 268	6 914
PAe000093	0.05	4 020	7 269	1 631	8 093	7 715	4 796	8 240
PAe000098	0.01	184	171	40	304	195	232	271
PAe000098	0.05	362	218	80	427	290	427	469
PAe000138	0.01	883	1 915	832	2 424	2 515	1 540	2 628
PAe000138	0.05	1 259	2 510	1 565	3 012	3 281	2 358	3 497
PAe000146	0.01	11 543	18 710	937	20 781	19 013	11 833	20 776
PAe000146	0.05	14 675	21 190	1 706	21 693	21 271	14 874	21 859
PAe000157	0.01	340	648	172	849	773	443	866
PAe000157	0.05	551	917	252	1 167	1 047	669	1 202
PAe000158	0.01	10 705	11 503	11 200	15 220	16 513	15 808	17 091
PAe000158	0.05	14 190	14 812	14 022	17 413	18 258	17 699	19 570
PAe000160	0.01	6 382	7 191	7 687	9 404	11 567	11 123	11 987
PAe000160	0.05	8 776	9 012	9 920	11 259	13 059	12 461	14 215
PAe000162	0.01	1 140	1 638	1 072	2 151	2 375	2 144	2 757
PAe000162	0.05	1 849	2 229	1 482	2 866	3 108	3 057	3 752
PAe000165	0.01	17 601	22 151	13 702	26 078	26 300	22 511	27 904
PAe000165	0.05	22 934	25 915	16 655	28 426	28 762	25 417	30 858
PAe000166	0.01	5 304	4 697	7 091	6 635	8 394	8 651	9 007
PAe000166	0.05	6 705	5 986	8 807	7 667	9 785	9 695	10 327
PAe000167	0.01	949	891	1 194	1 419	1 860	1 845	1 939
PAe000167	0.05	1 462	1 228	1 779	1 833	2 343	2 381	2 548
PAe000292	0.01	344	570	95	852	658	406	822
PAe000292	0.05	523	882	147	1 060	965	648	1 084

search engines, indicating (as might be predicted) that continuously adding additional search engines would have limited benefit.

### 3.2 Validation with standard datasets

The ABRF has generated an artificial mixture of 49 known proteins to allow proteome technologies to be validated in laboratories. Datasets generated by several laboratories have been released, and are publicly available, although detailed analyses of the datasets have yet to be published. From the different laboratories that have publicly released their data, we selected the MS/MS set in which the highest number of proteins were correctly identified with no false positives, to test that the software was correctly differentiating true and false positive peptide identifications. According to preliminary data from ABRF (<http://www.abrf.org> – Proteomics Standards Research Group), this was laboratory 12874 that correctly identified 45 out of 49 proteins with 0 false positive identifications.

The dataset was searched with the three search engines and *combined FDR Scores* calculated. At a threshold of *combined FDR Score* <0.01, 2451 peptide-spectrum matches were made from 6027 spectra, combining results from the three search engines as outlined above. There were ten decoy identifications within the list of 2451, which were clearly false positives and removed (following *FDR Method 2*). Of the remaining peptides, 16 are derived from proteins not expected in the analysis and can be identified as false positives. The peptide FDR can thus be estimated independently of the decoys at 0.0066 (16/2441). The results could potentially contain two types of false positives: (i) contaminants in the sample (correctly identified by the search engine) and (ii) incorrect identifications made by the search engine. It is not a simple process to distinguish the two types of error. However, even if we assume that all false positives are of type (ii), it is clear that setting a *combined FDR Score* <0.01 results in an acceptably low false positive rate (estimated independently at 0.0066). This is important to demonstrate since it could be argued, in the absence of validation, that the algo-

rithm is artificially favouring target identifications over decoys, and the additional target peptide identifications made are all false positives. The analysis of ABRF datasets demonstrates that this is not the case, since there is no significant increase in the number of false positives within the targets.

## 4 Discussion

Large-scale proteome analyses produce significant quantities of data, but they are time-consuming and costly to run. Running more technical replicates can lead to larger numbers of identifications, but it is not always practical or cost-effective. Furthermore, software has been developed to determine absolute protein quantitation by counting the frequency of peptides identified by mass spectra [15, 16]. Any methods that can increase the number of peptide identifications from a single study are therefore significant, as they can reduce the number of replicates required, and reduce the overall cost and time to run an experiment. It has been previously suggested that by using multiple search engines, a higher proportion of the proteome can be sampled [17], but such efforts have been hampered by the lack of consensus on a software independent score.

We have re-searched considerable volumes of data, downloaded from PeptideAtlas, with MASCOT, OMSSA and X!Tandem. In this work, we introduce the concept of an *FDR Score*, which reflects the predicted rate of false discovery for an identification made with a particular score, reported by a single search engine on a specific dataset. The *FDR Score* has a similar basis to the statistical measure of a *q*-value but uses a form of regression to maintain the local ordering of identification quality, which is lost in the calculation of *q*-values. Crucially, *FDR Scores* allow identifications from different search engines to be combined within the same scoring framework. Our implementation of the algorithm uses *e*-values to rank identifications produced by each search engine but other measures could also be appropriate such as X!Tandem's hyperscore or a MASCOT ion score. The crucial factor is finding the score that best differentiates correct and incorrect identifications. Further improvements in the results may be possible by using a combination of factors to rank identifications as effectively as possible, as discussed by Choi and Nesvizhskii [13].

It has previously been demonstrated that due to differences in how search engines score identifications, there are differences in the sets of peptides discovered. By analysing FDRs, we are able to demonstrate that if different search engines agree on identifications, the frequency of false positives is low. However, even in the sets of peptides identified by only a single search engine, true positives can still be found. The *combined FDR Score* allows results to be extracted from sets of identifications made by one, two or three search engines, maximising the number of peptide identifications for a controlled FDR. While clearly the method is somewhat

heuristic, we have demonstrated using standard datasets that it is robust, and does not appear to introduce an artificial bias of target identifications over decoys. It should be noted that the algorithm is intended for large datasets, at least several thousand spectra, since the errors associated with FDR calculation are increased when the total number of identifications is small [18].

The benefits of combining multiple search engine results has also been demonstrated by the Scaffold software [19]. Scaffold does not rely on rates of false discovery, but instead works on a related metric; the probability of a peptide being correct or incorrect, calculated using the PeptideProphet algorithm [8]. Identification probabilities can be calculated for each search engine, and Scaffold creates a combined probability of correct identification if more than one search engine has identified the same peptide from a spectrum. The relationship between identification probability (or *p*-values) and FDR has been examined recently by Käll *et al.* [20] and by Choi and Nesvizhskii [13]. Käll *et al.* argue that error probabilities are more relevant where the presence of a specific peptide or protein is being considered but that in large scale proteome scans, setting thresholds by FDR appears to be more successful at balancing the trade-off between false positives and sensitivity. Choi and Nesvizhskii demonstrate however, that *p*-values are connected to FDRs in the PeptideProphet implementation, and thus could be used in a similar way to control FDR.

In the data presented for Scaffold [19] (which calculates identification probabilities), a 33% gain in peptide identifications is reported over the best performing single search engine in an 18 protein sample, and a 14% increase in a more complex sample, at a 1% FDR. On average, our method identifies 35% more peptides than the best single search engine at 1% FDR, with gains of almost 70% in certain datasets, demonstrating that FDR-based scoring is effective in differentiating correct and incorrect identifications.

In summary, we present a proposal for a software-independent measure of the quality of an identification, the *FDR Score*, which can be assigned to all peptide identifications when a decoy database search has been performed. We have utilised the *FDR Score* to combine data across search engines, and demonstrated how the *FDR Score* can be reassessed to reflect the contribution of evidence from different search engines. The *combined FDR Score* is an effective discriminator between correct and incorrect identifications, allowing considerable gains in the number of peptides that can be identified for a fixed FDR. There are clearly benefits to FDR-based scoring, and the *FDR Score* would be a useful addition to existing search engine software. The algorithm for combining scores across search engines, and calculating *combined FDR Score*, provides a viable alternative to using probability-based scoring methods and can be implemented relatively simply for whichever search engines are available. The algorithm has been implemented as a Perl Script which we will make available on request.

Work in Manchester by A. R. J. was funded by a grant from the BBSRC. J. A. S., S. J. H., N. W. P. also acknowledge BBSRC funding on the ISPIDER grant, ref BBS/B/17204. The authors thank Julian Selley for assistance with MASCOT searches.

The authors have declared no conflict of interest.

## 5 References

- [1] Matthiesen, R., Methods, algorithms and tools in computational proteomics: A practical point of view. *Proteomics* 2007, 7, 2815–2832.
- [2] Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J., OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* 2003, 3, 1454–1463.
- [3] Fenyo, D., Beavis, R. C., A Method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 2003, 75, 768–774.
- [4] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. *et al.*, Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, 3, 958–964.
- [5] MacCoss, M. J., Wu, C. C., Yates, J. R., Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* 2002, 74, 5593–5599.
- [6] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20, 3551–3567.
- [7] Balgley, B. M., Laudeman, T., Yang, L., Song, T., Lee, C. S., Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol. Cell. Proteomics* 2007, 6, 1599–1608.
- [8] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, 74, 5383–5392.
- [9] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, 75, 4646–4658.
- [10] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, 4, 207–214.
- [11] Käll, L., Storey, J. D., MacCoss, M. J., Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* 2008, 7, 29–34.
- [12] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I. *et al.*, The PeptideAtlas project. *Nucleic Acids Res.* 2006, 34, D655–D658.
- [13] Choi, H., Nesvizhskii, A. I., False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* 2008, 7, 47–50.
- [14] Feng, J., Naiman, D. Q., Cooper, B., Probability-based pattern recognition and statistical framework for randomization: Modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* 2007, 23, 2210–2217.
- [15] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 2007, 25, 117–124.
- [16] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. *et al.*, Comparison of label free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, M500084–MCP500200.
- [17] Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A. *et al.*, An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* 2006, 5, 3475–3490.
- [18] Huttlin, E. L., Hegeman, A. D., Harms, A. C., Sussman, M. R., Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J. Proteome Res.* 2007, 6, 392–398.
- [19] Searle, B. C., Turner, M., Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* 2008, 7, 245–253.
- [20] Käll, L., Storey, J. D., MacCoss, M. J., Noble, W. S., Posterior error probabilities and false discovery rates: Two sides of the same coin. *J. Proteome Res.* 2008, 7, 40–44.