technical notes Journal of proteome research

# Accurate and Sensitive Peptide Identification with Mascot Percolator

**Markus Brosch, Lu Yu, Tim Hubbard, and Jyoti Choudhary***

*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom*

**Received November 12, 2008**

***Abstract:*** Sound scoring methods for sequence database search algorithms such as Mascot and Sequest are essential for sensitive and accurate peptide and protein identifications from proteomic tandem mass spectrometry data. In this paper, we present a software package that interfaces Mascot with Percolator, a well performing machine learning method for rescoring database search results, and demonstrate it to be amenable for both low and high accuracy mass spectrometry data, outperforming all available Mascot scoring schemes as well as providing reliable significance measures. Mascot Percolator can be readily used as a stand alone tool or integrated into existing data analysis pipelines.

**Keywords:** Peptide identification • database search algorithm • Mascot • data analysis • machine learning • SVM • Percolator

## Introduction

Technological advances in the field of mass spectrometry (MS) enable high-throughput shotgun proteomics experiments[1,2] producing thousands of tandem-MS spectra.[3−5] Database search engines are currently the method of choice for annotating spectra with peptide sequences, the most widely used being Sequest,[6] Mascot[7] and X!Tandem.[8] Database search algorithms calculate for every peptide spectrum match (PSM) a score that reflects the quality of the cross-correlation between the experimental and the computed theoretical peptide spectrum. The scored PSMs are ranked, and typically, only the best matches for each spectrum are reported. However, the top peptide match of a spectrum is not necessarily correct, and therefore, sensitivity and accuracy of peptide and protein identification are reliant on sound scoring schemes. Many alternative methods have been applied: from manual heuristic rules,[9] such as simple score thresholds, to more complex systems that score and classify PSM based on an ensemble of features, thereby exploiting information present in the search results that is otherwise not used.[10−14]

A crucial step forward in assessing the reliability of reported PSMs was the introduction of the target/decoy search strategy pioneered by Moore et al.:[15] data is not only searched against the standard sequence database (target), but also against a reversed,[15] randomized,[16] or shuffled[17] database (decoy). PSMs obtained from the decoy database can be used to estimate the number of incorrect target PSMs and directly enable the estimation of the well-established false discovery rate[18] (FDR). In this context, the FDR can be interpreted as the expected proportion of incorrect PSMs among the selected set of identifications.[19] Since FDRs are not a function of the underlying score, the *q*-value metric, proposed by Storey and Tibshirani[20] in the field of genomics, was applied to mass spectrometry by Käll et al.[19,21] The *q*-value can be understood as the minimal FDR at which a PSM is accepted, hence, enabling the association of specific *q*-values with any PSM in a data set.[19,21] However, it is important to note that the *q*-value measure is always a result of all PSMs in a data set; for example, with a *q*-value cutoff of 0.05, one would expect 5% incorrect PSMs in the data set.

With the advent of high accuracy instrumentation, it was anticipated that peptide identification specificity would improve, since peptide mass accuracy in the region of a few ppm reduces the search space by orders of magnitude.[22−24] However, our recent study, employing the target/decoy search strategy to evaluate the performance of Mascot, revealed that this is not necessarily the case.[25] Mascot reports a probability-based Mascot Identity Threshold (MIT) for each individual spectrum query, above which a PSM is considered to be a significant peptide assignment.[25] Our study demonstrated that the MIT was anticonservative (low specificity, but high sensitivity) for stringent peptide mass tolerance settings (small search space) and conversely very conservative (high specificity, but low sensitivity) for relaxed parameter settings. Mascot also reports an empirical Mascot Homology Threshold (MHT) at which a Mascot score can be considered a significant outlier from the score distribution of all peptide matches to a given spectrum. Overall, the MHT was shown to be more sensitive than the MIT, but is only reported for PSMs where sufficient peptide candidates are scored, for example, at relaxed search parameter settings. These findings led us to implement the Adjusted Mascot Homology Threshold (AMT), utilizing the MHT at relaxed search parameters that, combined with a peptide mass deviation filter (AMT/mass-filter) on mass error recalibrated data, was shown to be the most sensitive Mascot scoring method available for high accuracy data.[25]

However, a limitation of the AMT/mass-filtering strategy is that it requires a fixed mass tolerance filter in order to subsequently determine a score threshold that maintains a predefined FDR. A more flexible implementation would be to use both features, the score cutoff and the mass deviation, in combination for discrimination of correct and incorrect PSMs. This can be achieved using the recently introduced iterative

* To whom all correspondence should be addressed: E-mail: jc4@sanger.ac.uk.


10.1021/pr800982s CCC: $40.75 
Published on Web 04/01/2009

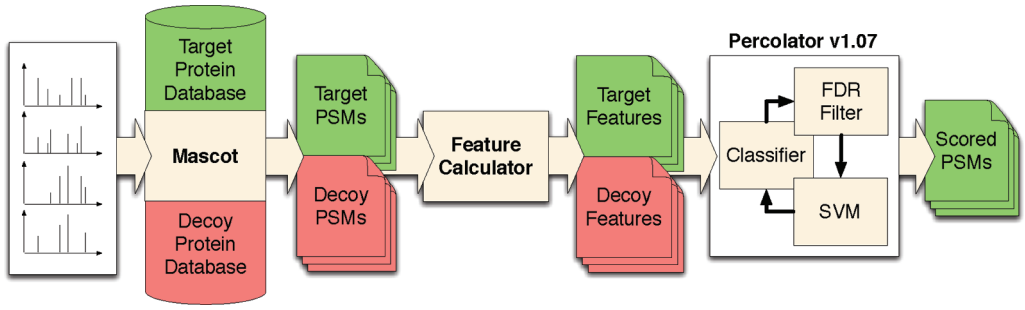


**Figure 1.** Illustration of the Mascot Percolator workflow.

machine learning method called Percolator[14] that utilizes target/decoy data.

For each target and decoy PSM, Percolator computes a vector of features that is related to the quality of the match (e.g., cross correlation scores or mass deviation). Subsequently, the set of target and decoy PSMs are discriminated by the most relevant feature (e.g., PSM score) and filtered to a fixed FDR (e.g., 1%). This subset (positive training set), together with all the decoy PSMs (negative training set), is used for training a support vector machine. The learnt classifier is then applied to all target/decoy PSMs, again followed by FDR filtering to continue the training procedure as before (Figure 1, percolator box). It was shown that, after a few iterations, the system converges and results in a robust classifier that is then used to rescore each PSM in the data set. For each PSM, the associated *q*-value, as well as the probability of the individual PSM being incorrect, is reported.[21,26] The whole process is fully automated and does not require any expert-driven or subjective decisions, thereby eliminating any artificial biases. The learnt classifier is specifically adapted and unique for each data set, thus, adapting to variations in data quality, protocols and instrumentation.

Although Percolator was originally designed for Sequest use only, the availability of a standard input format enables the use of Percolator as a generic machine learning algorithm where target/decoy data are available. We have therefore implemented a Mascot extension ("Mascot Percolator") that extracts and computes relevant features from the Mascot search results, trains Percolator, applies the resulting classifier to each PSM and writes a result file. We first assessed the AMT/mass-filtering approach with Mascot Percolator, but also extended this method with more features directly available from Mascot search results, such as Mascot scoring information, peptide and protein properties. Moreover, an extended feature set comprising information not directly accessible from Mascot search results, including ion matching statistics and intensity information, was explored. We have evaluated the performance of Mascot Percolator with high precursor mass accuracy LC-MS/MS data sets. We also benchmarked it with the low mass accuracy LC-MS/MS data set used in the original Percolator publication. In a final assessment, we validated the *q*-value accuracy reported by Percolator with a protein standard data set. Mascot Percolator is freely available at http://www.sanger.ac.uk/Software/analysis/MascotPercolator/ including databases, peak lists and results as presented in this article.

## Methods

**Samples.** Sample 1: A nuclear protein extract of murine embryonic stem cells (2 mg/mL) was reduced with 1 mM dithiothreitol (Sigma) at 70 °C for 10 min followed by alkylation with 2 mM iodoacetamide (Sigma) at room temperature (25 °C) for 30 min. Ten micrograms of total protein was separated on a NuPAGE Novex 4−12% Bis-Tris gel (Invitrogen). The gel was stained with colloidal Coomassie Blue (Sigma). The entire gel lane was excised into 48 bands, destained with 50% acetonitrile, and subsequently digested with sequencing grade trypsin (Roche) overnight. Peptides were extracted with 5% formic acid, 50% acetonitrile twice and vacuum-dried in a SpeedVac (Thermo Fisher Scientific). Peptides were redissolved in 0.5% formic acid and subjected to LC-MS/MS.

Sample 2: Yeast (*Saccharomyces cerevisiae* strain S288C) sample; see Käll et al.[14]

Sample 3: A standard protein set of 48 human proteins (Sigma, Universal Proteomics Standard Set UPS1) was reduced with Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), and alkylated with iodoacetamide as above, followed by digestion in solution with sequencing grade trypsin (Roche Applied Science) overnight. To minimize the chance of detection of low-abundance contaminants in the protein standard sample, a very low concentration of 10 fmol (per protein) was directly subjected to the LC-MS/MS.

**LC-MS/MS Analysis.** Peptides were analyzed using an online nano-LC-MS/MS system comprising an LTQ FT (Thermo Fisher Scientific), a hybrid linear ion trap and a 7-T Fourier transform ion cyclotron resonance mass spectrometer, coupled with an Ultimate 3000 Nano/Capillary LC System (Dionex). Samples were first loaded and desalted on a trap (0.3 mm inner diameter (i.d.) × 5 mm) at 20 $\mu$L/min with 0.1% formic acid for 5 min and then separated on an analytical column (75 $\mu$m i.d. × 15 cm) (both PepMap C18, LC Packings) over a 30-min linear gradient of 4−40% $CH_3CN$, 0.1% formic acid for sample 1. The flow rate through the column was 300 nL/min. For sample 3, the separation gradient was a 120-min gradient 4−32% $CH_3CN$/0.1% formic acid on a Atlantis C18 column (100 $\mu$m i.d. × 25 cm, Waters).

The LTQ FT mass spectrometer was operated in standard data-dependent acquisition mode controlled by Xcalibur 1.4 software. The survey scans were acquired on the FT-ICR ($m/z$ 400−2000 for sample 1, or 400−1500 for sample 3) at a resolution of 100 000 at $m/z$ 400, and one microscan was acquired per spectrum. For sample 1, the top three most abundant multiply charged ions with a minimal intensity at 1000 counts were subjected to MS/MS in the linear ion trap at an isolation width of 3 Th. For sample 3, the top 5 most abundant doubly and triply charged ions were subjected to MS/MS with the isolation width of 1.5 Th.

Precursor activation was performed with an activation time of 30 ms and activation $Q$ at 0.25. The normalized collision energy was set at 35%. The dynamic exclusion width was set at 5 ppm with two repeats and a duration of 30 s for sample 1, 10 ppm with 1 repeats and duration of 60 s for sample 3. To

achieve high mass accuracy, the automatic gain control target value was regulated at $4 \times 10^5$ (for sample 1) or $1 \times 10^6$ (for sample 3) for FT and $1 \times 10^4$ for the ion trap with a maximum injection time of 1000 ms for FT and 100 ms for the ion trap (sample 1) or 250 ms (sample 3). The instrument was externally calibrated using the standard calibration mixture of caffeine, a small peptide (sequence: MRFA) and Ultramark 1600.

**RAW Data Analysis.** LTQ FT MS raw data files were processed to peak lists with BioWorks 3.2 (Thermo Fisher Scientific). Processing parameters were as follows: precursor masses were set to 800−4500 Da, grouping was enabled allowing 50 intermediate MS/MS scans, precursor mass tolerance was set to 10 ppm, minimum ion count in the MS/MS was at 15. The number of minimum scans per group was set to 1. For sample 3, grouping was disabled.

LC-MS/MS analysis and RAW conversion for sample 2 has been previously performed and described by Käll et al.[14]

**MS/MS Database Searching.** Sample 1: Peak lists (38 058 spectra) were searched with Mascot 2.2 using the following parameters: enzyme = trypsin (allowing for cleavage before proline[27]); maximum missed cleavages = 2; variable modifications = carbamidomethylation of cysteine, oxidation of methionine; product mass tolerance = 0.5 Da. The International Protein Index (IPI) database version 337 (*Mus musculus*) was used as a protein sequence database. Common external contaminants from cRAP (a maintained list of contaminants, laboratory proteins and protein standards provided through the Global Proteome Machine Organisation, http://www.thegpm.org/crap/index.html, were appended. The compounded database contained 51 355 sequences and 23 635 027 residues. For FDR assessment, a separate decoy database was generated from the protein sequence database using the decoy.pl Perl script provided by Matrix Science. This script randomizes each entry, but retains the average amino acid composition and length of the entries.

Data was searched at 100 ppm peptide mass tolerance to evaluate the mass accuracy of the data set. After a correction[25] of a systematic mass deviation of 3 ppm, 90% and 99% of all PSMs with a Mascot score greater than 30 fell within a ±5 and ±20 ppm mass window, respectively. For the most stringent mass tolerance settings where Mascot thresholds are most sensitive, the data was searched at 20 ppm. Moreover, data was also searched at 500 ppm peptide mass tolerance to enable mass accuracy filtering combined with the adjusted MHT (Adjusted Mascot Threshold, AMT[25]). The mass deviation filter was set to 5 ppm, which was shown to be the most effective filter setting in combination with the AMT (Supporting Information Figure 1).

Sample 2: Peak lists (35 236 spectra) were searched with Mascot 2.2. against the same target and decoy databases that were used by Käll et al.[14] The following parameters were used: enzyme = trypsin; maximum missed cleavages = 2; fixed modification = carbamidomethylation of cysteine; peptide mass tolerance settings = 3 Da; product mass tolerance = 0.5 Da.

Sample 3: Peak lists (8190 spectra) were searched with Mascot 2.2 against human IPI (June 2007, 68 322 sequences, 28 806 780 residues) including common external contaminants from cRAP. Parameters used: enzyme = trypsin; maximum missed cleavages = 2; variable modifications = carbamidomethylation of cysteine, oxidation of methionine and deamidation of aspargine and glutamine; peptide mass tolerance = 20 ppm; product mass tolerance = 0.5 Da. Furthermore, 10 randomized versions of the sequence database were generated (using the decoy.pl script as described above) and were searched individually under the same conditions.

**Mascot Percolator.** Mascot Percolator was implemented with the Java programming language, ensuring platform independent operation. It utilizes the Mascot Java parser library provided by Matrix Science (http://www.matrixscience.com/msparser.html) and uses the generic interface to Percolator (Washington University, http://noble.gs.washington.edu/proj/percolator/). The latest Percolator version 1.07 was used for this study, which should be taken into account when comparing results of this study to the original publication of Percolator,[14] where version 1.01 was used.

Mascot Percolator performs the following operations for each run: it reads the Mascot results files, computes the scoring features as introduced in the Results and Discussion section and uses these for the Percolator training as was described in the Introduction. In a last step, the result file of Percolator and the input files are merged to comprise peptide, protein and scoring information (Figure 1).

Mascot Percolator was designed as a command line program to run either as a stand-alone application or as a component that can be embedded into existing data processing pipelines, allowing for streamlining data and automation. An example of executing the program follows for illustration: "java -cp MascotPercolator.jar cli.MascotPercolator -target 11026 -decoy 11027 -out 11026-11027". This command line reads the Mascot results from the files that are associated with the provided Mascot job IDs (11026. 11027), calculates the features used for the subsequent Percolator run and writes the results and logs into files prefixed with 11026-11027. Percolator was used with its default parameters. With the use of the basic and extended feature set, Mascot Percolator processes about 1500 and 75 PSMs/s (2.4 GHz AMD CPU), respectively.

**Data Analysis.** Receiver Operating Characteristics for Mascot Percolator were generated by varying the *q*-value cutoff values and reporting the corresponding number of true positives. The MIT, MHT and AMT were used as a reference for comparison. When no MHT was reported, the MIT was used instead, which is the default behavior of Mascot. Receiver Operating Characteristics for the MIT and MHT were generated by varying the Mascot significance threshold *p* (default 0.05) between $1 \times 10^{-5}$ to $1 \times 10^{-1}$, the latter representing the maximum allowed.

Percolator factors the percentage of target PSMs that are incorrect[19] into the *q*-value calculation (ref 14 supplementary methods 1.1.2). For consistency, the *q*-value calculations of MIT, MHT and AMT also take this factor into account and were determined using the software "qvality":[26] 0.55 (sample 1), 0.5 (sample 2), 0.77 (sample 3).

## Results and Discussion

**Mascot Percolator Using Peptide Mass Accuracy Features.** Sample 1 is a large data set acquired on a LTQ FT and is representative of a high mass accuracy proteomics experiment. For this data set, we previously showed that the AMT/mass-filtering method was the most sensitive Mascot scoring method available:[25] the data were searched at 500 ppm peptide mass tolerance, filtered to 5 ppm (Supporting Information Figure 1) and AMT thresholding was applied, resulting in 13 668 estimated true positive peptide identifications at a *q*-value of 1.0%. In comparison, the MIT and MHT at the same *q*-value only identified 10 385 and 12 338 true positives at the most restrictive (see Methods) peptide mass tolerance setting of 20 ppm (Figure 2, AMT, MIT, MHT).

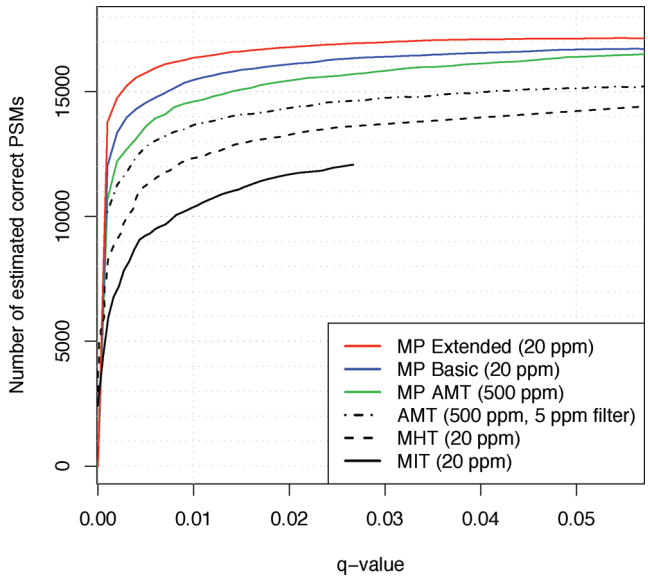


**Figure 2.** For the 20 ppm Mascot search, the basic and extended Mascot Percolator (MP), the Mascot Identity Threshold (MIT) and the Mascot Homology Threshold (MHT) performance were determined as a function of *q*-value cutoffs ranging from 0 to 0.06. Moreover, the performance of the mass-filtering (5 ppm) strategy, together with the Adjusted Mascot Threshold (AMT) and the emulated Percolator AMT method (MP AMT), is shown for the 500 ppm Mascot search. Note: if no MHT was reported, the MIT was used.[25]

**Table 1.** Features 1−9 Represent the Basic Feature Set and Features 1−18 Represent the Extended Feature Set As Used in Mascot Percolator[a]

| feature abbreviation | feature description |
|---|---|
| 1. mass | Calculated monoisotopic mass of the identified peptide. |
| 2. charge | Precursor ion charge |
| 3. mScore | Mascot score |
| 4. dScore | Mascot score minus Mascot score of next best nonisobaric peptide hit |
| 5. deltaM | Calculated minus observed peptide mass (in Dalton and ppm). |
| 6. absDeltaM | Absolute value of calculated minus observed peptide mass (in Dalton and ppm) |
| 7. isoDeltaM | Calculated minus observed peptide mass, isotope error corrected (in Dalton and ppm) |
| 8. uniquePeps | None (0), one (1), two or more (2) distinct peptide sequences match same protein |
| 9. mc | Missed tryptic cleavages |
| 10. totInt | Total ion intensity (log) |
| 11. intMatchedTot | Total matched ion intensity (log) |
| 12. relIntMatchedTot | Total matched ion intensity divided by total ion intensity |
| 13. binom | Peptide Score as described in ref 28 |
| 14. fragMassError | Mean fragment mass error (in Dalton and ppm) |
| 15. absFragMassError | Mean absolute fragment mass error (in Dalton and ppm) |
| 16. fracIonsMatched | Fraction of calculated ions matched (per ion series) |
| 17. seqCov | Sequence coverage of matched ions (per ion series) |
| 18. intMatched | Matched ion intensity (per ion series) |

[a] Further discussion of these features can be found in supplemental information 1 (Supporting Information).

A more flexible implementation would be to use both features, the score cutoff and the mass deviation, in combination for improved discrimination of correct and incorrect PSMs, for example, accepting PSMs with slightly larger mass deviation given the PSM scores are highly significant.

This can be achieved with a machine learning algorithm such as Percolator using features relevant to the AMT/mass-filtering strategy. Accordingly, the following features were calculated from the 500 ppm Mascot target and decoy searches and were used for Percolator training: MHT minus Mascot score, deviation of theoretical and observed peptide mass, and the absolute value of the mass deviation.

Mascot Percolator identified a total of 14 603 estimated true positive PSMs at a 1.0% *q*-value (Figure 2, MP AMT), clearly outperforming the AMT/mass-filtering approach by 7%. When Mascot Percolator was compared to the Mascot thresholds, it identified 41% (38%) and 18% (17%) more true positive (unique) peptides than the MIT and MHT, respectively, significantly outperforming both Mascot thresholds.

These results demonstrate that the combined use of the score threshold and the mass deviation features as a discriminator outperforms the AMT/mass-filtering strategy. It should be noted that the used features tackle systematic mass errors and random mass errors separately, therefore, simplifying the usability since postprocessing to remove systematic mass shifts is not required. These promising results, combined with the ability of the Percolator algorithm to handle any number of features, motivated the assessment of more comprehensive feature sets.

**Mascot Percolator Using Extended Feature Sets.** In addition to the mass deviation features described previously, features that can be directly extracted from the Mascot search results were added as inputs to Mascot Percolator, defining the "basic feature set" (Table 1, feature 1−9). Moreover, an "extended feature set" that comprises fragment ion matching statistics was also considered (Table 1, feature 1−18). However, these features are not readily available in the Mascot result files and were therefore computed by matching the observed spectra against the theoretical spectra. Some of these features (16−18) were calculated separately for each ion series (e.g., combinations of b/y series, singly/doubly charged series and neutral loss series).

With the use of the target/decoy Mascot search results for subsequent Percolator training with the basic and extended feature set, the peptide identification performance improved by 6% and 11%, respectively, as compared to the Mascot Percolator performance using only the AMT/mass-filtering features (Figure 2 and Supporting Information Figure 2). Since the same number of identifications were made for the 500 ppm and 20 ppm search, the basic and extended feature sets appear to effectively substitute the necessity for strong mass accuracy discriminators.

Therefore, Mascot Percolator with features that include Mascot scoring and peptide features as well as ion matching statistics identified more than 58% (52%) and 33% (29%) more true positive (unique) peptides than the MIT and MHT, respectively, at a 1.0% *q*-value with a standard 20 ppm search (Figure 2). This translates into 15% and 6% more protein identifications over the MIT and MHT, respectively.

Overall, these results are a significant improvement over all current Mascot scoring methods, including AMT, and eliminate the need to search high accuracy data at relaxed mass tolerances to improve sensitivity, as discussed in ref 25.

**Mascot Percolator Applied to Low Mass Accuracy Data.** The following evaluation is concerned with sample 2, a yeast data set acquired on a LTQ instrument that was used for the evaluation of Sequest Percolator. To enable comparison of Mascot Percolator and Sequest Percolator, the subsequent experiments were therefore not only based on the same data, but also on the same target/decoy databases and search

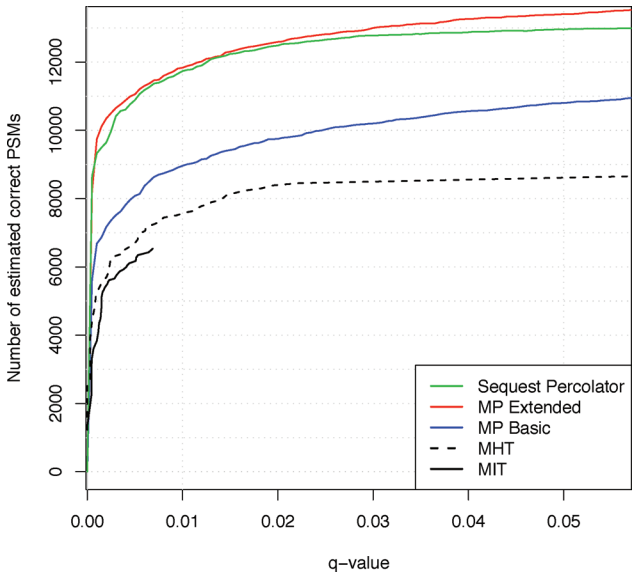


**Figure 3.** The number of estimated correct PSMs were determined for each *q*-value cutoff for the basic and extended Mascot Percolator (MP) runs, the Adjusted Mascot Threshold (AMT), the Mascot Identity Threshold (MIT) and Mascot Homology Threshold (MHT) as well as for the Sequest Percolator.

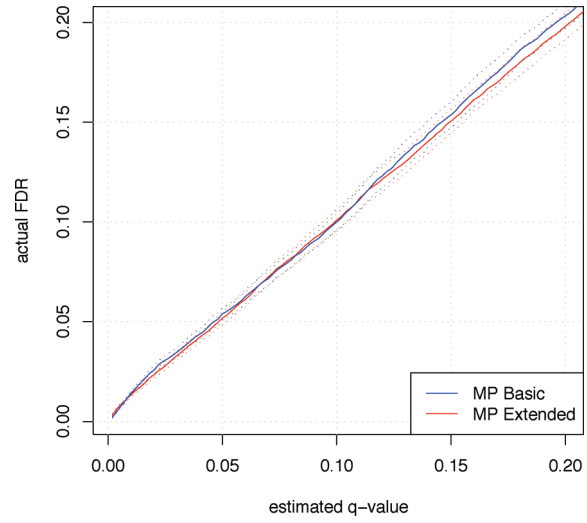


**Figure 4.** The estimated *q*-values were plotted against the false discovery rates as reported by the protein standard data set for the extended and the basic Mascot Percolator runs. The dotted lines represent the standard error.

parameters as described by Käll et al.,[14] with the only exception being the trypsin specificity parameter.

With the use of the MIT and MHT, 6426 and 7541 true positive identifications (Figure 3, MIT, MHT) were made at a *q*-value of 0.7% and 1.0%, respectively (the Mascot significance threshold is limited to 0.1, corresponding to a *q*-value of 0.7%). Using the basic feature set with Mascot Percolator improved sensitivity over MIT and MHT by more than 39% and 19%, respectively, at a 1.0% *q*-value (Figure 3, MP basic). Sensitivity was further boosted by more than 40% when the extended feature set was applied (Figure 3, MP extended). Compared to the MIT and MHT, this relates to a (unique) peptide identification gain of 84% (74%) and 57% (49%), respectively, at the standard 1.0% *q*-value, translating to 57% and 38% more protein identifications, respectively. Overall, these results further support the performance advantages of Mascot Percolator over the default MIT and MHT.

Moreover, the difference in performance of Mascot Percolator between the basic and extended feature set was significantly more prominent than it was with data from sample 1, highlighting that feature contribution can vary substantially for different data sets and demonstrating the dynamic and adaptive property of the Percolator algorithm (ref 14, supplement 2). It could be speculated that low accuracy data benefit from more discriminating features, while high accuracy data already reach the maximum sensitivity with the basic feature set due to the more restrictive search parameters and known charge states.

In addition, Käll et al. identified trypsin-specificity as a strong discriminating feature and consequently they searched without enzyme specificity.[14] However, this practice is significantly more CPU intensive due to the larger search space. Search times in Mascot are 1 order of magnitude slower when semitrypsin is specified instead of trypsin, and 2 orders of magnitude slower when no enzyme specificity is defined instead of trypsin (http://www.matrixscience.com/pdf/2006WKSHP1.pdf). Therefore, Mascot Percolator does not make use of any enzyme specificity related features, but maintains sensitivity with the extended feature set and performance is comparable to that of Sequest Percolator (Figure 3).

**Validation with a Standard Protein Data Set.** The robustness and precision of the *q*-value was validated in the supplemental material of the original Percolator publication.[14] The employed target/decoy search strategy for *q*-value estimation is a widely accepted approach, but various methods exist for generating the decoy databases. Therefore, we evaluated the accuracy of the *q*-value as a result of the Matrix Science decoy.pl script (see Methods) with a protein standard data set (sample 3). Ten Mascot searches were performed and analyzed with Mascot Percolator, using the same target but independently generated random databases. This enabled computation of the standard error for the *q*-value calculations. For every estimated *q*-value, the corresponding observed FDR was determined by counting the incorrect PSMs that did not match the expected protein sequences.

It was found that *q*-value estimates for both Mascot Percolator versions (basic and extended feature set) were in very good agreement with the results obtained by the expected protein sequences (Figure 4). This implies that the gain in sensitivity (Supporting Information Figure 3) with Mascot Percolator is limited to valid sequences within the expected error rates. Moreover, the same data set was used for a more demanding no-enzyme search and showed similar accuracy of the *q*-value estimates, demonstrating robust scoring (Supporting Information Figure 4).

Overall, the *q*-value evaluations have shown that none of the chosen features introduced any bias toward severe under- or overestimation of the *q*-values and that these can be seen as accurate and reliable estimates of the real error rates. This is a significant improvement over the standard Mascot results using the MIT or MHT, for which we have previously shown that the actual FDR can differ by several fold from the expected FDR.[25]

## Conclusion

The Percolator machine learning algorithm was recently introduced to rescore Sequest results and demonstrated significantly improved sensitivity for peptide and protein identification. Percolator learns a classifier independently for each

data set, thereby adapting to inherent variations between different data sets, such as changing analytical protocols or instrumentation. In this work, we have implemented and evaluated Mascot Percolator, a software package that interfaces Mascot with Percolator. It automatically extracts and computes relevant features from target/decoy Mascot search results, trains Percolator, applies the resulting classifier to each PSM and writes a result file. Mascot Percolator has been developed as a command line tool and can be readily integrated into existing pipelines or be used as a stand-alone application. A large number of features that are relevant to the quality of a PSM, such as Mascot scores, parent and fragment mass accuracy, peptide, protein, as well as ion matching statistics, among others, were explored.

We have shown that Mascot Percolator substantially outperforms previous Mascot scoring methods for high and low mass accuracy data, in the best case identifying 74% and 49% more unique peptides and 57% and 38% more proteins than using the MIT and MHT, respectively. This demonstrates the improved discrimination potential achieved when several factors that define the quality of a PSM are used collectively for scoring instead of only one metric. Furthermore, we have shown that the estimated *q*-values are in very good agreement with the actual FDRs and represent a significant improvement in accuracy as compared to the Mascot thresholds.

**Abbreviations:** MS, mass spectrometry; PSM, peptide spectrum match; MIT, Mascot identity threshold; MHT, Mascot homology threshold; AMT, adjusted Mascot threshold; FDR, false discovery rate.

**Acknowledgment.** We thank Lukas Käll for providing and supporting Percolator as well as for many helpful discussions, John Cottrell from Matrix Science for Mascot related support and feature suggestions, and Jenny Mattison and Mark Collins for critically reading the manuscript. This work was funded by the Wellcome Trust.

**Supporting Information Available:** Evaluation of the performance of the Adjusted Mascot Threshold (AMT) using mass deviation filter settings of 50, 25, 10, 5 and 3 ppm, the Mascot Percolator performance for the relaxed (500 ppm) and stringent (20 ppm) Mascot search, the Mascot Percolator, MIT and MHT performance for the protein standard data set using the basic and extended feature set over a range of *q*-values and without any enzyme constraints, supplementary feature information. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **1992**, *255*, 1261–1263.

(2) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001**, *73*, 5683–5690.

(3) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E. C.; Zhang, H.; Aebersold, R. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005**, *6*, R9.

(4) de Godoy, L. M.; Olsen, J. V.; de Souza, G. A.; Li, G.; Mortensen, P.; Mann, M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **2006**, *7*, R50.

(5) Foster, L. J.; de Hoog, C. L.; Zhang, Y.; Zhang, Y.; Xie, X.; Mootha, V. K.; Mann, M. A mammalian organelle map by protein correlation profiling. *Cell* **2006**, *125*, 187–199.

(6) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(7) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

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

(9) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76*, 3556–3568.

(10) 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.

(11) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2003**, *2*, 137–146.

(12) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *22*, 214–219.

(13) Ulintz, P. J.; Zhu, J.; Qin, Z. S.; Andrews, P. C. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol. Cell. Proteomics* **2006**, *5*, 497–509.

(14) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925.

(15) Moore, R. E.; Young, M. K.; Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378–386.

(16) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3*, 1454–1463.

(17) Klammer, A. A.; MacCoss, M. J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **2006**, *5*, 695–700.

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

(19) 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.

(20) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.

(21) 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.

(22) Zubarev, R. A.; Hakansson, P.; Sundqvist, B. Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Anal. Chem.* **1996**, *68*, 4060–4063.

(23) Zubarev, R.; Mann, M. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **2007**, *6*, 377–381.

(24) 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.

(25) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell. Proteomics* **2008**, *7*, 962–970.

(26) Käll, L.; Storey, J. D.; Noble, W. S. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008**, *24*, i42–48.

(27) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7*, 300–305.

(28) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24*, 1285–1292.

PR800982S