Chapter 3

Accurate and sensitive peptide identification with Mascot Percolator

3.1 Introduction

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 (Elias and Gygi, 2007; Zubarev and Mann, 2007; Zubarev *et al.*, 1996). However, in chapter 2 I have evaluated the performance of Mascot and demonstrated that this is not necessarily the case. I have shown that the Mascot Identity Threshold (MIT) was anti-conservative (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

peptide-spectrum matches (PSMs) where sufficient peptide candidates are scored, e.g. at relaxed search parameter settings. These findings led me to implement the Adjusted Mascot Homology Threshold (AMT), utilising 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 a sensitive Mascot scoring method for high accuracy data (see chapter 2).

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 cut-off and the mass deviation, in combination for discrimination of correct and incorrect PSMs. This can be achieved using the iterative machine learning method called Percolator (Käll *et al.*, 2007). See section 1.1.2.3 for more background information concerning Percolator.

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. I have therefore developed 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. I firstly 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 and peptide properties. Moreover, an extended feature set comprising information not directly accessible from Mascot search results, including ion matching statistics and intensity information, was explored. I have evaluated the performance of Mascot Percolator with high precursor mass accuracy LC-MS/MS datasets, but also benchmarked it with the low mass accuracy LC-MS/MS dataset used in the original Percolator publication. In a final assessment, I validated the q-value accuracy reported by Percolator with a protein standard dataset acquired on a range of

instruments. Mascot Percolator is freely available at `http://www.sanger.ac.uk/Software/analysis/MascotPercolator/` including databases, peak lists and results as presented in this chapter.

Parts of this chapter were published in the Journal of Proteome Research (Brosch *et al.*, 2009) by the author of this thesis (Markus Brosch) and my supervisors (Tim Hubbard, Jyoti Choudhary) as well as by Lu Yu who run the mass spectrometry experiments (specifically indicated in the relevant section).

Moreover, in collaboration with John Cottrell and David Creasy (Matrix Science, London) my method presented in this chapter is currently implemented into the official Mascot 2.3 software release (`http://www.matrixscience.com/workshop_2009.html`), and will be readily applicable by the proteomics community without the need of any third party software.

3.2 Methods

3.2.1 Datasets and experimental methods

- Dataset 1: LTQ-FT (Thermo Fisher Scientific) dataset from a nuclear protein extract of murine embryonic stem cells. Data and methods were described in detail and used in chapter 2.

- Dataset 2: Käll *et al.* (2007) provided us with their Yeast (*Saccharomyces cerevisiae*) dataset acquired on an LTQ (Thermo Fisher Scientific).

- Dataset 3: LTQ-FT dataset from a standard protein set comprising 48 human proteins (Universal Proteomics Standard Set UPS-1, Sigma). Data and methods were described in detail in chapter 2. In addition, the same sample was also acquired on a LTQ, LTQ-FT Ultra and Q-Tof Premier (Waters) by Lu Yu (Team 17, Wellcome Trust Sanger Institute), providing me a comprehensive set of protein standard data.

3.2.2 MS/MS database searching

Dataset 1: Peak lists of (38,058 spectra) were searched with Mascot 2.2 using the following parameters: enzyme = trypsin (allowing for cleavage before proline (Rodriguez *et al.*, 2007)); 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 (see 1.1.1.2). 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 randomises 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 data mass accuracy. After a correction of a systematic mass deviation of 3 ppm(Brosch *et al.*, 2008), 90% and 99% of all PSMs with a Mascot score greater than 30 fell within a ±5 ppm 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 (Brosch *et al.*, 2008), see chapter 2). The mass deviation filter was set to 5 ppm, which was shown to be the most effective filter setting in combination with the AMT (figure 3.1).

Dataset 2: Peak lists of (35,236 spectra) were searched with Mascot 2.2. against the same target and decoy databases that were used by Käll *et al.* (2007). 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.

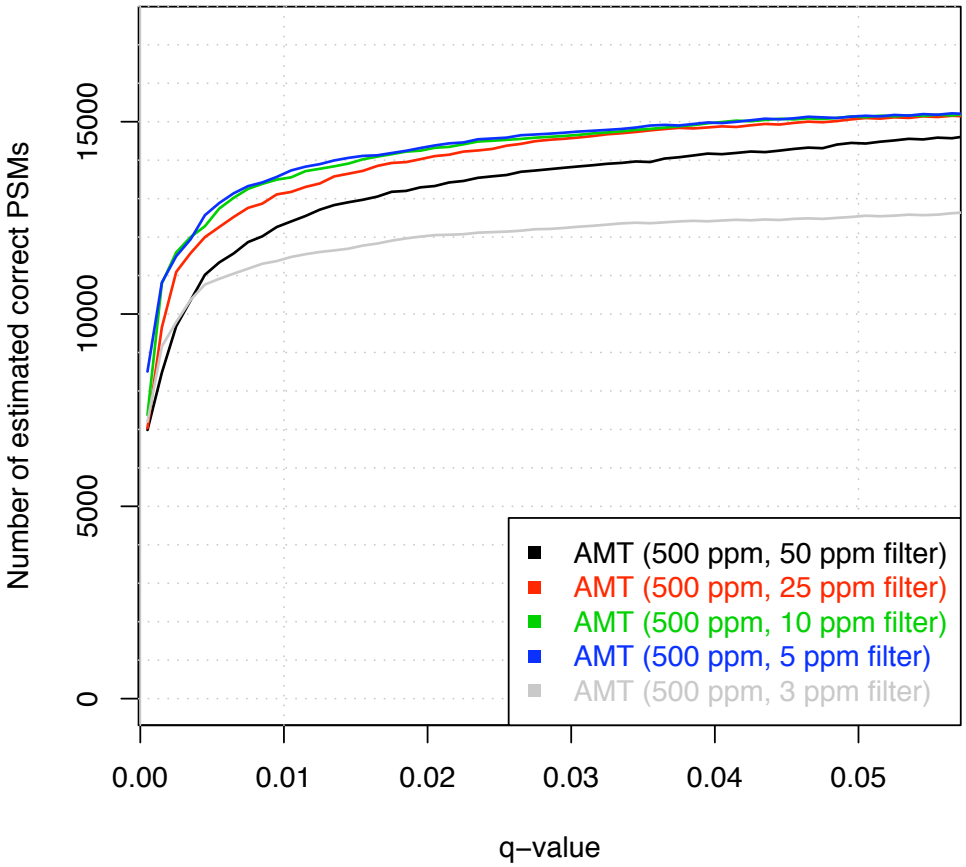


Figure 3.1: The performance of the Adjusted Mascot Threshold (AMT) was evaluated using mass deviation filter settings of 50, 25, 10, 5 and 3 ppm: for each, the number of estimated correct PSMs was determined across a range of q-values. These results show the trade-off between improving specificity with more stringent mass tolerance filters and conversely excluding potentially correct PSMs when the filters become too stringent. For this dataset the best mass filter was found to be 5 ppm.

Dataset 3: Peak lists (spectra count LTQ: 43,710, LTQ-FT: 45,289, LTQ-FT Ultra: 18,285, Q-Tof: 1206) 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/product mass tolerance = LTQ: 0.9 / 0.5 Da, LTQ-FT: 20, 50, 200 ppm / 0.5 Da, LTQ-FT Ultra: 10 ppm / 0.5 Da, Q-Tof: 30 ppm / 0.2 Da; 5 randomised versions (decoy.pl) of the sequence database were generated and searched individually under the same conditions.

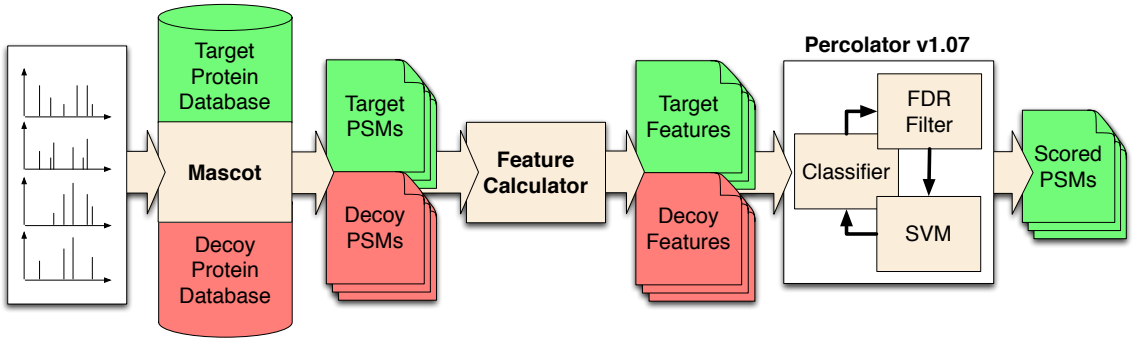


Figure 3.2: Illustration of the Mascot Percolator workflow.

3.2.3 Mascot Percolator implementation

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 1.12 using default parameters was used for this study, which should be taken into account when comparing results of this study to the original publication of Percolator (Käll *et al.*, 2007), where version 1.01 was used. Results in this chapter are based on Mascot Percolator version 1.09.

The Mascot Percolator implementation 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 described in section 1.1.2.3. In a last step, the result file of Percolator and the input files are merged to combine peptide, protein and scoring information (figure 3.2).

3.2.4 Data analysis

Receiver Operating Characteristics (ROCs) for Mascot Percolator were generated by varying the q-value cutoffs and reporting the corresponding number of estimated 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 behaviour of Mascot. ROCs for the MIT and MHT were generated by varying the Mascot significance threshold p (default 0.05) between 1×10^{-5} to 1×10^{1}, the latter representing the maximum allowed. Percolator factors the percentage of target PSMs that are incorrect (π_0) into the q-value calculation (see section 1.1.2.3). For consistency, the q-value calculations of MIT, MHT and AMT also take this factor into account and were determined using "Qvality": 0.55 (dataset 1), 0.5 (dataset 2), 0.77 (dataset 3).

3.3 Results and Discussion

3.3.1 Peptide mass accuracy features

Dataset 1 is representative of a large high mass accuracy proteomics experiment. For this dataset I previously showed that the AMT/mass-filtering method was the most sensitive Mascot scoring method available (see chapter 2). Therefore, the data were searched at 500 ppm peptide mass tolerance, filtered to 5 ppm (figure 3.1) 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 method section) peptide mass tolerance setting of 20 ppm (figure 3.3, AMT, MIT, MHT).

A more flexible implementation would be to use both features, the score cut-off 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

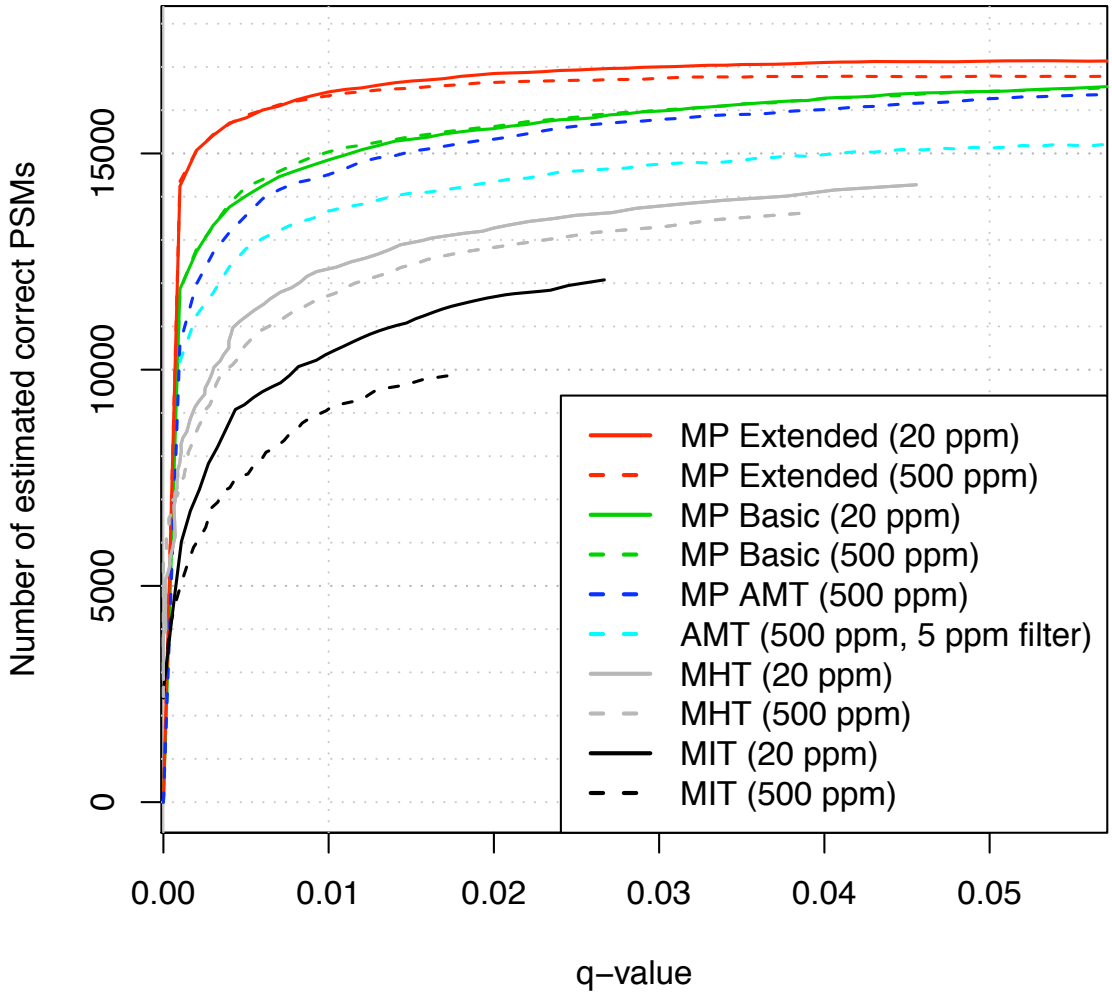


Figure 3.3: For the 20 ppm Mascot search, the basic and extended Mascot Percolator (MP), the Mascot Identity Threshold (MIT) and the Mascot Homology Threshold (MHT) performance was determined as a function of q-value cut-offs ranging from 0 to 0.06. Moreover, the performance of the mass-filtering (5 ppm) strategy together with the Adjusted Mascot Threshold (AMT), the emulated Percolator AMT method (MP AMT), the MIT and MHT are shown for the 500 ppm Mascot search. Note: if no MHT was reported, the MIT was used (default Mascot behaviour).

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 14,512 estimated true positive PSMs at a 1.0% q-value (figure 3.3, MP AMT), clearly outperforming the AMT/mass-filtering approach by 6.2%. When Mascot Percolator was compared to the Mascot thresholds, it identified 40% (37%) and 18% (17%) more true positive (unique) peptides than

the MIT and MHT, respectively, significantly improving performance upon both Mascot thresholds.

These results demonstrate that the combined use of the score threshold and the mass deviation features as a discriminator provide better performance than 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 post-processing to remove systematic mass shifts is not required. These promising results motivated the assessment of more comprehensive feature sets.

3.3.2 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, defining the "basic feature set"(table 3.1, feature 1-9).

The idea behind the additional chosen features: the native Mascot score is known to correlate well with the quality of a PSM (feature 1); the difference of the Mascot score between two non-isobaric peptide hits indicates the level of ambiguity between two competing matches (feature 2); the number of missed tryptic cleavages or variable modifications of a peptide may be indicative of whether the PSM matches the properties of the rest of the dataset (feature 8 and 9 respectively). Feature 3-4 are not expected to provide discrimination power by themselves, but they may correlate with other features and thereby improve discrimination.

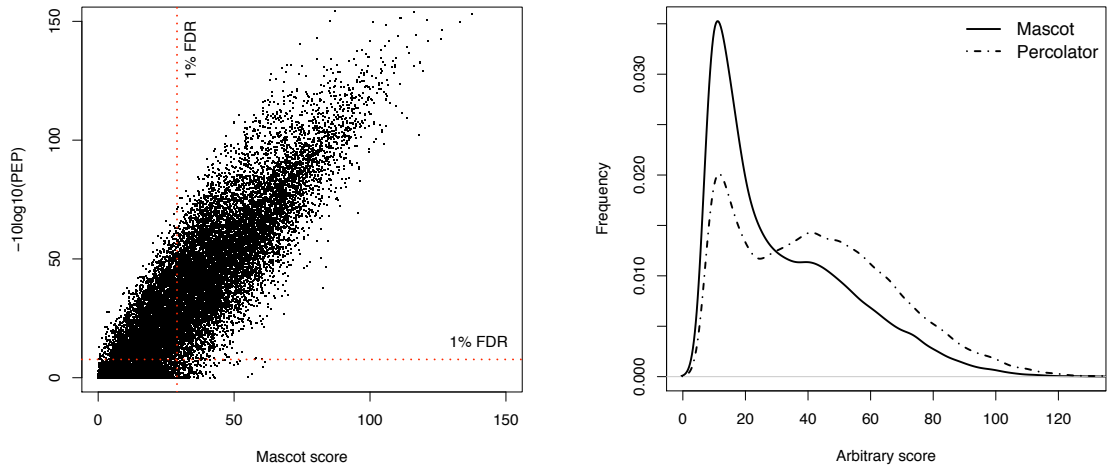
Moreover, an "extended feature set"(table 3.1, feature 1-17) that required re-matching the experimental spectra against the theoretical spectra was considered. The idea was to include fragment ion matching statistics, not readily available from the Mascot results: a higher total (matched) ion intensity can be indicative of better spectrum and peptide-spectrum match quality (feature 10-12 respectively); fragment mass error statistics is widely used to manually validate PSMs (feature

13-14); the longest consecutive ion series as well as the fraction of ions matched is another commonly used feature for manual validation (feature 15, 16) and lastly, the fraction of matched ion intensity relative to the total ion intensity may further aid discrimination. Importantly, features 15-17 are computed for each ion series separately, e.g. b and y series, doubly charged b and y series as well as the b and y series including derivatives such as neutral losses of ammonia or water, enabling Mascot Percolator to learn ion series preferences from the dataset at hand.

Using the target/decoy Mascot search results for subsequent Percolator training with the basic and extended feature set, the peptide identification performance improved by 2.5% and 13%, respectively, as compared to the Mascot Percolator performance using only the AMT/mass-filtering features (figure 3.3). Since about 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.

Table 3.1: Features 1-9 represent the "basic feature set" and features 1-17 represent the "extended feature set" as used in Mascot Percolator.

Feature No.	Short description
1	Mascot score
2	Mascot score of current peptide hit relative to 2nd best hit rank
3	Calculated monoisotopic mass of the identified peptide
4	Charge (1 to n)
5	Calculated minus observed peptide mass (in Dalton and ppm)
6	Feature No. 5, corrected for isotope error
7	Absolute value of feature No. 5
8	Number of missed tryptic cleavages
9	Number of variable modifications
10	Total ion intensity
11	Total ion intensity of matched ions
12	Relative ion intensity (Feature No. 11 / Feature No. 12)
13	Median of delta mass of fragment ions (in Dalton and ppm)
14	Interquartile range of delta mass of fragment ions (in Dalton and ppm)
15	Longest consecutive ion series (per ion series)
16	Fraction of ions matched (per ion series)
17	Relative ion intensity matched (per ion series)



(a) Mascot score plotted against the Mascot Percolator Posterior Error Probability score (log transformed). The 1% score cut-offs are indicated for each dimension.

(b) Mascot score overlaid with the Mascot Percolator Posterior Error Probability score (log transformed and scaled, for better visualisation only).

Figure 3.4: Both figures enable the visual comparison of the raw Mascot scores against the Mascot Percolator (posterior error probability) scores. Figure (a) highlights all the extra PSMs with weak Mascot score that Mascot Percolator accepted based on its discrimination power using all 17 features. Figure (b) compares the score distributions and the improved bi-modal distribution of the Percolator scores indicates the improved discrimination power.

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 3.3), clearly demonstrating the enhanced discrimination power when using an ensembl of features (figure 3.4). These improvements translate 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.

3.3.3 Mascot Percolator applied to low mass accuracy data

The following evaluation is concerned with dataset 2, a yeast sample acquired on a LTQ instrument that was used for the evaluation of Sequest Percolator. To enable

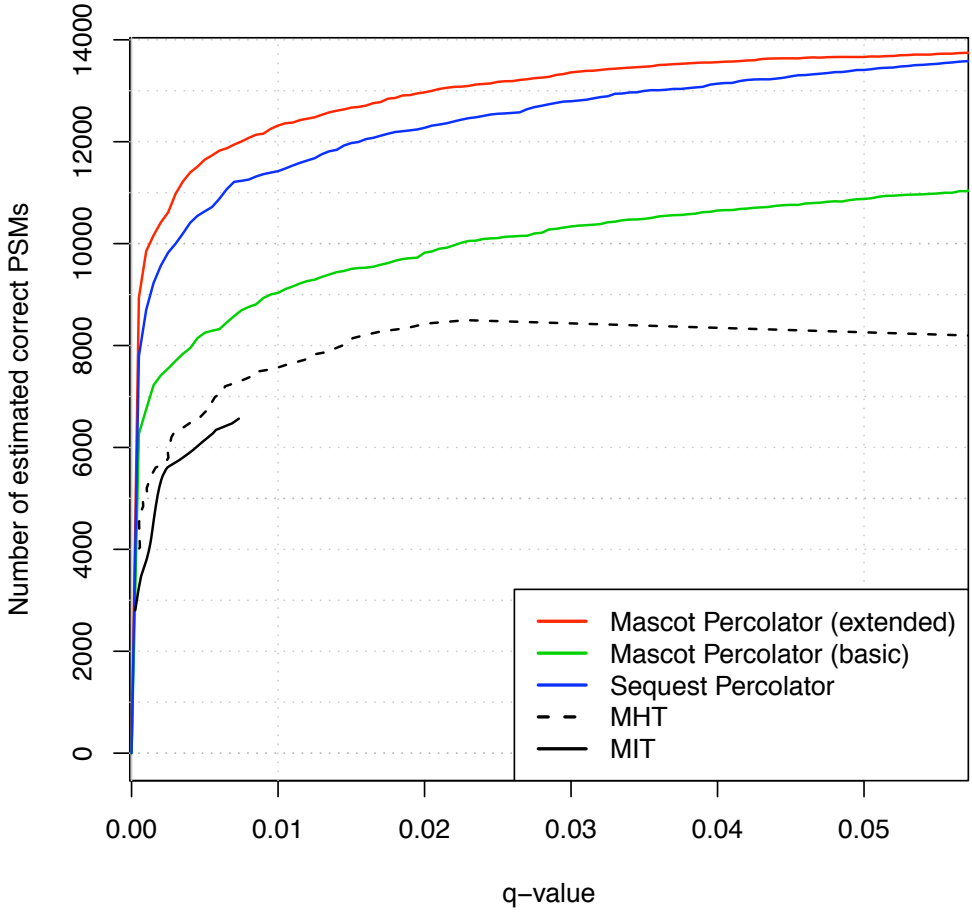


Figure 3.5: The number of estimated correct PSMs were determined for each q-value cut-off for the basic and extended Mascot Percolator (MP) runs, the Adjusted Mascot Threshold (AMT), the Mascot Identity Threshold (MIT) and Mascot Homology Threshold (MHT) as well as for the Sequest Percolator.

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 parameters as described by Käll *et al.* (2007), with the only exception being the trypsin specificity parameter.

Using the MIT and MHT, 6,379 and 7,541 true positive identifications (figure 3.5, 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 41% and 19%, respectively, at a 1.0% q-value (figure 3.5, MP basic). Sensitivity was further boosted by more than 50% when the extended feature set

was applied (figure 3.5, MP extended). Compared to the MIT and MHT, this relates to a (unique) peptide identification gain of 93% (82%) and 63% (55%), respectively, at the standard 1.0% q-value. 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 dataset 1, highlighting that feature contribution can vary substantially for different datasets and demonstrating the dynamic and adaptive property of the Percolator algorithm (Käll *et al.* 2007, supplement 2). It could be speculated that low accuracy data benefit from more discriminating features, while high accuracy data almost reaches the maximum sensitivity with the basic feature set due to the more restrictive search parameters and known charge states.

Käll *et al.* (2007) identified trypsin-specificity as a strong discriminating feature and consequently they searched without enzyme specificity in their study. However, this practice is significantly more CPU intensive due to the larger search space. Search times in Mascot are one order of magnitude slower when semi-trypsin is specified instead of trypsin, and two 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, yet improves upon the Sequest Percolator sensitivity with the extended feature set by about 8% at a 1% q-value (figure 3.5).

3.3.4 Validation with standard protein datasets

The robustness and precision of the q-value was validated in the supplemental material of the original Percolator publication (Käll *et al.*, 2007). The employed target/decoy search strategy for q-value estimation is a widely accepted approach, but various methods exist for generating the decoy databases (see section 1.1.2.3). Therefore, I extended this evaluation by assessing the accuracy of the q-value as

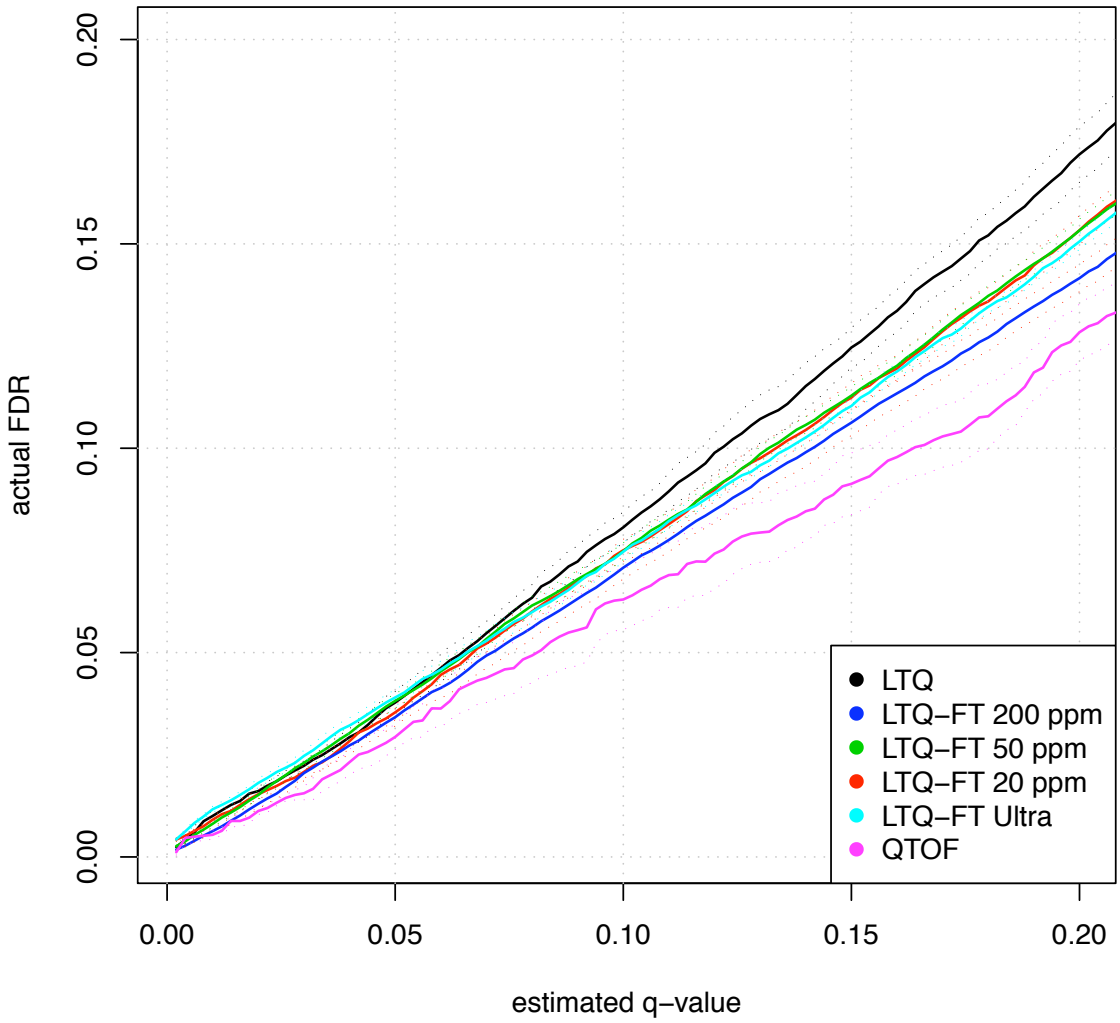


Figure 3.6: The estimated q-values were plotted against the false discovery rates as reported by the protein standard datasets for the extended and the basic Mascot Percolator runs. The dotted lines represent the standard error.

a result of the Matrix Science decoy.pl script (see methods) with a set of protein standard datasets (dataset 3). Five Mascot searches were performed and analysed with Mascot Percolator for each data, 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 were in very good agreement with the results

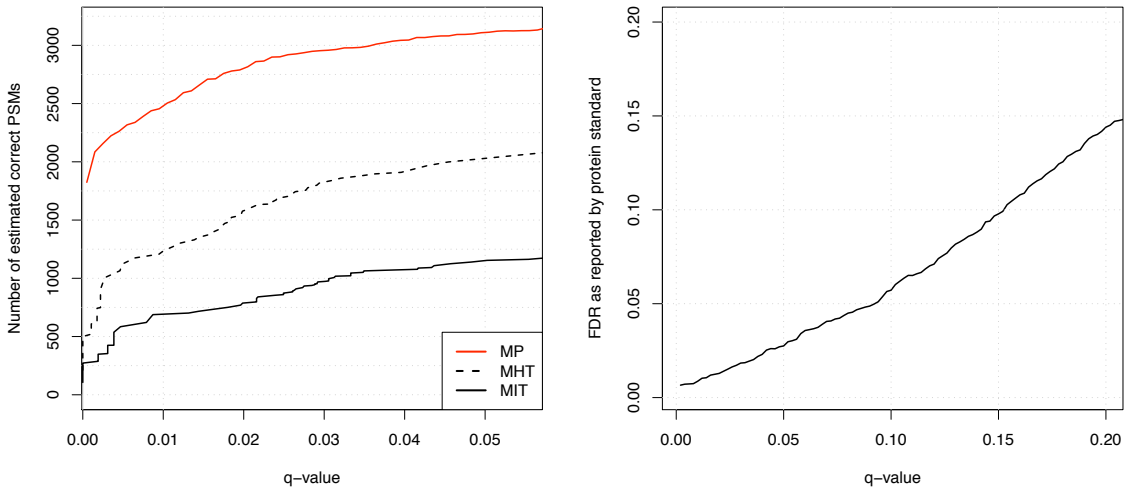


Figure 3.7: Performance of the Mascot Percolator, MIT and MHT were compared for a no enzyme search of the LTQ-FT protein standard dataset (left). Estimated q-values were also plotted against the FDR as reported by this protein standard dataset (right).

obtained by the expected protein sequences (figure 3.6), while the standard errors were negligible, particularly in the low FDR region. This implies that the gain in sensitivity with Mascot Percolator is limited to valid sequences within the expected error rates. These results demonstrate that none of the chosen features introduced any bias towards severe under- or overestimation of the q-values and that these can be seen as accurate and reliable estimates of the real error rates for a variety of analytical platforms. This is a significant improvement over the standard Mascot results using the MIT or MHT, for which I have previously shown that the actual FDR can differ by several fold from the expected FDR (Brosch *et al.*, 2008).

Moreover, the LTQ-FT dataset was also used for a more demanding no-enzyme search. The Mascot scoring scheme as well as the Mascot Percolator were evaluated with the protein standard dataset that was searched in Mascot without any enzyme constraints. Mascot Percolator identified 265% and 96% more peptides than using the MIT or MHT respectively, at a q-value of 1%. It should be noted that none of the features I use discriminate by enzyme specificity as pointed out earlier. Estimated q-values were validated against this protein standard dataset and showed good

accuracy, indicating that the identifications are limited to valid sequences within the expected error rates.

Overall these results demonstrate that Mascot Percolator can also be applied to more challenging conditions than standard tryptic searches, where the search space is increased by several orders of magnitude (`http://www.matrixscience.com/help/search_field_help.html`), such as searches without any enzyme constrains or excessive variable modification settings.

3.3.5 Mascot Percolator applied to a pool of 73 datasets

About 10 million tandem MS spectra from various sources were post-processed with Mascot Percolator in chapter 5 (see section 5.2.1 for details). This data enabled the evaluation of Mascot Percolator on a large scale. For each of the 73 datasets the increase in peptide identifications with Mascot Percolator over MHT, MIT and Mascot was determined, allowing to compute the median and interquartile range for each comparison: the median improvement at a 1% q-value were 54%, 109% and 99%, with an interquartile range (IQR) of 39%, 84% and 69% when the number of PSMs of Mascot Percolator were compared with MHT, MIT and the Mascot score respectively.

In a next step, the same data were searched against a database that was supplemented with gene predictions, resulting in about a 10-fold search space increase (see section 5.2.2 for details). Mascot Percolator identified 16% fewer peptides (median value) at a 1% q-value when compared with the searches against the smaller database, while the difference in performance between Mascot Percolator and the MIT, MHT and Mascot score also changed: the median improvement of peptide identifications with Mascot Percolator at a 1% q-value over the MHT, MIT and Mascot score were 65%, 197% and 155% respectively. While the improvement over the MHT was almost constant, the change over the MIT and raw score was considerable, resulting in about half the number of peptide identifications when compared to the search against the

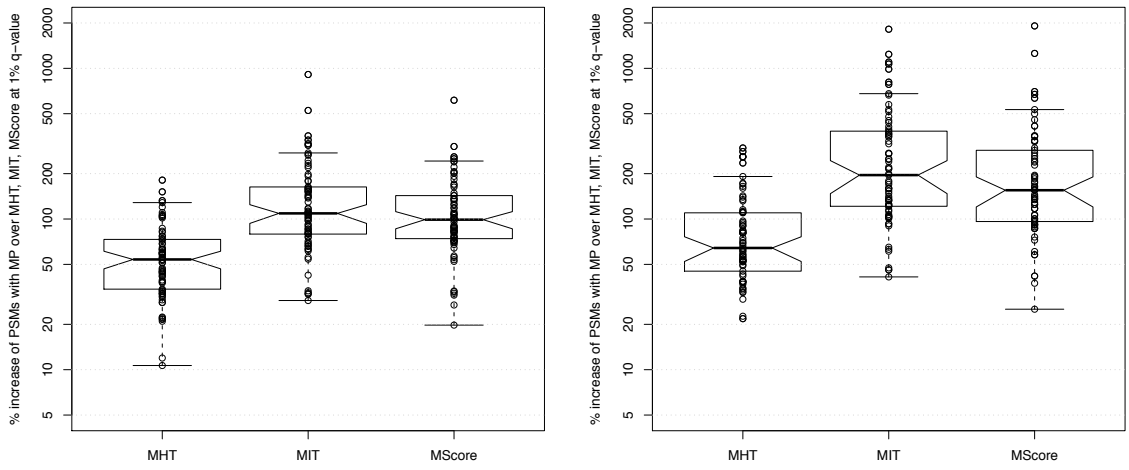


Figure 3.8: 73 datasets were processed with Mascot Percolator and the number of peptide identifications were compared to the identification rates of Mascot Identity and Homology thresholds as well as the raw Mascot score (left).
The same data was also searched against a 10-fold increased search space, followed by the same evaluation (right).

smaller database, further supporting the findings of chapter 2.

Overall this demonstrates on a large scale that Mascot Percolator shows a robust improvement over the native Mascot scoring, including a significant less severe drop in performance when search space inflates.

3.4 Mascot Percolator availability

Standalone package

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 -newDat". This command line triggers Mascot Percolator to parse the Mascot search results from the files that are associated with the provided Mascot job identifiers (11026, 11027) to subsequently calculate the features discussed above for the subse-

quent Percolator run. Results and logging files are written into files prefixed with "11026-11027". Moreover, the flag "newDat" directs Mascot Percolator to write a new Mascot results file (*.dat) that can be opened by the Mascot server just as if it was a standard Mascot result file. The differences however are as follows: the Mascot scores are replaced with the $-10log_{10}$ of the Posterior Error Probability (PEP); the expect value that is calculated on the Mascot results page directly corresponds to the PEP; the accepted FDR can be changed by setting the Probability values on the Mascot results page accordingly. A summary for illustration is shown in figure 3.9.

Mascot Percolator is available at `http://www.sanger.ac.uk/Software/analysis/MascotPercolator/`, where I also documented the more advanced command line options.

Distributed package

I was confronted with the problem to post-process the search results of 146 Mascot searches comprising a total of about 20 million spectra (2×10 million, see chapter 5). Moreover, I had to process these data as quickly as possibly and ultimately wanted to make use of our Mascot compute farm. This farm does not have a "Load Sharing Facility[1]" and hence there was a need to develop a distributed version of Mascot Percolator, which is now used by default in our lab.

The system is based on several components: a database server was implemented that runs independently of the system and logs every action of the distributed Mascot Percolator system. A Mascot Percolator Server was developed that after starting up connects to the database and triggers the status page for the intranet to be updated and listens for Mascot Percolator Nodes (figure 3.10). These nodes can be run on either Unix or Windows computers and automatically connect to the server. A script as well as a web-interface[2] were implemented that enable the submission of jobs in

[1] A commercial computer software job scheduler that can be used to execute batch jobs on networked Unix and Windows systems.

[2] The web-based submission interface was developed by Parthiban Vijayarangakannan, Team

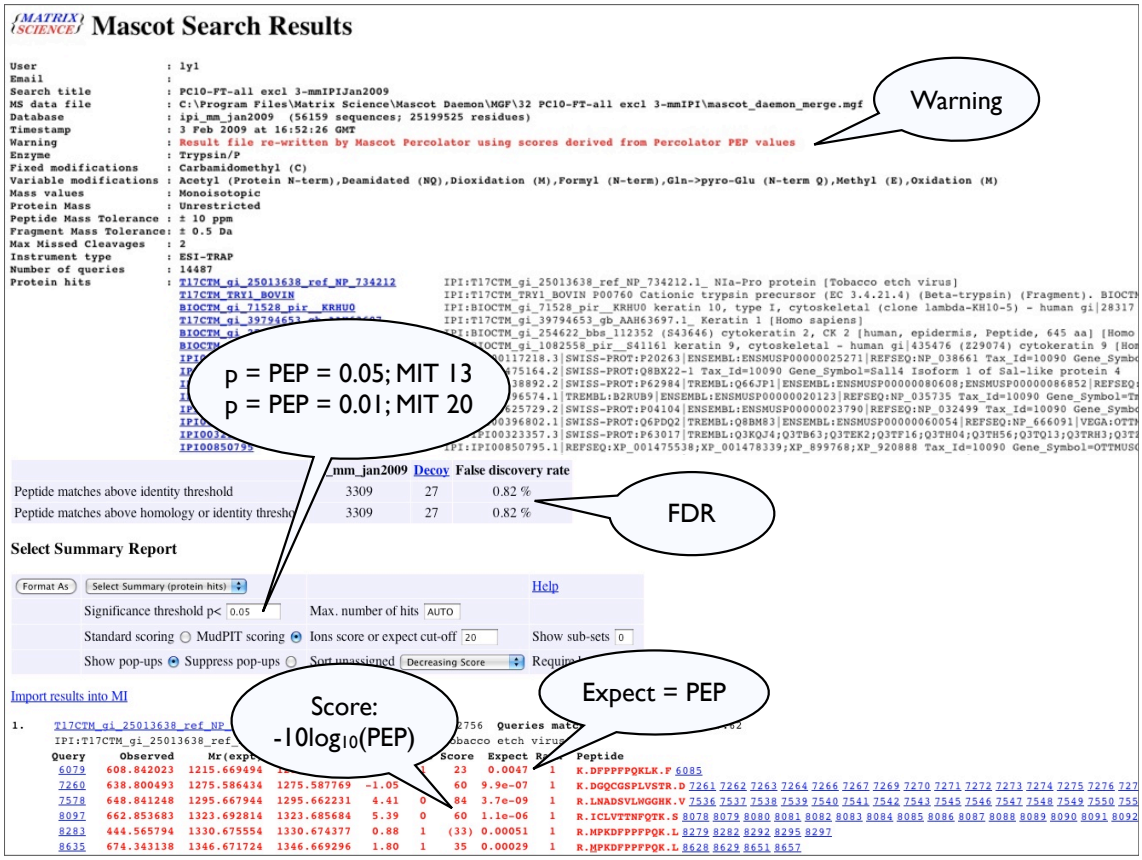


Figure 3.9: Screenshot of a Mascot results page that was generated by Mascot Percolator. The results are basically identical to a standard Mascot results page, can be opened by the Mascot server, but the scoring values are derived from the Mascot Percolator run. A warning at the top of the page states this very clearly to avoid confusion.

batches. The server then schedules the jobs and distributes these onto the different available nodes. When jobs complete, the status page updates and the result files can be browsed. The system has some more advanced options that are documented at `http://www.sanger.ac.uk/Software/analysis/MascotPercolator/`.

In conclusion this distributed system enabled me to process the large datasets reliably and efficiently in a fraction of the time when compared to the stand alone version of Mascot Percolator.

17, Wellcome Trust Sanger Institute

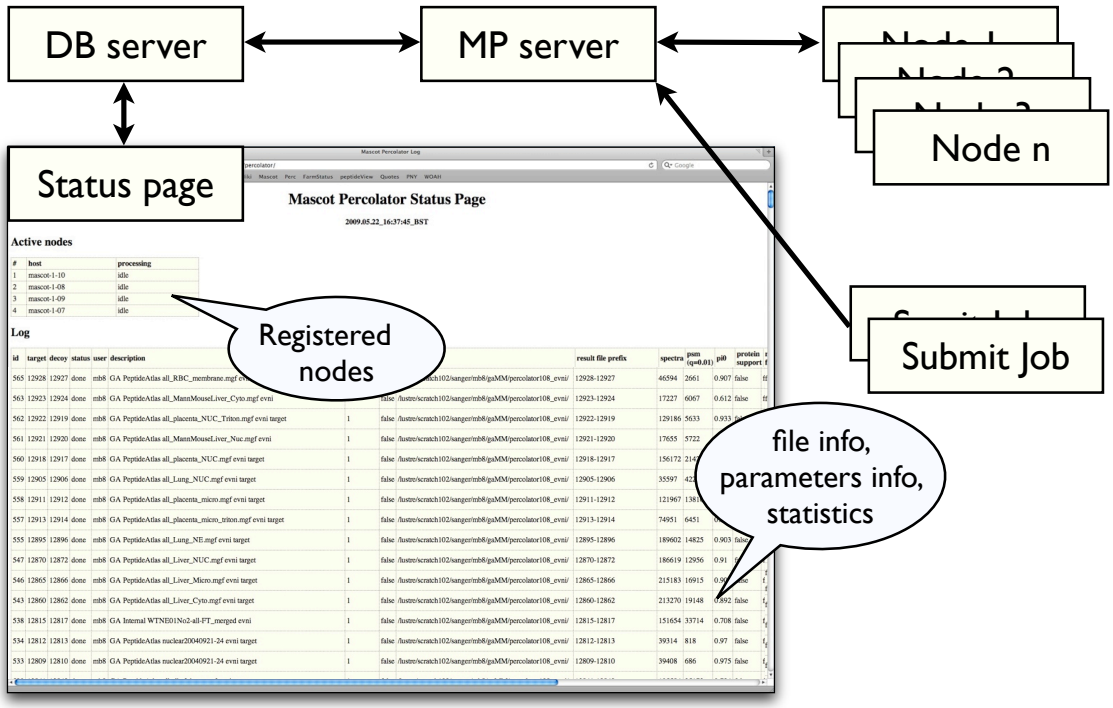


Figure 3.10: Schematic of the Distributed Mascot Percolator package and screenshot of the status webpage.

Official Mascot Percolator support by Matrix Science

In collaboration with John Cottrell, David Creasy (Matrix Science, London) and Lukas Käll (University of Stockholm), Mascot Percolator is currently implemented into the official Mascot release 2.3 (see `http://www.matrixscience.com/workshop_2009.html`). Features will be pre-computed for every Mascot search, cutting compute time and allowing streamlined processing through Percolator, without the need of a user to access the command line. This will ultimately expand the user group of Mascot Percolator significantly.

3.5 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 dataset, thereby

adapting to inherent variations between different datasets, such as changing analytical protocols or instrumentation.

In this work, I have developed 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 as well as ion matching statistics, amongst others, were incorporated.

I have demonstrated that Mascot Percolator substantially outperforms previous Mascot scoring methods for high and low mass accuracy data and applied it to a large ensembl of 73 datasets, where up to 65% and 197% more peptides than the MIT and MHT were identified with Mascot Percolator at a 1% q-value (median values). 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, I 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. Lastly, Percolator calculates both significance measures, the q-value and the posterior error probability. The latter is particularly important for my genome annotation efforts, where the significance of every peptide identification should be known.

Chapter 4

Development of a proteogenomics pipeline

4.1 Introduction

In chapter 3 the development and evaluation of Mascot Percolator was discussed, a powerful peptide scoring scheme with an implementation that can be automated and run in batch-mode, providing high-throughput capability. This system delivers sensitive peptide identification with accurate significance measures and thereby sets the foundation for a reliable proteogenomic pipeline. Mascot Percolator results can be written into a tab delimited text file or exported into the proprietary Mascot results file format. To use these data for genome annotation purposes, a proteogenomic pipeline is required that stores these data, maps the peptide identifications to the genome and enables comprehensive data analysis.

The currently available proteogenomics pipelines, Peptide Atlas (Desiere *et al.*, 2005) and GAPP (Shadforth *et al.*, 2006), which were described in detail in section 1.3, were found to be not suitable since these systems are highly specialised and the code bases are not in the public domain. The envisaged system should integrate peptide identifications available from Mascot Percolator, enable validation and