Chapter 2

Assessment of Mascot and X!Tandem and development of the Adjusted Mascot Threshold

2.1 Introduction

In the general introduction I have discussed the concept of sequence database searching, that is commonly used to assign sequence information to MS/MS spectra (section 1.1.1). This chapter focusses on the scoring schemes of database search algorithms, which are required to provide sound peptide assignment significance measures in order to minimising incorrect and maximising correct identification. Many different techniques have been applied in the past, from manual heuristic rules to machine learning algorithms that discriminate between correct and incorrect identifications (Anderson *et al.*, 2003; Jones *et al.*, 2009; Resing *et al.*, 2004; Ulintz *et al.*, 2006). The most popular database search engines to date, including Mascot (Perkins *et al.*, 1999) and X!Tandem (Craig and Beavis, 2004), provide theoretically or empirically derived statistical thresholds to help assess the significance of peptide identifications.

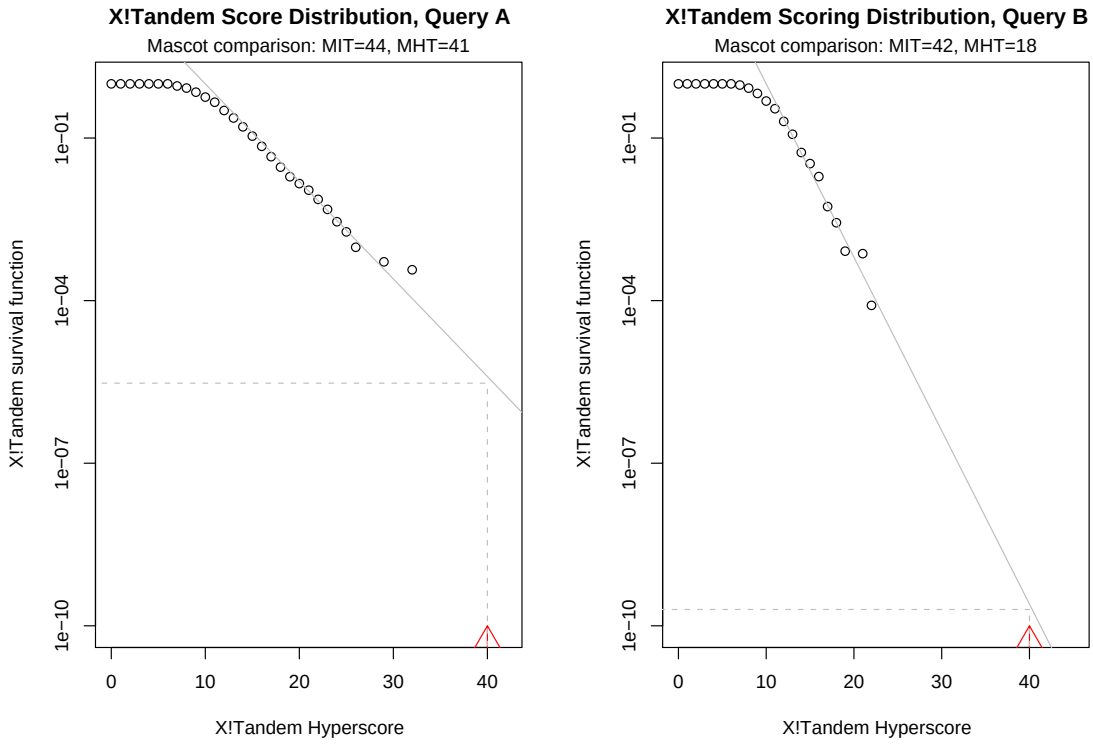


Figure 2.1: Exemplary survival functions from X!Tandem for two spectrum queries A and B. Although the number of peptide candidates for both queries is similar, there are apparent differences in the actual peptide score distributions. The survival functions were extrapolated for a score of 40 that corresponds to a probability of approximately 3×10^{-6} and 2×10^{-10} for query A and B respectively. Given the number of peptides scored were 1×10^5, the expectation value of the former would be 0.3 while the expectation value of the latter would be 2×10^{-5} (for a detailed explanation on how the survival function and expectation values are calculated, refer to Fenyo and Beavis, 2003). Therefore, at a significance level of 0.05 the same score would have been considered highly significant for query B, but not for query A.
In contrast, the MIT is inferred from the number of peptide candidates only, resulting in very similar thresholds of 44 and 42 for both queries. A hypothetical Mascot score of 40 would not have been considered significant for either query.
On the other hand, the empirically derived MHT was 41 for query A and 18 for query B, thus classifying the peptide hit for query B as significant which agrees with the X!Tandem extrapolation example. It should be noted that the absolute scores and threshold values of X!Tandem and Mascot are not directly comparable.

Mascot reports a probability-based Mascot Identity Threshold (MIT) for each individual spectrum query. A Mascot score above MIT is considered to be a significant peptide assignment. The MIT is defined as $-10 \times log_{10}(20 \times p \times n)$, where p is the probability of a random peptide match and n corresponds to the actual number of peptide candidates. For example, if a 1 in 20 chance of obtaining a false positive is acceptable ($p = 0.05$) and there are 10000, 1000, 100 and 10 peptide candidates for a given mass window in the sequence database, the MIT would be 40, 30, 20 and 10 respectively. For a peptide match with a score that equals the MIT ($p = 0.05$), the expectation value (E-value) of this hit is also 0.05, but if the score exceeds the MIT by e.g. 10, the E-value drops to 0.005. The E-value in Mascot is defined as $p \times 10^{(MIT-score)/10}$ and corresponds to the number of times one would expect this score by chance alone (`http://www.matrixscience.com/pdf/2005WKSHP4.pdf`). Therefore the MIT only reflects changes in search space, defined by the number of peptide candidates, and would be affected by various factors such as the maximum mass deviation (MMD) settings, the number of allowed missed cleavages, enzyme specificity and variable modifications.

Mascot also reports an empirical Mascot Homology Threshold (MHT). A Mascot score exceeding this threshold can be considered a significant outlier from the distribution of all candidate peptide-spectrum match scores, but an exact definition of the MHT was not published. Similarly, X!Tandem employs score distributions, but extrapolates empirical E-values to assess the significance of a peptide match (Craig and Beavis, 2004; Fenyo and Beavis, 2003). It is important to note that the E-values derived by Mascot and X!Tandem are based on completely different assumptions and may therefore lead to significantly different scoring results even for the same peptide spectrum match; as described above, the Mascot E-value is based on a theoretical statistical model, whereas the X!Tandem E-value is an empirical outlier determination. In figure 2.1 I illustrate the similarities and differences between the X!Tandem, MHT and MIT scoring scheme.

With high accuracy MMD settings the search space can decrease significantly leading to insufficient data points of the score distributions to reliably extrapolate E-values. To compensate for this, X!Tandem uses cyclic permutations of all peptide candidates that are scored and used to pad the score distribution (optional). In general, empirical scoring schemes that utilise the peptide candidate score distributions for thresholding or E-value extrapolation are more robust to changing MS/MS data quality such as signal to noise, mass accuracy or fragmentation quality.

It is anticipated that reducing the search space should improve the performance of algorithms for peptide identification (Zubarev and Mann, 2007). For example, with high mass accuracy data in the range of a few ppm, the search space can be reduced by orders of magnitude in comparison to low accuracy data acquired typically on ion trap instruments (Elias and Gygi, 2007).

Established database search algorithms, and in particular their scoring schemes, were not specifically developed for high mass accuracy data. Rudnick *et al.* (2005) evaluated the effects of MMD settings on Mascot performance and proposed an empirical Mass Accuracy based THreshold (MATH) that provided improved sensitivity at a user-defined false discovery rate (FDR). They applied a range of global cut-off thresholds and determined the associated FDRs. A linear regression over the logarithms of these FDRs and the cut-off values enabled an empirical threshold extrapolation at a predefined FDR. However, the Mascot evaluation was exclusively limited to the MIT. Savitski *et al.* (2005) have developed a database size independent scoring scheme for high accuracy data. This work is based on complementary fragmentation techniques, and cannot be applied solely on standard collision induced dissociation data (Biemann, 1988; Roepstorff and Fohlman, 1984). Gygi and co-workers proposed to exploit high accuracy MS data by searching at relaxed mass tolerance settings followed by mass accuracy filtering (Beausoleil *et al.*, 2006; Everley *et al.*, 2006). Combined with a moderate threshold on peptide-spectrum correlation scores, they found this strategy to serve as a good discriminator between correct and incorrect

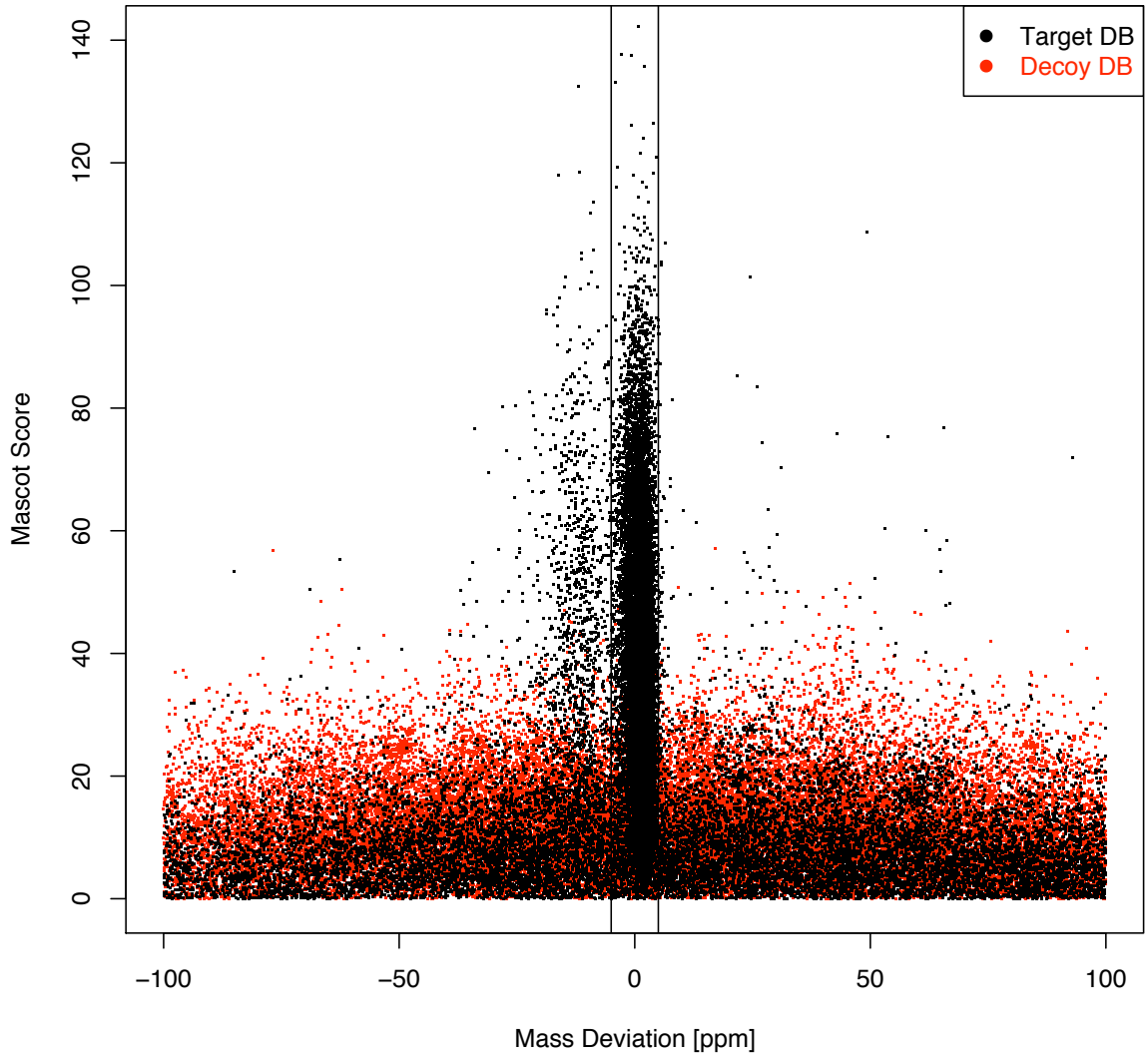


Figure 2.2: Distribution of all peptide matches obtained from a 1 Da MMD target and decoy database search of sample 1, showing the Mascot score and the mass deviations in ppm for a small window of ±100 ppm. Most mass deviations of high scoring peptide-spectra matches fell within the experimental mass errors that have been reported previously, 99% fell within ±20 ppm and 90% fell within ±5 ppm. The mass outliers between -5 and -20 ppm seem to be an experimental artefact for this particular sample.

peptide assignments. The rationale behind this is that the chance of finding a strong peptide match in a relaxed mass window with many peptide candidates is greater than for a very stringent mass window with only a few peptide candidates. A correct and strong match is likely to remain the same, regardless of the size of the search space. On the other hand, it is more likely for a weak match arising from a poor

spectra or from an incorrect peptide correlation to find a better alternative in a larger search space. A subsequent mass accuracy filtering step, which limits the matches to the experimental mass deviations, serves as useful discriminator between correct and incorrect matches. This is further illustrated in Figure 2.2 using the data of this study. Overall, these studies indicate that a more detailed evaluation and optimisation of established search algorithms for high accuracy mass spectrometry is still required.

In this chapter I have investigated the performance of Mascot and X!Tandem for varying MMD settings common for low and high accuracy MS. I show that the MIT is highly dependent on the search space and affects false discovery and identification rates. I also show that the empirical scoring scheme in X!Tandem is more robust across different mass tolerance settings. The Mascot equivalent empirical MHT outperforms X!Tandem for ion trap data, but is not comprehensively applicable for very stringent MMD settings. I demonstrate that searching high accuracy data at relaxed MMD windows followed by peptide mass accuracy filtering serves as a good discriminator between correct and incorrect assignments. I propose an alternative empirical Adjusted Mascot Threshold (AMT[1]), applicable to low accuracy data and, in combination with peptide mass accuracy filtering, also to high accuracy data. In addition, the AMT enables the user to freely select the best trade-off between sensitivity and specificity by defining the actual FDR.

Parts of this chapter were published in Molecular Cellular Proteomics (Brosch *et al.*, 2008) by the author of this thesis (Markus Brosch) and my supervisors (Tim Hubbard, Jyoti Choudhary) as well as by Sajani Swamy, who introduced me to the field of computational proteomics. Markus Brosch performed the work and wrote the manuscript. Lu Yu (acknowledgements) run the mass spectrometry experiments (specifically indicated in the relevant sections).

[1]Same abbreviation used for the accurate mass and time tag approach (Pasa-Tolic *et al.*, 2004)

2.2 Experimental Procedures

2.2.1 Sample preparation

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 20 mM iodoacetamide (Sigma) at room temperature for 30 min. 10 μg of total protein was separated on a NuPAGE Novex 4-12% Bis-Tris polyacrylamide gel (Invitrogen). The gel was stained with colloidal Coomassie Blue (Sigma). The entire gel lane was excised into 48 bands, de-stained 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. This work was carried out as part of my two month web-lab rotation and was guided by Mercedes Pardo (Team 17 at the Wellcome Trust Sanger Institute).

Sample 2: A standard protein set of 48 human proteins (Sigma, Universal Proteomics Standard Set UPS1) was reduced with Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), alkylated with iodoacetamide as above, followed by digestion in solution with sequencing grade trypsin (Roche Applied Science) overnight. To minimise 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. This work was carried out by Lu Yu (Team 17, Wellcome Trust Sanger Institute).

2.2.2 LC-MS/MS analysis

Peptides were analysed with on-line nanoLC-MS/MS on a LTQ FT (Thermo Fisher Scientific), a hybrid linear ion trap and a 7 Tesla 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 id x 5 mm) at 20 μL/min with 0.1% formic acid for 5 min then separated on an analytical column (75 μm id x 15 cm) (both PepMap C18, LC Packings) over a 30 min linear gradient of 4-40% CH_3CN/0.1% formic acid. The flow rate through the column was 300 nL/min. The LTQ FT mass spectrometer was operated in standard data dependent mode controlled by Xcalibur 1.4 software. The survey scans (m/z 400-2000) were acquired on the FT-ICR at a resolution of 100,000 at m/z 400 and one microscan was acquired per spectrum. The top three (top five for sample 2) most abundant multiply charged ions with a minimal intensity at 1000 counts were subject to MS/MS in the linear ion trap at an isolation width of 3 Th.

Precursor activation was performed with an activation time of 30 msec and the activation Q was set at 0.25. The normalised collision energy was set at 35%. The dynamic exclusion width was set at $\pm$5 ppm with 2 repeats and a duration of 30 sec. To achieve high mass accuracy, the automatic gain control (AGC) target value was regulated at 4E5 for FT and 1E4 for the ion trap, with a maximum injection time of 1000 ms for FT, and 100 msec for ion trap respectively. The instrument was externally calibrated using the standard calibration mixture of caffeine, MRFA and Ultramark 1600.

All LC and MS related work was carried out by Lu Yu (Team 17, Wellcome Trust Sanger Institute) and was used to introduce me to the basics of practical mass spectrometry during my wet-lab rotation project.

2.2.3 Raw data processing

LTQ FT MS raw data files were processed to peak lists with BioWorks 3.2 (Thermo Fisher Scientific). Parameters were as follows: precursor masses were set to 800-4500 Da, grouping was enabled allowing 50 intermediate scans, and a precursor mass tolerance setting of 10 ppm in BioWorks was applied. The number of minimum scans

per group was set to 1. For sample 2 grouping was disabled.

RAW data, peak lists (with and without mass error correction) and Mascot results for both samples are available through ftp under the address: `ftp://ftp.sanger.ac.uk/pub/mb8/mcp2008/`

2.2.4 Database search parameters

Sample 1: Mascot 2.1 (Matrix Science, London, UK) and X!Tandem 2007.07.01 (The Global Proteome Machine Organization) were used for analysing the data. Parameters used in Mascot and X!Tandem searches were: enzyme = trypsin; variable modifications = carbamidomethylation of cysteine, oxidation of methionine; maximum missed cleavages = 1; peptide mass tolerance settings/windows were as indicated in the individual experiments (between 2 Da and 5 ppm); product mass tolerance = 0.5 Da. Probability p of random matches for MIT calculations in Mascot was set to the default value of 0.05.

Specific X!Tandem parameters were: spectrum dynamic range was set to 1000, refinement was disabled, maximum valid E-value for reported peptides was set to 100 (E-values were limited in the data analysis steps) and cyclic permutations to compensate for small search spaces was enabled, with remaining parameters at default.

The protein sequence database used by Mascot and X!Tandem was built from an non-identical superset of Ensembl peptides, UniProtKB and RefSeq sequences for *Mus musculus*, including common external contaminants from cRAP (a maintained list of contaminants, laboratory proteins and protein standards provided through the Global Proteome Machine Organization, `http://www.thegpm.org/crap/index.html`) and contains 94,524 sequences and 42,765,694 residues. For false positive discovery assessment, a separate decoy database was generated from the target database using the Perl decoy.pl script provided by MatrixScience. This script randomises each entry, but retains the average amino acid composition and length of the entries. 0.1%

of sequences were common in both target and decoy database, including K/Q and L/I isoforms that are indistinguishable above 0.04 Da MMD.

Peak lists of sample 2 (8,190 spectra) were searched with Mascot and X!Tandem against human IPI (June 2007, 68,322 sequences, 28,806,780 residues) including common external contaminants from cRAP. To minimise unexpected contaminants from the protein standard set (Klimek *et al.*, 2007), a very low concentration of 10 fmol was used. Parameters used: enzyme = trypsin; variable modifications = carbamidomethylation of cysteine, oxidation of methionine and deamidation of aspargine and glutamine; maximum missed cleavages = 2; peptide mass tolerance = 1 Da; product mass tolerance = 0.5 Da. A random and a reversed version of the sequence database was generated and searched under the same conditions.

2.2.5 Data analysis

Mascot results ($p < 1.0$) were exported to pepXML using the Mascot export tool and X!Tandem results (E-value < 100) were stored as X!Tandem XML. An in-house Java tool was used for the data analysis. Results from Mascot and X!Tandem were imported and filters on score thresholds and mass tolerances were applied. Only doubly and triply charged ions and the first hit rank per spectrum were considered for analysis.

For FDR estimation I chose to search the target and decoy database separately to avoid affecting the MIT scoring by changing database size. The decoy database used was a randomised version of the target database, which was found to be the best approximation based on evaluations of sample 2 (see figure 2.3). All estimated FDRs in this work were calculated using the same target/decoy approach, enabling consistent comparison of results.

Estimated FDRs were calculated by counting all peptide assignments obtained from the decoy database (proxy for false positives, FP), divided by the number of peptide assignments that were obtained from the target database (TP+FP), given

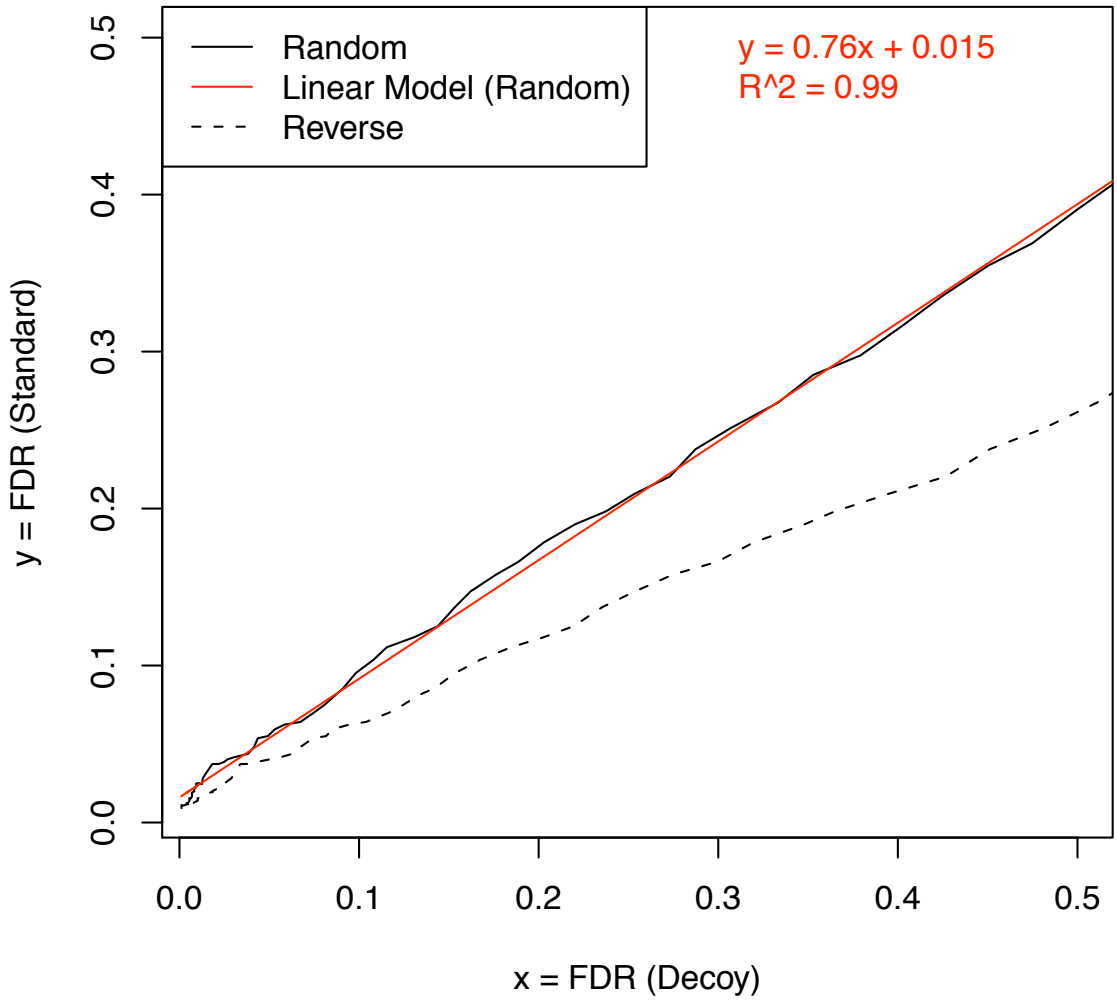


Figure 2.3: An experimental FDR, based on the known proteins of the set, can be determined as follows: any peptide hit that did not match against any of the 48 standard proteins or any of the external contaminants was considered a false positive hit. The FDR rates were determined for a range of Mascot score cutoffs (10-50). Similarly, the estimated FDRs based on target/decoy searching were determined for both the randomised and reversed database. This enabled a comparison of actual FDRs with estimated FDRs, which is interesting since there is no consensus in the proteomics community concerning the different decoy strategies (discussed in section 1.1.2.3). Nevertheless, both decoy strategies (randomized/reversed) tested in this work show a linear relationship between the FDR determined by the protein standard and the target/decoy estimation, validating the target/decoy approach. However, the FDRs derived by the random database were closer to what was reported by the protein standard, which let me to chose the random database as a decoy database for this study. The linear regression of the random database ($R^2 = 0.99$) indicates a small offset of 1.5% which can be explained by unexpected contaminations in the protein standard.

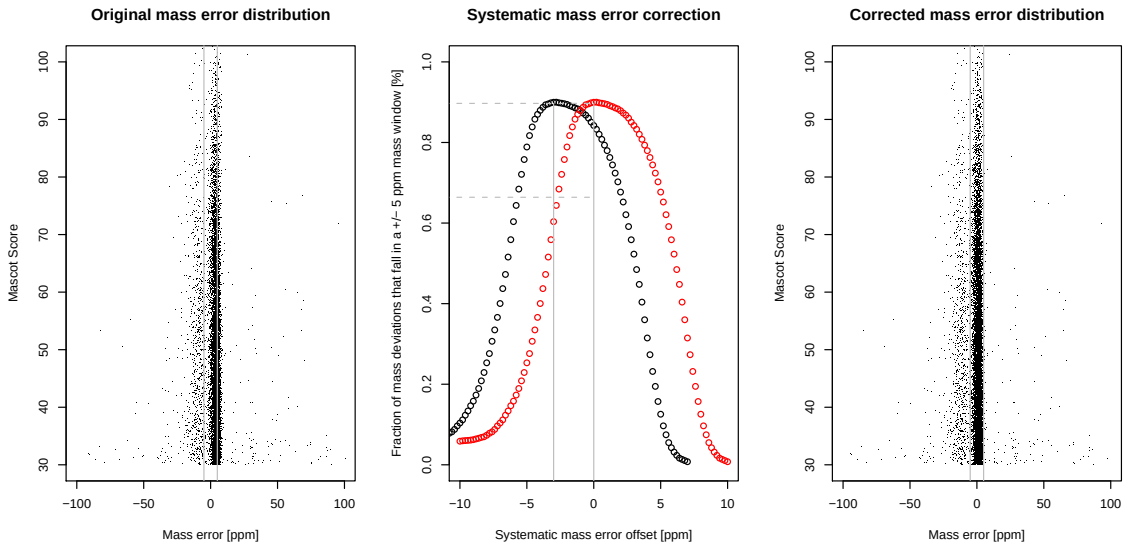
the same parameter and threshold settings. The estimated number of true positive hits (TP) was calculated by counting the number of all peptide hits against the target database minus all hits against the decoy database search. FDR assessment was limited to the peptide level only, since I was interested in the quality of matching individual spectra to peptide sequences. Furthermore it avoids comparison of protein inference strategies (Nesvizhskii and Aebersold, 2004), which is a separate issue.

2.2.6 Correction of systematic mass error

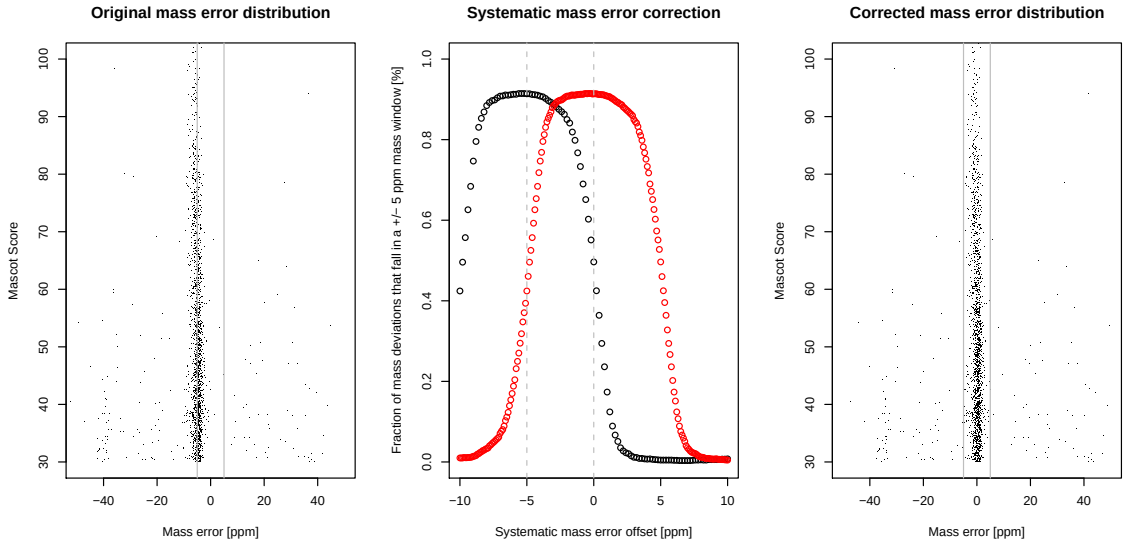
Data from sample 1 was searched in a first pass with Mascot at 100 ppm MMD in order to determine the mass accuracy for the experiment. Only peptide hits with a Mascot score greater than 30 were used for the mass accuracy assessment (10634 queries) to exclude mass deviations of incorrect matches. 99% of hits had mass deviations within a (3±20) ppm mass window (systematic mass error ± peptide mass error), while 90% of mass deviations fell within (3±5) ppm. In order to allow the best possible mass tolerance settings of (0±5) ppm in Mascot and X!Tandem, the precursor masses were corrected by 3 ppm (figure 2.4a). A similar mass error correction method was described by Zubarev and Mann (2007). The mass outliers between -5 and -20 ppm seem to be an experimental artefact for this particular sample. For this study I deliberately accepted a loss of identifications for 5 ppm MMD settings in order to study the effects of stringent mass settings on Mascot and X!Tandem. Mass error correction was applied in the same way to sample 2, where the peptide masses were corrected by 5 ppm (figure 2.4b).

2.3 Results and discussion

If not stated otherwise, all subsequent results are based on sample 1, which is a large complex dataset and representative of typical proteomics experiments.



(a) Sample 1



(b) Sample 2

Figure 2.4: Mass error determination and correction of systematic mass errors. Left: the original mass deviations of all highly significant peptide matches. Centre: Systematic mass error correction that maximises the peptide assignments within a 5 ppm mass window. Right: After correction of the systematic mass error.

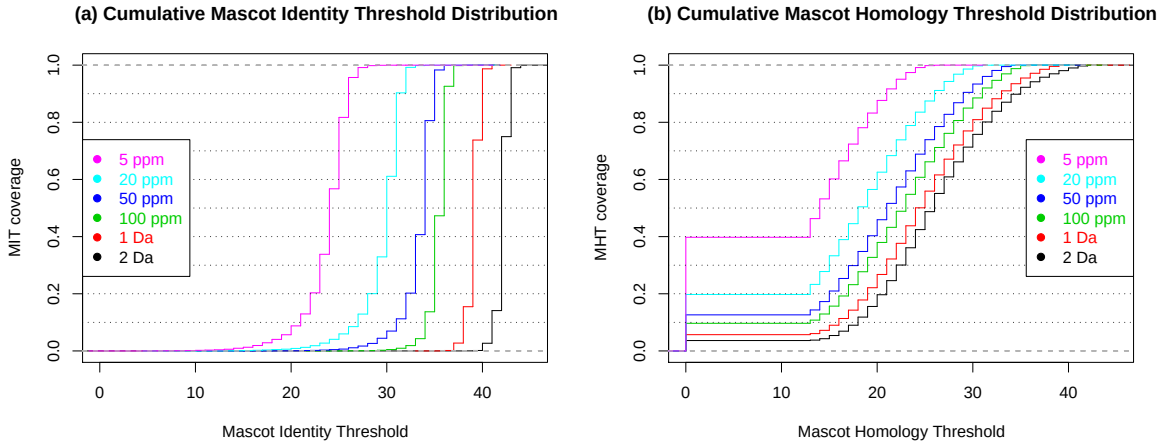


Figure 2.5: (a) Cumulative MIT distributions for different peptide mass tolerance settings. Only MITs from queries with a peptide assignment across all searches were used to enable comparison. With more stringent MMD settings, the MIT tends to decrease, accommodating for the smaller search space. *Vice versa* it increases for more relaxed MMD windows. (b) Cumulative MHT distributions over the range of MMD settings. The MHT is not reported for every query. All MHTs exceeding the MIT are omitted by Mascot and reported as 0 in the HTML and XML result files (personal communication, John Cottrell, Matrix Science). The minimum MHT reported by Mascot is 13 and the maximum MHT is limited by the corresponding MIT.

2.3.1 Performance of the Mascot Identity Threshold

Mass error corrected spectra were submitted to Mascot and searched at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm MMD settings, while all other parameters were fixed.

Spectra that were assigned across all searches (23,080 out of 38,058 queries) were used to draw the MIT distribution for each MMD setting (Figure 2.5a). From this analysis the median MIT values for relaxed MMD settings were 42 at 2 Da MMD and 39 at 1 Da MMD with an inter-quartile range of 1. Under more stringent settings (5 ppm) the MIT median decreased to 24 while the inter-quartile range increased to 2. These results suggest that the MIT adapts with changing search space and performs more like a global cut-off based on the narrow variation in thresholds.

To evaluate the effects of MIT adaptions on the peptide identifications performance at different MMD settings, the rates of incorrect and correct peptide-spectrum

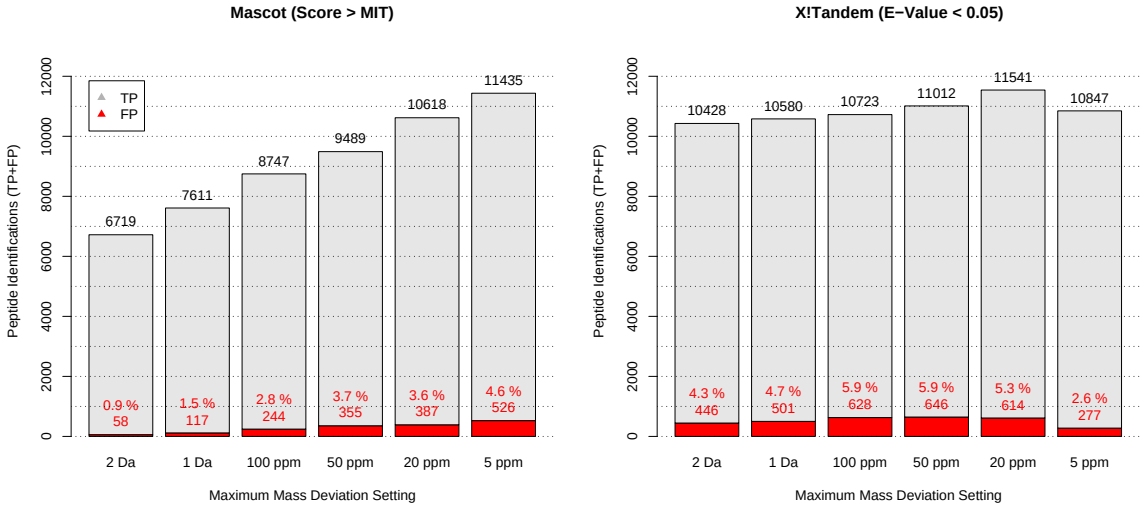


Figure 2.6: Comparative evaluation of Mascot and X!Tandem performance. Mascot and X!Tandem searches were performed against a target and decoy database at different MMD settings. The total number of identifications is reported, the estimated number of true identifications is indicated in grey, while the estimated number of incorrect assignments is highlighted in red.

matches were determined by target/decoy FDR estimations, under identical search and threshold parameters for all spectra (Figure 2.6, Mascot). Using the MIT as a score cut-off, 10,909 and 6,661 estimated TP peptide identifications were made at 2 Da and 5 ppm MMD settings respectively. Relative to the 5 ppm search, this suggests 4,248 (39%) false negative peptide assignments for the 2 Da search. For the same MMD settings, the FDR increased from 0.9% to 4.6% respectively, failing to maintain the specified (5%) rate of random (incorrect) assignments.

The MIT is based on a probabilistic model that attempts to maintain a constant rate of random (false) identifications and hence is dependent on search space. However, I found a correlation between FDRs and MMD settings, indicating that the MIT does not adhere to the predefined FDR. This trend is also mirrored in the number of correct identifications. At relaxed mass tolerances (large search space) used for ion trap data, the MIT tends to become very conservative resulting in excellent specificity but hindering sensitivity. With more stringent mass tolerances (smaller search space) sensitivity increases at the cost of specificity. The results reported

here represent a snapshot of many possible combinations of search parameters that directly affect the search space, for example: sequence database size, allowed variable modifications, allowed missed cleavages and enzyme specificity. This highlights the necessity to individually assess the FDR via a target/decoy database search.

2.3.2 Performance of the X!Tandem scoring scheme

Spectra were searched in X!Tandem using MMD settings as described in the previous section. FDRs were calculated on the basis of target and decoy database searches using identical search parameters.

Using an E-value cut-off value of 0.05, which is in-line with that used for the MIT evaluation discussed above, only moderate changes (9%) in sensitivity over all MMD settings were detected, varying between 9,982 TPs at 2 Da and up to 10,927 TP at 20 ppm (Figure 2.6, X!Tandem). A constant FDR for varying MMD settings was not delivered by X!Tandem. The FDRs increased from 4.3% to 5.9% between the 2 Da and 100 ppm MMD, and an inverse trend was observed below 100 ppm, with a minimum of 2.6% FDR at 5 ppm MMD. FDRs show no clear correlation with mass tolerance settings, suggesting no direct dependency. The E-Value distributions of these searches were very similar over the whole range, further supporting the robustness of the X!Tandem scoring (figure 2.7).

Overall, X!Tandem appears to maintain sensitive peptide identification at varying MMD settings. The FDRs were close to the defined E-Values, but were not constant over changing mass tolerance settings. However, there appears to be no direct correlation between the FDRs and search space. These results indicate that the empirical X!Tandem scoring, based on peptide-spectra match score distributions, is more robust over the search space dependent probabilistic scoring model of the MIT.

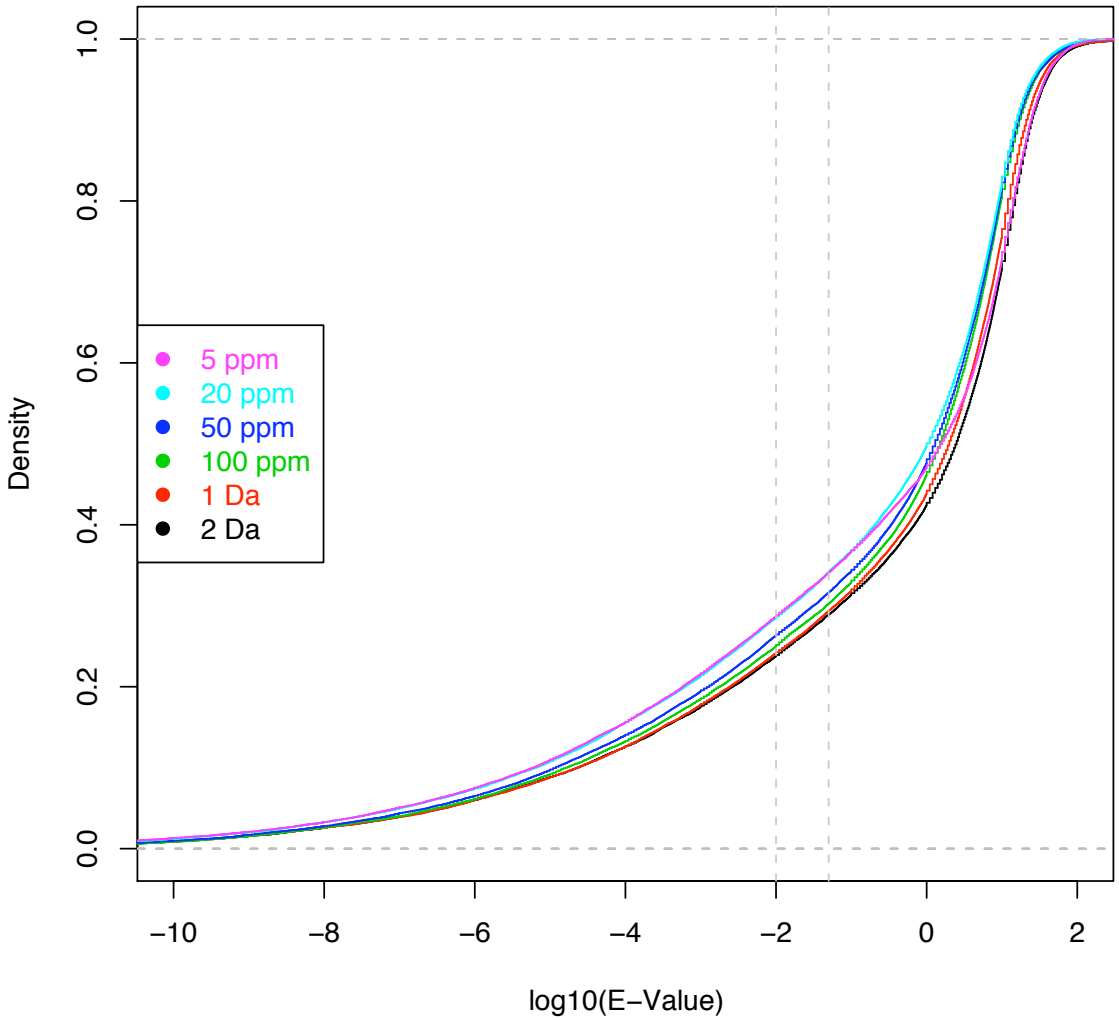


Figure 2.7: Spectra were searched with X!Tandem at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm MMD settings, while all other parameters were fixed. For each search the E-value distribution was drawn, indicating that the X!Tandem scoring is very robust over changes in search space. The E-values 0.01 and 0.05 are highlighted. The plot is in concordance with the ROC curve presented in the paper. Personal communication with Dr David Fenyo (The Rockefeller University) explained the robustness of the E-value distributions: Each E-value depends on the survival function and on the number of sequences scored (Fenyo and Beavis (2003), equation 2). For X!Tandem in its current format, the term "number of sequences scored" refers to the whole sequence database, regardless of the peptide mass tolerance setting and hence all variations seen in the E-values are the result of the slight differences in survival functions only.

2.3.3 Performance of the Mascot Homology Threshold

Similarly to X!Tandem, the empirical MHT also utilises peptide-spectra match score distributions. Using the results from the above Mascot searches, I plotted MHT distributions at different MMD settings (Figure 2.5b). Only spectra that were assigned across all searches (23,080 out of 38,058 queries) were used for comparison.

As stated earlier, the MHT is not always reported. A MHT value was reported for about 95% of the considered queries at relaxed MMD settings of 1 or 2 Da. For stringent MMD settings (5 ppm), MHTs were only reported for less than 60% of queries, limiting its applicability. The MHT median for a 1 Da MMD setting was 24, compared with a MIT median of 39 for the same setting, while the inter-quartile ranges were 9 and 1, respectively. The wide MHT variation observed would be reflective of a query specific thresholding.

Using the MHT as a cut-off score for a 1 Da MMD search, 11,315 TPs were identified at the given FDR of 3.1%. This corresponds to 51% more TP identifications than using the MIT at the given 1.5% FDR and 12% more TP identifications than X!Tandem at the given FDR of 4.7%.

Overall, I observe the MHT to be significantly more sensitive than the MIT and X!Tandem at the given FDRs. However, the FDR is pre-imposed and does not allow the user to select a fixed rate. Furthermore, Mascot omits any MHT which exceeds the MIT to prevent conservative thresholds that arise, for example from score distributions with insufficient data points. This effect is further compounded since the MIT values decrease for a smaller search space. Sufficient search space is required for the MHT to be comprehensively applicable, for example a larger or smaller database would need a more or less restrictive MMD setting to compensate for this effect.

2.3.4 Peptide mass accuracy filtering

An alternative approach for using high mass accuracy for peptide identification is to search under relaxed mass tolerance settings and subsequently apply mass accuracy filters. To evaluate this approach, data was searched with Mascot at a 1 Da MMD setting against the target and decoy databases, where approximately 95% of queries obtained a MHT.

As shown in figure 2.2, peptide-spectrum matches with high scores mostly lie within the experimental mass errors discussed previously, while low scoring matches were distributed evenly across the whole mass window. Mass accuracy filtering of the 1 Da search using 50, 20 and 5 ppm cut-offs, without imposing any other constraints such as MIT or MHT, limits the FDRs to 65%, 35% and 12% respectively. This clearly indicates that mass accuracy based filtering alone can reduce incorrect sequence assignments. However, the effectiveness of this discriminator is confined by experimentally derived mass error deviations. Significantly, 13,273 TP were identified with a 5 ppm mass filter, more than obtained by any method tested here, showing this to be a very sensitive approach for peptide identification with high accuracy data.

The 12% FDR observed at 5 ppm mass accuracy filtering suggests that even higher mass accuracy would be required for lower FDRs. An extrapolation from a regression over 10 data points ranging from 5 to 50 ppm ($r^2 = 0.99$) suggests a 5% FDR for 1.5 ppm, however this prediction would need to be verified experimentally. It should be noted that the use of ultra high mass accuracies cannot further improve FDR once mass accuracies resolve elemental compositions.

If mass accuracies cannot be achieved at this stringent level, an alternative would be to introduce a moderate thresholding on the peptide-spectrum match scores. I therefore tested mass accuracy filtering in combination with the MHT score cut-off. For this, data was searched at 1 Da MMD, then filtered at 5 ppm to exclude all peptide assignments with a larger mass deviation, and subsequently constrained by

the MHT. In instances where the MHT was not reported, the MIT was used. This two-step filtering identified 10,338 TP peptide assignments and reduced the FDR to only 0.2%, which is a 60-fold improvement over the mass accuracy filtering alone, although the TPs were reduced significantly (22%). In comparison with the Mascot search using 5 ppm MMD setting with the MIT score cut-off, where a FDR of 4.8% and 10,909 TP was previously reported, the two-step filtering improved the FDR by 23-fold, while the TPs were reduced by only 6%.

These results suggest that mass accuracy filtering on its own might be a valuable and very sensitive approach, however sub-ppm mass errors would be needed for highly specific identification. Alternatively, a combination with a threshold such as the MHT serves as a very strong discriminator between correct and incorrect peptide assignments. In comparison with a direct high accuracy Mascot search, the two-step filtering strategy leads to highly specific identifications without significantly compromising sensitivity. A less restrictive and adjustable thresholding would increase sensitivity for peptide identification from high accuracy data.

2.3.5 The Adjusted Mascot Threshold (AMT)

Applying either the MIT, MHT or the two-step filtering provides pre-imposed FDRs that are not directly adjustable by the user. However, it is often desirable to be able to select and fix the FDR.

To achieve this I have implemented the Adjusted Mascot Threshold (AMT). This is a similar strategy to the MATH threshold introduced by Rudnick *et al.* (2005), which uses a global threshold that defines a cut-off value for all queries. However, I favour the use of individual query specific thresholds based on the MHT, since I have found it to be very sensitive in my above evaluations. The AMT is defined as the sum of the query specific MHT and a global offset value. FDRs are determined for a range of offset values that are used to calculate a linear regression in order to approximate an offset value for a user defined FDR (Figure 2.8).

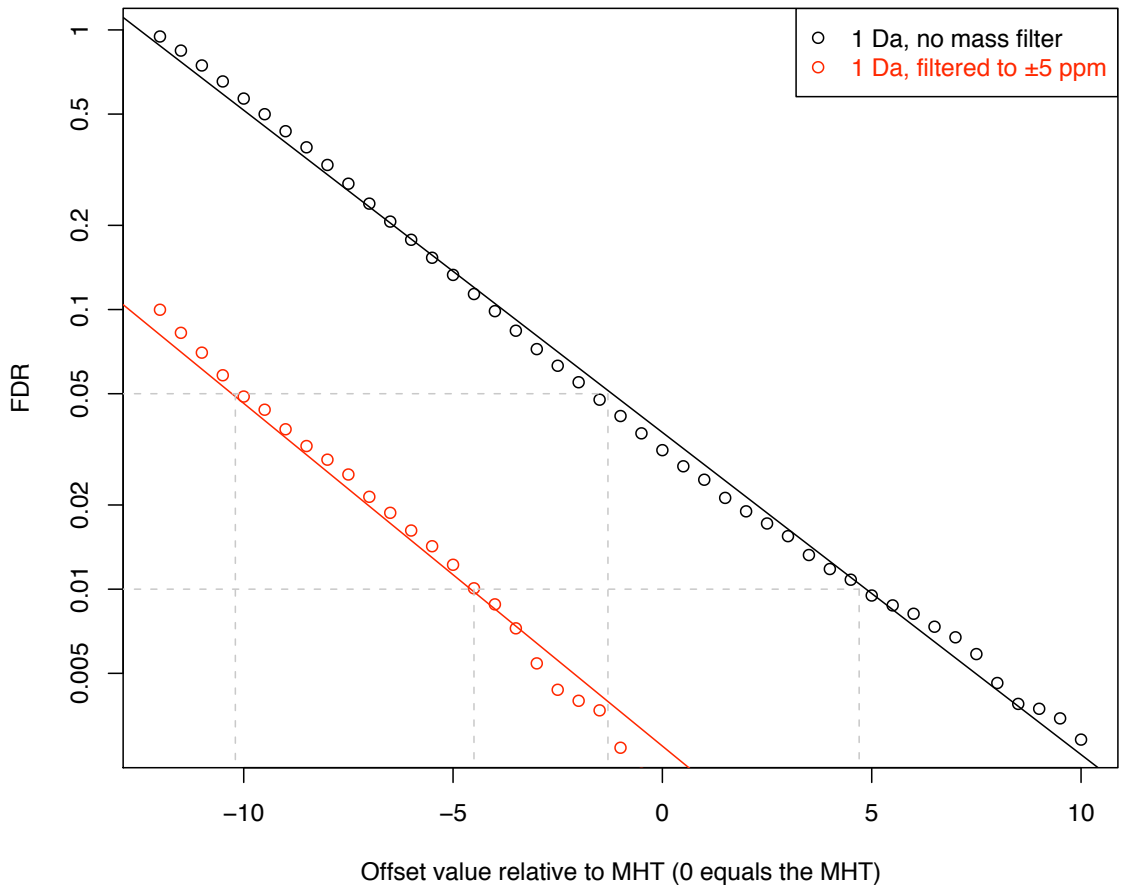


Figure 2.8: Regression for extrapolating the AMT thresholds. Data was searched at a 1 Da MMD setting against the target and decoy database. A range of offset values was applied that were added to the MHT and used as cut-off thresholds. For each new threshold the associated FDR was determined. A linear regression between the logarithm of the FDR and the offset values was calculated (r^2=0.99). The method was also applied to the mass accuracy filtered dataset (5 ppm). A new Adjusted Mascot Threshold can be extrapolated based on a user defined FDR for each dataset. The AMT adapts for the preceding mass accuracy filtering. The offset values for a FDR value of 1% and 5% are indicated as dashed lines.

For the 1 Da search, described in the previous section, the regression was calculated for an offset range of -12 to +10, indicating a strong linear correlation between the logarithm of FDRs and the offset values with a correlation coefficient of $r^2 = 0.99$. For the 5 ppm mass filtered dataset, a second regression was calculated ($r^2 = 0.99$). Offset values of 4.7 and -1.3 were reported for a target FDR of 1% and 5% using the 1 Da search data and for the 5 ppm mass accuracy filtered dataset these values were -4.5 and -10.2 respectively. The slope of both regressions was found to be very similar, but the difference between the offsets was approximately -9, which compensates for the inherent specificity of the mass accuracy filtered dataset by moderating these offset values.

Our proposed AMT is an adjustable and query specific cut-off value. It is calculated based on the MHT and a global offset value, the latter is derived from FDR estimates through target/decoy database searching and thus is no longer dependent on search parameters affecting search space. AMT can be extrapolated for either low or high accuracy (using mass filtered data), and combines the benefits of a highly sensitive MHT with a user defined FDR.

2.3.6 Comparison of the AMT with MIT, MHT, MATH and X!Tandem

I then tested the performance of the AMT. Search results obtained by application of AMT were compared to those from MIT, MHT, X!Tandem and MATH using a receiver operator characteristic (ROC) representing the number of true identifications at various FDRs. ROC curves (Figure 2.9) were calculated using varying thresholds of MATH (global cut-off value), X!Tandem (E-values) and AMT (offset values relative to MHT). Since the MIT and MHT are not variable, they define a single point in the diagram.

For low accuracy MMD settings (Figure 2.9a) applying the MIT identified 7,494

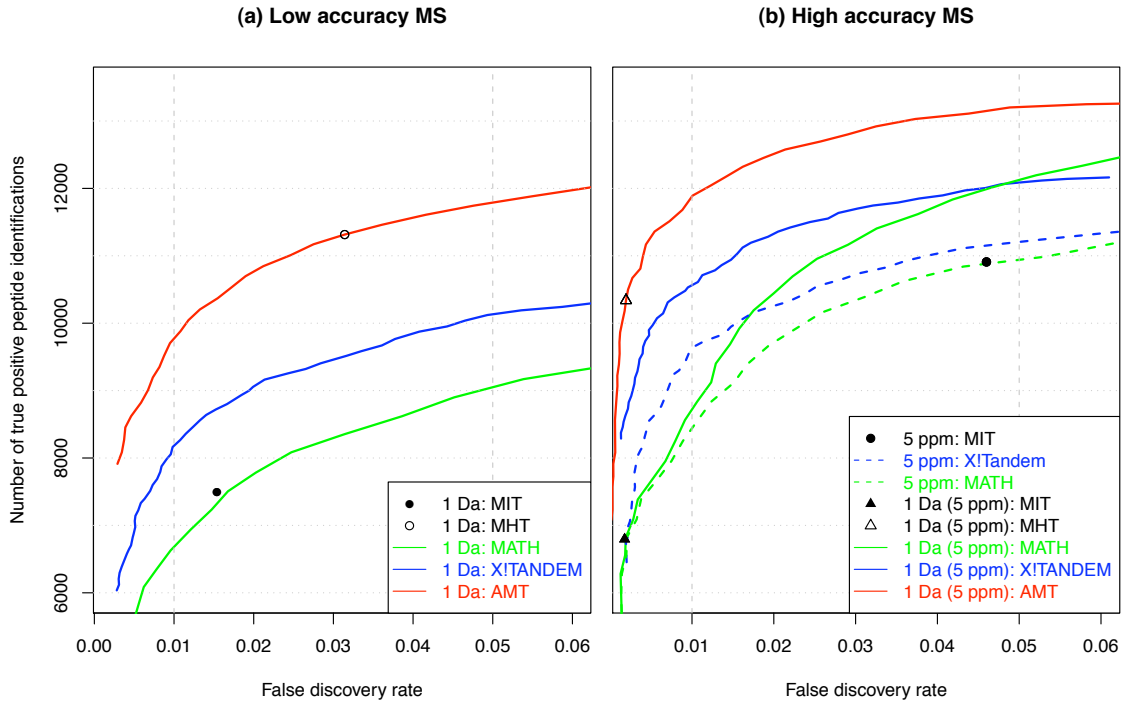


Figure 2.9: MIT, MHT, MATH, X!Tandem and AMT comparison for low and high accuracy mass tolerance settings. A 1 Da search (a), a 5 ppm search (b, dashed lines) and a 1 Da search with subsequent peptide mass accuracy filtering at 5 ppm (b, solid lines) were performed. The estimated number of TPs was determined as a function of the FDRs, represented in the receiver operator curve, enabling the user to choose where the best trade-off between sensitivity (TPs) and specificity (FDR).

TP with an inherent 1.5% FDR. MIT variation for these mass tolerance settings effectively acts as a global cut-off, hence MATH also identified a similar number at the same FDR. MATH however allows the user to freely select the target FDR, and at a 5% FDR it identified about 20% more TP peptides than at 1.5% FDR. X!Tandem empirical scoring outperformed both MIT (13% more TP at the same FDR of 1.5%) and MATH (between 10-15% more TP over the whole range of FDRs). The most striking observation is the MHT performance, identifying 11,315 TPs at the inherent FDR of 3.1%, improving correct identifications by 18% and 35% over X!Tandem and MATH at the same FDR. The AMT extends application of the MHT over the whole range of FDRs, improving the TP assignments by 18%, 39% and 42% over X!Tandem, MIT and MATH at 1.5% FDR, and by 16% and 30% over

X!Tandem and MATH at 5% FDR.

For the analysis of high accuracy data I have evaluated two strategies (I) searching high accuracy data at stringent mass tolerance settings (5 ppm) followed by peptide score thresholding (Figure 2.9b, dashed lines), and (II) searching high accuracy data at a relaxed mass window (1 Da) with subsequent peptide mass accuracy filtering (5 ppm) followed by peptide score thresholding (Figure 2.9b, solid lines).

(I) Using direct high mass accuracy searching at 5 ppm MMD setting, the number of expected true peptide identifications was similar, approximately 11,000, for MIT, X!Tandem and MATH at around 4.5% FDR. However, X!Tandem performed better for lower FDRs, e.g. at 1% X!Tandem identified about 1,000 more TPs than MATH. MHT was not assessed at these mass tolerance settings since it was absent for 40% of queries.

(II) The alternative mass filtering approach returned very conservative FDRs below 0.2% and identified 6,798 and 10,338 TP hits for the MIT and MHT respectively. Mass filtered X!Tandem results identified approximately 25% more peptides than the MIT and 18% less TP hits than with the MHT, at the corresponding FDRs. By relaxing the E-values of X!Tandem, 10,611 TP at 1% FDR and 12,100 TP at 5% FDR were identified. Using MATH, 6,821 TP assignments were made at the 0.2% FDR, which is again similar to MIT and significantly worse than X!Tandem or MHT. At a 1% FDR about 18% fewer identifications were made using MATH as compared to X!Tandem, while they performed similarly at 5% FDR. Significantly, the AMT identified 11,893 TP assignments at 1%, outperforming both MATH and X!Tandem by 35% and 12% at the same FDR.

Compared to the direct 5 ppm search strategy in (I), the mass accuracy filter approach in (II) was generally more sensitive, e.g. MATH and X!Tandem with mass filtering identified about 8-9% more TPs at 5% FDR than without mass filtering. The improvement of performance with X!Tandem can be seen throughout the whole range of FDRs, whereas for MATH sensitivity is only gained above a 1% FDR. By far the

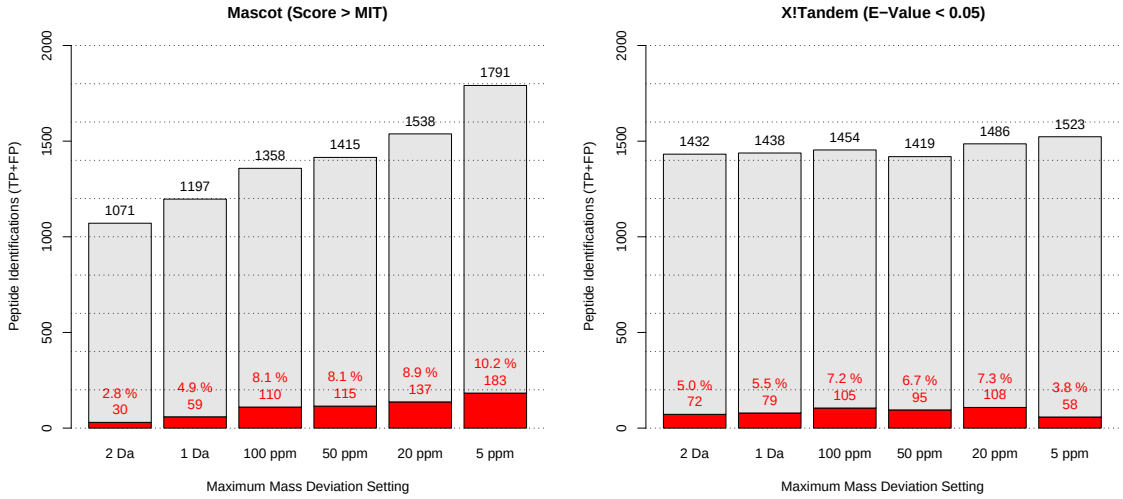
most sensitive approach at any given FDR was provided by mass accuracy filtering combined with the AMT. Against a direct 5 ppm search using MIT, MATH and X!Tandem, about 18-20% more TPs at a FDR of 4.6% were made, which corresponds to approximately 1,500 more unique peptides identifications.

In summary, application of MIT or MHT always results in a fixed pre-imposed FDR, while X!Tandem together with a target/decoy database search enables FDR adjustment using an appropriate E-value cut-off. MATH and AMT implement this target/decoy FDR estimation and directly deliver the defined FDRs. For low accuracy MS, MHT performed best at a fixed FDR, whilst this performance was extended to the whole FDR range by AMT. X!Tandem was significantly less sensitive than AMT, and MATH together with the MIT were the least sensitive thresholds. For direct high mass accuracy searching, MIT, MATH and X!Tandem performance was very similar and overall sensitivity improved over the low accuracy search. Exploiting high mass accuracy via mass filtering was the most sensitive search strategy at the corresponding FDRs. For this approach, AMT significantly outperformed X!Tandem, followed by MATH and MIT.

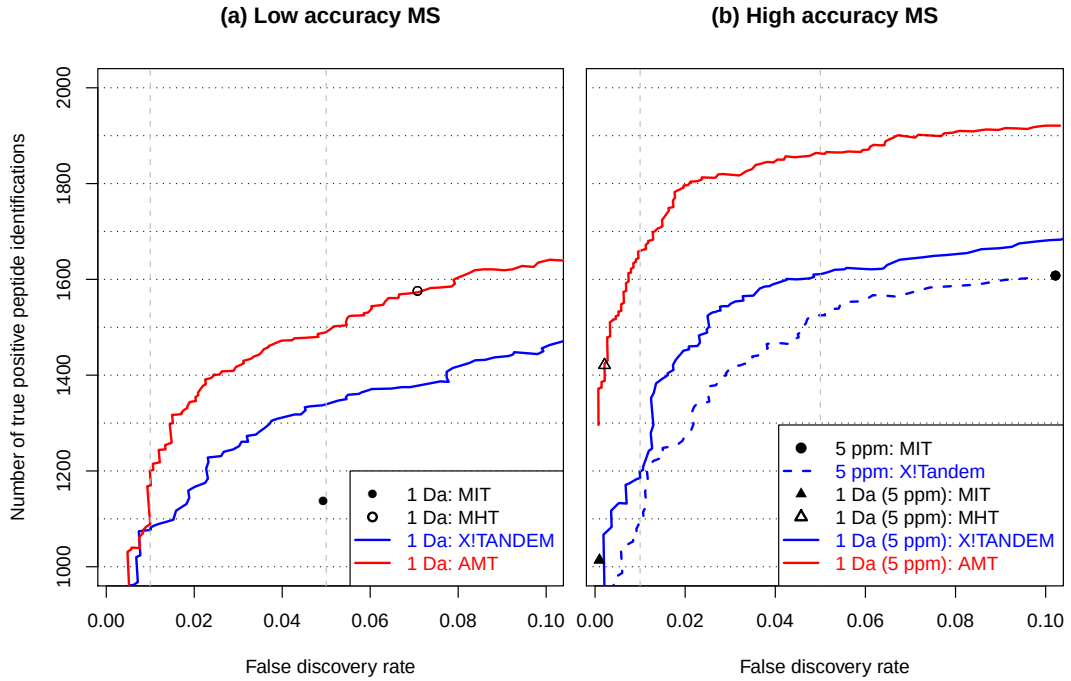
2.3.7 Validation with independent dataset

To validate the findings and the AMT performance, a standard mixture of 48 proteins (sample 2) was analysed in the same way as sample 1. First, data were searched against a 50 ppm peptide mass tolerance to identify any systematic mass error (Figure 2.4b), which was corrected (-5 ppm) subsequently.

Next, data were searched at 2 Da, 1 Da, 100 ppm, 50 ppm, 20 ppm and 5 ppm peptide mass tolerances and the FDRs were determined accordingly (figure 2.10a). The same FDR trends as for sample 1 (figure 2.5) were observed: using the MIT resulted in the FDR being dependent on the search parameters used, rising from 2.8% to over 10% when the mass tolerance window was narrowed from 2 Da to 5 ppm. However, X!Tandem was shown to be quite robust again, indicating little



(a) Comparative evaluation of Mascot and X!Tandem performance and FDR robustness for sample 2. Compare with figure 2.6.



(b) Comparison of the performance of X!Tandem and Mascot using the MIT, MHT and AMT for sample 2. Compare with figure 2.9. The vertical dashed lines correspond to commonly used 1% and 5% FDR values.

Figure 2.10: Validation of results on an independent protein standard dataset.

dependance of the search space on the scoring scheme.

ROC curves were compiled to enable comparison (Figure 2.10b) of the AMT and the standard Mascot thresholds as well as the X!Tandem performance. Again, the MHT was shown to be significantly more sensitive than the MIT, but the AMT scoring method clearly outperformed the MIT and MHT as well as X!Tandem, validating the findings of sample 1.

2.4 Conclusion

In this chapter I have investigated how MMD settings affect peptide identification using Mascot and X!Tandem and presented an alternative search strategy and an Adjusted Mascot Threshold (AMT) to enable sensitive identification of high accuracy data with Mascot.

I have demonstrated the correlation between the MIT and search space, which is for example affected by MMD settings. I have shown that the MIT can be very conservative for MMD settings commonly used for ion trap data, leading to very specific identifications at the expense of sensitivity, while it tends to become more optimistic for stringent MMD settings used for high accuracy data. The MHT was found to be significantly more sensitive for ion trap data, but is not comprehensively applicable to very stringent MMD settings commonly used for high accuracy data. However, the actual FDRs for both MIT and MHT are pre-imposed and deviate from the theoretically defined rate. Furthermore, my results indicate that X!Tandem is more robust than the MIT and MHT when faced with MMD changes and is equally applicable to both low and high accuracy MS data with a sensitivity that was better than using the MIT but worse than using the MHT.

I also investigated the use of mass accuracy filtering as the sole discriminator between correct and incorrect peptide assignments. Mass accuracy filtering served as a highly sensitive discriminator with limited specificity and sub-ppm mass errors

would be needed for more specific identifications. Alternatively, a two-step filtering strategy can be employed. I first searched the data at relaxed MMD settings, followed by applying mass accuracy filtering. The results demonstrate that combining peptide mass accuracy filtering with the MHT serves as a very strong discriminator, efficiently eliminating incorrect peptide assignments, although sensitivity was limited. To regain sensitivity I propose an Adjusted Mascot Threshold (AMT) that allows the user to freely select the best trade-off between sensitivity and specificity by having full control over the actual FDR. The AMT can easily be applied on top of any Mascot search where target/decoy searching is amenable. It is independent of search parameters affecting the search space and is expected to adjust with MS/MS data quality. AMT outperforms MIT and MHT, as well as MATH and X!Tandem for both low and high accuracy MS data.