Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

REGULAR ARTICLE

An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis

Eugene A. Kapp[1], Frédéric Schütz[1]*, Lisa M. Connolly[1], John A. Chakel[2], Jose E. Meza[2], Christine A. Miller[2], David Fenyo[3], Jimmy K. Eng[4], Joshua N. Adkins[5], Gilbert S. Omenn[6] and Richard J. Simpson[1]*

[1] Joint ProteomicS Laboratory, Ludwig Institute for Cancer Research (Melbourne Branch)/Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
[2] Agilent Technologies, Santa Clara, CA, USA
[3] GE Healthcare, Bio-Sciences, Piscataway, NJ, USA
[4] Institute for Systems Biology, Seattle, WA, USA
[5] Pacific Northwest National Laboratory, Richland, WA, USA
[6] University of Michigan Medical School, Ann Arbor, MI, USA

Received: March 3, 2005
Revised: April 22, 2005
Accepted: May 20, 2005

MS/MS and associated database search algorithms are essential proteomic tools for identifying peptides. Due to their widespread use, it is now time to perform a systematic analysis of the various algorithms currently in use. Using blood specimens used in the HUPO Plasma Proteome Project, we have evaluated five search algorithms with respect to their sensitivity and specificity, and have also accurately benchmarked them based on specified false-positive (FP) rates. Spectrum Mill and SEQUEST performed well in terms of sensitivity, but were inferior to MASCOT, X!Tandem, and Sonar in terms of specificity. Overall, MASCOT, a probabilistic search algorithm, correctly identified most peptides based on a specified FP rate. The rescoring algorithm, PeptideProphet, enhanced the overall performance of the SEQUEST algorithm, as well as provided predictable FP error rates. Ideally, score thresholds should be calculated for each peptide spectrum or minimally, derived from a reversed-sequence search as demonstrated in this study based on a validated data set. The availability of open-source search algorithms, such as X!Tandem, makes it feasible to further improve the validation process (manual or automatic) on the basis of "consensus scoring", *i.e.*, the use of multiple (at least two) search algorithms to reduce the number of FPs.

Keywords:
MASCOT / Mass Spectrometry / PeptideProphet / SEQUEST / Sonar / Spectrum Mill / X!Tandem

Correspondence: Professor Richard J. Simpson, Joint ProteomicS Laboratory, Ludwig Institute for Cancer Research, P.O. Box 2008, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia
E-mail: richard.simpson@ludwig.edu.au
Fax: +61-3-9341-3192

Abbreviations: FP, false-positive rate; **IPI**, International Protein Index; **NR**, nonredundant protein sequence database; **PPP**, Plasma Proteome Project; **TP**, true-positive rate

1 Introduction

A major goal of the HUPO Plasma Proteome Project (PPP) is a comprehensive analysis of the protein constituents of human plasma and serum [1]. The pilot phase of this project brought together submissions from 47 different laboratories, of which 18 laboratories submitted peptide and protein iden-

* These authors contributed equally.

tification tables based on MS/MS acquired in either an IT or Q-TOF-like mass spectrometers coupled to multidimensional LC. In order to maximize the discovery of low abundance and potentially interesting peptides and/or proteins, the HUPO-PPP committee emphasized the need for laboratories to submit peptide and protein identification tables along with corresponding protocols, and assigned identifications as either "high-" or "low-confidence". Although this approach is potentially flawed, since the error rate and number of false-positive (FP) protein identification submissions are unknown, it does at least allow the capture of all information; it is then up to informaticians to expertly curate the information such that a reliable analysis of the protein constituents of human plasma and serum can be reported.

It is well recognized that more extensive analysis of LC-MS/MS data is required if data from different experiments, instruments, and laboratories are to be compared [2, 3]. Recent guidelines [4] and issues [5] for the dissemination and publication of large proteomic data sets indicate a growing awareness that a significant number of published protein identifications are indeed incorrect. Hence, an appraisal of MS software and a more informed understanding of the scoring schemes employed by current industry standard MS/MS database search algorithms are warranted [6]. Future literature mining (*e.g.*, Anderson *et al.* [7]) and bioinformatic prediction tools rely heavily on expertly curated data sets so it is imperative that the level of reported FPs remains low, preferably below 1% level.

In MS/MS, gas-phase peptide ions (precursor ions) undergo CID with molecules of an inert gas, such as helium or argon. Under low-energy CID (<100 eV) conditions, typical for ITs, the precursor ion fragments along the peptide backbone bonds give rise to mainly *y*-, *b*-ions, and their neutral losses. Importantly, most of the intensity of the precursor ion is distributed amongst its product ions and depending on the peptides' composition and charge state, might also give rise to selective cleavages, such as enhanced cleavage *N*-terminal to a proline amino acid residue and/or oxidized methionine residues [8, 9], which might hinder its structural elucidation by both *de novo* sequencing and/or database search methods. If an MS/MS spectrum is acquired for a peptide, then its amino acid sequence can be determined by matching the MS/MS spectrum to a known *in silico* generated database of peptide spectra using search algorithms such as SEQUEST [10] and MASCOT [11] in an uninterpreted manner. The rate-limiting step in defining a proteome by these methods is not the capacity to correlate tryptic peptides in this manner, but rather the capacity to accurately interpret such data [12]. Ultimately, investigators aim to determine the protein or gene from which a peptide is derived. This problem is complicated by the fact that a peptide sequence usually does not uniquely define a protein [13]. To this end statistical approaches and models, which attempt to make tandem MS data analysis a consistent and transparent process across research groups, mass spectrometers, and even different MS/MS database search tools, have been developed [14, 15]. These models would undoubtedly benefit from a more informed understanding of the strengths, weaknesses, and limitations of current search algorithms.

Current MS/MS search algorithms scoring functions can essentially be classified into two categories. One category of search algorithms, referred to as heuristic algorithms, correlate the acquired experimental MS/MS spectrum with a theoretical spectrum and calculate a score based on the similarity between the two. These search algorithms are often based on the notion of "shared peak count" (SPC), which simply counts the number of peaks common to the two spectra. Examples of heuristic algorithms include SEQUEST, Spectrum Mill, X!Tandem, and Sonar. Probabilistic algorithms (*e.g.*, MASCOT), on the other hand, model to some extent the peptide fragmentation process (*e.g.*, ladders of sequence ions) and calculate the probability that a particular peptide sequence produced the observed spectrum by chance. A recent review by Sadygov *et al.* [16] provides a useful update and supplement regarding the different models of MS/MS database search algorithms.

1.1 Heuristic algorithms

SEQUEST [10] uses a preliminary scoring (Sp) algorithm, based on a variation of the SPC, to select the 500 best candidate peptide sequences for direct cross-correlation. To speed up computations, fast FTs are used to compute the cross-correlation (X_{corr}), but this does not have any influence on the score itself. For each candidate peptide sequence several scores and rankings are determined.

Spectrum Mill allows MS/MS spectra to be filtered prior to searching, which significantly reduces the number of spectra that need to be analyzed. Its scoring concept is similar to that of the SPC in that 25 of the most abundant fragment ions (above noise level) are matched. Bonus points are awarded depending on the ion type (*b* or *y*) as well as penalty points for unmatched peaks, which is inversely proportional to the relative peak intensity of the unmatched fragment ion. A "scored peak intensity" (SPI) is also calculated, which is the proportion of the TIC that has been assigned (values less than 70% represent a poor interpretation). Again, empirically determined thresholds are used to indicate the correctness of a match, which are applied in an automated fashion.

Sonar [17] (http://bioinformatics.genomicsolutions.com/service/prowl/sonar.html) ranks the proteolytic peptides from proteins in a sequence collection by calculating a score based on the dot product between the theoretical and experimental tandem mass spectra (similar to clustering approaches [18, 19]). The score is subsequently converted into an expectation value [2]. The expectation value is obtained by collecting statistics during the search to estimate the distribution of scores for random and false identifications. This distribution is hypergeometric, and the expectation value of high scoring peptides can therefore be obtained by extrapolation. The expectation value represents the number of peptides that are expected to get a certain score by random matching.

X!Tandem [20, 21] (http://www.proteome.ca/x-bang/tandem/tandem.html) is an open-source search engine that has been optimized for speed. It generates theoretical spectra for the peptide sequences using knowledge of the intensity patterns associated with particular amino acid residues. These spectra are then correlated with the experimental data using a dot product (similar to Sonar). Subsequently, an expectation value is calculated.

1.2 Probabilistic algorithms

Details of the probabilistic MASCOT scoring algorithm have not been published. However, the Matrix Science website (http://www.matrixscience.com) indicates that the MASCOT algorithm incorporates a probability-based implementation of the MOWSE scoring algorithm used for PMF [22] as well as, amongst other things, fragment ion series, mass accuracy, and peptide length. For each peptide, MASCOT reports a probability-based "Ions Score", which is defined as $-10*\log_{10}(p)$, where p is the probability that the observed match between experimental data and the database sequence is a random event. Knowing the size of the sequence database being searched, it becomes possible to provide an objective measure of the significance of a result. MASCOT V2.0 also reports an expectation value, which is similar to those reported by both Sonar and X!Tandem.

Since the majority of search algorithms will always return a score even if the peptide represented by the product ion spectrum is not in the database, it is useful to have an idea of the distribution of the scores for correct or incorrect hits to be able to assess the significance of a particular result. Empirically determined thresholds (filters) have been used [23–25] to indicate the correctness of a match. More recently, the PeptideProphet [14] rescoring algorithm uses Fisher's Linear Discriminant Analysis (LDA) to combine the different SEQUEST scores with other information (*e.g.*, mass difference). The Expectation-Maximization (EM) algorithm as well as Bayes theorem is then used to derive a probability that the peptide hit is correct.

In this paper, we explore the performance of the different MS/MS search algorithms, which were used by the participating HUPO-PPP laboratories, on IT data specially prepared by one of these laboratories. Overall, the main aim of the work is to accurately compare and benchmark the different MS/MS search algorithms based on a validated data set. The more detailed aims of the search algorithm analysis are: (i) to create an expertly curated reference data set that could be used for testing improved MS/MS scoring functions; (ii) to assess the strengths and weaknesses in terms of sensitivity and specificity of the different algorithms; (iii) to accurately benchmark the different algorithms at a specified FP rate; (iv) to assess the effect of database size and different search strategies (tryptic *vs.* nontryptic); (v) to determine the utility of reversed sequence database searches; and (vi) to assess the idea of consensus scoring by combining the results of multiple search algorithms.

2 Materials and methods

2.1 HUPO-PPP reference specimens

Two reference specimens from BD Diagnostics (citrated plasma (Cit-plasma) and serum) for each of three ethnic groups (B1-Caucasian-American, B2-African-American, and B3-Asian-American) were used in these studies [1]. The B1-serum and B1-Cit-plasma (Caucasian-American ethnic group) as well as B3-serum and B3-Cit-plasma (Asian-American ethnic group) were extensively analyzed including manual MS/MS spectrum validation.

2.2 Sample preparation and MS analysis

The HUPO-PPP samples were prepared for MS and run by the PNNL (Adkins and Pounds, see Acknowledgements) as described by Adkins *et al.* [26]. Briefly, serum and plasma were immunoglobulin (Ig) depleted, digested using modified trypsin (Promega), and conditioned by C18 SPE column (Supelco) clean-up. RP separation was performed with an Agilent 1100 capillary column (90 min gradient) interfaced to an LCQ Deca XP IT mass spectrometer (Thermo-Finnigan, San Jose, CA, USA) using ESI. The mass spectrometer was operated in the data-dependent mode to automatically switch between MS and MS/MS acquisition, selecting the three most intense precursor ions for fragmentation using CID.

2.3 Protein sequence databases

All tandem mass spectra were searched against two protein sequence databases and randomized versions of these databases (forward and reverse): a Ludwig Institute nonredundant database (NR, August 2003, ~1.5 million entries) [27] and the Human International Protein Index database (IPI, version 2.21 July 2003, ~56 000 entries, European Bioinformatics Institute, www.ebi.ac/uk/IPI/) [28]. The randomized versions of these databases were created by taking all protein sequence entries and reversing them, such that the original sequence length and composition were preserved.

2.4 MS/MS database search strategy

Since the majority of submissions by HUPO-PPP participating laboratories were based on IT-MS/MS data, it was deemed appropriate to restrict our analysis to search algorithms used by these individual laboratories. Peptide and protein identification lists submitted by the individual participating laboratories to the University of Michigan (central repository) were based on search results from MASCOT, SEQUEST, Sonar, X!Tandem, and Spectrum Mill. Four independent research groups with considerable experience in using one or more of these programs volunteered to analyze the MS data prepared by the PNNL. The JPSL group (Mel-

bourne, Australia) used SEQUEST and MASCOT. Independently, the Agilent team (Jose Meza, Christine Miller, and John Chakel) used Spectrum Mill, David Fenyo at GE Healthcare (formerly Amersham Biosciences) used Sonar and X!Tandem, and Jimmy Eng (ISB, Seattle) used SEQUEST and PeptideProphet to analyze the data. Each group independently decided on their choice of parameters (*i.e.*, data extraction and search parameters) as well as search strategy in order to maximize and optimize at the search algorithm level. Comparisons between search algorithms were carried out using only the subset of spectra common to all the searches (3952 MS/MS spectra, see Section 2.4.1). The MS data as well as protein sequence databases used (where appropriate) were identical for all groups.

2.4.1 SEQUEST and MASCOT workflow performed by the JPSL research group

The LCQ_DTA utility, obtained from ThermoFinnigan as part of the SEQUEST package of programs, was used to extract the MS/MS spectra from the raw instrument data files into individual spectra files (.dta file extension). Parameters used to extract MS/MS spectra were: 700–5000 (min–max mass); minimum of 35 peaks and minimum TIC of 1×10^5 counts. Spectra were not merged, and since doubly- and triply-charged precursor ions cannot accurately be distinguished using low-resolution ESI-IT MS, all spectra not calculated as singly charged were extracted as both doubly- and triply-charged spectra. This resulted in the analysis of 3952 MS/MS spectra for the B3-Cit-plasma (Asian-American) sample. Searches were carried out using both algorithms against the IPI and NR database in both forward and reverse directions using the following search parameters: trypsin-constrained (full with two missed cleavages) as well as no-enzyme (unconstrained) searches; no static or differential modifications; 3 Da precursor ion tolerance and 0.5 Da fragment ion tolerance using monoisotopic masses, and ESI-IT selected as instrument setting.

2.4.2 SEQUEST and PeptideProphet workflow performed by the ISB research group

ThermoFinnigan LCQ raw instrument data files were first converted to the mzXML file format using the ReAdW program [29]. The mzXML2Other program was used to extract individual spectra from the mzXML files into MS/MS files of the .dta format. For the reasons stated previously, all spectra not extracted as singly charged were extracted as both doubly- and triply-charged and no individual spectra were merged. After extraction, a filtering program, named dtafilter (http://sourceforge.net/projects/sashimi), was used to reduce the data set based on the following parameters: 600–4200 Da peptide mass range and a minimum of six peaks with a minimum intensity of 2. This resulted in the analysis of 5579 MS/MS spectra for the B3-Cit-plasma (Asian-American) sample.

SEQUEST database searches were performed on these spectra against the Human IPI protein sequence database (version 2.21). The search parameters were as follows: average masses used for both the peptide mass and fragment ion calculations, peptide mass tolerance set to 3.0, fragment ion tolerance set to 0.0, variable modification of +16.0 to methionine residues, and a sequence constraint of at least one tryptic cleavage site. All search results were passed to the PeptideProphet algorithm using default parameters for IT data. Based on multiple factors of the search results, including individual database search scores and distribution of peptides exhibiting expected cleavage rules, the PeptideProphet algorithm assigned a probability of being a correct identification to each search result.

2.4.3 Spectrum Mill workflow performed by the Agilent group

LCQ instrument data files (*.raw) were extracted with the Spectrum Mill Data Extractor using the following parameters: 600–5000 (min–max mass); sequence tag length on (>1) and off with no spectral merging for two separate sets of search results. Where spectral charge state cannot be determined, no charge state is assigned during extraction and both +2 and +3 charge states are considered during searches. Searches were carried out against the IPI database in both forward and modified reverse directions using the following search parameters: initial search in "multihomology" mode in which combinations of carbamylated lysine, oxidized methionine, and Pyro–Glu modifications were applied; trypsin specific with two missed cleavage; 2.5 Da precursor ion tolerance; 0.7 Da fragment ion tolerance; and ESI-IT as instrument. The initial results were also autovalidated using the following parameters for the "protein details" mode: SPI >70% for matches with score >8 for +1, >7 for +2, and >9 for +3; SPI >90% for score >6 on +1. A second autovalidation step was done in "peptide" mode using criteria of a score >13 and SPI >70%. In addition, both autovalidation steps required a forward–reverse score >1 for +1 and +2 and >2 for +3 peptides. The validated peptides were used to identify a set of proteins from which a result file was created. A second round of searches with unvalidated peptide spectra was performed against the set of proteins in this result file using a no-enzyme (unconstrained) search to identify possible nonspecific or semi-tryptic peptide fragments. All database matches above the threshold score of 3 were summarized and reported.

2.4.4 Sonar and X!Tandem workflow performed by David Fenyo

X!Tandem and Sonar searches were performed by grouping the MS/MS spectra (files with .dta extension) generated by the LCQ_DTA utility (ThermoFinnigan) into single files (with .pkl extension) to speed up the searches. The parameters used in the extraction of the MS/MS spectra were the

same settings as for the SEQUEST and MASCOT searches (see Section 2.4.1). The search parameters used were tryptic digestion with a maximum of two missed cleavage sites, parent ion tolerance of 3 Da, fragment ion tolerance of 0.5 Da, no complete or partial modifications, and the ESI-IT settings. The searches with both Sonar and X!Tandem were performed against IPI database in both forward and reverse directions. Perl scripts were written to automate searches and to parse the output of X!Tandem and Sonar. An expectation value cut-off of 1 was used to filter the results.

2.5 Web interface for data validation, integration, and cross annotation

Scripts written in Perl (version 5.8.4, http://www.perl.com) were used to manage the different data sets and the results of associated database searches obtained from the four independent groups. To assist with the process of manual validation, the Perl scripts also provided the following functionalities: (1) peptide hits with scores above user specified thresholds (cut-points) and/or accepted published cut-offs are highlighted as a visual aid to indicate that a hit is probably correct; (2) a protein summary view (list of inferred proteins) based on correctly identified peptides are sorted by number of matching peptide hits showing all assigned and unassigned peptide spectra matching a particular protein record; (3) options to autovalidate search results based on an already manually validated data set; and (4) highlight and detect inconsistencies between different search algorithms and/or results for the same data set (*i.e.*, same spectrum assigned with two different peptide sequences). Using the Apache web server, a web interface was assembled to allow easy access and manual validation of the data. The annotated spectra were displayed using a Java applet (see Fig. 1). The web interface also allows the user the ability to perform some simple statistics on the data sets, such as comparing numbers of peptide hits which are ranked first or in the top ten for different algorithms. These statistics can be viewed in the form of Venn diagrams and/or concordance plots. The FP and true-positive (TP) rates can also be calculated based on specified rules. For more sophisticated analysis, the validated data set (list of identifications with their scores) can be exported in tab-delimited format for import into spreadsheet packages (such as Excel) and the R statistical package [30] (http://www.r-project.org). Public access to the web interface, database, and associated search results as well as peaklists (.dta files) and supplementary material can be found at http://www.ludwig.edu.au/archive/.

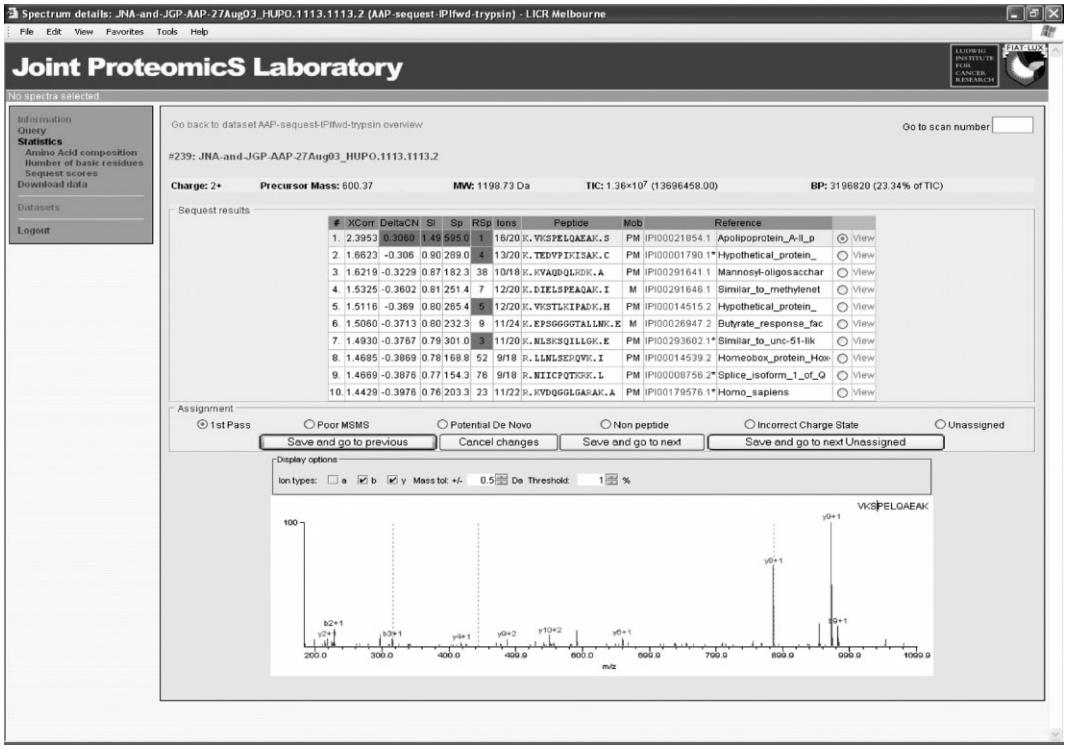

Figure 1. Web-interface for viewing and manually assigning tandem MS peptide identification results. The top ten SEQUEST search results (scores and ancillary information) for a particular spectrum are shown. The selected top hit is used to annotate the spectrum (java applet) showing matching *b* and *y* ions within a user defined threshold and tolerance. Clicking the *View* radio control selects the chosen peptide hit, which is saved in a temporary file if one of the *Save* buttons is selected. Traversing large lists of spectra is made simpler with the "Go to scan number" function at the top of the web page.

The informatics strategy employed to achieve accurately validated as well as unbiased data sets consisted of several phases. First, the four groups optionally validated all MS/MS spectra using their chosen search engine and/or analysis tools by a combination of automated as well as manual assignment. All the data sets in the form of spreadsheets were then collated and made accessible *via* the web interface. The SEQUEST and MASCOT search results (JPSL group) for the B3-Cit-plasma (Asian-American) sample consisting of 3952 MS/MS spectra were separately validated by two independent experts according to established protocols [8, 31]. For SEQUEST, cut-off filters, or thresholds developed by Yates *et al.* [23] and others [24, 25] were used as a guide to highlight probable correct identifications. For MASCOT, the peptides Ions Score, ranking, *E*-value as well as associated protein record were used as a guide to highlight probable correct identifications. All SEQUEST and MASCOT peptide identification search results were therefore independently classified and assigned, using the web interface, as either "1st Pass" (correct), "Poor" (spectra with few ions and/or poor S/N), or "Potential *de novo*" (good quality with many peaks above the noise level). The Perl scripts were then run to first detect inconsistencies (*i.e.*, same spectrum assigned with two different peptide sequences) between the SEQUEST and MASCOT search results. Second, to auto-validate the search results from the other groups based on the already validated SEQUEST and MASCOT assignments (*i.e.*, where peptide sequences were the same for a particular spectrum they were classified as 1st Pass (correct)). Third, peptide identification lists including scores and assignments for all the search algorithms were examined (sorted by descending score or probability) for unassigned spectra as well as conflicts between all search algorithms. Finally, all unassigned as well as conflicts (inconsistencies) were resolved by means of manual inspection by two independent MS experts (a detailed listing of all assignments and peptide sequences returned by the different algorithms for different search strategies can be obtained from http://www.ludwig.edu.au/archive/). An example listing (subset) with explanation is provided in Table 1S (supplementary material).

2.6 ROC curve generation

Receiver operating characteristic (ROC) plots were generated using the statistical package R (version 2.0.1) and used to measure the sensitivity (*i.e.*, the ability to make a correct identification irrespective of the quality of the data, see Eq. 1) and specificity (*i.e.*, the ability to calculate low-ranking scores for random (incorrect) matches, see Eq. 2) of all the MS/MS search algorithms used in this study. An ROC is a graphical plot of the TP rate *versus* the FP rate for a binary classifier system as its discrimination threshold is varied. For each search carried out using the forward protein sequence database, peptide hits were classified based on their score and whether they were correct or incorrect. If the score for a peptide hit was above the threshold, the hit was assigned as positive, and below the threshold they were assigned as negative. If a specific threshold value was selected, it was therefore possible to assign all peptides as either TP, true negative, FP- or false-negative hits.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

where TP is the number of "true positives" (correct hits with scores above threshold) and FN is the number of "false negatives" (correct hits with scores below threshold).

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (2)$$

where TN is the number of "true negatives" (incorrect hits with scores below threshold) and FP is the number of "false positives" (incorrect hits with scores above threshold).

3 Results and discussion

The Cit-plasma and serum samples analyzed as part of this study serve as excellent reference data sets because the acquired MS/MS peptide spectra originate from tryptic digests of plasma and/or serum proteins. Human plasma has a disproportional dynamic range of protein concentrations in that only 22 abundant proteins contribute ~99% of the total protein mass, while an unknown number of relatively low-abundance proteins make up to <1% of the total protein mass [32]. Currently available reference data sets are often mixtures of standard proteins of less dynamic range than that found in human specimens [33]. A particular challenge for MS/MS search algorithms and/or the validation process (automated and/or manual) is whether low-abundance peptides, which presumably originate from low-abundance proteins, are identifiable. Since the currently analyzed samples were not albumin depleted (the most abundant protein (40 mg/mL) in plasma) it is expected that the majority of peptides identified will belong to this protein. The capture and inclusion of lower scoring peptide hits (gray area between correct and incorrect hits), belonging to albumin, should enhance the quality of the reference data set. So even though each peptide hit is validated independently based on its score and annotated spectrum (whether automatically or manually), the inferred protein identity contributes to the overall subjective decision-making process. The inclusion of lower scoring peptide hits that match high-abundance proteins is therefore fundamental in determining the lower detection limits of current MS/MS search algorithms. More often than not, many low-abundance proteins are only identified by a single peptide (the so-called "one-hit wonders" [34]). Irrespective of how these peptides should be dealt with in terms of protein identification, it is important that their spectra are captured so as to facilitate future algorithmic improvements. Finally, all peptide hits were not only validated by a combination of automated as well as manual inspection, but were also cross-validated based on the results of the other search algorithms.

3.1 Comparison of MS/MS search algorithms

In order to compare the MS/MS search algorithms effectively one needs to calculate their coverage or sensitivity (*i.e.*, how many correct hits can be found irrespective of score) and their specificity (*i.e.*, whether the correct hit is significant relative to the other hits). Our reference data set enables accurate calculation of both of these metrics since all the hits returned by the various algorithms have been compared against each other as well as being validated by independent investigators.

3.1.1 Sensitivity and concordance between MS/MS search algorithms

The sensitivity of a search algorithm demonstrates its ability to make a correct identification using any data, irrespective of the quality of the data. Based on the B3-Cit-plasma reference data set and extensive validation and cross-checking/annotation between search algorithms, the overall number of correct peptide hits that were ranked first (irrespective of score) were tabulated in the form of concordance tables for trypsin-constrained (Table 1A) or no-enzyme (unconstrained) (Table 1B) searches of the IPI protein sequence database. The total number of correct hits for each search algorithm is indicated in bold text (diagonal line) (Table 1A and B) and ordered such that the search algorithm with the most hits appear first. For trypsin-constrained searches (Table 1A) it can be seen that SEQUEST identified 526 peptide hits, whilst Spectrum Mill (with tag >1 enabled) identified 397 peptide hits. Based on this observation it is clear that a large number of peptide spectra exhibit incomplete fragmentation patterns (*i.e.*, a less than ideal ladder of sequence ions due to fragmentation kinetics, *etc.*). Nevertheless, over 400 correct peptide hits are identified by at least four different search algorithms indicating reasonable concordance between the different algorithms. The fact that the SEQUEST/PeptideProphet combination (ISB group) identified slightly less hits than that of SEQUEST alone (JPSL group) can probably be attributed to differences in search parameters (*e.g.*, average *vs.* monoisotopic) and/or software versions. For no-enzyme (unconstrained) searches (Table 1B) it can be seen that SEQUEST and Spectrum Mill (when used in a less restrictive mode (*i.e.*, "no tag")) are better able to correctly identify peptides from poorer quality spectra (*i.e.*, higher sensitivity) and also identify a higher number of peptides compared with a trypsin-constrained search. All of the additional peptides identified in the no-enzyme mode were confirmed as belonging to already identified protein records (*e.g.*, albumin).

The overlap, between four of the MS/MS search algorithms, in terms of the number of correctly identified peptide hits that are ranked first is shown in the form of a Venn diagram (Fig. 2) for trypsin-constrained searches of the IPI protein sequence database. Out of a possible 608 hits from the four algorithms (union), 335 peptides are identified by all four algorithms (intersection), whilst 70 peptides are identified by a single algorithm. Almost 75% of these peptide hits are singly charged spectra and 46 of these were independently identified by Spectrum Mill. Upon further inspection, the majority of the 46 hits are either small peptides between 600 and 700 Da in mass or constitute modified peptides (two

Table 1A. Number of correctly identified peptide spectra that are ranked first based on trypsin-constrained searches against the Human IPI v2.21 protein sequence database

	SEQUEST	PeptideProphet	MASCOT	Spectrum Mill	Sonar	X!Tandem	Spectrum Mill(tag)
SEQUEST	**526**	463	463	402	443	424	338
PeptideProphet	463	**499**	453	390	435	416	327
MASCOT	463	453	**492**	389	443	431	324
Spectrum Mill	402	390	389	**476**	389	374	395
Sonar	443	435	443	389	**475**	422	324
X!Tandem	424	416	431	374	422	**457**	314
Spectrum Mill(tag)	338	327	324	395	324	314	**397**

Table 1B. Number of correctly identified peptide spectra that are ranked first based on no-enzyme (unconstrained) searches against the Human IPI v2.21 protein sequence database

	SEQUEST	Spectrum Mill	MASCOT	PeptideProphet	Spectrum Mill(tag)
SEQUEST	**531**	422	438	436	352
Spectrum Mill	422	**528**	388	375	436
MASCOT	438	388	**457**	388	327
PeptideProphet	436	375	388	**455**	321
Spectrum Mill(tag)	352	436	327	321	**438**

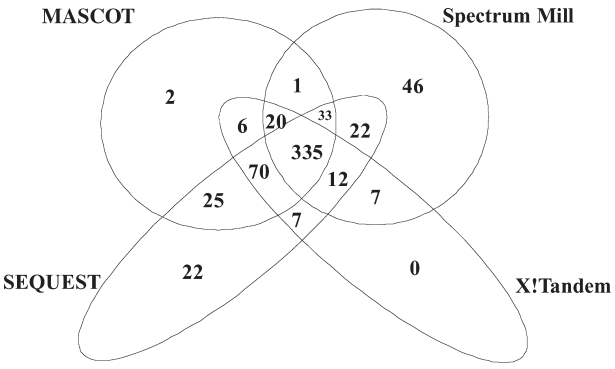

Figure 2. Four-way Venn diagram showing the overlap between four of the MS/MS search algorithms. The number of correctly identified peptides by one or more algorithms is indicated, *e.g.,* 335 peptide hits are correctly identified based on a consensus of all four algorithms (intersection), whilst 608 peptide hits are correctly identified by one or more algorithms (union).

with methionine oxidations, one with a pyroglutamic residue and three with internal carbamylated lysine residues). The majority of these matches were found to be highly credible upon closer inspection by two independent experts.

3.1.2 Specificity and discriminatory power of the primary score statistic for the different MS/MS search algorithms: Distribution of scores and ROC plots

Based on the B3-Cit-plasma reference data set, the distribution of the scores for top-ranking hits obtained from each of the MS/MS search algorithms was plotted for trypsin-constrained as well as no-enzyme (unconstrained) searches of the IPI and/or NR database in both forward and reverse directions (whichever was available). These plots (Fig. 3A–E) illustrate the distribution of scores (highest and lowest) as well as the potential overlap between scores of correct and incorrect peptide hits. For MASCOT searches (Fig. 3A) there is a clear distinction between correct (green) and incorrect (red) peptide hits, especially for trypsin-constrained searches. A search of the reverse databases gives 0 and 6 correct hits for the IPI and NR databases, respectively. The six peptide sequences identified from the reverse NR database are equivalent to real peptides that were also identified in the normal, "forward" search, and all were less than ten residues in length. For SEQUEST (Fig. 3B), it can be seen that there is more overlap between correct and incorrect peptide hits based on the X_{corr} score, especially for no-enzyme (unconstrained) searches (*i.e.,* lower specificity). As for the MASCOT search, a number of correctly identified peptides were obtained when searching the NR database in reverse order. The distribution of scores for Spectrum Mill (based on no-tag search mode) (Fig. 3C) appears to be similar to those of SEQUEST (*i.e.,* slightly more overlap when compared with MASCOT). The distribution of X!Tandem "hyperscores" and Sonar scores for trypsin-constrained searches is displayed on a log-scale (Fig. 3D and E, respectively). A comparison between all of the search algorithms suggests that MASCOT and X!Tandem demonstrate the highest specificity and therefore ability to calculate low-ranking scores for random (incorrect) matches.

ROC plots do not give an indication of the total number of correct hits nor do they illustrate the number of correct hits that might not be ranked first, but they do allow an overall comparison of the sensitivity and specificity of a search algorithm independent of a specific threshold. Based on the B3-Cit-plasma reference data set, ROC plots were generated for the different MS/MS search algorithms for trypsin-constrained (Fig. 4A) and no-enzyme (unconstrained) (Fig. 4B) search results. Since the ROC curve displays the sensitivities and FP rates at all possible cut-off levels, it can be used to assess the performance of the primary score (*i.e.,* X_{corr} for SEQUEST or Hyperscore for X!Tandem), independent of any decision threshold. Therefore, ideal behavior would be a curve that approaches a sensitivity of 1.0 without any FP (*i.e.,* 1-specificity is 0.0). This would indicate that a search algorithm is perfectly able to discriminate between correct and incorrect peptide hits, and the calculated area under the curve (AUC) would be 1.0. The AUC is a measure of the overall performance in terms of separating positives and negatives with values approaching 0.5 indicating random discrimination (*i.e.,* the diagonal line also called the chance diagonal). For trypsin-constrained searches (Fig. 4A) it can seen that the MASCOT Ion Score and SEQUEST/PeptideProphet combination perform better than X!Tandem and Sonar, which again perform better than SEQUEST and Spectrum Mill (tag >1 enabled). From Fig. 4B (no-enzyme searches), it can be seen that all the search algorithms, with the exception of the SEQUEST/PeptideProphet combination, perform worse when compared with the representative trypsin-constrained searches. The fact that the AUC improves slightly for the SEQUEST/PeptideProphet combination for no-enzyme searches (0.97 *vs.* 0.96) indicates that the number of tryptic termini is an important determinant in deriving the probability for a peptide hit. None of the individual search algorithms take this into account when classifying correct *versus* incorrect peptide hits.

3.1.3 Calculation of score thresholds based on specified FP identification error rates

Based on the B3-Cit-plasma reference data set and database search results of known validity, score thresholds (cut-offs) for the different MS/MS search algorithms were calculated at specified FP identification rates (0.1, 1, and 5%). These score thresholds were also calculated with regard to their charge state and various filtering criteria (*e.g.,* $R_{sp} < 5$ and $\Delta C_n >= 0.1$ for SEQUEST) for trypsin-constrained and/or no-enzyme (unconstrained) searches (see Table 2A–F). The criteria that give rise to the most TP hits at the 1% FP rate are

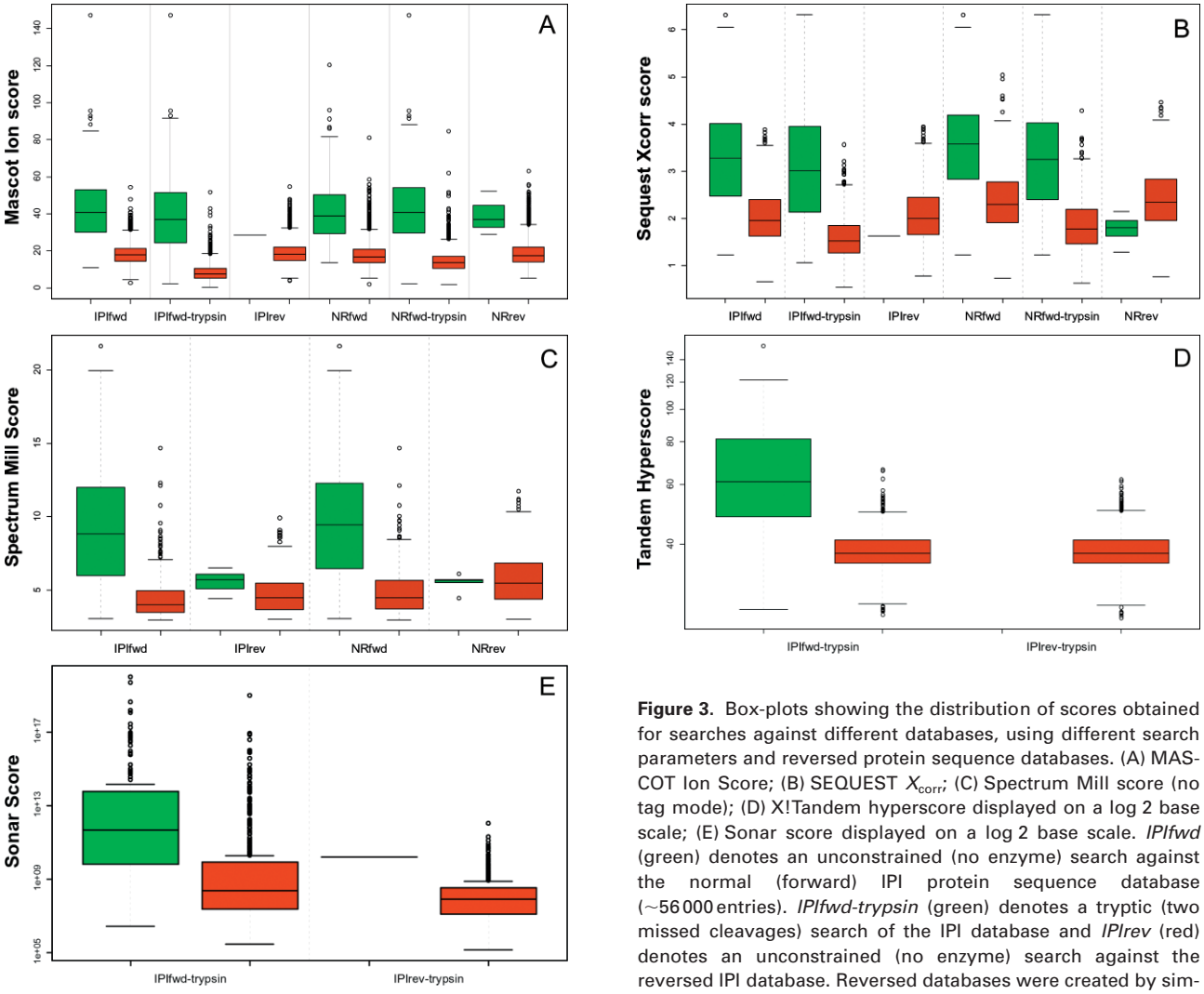

Figure 3. Box-plots showing the distribution of scores obtained for searches against different databases, using different search parameters and reversed protein sequence databases. (A) MASCOT Ion Score; (B) SEQUEST X_{corr}; (C) Spectrum Mill score (no tag mode); (D) X!Tandem hyperscore displayed on a log 2 base scale; (E) Sonar score displayed on a log 2 base scale. *IPIfwd* (green) denotes an unconstrained (no enzyme) search against the normal (forward) IPI protein sequence database (~56 000 entries). *IPIfwd-trypsin* (green) denotes a tryptic (two missed cleavages) search of the IPI database and *IPIrev* (red) denotes an unconstrained (no enzyme) search against the reversed IPI database. Reversed databases were created by simply reversing each individual protein sequence entry and as such maintaining the original sequence composition and length. The *NRfwd* (green) denotes a search against the normal (forward) NR protein sequence database (~1.5 million entries). Box-plots were automatically generated using the statistical package R, version 2.0.1 using default parameters (*i.e.*, outliers are scores >1.5X the interquartile range (75–25%), which are indicated by dots (o), whiskers represent the highest score not considered to be an outlier, and the box represents scores between 25 and 75% with median at 50%).

indicated in bold text in the tables. The calculated score thresholds can first be used to judge the usefulness of a specific criterion (*e.g.*, R_{sp} for SEQUEST) and whether or not its inclusion improves the overall specificity and sensitivity of a particular search algorithm. Second, a sense of what constitutes equivalence between search algorithms can be obtained if one compares score thresholds at a specified FP rate (*i.e.*, what MASCOT score is equivalent to an X!Tandem score, for example). Finally, the score thresholds can be used for autovalidation purposes (*i.e.*, assume all peptide hits to be correct if their scores are above the calculated thresholds) but with the following caveats: that the thresholds be applied to similar data sets (*i.e.*, LCQ-like data) obtained under similar experimental conditions and analyzed using the same search parameters (*i.e.*, searches are performed using 3 Da precursor ion tolerance and against similar sized protein sequence databases).

From Table 2A, for trypsin-constrained searches, it can be seen that the $\Delta C_n >= 0.1$ as well as $R_{sp} < 5$ criteria improve the overall specificity of the SEQUEST algorithm. The R_{sp} criterion has largely been ignored in published studies to date but it is clear from Table 2A that this filter should be included when analyzing search results from complex protein extracts such as cell lysates and tissues such as blood. For no-enzyme (unconstrained) searches (Table 2A), it can be seen that many random (incorrect) matches can be filtered by applying the strict trypsin rule (*i.e.*, peptide must be fully tryptic).

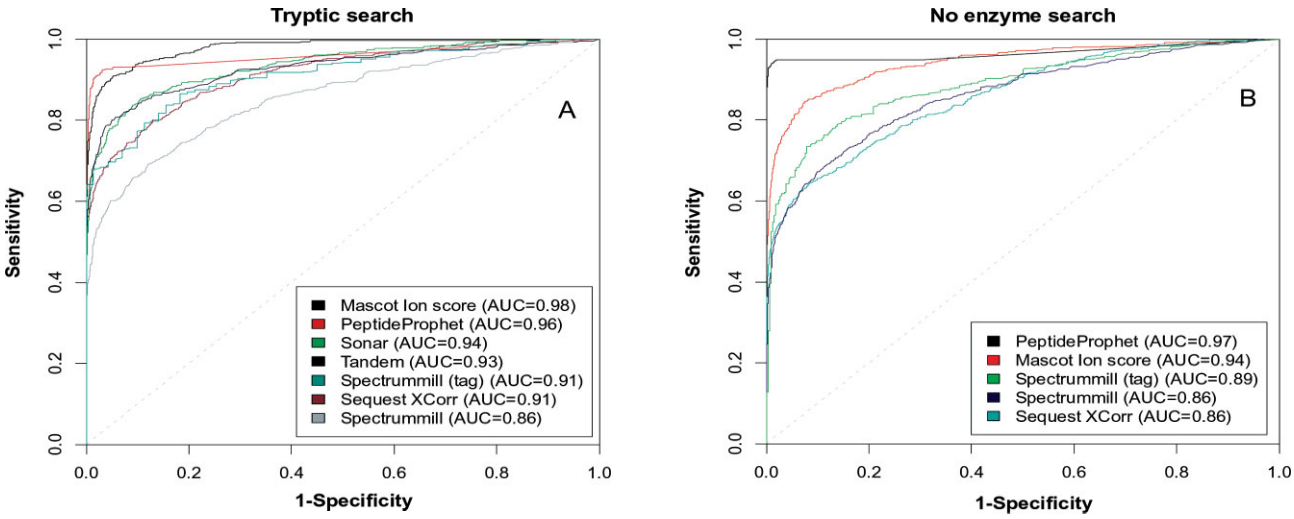

Figure 4. ROC plot for the different search algorithms based on searches against the IPI protein sequence database: (A) tryptic-constrained (two missed cleavages) and (B) unconstrained (no enzyme). Hundred percent discrimination between correct and incorrect peptide hits would be indicated by a sensitivity of 1.0 and 1-specificity of 0.0 (*i.e.,* a search algorithm is able to identify all TP hits without any FPs). The AUC is indicated for each search algorithm (values of 0.5 would be considered random also called the chance diagonal (dotted line)).

Table 2A. SEQUEST X_{corr} thresholds calculated based on different criteria at specified FP error rates for trypsin-constrained and no-enzyme searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin-constrained search %FP rate			No-enzyme search %FP rate		
		0.1	1	5	0.1	1	5
All	None	3.56 (37%)	2.64 (59%)	2.28 (71%)	3.89 (30%)	3.3 (49%)	2.95 (60%)
	ΔC_n[a)]	3.21 (44%)	2.55 (62%)	2.16 (73%)	3.89 (29%)	3.1 (53%)	2.4 (73%)
	$\Delta C_n + R_{sp}$[b)]	3.21 (44%)	**2.44 (65%[d)])**	1.68 (85%)	3.41 (43%)	2.05 (68%)	0 (70%)
	Tryptic[c)]	–	–	–	3.11 (50%)	**2.27 (73%)**	0 (89%)
	$\Delta C_n + R_{sp}$ + tryptic	–	–	–	3.01 (48%)	0[e)](63%)	0 (63%)
	No. of hits[f)]	**526**			**531**		
Singly-charged peptides (1+)	None	2.41 (18%)	2.05 (41%)	1.73 (75%)	2.82 (9%)	2.41 (28%)	2.11 (48%)
	ΔC_n	2.22 (27%)	1.86 (58%)	1.59 (76%)	2.29 (35%)	**2.01 (49%)**	1.38 (68%)
	$\Delta C_n + R_{sp}$	2.02 (35%)	**1.72 (58%)**	1.35 (64%)	2.03 (26%)	1.37 (31%)	0 (31%)
	Tryptic	–	–	–	2.69 (10%)	2.22 (33%)	1.71 (79%)
	$\Delta C_n + R_{sp}$ + tryptic	–	–	–	0 (24%)	0 (24%)	0 (24%)
	No. of hits	**168**			**127**		
Doubly-charged peptides (2+)	None	3.56 (50%)	2.44 (89%)	1.97 (98%)	3.13 (66%)	2.76 (81%)	2.41 (91%)
	ΔC_n	3.01 (70%)	2.35 (90%)	1.81 (97%)	3.08 (66%)	2.52 (84%)	2.07 (92%)
	$\Delta C_n + R_{sp}$	3.01 (70%)	**2.18 (93%)**	1.5 (97%)	3.01 (67%)	2.01 (84%)	0 (85%)
	Tryptic	–	–	–	3.01 (64%)	**2.22 (85%)**	0 (89%)
	$\Delta C_n + R_{sp}$ + tryptic	–	–	–	3.01 (61%)	0 (75%)	0 (75%)
	No. of hits	**281**			**331**		
Triply-charged peptides (3+)	None	3.21 (74%)	2.68 (83%)	2.42 (88%)	3.89 (58%)	3.43 (78%)	3.13 (84%)
	ΔC_n	3.21 (73%)	2.68 (83%)	2.34 (90%)	3.89 (56%)	3.3 (78%)	2.75 (86%)
	$\Delta C_n + R_{sp}$	3.21 (73%)	**2.53 (84%)**	2.02 (96%)	3.41 (64%)	2.38 (71%)	0 (71%)
	Tryptic	–	–	–	3.11 (82%)	**2.41 (95%)**	0 (99%)
	$\Delta C_n + R_{sp}$ + tryptic	–	–	–	2.38 (71%)	0 (71%)	0 (71%)
	No. of hits	**77**			**73**		

a) ΔC_n criteria >= 0.1
b) R_{sp} criteria < 5
c) True (full) tryptic criteria
d) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point f below)
e) Negligible score threshold (*i.e.,* almost zero)
f) Total number of correctly identified peptide hits

Table 2B. MASCOT Ions Score thresholds calculated based on different criteria at specified FP error rates for trypsin-constrained and no-enzyme searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin-constrained search %FP rate			No-enzyme search %FP rate		
		0.1	1	5	0.1	1	5
All	None	51.80 (25%[b)])	23.08 (78%)	16.20 (90%)	54.39 (23%)	34.63 (63%)	27.96 (80%)
	Tryptic[a)]	–	–	–	47.94 (33%)	**22.38 (79%)**	0[c)] (87%)
	No. of hits[d)]	**493**			**457**		
Singly-charged peptides (1+)	None	51.80 (2%)	**27.83 (35%)**	21.17 (59%)	42.79 (11%)	38.03 (22%)	31.21 (37%)
	Tryptic	–	–	–	42.79 (10%)	**32.36 (29%)**	18.46 (72%)
	No. of hits	**125**			**87**		
Doubly-charged peptides (2+)	None	39.02 (62%)	**20.00 (93%)**	13.87 (96%)	54.39 (31%)	33.12 (79%)	24.45 (91%)
	Tryptic	–	–	–	47.94 (43%)	**19.06 (85%)**	0 (88%)
	No. of hits	**299**			**311**		
Triply-charged peptides (3+)	None	16.06 (91%)	**20.84 (84%)**		40.92 (37%)	33.58 (64%)	28.26 (81%)
	Tryptic	–	–	–	40.77 (37%)	**19.39 (95%)**	0 (97%)
	No. of hits	**69**			**59**		

a) True (full) tryptic criteria
b) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point d below)
c) Negligible score threshold (*i.e.*, almost zero)
d) Total number of correctly identified peptide hits

Table 2C. X!Tandem score thresholds calculated based on specified FP error rates for trypsin-constrained searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin-constrained search %FP rate		
		0.1	1	5
All	None	66.4 (42%a))	50.7 (68%)	45 (81%)
	No. of hits[b)]	**457**		
Singly-charged peptides (1+)	None	54.4 (19%)	53.7 (20%)	44.6 (44%)
	No. of hits	**116**		
Doubly-charged peptides (2+)	None	66.4 (63%)	53.3 (82%)	46.3 (93%)
	No. of hits	**284**		
Triply-charged peptides (3+)	None	62.3 (21%)	48.1 (82%)	44.3 (89%)
	No. of hits	**57**		

a) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point b below)
b) Total number of correctly identified peptide hits

The MASCOT Ions Score thresholds for both trypsin and no-enzyme searches (Table 2B) indicate that thresholds are higher for singly-charged peptide ions compared with doubly- and triply-charged peptides. A comparison of the thresholds with the reported MASCOT "identity score ($p < 0.05$)" of 43 for trypsin-constrained and 60 for no-enzyme searches reveals the following: for trypsin-constrained searches a cut-off score of 43 gives an FP rate of 0.03% and TP rate of 38% whilst applying the reported homology score gives an FP rate of 0.23% and a TP rate of 70.38% (data not shown); for the no-enzyme searches a cut-off score of 60 gives an FP rate of 0% and a TP rate of 14.22% whilst applying the reported homology score gives an FP rate of 0.34% and TP rate of 60.39% (data not shown).

A comparison of trypsin-constrained and no-enzyme (unconstrained) searches for SEQUEST and MASCOT searches (Table 2A and B, respectively) indicates that score thresholds are considerably higher at all predefined FP rates for no-enzyme searches. Indeed for both SEQUEST and MASCOT, similar score thresholds are obtained for a no-enzyme search against the IPI protein sequence database compared with a trypsin-constrained search of the NR database (which is comprised of ~1.5 million entries) (data not shown). This

Table 2D. Sonar score thresholds calculated based on specified FP error rates for trypsin-constrained searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin-constrained search %FP rate		
		0.1	1	5
All	None	$5.3e^{13}$(21%[a)])	$3.6e^{9}$ (63%)	$2.4e^{8}$ (78%)
	No. of hits[b)]	**475**		
Singly-charged peptides (1+)	None	$3.7e^{12}$ (12%)	$4.5e^{10}$ (33%)	$1.6e^{9}$ (60%)
	No. of hits	**129**		
Doubly-charged peptides (2+)	None	$6.7e^{10}$ (54%)	$3.4e^{8}$ (77%)	$5.3e^{7}$ (87%)
	No. of hits	**281**		
Triply-charged peptides (3+)	None	$5.3e^{13}$ (38%)	$4.1e^{9}$ (77%)	$3.4e^{8}$ (85%)
	No. of hits	**65**		

a) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point b below)
b) Total number of correctly identified peptide hits

Table 2E. Spectrum Mill (tag >1) score thresholds calculated based on different criteria at specified FP error rates for trypsin and hierarchical iterative searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin and no-enzyme iterative search %FP rate		
		0.1	1	5
All	None	14.68 (13%[b)])	9.42 (53%)	7.96 (66%)
	Tryptic[a)]	8.46 (58%)	**7.67 (63%)**	5.39 (79%)
	No. of hits[c)]	**438**		
Singly-charged peptides (1+)	None	8.6 (16%)	8.6 (16%)	5.77 (51%)
	Tryptic	7.65 (21%)	**7.65 (21%)**	5.39 (46%)
	No. of hits	**104**		
Doubly-charged peptides (2+)	None	10.78 (50%)	10.78 (50%)	8.66 (71%)
	Tryptic	8.46 (69%)	**8.46 (69%)**	7.96 (73%)
	No. of hits	**268**		
Triply-charged peptides (3+)	None	14.68 (23%)	12.13 (45%)	7.95 (85%)
	Tryptic	7.51 (86%)	**6.38 (91%)**	4.03 (97%)
	No. of hits	**66**		

a) True (full) tryptic criteria
b) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point c below)
c) Total number of correctly identified peptide hits

Table 2F. SEQUEST/PeptideProphet thresholds calculated based on specified FP error rates for trypsin-constrained searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	Criteria applied to peptide hit	Trypsin-constrained search %FP rate		
		0.1	1	5
All	None	0.96 (56%[a)])	0.11 (88%)	0 (93%)
	No. of hits[b)]	**499**		
Singly-charged peptides (1+)	None	0.49 (57%)	0.29 (67%)	0 (76%)
	No. of hits	**126**		
Doubly-charged peptides (2+)	None	0.96 (76%)	0.17 (94%)	0.01 (99%)
	No. of hits	**301**		
Triply-charged peptides (3+)	None	0.86 (68%)	0 (93%)	0 (97%)
	No. of hits	**72**		

a) %TP peptide identifications based on the specified criteria and total number of correctly identified peptide hits (see point b below)
b) Total number of correctly identified peptide hits

clearly highlights the "distraction effect" as a result of an effective increase in database size due to the increased number of peptides that must be queried.

The hyperscore thresholds for X!Tandem and Sonar score thresholds, based on trypsin-constrained searches, are shown in Table 2C and D, respectively. For X!Tandem (Table 2C), the thresholds are constant across all charge states, indicating that singly-charged spectra do not have such a negative effect on the X!Tandem scoring function. Also, based on these calculations, a score of 50 (~1% FP) would be more appropriate than previously suggested (unpublished) score cut-offs of 45 which equates to ~5% FP rate under these conditions. The Spectrum Mill (tag >1 mode) results (see Table 2E) are based on a five-phase iterative search strategy. Again (similar to SEQUEST), it can be seen that the scores are dependent on the charge state of the precursor ion. Finally, the probability thresholds at the different FP rates for SEQUEST/PeptideProphet are shown in Table 2F.

It is clear from Table 2A–F that the number of TP peptide identifications is lowest for singly-charged peptide spectra. This is perhaps not surprising when one considers that singly-charged precursor ions are inherently smaller, fragment in a less predictable manner, and generate less fragment ions. Approximately, 30% of the low-mass ions are not observed on a 3-D IT due to the low-mass cut-off. Doubly- and triply-charged peptide spectra are less affected by the low-mass cut-off, and since the majority of tryptic peptide spectra are doubly-charged under electrospray conditions and have a mobile proton [35], more ideal fragmentation is facilitated and hence identified by current search algorithms.

3.1.4 Benchmarking of the different MS/MS search algorithms at 1% FP error rate

Based on the results from Table 2A–F, Table 3 provides an overall comparison and accurate benchmark of the search algorithms evaluated in this study in terms of the number of correctly identified peptide spectra (TP) at 1% FP rate. Overall, taking into account all charges (first row Table 3) it can be seen that PeptideProphet when applied to SEQUEST results identifies 439 peptides whilst Spectrum Mill (used with tag >1) identifies 276 peptides. However, at the individual charge state level, especially singly-charged, there appears to be much variation between the different search algorithms.

3.1.5 Effect of database size and search strategy

In order to investigate the effect of database size as well as optimal search strategy, the total number of correct hits (ranked first or in the top ten) reported by both MASCOT and SEQUEST was tabulated based on searches against the IPI and NR databases using trypsin-constrained as well as no-enzyme (unconstrained) searches (see Table 4). First, it can be seen that the search algorithms lose sensitivity as the search space is increased (*i.e.*, more peptides have to be queried) and that MASCOT is affected more than that of SEQUEST since the correct peptide hit appears more often in the top ten hits rather than being ranked first. This indicates that the SEQUEST scoring function is slightly more sensitive (*i.e.*, better able to rank poorer quality peptide spectra) compared with that of MASCOT, especially when large protein sequence databases are used and/or unconstrained searches are carried out. Second, of the 581 correctly identified peptides (top ten considered, no-enzyme search) for SEQUEST, 89% are true-tryptic, 11% are semitryptic, and none are nonspecific, whereas of the incorrectly identified peptides, 2% are true-tryptic, 20% are semitryptic, and 78% are nonspecific. These values are in close agreement with those calculated by Keller *et al.* [33] based on tryptic digestion of an 18 standard protein mixture. Our findings therefore support and confirm the observation that trypsin is a very specific protease [36, 37]. In fact, the majority of semitryptic peptides identified in this analysis were derived from human albumin, the most abundant protein in these samples.

3.1.6 Utility of reversed sequence searches

The utility of reversed sequence searches to restrict the number of FP peptide identifications has been explored by various groups [25, 38, 39]. The idea is to analyze a particular data set and identify peptides using both the "normal" forward and "random" reversed protein sequence database searches. The random database could be appended to the normal database or searched separately. Our protocol consisted of the following steps: (1) reversed sequence searches were carried out separately; (2) the search results were then filtered so as to remove correct matches based on the validated normal forward search (see Section 3.1.2); (3) the

Table 3. Number of correctly identified peptide spectra (TP rate) based on a 1% FP rate (benchmark) for the different search algorithms for trypsin-constrained searches against the Human IPI v2.21 protein sequence database

Charge state of precursor ion	SEQUEST/ PeptideProphet	MASCOT	SEQUEST ($\Delta C_n + R_{sp}$)	X!Tandem	Sonar	Spectrum Mill (tag >1)
All	439	385	342	311	299	276
Singly-charged peptides (1+)	84	44	97	23	43	22
Doubly-charged peptides (2+)	283	278	261	233	216	185
Triply-charged peptides (3+)	67	58	65	47	50	60

Table 4. Number of correctly identified peptide spectra for SEQUEST and MASCOT based on different search strategies and protein sequence databases

		Top hit[a]		Top ten hits[b]	
		Trypsin[c]	No-enzyme[d]	Trypsin	No-enzyme
SEQUEST	IPI[e]	526	531	535	581
	NR[f]	498	418	552	481
MASCOT	IPI	492	457	539	526
	NR	425	363	508	446

a) Only the correct peptide hits that are ranked first are considered
b) Correct peptide hits ranked amongst the top ten are considered
c) Trypsin-constrained search (full tryptic) with two missed cleavages
d) No-enzyme (unconstrained) search
e) Human IPI v2.21 protein sequence database comprising ~56 000 entries
f) Ludwig Institute NR (nonredundant) protein sequence database comprising ~1.5 million entries

scores were then sorted in descending order and the threshold determined based on the *n*th ranked score depending on the specified (acceptable) FP rate. For example, if the FP rate is 1% and 1000 peptide spectra are scored, the tenth highest score would be the score threshold. Our findings indicate that similar score thresholds, albeit slightly higher thresholds, were obtained compared with those from the normal forward search (Table 2A–F). This appears to be in agreement with others regarding the estimation of FP rates based on the reverse database model [39]. In order for this approach to be effective it would have to be repeated for each experiment. The obvious disadvantage of the reverse database model is the number of false-negative peptide hits (*i.e.*, the correct peptide identifications below the threshold) but it demonstrates an improvement on empirically derived published score cut-offs.

3.1.7 Consensus scoring between MS/MS search algorithms

The idea of consensus scoring has previously been raised [40] and briefly explored here. The basic idea is to merge search results from two or more algorithms and combine the scores for peptide spectra where there is consensus between different algorithms. The top ranking peptide hit or top ten peptide hits for each spectrum, from the different algorithms, could be considered. Interestingly, based on the data sets used in this study, when one compares all the top ranked peptide sequences returned by both MASCOT (trypsin search) and SEQUEST (trypsin search), 646 peptide sequences are found to be identical, and of these, 465 have been validated as correct. However, when one compares the top ranked peptide sequences returned by both MASCOT (trypsin search) and SEQUEST (no-enzyme search), 470 peptide sequences are found to be identical, and of these, 450 have been validated as correct (data available from website, see Section 2.5). A closer inspection of these 20 peptide spectra (identical sequences but not 1st Pass) reveals that they are mostly poorer quality (singly-charged) spectra with low scores and exhibiting less than ideal ladders of sequence ions. Further examination and observation regarding consensus amongst at least three algorithms reveal that the MASCOT scoring function generally performs poorer on singly-charged spectra and/or spectra exhibiting few ions or spectra exhibiting many ions but with a few very intense peaks. Indeed when one compares all the top ranked peptide sequences returned by all the search algorithms and filter out nonidentical sequences, we find that the remaining peptides have all been classified by the investigators as correct. This suggests that the consensus approach based on multiple scoring functions definitely has merit and that the scores could be considered as independent and orthogonal. Further work needs to be carried out to determine exactly how many search algorithms (or independent scoring functions) are required so as to allow confident and automated validation of peptide identifications and therefore accurate protein identifications.

4 Concluding remarks

Our aims in this paper were to assess the strengths and weaknesses of different MS/MS search algorithms on IT data, and to provide guidelines to help assess the significance of peptide identification results obtained from the individual HUPO-PPP participating laboratories. Important considerations when carrying out MS/MS database searches are the specified search parameters (*i.e.*, mass tolerance which is dependent on the instrument and calibration), search strategy (*i.e.*, semitryptic *vs.* tryptic), chosen protein sequence database to query (*i.e.*, IPI *vs.* NCBI NR which is dependent on the particular experiment), and chosen search engine. The choice of search engine should not only be guided by the range of mass spectrometers available but also whether or not it is restrictive regarding the above choices as well as its overall sensitivity and specificity, which we have addressed in this study.

It is clear from this study that the number of correctly identified peptides that are ranked first by the different algorithms decreases (less sensitive) as the search space is increased (*i.e.*, no-enzyme search and/or large protein sequence database). This is particularly notable for MASCOT compared with SEQUEST, on the basis of the number of correctly identified peptides that are no longer ranked first but appear in the top ten. SEQUEST and Spectrum Mill (using no tag filter) are more sensitive than the other algorithms but MASCOT, Sonar, and X!Tandem are more specific (*i.e.*, better able to discriminate between correct and incorrect

peptide hits). Overall, calculating the TP rate at a specified FP rate shows that MASCOT performs better than the other algorithms used in this study. Application of a rescoring algorithm, such as PeptideProphet, improves the specificity of the SEQUEST algorithm and based on these results should also improve the results of the other algorithms.

Score thresholds, if used, can be determined based on reverse sequence searches as demonstrated in this study. For high-confidence peptide identifications these thresholds could be combined with orthogonal scoring information, such as scores from other search algorithms. The availability of open-source algorithms, such as X!Tandem as well as OMSSA [41] make this process feasible. In this respect, an algorithm that demonstrates high sensitivity should be used in conjunction with an algorithm that demonstrates high specificity. Thresholds, if used, should also be calculated on a *per*-experiment basis because the number of spectra generated and the detectable dynamic range of proteins have a major influence on the number of potential FP identifications. For example, at a predefined score threshold, the number of FP identifications will be higher if a large number of spectra are generated that do not correctly match anything in the protein sequence database. This scenario is typical of human specimens, such as plasma, which exhibits a disproportional dynamic range of protein concentrations.

The MS data, generated for this study, were performed in the Environmental Molecular Sciences Laboratory, a US national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. We thank Joel Pounds, Dick Smith, and Ron Moore for access to the MS data; James Eddes for the mass spectrum applet used in the web interface; Robert Moritz for access to the JPSL MASCOT server. Funding was provided, in part, by the HUPO-PPP and by the Australian National Health and Medical Research Council (program grant no. 280912).

5 References

[1] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[2] Fenyo, D., Beavis, R. C., *Anal. Chem.* 2003, *75*, 768–774.

[3] Nesvizhskii, A. I., Aebersold, R., *Drug Discov. Today* 2004, *9*, 173–181.

[4] Clauser, K., Negvizhskii, A., Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[5] Baldwin, M. A., *Mol. Cell. Proteomics* 2004, *3*, 1–9.

[6] Simpson, R. J., *Eur. J. Pharm. Sci. Rev.* 2004, *9*, 25–36.

[7] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 311–326.

[8] Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S. *et al.*, *Anal. Chem.* 2003, *75*, 6251–6264.

[9] Reid, G. E., Roberts, K. D., Kapp, E. A., Simpson, R. J., *J. Proteome Res.* 2004, *3*, 751–759.

[10] Eng, J. K., McCormack, A. L., Yates III, J. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[11] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[12] Marshall, J., Jankowski, A., Furesz, S., Kireeva, I. *et al.*, *J. Proteome Res.* 2004, *3*, 364–382.

[13] Kearney, P., Thibault, P., *J. Bioinform. Comput. Biol.* 2003, *1*, 183–200.

[14] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[15] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *Anal. Chem.* 2003, *75*, 4646–4658.

[16] Sadygov, R. G., Cociorva, D., Yates III, J. R., *Nat. Methods* 2004, *1*, 195–202.

[17] Field, H. I., Fenyo, D., Beavis, R. C., *Proteomics* 2002, *2*, 36–47.

[18] Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., Yates, J. R., 3rd, *Anal. Chem.* 2003, *75*, 2470–2477.

[19] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[20] Craig, R., Beavis, R. C., *Bioinformatics* 2004, *20*, 1466–1467.

[21] Craig, R., Beavis, R. C., *Rapid Commun. Mass Spectrom.* 2003, *17*, 2310–2316.

[22] Pappin, D. J., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[23] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al.*, *Nat. Biotechnol.* 1999, *17*, 676–682.

[24] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, *Nat. Biotechnol.* 2001, *19*, 242–247.

[25] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P., *J. Proteome Res.* 2003, *2*, 43–50.

[26] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 947–955.

[27] Moritz, R. L., Ji, H., Schutz, F., Connolly, L. M. *et al.*, *Anal. Chem.* 2004, *76*, 4811–4824.

[28] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[29] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[30] R Developement Core Team, *A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna, Austria 2005, http://www.R-project.org.

[31] Simpson, R. J., Connolly, L. M., Eddes, J. S., Pereira, J. J. *et al.*, *Electrophoresis* 2000, *21*, 1707–1732.

[32] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[33] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S. *et al.*, *Omics* 2002, *6*, 207–212.

[34] Veenstra, T. D., Conrads, T. P., Issaq, H. J., *Electrophoresis* 2004, *25*, 1278–1279.

[35] Wysocki, V. H., Tsaprailis, G., Smith, L. L., Breci, L. A., *J. Mass Spectrom.* 2000, *35*, 1399–1406.

[36] Keil, B., *Specificity of Proteolysis*, Springer-Verlag, Berlin 1992.

[37] Olsen, J. V., Ong, S. E., Mann, M., *Mol. Cell. Proteomics* 2004, *3,* 608–614.

[38] Cargile, B. J., Bundy, J. L., Stephenson, J. L., Jr., *J. Proteome Res.* 2004, *3,* 1082–1085.

[39] Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F. *et al.*, *J. Proteome Res.* 2005, *4,* 53–62.

[40] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D. *et al.*, *Anal. Chem.* 2004, *76,* 3556–3568.

[41] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. *et al.*, *J. Proteome Res.* 2004, *3,* 958–964.

