EuPA OPEN PROTEOMICS 5 (2014) 21–31

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis

Andreas Quandt[a], *Lucia Espona*[a,b], *Akos Balasko*[c], *Hendrik Weisser*[a], *Mi-Youn Brusniak*[d], *Peter Kunszt*[b,1], *Ruedi Aebersold*[a,e], *Lars Malmström*[a,*,1]

[a] *Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Switzerland*
[b] *SyBIT, SystemsX.ch, Switzerland*
[c] *MTA SZTAKI, Laboratory of Parallel and Distributed Systems, Budapest, Hungary*
[d] *Institute for Systems Biology, Seattle, USA*
[e] *Faculty of Science, University of Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 4 November 2013
Received in revised form
17 April 2014
Accepted 6 October 2014
Available online 28 October 2014

Keywords:
Mass spectrometry
Data analysis
Classical database search
Synthetic peptides
Search engine

ABSTRACT

Tandem mass spectrometry and sequence database searching are widely used in proteomics to identify peptides in complex mixtures. Here we present a benchmark study in which a pool of 20,103 synthetic peptides was measured and the resulting data set was analyzed using around 1800 different software and parameter set combinations. The results indicate a strong relationship between the performance of an analysis workflow and the applied parameter settings. We present and discuss strategies to optimize parameter settings in order to significantly increase the number of correctly assigned fragment ion spectra and to make the analysis method robust.

© 2014 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

Tandem mass spectrometry (MS/MS) is the method of choice for identifying and quantifying proteins in complex mixtures because of its high throughput, sensitivity and relative ease of use. However, the optimal analysis of the resulting mass spectrometry data is complex and the subject of continuous research. In the most frequent data analysis workflow, fragment ion spectra generated from selected peptide ions

Abbreviations: CPM, classical parametric model; FDR, false discovery rate; FME, fragment mass error; I, MS2Deisotope; LTQ, linear trap quadrupole; FT, Fourier transform; M, Mascot; MC, missed cleavage; N, MS2Denoise; O, OMSSA; PM, parametric model; PMC, parametric model with correction of the negative distribution based on decoy hits; PME, parent mass error; PSMs, peptide spectrum matches; PTMs, post-translational modifications; R, precursorRefine; SPM, the semi-parametric model; TPP, Trans-Proteomic Pipeline; UIPs, uniquely identified peptides; X, X! Tandem.

* *Corresponding author*. Tel.: +41 446332195.
E-mail address: lars@imsb.biol.ethz.ch (L. Malmström).
[1] Current address: S3IT, University of Zurich, Switzerland.

http://dx.doi.org/10.1016/j.euprot.2014.10.001
2212-9685/© 2014 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

are assigned to their corresponding peptide sequences using software tools commonly referred to as database search engines. Numerous search engines have been developed, each one using a different algorithm to maximize the number of peptide-spectrum matches (PSMs) and to assess confidence in the correctness of their assignments [1,2]. Search engines compute a score for each PSM that reflects the quality of the assignment; the user defines a cutoff that optimally separates correct from incorrect assignments. In more recent studies, the score cutoff is selected by a target-decoy strategy [3] to achieve a specific false discovery rate (FDR) [4]. PSMs above the cutoff can be either true positives or false positives, and PSMs below the cutoff can be either false negatives or true negatives. Most search engines use a protein database to define which proteins are expected in the sample, thereby reducing the search space significantly. De novo sequencing algorithms [5] and spectral library search engines [6] use no database or a spectral library database instead of a protein database. We did not use these types of search engines in this study since they are used less compared to search engines that rely on protein databases. Although database search engines use variations on the same principle, matching a measured to a theoretical spectrum, their respective search results differ even if the same data set is searched against the same sequence database [7]. Search engines provide different results because they generate different fractions of correct and incorrect PSM-assignments. Alternatively, this is probably the result of search engines differing in the number and type of correct assignments they make. Determining correctly identified peptides and wrongly identified peptides in each data set, respectively, is therefore important to evaluate the performance of a data analysis workflow. Most workflows rely on either manual, expert inspection of the search results or on software tools to estimate the proportion of false identifications. The manual assessment of the quality of PSMs is error-prone, dependent on the level of experience of the evaluator, inconsistent between evaluators and time-consuming [8]. For computer-based assessments of the quality of PSMs, there are two principal strategies that are primarily applied. The first uses statistical models as exemplified by PeptideProphet in the Trans-Proteomic Pipeline (TPP) [7,9] or Percolator [10], and the second uses a target-decoy strategy [3]. PeptideProphet relies on mixture models to integrate different types of information, such as the distribution of search engine scores, the likelihood that assigned peptides are present in the sample or the score difference between the best and second-best assignment of a spectrum. The mixture models are used to convert this information into search engine-independent scores, reflecting the probability that a particular PSM has been correctly assigned [11]. The principle behind a target-decoy strategy is based on the calculation of FDRs using the decoy part of the search database [4] to estimate how many false assignments are expected among the hits in the target-part of the database at some score cutoffs. The consistent determination of the FDR for different data sets provided either by statistical models or by target-decoy strategies is critical in making meaningful comparisons of different search engines and parameter sets [6,12–18]. To increase the fraction of correctly assigned spectra and to increase confidence in the reported results the output of multiple search engines was combined [19–22].

Although both methods for assessing the quality of search results are widely used, it still remains challenging to objectively evaluate the different analysis workflows, search tools and parameter sets and to prove that one has a better performance than the other. The difficulties arise for two major reasons. The first is the absence of a complex sample set of known composition although this difficulty is increasingly mitigated by the falling prices to create synthetic peptides. The second is the inability to systematically assess the influence of various perturbations to the analysis workflow. Most studies that present a new workflow use specific biological samples [23], or so-called spike-in samples [24], and increasingly complex synthetic samples [25,26] to evaluate its performance. Using a biological sample such as a digested cell lysate has the advantage that peptide-to-spectrum matching is carried out under realistic conditions, i.e. on a sample that contains thousands of peptides covering a wide range of signal intensities. However, these studies are limited because the true peptide composition of such samples is unknown, partly due to the presence of peptides that have post-translational modifications (PTMs) or that are the result of non-specific and missed cleavages [27]. It is often not possible to control these biological events preventing a reliable comparison of the search results. PSMs cannot be categorized as correct or incorrect with confidence because there is no evidence that the matching peptide truly exists in the measured sample and it is difficult to estimate how closely the generated set of identified peptides matches the maximally achievable set. An alternative to biological samples is the use of spike-in samples. Spike-in samples usually consist of a mix of a few dozen purified recombinant proteins of known sequence and quantity that are digested with trypsin to generate a peptide mixture of known composition. Such samples are then analyzed either by themselves, or in a complex background sample of unknown composition. The presence of PTMs is no longer a concern [28]. However, non-specific and missed cleavages, artifactual modifications generated during the sample processing and the presence of proteins introduced as minor contaminants of the purified reference proteins are still possible and, therefore, the peptide composition of such samples is still unknown, even if they are analyzed without added background. In addition, if analyzed without added background, the complexity of such samples does not match the complexity and intensity distribution of most biological samples. The spectra produced often contain fewer signals from peptides co-fragmented with the target peptides and, if these do occur, they show a lower signal intensity compared to that of the usual biological samples. These factors affect the complexity of the spectrum pattern and lower the threshold at which targeted peptides are correctly matched using a search engine. The second problem, that is, the influence of perturbations – such as variations in the parameter set – on the search results, is often not systematically addressed in the related studies because most data analysis workflows are not automated to a level where many different search parameter sets can be easily tested and compared.

Here we present a study on systematically varying parameters and search engines, in which we investigate the impact

these have on the sensitivity of the analysis of a dataset generated from a complex synthetic sample of more than 20,100 peptides, previously observed my MS in human samples. The complexity of the synthetic sample is sufficient for a realistic test and allowed us to accurately estimate both the sensitivity and specificity of the search results. The sample does however not mimic biological samples in terms of dynamic range. We propose a strategy to find optimal search parameters and present detailed information on how the various parameters influence the results. The peak list files and the identification results are publicly available in the PeptideAtlas repository (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?identifier=PASS00090).

2. Methods

2.1. Preparation of the benchmark sample

20,103 unique peptides were synthesized by JPT Peptide Technologies GmbH using SPOT-synthesis [28] (suppl. synthetic_peptide_sequences.txt) and crude synthesis products were aliquoted at a concentration of approximately 60 nmol/μl per peptide in 96 well microtiter plates. 5 μl from each well was used to create intermediate pools that were subsequently used to create a pool of all peptides, each at an estimated final concentration of about 3 pmol/μl per peptide.

2.2. Mass spectrometry

The synthetic peptide pool was measured on two liquid chromatography–tandem mass spectrometers (LC–MS/MS), an LTQ-FT Ultra (Thermo Fischer Scientific) coupled to a Tempo NanoLC (Applied Biosystems) and an LTQ-Orbitrap XL (Thermo Fischer Scientific) with a 1D-NanoLC-Ultra system (Eksigent). Both systems were equipped with a standard nanoelectrospray source and the chromatographic separation was performed with the same buffer system: 97% water, 3% acetonitrile and 0.1% formic acid constituted mobile phase A, while mobile phase B comprised 3% water, 97% acetonitrile and 0.1% formic acid. For each LC–MS/MS run, 2 μl of the peptide pool were injected onto an 11 cm × 0.075 mm I.D. column packed in-house with Magic C18 material (3 μm particle size, 200 Å pore size, Michrom Bioresources). The peptides were loaded onto the LC column at a flow rate of 300 nl/min and eluted with either of the two following gradients: Gradient #1: 0–5 min = 5% phase B solution, 5–95 min = linear gradient from 5–35% phase B solution, 95–97 min = linear gradient from 35–95% phase B solution and 97–107 min = 95% phase B solution. Gradient #2: 0–5 min = 5% phase B solution, 5–125 min = linear gradient from 5–35% phase B solution, 125–127 min = linear gradient from 35–95% phase B solution and 127–137 min = 95% phase B solution. The ion source and transmission settings for both mass spectrometers were: Spray voltage = 2 kV with capillary temperature = 200 °C, capillary voltage = 60 V and tube lens voltage = 135 V. All measurements of the synthetic peptide mix, on both the FT and Orbitrap instruments, were acquired in data-dependent mode, selecting up to five precursors from an MS1 scan (resolution for the FT: 100,000; resolution for the Orbitrap: 60,000) in a range of 350–1600 *m/z* for collision-induced dissociation (CID). The ion target values were 1,000,000 (or maximum 500 ms fill time) for full scans and 10,000 (or maximum 200 ms fill time (Orbitrap) and 300 ms (FT)) for fragment ion scans, respectively. Ions with a single or unknown charge state were automatically rejected. The synthetic peptide mix was measured in duplicates on either mass spectrometer, using each of the two specified gradients, resulting in a total of eight LC–MS/MS data sets. The mass spectrometers were equilibrated using a standard mixture before each sample injection.

2.3. Data analysis

The data analysis workflows rely on searching the acquired fragment ion spectra against a protein sequence database. We therefore derived a human subset from UniprotKB/Swiss-Prot, version 57.1[29], performing the following steps. Firstly, the complete UniprotKB/Swiss-Prot database was converted from its original DAT format into a FASTA file including all known splice variants and isoforms. For this process, we used uniprotdat2fasta.pl, which is part of InSilicoSpectro, a bioinformatics tools collection used for mass spectrometry [30]. Secondly, the subset of human protein sequences was extracted with subsetdb, which is part of the TPP [7]. In the final step, a target-decoy database with reverse decoys was generated using fasta-decoy.pl, another tool included in InSilicoSpectro. After generating the database, the data acquired from the mass spectrometers were converted into peak lists of the format mzXML [31]. This process was accomplished using ReAdW.exe, which is part of the TPP. Next, we created seven workflows in order to search these peak lists against the previously generated database. For three workflows this was done using a single database search engine, specifically Mascot [16] (version 2.3) (M), OMSSA [18] (version 2.1.7) (O) and X! Tandem [17] (version 2009.04.01, k-score plugin) (X); for a further three workflows two database search engines were combined, specifically Mascot-OMSSA (MO), Mascot-X! Tandem (MX) and X! Tandem-OMSSA (XO); and for the final workflow the three database search engines tested were combined (MXO). In each workflow, the search engine output was converted into the pepXML format and scored using PeptideProphet [7]. The final score distribution model was calculated with iProphet [20] and the result converted from its pepXML format into a simple text format (CSV) using pepxml2csv [32]. In addition to the iProphet score, we applied the target-decoy strategy and calculated the corresponding FDR values with fdr2probability [32]. The resulting data matrix was then imported into a MySQL database [33] to evaluate the performance of each data analysis based on the number of uniquely identified peptides (UIP) that matched a synthetic peptide sequence at 1% FDR.

Each of the seven workflows was tested with 54 different parameter settings to investigate the influence of the following search parameters: the precursor mass error (PME), the fragment mass error (FME), the number of allowed missed cleavages (MC) and the different scoring models of PeptideProphet. Values for the PME were set to 25 ppm, 15 ppm and 5 ppm, respectively. We chose values of 0.8 Da, 0.6 Da and 0.4 Da for the FME, respectively (Fig. 1). We also allowed a maximum of one or two missed cleavage events, respectively, and defined carbamidomethylation of cysteine as a

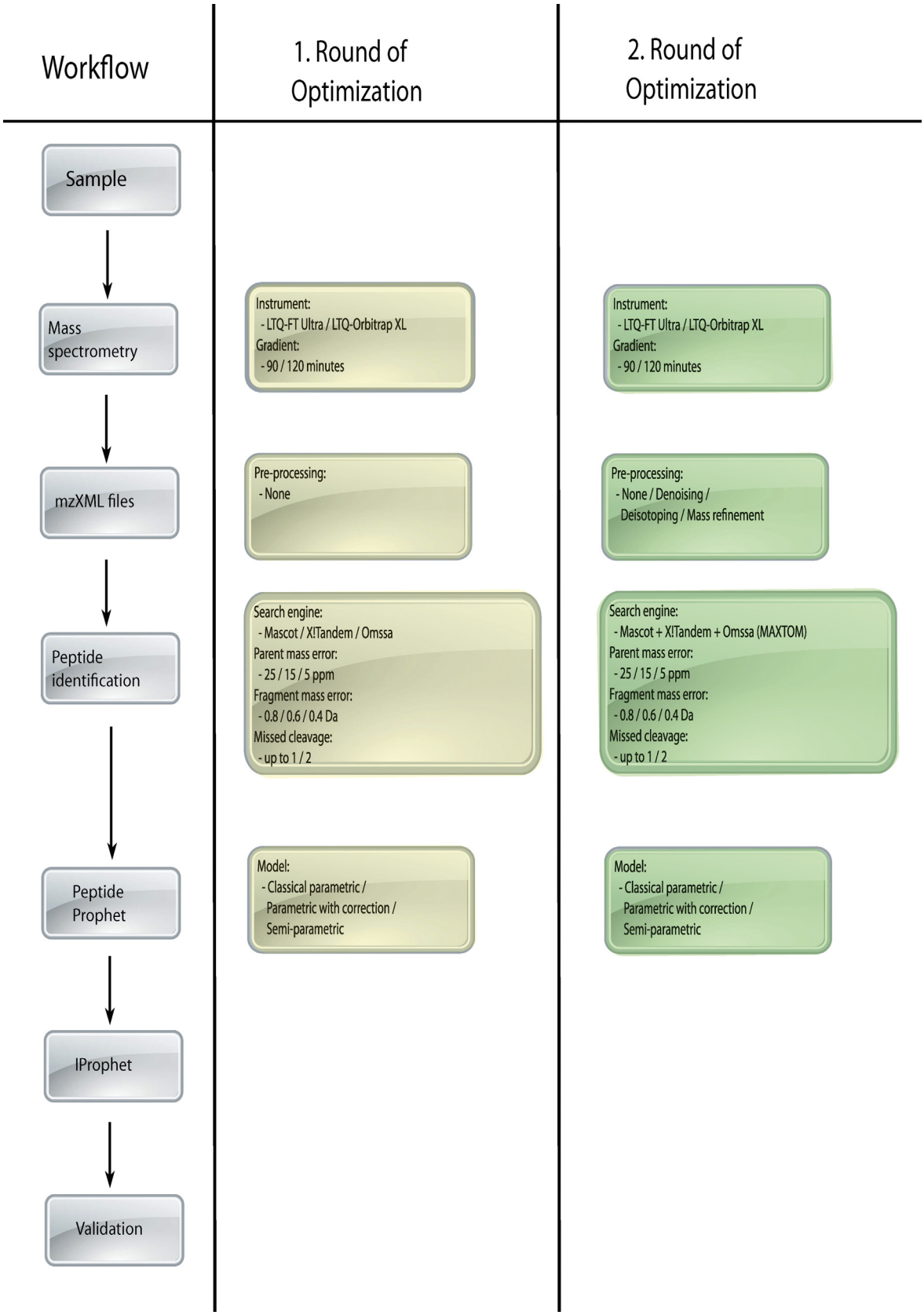

Fig. 1 – Overview of the experimental design. The diagram shows the steps that were systematically varied. The spectral pre-processing was only carried out on the optimal combination of search engines.

fixed modification. For PeptideProphet, the following statistical models were tested: the classical parametric model (CPM), the parametric model (PM) with correction of the negative distribution based on decoy hits (PMC) and the semi-parametric model (SPM). In total, 756 data analyses were produced in order to study the performance variation of single and multi-search engine workflows induced by changing parameter settings (suppl. Fig. S1) (Supplementary figures are available free of charge via the Internet at http://pubs.acs.org).

In order to investigate peak list pre-processing, we studied the effects of the following filters: MS2Deisotope (I), which deisotopes; MS2Denoise (N), which denoises; and precursorRefine (R), which refines the precursor ion mass; all operating on MS2 spectra. All three were implemented in msconvert (revision 2238), which forms part of ProteoWizard [34]. Since these filters have to be applied in the process of converting the original RAW format to a peak list file, we used msconvert instead of ReAdW.exe for this investigation. Based on the original instrument files, we generated peak list files in eight different filter combinations, one using no filter, three using a single filter, three where two filters were combined and one where all three filters were combined. The peak list files of each filter combination were tested with the MXO-workflow using the same sequence database and the same 54 parameter settings as were used for the previous data analyses, plus an additional set of parameter combinations (suppl. Figs. S2 and S10). In total, 1074 searches were performed to investigate the effect of pre-processing the peak lists on workflow performance.

Detection of peptide synthesis by products was carried out by generating a database with all permutations of single amino acids missing. This database was then used as described above.

OpenMS MS1 feature detection and peptide quantification was carried out as described by Weisser et al. [35].

All the data analyses presented in this study were automated and executed using the workflow system P-GRADE [36].

3. Results

The benchmark study we present was performed in three steps. Firstly, we explored the influence of critical search parameters on the performance of individual search engines. Secondly, we tested how the combinations of the results from different search engines affected the identification performance. Thirdly, we measured the influence of pre-processing the peak lists on search performance. The results of the analyses we performed were compared based on the number of true positive UIPs at 1% FDR. A schematic of the workflow used to carry out all three steps is presented in Fig. 1 and a summary of all the results is given in Table 1.

3.1. *Single search engine performance*

We generated three workflows each incorporating a single search engine, one based on Mascot (M), one on OMSSA (O) and one on X! Tandem (X). The X! Tandem K-score plug-in was used in combination with X! Tandem throughout. We assessed each respective performance by changing parameters such as the instrument type, mass errors for the fragment ion (FME) and for the parent ion (PME; see Fig. 1).

Across all the parameter settings tested, the number of identified peptides was consistently higher for data acquired on the Orbitrap instrument compared to data from the FT (suppl. Figs. S3–5). This observation should not be interpreted as one type of instrument being superior to another, it rather simply means that one instrument was better optimized compared to the other at the time of acquisition. The other parameters that had a major influence on search performance were the mass errors FME and PME. The results indicated that lowering the FME increased the search performance, but that reducing the PME below 15 ppm caused a loss of performance (suppl. Figs. S6 and S7). The data for the single database search engine workflows are shown in Fig. 2. Each boxplot in Fig. 2 consists of 54 independent search results, one per parameter set. The figure shows the number of correct UIPs. Two out of the 756 data analyses failed because the qualitative requirements for modeling the negative peptide distribution in PeptideProphet were not fulfilled.

The best performance among the single search engine workflows was achieved with OMSSA (Table 1). A total of 6489 correct UIPs were found when a small FME of 0.4 Da and a moderate PME of 15 ppm were used. Non-optimal parameter settings, such as larger mass errors (FME: 0.8 Da, PME: 25 ppm), caused a drop in the number of correct UIPs to 614, the lowest identification rate in this study. In addition, the use of a smaller mass error is sub-optimal, as using a PME of 5 ppm resulted in 5900 correct UIPs, compared to the 6489 identifications at 15 ppm PME. Another parameter with a noticeable impact on performance was the type of mass spectrometer used. Applying the same mass error settings resulted in a drop from 6489 correct UIPs (Orbitrap) to 6097 correct UIPs (FT).

The Mascot (M) workflow identified a maximum of 6401 correct UIPs (Table 1) over all parameters tested. In contrast to OMSSA, Mascot performed better with a less strict PME of 25 ppm in conjunction with a small FME of 0.4 Da, although the performance difference between a PME of 15 ppm and 25 ppm was insignificant, with 6401 compared to 6391 correct UIPs, respectively. Other parameters had a smaller effect on the performance.

The third search engine we tested was X! Tandem. Compared to the O- and M-workflows, X! Tandem was more robust when there were changes of the mass errors FME and PME. In particular, varying FME did not have a significant impact. For example, the maximal performance of 6219 correct UIPs was achieved when applying a PME of 15 ppm, but did not change when the FME was varied between 0.4, 0.6 and 0.8 Da. The same finding applied to data analyses with a PME of 25 ppm and 5 ppm, which resulted in 6170 and 5856 correct UIPs, respectively, and these were unaffected by variation of the FME within the range tested (Table 1).

3.2. *Performance of multiple search engine searches*

We combined the output of multiple search engines to investigate if the combined output could improve search performance. All two-way combinations (MO, MX, XO) and the combination of all three (MXO) were tested using the same parameter settings used for the single search engines. The

Table 1 – Performance of the various tool combinations. Each tool combination was tested using several parameter sets.

Search engines	Abbreviation	Max/Min of correct UIPs	Max of correct Ups (%)	Most influential parameter	Best parameters PME/FME/MC/PM	Note
OMSSA	O	6489/614	32.28	FME	15/0.4/1/SPM	Best single engine
MASCOT	M	6401/5330	31.84	PME/FME	25/0.4/1/SP	Most robust engine w.r.t. MS
X! Tandem	X	6219/5244	30.94	MS	15/0.4–0.8/1/CPM	Most robust engine w.r.t. PME/FME
Mascot/OMSSA	MO	6674/5240	33.20	PME/FME	15/0.4/1/CPM	
Mascot/X! Tandem	MX	6595/5330	32.81	PME/FME	25/0.4/1/PMC	
X! Tandem/OMSSA	XO	6769/5510	33.67	PME/FME	15/0.4/1/CPM	
Mascot/X! Tandem/OMSSA	MXO	6814/5846	33.90	PME/FME	25/0.4/1/CPM	Best search combination
Denoise	N	6807/5890	33.86	PME/FME	15/0.4/2/CPM	
Deisotop	I	6802/5844	33.84	PME/FME	15/0.4/1/CPM	
Refine	R	6910/5821	34.37	PME/FME	5/0.4/2/PMC	Effective for LTQ-Orbitrap
Denoise/Deisotop	NI	6828/5919	33.97	PME/FME	15/0.4–0.6/1/CPM	
Refine/Denoise	NR	6910/5916	34.37	PME/FME	5/0.4/2/PMC	Best preprocessing for FT
Refine/Deisotop	IR	6909/5869	34.37	PME/FME	5/0.4/1/PMC	
Refine/Deonise/Deisotop	NIR	6938/5945	34.51	MS	5/0.4–0.6/1/CPM	Best preprocessing for LTQ-Orbitrap

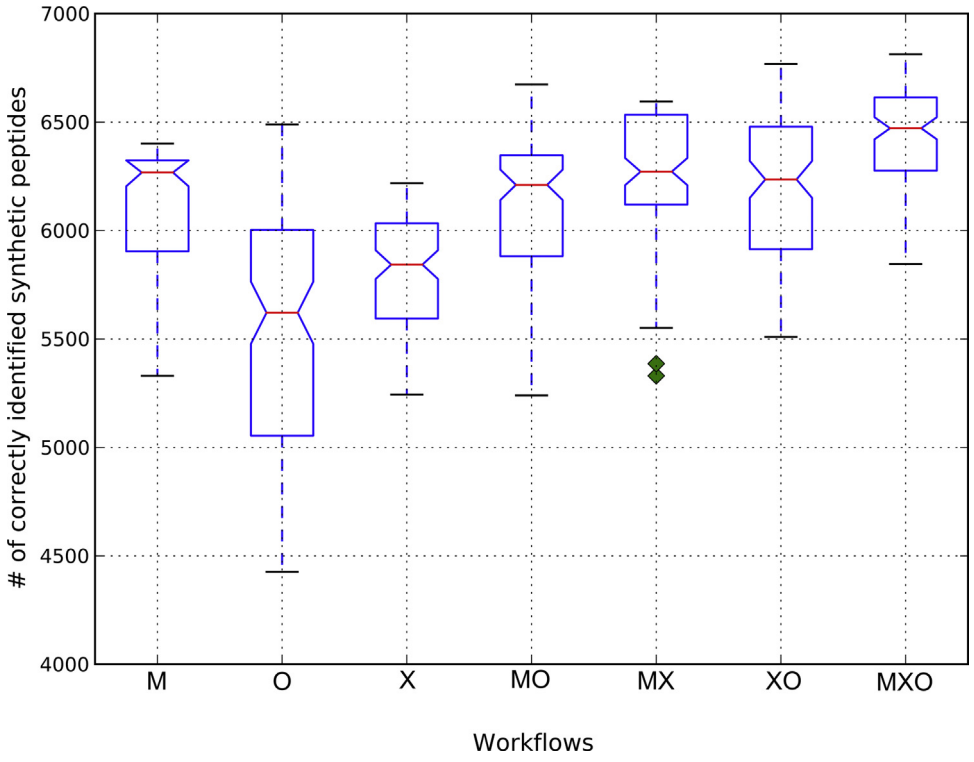

Fig. 2 – The number of correctly identified peptides per workflow is shown in a box plot representation. A total of 54 parameter sets and two mass spectrometer types were used for each workflow. Different single database search engines were used, Mascot (M), OMSSA (O) or X! Tandem (X), respectively, in addition to combinations of two or three search engines: Mascot-OMSSA (MO), Mascot-X! Tandem (MX), X! Tandem-OMSSA (XO) and Mascot-X! Tandem-OMSSA (MXO). The upper whisker indicates the number of peptides identified using an optimal parameter set and the red line marks the mean number of peptides identified for the parameter sets tested within a workflow. The box itself circumscribes the search results between the first and the third quartile. The larger the spread, the more sensitive was the search to the parameters. The green daggers mark measurements which are outside of the range between first and third quartile. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

search engine results were combined using iProphet [20]. The results are presented in Fig. 2 and Table 1. The data showed an improved performance for multi-search engine workflows compared to single-engine workflows.

The effects of combining search results are apparent from the XO-workflow, which combined the fairly robust X! Tandem engine with the more sensitive OMSSA tool. The combined workflow outperformed the single engine results of OMSSA (maximal 6489 correct UIPs) and X! Tandem (maximal 6219 correct UIPs), with a maximum of 6769 correct UIPs under optimal search parameters (PME = 15 ppm and FME = 0.4). The XO-workflow achieved 5510 correct UIPs using the least optimal parameters tested, which was better than both the X-workflow with 5254 correct UIPs and the O-workflow with 614 correct UIPs (Table 1). In the XO-workflow, X!Tandem largely compensated for the poor performance of OMSSA in cases in which sub-optimal parameters were being used, and the performance increased above the level of a single search engine when optimal parameter settings were applied. Similar trends were observed for the two other two-engine combinations.

We also tested the combination of the three database search engines in a single workflow (MXO), which resulted in 6814 correct UIPs, the highest identification rate of the seven workflows, if a moderate PME of 15 ppm and a small FME of 0.4 Da was used (Table 1). Additionally, with sub-optimal search parameter settings, such as a PME of 5 ppm and a FME of 0.8 Da, a minimum of 5846 correct UIPs was scored. This is significantly higher when compared to the other workflows. The resulting spread of 14.2% was the lowest of all workflows and therefore indicated that the MXO-workflow was the least dependent on the search parameter settings. Fig. S12 displays pseudo-receiver operating characteristic curves for the optimal parameter settings for each of the seven search engine combinations. Fig. S11 show a VENN diagram comparing the results of three individual search engines with the MXO-workflow.

3.3. *Effect of peak list pre-processing*

We investigated three types of peak list pre-processing: deisotoping (filter = MS2Deisotope), denoising (filter = MS2Denoise) and refining the mass of the precursor ion (filter = precursorRefine). We tested the performance of these filters using a single filter, the combination of two filters and combining all three filters, and compared these results to results generated without a filter. The data was processed with the MXO-workflow and the same 54 parameter settings we used previously (suppl. Figs. S1 and S2) plus additional parameter sets investigating more narrow parent and fragment mass errors (suppl. Fig. S10). The results of the pre-processing benchmark are presented in Fig. 3.

Pre-processing with MS2Denoise and MS2Deisotope showed no performance improvement compared to the 'no filter' setting. This is supported by the 6807 and 6802 correct UIPs achieved at PME = 15 ppm and FME = 0.4 Da, respectively, in comparison with the 6805 correct UIPs obtained with no filter. However, the data indicated that both filters partly compensated for sub-optimal settings. The minimal number of correct UIPs increased if a PME of 5 ppm with a FME of 0.4 Da was used (for no filter: $n = 5812$, for MS2Denoise: $n = 5890$ and for MS2Deisotope: $n = 5844$) (Fig. 3, Table 1). The same was observed if both filters were combined: The maximal identification rate of 6828 correct UIPs (PME = 15 ppm, FME = 0.6 Da) was no improvement compared to the 'no filter' settings. In contrast, the lowest identification rate, for which a PME of 5 ppm and a FME of 0.4 Da was applied, increased from 5812 correct UIPs (no filter) to 5919 correct UIPs (MS2Deisotope + MS2Denoise).

The use of precursorRefine had a noticeable impact on the performance (Fig. 3, suppl. Fig. S8), as the maximal number of correct UIPs increased from 6805 (no filter) to 6910 (precursorRefine) (Table 1). Moreover, the top identification rate was achieved when small mass errors were used (PME = 5 ppm, FME = 0.4 Da, instrument = Orbitrap). The lowest identification rate, 5821 correct UIPs, was also achieved by using small mass errors (PME = 5 ppm, FME = 0.6 Da, instrument = FT). For data from the Orbitrap, identification rates increased and the impact of parameter settings was reduced. Specifically, the range of 3.2% was reduced by over 50% compared to the wider range of 8.13% for the 'no filter' settings. We did not observe the same positive effect for data obtained from the FT, where the results were more comparable to those obtained when no filters were used (suppl. Fig. S9).

We tested the combination of all three pre-processing filters and this combination performed best, with 6938 correct UIPs, when using PME = 5 ppm and FME = 0.4 Da (Table 1). Using smaller mass errors did not lead to further improvements.

3.4. *Benchmark sample*

We pooled 20,103 synthetic peptides to create the test sample (see Section 2). Analyzing the sample on the FT and on the Orbitrap generated about 11,500 spectra per file for a 90-min gradient and around 13,000 spectra per file for a 120-min gradient, respectively (Table 2).

The use of a synthetic pool allowed us to identify assignments above the chosen cutoff as correct if they matched the synthetic peptide sequences. However, the type of synthetic peptides we used limited us in drawing further conclusions about potentially incorrect PSMs. Although we could certify a PSM as being correct, it was not possible to reject PSMs as truly false when they did not match a synthetic peptide sequence. Each of the 20,103 crude synthetic peptides that we pooled in order to generate our sample data set were produced using SPOT synthesis, which means that traces of by-products could also be found in the pool. By-products are dominated by peptide sequences which have one or more gaps (missing amino acids) at certain positions in the targeted peptide sequence [37]. The quantity of these by-products is usually significantly lower than the amount of targeted peptide sequences [37]. This does not mean, however, that these peptide variants cannot be detected by mass spectrometry. Depending on the workflow, up to 9% of all uniquely identified peptides (UIPs) on a 1% FDR level matched one of these by-products (up to 6.15% matched a one-gap by-product and up to 2.7% matched a two-gap by-product). We performed two label-free quantification experiments using OpenMS [38], one with the default parameters and one with relaxed quality criteria for feature detection. In both cases, we detected features for the original synthetic

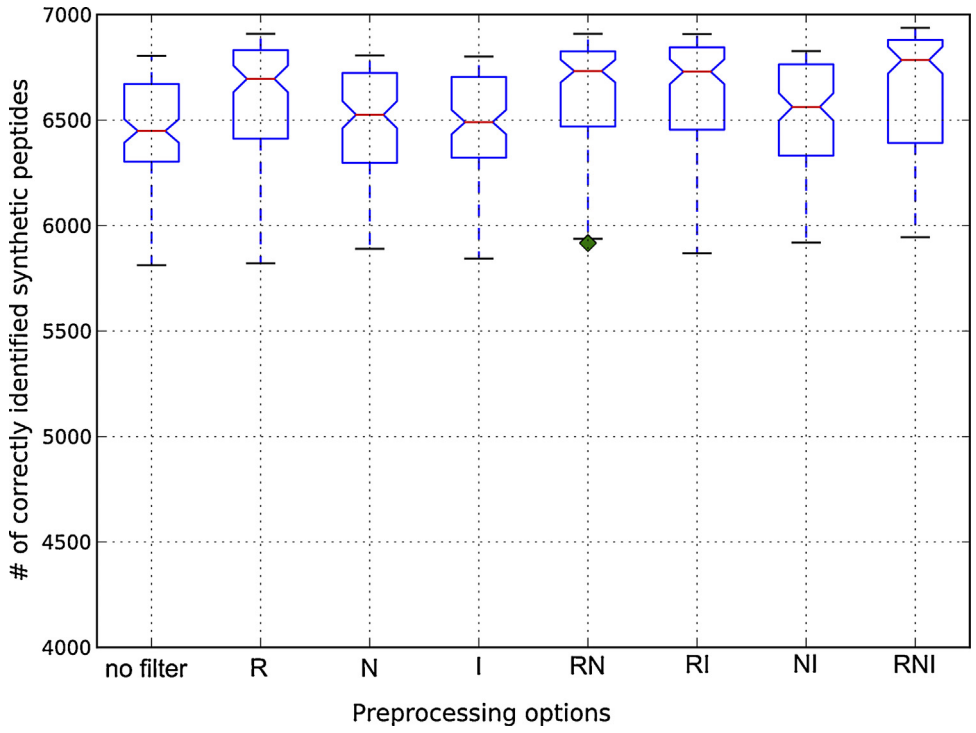

Fig. 3 – Box plots of the search results (UIPs) for different pre-filtering options. Each box plot contains the results from 108 database searches (54 different parameter settings, each for data files acquired with two different mass spectrometers). Pre-filtering tools were MS2Denoising (N), MS2Deisotoping (I) and precursorRefine (R), which were applied either individually or in various combinations (NI, NR, IR, NIR). The green daggers represent data points outside the range between the first and third quartile. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

peptides (default parameter: $n = 5843$; less restrictive parameter: $n = 6000$) and by-products with up to two gaps (default parameter: $n = 6027$; less restrictive parameter: $n = 6188$).

4. Discussion

Here, we present a large-scale study investigating strategies to improve MS/MS peptide identification and to make database searches more robust. A complex pool of synthetic peptides was created in order to comprehensively benchmark the different analysis workflows based on the number of correctly identified peptide sequences from this well characterized, but complex, sample. The use of synthetic peptides as a reference sample addressed issues of other sample types, such as complexity and certainty about the true positive matches. This allowed an improved interpretation of the benchmark results that was independent of estimations based on decoy databases. We investigated the impact of 108 different search parameter settings on the identification performance of three search engines, Mascot, X! Tandem and OMSSA, the effect of combining multiple search engines in a single approach and the processing of ion mass spectra prior to database searching, totaling around 1800 distinct combinations.

Firstly, we tested and optimized the parameter settings for each search engine individually. Overall, the data showed that Mascot, X! Tandem and OMSSA performed at a comparable level if optimal search parameters were used. However, there were distinct performance differences using other parameter sets. OMSSA achieved the best performance of the search engines tested but was most sensitive to the parameter settings. Mascot was more robust in the face of parameter changes

Table 2 – Sample properties for measurements of the synthetic peptide data set.

Measurement	Mass spectrometer	LC-gradient (min)	MS1 scans	MS2 scans
1	LTQ-FT Ultra	90	4899	11,491
2	LTQ-FT Ultra	90	4890	11,752
3	LTQ-FT Ultra	120	6570	12,963
4	LTQ-FT Ultra	120	6509	12,760
5	LTQ-Orbitrap XL	90	4036	11,445
6	LTQ-Orbitrap XL	90	4040	11,244
7	LTQ-Orbitrap XL	120	5228	13,819
8	LTQ-Orbitrap XL	120	5206	13,795

but was negatively affected when a small PME was selected and did not benefit to the same extent from the increased sensitivity of an Orbitrap compared to an FT. X! Tandem benefited most from the increased instrument performance and was least affected by parameter changes. It did, however, identify fewer correct UIPs under optimal conditions than the other search engines. This is likely a result of X! Tandems algorithm that benefits from finding more than one peptide from each protein. Generally, the choice of optimal parameter settings was more important and had a larger influence on the number of correctly identified PSMs than did the choice of the search engine itself.

We continued our benchmark study by testing combinations of the different search engines in a single workflow. Here, the data supported two main conclusions. Firstly, integrating the results from multiple search engines using iProphet reduced the impact of sub-optimal parameter settings and, secondly, it helped to increase the number of UIPs identified. The MXO-combination performed better than the other workflows tested, independently of the applied parameter sets. However the data also indicated that performance does not increase linearly with the number of search engines added to a workflow. Under optimal conditions, the XO-workflow performed almost as well as the MXO-workflow, suggesting that these combinations, used optimally, identify close to the maximal number of detectable UIPs in a data set using the current technology. In general, we observed that each workflow reached maximum identification rates when a PME of 15–25 ppm was used in conjunction with a small FME of 0.4 Da. Lowering the PME, i.e. to 5 ppm, led to a decrease in the number of correct identifications for all the workflows tested.

After identifying the optimal combination of search engines, we investigated the potential benefit of pre-processing peak lists prior to their processing with search engines. We postulated that pre-processing of ion mass spectra leads to further improvements in the identification performance while significantly decreasing the impact of any sub-optimal parameter settings. Data from the LFQ-FT Ultra were best analyzed with a combination of MS2Denoise and MS2Deisotope in conjunction with a moderate PME of 15 ppm and a FME of 0.6 Da. The optimal pre-processing strategy for data obtained from the Orbitrap was to apply all three filters, namely, MS2Denoise, MS2Deisotope and precursorRefine, together with a small PME of 5 ppm and a FME of 0.4 or 0.6 Da. Lowering the errors for the parent mass or the fragment mass did not further increase the performance.

The peptides used in this study are the so-called crude peptides, in which some peptide-based by-products are present at a high enough concentration to be detectable by MS. X! Tandem detected a higher percentage of by-products than did Mascot or OMSSA because its scoring algorithm does not rely on information about the peak intensity of the precursor ions. Therefore, spectra with weaker signals (as expected for by-products) are not and are therefore reported as PSMs. That these PSMs are not false positive assignments is supported by two facts: Firstly, the other two identification tools, Mascot and OMSSA, also detected some of the by-products at 1% FDR on the PSM-level and, secondly, we were able to quantify some of these by-products by label-free quantification with OpenMS. Although we demonstrated that a small proportion of the synthesis by-products is detectable and is probably correctly assigned, we could not determine whether these matches were truly correct. This is why we only considered PSMs to be correct when they matched one of the 20,103 synthetic peptide sequences.

Several recent publications have demonstrated that the results from pure decoy databases and the use of entrapment databases differ [39,40]. As it is still unclear how to best select the entrapment database, both in terms of number of sequences but also which organisms that are sufficiently diverged from the organism under study. We have refrained from using entrapment databases and opted to use the more traditional reverse decoy database [3]. It is also possible to use much wider parent mass windows and instead rely on filtering the results [41]. This option was not explored. Some parameters had no impact on the results as exemplified by the peptideProphet models despite previous reports on the topic [42]. It is still unclear why our results are not in line with the literature and should be explored further. Both mass spectrometers used in this study have relatively similar specifications and it is unclear how the findings here would hold for other types of instruments [43].

In summary, the results of our benchmark study showed that the correct choice of parameter settings has a large influence on the identification performance of the search engines we evaluated. The search parameter combinations tested led to identification rates of between 36% and 93% on the PSM-level and between 14% and 90% on the UIP-level. We assessed the influence exerted by each of the parameters and found that the mass spectrometer type and the allowed mass errors (PME, FME) have the largest impact on variations in the results. There are other untested parameters such as the size of the protein database, which can influence the result. In general, the search engines performed better when the mass spectrometer had a higher sensitivity, such as the Orbitrap, and when mass errors of 15 ppm (PME) and 0.4 Da (FME) were used. Reducing both mass errors further, especially the PME, led to a decreased identification rate. Other search parameters, such as the number of allowed missed cleavages and the statistical models of PeptideProphet, only had a minimal impact on improving the results from the database search engines. These results demonstrated that using advanced identification workflows was the key to successfully improving the understanding of the measured data while keeping the false-positive hits to a minimum.

Conflict of interest

The authors declare no competing financial interests.

Transparency document

The Transparency document associated with this article can be found in the online version.

Authors' contribution

AQ and LM conceived the project and wrote the initial manuscript. AQ undertook the experiment planning, sample

preparation, workflow definition and data analysis. LE wrote prototypes of the nodes used in the tested workflows. LE also wrote the post-processing tools applied in the pipeline. AB implemented the concept of super-workflows into P-GRADE, which was used to perform the data analyses used in this study. HW prepared the OpenMS processing workflow used for the label-free quantification of the benchmark sample. MB provided early stage access to the pre-processing algorithms in msconvert and wrote parts of the Methods section explaining these filters. PK supported the project with resources and wrote parts of the manuscript. RA supported and financed this project, gave valuable scientific input and wrote major parts of this manuscript.

Acknowledgments

The authors would like to thank Christopher Paulse for implementing the pre-processing filters in msconvert. We would also like to thank Oliver Rinner for helping prepare the crude peptides mixture used in this paper, as well as Alexander Leitner for fruitful discussions regarding the instrument settings and handling and Paola Picotti for discussing matters regarding the synthetic peptides. We would also like to extend thanks to the SyBIT project of the SystemsX.ch intiative and the Brutus system administrators for support with computing infrastructure and other IT-related resources.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.euprot.2014.10.001.

REFERENCES

[1] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4(10):787–97.
[2] Matthiesen R. Methods, algorithms and tools in computational proteomics: a practical point of view. Proteomics 2007;7(16):2815–32.
[3] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4(3):207–14.
[4] Benjamini YHY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Met 1995;57(1):289–300.
[5] Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 2005;77(4):964–73.
[6] Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 2007;7(5):655–67.
[7] Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 2005;1:0017.
[8] Cottingham K. Manual validation is a hot proteomics topic. Anal Chem 2005;77(5):92.
[9] Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, Carapito C, et al. Trans-proteomic pipeline supports and improves analysis of electron transfer dissociation data sets. Proteomics 2010;10(6):1190–5.
[10] Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007;4(11):923–5.
[11] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74(20):5383–92.
[12] Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005;77(14):4626–39.
[13] Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994;5(11):976–89.
[14] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003;3(8):1454–63.
[15] Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 2007;6(2):654–61.
[16] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20(18):3551–67.
[17] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20(9):1466–7.
[18] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. J Proteome Res 2004;3(5):958–64.
[19] Quandt A, Masselot A, Hernandez P, Hernandez C, Maffioletti S, Appel RD, et al. SwissPIT: an workflow-based platform for analyzing tandem-MS spectra using the Grid. Proteomics 2009;9(10):2648–55.
[20] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 2011;10(12). M111.007690.
[21] Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS. Rapid and accurate peptide identification from tandem mass spectra. J Proteome Res 2008;7(7):3022–7.
[22] Nahnsen S, Bertsch A, Rahnenfuhrer J, Nordheim A, Kohlbacher O. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. J Proteome Res 2011;10(8):3332–43.
[23] Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res 2008;7(9):3838–46.
[24] Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, et al. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. J Proteome Res 2008;7(1):96–103.
[25] Ivanov AR, Colangelo CM, Dufresne CP, Friedman DB, Lilley KS, Mechtler K, et al. Interlaboratory studies and initiatives developing standards for proteomics. Proteomics 2013;13(6):904–9.
[26] Marx H, Lemeer S, Schliep JE, Matheron L, Mohammed S, Cox J, et al. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat Biotechnol 2013;31(6):557–64.
[27] Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 2007;6(9):1589–98.

[28] Frank R. The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports – principles and applications. J Immunol Methods 2002;267(1):13–26.
[29] The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 2009;37(Database issue):D169–74.
[30] Colinge J, Masselot A, Carbonell P, Appel RD. InSilicoSpectro: an open-source proteomics library. J Proteome Res 2006;5(3):619–24.
[31] Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004;22(11):1459–66.
[32] SyBIT http://www.sybit.net
[33] MySQL http://www.mysql.com
[34] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24(21):2534–6.
[35] Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, Brauer H, et al. An automated pipeline for high-throughput label-free quantitative proteomics. J Proteome Res 2013;12(4):1628–44.
[36] Farkas ZKP. P-GRADE portal: a generic workflow system to support user communities. Future Gener Comput Syst 2011;27(5):454–65.
[37] Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods 2010;7(1):43–6.
[38] Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS – an open-source software framework for mass spectrometry. BMC Bioinform 2008;9:163.
[39] Granholm V, Noble WS, Kall L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J Proteome Res 2011;10(5):2671–8.
[40] Vaudel M, Burkhart JM, Breiter D, Zahedi RP, Sickmann A, Martens L. A complex standard for protein identification, designed by evolution. J Proteome Res 2012;11(10):5065–71.
[41] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24(10):1285–92.
[42] Ma K, Vitek O, Nesvizhskii AI. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. BMC Bioinform 2012;13(Suppl. 16):S1.
[43] Colaert N, Degroeve S, Helsens K, Martens L. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 2011;10(12):5555–61.