When Less Can Yield More - Computational Preprocessing of MS/MS Spectra for Peptide Identification

Technical Report

Bernhard Y. Renard[1,2], Marc Kirchner[1,2,3]*, Flavio Monigatti[2,3]*, Alexander R. Ivanov[4], Juri Rappsilber[5], Dominic Winter[2,3], Judith A. J. Steen[2,6], Fred A. Hamprecht[1,2], Hanno Steen[2,3,#]

[1] Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany
[2] Proteomics Center, Children's Hospital, Boston, MA, USA
[3] Dept. of Pathology, Harvard Medical School and Children's Hospital, Boston, MA, USA
[4] Dept. of Genetics and Complex Diseases, Harvard School of Public Health, Boston, MA, USA
[5] Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, UK
[6] Dept. of Neurobiology, Harvard Medical School and Dept. of Neurology, Children's Hospital, Boston, MA, USA
* authors contributed equally
# corresponding author: Hanno Steen, Children's Hospital Boston, Department of Pathology, Enders 1130, 320 Longwood Avenue, Boston, MA 02115, Fax: +1-617-730-0168, email: hanno.steen@childrens.harvard.edu

Summary

The effectiveness of database search algorithms such as Mascot, Sequest and ProteinPilot is limited by the quality of the input spectra: spurious peaks in MS/MS spectra can jeopardize the correct identification of peptides or reduce their score significantly. Consequently, efficient preprocessing of MS/MS spectra can increase the sensitivity of peptide identification at reduced file sizes and run time without compromising its specificity. We investigate the performance of 25 MS/MS preprocessing methods on various data sets and make software for improved preprocessing of mgf/dta-files freely available from http://hci.iwr.uni-heidelberg.de/mip/proteomics or http://www.childrenshospital.org/research/steenlab.

Mass spectrometry (MS)-based protein identification is one of the key elements for the understanding of biological systems. Currently applied protein identification methods compare experimental spectra to theoretical spectra created from sequence databases [1, 2]. Commonly used methods include various commercial implementations, in particular Mascot [3], (Sorcerer-) Sequest [4] and ProteinPilot [5].

Only relatively few publications have addressed the preprocessing of MS/MS spectra prior to their submission to a database search and three groups can be distinguished: i) spectral quality scoring, ii) precursor preprocessing and iii) MS/MS spectra preprocessing.

i) Spectral quality scoring methods select high quality spectra for further processing, but do not change the selected spectra themselves [6]. Such spectral quality scoring is generally feature- or clustering based (e.g. [7, 8, 9, 10]).

ii) Precursor preprocessing approaches focus on enhancing the MS1 information e. g. by identifying the precursor charge state [11, 12, 13, 14]. Gentzel et al. [11] also address the problems of centroiding, spectra joining and filtering as well as automatic calibration. Mascot Distiller, MS Cleaner [15] and DTASuperCharge [14] additionally offer the removal of multiply charged ions, deisotoping and background noise removal.

iii) For the MS/MS spectra preprocessing, different problem-specific approaches have been developed to identify a subset of peaks in a given MS/MS spectrum that is worth submitting to further searches: Tailored filters for peaks were suggested for the OMSSA [16] and InSpect [17] packages, rudimentary preprocessing by intensity-based cut off thresholding for database searches is included in mzStar [18] and wiff2data [19]. Another approach is realized by MaxQuant [20] which selects the 6 most intensive peaks within 100 *m*/z intervals in its default setting besides identifying the charge state and correcting the monoisotopic masses of the precursors.

MS/MS spectra normally show more ions than expected from the fragmentation processes [15], and thus do not have an optimal signal-to-noise ratio [6]. Preprocessing of MS/MS spectra themselves – with regard to the question which peaks are submitted to the search – can improve the signal-to-noise ratio by removing peaks which most likely do not belong to the expected b- or y-fragment ion series. A removal of such peaks reduces the risk of false identification, possibly increases the score for correct identification and at the same time decreases run time and file sizes. Hence, peak elimination methods are an extreme case of peak intensity modification approaches [8] and can provide the benefit of increased identification quality at significantly reduced file sizes and run time.

Many researchers recur to individually found heuristics such as submitting only a pre-specified number of the highest intensity peaks from the MS/MS spectrum to the database search (e. g. [21]) whereas other labs usually run searches without any preprocessing.

This study describes the testing of various procedures for the MS/MS preprocessing and rigorous comparison of their performances with regard to the number of peptides identified on various datasets using Mascot, Sequest and ProteinPilot. This allows us to identify an optimized preprocessing procedure for each search engine. The goal of this study is to focus on the comparison of MS/MS preprocessing methods, but not to compare the performance of the search engines (see e. g. [22, 23] for such comparisons) or precursor preprocessing approaches. Since preprocessing methods result in different signal-to-noise characteristics and algorithms weight these characteristics in varying – and often unknown – ways, preprocessing can only be empirically and separately

optimized for each search algorithm. This study is directed towards the optimization of MS/MS preprocessing of low resolution data as they are generated by quadrupole or quadrupole ion trap instruments. The preprocessing of high resolution/high accuracy MS/MS data is significantly different as noise peaks can easily be determined based on accurate mass increments and charge state information. Similarly, modification that give rise to characteristic fragmentation patterns such phosphopeptides and their distinctive neutral loss of phosphoric acid might benefit from a more problem-tailored preprocessing that can be derived from the proposed strategy.

The unifying idea behind MS/MS preprocessing approaches is that in common collision-activated dissociation MS/MS spectra standard fragment ions like b- or y-fragment ions tend to have higher intensities than other fragment ions in their neighborhood or noise signals [16, 17]. Furthermore, within a certain window around a given high intensity peak, only a limited number of fragment ions can be present. This fact can be taken advantage of by disregarding further low intensity peaks.

The MS/MS spectra preprocessing methods used in this study can be grouped into the following categories (also see **Figure 1**).

'Top X intensity' approaches: The simplest MS/MS preprocessing method (e. g. [21]), 'Top X intensity', sorts all ions in a MS/MS scan by decreasing intensity and only keeps the first X ions. If there are less than X ions, all existing ions are selected (**Figure 1a**).

'Top X intensity in Y regions' approaches: To alleviate the problem that high intensity peaks cluster in one part of the spectrum (e.g. around the precursor) dominate the preprocessing, we use a 'Top X intensity in Y regions' approach. There, the spectrum is first split into Y equal sized regions in the *m/z* domain (with a 2.5 *m/z* overlap) and for each of the resulting regions, a 'Top X' approach is applied. The resulting peak lists are merged and possible duplicates in the lists resulting from the overlaps are removed (**Figure 1b**). A parameterization of (X,Y) refers to 'Top X intensities in Y regions'.

'Top X intensity in a window of ± Z' approaches: These approaches [16, 17] sort all peaks by decreasing intensity. Starting with the highest intensity peak, a window of **±** Z *m/z* to the left and right of that peak is computed. Among all peaks within this window, only the top X most intense peaks are retained for further analysis, whereas the peaks remaining within the window are excluded from further analysis. This process is repeated until all peaks have either been selected or discarded (**Figure 1c**). The corresponding parameterization is given as (X,Z). Since the window is defined to the left and right of the peak, the resulting interval has twice the size of the window. InSpect [17] utilizes such an approach with a window of **±** 25 *m/z* and a selection of the six most intensive peaks. Similarly, OMSSA [16] selects the one or two most intensive peaks within a window of **±** 27 *m/z* (for a precursor of charge 2) or **±** 14 *m/z* (for a precursor of charge 3).

For all three approaches, sets of parameters X, (X,Y), (X,Z) were spread on a wide grid with values determined from literature [16, 17, 20, 21] and preliminary experiments on different datasets. We then sampled with a finer grid around maxima to pinpoint the exact position.

Starting values for the parameterization were chosen to be 100, 150 and 200 for the 'Top X' approach, (50,3), (30,5), (25,6), (20,8) for the 'Top X in Y regions' approach and (4,40) , (4,50), (4,60), (6,30), (6,40), (6,50), (6,60), (8,40), (8,50), (8,60) for the 'Top X in window of **±** Z' approaches. We also

included a parameterization inspired by InSpect (6,25) and OMSSA (1,27 or 2,14 depending on the availability of precursor charge state information, but always denoted as Top 1 in window 27 in the following). For comparison of the additional effects of preprocessing, spectra without external preprocessing ('no preprocessing') were analyzed as well; in this case unadulterated .mgf or .dta files containing all peaks were submitted to the search engines and their internal filtering routines.

DTASuperCharge v. 1.01, MaxQuant and Mascot Distiller were run in their respective default settings (for LTQ and LTQ-Orbitrap). All three approaches are based on both, precursor and MS/MS spectra preprocessing; the effects of these two steps cannot be easily separated. Thus, the results reported here for these two steps together can be regarded as an upper limit of their MS/MS spectra preprocessing performance alone.

The data files used in this analysis were derived from samples analyzed in the Proteomics Center at Children's Hospital Boston. 'Yeast', 'Mouse' and 'Human 1' are unfractionated whole cell lysates derived from the respective organism. 'Human 2' corresponds to one fraction of a human body fluid proteome separated by SDS-PAGE into 17 fractions. The datasets contained 3521 ('Yeast'), 8884 ('Mouse'), 8457 ('Human 1') and 7868 ('Human 2') spectra respectively. 'Yeast' and 'Human 1' were analyzed on an LTQ-Orbitrap (Thermo Scientific) equipped with a nanoflow HPLC system (Eksigent). For the analysis of 'Mouse' and 'Human 2' an LTQ equipped with a microscale capillary HPLC system was used (both Thermo Scientific). Data were acquired in data dependent acquisition mode with the 6 most intensive signals being selected for fragmentation after each survey scan (more details are given in the supplementary material).

The comparison for each search engine was carried out on the peptide level. Each preprocessing parameterization was searched against the respective combined forward-reverse database and the local false discovery rate (local FDR) was determined using the provided PSPEP-Software [24] (for ProteinPilot) or an in-house R implementation of PSPEP (for Mascot and Sequest), which is also freely available from http://hci.iwr.uni-heidelberg.de/mip/proteomics or http://www.childrenshospital.org/research/steenlab. For all analyses, a local FDR cut-off of 1% for peptides was applied. The results for a global FDR cut-off of 1‰, which result in a similarly restrictive number of identified peptides as a local FDR of 1%, are discussed in the supplementary materials.

For the analysis, we define x_{ij} to be the number of local FDR controlled peptide hits of parameterization i on dataset j and n to be the overall number of datasets (in our case n=4). We used two measures to evaluate the results. The sum (S) of the number of the identified peptides is given by

$$S_{\mathrm{i}} = \sum_{j} x_{ij}$$

Large values indicate overall good performance. However, this measure is slightly biased towards methods which perform well on larger datasets with more identified peptides.

Thus, we also used the mean proportion (MP) of the best result which is defined by taking the ratio of the number of peptides identified by an approach on a given dataset and the maximum number of hits identified by any approach on that dataset and averaging the ratio over all data sets:

$$\mathrm{MP_i} = \frac{1}{n}\sum_j \frac{x_{ij}}{\max_k x_{kj}}$$

A value close to 1 suggests competitive performance of a method across all datasets.

Results for all search engines and datasets show that preprocessing has a sizeable effect on peptide identification. The consequences of the preprocessing at the spectral level are exemplified in **Figure 2**: preprocessing can remove noise peaks without changing the relevant b- and y-fragment ions and thus improve peptide identification, but sometimes also sequence-revealing fragment ions are removed. Among the search engines, we notice improvements of 15-33% in the MP for the best tested preprocessing method in comparison to the original spectra with no preprocessing. Equally important is the fact, that inappropriate MS/MS preprocessing can have detrimental effects on the database searches resulting in fewer identified spectra. For instance, we observe that Mascot Distiller with standard settings reduces the MP by 12% relative to no preprocessing. This clearly indicates that preprocessing must be chosen with great care and adapted to the respective search algorithm.
The effect of the individual preprocessing methods varied from data set to data set. However, the overall most favorable methods were always among the best performing methods for all four data sets.

Mascot

Detailed results for the number of identified peptides using Mascot and all preprocessing methods and parameterizations described earlier are given in **Figure 3a** and in the supplementary material (Table 1). 'Top200' as the best performing approach shows a 15% increase in the MP, whereas we observe a 12% decrease in the MP for Mascot Distiller which does MS1 and MS/MS preprocessing; both values are in comparison to the MP without preprocessing.

Sequest

For Sequest, results for all preprocessing methods and parameterizations are displayed in **Figure 3b** and in the supplementary material (Table 2). Not all preprocessing methods improved the number of identified peptides, this was particularly true for 'Top X' and 'Top X in Y regions' approaches for which the numbers of identified spectra decreased by up to 5 %. Best results were obtained for 'Top 6 in window **±** 30' with an increase of 16% in the mean proportion measure compared to spectra without preprocessing.

ProteinPilot

Figure 3c and the supplementary material (Table 3) display the results for Protein Pilot with regard to all preprocessing methods and parameterizations. With the single exception of the 'Top 1 in window **±** 27' approach, all preprocessing methods result in a strong improvement in the number of identifications in comparison to the spectra without any preprocessing. In general, 'Top X in Y region' approaches and some of the 'Top X in window **±** Z' approaches perform better than 'Top X' approaches, with the overall best results for 'Top 6 in window **±** 30' with an increase 33% in the mean proportion measure compared to the original spectra without preprocessing.

Preprocessing the MS/MS data had two other advantages in addition to identifying more peptides: Firstly, the various preprocessing methods reduce the file size by 60-95% compared to the original MS/MS spectra (see supplementary material) without compromising the information content. Secondly, on our computer systems, we also observe significant decreases in the run times. For instance, for the 'Human 2' dataset and 'Top 200' we observe a 16% decrease in run time with ProteinPilot compared to the spectra without preprocessing and a 61% decrease in run time with Mascot. Detailed run time comparisons strongly depend on the computer systems involved and are beyond the scope of this manuscript. The run time of our software tool itself is negligible with 0.25-1s per megabyte of the original data on a standard PC.

Our interpretation of the results for Mascot is that the preprocessing impacts the internal preselection within Mascot. Since Mascot iterates its scoring over increasingly larger sets of the most intense ions [3], the 'Top 200' sets an upper limit to the number of ions included in the iteration and thus reduces the risk of overfitting sequences to noise peaks in spectra which might lead to false positives. This also explains the reduction in search time since fewer iterations are possible. For Sequest, which internally selects the maximum intensity ions and then splits the spectrum into parts which are normalized individually [25], the window-based approach reduces the risk that high intensity peaks clustered together dominate the internal normalization. Same as for ProteinPilot, which does not preprocess ions internally if .mgf files are used as input and therefore benefit from most sensible preprocessing approaches, 'Top 30 in Window ± 30' shows best results. Building up on the motivation of the preprocessing of OMSSA [16], this can be interpreted as accepting at least two noise ions for each b- and y-ion in a MS/MS spectrum, which is restrictive enough to avoid overfitting without losing significant information, even when some noise ions show high intensity.

Since the increase in the number of identified peptides is achieved on the spectral level, there is no bias towards specific proteins. Therefore, the number of identified proteins as well as their sequence coverage generally increase and additional identifications otherwise based on a single peptide hit are found. For instance, for the 'Yeast' dataset (3521 spectra) and a local FDR of 0.01 on the peptide level based on the Mascot results, we observe for 'Top 200' in comparison to spectra without preprocessing an increase from 141 to 165 identified proteins. Instead of 69, now 78 proteins are based on multiple identified peptides and the average number of spectra per protein increases slightly from 3.92 to 4.28.

Equally important, the increase in the number of peptides identified comes at virtually no price. Since peptides from the forward and reverse database are equally affected by the removal of ions from the MS/MS spectrum, a local FDR-control allows preserving the specificity while benefitting from the increased sensitivity and reductions in run time. Furthermore, our preprocessing does not change the structure of the data and our software generates .mgf or .dta-files which can be directly read by Mascot, Sequest and ProteinPilot. Thus, existing workflows for peptide and protein identification only have to undergo minimal changes to incorporate the proposed preprocessing.

The best performing approaches also showed good robustness across different datasets and charge states. They were among the best performing approaches in all datasets studied (see tables 1-3 in the supplementary materials) and we could not see any indication that the best performing

approaches favored specific charge states, since the ratio of charge 2 and charge 3 peptides remained constant throughout the preprocessing steps (see supplementary material).

Here, we focused solely on optimizing MS/MS preprocessing. However, there is high potential in combining all three preprocessing steps, i.e. quality scoring, precursor preprocessing and MS/MS preprocessing. Since these steps focus on complementary information resulting from different measurements, there is good reason to assume that they can be individually optimized such that their combined usage will strongly improve results. When combining the parent ion mass-to-charge and charge information as implemented in MaxQuant, Mascot Distiller or given by a state-of-the-art peak picking approach such as NITPICK [26] with optimized MS/MS preprocessing, we see significant improvement in the mean proportion measure in comparison to only using the precursor preprocessing on its own. For instance, the precursor preprocessing of MaxQuant was coupled to 'Top 200', the MS/MS preprocessing method empirically found as the best in this study for Mascot, and resulted in a 34% increase in the mean proportion measure in comparison to no preprocessing, which corresponds to a 18% increase over using 'Top 200' alone.

It should be noted that the results do not indicate a generally optimal preprocessing. Due to the different nature of the three search engines tested, a preprocessing method may increase the number of identified spectra for one search engine, but decrease the number for the case of another search engine. Hence, for each search engine, a separate comparison study is necessary. In this article, we focused on three commonly used peptide identification procedures (Mascot, Sequest, ProteinPilot) and identified the empirically best preprocessing for each of the three investigated search engines: 'Top 200' for Mascot and 'Top 6 in Window ± 30' for Sequest and ProteinPilot. These methods resulted in 15-33% increases in spectral identifications (at constant local FDR of 1%) with concomitant reduction in file sizes of up to 75% relative to the unprocessed files which in turn significantly reduced search time, i.e., the increase in spectral identification comes at negligible cost.

The authors would like to thank Michael Hanselmann and Xinghua Lou (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany) for comments, suggestions, and fruitful discussions. We gratefully acknowledge financial support by the DFG under grant no. HA4364/2-1 (B.Y.R., M.K., F.A.H.) and Robert Bosch GmbH (F.A.H.).

The authors have declared no conflict of interest.

References

[1] McHugh L, Arthur J W. Computational methods for protein identification from mass spectrometry data *PLoS Computional Biology.* 2008;4:e12.

[2] Nesvizhskii A I, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry *Nature Methods.* 2007;4:787–797.

[3] Perkins D N, Pappin D J, Creasy D M, Cottrell J S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999;20:3551-67.

[4] Eng J K, McCormack A L, Yates J R. An Approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database *Journal of the American Society for Mass Spectrometry.* 1994;5:976–989.
[5] Shilov I V, Seymour S L, Patel A A, et al. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra *Mol. Cell Proteomics.* 2007;6:1638–1655.
[6] Salmi J, Nyman T A, Nevalainen O S, Aittokallio T. Filtering strategies for improving protein identification in high-throughput MS/MS studies *Proteomics.* 2009;early.
[7] Koenig T, Menze B H, Kirchner M, et al. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics *J. Proteome Res..* 2008;7:3708–3717.
[8] Na S., Paek E.. Quality assessment of tandem mass spectra based on cumulative intensity normalization *J. Proteome Res..* 2006;5:3241–3248.
[9] Frank A. M., Bandeira N., Shen Z., et al. Clustering millions of tandem mass spectra *J. Proteome Res..* 2008;7:113–122.
[10] Tabb D. L., Thompson M. R., Khalsa-Moyers G., VerBerkmoes N. C., McDonald W. H.. MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra *J. Am. Soc. Mass Spectrom..* 2005;16:1250–1261.
[11] Gentzel M, Kocher T, Ponnusamy S, Wilm M. Preprocessing of tandem mass spectrometric data to support automatic protein identification *Proteomics.* 2003;3:1597–1610.
[12] Mayampurath A M, Jaitly N, Purvine S O, et al. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra *Bioinformatics.* 2008;24:1021-1023.
[13] Sadygov R G, Hao Z, Huhmer A F R. Charger: Combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra *Analytical Chemistry.* 2008;80:376–386.
[14] Schulze W X, Mann M. A novel proteomic screen for peptide-protein interactions *J. Biol. Chem..* 2004;279:10756–10764.
[15] Mujezinovic N, Raidl G, Hutchins J R A, Peters J-M, Mechtler K, Eisenhaber F. Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics.* 2006;6:5117-31.
[16] Geer L Y, Markey S P, Kowalak J A, et al. Open mass spectrometry search algorithm *Journal of Proteome Research.* 2004;3:958–964.
[17] Tanner S, Shu H J, Frank A, et al. InsPecT: Identification of posttransiationally modified peptides from tandem mass spectra *Analytical Chemistry.* 2005;77:4626-4639.
[18] Pedrioli P G, Eng J K, Hubley R, et al. A common open representation of mass spectrometry data and its application to proteomics research *Nat. Biotechnol..* 2004;22:1459–1466.
[19] Boehm A M, Galvin R P, Sickmann A. Extractor for ESI quadrupole TOF tandem MS data enabled for high throughput batch processing *BMC Bioinformatics.* 2004;5:162.
[20] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification *Nat. Biotechnol..* 2008;26:1367–1372.

[21] Hansen K C, Schmitt-Ulms G, Chalkley R J, Hirsch J, Baldwin M A, Burlingame A L. Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography *Mol. Cell Proteomics.* 2003;2:299–314.
[22] Brosch M, Swamy S, Hubbard T, Choudhary J. Comparison of mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold *Mol. Cell Proteomics.* 2008;7:962–970.
[23] Balgley B M, Laudeman T, Yang L, Song T, Lee C S. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy *Mol. Cell Proteomics.* 2007;6:1599–1608.
[24] Tang W H, Shilov I V, Seymour S L. Nonlinear fitting method for determining local false discovery rates from decoy database searches *Journal of Proteome Research.* 2008;7:3661–3667.
[25] Yates J R, Eng J K. Identification of nucleotides, amino acids, or carbohydrates by mass spectrometry 2000. United States Patent US 6017693.
[26] Renard B Y, Kirchner M, Steen H, Steen J A J, Hamprecht F A. NITPICK: Peak Identification for Mass Spectrometry Data *BMC Bioinformatics.* 2008;9:355.

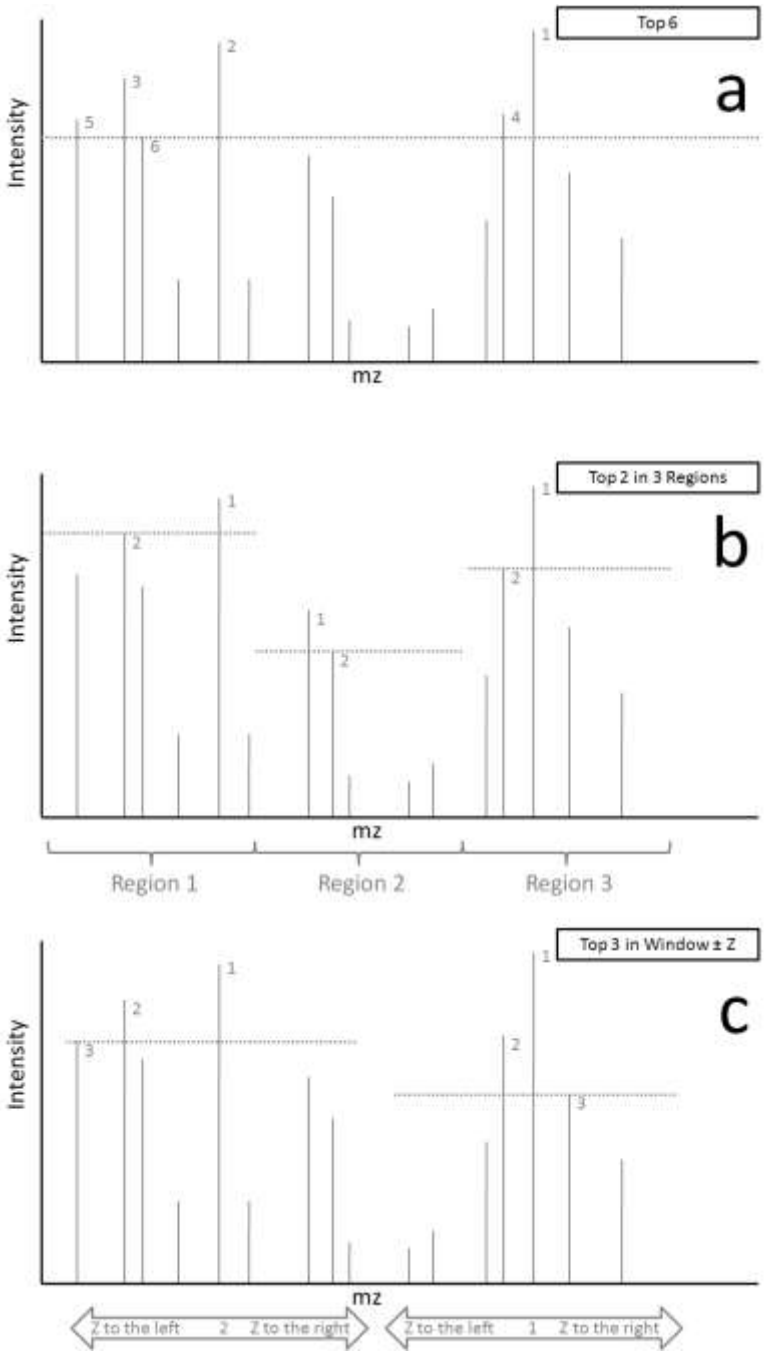


Figure 1: Visualization of the three preprocessing categories: (a) displays a 'Top 6' approach, peaks are ordered by their intensities and only the six most intensive peaks are chosen. (b) shows an example of a 'Top 2 in 3 regions' approach: the spectrum is first partitioned into three equal sized regions. In each region, the two most intensive peaks are selected. (c) shows a 'Top 3 in a window **±** Z' approach. The most intensive peak is identified and among all peaks within a window of Z Da to the left and right the two most intensive peaks are chosen. This procedure is then repeated for the second most intensive peak and a corresponding window is defined. Within the window, again the three most intensive peaks are selected. The procedure stops when no more peaks are available. Even though all three approaches result in six peaks, they only agree on the four most intensive peaks and differ in the remaining peaks. 'Top 6' selects more peaks clustered closely together than the other two approaches.

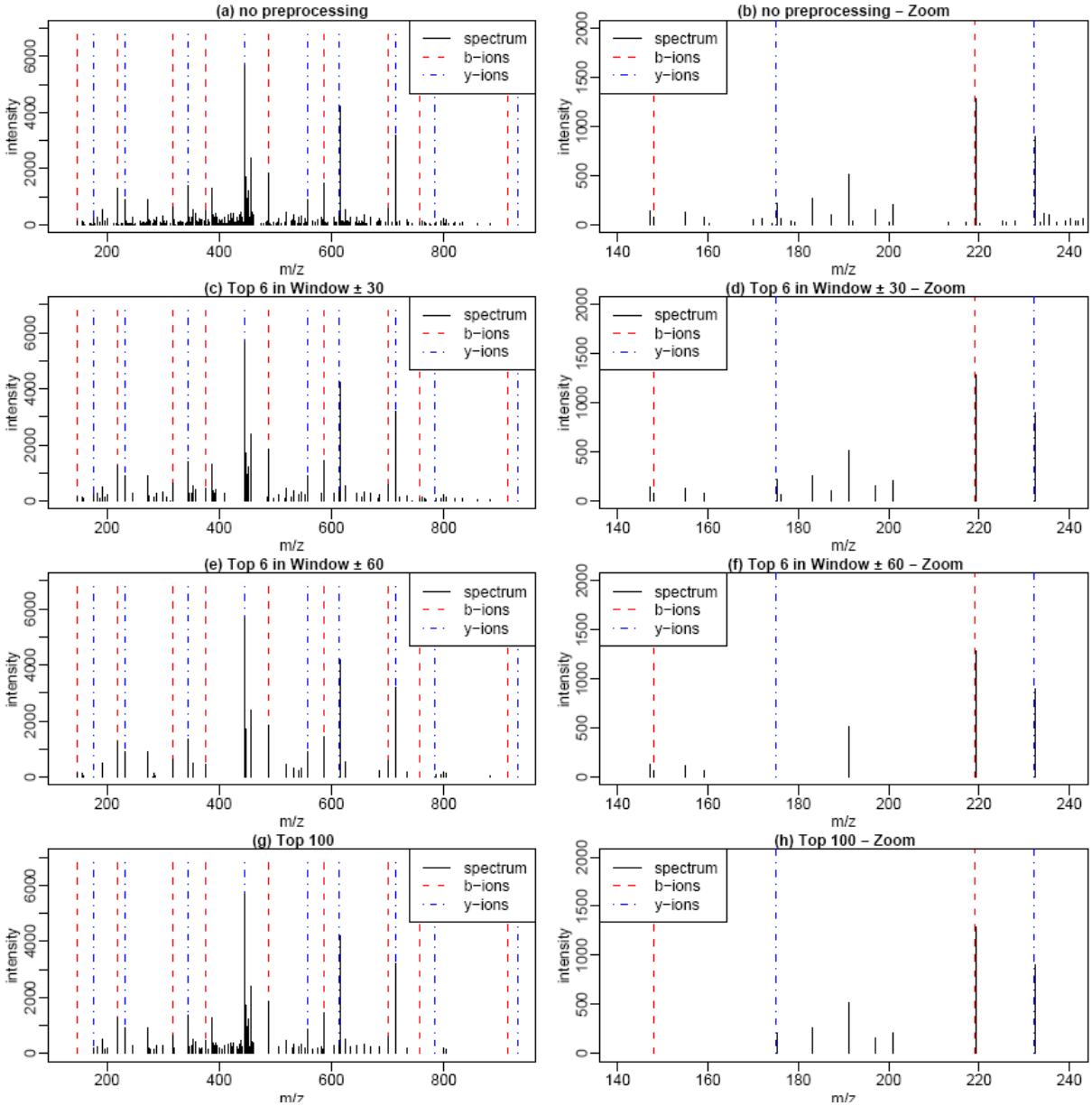
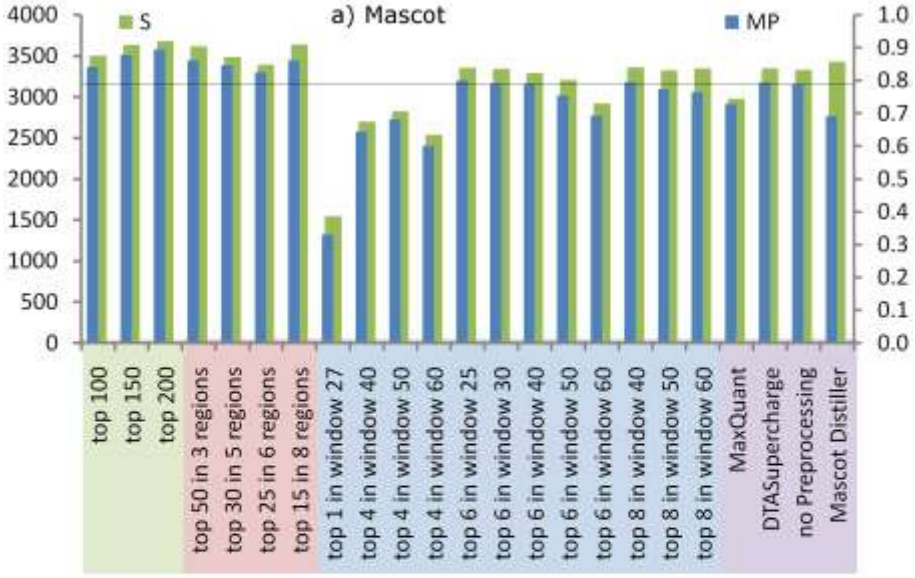
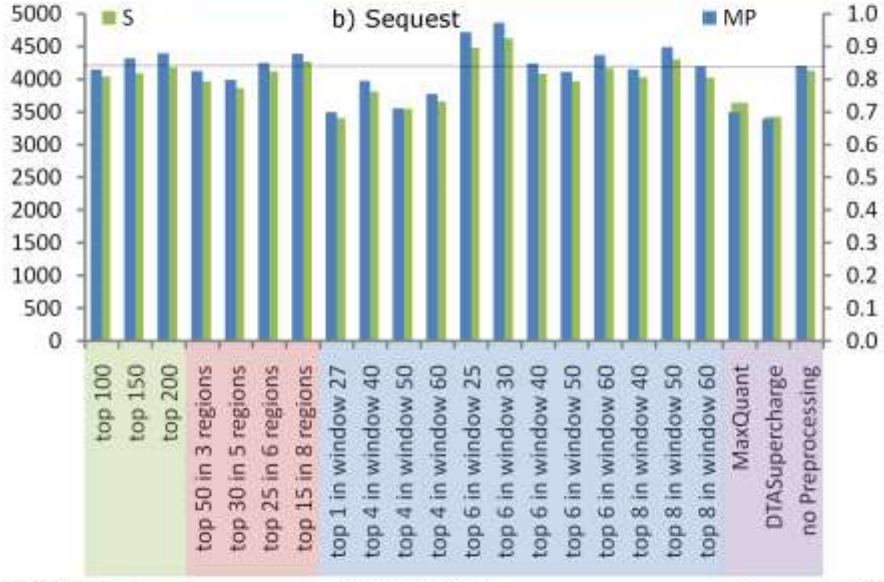


Figure 2: Visualization of a spectrum and its respective theoretical b- and y-ions from the 'Human 1'dataset (a, b) before preprocessing, (c, d) after preprocessing with a 'Top 6 in window ± 30' approach, (e, f) after preprocessing with a 'Top 6 in window ± 60' approach as well as (g, h) after preprocessing with a 'Top 100' approach. Whereas the 'Top 6 in window ± 30' preprocessing reduces the number of fragment ions by a factor of 3 without removing a single b- or y-ion, the 'Top 6 in window ± 60' as well as the 'Top 100' approach are too aggressive and also remove some b- and y-ions. The zooms in the 140-240 Da region on the right hand side show the abundance of noise in the spectrum without preprocessing (b) and the removal of signal carrying ions in the aggressive 'Top 6 in window ± 60' (f) and 'Top 100' (h) preprocessing, whereas the 'Top 6 in window ± 30' preprocessing (d) shows a reasonable balance of removing noise peaks while preserving the signal. In consequence, the confidence score for ProteinPilot ranges from 0.84 (no preprocessing) to 0.92 ('Top 6 in window ± 60') to 0.98 ('Top 100') to 0.99 ('Top 6 in window ± 30').


a) Mascot
S
MP
4000
3500
3000
2500
2000
1500
1000
500
0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
top 100
top 150
top 200
top 50 in 3 regions
top 30 in 5 regions
top 25 in 6 regions
top 15 in 8 regions
top 1 in window 27
top 4 in window 40
top 4 in window 50
top 4 in window 60
top 6 in window 25
top 6 in window 30
top 6 in window 40
top 6 in window 50
top 6 in window 60
top 8 in window 40
top 8 in window 50
top 8 in window 60
MaxQuant
DTASupercharge
no Preprocessing
Mascot Distiller


b) Sequest
S
MP
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
top 100
top 150
top 200
top 50 in 3 regions
top 30 in 5 regions
top 25 in 6 regions
top 15 in 8 regions
top 1 in window 27
top 4 in window 40
top 4 in window 50
top 4 in window 60
top 6 in window 25
top 6 in window 30
top 6 in window 40
top 6 in window 50
top 6 in window 60
top 8 in window 40
top 8 in window 50
top 8 in window 60
MaxQuant
DTASupercharge
no Preprocessing

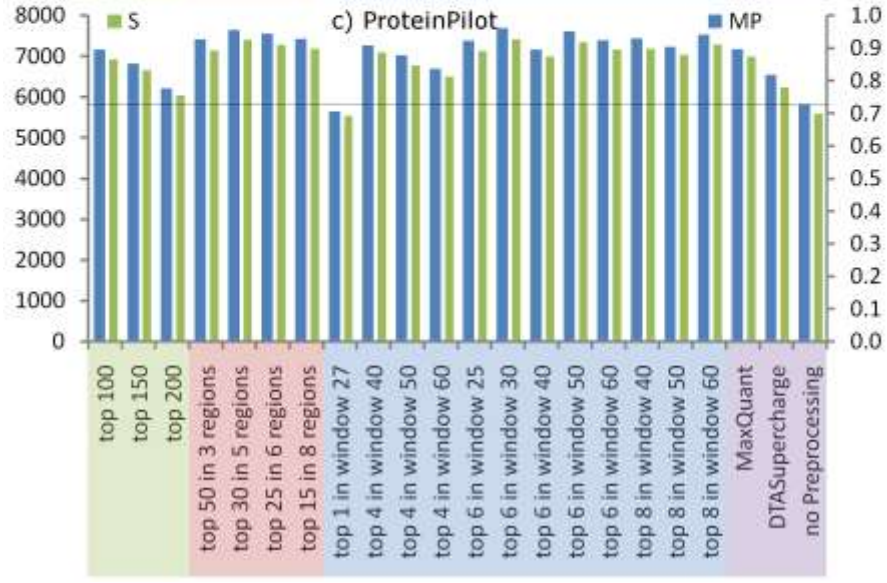

c) ProteinPilot
S
MP
8000
7000
6000
5000
4000
3000
2000
1000
0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
top 100
top 150
top 200
top 50 in 3 regions
top 30 in 5 regions
top 25 in 6 regions
top 15 in 8 regions
top 1 in window 27
top 4 in window 40
top 4 in window 50
top 4 in window 60
top 6 in window 25
top 6 in window 30
top 6 in window 40
top 6 in window 50
top 6 in window 60
top 8 in window 40
top 8 in window 50
top 8 in window 60
MaxQuant
DTASupercharge
no Preprocessing

Figure 3: Comparison of preprocessing results on all datasets for a) Mascot, b) Sequest and c) ProteinPilot: For Mascot, a strong variation can be observed in both measures with regard to the different preprocessing methods. 'Top 200' shows the highest number of overall correctly identified peptides (S) as well as the highest mean proportion (MP) value with a 15% increase over no preprocessing. For Sequest, 'Top 6 in a window of **±** 30' clearly outperforms any other preprocessing method in both measures. We observe an increase of 16% in comparison to no preprocessing. The worst results are obtained for DTASupercharge with a decrease of 30% in comparison to no preprocessing. For ProteinPilot, all preprocessing methods show improved peptide identification with the single exception of 'Top 1 in a window of **±** 27' where we observe a 3% decrease in the number of identified peptides. Best results are obtained for 'Top 6 in a window of **±** 30' with a 33% increase in the number of identified peptides.

Supplementary Material

Experimental setup

The two LTQ-Orbitrap samples were acquired with the following instrument settings: Full scan MS1 in the Orbitrap mass analyzer, 1 microscan, 50 msec maximum fill time at a target value of 1e6; MS2 scan in the LTQ mass analyzer, 1 microscan, 150 msec maximum fill time at a target value of 1e4.

The two LTQ samples used the following instrument settings: Full scan MS1 – 1 microscan, 75 msec maximum fill time at a target value of 4e4; MS2 – 1 microscan, 200 msec maximum fill time at a target value of 1e4.

In the case of missing precursor charge state information resulting from the low-resolution MS1 scan, charge states were set to either 1, 2, 3, 4 (Mascot, ProteinPilot) or 2,3 and searched against the database, only the best fitting sequence resulting from these multiple searches was regarded for the further local-FDR computation and analysis.

Since the goal of the study was not to compare search engines, but to identify the empirically optimal preprocessing, parameterization of the search engines was not standardized, but we rather referred to the standard settings we usually apply as described below. Since different databases were used for the varying search engines, results are only comparable for a specific search engine, but not across different search engines.

Mascot

Mascot [3] was run in version 2.2.04 with Carbamidomethyl (C) as fixed modification and Deamidated (NQ), Gln->pyro-Glu (N-term Q), Glu->pyro-Glu (N-term E) as well as Oxidation (M) as variable modifications. We only considered fully tryptic peptides with a maximum of one missed cleavage. Fragment ion tolerance was set at 0.8 Da and protein tolerance at 1.5 Da (LTQ) or 10 ppm (Orbitrap) respectively. For the human samples and the mouse sample the respective RefSeq database [2] (July 11th, 2008) were chosen, whereas the SGD project database [3] was used for the yeast dataset (May 7th, 2006).

Sequest

The Sequest-Sorcerer algorithm (SageN Research, CA) (Version: 3.5 RC2; 4.0.4.) [4] was used as a part of the Sorcerer IDA II platform. The searches were run with parent tolerance set to 1.5 Da for the LTQ samples and to 10 ppm for the Orbitrap samples; fragment ion tolerance was set to 1 Da. Carbamidomethyl (C) was searched as fixed modification for searches of both the Orbitrap and the LTQ data. Deamidation(NQ), Oxidation(M) and Gln->pyro-Glu (N-term Q) were set as variable modifications for the Orbitrap data. One missed cleavage was allowed at maximum. The IPI databases version 3.09 [5] were used for the mouse and human samples whereas the SGD project database [3] (May 7th, 2006) was used for the yeast sample. PeptideProphet [6] was used to rank the peptides according to their computed probability.

ProteinPilot

Protein Pilot [5] version 2.0.1 (software revision number 65587) was run with its standard LTQ or Orbitrap settings with Iodoacetamide as a Cys. Alkylation, tryptic digestion and search effort set to 'thorough'. The IPI databases version 3.09 [5] were used for mouse and human samples whereas the UniProt Combined Panther (version 5) [8] with specification *Saccharomyces cerevisiae* was used for the yeast samples.

Preprocessing Software

The preprocessing software was developed in C++ and is available as a standalone application for windows and linux. It accepts a mgf file or a folder of dta-files as well as the choice of the parameterization ('Top X'/'Top X in Y regions'/'Top X in window of **±** Z' as well as X, Y and Z respectively) as an input and outputs a new mgf file or folder of dtas with preprocessed peak lists according to the chosen parameterization. The software is freely available from http://hci.iwr.uni-heidelberg.de/mip/proteomics or http://www.childrenshospital.org/research/steenlab.

Measures

Several measures were used to obtain an unbiased view on the performance of the various parameterizations on the four datasets, and to reduce the influence of other factors including the overall number of peptides in a given dataset. Therefore, we define x_{ij} to be the number of local-FDR controlled peptide hits of parameterization i on dataset j and n to be the overall number of datasets (in our case n=4).

Mean Proportion

The mean proportion (MP) of the best result is defined by taking the ratio of the number of peptides identified by an approach on a given dataset and the maximum number of hits identified by any approach on that dataset and averaging the ratio over all data sets:

$$\mathrm{MP_i} = \frac{1}{n}\sum_j \frac{x_{ij}}{\max_k x_{kj}}$$

A value close to 1 therefore suggests competitive performance of a method across all datasets.

Mean Relative Squared Error

Similarly, the mean relative squared error (MRSE) is defined by computing the squared distance between the number of identified peptides of one approach and the maximum number of identifications for a given dataset. The squared distance is weighted by the squared maximum number of identifications to reduce the dependence on the overall number of identified peptides.

$$\mathrm{MRSE_i} = \frac{1}{n}\sum_j \frac{(x_{ij} - \max_k x_{kj})^2}{\max_k x_{kj}^2}$$

Large values indicate strong departure from the best method in at least one dataset.

Sum of all datasets

The simplest (and most biased with regard to the number of peptides actually included in a dataset) measure is the sum (S) of the number of the identified peptides.

$$S_{\mathrm{i}} = \sum_{j} x_{ij}$$

Here, large values indicate overall good performance, but might be slightly biased towards methods which perform well on the datasets with overall more identifications.

Local FDR and Global FDR

Tang et al [24] point out that for proteomics, the local FDR reports how likely a specific protein or peptide is incorrect, rather than the overall error rate for the set of proteins or peptides it is a member of and that it is thus more meaningful than the global FDR. We performed additional experiments using a global FDR approach with similar restrictiveness to investigate whether results are dependent on the choice of postprocessing. Therefore, all results obtained from Sequest for all datasets and for all preprocessing methods were reanalyzed using a global FDR (1 ‰). Results for the sum (S) of identified peptides across all datasets are given in Figure 4. Overall, results for both approaches show a high similarity. The local FDR shows a higher variability and sharper drop offs to low numbers for poorly performing methods compared to the global FDR. In contrast, for the overall best performing methods ('Top 6 in window ±30', 'Top 6 in window ±25', 'Top 4 in window ±25') , local FDR and global FDR show identical ordering and highly similar numbers. These results indicate that the choice of local and global FDR does not influence the selection of the best performing methods if a similar level of restrictiveness is chosen.

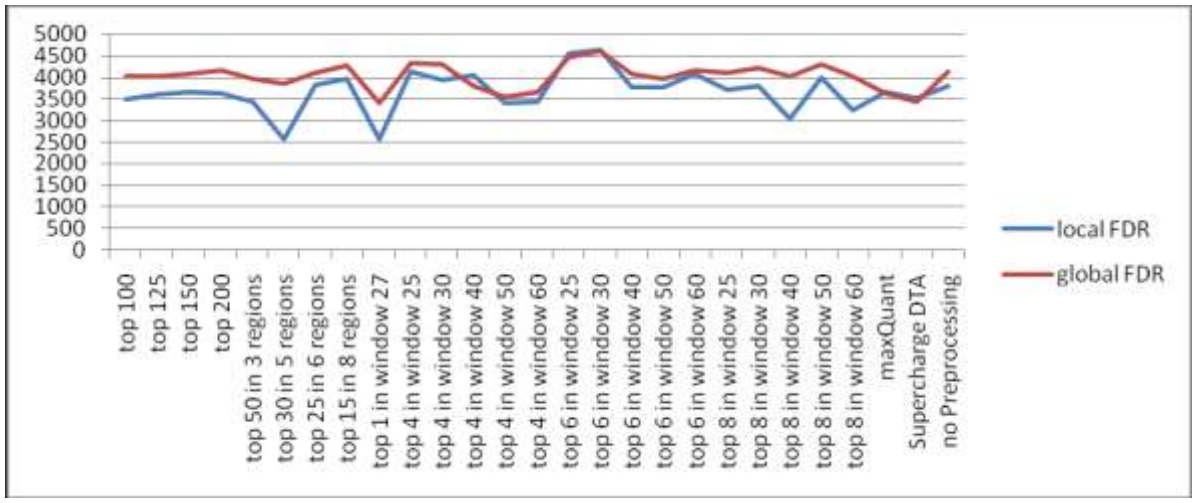


Figure 4: Sum (S) of identified peptides for Sequest using a local and a global FDR of similar restrictiveness. While for poorly performing preprocessing schemes, the local FDR shows a more extreme drop off, the two FDRs result in very similar numbers for well performing methods. This indicates that the selection of the optimal preprocessing does not depend on the choice of the local FDR as a postprocessing step.

File Sizes

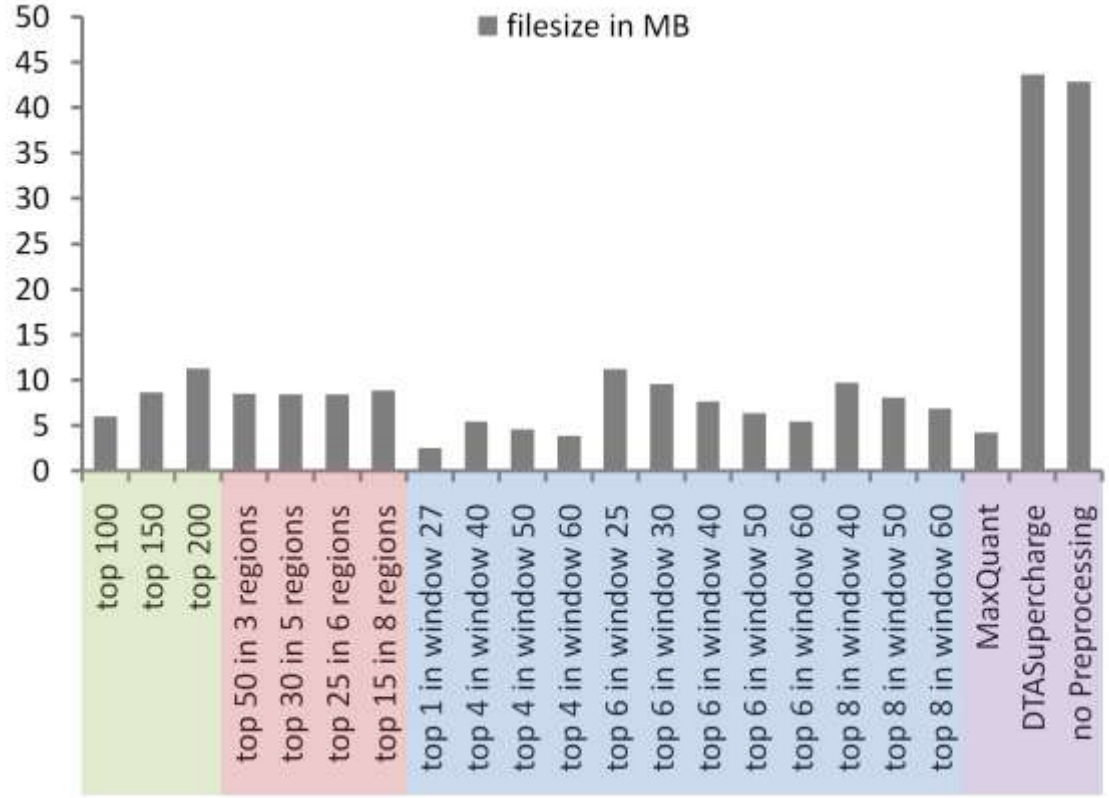


Figure 5: Comparison of the file sizes of the preprocessing approaches for the 'Yeast' dataset: With the single exception of DTASupercharge, all preprocessing methods drastically reduce the file size in comparison to the original spectra without preprocessing. The smallest file sizes are achieved for 'top 1 in window of ± 27' with a reduction of 94%; for the other approaches the reduction ranges between 75% and 90%.

Influence of the Charge State

If the charge state were influential for the preprocessing, we would assume to see a different proportion of charge 2 to charge 3 spectra within a preprocessed dataset compared to the case of no preprocessing. Therefore, we computed the ratio r_i as

$$r_i = \frac{\frac{\#\text{charge 2 peptides with preprocessing i}}{\#\text{charge 3 peptides with preprocessing i}}}{\frac{\#\text{charge 2 peptides with no preprocessing}}{\#\text{charge 3 peptides with no preprocessing}}}$$

where the number of peptides of a certain charge state was computed as all peptides of that charge having a score above the corresponding local FDR cut off.

If the charge has no influence on the preprocessing i, r_i should not be significantly different from 1 since then the ratio coincides with the ratio obtained for no preprocessing and the charge distribution remains unchanged. Departures to smaller numbers indicate a preference of charge 3

ions through the preprocessing and larger numbers a preference of charge 2 ions. In these cases, we could conclude that the preprocessing is better suited to the corresponding more often observed charge state.

For all preprocessing methods and for all datasets, we computed the r_i values based on the Sequest identifications. Further, we computed the mean of the ratios and 95% confidence intervals across all for datasets for each preprocessing approach (using a log transformation to account for the asymmetric metric introduced by the ratio in the computation of the mean and standard deviation). Results are given in Figure 6. With the exception of MaxQuant on the 'Human 1' dataset, all values are close to 1. Still a large variation can be observed for the ratios of a single preprocessing between the different datasets often showing values below and above 1. With the single exception of 'Top 6 in Window of ±50', all methods have confidence intervals for the mean ratio which include 1. Therefore, it can be concluded from these findings that for these preprocessing methods no significant influence of the charge state on the preprocessing can be determined.

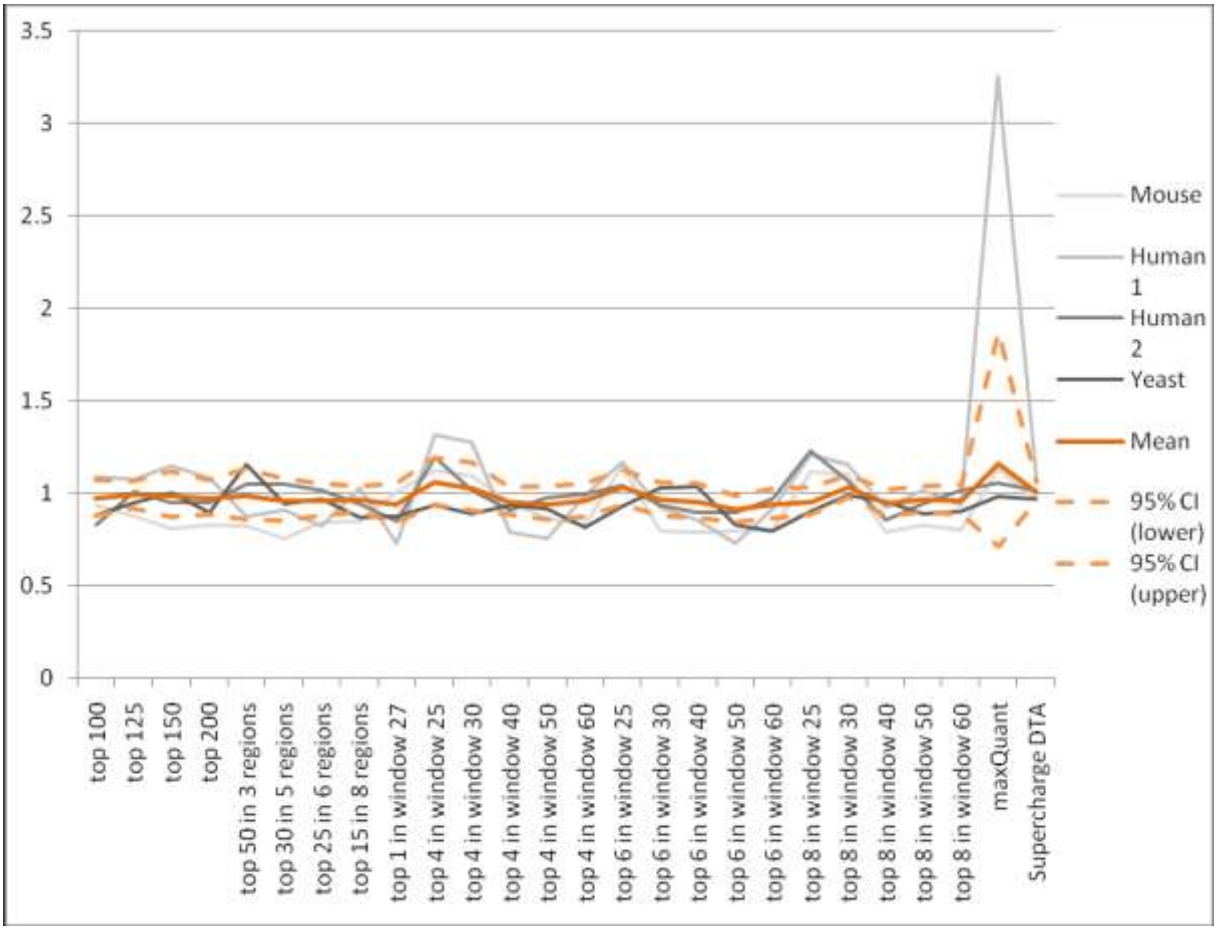


Figure 6: Charge state ratios r_i and their mean and confidence intervals for the preprocessing methods. With the exception of MaxQuant for the 'Human 1' dataset, all ratios for all preprocessing methods are in the vicinity of 1, but display a large variance across datasets even within a single method. Only 'Top 6 in Window of ±50' shows a statistical significant departure from 1 in its mean, whereas for all other methods, no dependence between the charge distribution and the preprocessing method can be concluded.

Mascot local FDR 1%		DataSet					Measure		
Preprocessing type		'Human 1'	'Mouse'	'Yeast'	'Human 2'		MP	MRSE	S
Instrument		LTQ-Orbitrap	LTQ	LTQ-Orbitrap	LTQ				
Number of spectra		8457	7868	3521	8884				
TopX									
Top 100		1179	**920**	665	742		0.842	0.04869	3506
Top 125		1159	896	**709**	824		0.871	0.04438	3588
Top 150		1191	891	699	855		0.879	0.04022	3636
Top 200		1173	910	706	891		**0.895**	0.04138	**3680**
Top 250		1082	906	681	667		0.810	0.06748	3336
Top X in Y regions									
Top 50 in 3 regions		1259	848	682	827		0.862	**0.03616**	3616
Top 30 in 5 regions		1135	803	703	848		0.848	0.04999	3489
Top 25 in 6 regions		1103	779	698	817		0.827	0.05652	3397
Top 15 in 8 regions		1314	788	676	865		0.861	0.03402	3643
Top X in window **±** Z									
Top 1 in window **±** 27		871	55	452	167		0.331	0.49716	1545
Top 4 in window **±** 40		951	667	553	528		0.645	0.13984	2699
Top 4 in window **±** 50		974	616	631	601		0.682	0.12115	2822
Top 4 in window **±** 60		934	578	501	528		0.600	0.16714	2541
Top 6 in window **±** 25		1150	781	589	837		0.800	0.05753	3357
Top 6 in window **±** 30		1179	791	561	815		0.791	0.05836	3346
Top 6 in window **±** 40		1100	780	581	832		0.789	0.06423	3293
Top 6 in window **±** 50		1175	721	570	601		0.754	0.06934	3209
Top 6 in window **±** 60		1027	665	521	711		0.694	0.10471	2924
Top 8 in window **±** 40		1182	772	576	829		0.795	0.05690	3359
Top 8 in window **±** 50		1244	764	549	770		0.775	0.05890	3327
Top 8 in window **±** 60		**1342**	722	528	756		0.764	0.05940	3348
MaxQuant		965	689	684	638		0.729	0.10174	2976
DTASupercharge		1153	819	555	824		0.795	0.05965	3351
Mascot Distiller		1164	719	569	**1977**		0.692	0.18858	3428
no Preprocessing		1164	783	553	834		0.789	0.06095	3334

Table 1: Comparison of number of peptide hits for a local FDR of 0.01 using the Mascot search engine. Bold values indicate the column-wise best results. Overall, preprocessing can increase the number of identified peptides by up to 14% with the 'top 150' and 'top 200' approaches as the best performing methods.

Sequest local FDR 1%		DataSet					Measure		
Preprocessing type		'Human 1'	'Mouse'	'Yeast'	'Human 2'		MP	MRSE	S
Instrument		LTQ-Orbitrap	LTQ	LTQ-Orbitrap	LTQ				
Number of spectra		8457	7868	3521	8884				
Top X									
Top 100		598	1571	910	960		0.830	0.04160	4039
Top 125		730	1486	884	920		0.841	0.03047	4020
Top 150		829	1426	815	1023		0.864	0.02366	4093
Top 200		794	1439	872	1077		0.881	0.01571	4182
Top X in Y regions									
Top 50 in 3 regions		630	1439	947	953		0.826	0.03663	3969
Top 30 in 5 regions		541	1357	912	1053		0.799	0.05045	3863
Top 25 in 6 regions		619	1502	862	1135		0.849	0.03121	4118
Top 15 in 8 regions		628	**1573**	884	1187		0.879	0.02602	4272
Top X in window Z									
top 1 in window **±** 27		910	1228	840	428		0.694	0.10729	3406
top 4 in window **±** 25		1238	1470	863	770		0.902	0.01379	4341
top 4 in window **±** 30		1181	1557	850	715		0.885	0.01824	4303
top 4 in window **±** 40		884	1377	939	612		0.788	0.05123	3812
top 4 in window **±** 50		1085	1367	759	334		0.704	0.12452	3545
top 4 in window **±** 60		1182	1226	738	516		0.750	0.07924	3662
top 6 in window **±** 25		1242	1517	860	**865**		0.936	0.00986	4484
top 6 in window **±** 30		1208	1545	**1058**	814		**0.966**	**0.00167**	**4625**
top 6 in window **±** 40		1069	1475	857	683		0.842	0.02712	4084
top 6 in window **±** 50		1013	1437	918	603		0.816	0.03922	3971
top 6 in window **±** 60		1075	1442	898	753		0.868	0.01765	4168
top 8 in window ± 25		**1243**	1555	715	615		0.836	0.04762	4341
top 8 in window ± 30		1090	1557	849	672		0.864	0.02599	4128
top 8 in window ± 40		1082	1475	892	583		0.824	0.03905	4236
top 8 in window ± 50		1200	1456	953	694		0.891	0.01524	4303
top 8 in window ± 60		962	1512	821	732		0.832	0.03244	4027
MaxQuant		1109	1558	839	134		0.700	0.19258	3640
DTASupercharge		129	1269	971	1054		0.680	0.19790	3423
no Preprocessing		517	1576	865	1175		0.840	0.04951	4133

Table 2: Comparison of number of peptide hits for a local FDR of 0.01 using Sequest. Bold values indicate the column-wise best results. Overall, preprocessing increases the number of identified peptides by up to 16% with a 'Top 6 in window of **±** 30' approach showing the empirically best results in all three measures.

ProteinPilot local FDR 1%		DataSet					Measure		
Preprocessing type		'Human 1'	'Mouse'	'Yeast'	'Human 2'		MP	MRSE	S
Instrument		LTQ-Orbitrap	LTQ	LTQ-Orbitrap	LTQ				
Number of spectra		8457	7868	3521	8884				
TopX									
Top 100		2468	1380	1091	1983		0.896	0.01102	6922
Top 150		2476	1335	975	1868		0.853	0.02339	6654
Top 200		2229	1156	979	1679		0.776	0.05115	6043
Top X in Y regions									
Top 50 in 3 regions		2631	1466	1156	1886		0.927	0.00697	7139
Top 30 in 5 regions		**2717**	1398	1201	2088		0.956	0.00374	7404
Top 25 in 6 regions		2674	1393	**1223**	1982		0.944	0.00560	7272
Top 15 in 8 regions		2708	1343	1198	1940		0.929	0.00873	7189
Top X in window Z									
Top 1 in window **±** 27		2041	1163	737	1597		0.706	0.09028	5538
Top 4 in window **±** 20		2503	1308	1151	2044		0.907	0.01065	7006
Top 4 in window **±** 25		2537	1466	1112	2063		0.929	0.00516	7178
Top 4 in window **±** 30		2473	1437	1050	2034		0.903	0.01017	6994
Top 4 in window **±** 40		2597	1474	990	2033		0.908	0.01184	7094
Top 4 in window **±** 50		2345	1491	1000	1940		0.879	0.01693	6776
Top 4 in window **±** 60		2170	1417	886	2024		0.837	0.03302	6497
Top 6 in window **±** 20		2591	1136	1174	1691		0.852	0.03324	6592
Top 6 in window **±** 25		2684	1360	1190	1893		0.922	0.00941	7127
Top 6 in window **±** 30		2695	1452	1212	2053		**0.961**	**0.00250**	**7412**
Top 6 in window **±** 40		2641	1343	1059	1949		0.895	0.01308	6992
Top 6 in window **±** 50		2610	1496	1138	2105		0.951	0.00258	7349
Top 6 in window **±** 60		2516	**1574**	1006	2068		0.923	0.01000	7164
Top 8 in window **±** 20		2543	1170	1040	1642		0.820	0.03870	6395
Top 8 in window **±** 25		2429	1186	1013	1644		0.807	0.04087	6272
Top 8 in window **±** 30		2487	1309	1167	1772		0.878	0.01847	6735
Top 8 in window **±** 40		2624	1374	1185	2000		0.930	0.00643	7183
Top 8 in window **±** 50		2553	1364	1064	2053		0.904	0.01055	7034
Top 8 in window **±** 60		2519	1464	1113	**2189**		0.942	0.00457	7285
MaxQuant		2533	1483	975	2003		0.897	0.01407	6994
DTASupercharge		2196	1205	1115	1721		0.818	0.03631	6237
no Preprocessing		2243	956	1059	1336		0.727	0.08861	5594

Table 3: Comparison of number of peptide hits for a local FDR of 0.01 using ProteinPilot. Bold values indicate the column-wise best results. Overall, preprocessing increases the number of identified peptides by up to 33% with a 'Top 6 in window of **±** 30' approach.

References to the Supplementary Materials

[1] Perkins D N, Pappin D J, Creasy D M, Cottrell J S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999;20:3551-67.

[2] Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007;35:D61-5.

[3] SGD project . Saccharomyces Genome Database 2008.

[4] Eng J K, McCormack A L, Yates J R. An Approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database *Journal of the American Society for Mass Spectrometry.* 1994;5:976–989.

[5] Kersey P J, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: An integrated database for proteomics experiments. *Proteomics.* 2004;4:1985–1988.

[6] Keller A., Nesvizhskii A. I., Kolker E., Aebersold R.. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search *Anal. Chem..* 2002;74:5383–5392.

[7] Shilov I V, Seymour S L, Patel A A, et al. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra *Mol. Cell Proteomics.* 2007;6:1638–1655.

[8] Mi H, Lazareva-Ulitsky B, Loo R, et al. The PANTHER database of protein families, subfamilies, functions and pathways *Nucl. Acids Res..* 2005;33:D284-288.

[9] Tang W H, Shilov I V, Seymour S L. Nonlinear fitting method for determining local false discovery rates from decoy database searches *Journal of Proteome Research.* 2008;7:3661–3667.