Ranking Tandem Mass Spectra

and the Impact of Database Size and Scoring Function on Peptide Spectrum Matches

Canan Has, Cemal Ulaş Kundakcı,
Molecular Biology and Genetics
Izmir Institute of Technology
Gulbahce, Urla, Izmir, Turkey
cananhas@gmail.com

Aybuge Altay, and Jens Allmer
Molecular Biology and Genetics
Izmir Institute of Technology
Gulbahce, Urla, Izmir, Turkey
jens@allmer.de

Abstract—**Proteomics is currently driven by mass spectrometry. For the analysis of tandem mass spectra many computational algorithms have been proposed. There are two approaches, one which assigns a peptide sequence to a tandem mass spectrum directly and one which employs a sequence database for looking up possible solutions. The former method needs high quality spectra while the latter can tolerate lower quality spectra. Since both methods are computationally expensive, it is sensible to establish spectral quality using an independent fast algorithm. In this study, we first establish proper settings for database search algorithms for the analysis of spectra in our gold benchmark dataset and then analyze the performance of ScanRanker, an algorithm for quality assessment of tandem MS spectra, on this ground truth data. We found that OMSSA and MSGFDB have limitations in their scoring functions but were able to form a proper consensus prediction using majority vote for our benchmark data. Unfortunately, ScanRanker's results do not correlate well with the consensus and ScanRanker is also too slow to be used in the capacity it is supposed to be used.**

Keywords—mass spectrometry; false discovery rate; database size; spectrum quality; scoring algorithm; ScanRanker

I. INTRODUCTION

Proteomics investigates the proteins that make up an organism. For the identification of proteins, their sequencing, quantification and other tasks, mass spectrometry is currently the tool of choice [1]. In short, proteins are first digested into peptides since smaller compounds are easier transferred into gas phase and accelerated in a mass spectrometer (MS). The peptides are then channeled to the MS using liquid chromatography (LC), usually via a reverse phase column. The first measurement resolves the mass to charge ration (m/z) of the peptides that elute into the MS. Following fragmentation of the peptides by, for example, collision induced dissociation (CID) [2] a second stage of MS (tandem MS, MS/MS, MS^2) resolves the m/z of the peptide fragments. These can then, similar to Sanger sequencing for nucleotides, be analyzed computationally to reveal the complete peptide's sequence.

This process depends on computational algorithms and they come in two flavors. One, *de novo* sequencing, assigns a sequence to the MS/MS spectrum with no additional information [3], while the other (database search) uses a sequence database to select the best matching peptide from a list of expected proteins. Many algorithms have been proposed for both *de novo* sequencing [3] and database search [4], [5]. In database search, OMSSA [6], X!Tandem [7], and MSGFDB [8] are prominent free tools. These algorithms are routinely used in many laboratories to assign peptides to mass spectra and are part of computational pipelines like TOPP [9] and TPP [10]. Unfortunately, the identifications of different tools cannot easily be compared and therefore a population-based statistic, the false discovery rate (FDR), is widely employed to assign a confidence to peptide spectrum matches (PSMs). An early example of the use of 5% FDR is given in [11]. Currently, FDR is a controversial topic and is being investigated more closely for example by [12].

Database search algorithms depend on a database which contains the expected sequences for assigning correct PSMs. *De novo* sequencing tools on the other hand depend on high quality spectra. Both methods are computationally expensive and therefore tools to assess the quality of mass spectra to avoid unnecessary calculations have been proposed and the latest addition has been ScanRanker [13]. In addition to avoiding unnecessary calculations, knowing the quality of a spectrum can be useful to submit it to *de novo* sequencing instead of database search in case the sequence does not exist in the database for which there are many reasons [14] and if the spectral quality is judged high enough. Obviously, the quality assessment methodology must be significantly faster than downstream tools since otherwise it is merely imposing a computational overhead without tangible benefit.

Here we analyze a ground truth benchmark [15] LTQ dataset created by directly injecting synthetic peptides into the mass spectrometer and repeatedly measuring the MS/MS spectra which we previously created for a different purpose. We used this dataset to optimize the settings of OMSSA, MSGFDB and X!Tandem searches by varying fragmentation tolerance and database size. We combined the database search results into a simple consensus score which is useful for spectral quality assessment. In this process peculiarities with OMSSA and MSGFDB scoring functions were uncovered which we will also discuss in this paper. Nonetheless, we were able to establish a consensus scoring system based on majority vote creating a criterion to measure spectral quality. It has been shown by [16], [17], and [18] that the calculation of a consensus is superior to the results from individual algorithms. Our consensus method is based on the number of database

This work was in part supported by an award for outstanding young scientist by TÜBA, awarded to Jens Allmer, and by the TÜBITAK grant 111E139.

978-1-4799-0701-4/13/$31.00 ©2013 IEEE

September 25-27, 2013
Ankara, Turkey

search algorithms (OMSSA, MSGFDB, and X!Tandem) and their respective scores.

We hypothesized that ScanRanker is useful if its score correlates with this assessment since spectra that are consistently well identified by all three algorithms must be of high quality whereas spectra that are identified by none of the algorithms correctly must be of low quality. We analyzed ScanRanker using this dataset in order to find out whether its speed and accuracy warrant its use in computational pipelines for mass spectrometry-based proteomics. Unfortunately, ScanRanker does not correlate with our quality assessment and its speed is similar to OMSSA's so that we had to conclude that its use in computational pipelines is not warranted.

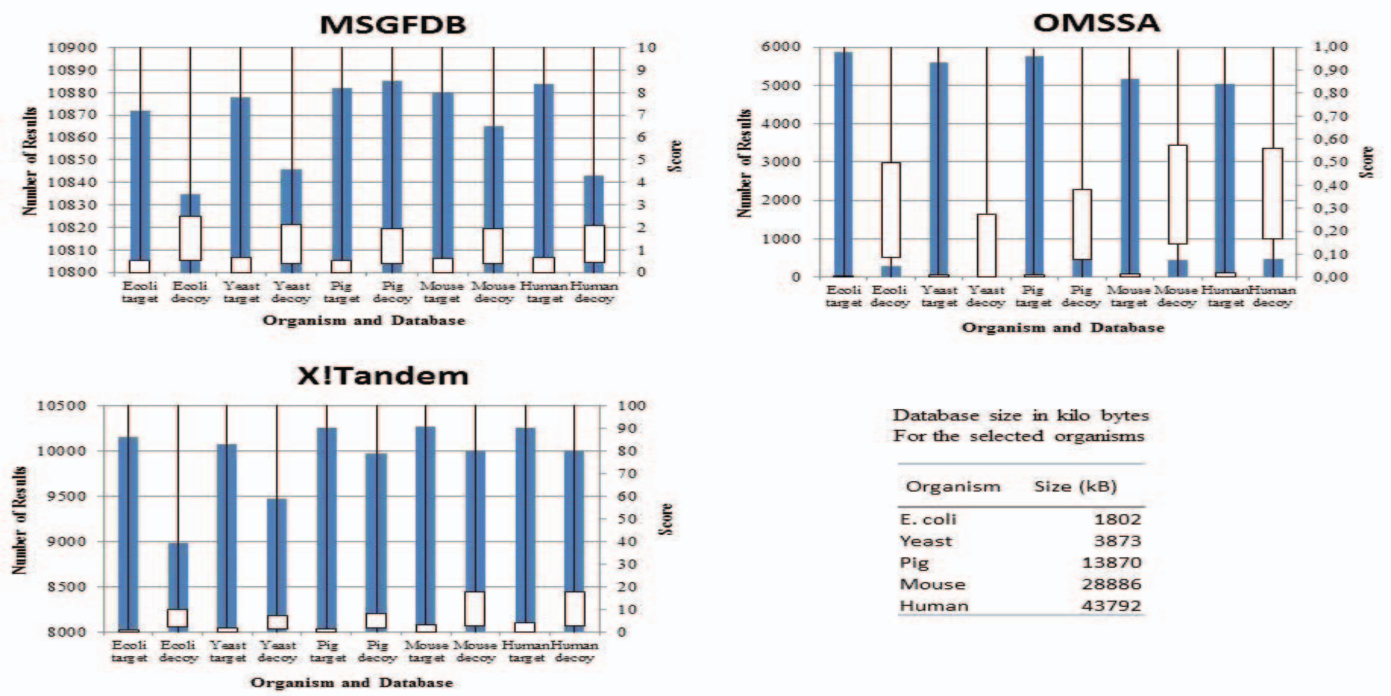
II. Materials and Methdos

A. Dataset

The dataset is composed of 11065 MS/MS spectra from 45 synthetic peptides derived from five different proteins which are cytochrome c (ACN P00004), bovine serum albumin (ACN P02769), oval albumin (ACN P01012), myoglobin (ACN P68082) and lysozyme C (ACN P61626) (GL Biochem Ltd, Shanghai, China). MS/MS analysis was carried out via Thermo Scientific LTQ XL Linear Ion Trap ESI with CID fragmentation at the Izmir Institute of Technology. CID energy and activation times were varied to collect many MS/MS spectra of different qualities and charges between +1 to +3.

B. Spectrum Identification

Raw data was converted to mzXML format by TransProteomicPipeline [19]. Afterwards, all files were converted to mgf format by using our in-house Java library. OMSSA (Version 2.1.9) and X!Tandem (released in 2013.02.01) were run through SearchGUI (1.12.2) [20] and MSGFDB (Plus-2012) was run individually.



Database size in kilo bytes
For the selected organisms

Organism	Size (kB)
E. coli	1802
Yeast	3873
Pig	13870
Mouse	28886
Human	43792

Fig. 1. Number of identifications (blue bars; left vertical axis) and quality distribution of the scores (box whisker plots, right vertical axis) for MSGFDB (top left), OMSSA (top right), and X!Tandem (bottom left). Database sizes are given on the bottom right and database size are ordered increasingly in the given plots. For this analysis fragment tolerance was 0.3 Da and precursor mass tolerance was 1.4 Da.

Thermo Scientific, the manufacturer of the LTQ mass spectrometer used for measurements, recommends users to use 1.4 Dalton (Da) precursor mass tolerance and around 0.4 Da fragment tolerance. In order to decide the proper settings for each algorithm, 2 Da as default by most algorithms and 1.4 Da precursor mass tolerances as recommended were used. For each precursor mass tolerance, fragment tolerance was varied from 0.1 Da to 1 Da by 0.1 Da increments.

In order to obtain maximal number of prediction and ignore the bias of sequence homology between peptide source proteins and proteins in the used database, we ran the algorithms for each precursor mass tolerance-fragment tolerance pair on different databases with increased exponential size. Other settings were as following; miss-cleavage was set to 1, peptide charge was between +1 to +3. Carboamidomethylation of Cys was set as fixed modification; number of results to be reported was limited to 10. Spectra were searched against 5 different databases with added source peptides and their decoy versions created by reversal of proteins in the target database. The proteomes of *Escherichia coli*, Yeast, Pig, Mouse, and Human were used in this study. Current releases of all databases were obtained from UniProtKB on March, 2013.

C. *Calculation of Consensus*

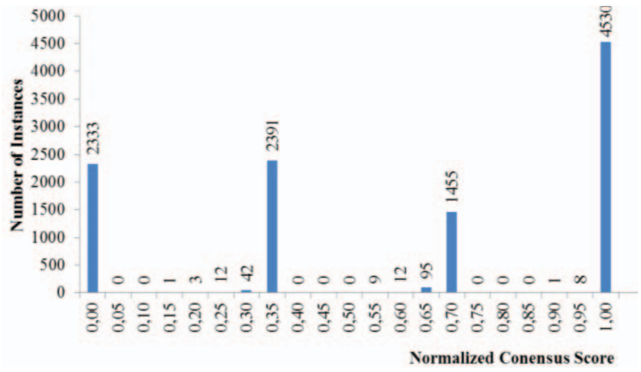


Fig. 2. The distribution of the normalized consensus score for all MS/MS data used in this study

For the calculation of the consensus score the three selected database search algorithms are used to predict the sequence of a mass spectrum and the best 10 results are stored. For each MS/MS spectrum the rank of the correct sequence was divided by 10 to get a normalized rank. Then the normalized ranks of the three algorithms were added up, using 1.5 in case the expected sequence was not among the best 10 results. The overall score was then normalized to fit into the range 0-1 and inverted such that the best score is 1 and the worst score is 0 which means that the correct peptide was not assigned to the tandem-MS spectrum by any of the database search engines employed. Fig. 2 shows the distribution of this normalized score for our data.

D. *Spectral Quality Assessment with ScanRanker*

The quality assessment of the spectra was performed by running the ScanRanker algorithm. According to the number of identifications for each algorithm, the largest number of predictions was observed at 1.4 Da precursor mass and 0.3 Da fragment tolerances. Thus, the ScanRanker algorithm was run on all spectra with the following settings: Precursor mass tolerance 1.4 Da, fragment mass tolerance 0.3 Da, usage of monoisotopic mass was selected and tag sequence length was set to 3. Other settings were left at their default settings and Spectral Removal option was turned off. ScanRanker scores were normalized to the range of 0 – 1 with better scores mapped to larger values.

E. *Speed Comparison*

For speed comparison of the selected database search tools and ScanRanker 1066 spectra were arbitrarily selected and all test were run on the same PC with all background programs turned off. The PC has 6 GB RAM and the processor is an Intel™ i3 running at 2.53 GHz.

III. Results and Discussion

A. *Varying Database Size*

We used the spectra from our benchmark dataset using a target decoy strategy to ensure that the mass spectra we measured can be consistently identified by the database search algorithms employed despite increasing database sizes. For this assessment only the best prediction per spectrum and search algorithm was retained. Fig. 1 shows that with an increase in database size the number of spurious identifications increases. This can be deduced from more identification in the decoy database where we do not expect any true positive identification. The number of false positive identifications in the decoy database to the true positive ones in the target database is then used to calculate the certainty of the identification using FDR calculation. Furthermore, a larger

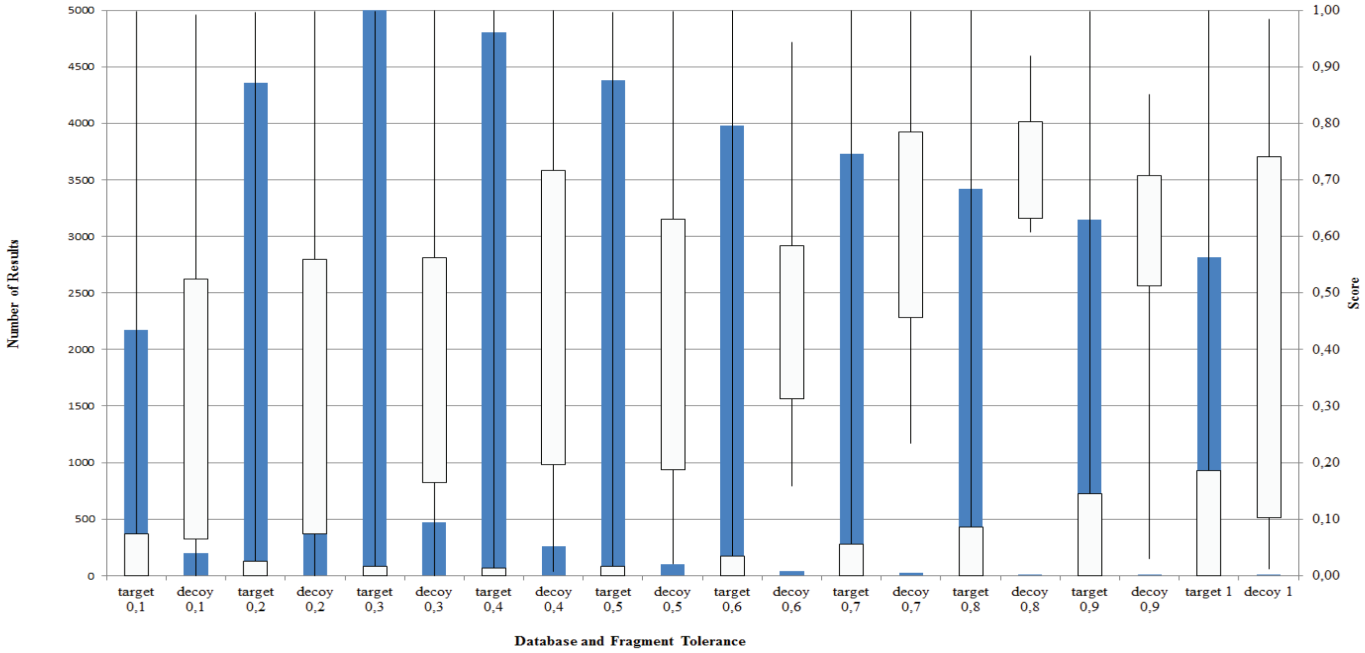


Fig. 3. Number of results in target and decoy databases and their associated score distribution for changing fragment tolerance in OMSSA on human protein database. Blue bars show the number of results and box and whisker plot shows the associated score distribution.

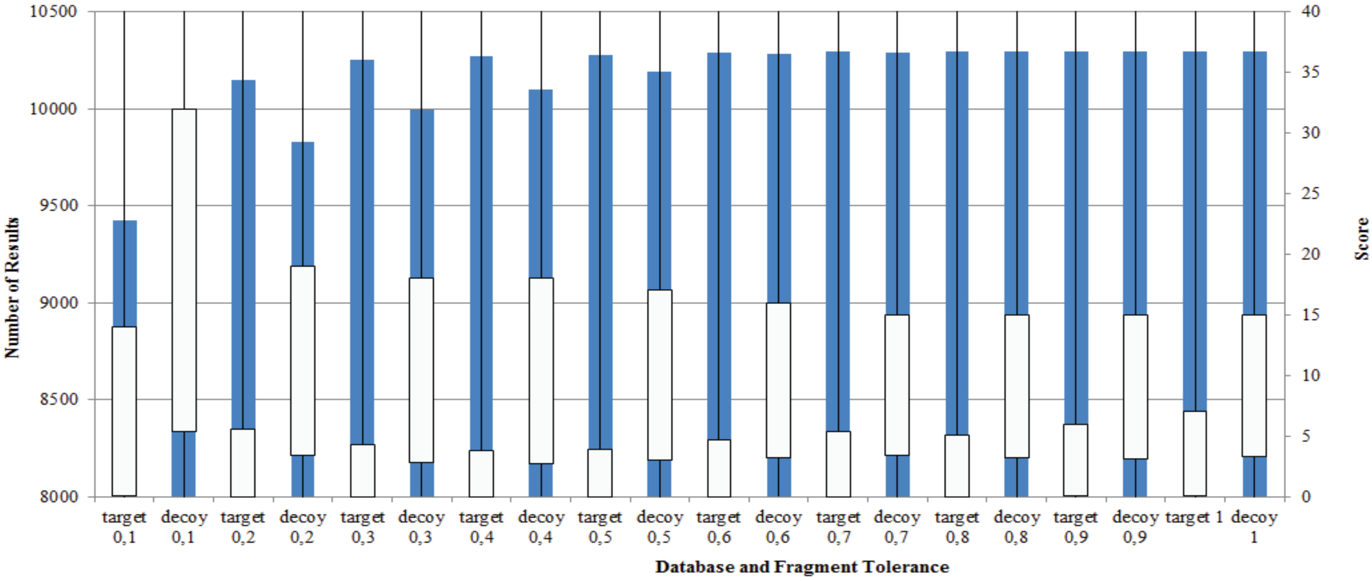


Fig. 4. Number of results in target and decoy databases and their associated score distribution for changing fragment tolerance in X!Tandem on human protein database. Blue bars show the number of results and box and whisker plot shows the associated score distribution.

overlap between score distribution for larger databases might have an impact on number of spurious identifications (Fig. 1). It also becomes clear, that the impact is not significant even for the largest database (human) used in this study the score overlap is not large enough to be detrimental for any of the database search algorithms. This result is somewhat contrary to what was shown in [12], but the size of the human proteome used here is insignificant compared to the database sizes that lead to problems in [12]. These results confirm that the consensus approach can be performed using the human database as a basis for the three database search algorithms.

Initial attempts with larger fragment tolerances than the one used to in Fig. 1 and that are commonly used in proteomics, lead to unexpected results (data not shown) which indicated the need to investigate the influence of fragment tolerance on the database search algorithms.

B. Varying Fragment Tolerance

In this assessment the human proteome was used as the database to be searched by the algorithms and the only variable was the fragment tolerance. We expect that with an increase in fragment tolerance more results will be found in the target and decoy databases. For OMSSA this is true up to a fragment tolerance of 0.3; but thereafter the number of results in the decoy database quickly diminishes to 0 and the results in the target database also decrease with an increase in fragment tolerance (Fig. 3) which is completely counter intuitive for how database search algorithms are expected to perform. A closer inspection of OMSSA revealed that this can be attributed to the scoring function which penalizes fragment tolerance. In contrast to OMSSA, X!Tandem behaves as expected (Fig. 4) Unlike the other two tools, MSGFDB does not offer an option to adjust fragment tolerance and therefore it was not assessed in this manner.

C. Consensus Scoring and ScanRanker Performance

As can be seen from Fig. 2 the consensus scoring method provides almost discrete scores which can either be due to the measurement, due to the nature of the algorithms employed or due to the design of our consensus method. According to the distribution (Fig. 2) four distinct groups of spectral quality were defined: ≤ 0.15, > 0.15 - 0.5, > 0.5 - 0.8, and > 0.8 and were named 0, 1, 2, and 3 consensus, respectively. The names coincide with the number of algorithms that agree for a given consensus.

Fig. 2 shows that the largest individual fraction of data we have is of high quality (~41%), which is to be expected since the measurement was of directly injected synthetic peptides. But it also becomes clear that there is a good distribution of data of lower and even low quality and an almost equal amount of examples are below 0.5 (~44%) and above a score of 0.5 (~56%). Therefore, we believe that this dataset is well suited to benchmark spectral quality assessment tools.

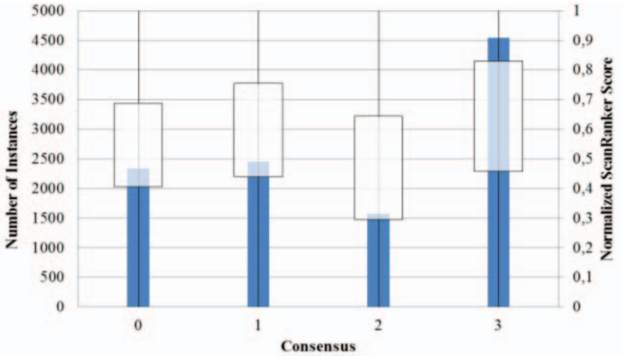


Fig. 5. Normalized ScanRanker score against the four identified consensus groups. The ScanRanker score distribution is presented as box and whisker plots using the right vertical axis. The number of instances for the groups are displayed as blue bars and are associated with the left vertical axis.

There are very complicated consensus scoring methods available, but here we do not aim to provide a consensus prediction of a PSM, but only want to assign a quality score to a spectrum based on its potential to be correctly identified by database search engines. The aim to assign a quality score to a spectrum is to define its potential to be useful for database search, *de novo* sequencing, or more advanced problems like blind detection of post translational modifications (PTMs). Therefore, we expected ScanRanker to give scores that approximately correlate with our consensus score. However, ScanRanker only shows a very slight trend from highest consensus to lowest consensus (Fig. 5). The boxes in the box and whisker plots significantly overlap for all consensus groups which means that the ScanRanker scoring is not discriminative for the LTQ data we used in this study.

The score distribution calculated by ScanRanker varies less for lower quality spectra and more for higher quality spectra (length of boxes in Fig. 5) while we would expect the opposite. Spectra where 2 algorithms agree on the consensus, result in the worst ScanRanker score distribution which may in part be due to the low number of such examples.

D. ScanRanker Speed

A quality scoring algorithm that aims to help decide whether a spectrum can and should be analyzed by more computationally costly downstream algorithms should not be computationally expensive itself. If it was computationally as involved as the downstream tools it would defeat its purpose. ScanRanker needs 28 ms per MS/MS spectrum on our dataset and the slowest database search algorithm (MSGFDB) needs 242 ms per MS^2 spectrum (Table I) which is likely due to it preparing the sequence database on the fly during these experiments. Overall, ScanRanker is too slow to be useful as a tool to decide spectral quality before employing other analysis algorithms since both OMSSA and X!Tandem are faster than ScanRanker (Table I).

TABLE I. SPEED COMPARISION OF DATABASE SEARCH ALGORITHMS AND SCANRANKER ON 1066 TANDEM-MS SPECTRA. RUNTIME IS PRESENTED IN MILLISECONDS (MS) PER MS/MS SPECTRUM. TABLE IS DECREASINLY SORTED BY RUNTIME.

Algorithm	***Speed per Spectrum [ms]***
MSGFDB	242
ScanRanker	28
OMSSA	24
X!Tandem	13

IV. CONCLUSION

Here we first show that it is acceptable to use the human proteins with the expected sequences appended to it for database search with OMSSA, MSGFDB, and X!Tandem since the database size is large enough to allow false positives but small enough to not force false negative identifications -. We further show that setting the fragment tolerance is crucial for OMSSA and that it must not be larger than 0.3 for our dataset.

Although other consensus methods have been proposed a simple consensus is completely sufficient in this study as can be deduced from Fig. 2 where it is seen that employing more complicated scoring, for instance including the normalized rank of the correct result within the result list, has a negligible effect. With these preliminaries, we then assessed the quality of ScanRanker spectral quality assessment.

While ScanRanker is not useful to actually rank the quality of MS/MS spectra for LTQ data, it may be useful for spectra from other instruments. Unfortunately, even if the quality assessment would suggest ScanRanker's usefulness, its runtime is too high to warrant its use for preprocessing data since it is slower than some of the downstream tools employed to process the data which defeats its purpose.

We propose that similar assessments be made on similar datasets from different mass spectrometers and using other proposed algorithms.

ACKNOWLEDGMENT

The authors would like to thank Talat Yalçın, Şule Yılmaz, and the Biological Mass Spectrometry and Proteomics facility at the Izmir Institute of Technology for their support in measuring the dataset.

REFERENCES

[1] M. Mann, R. C. C. Hendrickson, and A. Pandey, "Analysis of proteins and proteomes by mass spectrometry.," *Annual review of biochemistry*, vol. 70, pp. 437–73, Jan. 2001.

[2] J. M. Wells and S. A. McLuckey, "Collision-induced dissociation (CID) of peptides and proteins," *Methods in Enzymology*, vol. 402, pp. 148–185, 2005.

[3] J. Allmer, "Algorithms for the de novo sequencing of peptides from tandem mass spectra.," *Expert review of proteomics*, vol. 8, no. 5, pp. 645–57, Oct. 2011.

[4] E. A. Kapp, F. Schütz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn, and R. J. Simpson, "An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.," *Proteomics*, vol. 5, no. 13, pp. 3475–90, Aug. 2005.

[5] I. Shadforth, D. Crowther, and C. Bessant, "Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines," *Proteomics*, vol. 5, no. 16, pp. 4082–95, Nov. 2005.

[6] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–964, Oct. 2004.

[7] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, Jun. 2004.

[8] S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck, and P. A. Pevzner, "The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search.," *Molecular & cellular proteomics : MCP*, vol. 9, no. 12, pp. 2840–52, Dec. 2010.

[9] O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm, "TOPP--the OpenMS proteomics pipeline.," *Bioinformatics (Oxford, England)*, vol. 23, no. 2, pp. e191–7, Jan. 2007.

[10] P. G. A. Pedrioli, "Trans-proteomic pipeline: a pipeline for proteomic analysis.," *Methods in molecular biology (Clifton, N.J.)*, vol. 604, pp. 213–38, Jan. 2010.

[11] N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith, and P. A. Pevzner, "Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation.," *Genome research*, vol. 17, no. 9, pp. 1362–77, Sep. 2007.

[12] M. Vaudel, J. M. Burkhart, D. Breiter, R. P. Zahedi, A. Sickmann, and L. Martens, “A complex standard for protein identification, designed by evolution.,” *Journal of proteome research*, vol. 11, no. 10, pp. 5065–71, Oct. 2012.

[13] Z.-Q. Ma, M. C. Chambers, A.-J. L. Ham, K. L. Cheek, C. W. Whitwell, H.-R. Aerni, B. Schilling, A. W. Miller, R. M. Caprioli, and D. L. Tabb, “ScanRanker: Quality Assessment of Tandem Mass Spectra via Sequence Tagging.,” *Journal of proteome research*, vol. 10, no. 7, pp. 2896–904, Jul. 2011.

[14] J. Allmer, C. H. Markert, E. J. Stauber, and M. Hippler, “A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases.,” *FEBS Letters*, vol. 562, no. 1–3, pp. 202–206, Mar. 2004.

[15] J. Allmer, “A Call for Benchmark Data in Mass Spectrometry-Based Proteomics,” *Journal of Integrated OMICS*, 2012.

[16] T. Sultana, R. Jordan, and J. Lyons-Weiler, “Optimization of the Use of Consensus Methods for the Detection and Putative Identification of Peptides via Mass Spectrometry Using Protein Standard Mixtures.,” *Journal of proteomics & bioinformatics*, vol. 2, no. 6, pp. 262–273, Jun. 2009.

[17] R. K. Dagda, T. Sultana, and J. Lyons-Weiler, “Evaluation of the Consensus of Four Peptide Identification Algorithms for Tandem Mass Spectrometry Based Proteomics.,” *Journal of proteomics & bioinformatics*, vol. 3, pp. 39–47, Feb. 2010.

[18] G. Alves, W. W. Wu, G. Wang, R.-F. Shen, and Y.-K. Yu, “Enhancing peptide identification confidence by combining search methods.,” *Journal of proteome research*, vol. 7, no. 8, pp. 3102–13, Aug. 2008.

[19] E. W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, J. K. Eng, D. B. Martin, A. I. Nesvizhskii, and R. Aebersold, “A guided tour of the Trans-Proteomic Pipeline.,” *Proteomics*, vol. 10, no. 6, pp. 1150–9, Mar. 2010.

[20] M. Vaudel, H. Barsnes, F. S. Berven, A. Sickmann, and L. Martens, “SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches.,” *Proteomics*, vol. 11, no. 5, pp. 996–9, Mar. 2011.