COSS: A fast and user-friendly tool for spectral library searching

Genet Abay Shiferaw [1,2], Elien Vandermarliere [1,2], Niels Hulstaert [1,2], Ralf Gabriels [1,2], Lennart Martens [1,2], Pieter-Jan Volders [1,2,3]

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium

[3] Cancer Research Institute Ghent, Ghent University, 9000 Ghent, Belgium

Corresponding Author

*Prof. Dr. Lennart Martens, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. E-mail: lennart.martens@vib-ugent.be , Tel: +3292649358

ABSTRACT: Spectral similarity searching to identify peptide-derived MS/MS spectra is a promising technique, and different spectrum similarity search tools have therefore been developed. Each of these tools, however, comes with some limitations, mainly due to low processing speed and issues with handling large databases. Furthermore, the number of spectral data formats supported is typically limited, which also creates a threshold to adoption. We have therefore developed COSS (CompOmics Spectral Searching), a new and user-friendly spectral library search tool supporting two scoring functions. COSS also includes decoy spectra generation for result

validation. We have benchmarked COSS on three different spectral libraries and compared the results with established spectral searching tools and sequence database search tool. Our comparison showed that COSS more reliably identifies spectra, is capable of handling large datasets and libraries and is an easy to use tool that can run on low computer specifications. COSS binaries and source code can be freely downloaded from https://github.com/compomics/COSS.

KEYWORDS: tandem mass spectrometry, peptide identification, spectral library searching.

INTRODUCTION

Tandem mass spectrometry (MS/MS) is a commonly used method to analyze and identify peptides and proteins. Typically, MS/MS analysis and identification consists of several steps[1]. First, an unknown protein mixture is digested into peptides with the aid of a protease, and the resulting peptides are then separated in time by liquid chromatography (LC). This LC is coupled directly to a mass spectrometer's source where the eluting peptides are detected, selected, and fragmented. The resulting fragment ions are then analyzed by a second stage of mass spectrometry to acquire an MS/MS spectrum. These MS/MS spectra can then be subjected to different computational approaches to match them to peptide sequences.

Commonly used approaches are *de novo* sequencing, sequence database searching, and spectral library searching. *De novo* sequencing[2] algorithms directly infer the amino acid sequence from the experimental spectrum. In this technique, the quality of the spectrum affects the success of the inference process and hence the identification result. Therefore, the identification rate of such algorithms in practice is typically limited[3], in turn limiting their use. In sequence database searching, an *in silico* digest of a protein sequence database produces a list of peptide sequences,

each of which is then used to generate theoretical mass spectra. These theoretical spectra are subsequently compared with experimental spectra using a similarity scoring function. Due to their performance, sequence database search engines are the most widely used approach to analyze MS/MS data. Nevertheless, despite its popularity, database searching comes with some drawbacks[4]. The first problem with database searching is the computational complexity imposed when working with large databases. As the algorithm needs to consider all possible peptides derived from a protein sequence, the resulting databases will grow exponentially when taking into account multiple missed cleavages and a variety of potential post-translational modifications (PTMs)[3]. Another important disadvantage of database searching is the lack of peak intensity information and information on non-canonical fragments in the generated theoretical spectra, which limits the sensitivity of the scoring function.

Spectral library searching seeks to correct for these two issues, by comparing experimental spectra to a spectral library built from previously identified spectra[5]. Nowadays, this spectral library searching approach is gaining more attention due to a number of advantages[6]. Because the search space is confined to previously observed and identified peptides, the computational complexity is reduced[7]. Moreover, spectral searching can take advantage of all spectral features, including actual peak intensities and the presence of non-canonical fragment ions[8], to determine the best possible peptide match. As a result, this technique often yields improved sensitivity[9].

Different tools to apply spectral library searching have been developed over the past years, with SpectraST[10], the National Institute of Standards and Technology (NIST) MS-Search[11] and MSPepSearch (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch), ANN-SoLo[12] and X!Hunter[13] as notable examples. Each of these tools, however, comes with some limitations, such as some of them run with a low processing speed, issues with handling large

databases, and operational complexity. Furthermore, these tools typically support only specific spectral data formats, which also creates a threshold to adoption if the desired library is not presented in a compatible format. Taken together, these issues have prevented widespread adoption of the spectral library searching approach in proteomics.

We have therefore developed COSS (CompOmics Spectral Searching), a new, fast, and user-friendly spectral library search tool capable of processing large databases and supporting different file formats. Two scoring functions are available in COSS, namely MSROBIN, which relies on probabilistic scoring, and the cosine similarity score. COSS also offers an intuitive graphical user interface, allowing it to be adopted easily. To control the false discovery rate, a built-in mechanism to generate decoy spectral libraries has been provided as well. We have benchmarked COSS on three different spectral libraries and our results show that, compared to established tools, COSS delivers more reliable identification. At the same time, COSS requires a reasonable lower computation time than most other algorithms and has a much-reduced memory footprint, eliminating the requirement for high performance and costly equipment, and further lowering the threshold to adoption of the spectral library searching approach.

MATERIALS AND METHODS

Implementation

COSS is developed in Java in a modular fashion so that its code is reusable and future-proof. Separate modules have been developed for key tasks such as indexing, filtering, matching, and decoy generation. To ensure maximal compatibility with input formats, the spectrum reader has been developed as a separate subsystem. COSS supports mzXML[14], mzML[15], ms2 and dta input formats through the mzIdentML[16] library, while support for the msp and mgf formats is included through an in-house implementation. The compomics-utilities[17] library was used for spectra visualization. Because COSS is completely developed in Java, it is platform independent, allowing users to run the software in their own preferred environment (e.g., Windows, Linux, or MacOS).

Scoring function

COSS implements two scoring functions: MSROBIN, which is based on the probabilistic scoring function of Yilmaz et al.[18], itself a derivative of the Andromeda scoring function[19] and the cosine similarity score. The scoring procedure consists of two main steps. First, both the query and library spectrum are divided into 100 Da windows and within each window, the q peaks with the highest intensity are selected. Next, the score is calculated for q varying from 1 to 10 and the highest score is retained. The MSROBIN scoring function consists of two parts, an intensity part and a probability part. The probability scoring part is as follows:

$$Pscore(k,p,n) = \sum_{j=k}^{n} \binom{n}{k} p^{j}(1 \quad p)^{n-j}$$

Where n is the number of peaks, *k* is number of matched peaks, and *p* is the probability of finding a match for a single matched peak, calculated by dividing the number of retained high intensity peaks by the mass window size which we set at 100 Da.

The second part is the intensity scoring which is calculated as:

$$Iscore = \frac{\sum_{x \quad X} I_x}{\sum_{y \quad Y} I_y}_{ExpSpec} \times \frac{\sum_{x \quad X} I_x}{\sum_{y \quad Y} I_y}_{LibSpec}$$

Here, *I* is the peak intensity, *X* is the set of matched peaks and *Y* is the set of intense peaks selected from each 100 Da window. The final score is then computed as:

$$Score = [\ \ 10 \times \log_{10} \ Pscore \ \ ] \times Iscore$$

We have calculated the cosine similarity scoring function as follows:

$$Score = \frac{\sum_{i=0}^{N} (Q_i \quad L_i)}{\sqrt{\sum_{i=0}^{N} Q_i^2} \quad \sqrt{\sum_{i=0}^{N} L_i^2}} \quad \text{N}$$

Where Q is the intensity of the matched peak found in the query spectrum, L the intensity of the matched peak found in the library spectrum and N is the total number of matched peaks between query and library spectra under comparison. The score is weighted by the number of matched peaks.

False discovery rate estimation

Erroneous peptide assignments can occur due to poor spectrum quality or limitations in the scoring function. Validation of the obtained results is therefore a key step in peptide identification, and typically takes the form of false discovery rate (FDR) control[20]. For this purpose, COSS implements a decoy spectral library strategy, which can generate a number of decoy spectra equal to the size of the original spectral library using reverse and random sequence decoy generation technique as described in Zhang et al.[21]. Briefly, the sequence of each spectrum is reversed, leaving the last amino acid in place. Based on this sequence, the masses of the *a*, *b* and *y* ions are calculated and the corresponding annotated peaks in the spectrum are moved on the m/z axis accordingly leaving the unannotated peaks in place.

The generated decoy spectra are concatenated to the original spectra in the library, and the search is run against this concatenated target-decoy spectral library. The corrected FDR value is then calculated as described previously in Sticker et al.[20].

To evaluate whether the generated decoys accurately control FDR, we have used a modified entrapment method[22]. For this, we have obtained a tandem mass spectrometry dataset of *Pyrococcus furiosus* (ProteomeXchange[23] ID PXD001077) which contains a total of 15,615 spectra. Unlike the original entrapment method, where *Pyrococcus furiosus* spectra are appended to the library spectra, we have appended it to our query spectra and then ran the search against the NIST library concatenated with the generated decoy spectra. FDR and false discovery proportion (FDP) are then calculated as follows:

$$FDR = \frac{\#decoy}{\#target}$$

$$FDP = \frac{\frac{\#target_{Pyrococcus}}{f}}{\#target_{Sample}}$$

With *f* the fraction of *Pyrococcus furiosus* spectra over non-*Pyrococcus* spectra in the search.

Benchmarking datasets and spectral libraries

We obtained raw data files from eleven runs from the Human Proteome Map[24] (ProteomeXchange[23] ID PXD000561) and ten runs from the deep proteome and transcriptome abundance atlas[25] dataset (ProteomeXchange ID PXD010154) as benchmarking data sets (Supplementary Table S-1). All these 21 raw files were converted to Mascot Generic Format (mgf) format using the *msconvert* tool (ProteoWizard[26]), with the peak picking algorithm activated. Benchmarking was performed using three distinct spectral libraries (Table 1), obtained from PRIDE[27], NIST and MassIVE[26].

Table 1. Spectral libraries used to benchmark COSS.

Spectral Library	Total number of spectra	URL	Version
PRIDE Cluster[27] (Human)	789,745	https://www.ebi.ac.uk/pride/cluster/#/libraries	2015-04
NIST (human HCD library)	1,127,970	https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503	July 25, 2018
MassIVE[26] (Human HCD Spectral Library)	2,154,269	http://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp	1.3.3

Running searches

All benchmarking is performed on the same virtual machine, equipped with dual Xeon E5-242016 processors at 1.90GHz, 28 GB of RAM, and running the Microsoft Windows 10 operating system. To run SpectraST, we used the Trans Proteomic Pipeline (TPP v5.2.0-b1) software for Windows. SpectraST was run in command line mode according to the user manual (http://tools.proteomecenter.org/wiki/index.php?title=Software:SpectraST). The spectral libraries, originally in msp format, were first converted to the splib file format, and then a consensus spectrum from the generated splib file was created. Quality control was applied on this consensus file, and finally decoy spectra were generated and appended to the consensus file. Search settings were a precursor m/z tolerance of 0.01 Th, with the rest of the settings left at their defaults (Supplementary methods).

Since MSPepSearch (version 0.96) does not include decoy generation functionality, we first generated the decoy appended library using COSS. Next, using Lib2NIST (version 1.0.6.5) we have converted the msp file to MSPepSearch's binary file format and performed the searches using a precursor tolerance of 10 ppm and fragment tolerance 0.05 Da.

For ANN-SoLo, libraries were generated as recommended on the ANN-SoLo wiki page (https://github.com/bittremieux/ANN-SoLo/wiki/Search). In brief, SpectraST was used for library preparation. First, libraries were converted from the msp format to the splib file format. Next, a consensus spectral library was constructed from the generated splib file followed by a quality control step using through SpectraST's quality filters. Finally, decoy spectra were generated and appended to the consensus file. After library preparation, the search is performed in ANN-SoLo using a precursor tolerance of 10 ppm, a fragment tolerance of 0.05 Da. and an FDR threshold of 0.01.

Sequence database searches using MS-GF+[28] (version v2018.04.09), have been performed through SearchGUI[29] (version 3.3.16) and PeptideShaker[30] (version 1.16.42). The search database is constructed from the human proteome (UP000005640) as obtained from UniProt[31] (consulted on 9/10/2018). Carbamidomethylation of cysteine and oxidation of methionine are used as fixed, and as variable modification respectively. Trypsin is used as protease and a maximum of two missed cleavages was allowed. Precursor m/z tolerance was set to 10 ppm, and fragment tolerance to 0.05 Da. Precursor charges from 2 to 4 are considered.

To run COSS, decoy spectra were generated using the reversed sequence decoy generation technique. For the MassIVE and PRIDE libraries, where spectra are not (fully) annotated, we have first annotated the library using the built-in spectra annotator in COSS and a fragment tolerance of 0.05 Da. In this way, *a*, *b* and *y* ions where annotated, taking into account +1, +2 and +3 as possible charges and H2O and NH3 as possible neutral losses. The generated decoys were appended to the original spectral library and searches were performed using a precursor m/z tolerance of 10 ppm and a fragment m/z tolerance of 0.05 Da.

RESULTS AND DISCUSSION

Graphical user interface

COSS comes with a user-friendly interface that allows the user to set all parameters (Supplementary Figure S-1) needed for spectral similarity search. COSS supports most common MS/MS spectrum formats (including mgf, msp, ms2, mzML, mzXML, and dta). The user can generate decoy spectra for their spectral library using two types of decoy generation techniques that are implemented and integrated in COSS. COSS also provides an intuitive interface to visually inspect the obtained results (Figure 1). This interface reports all experimental spectra with matches in the spectral library in an interactive table, sorted by descending match score. When a query spectrum is selected, the top 10 matched spectra from the spectral library are displayed in the bottom table. For each match, the query spectrum and the matched library spectrum can be visually inspected. The results can be exported in tab-delimited text format, comma-delimited text format (CSV) and Microsoft Excel format (xlsx) for further processing and reporting. In addition to the graphical user interface, COSS also comes with a documented command-line interface to easily deploy the software on servers and high-performance clusters. The flexibility of COSS is further enhanced by its ability to run on all common operating systems.

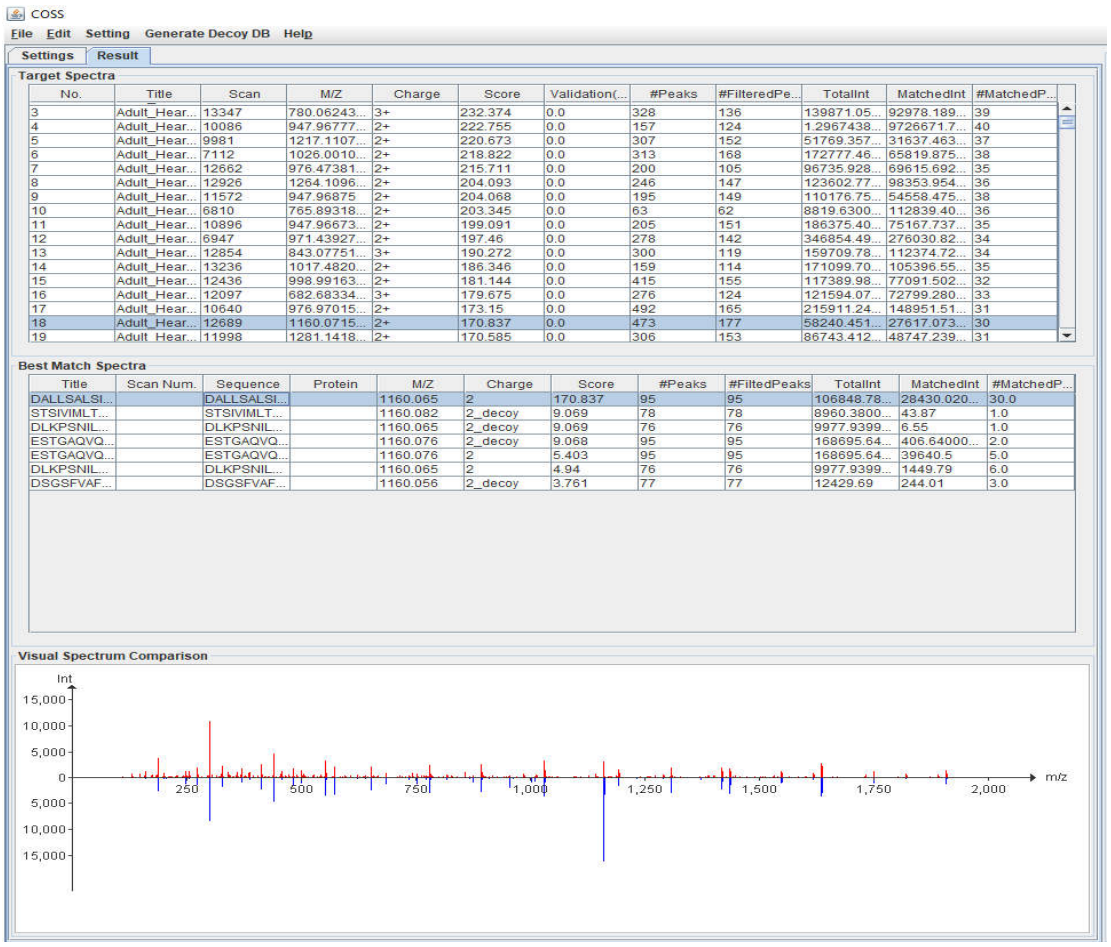

Figure 1. Search result interface of COSS: the upper table lists the experimental spectra while the lower table lists the top 10 matched spectra for the selected experimental spectrum. An interactive spectrum comparison view is presented at the bottom with the selected experimental spectrum (red) mirrored with the selected matched library spectrum (blue).

Evaluation of false discovery rate estimation

To evaluate the ability of COSS to accurately assess the FDR based on the implemented decoy generation technique, we used the modified entrapment approach using *Pyrococcus furiosus* spectra[22]. Our results show that while SpectraST underestimates the FDR (at 1% estimated FDR, the actual FDP as measured by *Pyrococcus* identifications is 2.57%) while COSS and MSPepSearch assess it quite accurately (at 1% estimated FDR, actual FDP is 1.8% and 1.6% respectively, Figure 2, Supplementary Figure S-2). Of note: ANN-SoLo is not included in this comparison since decoy hits are not reported in the output file.

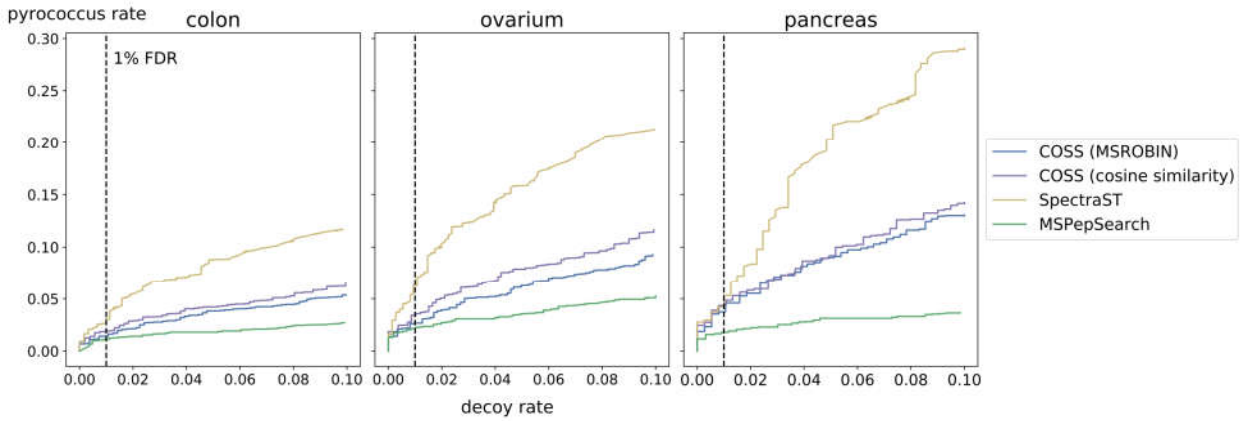

Figure 2. COSS accurately controls the FDR. Here, decoy rate (FDR) is shown in conjunction with the *Pyrococcus* rate (FDP) for search results from three runs from the Human Proteome Map. Shown are the FDR and FDP from COSS with the MSROBIN scoring function, COSS with cosine similarity scoring function, SpectraST and MSPepSearch. At the 1% FDR level, both scoring functions from COSS and MSPepSearch accurately assess the FDR while SpectraST underestimates it. Interestingly, MSPepSearch actually slightly overestimates the FDR at higher FDR values while the estimates from COSS remain very close to the actual FDR.

Search result comparison

Figure 3 shows the identification rate of COSS (MSROBIN and cosine similarity scoring functions), SpectraST, MSGF+, MSPepSearch and ANN-SoLo at 1% FDR (Supplementary Table S-2). The overall identification rates of COSS with the MSROBIN scoring function are 9.4%, 13.4% and 16.1% against the PRIDE, NIST, and MassIVE libraries respectively, while the cosine scoring function of COSS obtains 8.8%, 12.7% and 15.5% against PRIDE, NIST and MassIVE respectively. The identification rates thus nicely follow the size of the spectral library, with PRIDE being the smallest and MassIVE the largest, showing that the larger spectral libraries offer a more complete coverage of the proteome. It is important to note that the PRIDE spectral library is the oldest and also includes data with lower resolution fragment ions, which could further explain the lower identification rates observed with this library. The identification rates of SpectraST are very similar at 9.0% and 16.0% against PRIDE and NIST respectively. It should be noted that the default SpectraST parameters may not be ideal for high-resolution spectra and we did not perform extensive optimization to evaluate whether the identification rate could be improved. SpectraST's performance against the MassIVE database could not be assessed as SpectraST could not handle a library of this size, even on our server with 28GB of RAM. MSPepSearch results are comparable to COSS and SpectraST for the PRIDE and NIST libraries with 8.9% and 11.3% respectively. However, the performance against the MassIVE library is very poor at 2.3% compared to the 16.1% achieved with COSS. ANN-SoLo identifies 4.1% and 12.9% against PRIDE and NIST libraries. Since ANN-SoLo relies on SpectraST to generate the indexed libraries, we faced the same issues in the decoy generation step and were unable to produce results for the MassIVE spectral library.

Overall, the identification rate of COSS is higher compared to the other three spectral library searching tools and the identification rate scales consistently with the size of the spectral library. When comparing to the sequence database search approach (MS-GF+, 15.3% identification rate) we observe overall lower identification rates for the smaller spectral libraries (NIST and PRIDE), while the identification rates against MassIVE are very much in line.

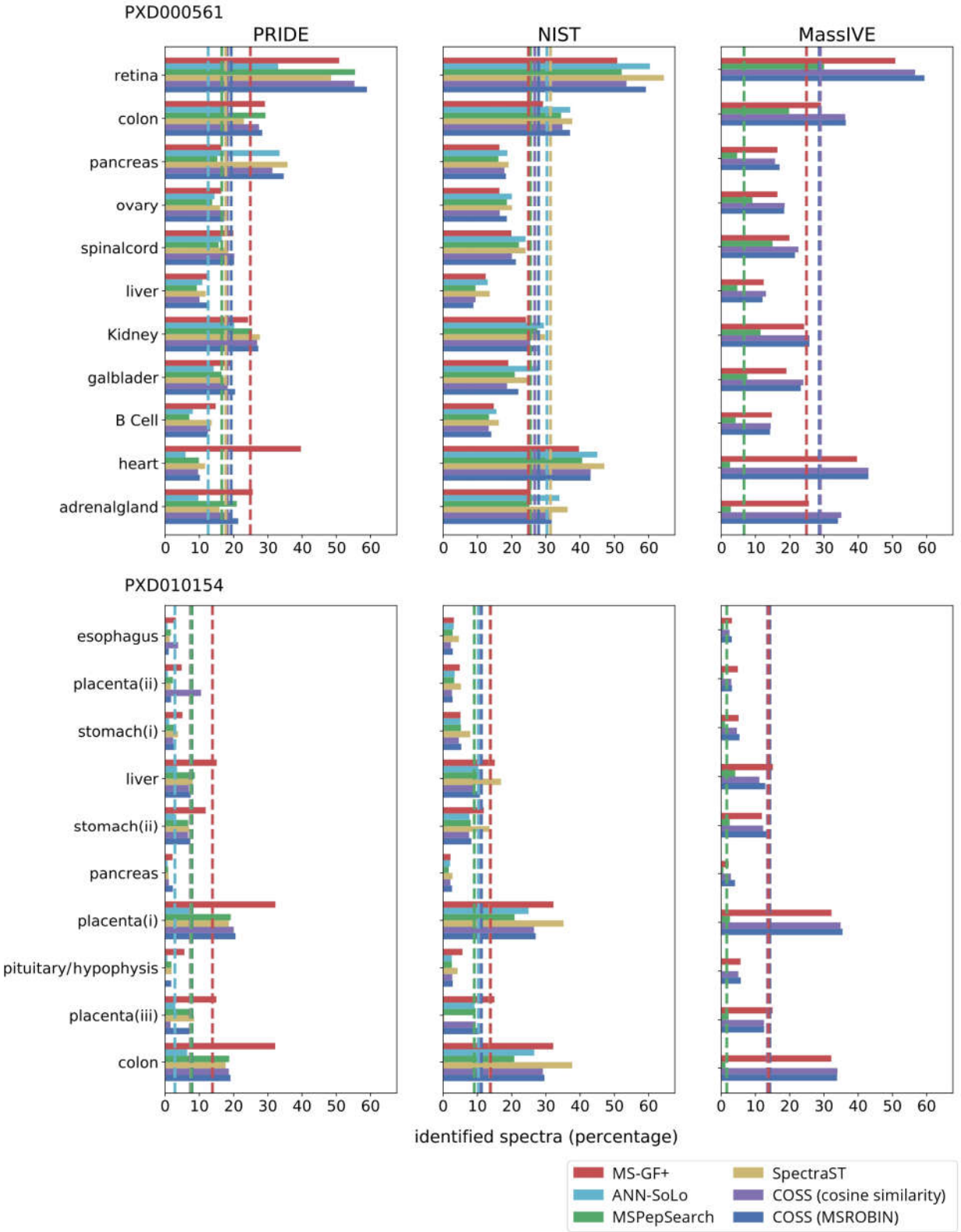
PXD000561
PRIDE
NIST
MassIVE
retina
colon
pancreas
ovary
spinalcord
liver
Kidney
galblader
B Cell
heart
adrenalgland
PXD010154
esophagus
placenta(ii)
stomach(i)
liver
stomach(ii)
pancreas
placenta(i)
pituitary/hypophysis
placenta(iii)
colon
0 10 20 30 40 50 60
identified spectra (percentage)
MS-GF+
ANN-SoLo
MSPepSearch
SpectraST
COSS (cosine similarity)
COSS (MSROBIN)

Figure 3. COSS performance evaluation against SpectraST, MSPepSearch, ANN-SoLo and sequence database searching in terms of identification rate. Shown here is the identification rate at 1% FDR against the PRIDE Cluster, NIST and MassIVE spectral libraries for COSS and SpectraST, and against the human proteome sequence database for MS-GF+. The dashed vertical lines represent the overall identification rate for that tool and library result across all samples in the dataset. Due to excessive memory requirements, SpectraST and ANN-SoLo could not run the MassIVE spectral library on our server with 28GB of RAM.

The identifications of the spectral library tools show a good overlap in terms of the identified peptides (Figure 4, Supplementary Figures S-3, S-4 and S-5). While COSS (MSROBIN score), MsPepSearch and ANN-Solo have a high degree of overlap (Figure 4), SpectraST has more unique identified peptides. This can probably be attributed to an underestimation of the FDR by SpectraST (Figure 2 and Supplementary Figure S-2). The identification overlap between the two scoring functions of COSS is very large in all the three different libraries (Supplementary Figure S-6).

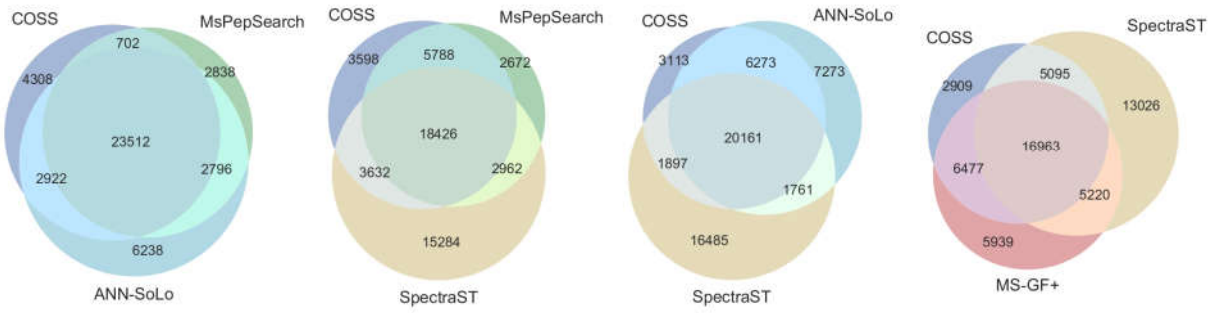

Figure 4. Search result overlap of COSS (MSROBIN score), SpectraST, MSpepSearch, and ANN-SoLo against the NIST spectral library. Intersections represent spectra with identical peptide identification. The results represent the union of the identifications in both datasets (PXD000561 and PXD010154) at 1% FDR. With the exception of SpectraST, the sequence level identifications of the different tools are very much in agreement. The discrepancy and higher identification rate observed for SpectraST can probably be attributed to an underestimation of the FDR (Figure 2 and Supplementary Figure S-2).

Execution time comparison

To evaluate the computational efficiency of the algorithm, we ran COSS, SpectraST, MSPepSearch and ANN-SoLo on the same data sets using the same virtual machine and recorded the execution time for each algorithm. The results of the comparison are shown in Figure 5. While the size of the query dataset and the spectral library both clearly influence the executing time, we found that COSS outperforms SpectraST, yet ANN-SoLo and MSPepSearch are faster in all cases. Here again, there is no runtime information for SpectraST and ANN-SoLo for the MassIVE library due to the inability to run them against this library on our server.

It should be noted that only the time required for searching is considered here and library preparation is not taken into account. Library format conversion, decoy generation and indexing take a substantial amount of time. For the NIST library, SpectraST took 12+ hours to prepare the library and generate and append decoy spectra. MSPepSearch and ANN-SoLo both rely on SpectraST for decoy generation, yet after decoy generation MSPepSearch libraries need to be converted to a binary format which took an additional four hours. For COSS, all these steps combined took around 20 minutes for the NIST library on the same machine.

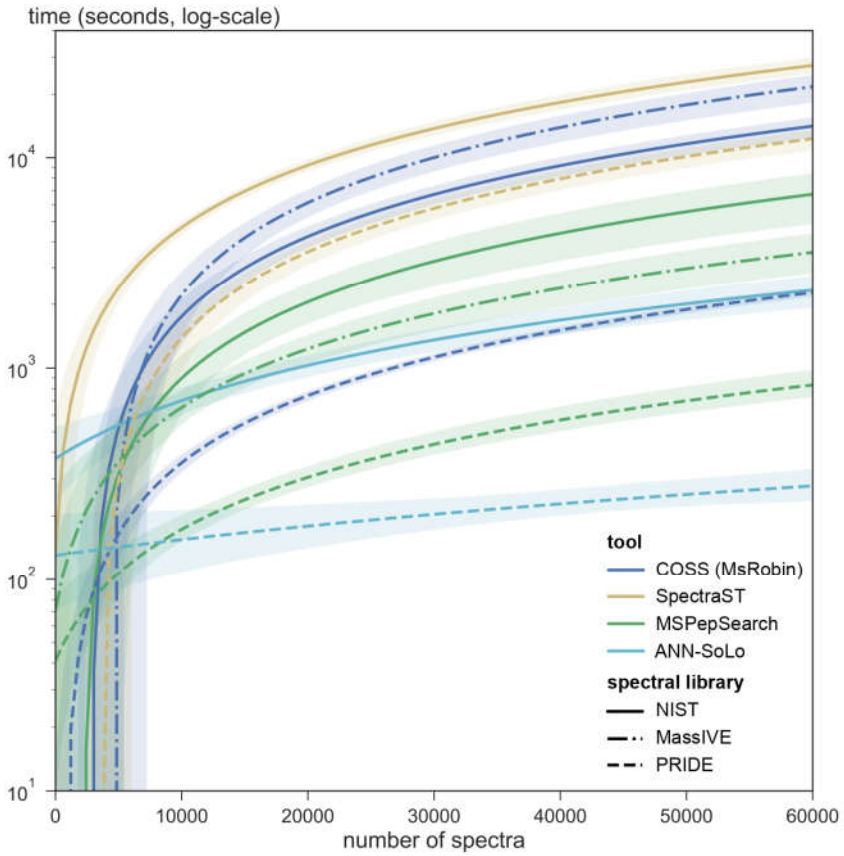

Figure 5. Execution times of COSS (using MSROBIN score), SpectraST, MSPepSearch and ANN-SoLo. While COSS is faster than SpectraST, MSPepSearch and ANN-SoLo outperform COSS. The execution time however only considers the time required for searching and does not include spectral library preparation which takes a considerable amount of time for SpectraST, MSPepSearch and ANN-Solo. Interestingly, the runtime for MSPepSearch against the MassIVE library is much lower than against the NIST library despite the former being much larger in size. This is in line with the low identification rate observed for MSPepSearch on the MassIVE library (Figure 3). There is no runtime information for SpectraST and ANN-SoLo for the MassIVE library due to the inability to run them against this library on our server.

CONCLUSIONS

There is a need for spectral library search tools that can easily analyse data from today's high-throughput mass spectrometry-based proteomics experiments, and that can match tens of thousands of acquired spectra against proteome-wide spectral libraries. A few such search

algorithms like SpectraST, MSPepSearch and ANN-SoLo have already been developed but come with important limitations: an inability to handle very large spectral libraries, limited input file format support and only usable by advanced users due to lack of graphical user interface. Here we present COSS, a user-friendly spectral library search tool that is fast, can handle large datasets, and supports the most commonly used MS/MS data formats. COSS also includes library and decoy generating in a single step making COSS the "all-in-one" tool to perform spectral library searching very easily. COSS offers both a graphical as well as a command-line interface, enabling users to perform anything from small-scale analyses on laptops to automated, large-scale data reprocessing on high-performance compute clusters. Because COSS is developed in Java, it is also platform independent, allowing it to run seamlessly on all commonly used operating systems. Furthermore, COSS's modular architecture and open-source code invites and facilitates future development by the community at large. We have compared COSS to SpectraST, MSPepSearch, ANN-SoLo and a sequence database search algorithm, MS-GF+, in terms of identification performance, and found that COSS offers a more reliable identification and a reasonable runtime even with a large query dataset and large spectral libraries. While MSPepSearch and ANN-SoLo offer faster searches, library generation is much slower and more cumbersome compared to COSS. The identification rates for the PRIDE and NIST libraries are comparable between the tools, yet COSS outperforms all the other tools on the large MassIVE library which contains 2+ million spectra. Combined, these properties make COSS a user-friendly tool, highly suitable for large-scale analyses against ever expanding spectral libraries, including those that aim to cover an entire proteome of an organism.

AVAILABILITY

The software and its source code can be freely downloaded from https://github.com/compomics/COSS and is licensed under the permissive, open-source Apache License, version 2.0.

ACKNOWLEDGMENTS

This project is supported by the National Institute of Health (NIH) [NCI-ITCR grant number 1U24CA199347 to G.A.S.], Research Foundation - Flanders (FWO) [3E023815 to E.V., grant number 1S50918N to R.G., grant number G042518N to L.M.], the Horizon 2020 programme of European Union project EPIC-XS [grant number 823839], and Kom op tegen Kanker (Stand up to Cancer), the Flemish cancer society [to P.V.]. We would like to thank Zheng Zhang from NIST, USA for providing us with the R code for generating decoy spectra libraries. We would like to thank Wout Bittremieux for helping us in running the ANN-SoLo searches. We would also like to thank all the CompOmics group members for their ideas, discussions and support.

REFERENCES

(1) Ingvar, Eidhammer; Kristian, Flikka; Lennart, Martens; Svein-Ole, M. Protein Identification and Characterization by MS. In *Computational Methods for Mass Spectrometry Proteomics*; **2007**; p 97,98.

(2) Hughes, C.; Ma, B.; Lajoie, G. A. De Novo Sequencing Methods in Proteomics BT - Proteome Bioinformatics; Hubbard, S. J., Jones, A. R., Eds.; Humana Press: Totowa, NJ, **2010**; pp 105–121.

(3) Costa, E.; Menschaert, G.; Luyten, W.; Grave, K. De; Ramon, J. Peptide Identification Using Tandem Mass Spectrometry Data. *Tech. Rep.* **2013**.

(4) Yen, C.-Y.; Stephane, H.; G, A. N.; Old; William M. Spectrum-to-Spectrum Searching Using a Proteome-Wide Spectral Library. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.007666.

(5) Lam, H.; Aebersold, R. Using Spectral Libraries for Peptide Identification from Tandem Mass Spectrometry (MS/MS) Data. *Curr. Protoc. Protein Sci.* **2010**, *2010*.

(6) Lam, H.; Aebersold, R. Building and Searching Tandem Mass (MS/MS) Spectral Libraries for Peptide Identification in Proteomics. *Methods*. **2011**.

(7) Ahrné, E.; Masselot, A.; Binz, P. A.; Müller, M.; Lisacek, F. A Simple Workflow to Increase MS2 Identification Rate by Subsequent Spectral Library Search. *Proteomics* **2009**, *9* (6), 1731–1736.

(8) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. SPECIAL FEATURE : Mobile and Localized Protons : A Framework for Understanding Peptide Dissociation. *J. Mass*

Spectrom. **2000**, *1406* (September), 1399–1406.

(9) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the Improved Sensitivity of Spectral Library Searching over Sequence Database Searching in Proteomics Data Analysis. *Proteomics* **2011**, *11* (6), 1075–1085.

(10) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667.

(11) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994**.

(12) Bittremieux, W.; Meysman, P.; Sta, W.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **2018**, *17*, 3463–3474.

(13) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.* **2006**, *5* (8), 1843–1849.

(14) Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; et al. A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research. *Nat. Biotechnol.* **2004**, *22* (11), 1459–1466.

(15) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Ro, A.; Neumann, S.; Pizarro, A. D.; et al. MzML — a Community Standard for Mass Spectrometry Data*. *Mol. Cell. Proteomics* **2011**, 1–7.

(16) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; et al. The MzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Mol. Cell. Proteomics* **2012**, *11* (7), M111.014381.

(17) Barsnes, H.; Vaudel, M.; Colaert, N.; Helsens, K.; Sickmann, A.; Berven, F. S.; Martens, L. Compomics-Utilities: An Open-Source Java Library for Computational Proteomics. *BMC Bioinformatics* **2011**, *12* (1), 70.

(18) Yllmaz, Ş.; Victor, B.; Hulstaert, N.; Vandermarliere, E.; Barsnes, H.; Degroeve, S.; Gupta, S.; Sticker, A.; Gabriël, S.; Dorny, P.; et al. A Pipeline for Differential Proteomics in Unsequenced Species. *J. Proteome Res.* **2016**, *15* (6), 1963–1970.

(19) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10* (4), 1794–1805.

(20) Sticker, A.; Martens, L.; Clement, L. Mass Spectrometrists Should Search for All Peptides, but Assess Only the Ones They Care About. *Nat. Methods* **2017**, *14*, 643.

(21) Zhang, Z.; Burke, M.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Markey, S. P.; Yu, W.; Chaerkady, R.; Hess, S.; Stein, S. E. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *J. Proteome Res.* **2018**, *17* (2), 846–857.

(22) Vaudel, M.; Burkhart, J. M.; Breiter, D.; Zahedi, R. P.; Sickmann, A.; Martens, L. A Complex Standard for Protein Identification, Designed by Evolution. *J. Proteome Res.*

2012, *11* (10), 5065–5071.

(23) Rosenberger, G.; Navarro, P.; Gillet, L.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Aebersold, R.; Diseases, M. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.

(24) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A Draft Map of the Human Proteome A. *Nature* **2014**, *509* (7502), 575–581.

(25) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; et al. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, 1–16.

(26) Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Frewen, B.; Baker, T. A.; Brusniak, M.; Paulse, C.; Lefebvre, B.; Kuhlmann, F.; et al. HHS Public Access. *Nat. Biotechnol.* **2013**, *30* (10), 918–920.

(27) Vizca, J. A.; Csordas, A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.; Wang, R.; et al. 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res.* **2016**, *44* (November 2015), 447–456.

(28) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(29) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17* (7), 2552–2555.

(30) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.;

Martens, L.; Barsnes, H. PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets: To the Editor. *Nat. Biotechnol.* **2015**, *33* (1), 22–24.

(31) The Uniprot Consortium. UniProt : A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (November 2018), 506–515.