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

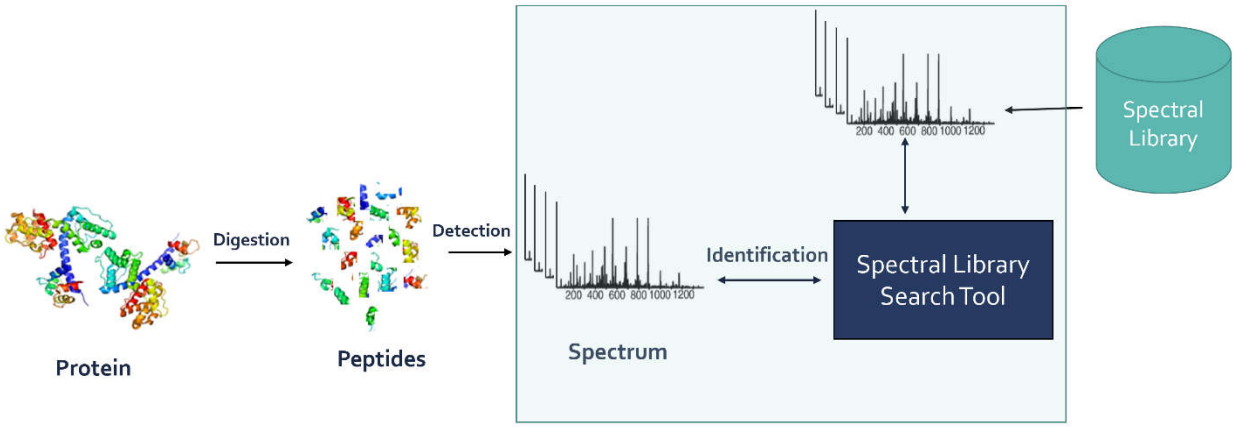
[3] Cancer Research Institute Ghent, Ghent University, 9000 Ghent, Belgium

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 that relies on a probabilistic scoring function, and that includes decoy spectra generation for

result validation. We have benchmarked COSS on three different spectral libraries and compared the results with established spectral search and sequence database search tools. Our comparison showed that COSS identifies more peptides, and is faster than other tools. 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 been developed over the past years, with SpectraST[10], the National Institute of Standards and Technology (NIST) MS-Search[11] tool and X!Hunter[12] as notable examples . Each of these tools, however, comes with some limitations, such as 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. 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 an overall improved identification rate. At the same time, COSS requires drastically lower computation time 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[13], mzML[14], MS2 and dta input formats through the mzIdentML[15] library, while support for the MSP and MGF formats is included through an in-house implementation. The compomics-utilities[16] 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 the MSROBIN scoring function, which is based on the probabilistic scoring function of Yilmaz et al.[17], itself a derivative of the Andromeda scoring function[18]. 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 scoring function itself 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 set of matched spectra and *Y* is set of filtered spectra. The final score is then computed as:

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

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[19]. 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 one of two techniques. The first technique shifts all peaks by a constant m/z value of 20 Da. The second technique shifts both the mass-over-charge (m/z) and intensity of each peak by random values. In the second technique, the new m/z value is contained between the minimum and maximum m/z range of the original spectrum, while the new intensity is set larger than zero and less than the maximum intensity of the original range. Both decoy generation techniques yield highly similar results (Supplementary Figure S-1). We have used the second method to generate decoys for benchmarking COSS.

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.[19].

Benchmarking datasets and spectral libraries

We obtained raw data files from eleven runs from the Human Proteome Map[20] (ProteomeXchange[21] ID PXD000561) as benchmarking data sets (Table 1). All data sets were originally generated using the Thermo Scientific™ LTQ Orbitrap Elite instrument, and each data set represents the run with the largest number of spectra for that tissue. These eleven raw files were converted to Mascot Generic Format (mgf) format using the *msconvert* tool (ProteoWizard[22]), with the peak picking algorithm activated.

Benchmarking was performed using three distinct spectral libraries (Table 2), obtained from NIST, PRIDE[23] and MassIVE[22].

Table 1. Benchmarking data sets used to test COSS. For each of the eleven randomly chosen tissues from the Human Proteome Map (ProteomeXchange ID PXD000561), the largest raw file was selected.

Raw file	Number of spectra
Adult_Adrenalgland_Gel_Elite_49_f22.raw	14954
Adult_Bcells_Gel_Elite_76_f01.raw	7496
Adult_Colon_bRP_Elite_50_f24.raw	1811
Adult_Gallbladder_Gel_Elite_52_f24.raw	6146
Adult_Heart_Gel_Elite_54_f05.raw	12216
Adult_Kidney_Gel_Elite_55_f01.raw	5267
Adult_Liver_Gel_Elite_83_f23.raw	4615
Adult_Ovary_Gel_Elite_58_f02.raw	3885
Adult_Pancreas_Gel_Elite_60_f02.raw	2156
Adult_Retina_bRP_Elite_64_f15.raw	1770
Adult_Spinalcord_Gel_Elite_67_f01.raw	4105

Table 2. Spectral libraries used to benchmark COSS.

Spectral Library	Total number of spectra	URL
MassIVE[22] (Human HCD Spectral Library)	2,154,269	http://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp
NIST (human HCD library)	1,127,970	https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503
PRIDE Cluster[23] (Human)	789,745	https://www.ebi.ac.uk/pride/cluster/#/libraries

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 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).

Sequence database searches have been performed using a combination of four tools: OMSSA[24] (version 2.1.9), X!Tandem[25] (version 2015.12.15.2), MS-GF+[26] (version v2018.04.09), and Andromeda[18] (version 1.5.3.4) through PeptideShaker[27] (version 1.16.32) and SearchGUI[28] (version 3.3.6). The search database is constructed from the human proteome (UP000005640) as obtained from UniProt[29] (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 random m/z and intensity values (see section on FDR estimation) and these decoys were appended to the original spectral library. 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-2) 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.

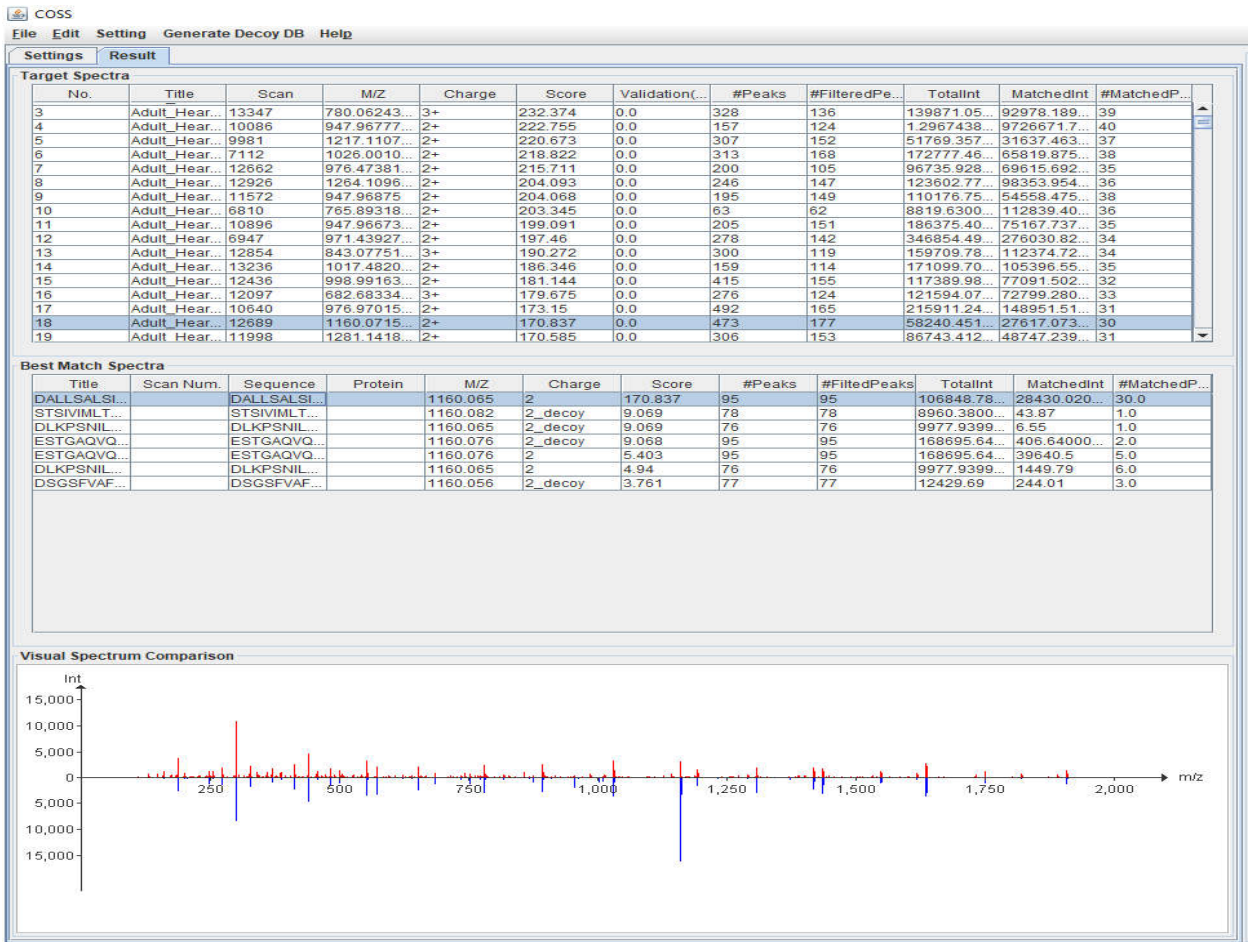


Figure 1. Search result interface of COSS: the upper table lists the experimental spectra while the lower table lists the top 10 matched spectra for the selected experimental spectrum. An interactive spectrum comparison view is presented at the bottom with the selected experimental spectrum (red) mirrored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

Search result comparison

As shown in Figure 2, COSS overall outperforms both SpectraST and the combination of sequence database search engines in terms of identification rate for the NIST and MassIVE spectral libraries. The overall identification rates of COSS are 44.51%, 17.88% and 55.86% against NIST, PRIDE, and MassIVE libraries, respectively, while SpectraST's are 31.257% and 17.69% against NIST and PRIDE. Note that SpectraST's performance against the MassIVE database could not be assessed as SpectraST could not handle a library of this size on our server with 28GB of RAM. The combined sequence database search algorithms identified 29.23% of the spectra when searching the human proteome database. The identification rate of COSS on the PRIDE Cluster spectral library, however, is lower than that of the combined sequence database searching algorithms, most likely due to the incomplete coverage of this library (consisting of 189,400 unique peptides, compared to the 1,500,000 unique peptides derived from the *in silico* tryptic digest of the human proteome sequence database). Overall, the identifications of the three approaches show a good overlap in terms of the identified spectra and peptides (Supplementary Figure S-3).

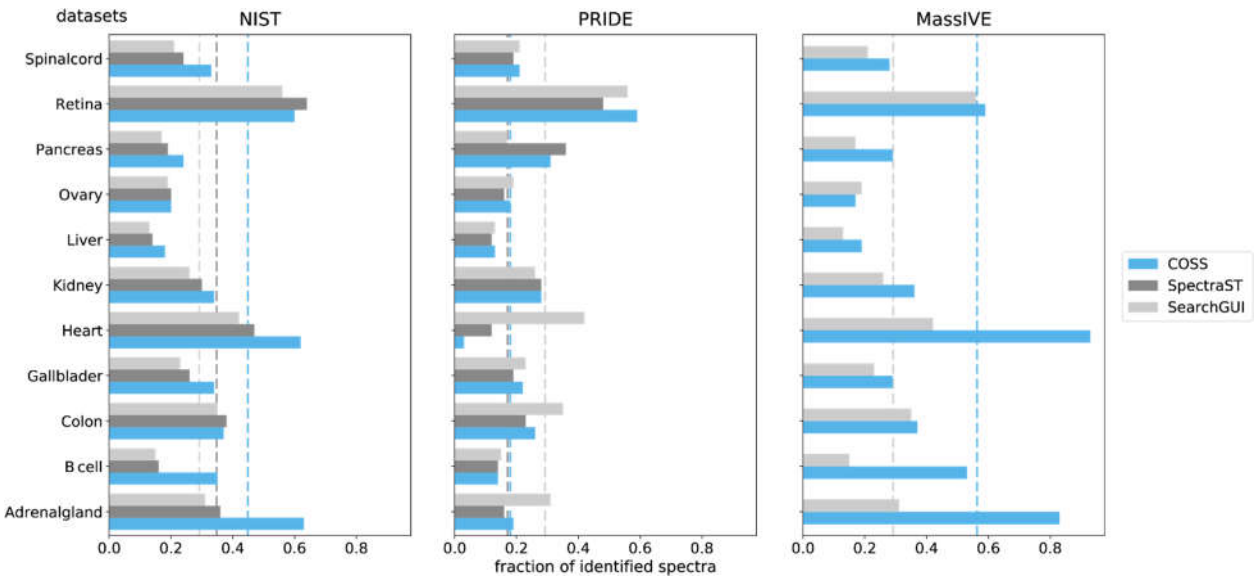


Figure 2. COSS performance evaluation against SpectraST and sequence database searching in terms of identification rate. Shown here is the identification rate against the NIST, PRIDE Cluster and MassIVE spectral libraries for COSS and SpectraST, and against the human proteome sequence database for the combination of sequence database search engines. Due to the excessive memory requirements, SpectraST could not run the MassIVE spectral library on our server with 28GB of RAM.

Running time comparison

To evaluate the computational efficiency of the algorithm, we ran COSS and SpectraST 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 3. While the size of the query dataset and the spectral library both clearly influence the executing time, we found that COSS drastically outperforms SpectraST in all cases. Here again, there is no data for SpectraST for the MassIVE library due to the inability to run SpectraST on this library on our server.

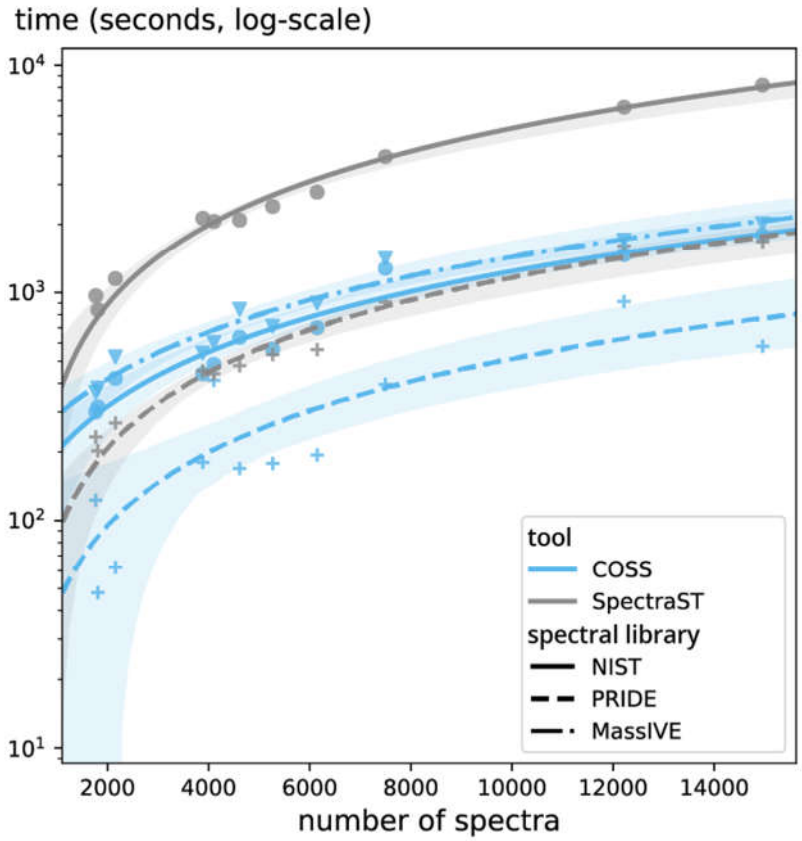


Figure 3. Execution times of COSS and SpectraST. While execution time increases with experimental data set and spectral library sizes, COSS clearly outperforms SpectraST in every case. Even for large data sets and spectral libraries, results for COSS are retrieved in well under half an hour.

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 have already been developed but come with important limitations: long search times, an inability to handle very large spectral libraries, and limited input file format support. Here we present COSS, a user-friendly spectral library search tool that is fast, that can handle large datasets, and that supports the most commonly used MS/MS data formats. 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 and a combination of state-of-the-art sequence database search algorithms in terms of identification performance, and found that COSS offers an overall improved identification rate, while also drastically outperforming SpectraST in running time. These properties make COSS far more suitable for large-scale analyses against expansive spectral libraries, such as those that aim to cover the entire human proteome.

AVAILABILITY

The software and its source code can be freely downloaded from https://github.com/compomics/COSS and is licensed under the 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) [grant number (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 all the CompOmics group members for the ideas, discussions and support.

REFERENCES

(1) Ingvar Eidhammer Kristian Flikka Lennart Martens Svein-Ole Mikalsen. 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.; Houel, S.; Ahn, N. G.; Old, W. 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. **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) 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.

(13) 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. **2004**, *22* (11), 1459–1466.

(14) 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 *. **2011**, 1–7.

(15) 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.

(16) 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.

(17) 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.

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

(19) 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.

(20) 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.

(21) 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. **2014**, *32* (3), 223–226.

(22) 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. **2013**, *30* (10), 918–920.

(23) Griss, J.; Foster, J. M.; Hermjakob, H.; Vizcaíno, J. A. Europe PMC Funders Group Europe PMC Funders Author Manuscripts PRIDE Cluster : Building the Consensus of Proteomics Data. **2013**, *10* (2), 95–96.

(24) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **2004**, *3* (5), 958–964.

(25) Craig, R.; Beavis, R. C. TANDEM : Matching Proteins with Tandem Mass Spectra. **2004**, *20* (9), 1466–1467.

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

(27) 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.

(28) 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.

(29) Consortium, T. U. UniProt : A Worldwide Hub of Protein Knowledge. **2019**, *47* (November 2018), 506–515.