Published in final edited form as:

Proteomics. 2008 June ; 8(12): 2376–2378. doi:10.1002/pmic.200701126.

OMSSAGUI: An open-source user interface component to configure and run the OMSSA search engine

Ravi Tharakan^{1,*}, Lennart Martens^{2,*}, Jennifer E. Van Eyk¹, and David R. Graham^{1,3}

¹ Department of Medicine, Bayview NHLBI Proteomics Center, Johns Hopkins School of Medicine Bayview Campus, Baltimore, MD, USA

² EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

³ Retrovirus Laboratory, Department of Molecular and Comparative Pathobiology, Johns Hopkins School of Medicine Bayview Campus, Baltimore, MD, USA

Abstract

We here present a user-friendly and extremely lightweight tool that can serve as a stand-alone front-end for the Open MS Search Algorithm (OMSSA) search engine, or that can directly be used as part of an informatics processing pipeline for MS driven proteomics. The OMSSA graphical user interface (OMSSAGUI) tool is written in Java, and is supported on Windows, Linux, and OSX platforms. It is an open source under the Apache 2 license and can be downloaded from <http://code.google.com/p/mass-spec-gui/>.

Keywords

Bioinformatics; Protein identification

MS/MS is the primary method for the identification of proteins in high-throughput proteomics [1]. In this method, proteins or complex mixtures of proteins are proteolytically digested with an enzyme such as trypsin resulting in a more complex mixture of peptides. The peptides are then separated by 1- or 2-D LC and ionized online by ESI or offline by MALDI, for the analysis by MS. MS/MS is then usually carried out by first recording the mass over charge (m/z) values of the ionized peptides and subsequently selecting one peptide ion at a time for the fragmentation analysis. Peptides can be fragmented by a variety of methods, including PSD, CID, electron capture dissociation (ECD), and electron transfer dissociation (ETD). The m/z values of the resulting fragment ions are recorded in a fragmentation spectrum. A specialized algorithm is then typically used to search these fragmentation spectra against theoretical spectra derived from the predicted fragment ions of peptides obtained after an *in silico* digest of a protein sequence database.

In order to differentiate correct spectrum matches from incorrect ones, many different algorithms are used for this step, such as MASCOT [2,3], SEQUEST [4], SpectrumMill (Agilent Technologies), open MS search algorithm (OMSSA) [5], and X!Tandem [6]. These algorithms typically assign one or more scores to these matches. Because these scoring

Correspondence: Dr. David Graham, Johns Hopkins School of Medicine, 733 N. Broadway, BRB 835, Baltimore, MD, USA, dgraham@jhmi.edu, Fax: +1-410-955-9823.

*These authors contributed equally to this work.

The authors have declared no conflict of interest.

functions vary widely between algorithms, both in the parameters which they consider and the overall methodology used to estimate false positive rates, recent work by Balgley *et al.* [7] has relied upon a target-decoy search strategy [8] to directly compare different algorithms. In this particular study, OMSSA performed best by almost all measures, and especially in terms of the number of fragmentation spectra (MS² events) directly assigned *versus* false positive rate [7]. This increased sensitivity of the OMSSA algorithm makes it an ideal candidate for a “second-pass” search approach, in which a more specific search algorithm such as MASCOT or SEQUEST performs an initial search of the data, and OMSSA is used in a second pass to increase coverage.

One obstacle to the quick adoption of OMSSA in this way is the tendency of proteomics laboratories to use home-grown and often complex chains of software tools in the analysis of data. For example, one software tool may be employed to extract a list of peaks from the original spectra, a second algorithm performs the database searching, another tool is then used to parse and display the search results, and a final step could consist of a program that exports the data to an archive database. If a research group has additional aims in analyzing the data besides the straightforward identification of proteins, such as polymorphism discovery, more tools may be necessary. Because of the labor required to run these long data “pipelines,” custom software is often written specifically to automate data management, for example, MASPECTRAS [9], MassSieve [10], ISB TPP [11], and Scaffold (Proteome Software). Finally, a data management pipeline will always be dependent on how the proteomic experiment was carried out, subject to variables such as digestion enzyme, ionization type, the instrument used, and so on. For these reasons it is often difficult to quickly integrate a new search program like OMSSA into an already existing analysis pipeline, especially when the user needs to be provided with a convenient and intuitive interface to these sophisticated software tools.

To solve this problem for OMSSA, we here present OMSSA graphical user interface (OMSSAGUI), a simple, robust, and a very lightweight tool designed for easy integration of OMSSA into existing pipelines. It features an easy to use interface, allows users to save and load search settings, and provides the ability to easily search against a customized sequence database. Written entirely in Java, OMSSAGUI is completely portable, and is fully supported on Windows, Mac OSX, and Linux operating systems. The software, installation instructions, user manual, and source code, can be obtained from <http://code.google.com/p/mass-spec-gui/>.

We expect that the ability to save search settings for later use will allow automated use of OMSSA in a pipeline context while maintaining consistency in parameters throughout the data analysis process for any given type of MS/MS experiment. For example, if a laboratory uses a given MS instrument with a certain mass tolerance, search settings could be simply be automatically written to the settings file for the GUI. If the output of the search is then to be uploaded to a databasing archive, such as PRIDE, information for settings could be read automatically from the settings file for that upload. This can help to minimize the occurrence of the error of having inconsistent information about an experiment throughout the pipeline, both upstream and downstream of the OMSSA search run. The ability to search on a customized database will also allow OMSSA to be integrated into pipelines in which the search database is to be limited in some way, for example, to screen for specific proteins, if modifications are to be coded directly into a database, or if OMSSA is used as a second-pass algorithm. Finally, the interface, which is simple and intuitive and can be used on its own, will enable individual users to quickly begin using OMSSA, even outside the scope of processing pipelines.

The OMSSAGUI interface, shown in Fig. 1, utilizes text boxes for selecting input file locations (spectra, sequence databases) and output file locations (the OMSSA results file). Fixed and variable modifications are selected from lists, which are populated directly from the OMSSA

configuration files upon start-up. A new interface feature we introduce in these lists conveniently displays the number of modifications selected. The actual search parameters can be customized in the next part of the interface. The results are output in the XML format created by the original OMSSA developers, omx, and may be viewed in OMSSA Browser, which is free and available for download, or can be integrated into pipeline software such as MASPECTRAS, MassSieve, or Scaffold. Finally, the NCBI toolkit also contains libraries for developers to parse the OMSSA results files directly.

We have presented a lightweight yet powerful and extremely user-friendly tool for performing OMSSA searches, which is aimed at lowering the threshold for implementing OMSSA, and will allow easy integration of this search engine in pipeline projects. OMSSAGUI is open source under the Apache 2.0 license, and is reliable, production-grade software. We certainly hope that, together with the increasing availability of similar, modular tools [12,13], the fast and flexible implementation of sophisticated multistep informatics backbones for MS proteomics data processing should come into the reach of every laboratory in the field.

Acknowledgments

L. M. would like to thank Henning Hermjakob and Rolf Apweiler for their support. This work was supported by the Dean's Office of John Hopkins University (startup funds to D. G.) and by the National Heart Lung Blood Institute Proteomic Initiative (contract NO-HV-28120).

Abbreviations

OMSSA	open MS search algorithm
OMSSAGUI	open MS search algorithm graphical user interface

References

1. Cox J, Mann M. Is proteomics the new genomics? *Cell* 2007;130:395–398. [PubMed: 17693247]
2. Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993;3:327–332. [PubMed: 15335725]
3. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
4. Eng JK, McCormack AL, Yates JR. III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Am Soc Mass Spectrom* 1994;5:976–989.
5. Geer LY, Markey SP, Kowalak JA, Wagner L, et al. Open mass spectrometry search algorithm. *J Proteome Res* 2004;3:958–964. [PubMed: 15473683]
6. Craig R, Beavis RC. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467. [PubMed: 14976030]
7. Balgley BM, Laudeman T, Yang L, Song T, Lee CS. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics* 2007;6:1599–1608. [PubMed: 17533222]
8. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4:207–214. [PubMed: 17327847]
9. Hartler J, Thallinger GG, Stocker G, Sturn A, et al. MASPECTRAS: A platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics* 2007;8:197. [PubMed: 17567892]
10. Slotta DJ, McFarland MA, Makusky AJ, Markey SP. Mass Sieve: A new tool for mass spectrometry-based proteomics. *J Biomol Tech* 2007;18:7.
11. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 2005;1:0017. [PubMed: 16729052]

12. Helsens K, Martens L, Vandekerckhove J, Gevaert K. MascotDatfile: An open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* 2007;7:364–366. [PubMed: 17203510]
13. Flikka K, Meukens J, Helsens K, Vandekerckhove J, et al. Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *Proteomics* 2007;7:3245–3258. [PubMed: 17708593]

Figure 1.
Screenshot of the OMSSAGUI component.