SearchGUI: a highly adaptable common interface for proteomics search and *de novo* engines

Harald Barsnes[1,2,*] *and Marc Vaudel*[3,4,*]

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway
[2] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway
[3] KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway
[4] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[*] To whom correspondence should be addressed.

E-mail: marc.vaudel@uib.no

Phone: +47 55 58 66 65

Abstract

Mass spectrometry-based proteomics has become the standard approach for identifying and quantifying proteins. A vital step consists of analyzing experimentally generated mass spectra to identify the underlying peptide sequences, for later mapping to the originating proteins. We here present the latest developments to SearchGUI, a common open source interface for the most frequently used freely available proteomics search and *de novo* engines, which has evolved into a central component in numerous bioinformatics workflows.

Keywords: Bioinformatics, protein identification, search engines, de novo algorithms

Introduction

Peptide identification is a cornerstone in most modern proteomics data analysis pipelines, either by mapping the experimentally generated mass spectra to potential peptide sequences from a protein sequence database(1) or against a set of known spectra from a spectral database(2), or by analyzing the spectra directly using *de novo* sequencing or tagging(3). Numerous algorithms have been developed to help with this processing, both commercial and open source. Using these algorithms may however require advanced computational skills and/or paying for, often closed source, commercial software packages.

Here we present a completely redesigned version of SearchGUI, a common open source interface for the most frequently used freely available proteomics search and *de novo* engines, which makes it straightforward to carry out peptide identification, either by using an intuitive graphical user interface or *via* the command line, *e.g.* for running in the cloud or on a cluster(4). Since the first release back in 2010(5), SearchGUI has been under continuous development, while at the same time gaining a growing userbase.

Accurate usage statistics are difficult to obtain given that SearchGUI downloads are automated without user tracking, and bioinformatics tools are often run in protected environments that prevent software from accessing the Internet. However, based on the users who enabled automated updates, over 60,000 sessions were started over the past year, showing good adoption by the community.

This manuscript highlights the most important recent improvements and demonstrates the increasingly wider array of use cases in which SearchGUI can be employed.

Results

This section provides an overview of the major new features added since the initial release of SearchGUI. It is not exhaustive in terms of new features and code changes. Readers wanting more details are referred to the release notes section at the SearchGUI website (compomics.github.io/projects/searchgui).

Originally, SearchGUI only supported two search engines: OMSSA (6) and X! Tandem (7). In the current version, a total of eight search engines are supported, namely OMSSA, X! Tandem, MyriMatch (8), MS Amanda (9), MS-GF+ (10), Comet (11), Tide (12), and Andromeda (13). All can be set up and run from the same graphical user interface or command line, ensuring that a single set of parameters is translated to the specific settings required by the individual search engines. This makes it straightforward for any user to run multiple search engines with the same input files and search settings, reducing the need to set up and configure each search engine separately. The advanced settings for each search engine are also easily available for fine-tuning by expert users.

In addition to the eight proteomics search engines, SearchGUI now also supports two common *de novo* engines in DirecTag (14) and Novor (15). The first being a tag-based algorithm, while the second provides complete peptide sequences. Being able to run both standard search and *de novo* sequence algorithms on the same data sets gives more flexibility for workflow design and will hopefully increase the number of total spectra that can be identified. Note that the same parameters used for the search engines are also used for *de novo*, albeit specific *de novo* settings can also be edited.

The standard spectral input format in peptide identification is the Mascot Generic Format (matrixscience.com/help/data_file_help.html#GEN), known as mgf. However, SearchGUI now supports preprocessing and conversion of raw data formats through the embedded use of msconvert from the ProteoWizard (16) library. All the user has to do is refer to the location of the ProteoWizard installation and the conversion will be done automatically. The advanced msconvert parameters can also be

accessed directly from SearchGUI, *e.g.* to tune signal processing as part of the conversion. This addition removes the need for a separate step to convert raw files, thus simplifying the data processing pipeline.

In addition, the output of all the search and *de novo* engines can now be forwarded seamlessly to PeptideShaker (17) for joint analysis. All the PeptideShaker processing and filtering parameters can be accessed directly from SearchGUI when setting up the workflow, removing the need for an extra manual step in the pipeline to carry out the protein identification. Being able to run both SearchGUI and PeptideShaker from the same interface greatly simplifies the task of protein identification, and at the same time enables users to easily compare the output from individual algorithms, increasing the overall number of peptide identifications (18). SearchGUI has also been incorporated into the *Reshake* feature of PeptideShaker, enabling the reprocessing of public data sets in the PRIDE repository (19) using all supported algorithms.

As illustrated in Figure 1, to accommodate all the new features, the SearchGUI graphical user interface has been entirely redesigned. As a part of this process the user-friendliness has also been dramatically improved, and more of the advanced settings made available to users, which could previously only be altered from the command line. This includes making it possible to alter all of the more than 200 advanced search and *de novo* parameters directly from the graphical user interface, making it as customizable as the command line. Better support for setting up user-defined post-translational modifications has also been added, plus a simpler way of interacting with the numerous settings required to control the postprocessing identification pipeline.

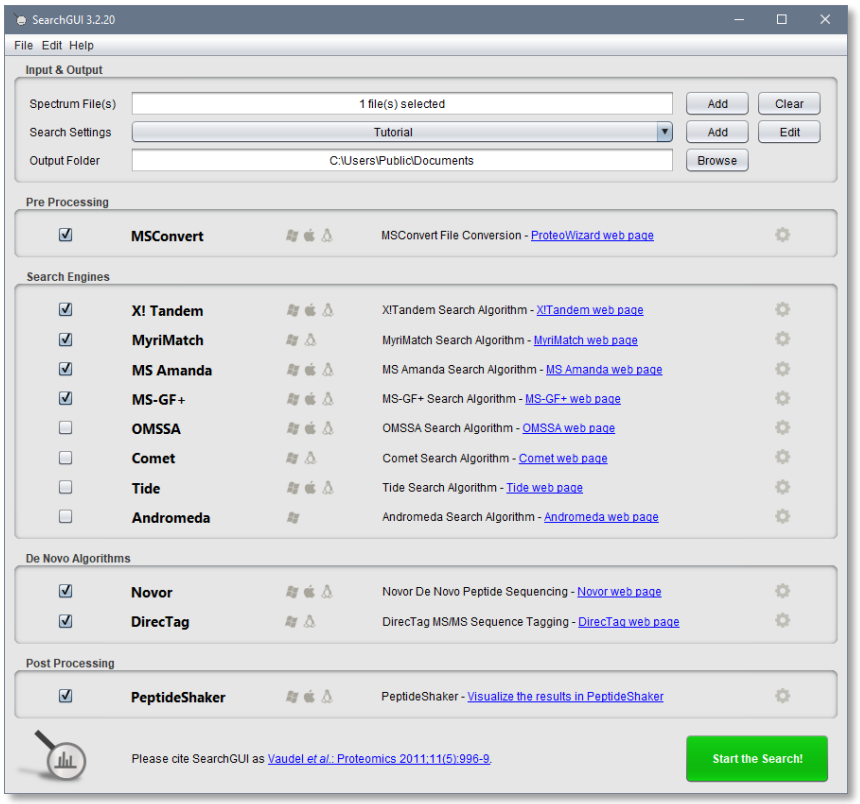

Figure 1: *The new interface of SearchGUI makes it possible to set up and run eight proteomics search engines and two de novo engines, including preprocessing of raw data* via *ProteoWizard and postprocessing of the identifications* via *PeptideShaker, all from the same interface. Tool-specific settings are available by clicking the cogwheels.*

At the same time, the command line support for SearchGUI has been significantly extended, making it possible to run all ten engines *via* a single command line, with access to all the individual advanced parameters. The file management has been simplified and control of intermediate files expanded. These changes have allowed integration of SearchGUI in Galaxy (20), Bioconda (bioconda.github.io) and BioContainers (21), see compomics.github.io/projects/searchgui/wiki/searchcli for more details.

Together with PeptideShaker, SearchGUI has been used at a long list of international courses and workshops. This has resulted in extensive tutorial material detailing both how to use the tools, but equally important how to understand and optimize the long list of parameters required to set up a protein identification

pipeline (22). All of the tutorial material is freely available at compomics.com/bioinformatics-for-proteomics.

Conclusions

The highly adaptable nature of SearchGUI is its strongest feature, as indicated by the fact that it is already being used in a wide range of contexts and across various technical solutions, as illustrated in Figure 2. The development of SearchGUI is still an ongoing process, with future plans including the addition of spectral library search engines and algorithms optimized for open modification searches. Through long-term development and user support, we believe that it will become accessible to even more users, hopefully resulting in better and more user-friendly protein identification pipelines.

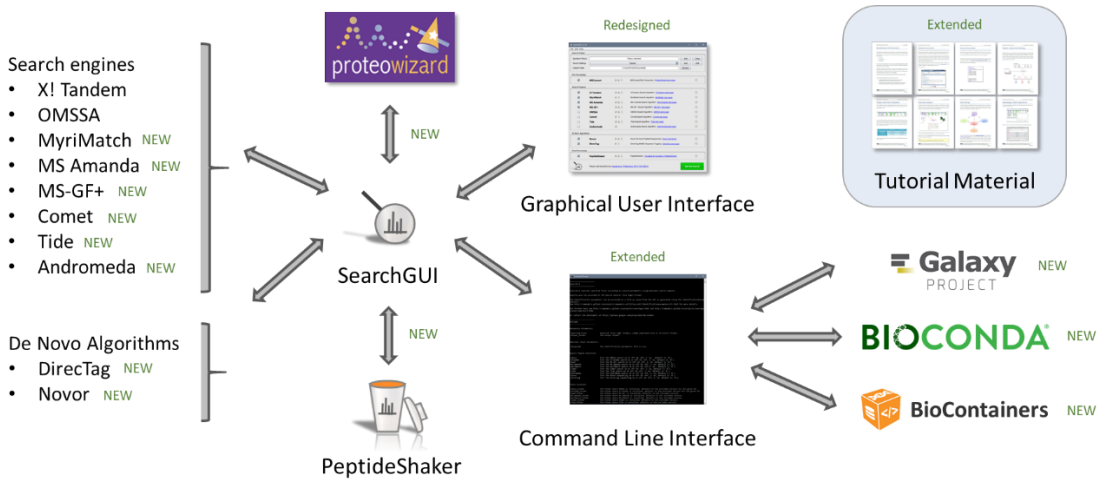

Figure 2: *Overview of the SearchGUI environment, showcasing its adaptability to a wide array of use cases, and highlighting all the additions and extensions since the initial release.*

SearchGUI is developed in Java and is open source under the very permissive Apache2 license. It is supported on Windows, Mac OS X and Linux (with individual restrictions for third-party software). Cross-platform executable binaries, source code, documentation and support are available at compomics.github.io/projects/searchgui.

Acknowledgments

H.B. is supported by the Research Council of Norway and the Bergen Research Foundation. The authors are grateful to the developers of all the search and *de novo* engines for their cooperation and support in making their software available for use in SearchGUI. Finally, the authors would like to thank all the SearchGUI users and developers of workflows and bioinformatic environments for their thorough testing and for coming up with valuable suggestions for improvements.

Author contributions

H.B. and M.V. did all of the programming and together wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1. Verheggen, K.; Raeder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M., Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev* **2017**. doi: 10.1002/mas.21543. [Epub ahead of print]
2. Shao, W.; Lam, H., Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass Spectrom Rev* **2017,** 36, (5), 634-648.
3. Medzihradszky, K. F.; Chalkley, R. J., Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrom Rev* **2015,** 34, (1), 43-63.
4. Verheggen, K.; Barsnes, H.; Martens, L., Distributed computing and data storage in proteomics: many hands make light work, and a stronger memory. *Proteomics* **2014,** 14, (4-5), 367-77.
5. Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L., SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011,** 11, (5), 996-9.
6. 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), 958-64.
7. Fenyo, D.; Beavis, R. C., A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* **2003,** 75, (4), 768-74.
8. Tabb, D. L.; Fernando, C. G.; Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **2007,** 6, (2), 654-61.
9. Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K., MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *J Proteome Res* **2014,** 13, (8), 3679-84.
10. Kim, S.; Pevzner, P. A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **2014,** 5, 5277.
11. Eng, J. K.; Jahan, T. A.; Hoopmann, M. R., Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013,** 13, (1), 22-4.
12. Diament, B. J.; Noble, W. S., Faster SEQUEST searching for peptide identification from tandem mass spectra. *J Proteome Res* **2011,** 10, (9), 3871-9.
13. 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), 1794-805.
14. Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C., DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* **2008,** 7, (9), 3838-46.
15. Ma, B., Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* **2015,** 26, (11), 1885-94.
16. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly,

K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** 30, (10), 918-20.
17. 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. *Nat Biotech* **2015,** 33, (1), 22-24.
18. Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W., Combining results of multiple search engines in proteomics. *Mol Cell Proteomics* **2013,** 12, (9), 2383-93.
19. Vizcaino, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q. W.; Wang, R.; Hermjakob, H., 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **2016,** 44, (22), 11033.
20. Chambers, M. C.; Jagtap, P. D.; Johnson, J. E.; McGowan, T.; Kumar, P.; Onsongo, G.; Guerrero, C. R.; Barsnes, H.; Vaudel, M.; Martens, L.; Gruning, B.; Cooke, I. R.; Heydarian, M.; Reddy, K. L.; Griffin, T. J., An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res* **2017,** 77, (21), e43-e46.
21. da Veiga Leprevost, F.; Gruning, B. A.; Alves Aflitos, S.; Rost, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; Bai, M.; Jimenez, R. C.; Sachsenberg, T.; Pfeuffer, J.; Vera Alvarez, R.; Griss, J.; Nesvizhskii, A. I.; Perez-Riverol, Y., BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **2017,** 33, (16), 2580-2582.
22. Vaudel, M.; Venne, A. S.; Berven, F. S.; Zahedi, R. P.; Martens, L.; Barsnes, H., Shedding light on black boxes in protein identification. *Proteomics* **2014,** 14, (9), 1001-5.