Journal of proteome research

Technical Note

pubs.acs.org/jpr

Terms of Use

DeNovoGUI: An Open Source Graphical User Interface for *de Novo* Sequencing of Tandem Mass Spectra

Thilo Muth,[†] Lisa Weilnböck,[‡] Erdmann Rapp,[†] Christian G. Huber,[‡] Lennart Martens,[§,∥] Marc Vaudel,*[,⊥] and Harald Barsnes[⊥]

[†]Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany

[‡]Department of Molecular Biology, University of Salzburg, 5020 Salzburg, Austria

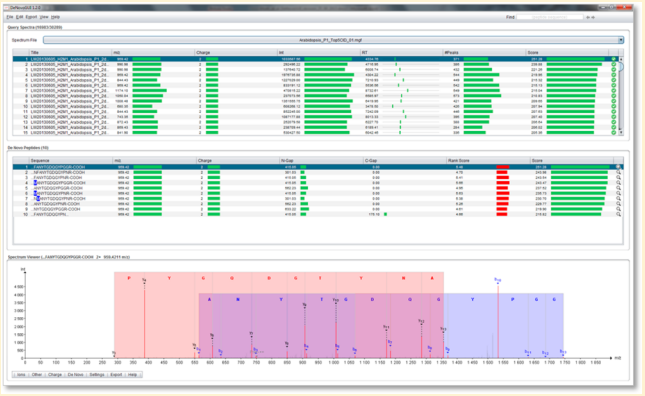
[§]Department of Biochemistry, Ghent University, 9000 Ghent, Belgium

[∥]Department of Medical Protein Research, VIB, 9052 Ghent, Belgium

[⊥]Proteomics Unit, Department of Biomedicine, University of Bergen, Jones Liesvei 91, N-5009 Bergen, Norway

Supporting Information

ABSTRACT: *De novo* sequencing is a popular technique in proteomics for identifying peptides from tandem mass spectra without having to rely on a protein sequence database. Despite the strong potential of *de novo* sequencing algorithms, their adoption threshold remains quite high. We here present a user-friendly and lightweight graphical user interface called DeNovoGUI for running parallelized versions of the freely available *de novo* sequencing software PepNovo+, greatly simplifying the use of *de novo* sequencing in proteomics. Our platform-independent software is freely available under the permissible Apache2 open source license. Source code, binaries, and additional documentation are available at http://denovogui.googlecode.com.



KEYWORDS: *bioinformatics, de novo, mass spectrometry, PepNovo+, peptide identification*

Mass spectrometry (MS)-based proteomics is an efficient high-throughput method for the analysis of peptides and proteins.[1,2] However, in a typical tandem mass spectrometry (MS/MS) experiment, a high proportion of the mass spectra remain unidentified when matched against *in silico*-generated spectra, derived from peptides obtained through *in silico* proteolytic digestion of known protein sequences.[3] Some of these unidentified spectra derive from contaminants and low-quality spectra, but the rest are likely to contain unexpected peptides.[4] One obstacle for the successful identification of such peptides is the fact that protein sequence databases are incomplete, as many organisms have not yet been sequenced, an issue that is particularly strongly felt in challenging fields such as metaproteomics[5] or plant proteomics.[6] Another common issue is the presence of unknown or unexpected modifications on the peptide precursors.[4] *De novo* sequencing constitutes a powerful technique for overcoming such issues and successfully assigning high-quality unidentified spectra to peptides. Moreover, *de novo*-derived peptide sequences can be used for the validation of insignificant database search results, for instance, proteins backed merely by a single peptide identification, so-called "one-hit wonders".[7]

Several *de novo* algorithms have been described and evaluated in the literature,[8,9] including the commercial PEAKS[10] software suite. PepNovo+,[11] on the other hand, is a powerful, freely available software tool. However, as with most open source *de novo* algorithms, it comes with several shortcomings. (i) It is distributed only with a command line interface, thus requiring advanced computational skills to operate. (ii) Modifications need to be configured manually for every search and are not based on the standardized PSI-MOD[12] controlled vocabulary. (iii) The search is not parallelized when multiple cores are available. (iv) The output of the algorithm is a text file containing only the derived sequences and their scores, thus omitting additional useful information such as fragment ions and spectrum annotation. Because of these issues, user validation of the results is cumbersome, and standardized dissemination of results is quite difficult.

Here, we describe an intuitive, end-user-oriented front end to the PepNovo+ algorithm called DeNovoGUI, which aims to solve the aforementioned problems. Similar to the SearchGUI[13] software for the OMSSA[14] and X!Tandem[15] database search algorithms, DeNovoGUI provides a self-contained and easily adopted solution for convenient and efficient *de novo* sequencing using the PepNovo+ algorithm. The processing of

Received: August 9, 2013
Published: December 2, 2013

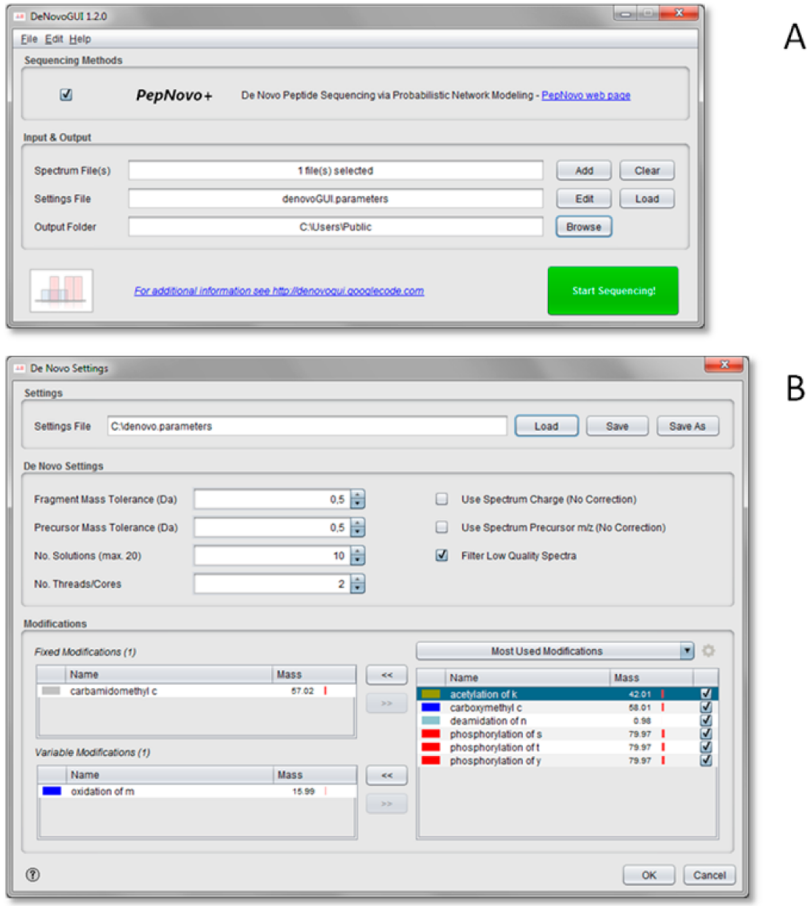


Figure 1. (A) Main DeNovoGUI interface that allows the user to input the spectrum files, the settings, and the output folder for the results. (B) *De novo* sequencing settings dialogue that allows the user to specify the fragment ion and precursor mass tolerances, and the fixed and variable post-translational modifications. Additional settings for fine-tuning the PepNovo+ algorithm can also be configured.

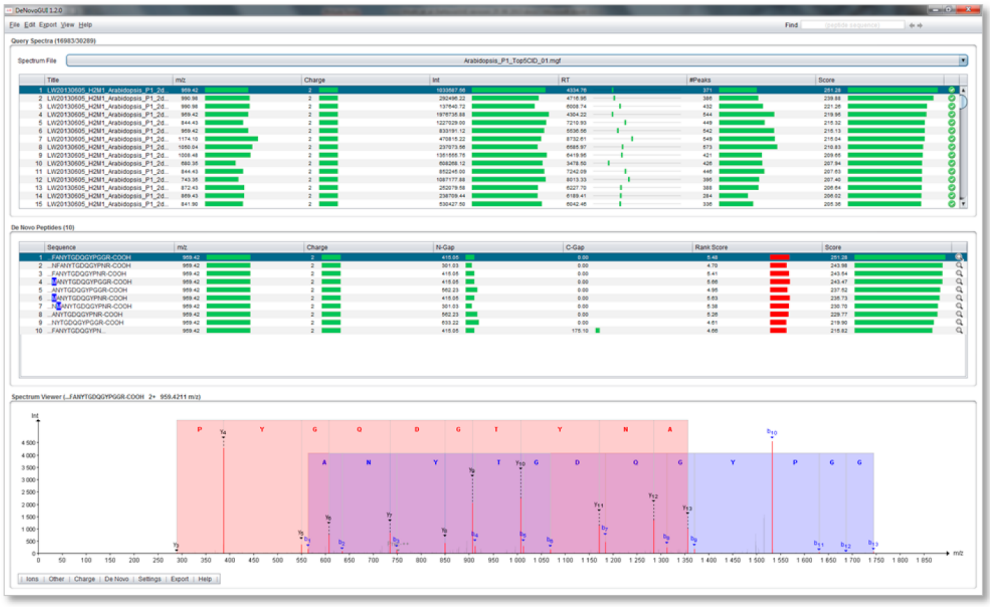


Figure 2. DeNovoGUI *de novo* results viewer that shows the currently selected *de novo* peptide solution and its corresponding fragment ion annotations on the selected spectrum. The 'Query Spectra' section at the top allows the user to browse through the input spectra, while the 'De Novo Peptides' section below provides sequence and scoring information for all peptide solutions for the currently selected input spectrum.

a large amount of spectra has been accelerated by automated and completely transparent parallelization across multiple cores, a crucial feature for modern computers that typically come equipped with two to eight (hyperthreaded) cores.

dx.doi.org/10.1021/pr4008078 | *J. Proteome Res.* 2014, 13, 1143−1146

DeNovoGUI can be installed with minimal effort by downloading the latest release from the tool Web site (http://denovogui.googlecode.com), subsequently unzipping the downloaded file, and then double clicking the DeNovoGUI jar file. To start the *de novo* procedure, the user has to provide only the spectrum files to analyze (in the standard mgf format), the settings to use, and the desired output folder in which to store the *de novo* results (Figure 1A). The settings include the fragment and precursor ion mass tolerances, as well as the fixed and variable post-translational modifications to consider. Furthermore, additional settings for fine-tuning the PepNovo + algorithm can be specified, including the number of *de novo* solutions to provide for each spectrum. Figure 1B provides a complete overview of the available settings. Importantly, the handling of modifications is greatly simplified by the graphical user interface: user-defined modifications can easily be created from the Edit menu in the mainframe. Note that DeNovoGUI allows all settings to be saved for later reuse or batch entry.

As soon as the settings and input files have been provided, the *de novo* sequencing can be initiated by clicking the "Start Sequencing!" button in the main DeNovoGUI interface. While the PepNovo+ algorithm is running, the user is continuously informed about the status of the sequencing and a progress bar is displayed to indicate overall progress. When the process is complete, the *de novo* sequencing results are stored in the provided output folder in a simple text-based format, and the detailed results can be visualized in the DeNovoGUI interface (see Figure 2). At the top, the user can browse through all the input spectra in the 'Query Spectra' table, and through the *de novo* peptide matches for the selected spectrum in the 'De Novo Peptides' table. The 'Query Spectra' table provides information collected from the original spectra, such as title, precursor m/z, charge, and identification state, while the 'De Novo Peptides' table lists details obtained from the *de novo* sequencing results: peptide sequence, scores, and terminal gaps and precursor m/z and charge. At the bottom, a spectrum viewer[16] shows the currently selected spectrum with the fragment ion annotation corresponding to the selected *de novo* peptide solution. A sequence overlay is also presented on the spectrum by default, aiding the efficient validation of the proposed peptide solution. The *de novo* results can be validated using BLAST, either by clicking the BLAST option at the end of a given line in the 'De Novo Peptides' table or by exporting a list of matches in a BLAST compatible format via the Export menu. Peptide matches can also be exported in a simple text-based format from the same menu.

A reference data set is provided as an example in DeNovoGUI and can be opened easily from the main menu. It consists of 30289 MS/MS spectra from an *Arabidopsis thaliana* whole leaf proteome. The obtained tryptic peptides were separated via ion-pair reversed-phase high-performance liquid chromatography on a poly(styrene/divinylbenzene) monolithic column[17] using a 5 h gradient and were measured on an LTQ Orbitrap XL mass spectrometer using high-resolution precursor ion selection followed by CID fragmentation. This reference data set of a well-established plant model system represents a realistic study case for plant proteomics and is thus ideally suited for the benchmarking of *de novo* sequencing algorithms. For further details about the data set, see the Supporting Information.

Because of its ability to spread the *de novo* task across multiple compute cores and/or hyperthreads, DeNovoGUI substantially reduces the time required to analyze large amounts of spectra using PepNovo+. Indeed, while the analysis of our 30289 MS/MS spectra took ∼7 h using only a single thread, the running time was reduced to approximately 3 h using four threads and to approximately 1.5 h using eight threads. To obtain comparable and consistent results, running times were measured on identical virtual machines with the desired number of cores set (Intel Xeon CPU X5660 at 2.80 GHz). These tests clearly show that the multithreading capability of DeNovoGUI results in substantial reductions in processing time on today's multithreading, multicore laptop and desktop computers.

Upon being downloaded, DeNovoGUI comes with the latest version of PepNovo+ included, and apart from the unzipping of the downloaded DeNovoGUI zip file, no further installation is required to run the software. DeNovoGUI is written in the Java programming language and is freely available as open source under the permissive Apache2 license. Documentation, source files, and binaries can be downloaded from http://denovogui.googlecode.com.

■ ASSOCIATED CONTENT

Supporting Information

Sample preparation, liquid chromatography, and mass spectrometry. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Proteomics Unit, Department of Biomedicine, University of Bergen, Jones Liesvei 91, N-5009 Bergen, Norway. E-mail: marc.vaudel@biomed.uib.no. Telephone: +47 55 58 63 78.

Author Contributions

T.M. and L.W. contributed equally to the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

T.M. and E.R. acknowledge the support of the Max Planck Society. L.W. and C.G.H. appreciate the support by the Austrian Science Funding Agency under Project W_01213. L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), the PRIME-XS project (Grant 262067), and the 'ProteomeXchange' project (Grant 260558), both funded by the European Union 7th Framework Program. H.B. is supported by the Research Council of Norway. We thank Robert Behmüller and Raimund Tenhaken for providing the *Arabidopsis thaliana* plants, as well as An Staes, Evy Timmerman, and Davy Maddelein for their support.

■ REFERENCES

(1) Gevaert, K.; Van Damme, P.; Ghesquiere, B.; Impens, F.; Martens, L.; Helsens, K.; Vandekerckhove, J. A la carte proteomics with an emphasis on gel-free techniques. *Proteomics* **2007**, *7* (16), 2698−2718.

(2) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(3) Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: Toward more efficient identification of post-

translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **2006**, *5* (4), 652−670.
(4) Flikka, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K.; Eidhammer, I. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* **2006**, *6* (7), 2086−2094.
(5) Muth, T.; Benndorf, D.; Reichl, U.; Rapp, E.; Martens, L. Searching for a needle in a stack of needles: Challenges in metaproteomics data analysis. *Mol. BioSyst.* **2013**, *9*, 578−585.
(6) Castellana, N. E.; Payne, S. H.; Shen, Z.; Stanke, M.; Bafna, V.; Briggs, S. P. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (52), 21034−21038.
(7) Veenstra, T. D.; Conrads, T. P.; Issaq, H. J. What to do with "one-hit wonders"? *Electrophoresis* **2004**, *25* (9), 1278−1279.
(8) Allmer, J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteomics* **2011**, *8* (5), 645−657.
(9) Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X. Performance evaluation of existing de novo sequencing algorithms. *J. Proteome Res.* **2006**, *5* (11), 3018−3028.
(10) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−2342.
(11) Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.
(12) Montecchi-Palazzi, L.; Beavis, R.; Binz, P. A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S. The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.* **2008**, *26* (8), 864−866.
(13) 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), 996−999.
(14) 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.
(15) Craig, R.; Beavis, R. C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(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 Bioinf.* **2011**, *12*, 70.
(17) Walcher, W.; Oberacher, H.; Troiani, S.; Hölzl, G.; Oefner, P.; Zolla, L.; Huber, C. G. Monolithic Capillary Columns for Liquid Chromatography-Electrospray Ionization Mass Spectrometry in Proteomic and Genomic Research. *J. Chromatogr., B* **2002**, *782*, 111−125.