Journal of proteome research

Letter

pubs.acs.org/jpr

# Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis

Sean McIlwain,[†] Kaipo Tamura,[‡] Attila Kertesz-Farkas,[‡] Charles E. Grant,[‡] Benjamin Diament,[‡] Barbara Frewen,[‡] J. Jeffry Howbert,[‡] Michael R. Hoopmann,[§] Lukas Käll,[∥,⊥] Jimmy K. Eng,[‡] Michael J. MacCoss,[‡] and William Stafford Noble*[,‡,#]

[†]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, 1552 University Avenue, Madison, Wisconsin 53726, United States

[‡]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, United States

[§]Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109, United States

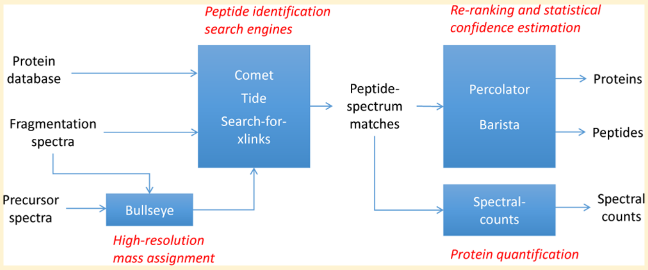
[∥]Science for Life Laboratory, School of Biotechnology, KTH—Royal Institute of Technology, Tomtebodavägen 23A, Solna 17165, Sweden

[⊥]Swedish e-Science Research Centre, KTH—Royal Institute of Technology, 17121 Solna, Sweden

[#]Department of Computer Science and Engineering, University of Washington, 185 Stevens Way, Seattle, Washington 98195, United States

Ⓢ *Supporting Information*

**ABSTRACT:** Efficiently and accurately analyzing big protein tandem mass spectrometry data sets requires robust software that incorporates state-of-the-art computational, machine learning, and statistical methods. The Crux mass spectrometry analysis software toolkit (http://cruxtoolkit.sourceforge.net) is an open source project that aims to provide users with a cross-platform suite of analysis tools for interpreting protein mass spectrometry data.



Modern mass spectrometers produce massive amounts of data. For example, a Thermo Fusion mass spectrometer produces >24 GB of compressed data per day. Keeping pace with such a machine requires balancing three competing needs: analysis software must be *robust*, ensuring that the program executes successfully and that the results are valid, *efficient* to keep pace with the rapid rate of data acquisition, and *state-of-the-art*, gleaning as much information as possible from the data by bringing to bear the latest algorithms and statistical methods.

To simultaneously address these three needs, we created an open source software toolkit called Crux (http://cruxtoolkit.sourceforge.net, Figure 1) that is capable of efficiently and accurately analyzing a variety of types of shotgun proteomics data. Originally, Crux consisted of a single search engine.[1] In Crux v2.0, the original search engine has been replaced by two search engines, Comet[2] and Tide,[3] both of which implement SEQUEST-style searching.[4] In addition, a specialized search engine provides the capability to identify cross-linked peptides.[5] The Bullseye preprocessor assigns high-resolution masses to fragmentation spectra,[6] and the Percolator[7] and Barista[8] post-processors use machine learning techniques to identify and assign statistical confidence estimates to spectra, peptides, and proteins. Peptide and protein quantification can be carried out using a spectral counting tool.[9]

Robust parsing of diverse file formats is an ongoing challenge in computational proteomics. Accordingly, we have adopted the open source ProteoWizard library,[10] which enables Crux to parse a wide variety of file formats (Table 1). In particular, ProteoWizard allows the parsing of vendor-specific raw files when Crux runs under Windows. Furthermore, support for various open file formats allows interoperability between Crux and other search engines as well as toolkits such as the Trans-Proteomic Pipeline[11] and MSDaPl[12] that provide summarization and visualization functionality.

A variety of other mass spectrometry analysis toolkits have been produced, including commercial products (Scaffold, LabKey Server, Mascot tools) and academic software (pFind Studio,[13] Bumbershoot,[14] the Trans-Proteomic Pipeline,[11] MaxQuant,[15] OpenMS,[16] the Global Proteome Machine,[17] and the Central Proteomics Facilities Pipeline[18]). Each of these toolkits offers distinct features (Table 2). Crux offers extensive confidence estimates, including false discovery rate and posterior error probability estimates at the spectrum, peptide, and protein levels and has recently added functionality (to the Tide

**Received:** July 15, 2014
**Published:** September 3, 2014

dx.doi.org/10.1021/pr500741y | *J. Proteome Res.* 2014, 13, 4488−4491

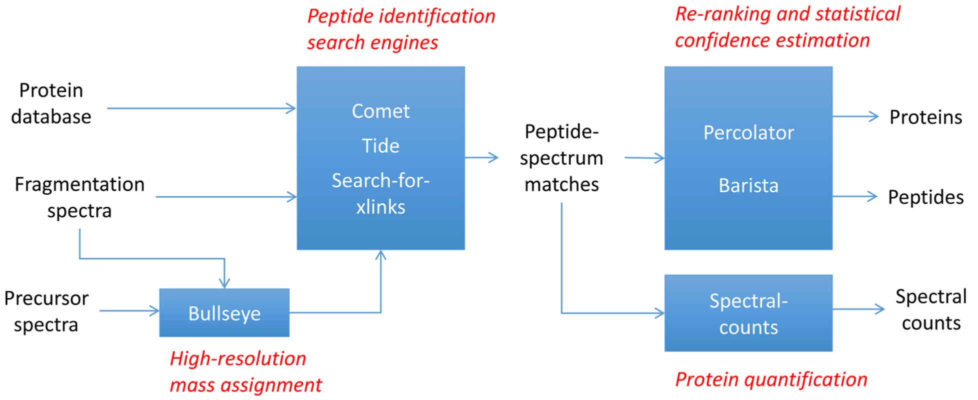


**Figure 1.** Crux analysis workflow and sample results. Crux provides tools for identifying spectra derived from single peptides or from cross-linked peptides as well as tools for postprocessing the resulting identifications to yield peptide- and protein-level identifications.

### Table 1. File Formats in Crux

| command | MS1[a] | MS2 | various[a,b] | FASTA | Tide index | TSV | pepXML | PIN | mzIdentML | SQT | Barista XML |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bullseye | in | in/out | in | | | | | | | | |
| Tide index | | | | in | out | | | | | | |
| Tide search | | in | in | | in | out | out | out | out | out | |
| Comet | | in | in | in | | out | out | out | out | out | |
| Percolator | | | | | | in/out | in/out | in | out | in | in/out |
| Barista | | in | in | | | in/out | out | | | in | out |
| spectral counts | | in | in | | | in/out | in | | in | in | |

[a] Additional vendor proprietary formats for MS1 and MS2 data are supported on Windows: Agilent MassHunter .d, Waters RAW, Thermo RAW, Applied Biosciences Wiff, and Bruker Compass .d/YEP/BAF/FID. [b] Supported open MS2 file formats include BMS2, CMS2, MGF, mzML, and mzXML.

### Table 2. Comparison of Mass Spectrometry Analysis Toolkits[a]

| feature | TPP | MaxQuant | OpenMS | GPM | CPFP | Scaffold | LabKey Server | pFind Studio | Bumbershoot | Mascot tools | Crux |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Tools | | | | | | | | | | | |
| high-res mass assignment | × | × | × | × | | × | | × | | × | × |
| peptide database search | × | × | × | × | × | × | × | × | × | × | × |
| machine learning postprocessor | × | × | | | × | × | × | | × | × | × |
| protein cross-link searching | | | | | | | | × | | | × |
| RNA cross-link searching | | | × | | | | | | | | |
| spectral counting | | × | | × | | × | × | | | × | × |
| isobaric tag quantification | × | × | × | | | × | | × | | × | × |
| peak area quantification | × | × | × | | | × | | × | | × | |
| Statistical Confidence Estimates | | | | | | | | | | | |
| decoy-based estimates | × | × | × | × | × | × | × | × | | × | × |
| parametric PSM $p$ values | | | | × | | × | | × | | | × |
| exact PSM $p$ values | | | | | | | | | | | × |
| PSM $q$ values | × | | × | | × | × | × | × | × | × | × |
| PSM PEPs | × | × | × | | | × | | | | × | × |
| peptide $q$ values | × | | | | × | × | × | × | × | | × |
| peptide PEPs | × | × | | | | | | | | × | × |
| protein $q$ values | × | | | | × | × | × | | × | × | × |
| protein PEPs | × | × | | | | × | | | | | × |
| Input Spectrum File Formats | | | | | | | | | | | |
| Thermo.RAW | × | × | × | | | | × | × | × | × | × |
| Waters.RAW | × | | × | | | | × | | × | × | × |
| MDS/Sciex.wiff | × | × | × | | | | × | | × | × | × |
| Agilent.d | × | | × | | | | × | | × | × | × |
| Bruker.d | × | | × | | | | × | | × | × | × |
| MS1 | | | | | | | | | × | | × |
| MS2 | | | × | | | | | × | × | | × |
| mzML | × | | × | × | × | | × | | × | × | × |
| mzXML | × | × | × | × | | | × | | × | × | × |
| MGF | | | × | × | × | | | × | × | × | × |

Table 2. continued

| feature | TPP | MaxQuant | OpenMS | GPM | CPFP | Scaffold | LabKey Server | pFind Studio | Bumbershoot | Mascot tools | Crux |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Input PSM File Formats | | | | | | | | | | | |
| PepXML | × | | × | | | | × | | | | × |
| mzIdentML | × | | × | | | × | | | × | | |
| mzQuantML | | | × | | | | | | | | |
| .dat (Mascot) | × | | | | | × | | | | | |
| .out (SEQUEST) | × | | | | | × | | | | | |
| .sqt (SEQUEST) | × | | | | | × | | | | | × |
| .srf (SEQUEST) | | | | | | × | | | | | |
| other tool-specific formats | | | | | | × | | | | | |
| Output File Formats | | | | | | | | | | | |
| tab-delimited | × | × | × | × | | × | × | × | × | × | × |
| mzTab | | × | × | | | | | | | × | |
| PepXML | × | | × | | × | | | | | × | × |
| ProtXML | × | | | | | × | | | | | |
| mzIdentML | × | | × | | | × | | | × | × | × |
| mzQuantML | | | × | | | | | | | | |
| Implementation | | | | | | | | | | | |
| free | × | × | × | × | × | | × | × | × | | × |
| source code available | × | | × | × | × | | × | | × | | × |
| open source license | × | | × | × | × | | × | | × | | × |
| Linux binaries | | | × | | | × | × | × | × | × | × |
| MacOS binaries | | | × | | | × | × | | | | × |
| native Windows binaries | × | × | × | | | × | × | × | × | × | × |
| command line interface | × | × | × | × | | × | | | × | × | × |
| graphical user interface | × | × | × | × | × | × | × | × | × | × | |
| application programming interface | | | × | | | | × | | | × | |

[a] "Mascot tools" refers to Mascot Server and Mascot Distiller, which are licensed separately. GPM is Perl-based, so no binaries are needed. Scaffold parses tool-specific PSM formats produced by Proteome Discoverer, MS Amanda, Byonic, OMSSA, MaxQuant, SpectrumMill, X!Tandem, Waters Identity E, and Phenyx. Note that as of August 2014 CPFP is no longer actively maintained.

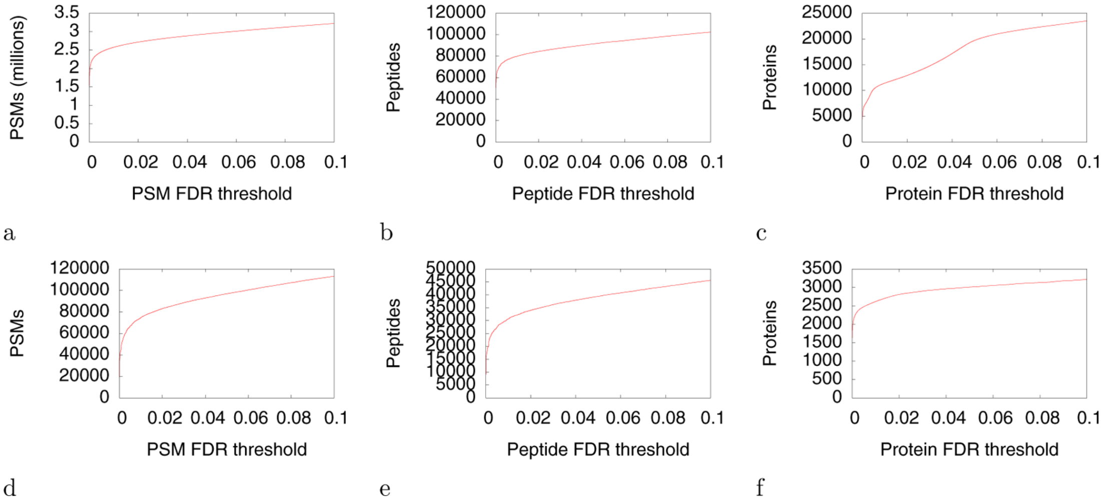


**Figure 2.** (a−c) We used Tide+Percolator to analyze 9 092 380 fragmentation spectra from 95 different human samples. The figure plots the number of spectra, peptides and proteins identified as a function of false discovery rate threshold. (d−f) Each panel plots, from Comet+Percolator analysis of 348 157 *Plasmodium falciparum* fragmentation spectra, the number of (respectively) spectra, peptides and proteins identified as a function of false discovery rate threshold. Total analysis time was 61.2 m (34.4 m for Comet and 26.8 m for Percolator). The number of proteins identified at 1% FDR (2618) by Comet+Percolator compares favorably with the published analysis (2767 proteins).

search engine) to compute exact *p* values using a dynamic programming approach.[19]

Crux supports a variety of workflows, providing users with flexibility to tailor their analysis to their experimental goals. The choice of search engine—Comet versus Tide—is a matter of personal preference and processing considerations and is not likely to substantially affect the final results. Tide is faster on a single thread, but, unlike Comet, does not yet operate in multithreaded mode. Exact *p* values, which are only available in Tide, provide significantly improved statistical power at the expense of some computational overhead (roughly 0.2 s per spectrum). The two primary postprocessors, Percolator and Barista, offer more substantial differences. Both use a target-decoy machine learning approach. However, Percolator first learns to rerank peptide-spectrum matches (PSMs) and then performs a probabilistic protein-level inference,[20] whereas Barista formulates both tasks jointly in a single discriminative learning procedure. Which approach performs better in practice is an open question that deserves further exploration.

To demonstrate the efficiency and accuracy of our software, we downloaded 224 GB of compressed data from a recent study of genetic control of protein abundance in humans[21] (details in the Supporting Information). Searching these >9 million fragmentation spectra using Tide against a human protein database containing ~90,000 proteins and a matched set of decoys required 20.2 h of CPU time on a single thread, for a rate of 121 spectra/s. Postprocessing with Percolator required an additional 20.5 min. At 1% false discovery rate (FDR) thresholds for PSMs, peptides, and proteins, respectively, this analysis identified 2 576 283 PSMs, 79 976 peptides, and 11 432 proteins (Figure 2a–c). These results are comparable to the published analysis, which reported 2 726 242 PSMs corresponding to 71 800 distinct peptides at a 1% peptide-level FDR threshold. We also used Comet and Percolator to analyze a collection of 348 157 high-resolution spectra from the erythrocytic cycle of the malaria parasite *Plasmodium falciparum*,[22] identifying at 1% FDR 74 974 PSMs, 30 640 peptides, and 2618 proteins (Figure 2d–e).

Crux is a command line tool, written in C++ and distributed as a single binary executable supporting a variety of commands. Users wishing to compile their own version of Crux can download the source code, which is covered by an Apache license. All Crux code undergoes code review and revisions to reflect our documented coding standards, and the software is automatically tested using a continuous integration system, which compiles Crux on three operating systems—Windows, MacOS and Linux—thereby providing up-to-date binary executables. Crux is under active development, with several important improvements and additions planned for the near future. In addition, we encourage community members to contribute to the toolkit.

## ■ ASSOCIATED CONTENT

### Ⓢ Supporting Information

Sample analyses. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: william-noble@uw.edu. Tel: (206) 221-4973. Fax: (206) 685-7301.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank the authors of the various software toolkits—Jürgen Cox, Oliver Kohlbacher, Ron Beavis, Audrey Lester, Brian Searle, Chao Liu, and John Cottrell—for their assistance in ensuring the accuracy of Table 2.

## ■ REFERENCES

(1) Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. P.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **2008**, *7*, 3022–3027.

(2) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics* **2012**, *13*, 22–24.

(3) Diament, B.; Noble, W. S. Faster sequest searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **2011**, *10*, 3871–3879.

(4) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(5) McIlwain, S.; Draghicescu, P.; Singh, P.; Goodlett, D. R.; Noble, W. S. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *J. Proteome Res.* **2010**, *9*, 2488–2495.

(6) Hsieh, E.; Hoopmann, M.; Maclean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138–1143.

(7) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925.

(8) Spivak, M.; Weston, J.; Tomazela, D.; MacCoss, M. J.; Noble, W. S. Direct maximization of protein identifications from tandem mass spectra. *Mol. Cell. Proteomics* **2012**, *11*, M111.012161.

(9) McIlwain, S.; et al. Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinf.* **2012**, *13*, 308.

(10) Chambers, M. C.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920.

(11) Deutsch, E. W.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150–1159.

(12) Sharma, V.; Eng, J. K.; MacCoss, M. J.; Riffle, M. A mass spectrometry proteomics data management platform. *Mol. Cell. Proteomics* **2012**, *11*, 824–831.

(13) Li, D.; et al. pfind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **2005**, *21*, 3049–3050.

(14) Holman, J. D.; Ma, Z.-Q.; Tabb, D. L. Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker. *Curr. Prot. Bioinf.* **2012**, *13*, 10.1002/0471250953.bi1317s37.

(15) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

(16) Sturm, M.; et al. OpenMS– an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.

(17) Craig, R.; Cortens, J. P.; Beavis, R. C. An open source system for analyzing, validating and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234–1242.

(18) Trudgian, D. C.; et al. CPFP: a central proteomics facilities pipeline. *Bioinformatics* **2010**), *26*, 1131–1132.

(19) Howbert, J. J.; Noble, W. S. Computing exact *p* values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **2014**, 10.1074/mcp.O113.036327.

(20) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **2010**, *9*, 5346–5357.

(21) Wu, L.; et al. Variation and genetic control of protein abundance in humans. *Nature* **2013**, *499*, 79–82.

(22) Pease, B. N.; et al. Global analysis of protein expression and phosphorylation of three stages of Plasmodium falciparum intraerythrocytic development. *J. Proteome Res.* **2013**, *12*, 4028–4045.