technical notes Journal of proteome research

# MGFp: An Open Mascot Generic Format Parser Library Implementation

**Marc Kirchner,[†,‡] Judith A. J. Steen,[§] Fred A. Hamprecht,[||] and Hanno Steen*[,†,‡]**

*Proteomics Center, Children's Hospital Boston, Boston, Massachusettes, Departments of Pathology, Harvard Medical School and Children's Hospital Boston, Boston, Massachusettes, Department of Neurobiology, Harvard Medical School and T. M. Kirby Neurobiology Center, Children's Hospital, Boston, Massachusettes, and Multidimensional Image Processing Group, Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany*

**Received February 6, 2010**

***Abstract:*** Despite the efforts of the mass spectrometry (MS) community to migrate data representation toward modern file formats, legacy text formats still play an important role in MS data processing workflows. We provide a formal grammar and a portable, efficient C++ implementation for a Mascot Generic Format (MGF) parser. Software and technical documentation are available from http://software.steenlab.org/mgfp/.

**Keywords:** mass spectrometry • Mascot Generic Format • MGF

Despite significant efforts in the mass spectrometry community to define, implement, and motivate XML-based open data formats such as mzXML,[1] mzML,[2] pepXML,[3] protXML,[4] and mzIdentXML,[5] legacy text formats are still widely used. A particular example is the Mascot Generic Format (MGF).[6] MGF is the *de facto* standard for MS2 data storage and peptide/protein search submission in existing mass spectrometry data analysis workflows. As a consequence, the complete transition from MGF to a modern file format is likely to be a lengthy process during and after which it will be necessary to maintain backward compatibility for tools and processing pipelines.

Given the importance of MGF, it is surprising to see that: (i) MGF is not a rigorously defined format, that is, there is no formal grammar;[6] (ii) there exist no efficient, openly available parser implementations for compiled languages (e.g., C++ or the .NET environment). Current analysis software needs to include *ad hoc* MGF parser implementations that support the MGF subset necessary for the analysis task at hand.

As illustrated in Figure 1, we have derived a formal grammar from the MGF format description and have implemented the grammar using the bison/flex parser generators.[7,8] We have defined and implemented an efficient, intuitive C++ library interface which can easily be integrated into existing C++ projects, adapted to managed code environments (e.g., .NET) or bridged to scripting languages such as Python[9] or R.[10] The library is portable and has been tested successfully on Linux, MacOS, and MS Windows platforms. Binaries and source code are freely available under a BSD license and can be downloaded from http://software.steenlab.org/mgfp.

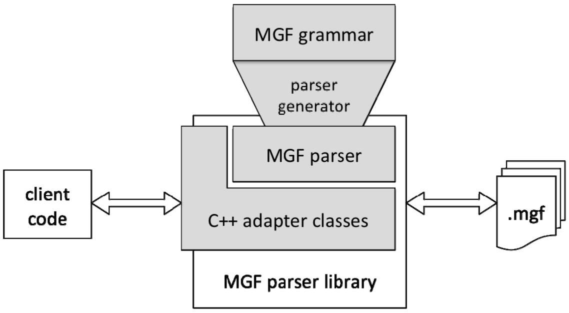

**Figure 1.** Schematic overview of the MGFp library components. The formal MGF grammar is translated into a bison/flex[7,8] parser which is encapsulated in a C++ interface. The library is capable of reading and writing MGF files and provides convenient data access for client code.

**Acknowledgment.** We thank Bernhard X. Kausler, and Thorben Kroeger for code contributions and Anna Kreshuk, Wiebke Timm, Bernhard Y. Renard, Michael Hanselmann, Xinghua Lou, and Ullrich Koethe for helpful discussions.

## References

(1) Pedrioli, P. G. A.; et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004**, *22* (11), 1459–1466.
(2) Chambers, M. Significant improvements to the PSI mass spectrometer data file standard: mzML, 1.1. *Annual Conference of the American Society for Mass Spectrometry*, 2009.
(3) Seattle Proteome Center, Open Formats Information for pepXML (http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML, accessed 2010-03-29).
(4) Seattle Proteome Center, Open Formats Information for protXML (http://tools.proteomecenter.org/wiki/index.php?title=Formats:protXML, accessed 2010-03-29).

* To whom correspondence should be addressed. Hanno Steen, Proteomics Center, Children's Hospital Boston, 320 Longwood Avenue, Boston, MA 02115, E-mail: hanno.steen@childrens.harvard.edu.

[†] Proteomics Center, Children's Hospital Boston.

[‡] Departments of Pathology, Harvard Medical School and Children's Hospital Boston.

[§] Department of Neurobiology, Harvard Medical School and T. M. Kirby Neurobiology Center, Children's Hospital.

[||] University of Heidelberg.

10.1021/pr100118f

(5) HUPO Proteomics Standards Initiative, Proteomics Informatics Standards Group (PSI-PI) Online information (http://www.psidev.info/index.php?q=node/40, last accessed 2010-03-29).

(6) MatrixScience, Mascot Generic Format Documentation (http://www.matrixscience.com/help/data_file_help.html, last accessed 2010-03-29).

(7) Bison GNU parser generator (http://www.gnu.org/software/bison/, last accessed 2010-03-29).

(8) flex: The Fast Lexical Analyzer (http://flex.sourceforge.net/, last accessed 2010-03-29).

(9) Python Programming Language Official Website (http://www.python.org/, last accessed 2010-03-29).

(10) R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2004).

PR100118F