Published in final edited form as:
Nat Methods. 2009 April ; 6(4): 240–241. doi:10.1038/nmeth0409-240.

mzAPI: a new strategy for efficiently sharing mass spectrometry data

Manor Askenazi[1,3], **Jignesh R. Parikh**[2], and **Jarrod A. Marto**[1,**]

1*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston University, Boston, MA 02115-6084*

2*Bioinformatics Program, Boston University, Boston, MA 02115-6084*

3*Department of Biological Chemistry, The Hebrew University of Jerusalem, Israel*

To The Editor: The call for data access standards in mass spectrometry-based proteomics has led to proposals focused on the *extraction* of native data to XML-based formats. [1,2] While self-describing and human-readable formats represent laudable goals, particularly for archival purposes, they are not well suited to large numeric datasets. Consequently, while metadata in mzML[2] remain human-readable, the vast majority of the file is devoted to a hexadecimal representation of the mass spectra. Moreover, the transition from mzXML to a true XML format (mzML2) eliminates embedded indexing schemes; consequently, extracted files are compromised in both content and access efficiency.[1,3]

Based on similarities in data structure and access patterns, we suggest that fields such as astronomy are better models for proteomics data analysis (Figure 1). These fields also rely on common formats, but typically utilize binary standards such as HDF5[4] or netCDF[5]. By contrast, the commercial nature of mass spectrometry has led to the evolution of proprietary binary file formats. In light of these observations, we propose that a common and redistributable application programming interface (API) represents a more viable approach to data access in mass spectrometry. In effect, we propose to shift the burden of standards compliance to the manufacturers' existing data access libraries.

While our proposal for *abstraction* through a common API represents a clear departure from current attempts to define a universal file format, it is by no means unique within the broader scientific community. For example, standardized APIs have enabled the development of portable applications in such diverse areas as computer graphics (OpenGL7) and parallel processing (Message Passing Interface, MPI[8]). More importantly, we believe that a common API will benefit all stakeholders. For example, free redistribution of compiled, vendor-supplied dynamically linked libraries (DLLs) will protect the proprietary layout of native files and provide users with direct and flexible access to data system- and instrument-specific functionality which are typically ignored by lowest common denominator export routines. In addition, we note that mzAPI naturally supports the FDA's 21 CFR part 11 regulatory requirements for electronic records[9] Finally, a community-driven API standard will facilitate manufacturer support of UNIX, in addition to Windows, by highlighting the subset of procedures, from each data system (Xcalibur™, Analyst™, etc.), which need to be ported.

Motivated originally by our desire to provide a more intimate environment for flexible and in-depth exploration of mass spectrometry data, particularly from low-throughput experiments,

**Email: E-mail: jarrod_marto@dfci.harvard.edu.
Department of Cancer Biology and Blais Proteomics Center, Dana-Farber Cancer Institute

we developed a preliminary common API (mzAPI) – consisting of just five procedures (http://blais.dfci.harvard.edu/mzAPI). To maximize accessibility we exposed mzAPI in the form of a Python library within a flexible, mass-informatics desktop framework called multiplierz (http://blais.dfci.harvard.edu/multiplierz). We are encouraged by results from this test harness, in particular how well mzAPI and our desktop environment support a variety of data analytic operations. Equally impressive is how quickly non-programmers can customize scripts for their specific tasks. Despite success to date in our own lab, we recognize that mzAPI will benefit from further refinement and stress testing. Accordingly, we are actively seeking input from the research community with respect to both concept and implementation of a comprehensive API-based standard for mass spectrometry data access and analysis.

ACKNOWLEDGMENTS

The authors thank Yi Zhang and Scott Ficarro for valuable discussion and input, and Eric Smith for preparation of Figure 1. This work was supported by the Dana-Farber Cancer Institute and the National Human Genome Research Institute (P50HG004233).

REFERENCES

1. Pedrioli PG, Eng JK, Hubley R, et al. Nature biotechnology 2004;22:1459.
2. Deutsch E. PROTEOMICS 2008;8:2776. [PubMed: 18655045]
3. Lin SM, Zhu L, Winter AQ, et al. Expert Rev Proteomics 2005;2:839.
4. Folk M, Cheng A, McGrath RE. Astronomical Data Analysis software and Systems VIII Proceedings 1999;172
5. Rew R, Davis G. Computer Graphics and Applications, IEEE 1990;10:76.
6. Lorimer DR, Bailes M, McLaughlin MA, et al. Science 2007;318:777. [PubMed: 17901298]
7. http://www.opengl.org/
8. http://www-unix.mcs.anl.gov/mpi/
9. http://www.fda.gov/ora/compliance_ref/Part11/

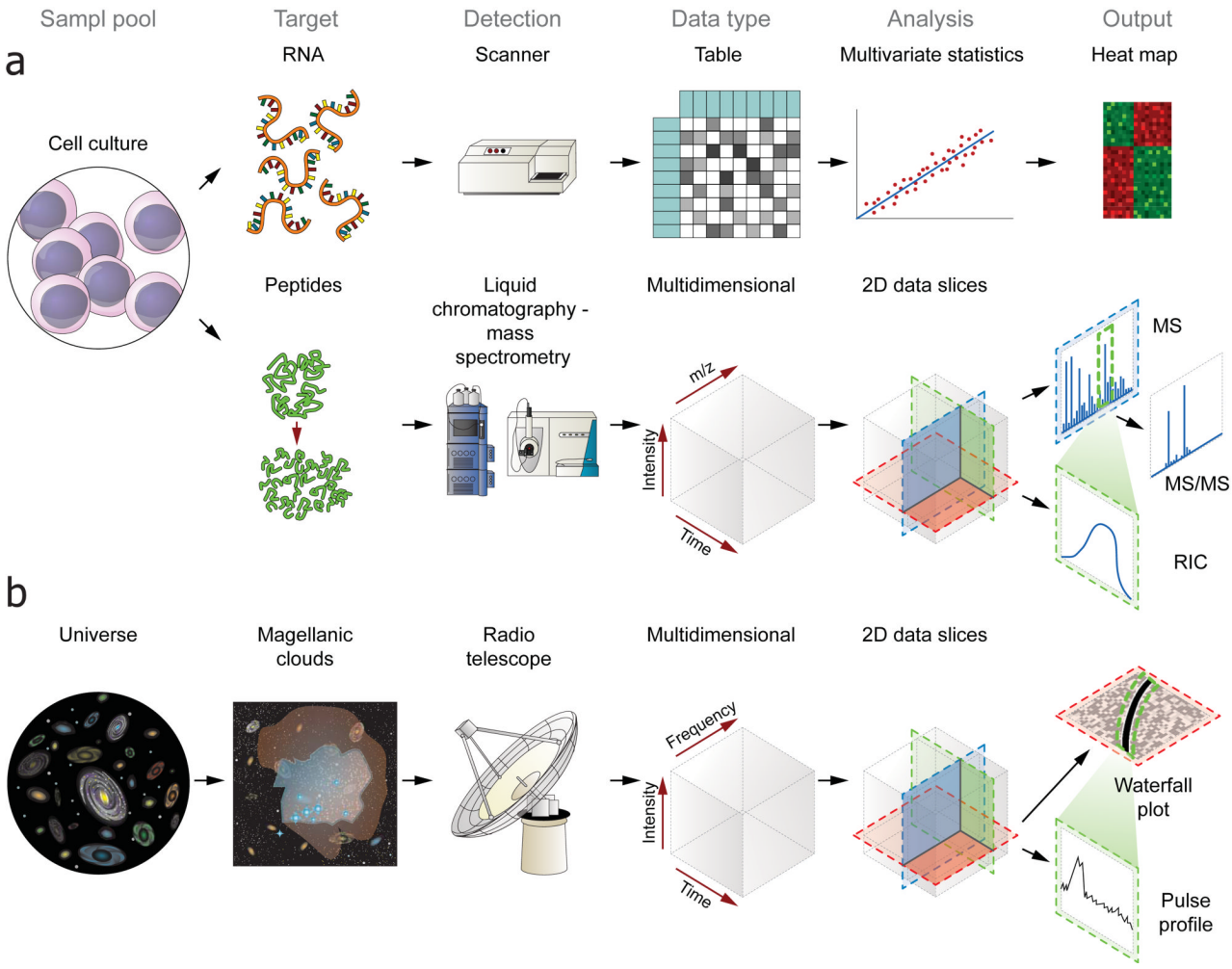

Figure 1. Array Scanners, Telescopes, and Mass Spectrometers: XML, HDF, or API?

(a) While biological samples provide both protein and RNA, subsequent large-scale analyses (microarray or proteomics) yield data structures that diverge with respect to typical access patterns. For example **(a, top)**, features detected by array scanners can be exported to a tabular format, immediately suited for clustering (or other multivariate analysis). In contrast **(a, bottom)**, LC-MS experiments for proteomics generate complex multidimensional data, where feature characterization is itself, still an active area of research. The underlying raw data is repeatedly accessed as 2-dimensional slices, reconstructed ion chromatograms (RIC) for example, and hence requires inclusion of indices in file formats designed to accommodate large-scale experimental results. (**b**) A similar situation exists in fields outside biomedical research, where, for example, slices taken through radio frequency data yield waterfall plots and pulse profiles which are used to characterize signals of astrophysical origin. [6]