The mzML vendor-neutral data format for chromatography-mass spectrometry

Seán O'Callaghan[1]; Steffen Neumann[2]; Nils Hoffmann[3]; Eric Deutsch[4]; Vladimir A. Likić[1]

[1]Metabolomics Australia, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Australia; [2]Department for Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany; [3]Faculty of Technology, AG Genominformatik, Bielefeld University, Germany; [4]Institute for Systems Biology, Seattle, WA

Introduction

Mass spectrometry produces a huge amount of raw data that is not manageable without computerized automation. Natively mass spectrometers store output in a variety of proprietary formats, which hinders data sharing and makes writing vendor-neutral software difficult.

The purpose of this work is to develop a vendor-neutral open format that all vendor software can adopt, thereby facilitating data exchange between vendor software and vendor neutral software packages.

A suitable data format for mass spectrometry should capture both the spectra and corresponding metadata. XML is designed to describe hierarchically structured data in a textual data format. It was designed to facilitate simplicity, generality, and usability for electronic data exchange. It is easily parsed by software and easy to read.

Hence XML appears to be the natural choice for the storage of mass spectrometry data.

History

Developed by an international proteomics consortium including Human Proteome Organisation (HUPO) and Proteomics Standards Initiative (PSI).

The mzML format has replaced two previous formats with the same purpose developed by the proteomics community (mzXML and mzData).

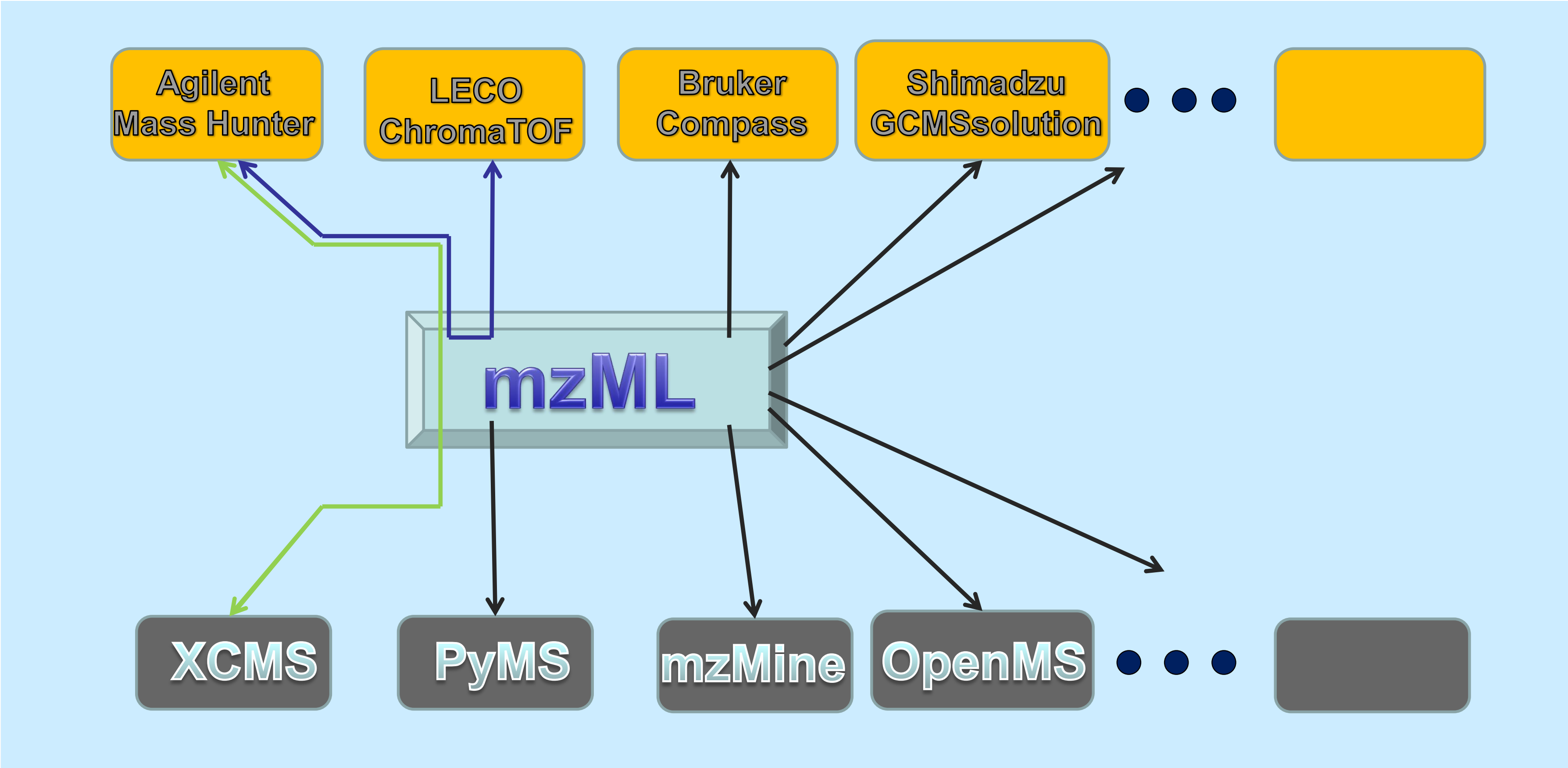
mzML combines the best features of mzXML and mzData with the specific purpose to provide a unified standard data format for mass spectrometry based proteomics.

The metabolomics community currently lacks standard data format for the storage and exchange of mass spectrometry data. The mzML format specification provides a unique opportunity to address this challenge by building on the experience of the proteomics community, and the success of mzML.

Fig1: mzML as a Conduit

mzML can provide a common language between :

- instruments
- vendor software
- vendor neutral software
- open source software



Challenges

- Few examples of GC-MS data mzML data sets currently available.
- Possible that changes to the format are necessary to achieve good encapsulation of GC-MS data.
- Some metabolomics specific meta-data needs may have to be addressed.

Aims

- To (1) test, (2), validate, (3) recommend, and in the long-term (4) promote the use of mzML format as a standard vendor-neutral format for mass spectrometry-based metabolomics
- Achieve a consensus view on the suitability of mzML for mass spectrometry based metabolomics after testing and validation

Future Work

- Seek further engagement with instrument vendors & the wider metabolomics community
- Produce tools and information to accelerate the uptake of mzML within the metabolomics community

Results

LC-MS:

- A number of software tools can take advantage of mzML already exist, e.g. XCMS (metlin.scripps.edu/xcms/), OpenMS (www.openms.de)

GC-MS:

- First GC-MS single quad data set translated into mzML (Thermo ISQ single quad system)
  - Using Proteowizard tool (proteowizard.sourceforge.net)
- First GCxGC data set translated (LECO) into mzML
- Ability to convert NetCDF to mzML
  - Both using OpenMS file converter

CE-MS:

- Future work

Additionally, some manufacturers, notably LECO and Shimadzu are producing MS machines/software natively capable of writing mzML

Advisors

Christoph Steinbeck (European Bioinformatics Institute), Tony Bacic (University of Melbourne), Steve Fischer (Agilent Technologies), Malcolm McConville (University of Melbourne), Joachim Kopka (Max Planck Institute).

Bioplatforms Australia

THE UNIVERSITY OF MELBOURNE

SCHOOL OF BOTANY

bio21 institute

ACPFG AUSTRALIAN CENTRE FOR PLANT FUNCTIONAL GENOMICS PTY LTD

State Government Victoria — Department of Innovation, Industry and Regional Development

An Australian Government Initiative
National Collaborative Research Infrastructure Strategy

Web page at http://bioinformatics.bio21.unimelb.edu.au/mzML.html