

Properties of mzXML and pepXML

Joris Borgdorff

February 15, 2006

1 Sashimi

The XML-formats of mzXML and pepXML belong to a project making open source software tools for mass spectrometric data, in particular for proteomics. This project is called Sashimi and can be found on <http://sashimi.sourceforge.net/>.

One of the goals of Sashimi is to create standards for data for proteomics.

Via their contact page there is a forum which could answer some of the questions we will run into. It is mildly active, with 4 threads à 4 posts per month.

All tools can be found at http://sourceforge.net/project/showfiles.php?group_id=69281&release_id=200568

2 mzXML

Different mass spectrometers produce different output (MS data). One can convert this MS data to mzXML. This way further processing can be done on a standard file format, instead of several different formats.

MS data is first compressed with Base64 encoding and then put in a mzXML schema. The conversion with Base64 is done to reduce file size and processing time. Base64 encoding is the closest thing to binary data an ASCII file can get.

Sashimi offers several tools to deal with this mzXML:

- validateXML,
- mzXML2Other,
- Random Access Parser (RAP).

- `mzXMLViewer`

As these tools fulfill our needs we probably will not need the exact XML schema for `mzXML`. This is available though, at http://sashimi.sourceforge.net/software_glossolalia.html#mzXMLSchema, along with some documentation and a tutorial about the schema. This could become very useful as this tutorial is the only one available on the site. This means we will need it for figuring out the other tools as well.

Example `mzXML` files can be found in the data repository at Sashimi.

2.1 ValidateXML

The tool `ValidateXML` simply validates the `mzXML`. This is very useful if data was transferred over a network as `ValidateXML` also detects data corruption.

2.2 mzXML2Other

Some applications, like databases, need a special format for their MS data. With `mzXML` these formats can be generated.

2.3 RAP

In most programming languages there are already tools available for parsing XML, notably SAX (Simple API for XML), available at <http://www.saxproject.org>. The main disadvantage of SAX is that it parses XML data sequential, while applications often require their data in a non-sequential method.

This can be solved by indexing the XML-document. When the `mzXML` is indexed random access parsing becomes possible. This led to the creation of the RAP, which is available in C, C++ and Java. They are called RAMP, RAP and JRAP respectively where JRAP is available as jar-file.

RAMP stands for Random Access Minimal Parser which is simpler and faster than the other two parsers but has less control options. There is also a C++ wrapper available for RAMP, called `cRAMP`.

The documentation concerning the parsers is minimal. JRAP has documentation generated by javadoc, while documentation about RAP and RAMP have to be searched in the source.

Sources can be found on http://sashimi.sourceforge.net/software_glossolalia.html#RAP.

2.4 mzXML Viewer

There is also a simple mzXMLViewer. If anything, the source can teach us something about how RAP works.

3 PepXML

The pepXML format is generated after a database search. It is part of the Trans Proteomic Pipeline (TPP). The TPP includes all steps of the MS/MS analysis pipeline after results of database search: Peptide validation, Peptide quantitation, Protein identification, and Protein quantitation.

There does not seem to be a parser available for pepXML. There is an XML scheme with some documentation. Quite some studying will be necessary to use this. The scheme can be found at http://sashimi.sourceforge.net/software_tpp.html#pepXML.