RpepXML: an R interface to the pepXML format for peptide identificaiton.

Laurent Gatto
lg390@cam.ac.uk
Cambridge Center for Proteomics – EBI

Abstract

This vignette describes the data structures and associated methods and functions to import peptide identificaion data in the `pepXML` format.

Keywords:˜mass spectrometry, proteomics, MSMS, identification, XML.

1. Introduction

From the `pepXML` webpage[1]:

> `pepXML` is an open data format developed at the SPC/Institute for Systems biology[2] for the storage, exchange, and processing of peptide sequence assignments of MS/MS scans. `pepXML` is intended to provide a common data output format for many different MS/MS search engines and subsequent peptide-level analyses. Several search engines already have native support for outputting `pepXML` and converters are available to transform output files to `pepXML`.
>
> Note that the HUPO Proteomcs Standards Initiative[3] (PSI) has also develop an exchange standard for database search results, called `mzIdentXML`[4]. It is not yet as widely used as `pepXML` but should supersede the latter. A `RmzIdentML` will be developed when at a later stage.
>
> Currently, only the generic MSMS identification results of the `msms_pipeline_analysis` is implemented. This package has currently been tested and developped around Mascot[5] search results.

2. Data structure

The classes implemented mimic a simplified version of the XML structure for the `msms_pipeline_analysis` element (see the docs[6], for a browsable desciption). The structure of the 4 classes is described below, starting with the main, high-level, data structure. Each class and slots are described in the respective on-line manuals.

3. Session information

[1] http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML
[2] http://tools.proteomecenter.org/wiki/index.php?title=Main_Page
[3] http://www.psidev.info/
[4] http://www.psidev.info/index.php?q=node/319
[5] http://www.matrixscience.com/
[6] http://sashimi.sourceforge.net/schema_revision/pepXML/Docs/pepXML_v18.html

- R version 2.12.0 Under development (unstable) (2010-08-22 r52792), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_GB.utf8`, `LC_NUMERIC=C`, `LC_TIME=en_GB.utf8`, `LC_COLLATE=C`, `LC_MONETARY=C`, `LC_MESSAGES=en_GB.utf8`, `LC_PAPER=en_GB.utf8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_GB.utf8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Loaded via a namespace (and not attached): tools˜2.12.0