Clinical Proteomic Tumor Analysis Consortium Data Coordinating Center (CPTAC-DCC) mzIdentML Format Peptide-Spectrum-Matches

Summary

The Data Coordinating Center (DCC) converts peptide-spectrum-matches (PSMs) computed by the common data analysis pipeline (CDAP) at NIST or by the protein characterization centers (PCCs) to a common, consistent, HUPO Proteome Standards Initiative (PSI) compliant mzIdentML format for use by (CPTAC-program) internal and external data-consumers. The resulting mzIdentML format PSMs are intended (by the DCC) to provide a stable, consistent data-format for third-party software developers.

Authors

Nathan J. Edwards
Georgetown University
Department of Biochemistry and Molecular & Cellular Biology
(nje5@georgetown.edu)

Table of Contents

Versions

The version numbers of the software used to compute the PSMs and carry out the conversion are listed in the AnalysisSoftwareList element of the mzIdentML document. Where PSI-MS terms for the software are available, they are used; otherwise the “analysis software” designation is used. Where no official version is available, the SVN repository revision or an md5 checksum for the binary is provided.

Also provided in the AnalysisSoftwareList element of the mzIdentML document is a CPTAC-DCC:mzIdentML element, designated as “file format”, which provides a version number for the mzIdentML file-format itself. This version number will be incremented for any significant change to the generated XML.

Peptide-Spectrum-Matches (PSMs)

The peptide-spectrum-matches determined by the CDAP or by the PCCs for CPTAC spectra are computed, scored, and filtered using a variety of tools – these are described in the appropriate accompanying methods documents. The DCC conversion to mzIdentML does not change the matches or their scores – but merely semantically transforms and reformats them to a common, consistent format. The intent is to accurately describe the PSMs without necessarily describing the method used to find them.

The PSMs, prior to conversion at the DCC, are filtered to remove runner up peptide hits and decoys and are filtered for statistical significance. While there is generally only one PSM per spectrum, more than one PSM may receive the best score, resulting in a tie. In this case, all PSMs with the best score are output. See the appropriate PSM methods documents for more details.

Each PSM links an identifier for the spectrum, the peptide sequence, any post-translational modifications on the peptide, and a list of identifiers for the protein sequences found to contain the peptide sequence. In addition, depending on the analysis pipeline, PSMs may be annotated with additional information, such as iTRAQ reporter ion intensities and post-translational modification localization scores.

During format conversion, the input PSM spectral identifiers are mapped to unambiguous mzML nativeIDs, and checked against the corresponding mzML spectra data file. In the process, each spectrum’s precursor m/z value, retention time, and precursor intensity (where available) are extracted for insertion into the resulting mzIdentML.

Post-translational modifications are mapped to their corresponding PSI-MOD terms and PSI conventions for describing N-terminal modifications are applied.

All theoretical masses and m/z values are recomputed from the elemental composition of the provided peptide sequences and their PSI-MOD modifications using ProteoWizard APIs.

Where possible, search engine scores, metrics, and parameters of the input PSMs are providing using PSI-MS terms, otherwise a prefix is added to the score name indicating the source. For CPTAC CDAP pipeline scores, the prefix "CPTAC-CAP:" (CPTAC-DCC:mzIdentML version 1.0) or "CPTAC-CDAP:" (CPTAC-DCC:mzIdentML version 1.1) is used. See the appropriate methods documents for more details on these scores. The score used for filtering the PSMs and the filter threshold (if available) is declared in the SpectrumIdentificationProtocol element.

Peptide Alignments

Identified peptide sequences of the PSMs are realigned to organism specific RefSeq and UniProt protein sequences, generally human, or human and mouse for human in mouse tumor xenografts.

RefSeq protein sequences are downloaded using the URLs:

HUMAN: ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz
MOSUE: ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.protein.faa.gz

UniProt protein sequences are downloaded using the URLs:

HUMAN:
http://www.uniprot.org/uniprot/?query=taxonomy%3a9606+AND+keyword%3a1185&force=yes&format=fasta&include=yes
MOUSE:
http://www.uniprot.org/uniprot/?query=taxonomy%3a10090+AND+keyword%3a1185&force=yes&format=fasta&include=yes

UniProt sequences are those tagged with the keyword "Reference Proteome" and include enumerated isoforms.

In each case, the most current release of RefSeq and UniProt at the time of conversion is used, and the release/version is documented in the SearchDatabase element.

After alignment, peptide start and end positions and amino-acids to the left and right of the peptide are populated in PeptideEvidence elements, and DBSequence elements are populated with accessions in a consistent format and human readable descriptions. Protein descriptions include the accession fields, to facilitate alternative accession extraction (UniProt ID, gi number), and contain other useful content such as the gene name (GN keyword of UniProt descriptions).

The DBSequence elements reference SearchDatabase element ids, which provide a namespace and (where possible) organism for each accession e.g. RefSeq:Human, UniProt:Mouse. SearchDatabase elements also provide an explicit release string where available.

Where protein accessions from the input PSMs cannot be recapitulated by realignment to UniProt and RefSeq protein sequences to obtain peptide start and end positions, amino-acid context, consistent accessions, or protein descriptions, these additional fields are populated if provided and left blank otherwise. Often, these accessions represent retired protein sequences, but they are retained for completeness. Occasionally, these accessions provide the sole evidence for a specific peptide sequence. These accessions will reference a SearchDatabase that indicates the namespace of the accessions, such as RefSeq or UniProt.

XML Element IDs

The mzIdentML format uses element ids for spectra results (SpectrumIdentificationResult), PSMs (SpectrumIdentificationItem), peptides (Peptide), peptide alignments (PeptideEvidence), and protein sequences (DBSequence); these ids enable cross-element references. Where possible the CPTAC DCC mzIdentML format uses “information-rich” ids. These identifiers can be used without semantic interpretation as unique identifiers by PSI compliant mzIdentML parsers, but they also provide a convenient short-cut for ad-hoc analyses. These ids make it possible to extract PSMs from the mzIdentML without having to randomly address the mzIdentML sequence collections, in particular.

The following table describes the format of the ids for each of the above elements:

SpectrumIdentificationResult	<spectral-filename>|<scan>
SpectrumIdentificationItem	<spectral-filename>|<scan>|<psm-ordinal>
Peptide	<peptide-sequence>(|<mod>)*
PeptideEvidence	<sequence-accession>|<left-aa>|<start>|<peptide-sequence>|<end>|<right-aa>(|<mod>)*
DBSequence	<sequence-accession>

The “<mod>” value specifies the amino-acid modification(s) of the peptide: “<amino-acid-symbol><position>:<signed-delta>”, with N-terminal modifications specified “0:<signed-delta>”. “(|<mod>)*” indicates zero, one, or more amino-acid modifications separated by “|”. The specific interpretation of these informally useful ids is readily apparent by following the formal mzIdentML references. Where values are unavailable (PeptideEvidence start and end, for example), these are left blank, but all delimiters remain.

Search Engine Parameters

The separation of roles and responsibilities between the DCC and those computing the PSMs and the many and varied tools or pipelines for computing, filtering, and annotating PSMs mean that many of the parameters used to conduct the peptide identification analysis are not available in the provided PSM output. As indicated earlier, the intent is to accurately describe the PSMs without necessarily describing the method used to obtain them. For specific details on the experimental protocols and workflows used, interested parties are referred to the relevant protocol and meta-data resources available at the DCC. We do populate a list of post-translational modifications used, as a variable SearchModification, in the SpectrumIdentificationProtocol element and document the filtering threshold criterion in the Threshold element. The Enzyme element is provided, as required, but indicates the PSI-MS term NoEnzyme (CPTAC-DCC:mzIdentML version 1.0) or unspecific cleavage (CPTAC-DCC:mzIdentML version 1.1) as required when the proteolytic agent is not known. These values are intended only to ensure that the provided PSMs are not excluded and placate down-stream mzIdentML software that expect XML elements typically populated in search engines' results.

Document Version History

Version 1.0 (May 6, 2013)
Initial CPTAC DCC mzIdentML format methods document.

Version 1.0.1 (June 10, 2013)
Revision based on internal feedback.

- Major change – correction of URLs for downloading RefSeq data.
- Minor language tweaks to clarify whether NIST, PCCs, the proteomics community, or DCC are responsible for, or proposing particular elements of the method.

Version 1.0.2 (August 13, 2013)
Minor edits and formatting in preparation for initial release of PSMs.

Version 1.1.0 (Jan 6, 2013)
Changes in support of name change for central analysis pipeline to common data analysis pipeline. Description of the use of

Format Version History

Version 1.0.0 (May 6, 2013)
Initial CPTAC DCC mzIdentML format release.

Version 1.1.0 (Jan 6, 2014)
Change score/parameter prefix to CPTAC-CDAP. Update version numbers of analysis software elements. The length attribute for DBSequence element is now populated. Use of the "unspecific cleavage" PSI-MS

term for Enzyme element, in place of NoEnzyme, as NoEnzyme is deprecated. The location attribute of the SpectralData element is now populated to refer to the accompanying mzML file. The FileFormat element of the SpectralData element is now populated with “mzML file” term, rather than “Thermo RAW file” or similar.