A Description of the CPTAC Common Data Analysis Pipeline (CDAP)

v. 02/25/2014

Summary

The purpose of this document is to describe the software programs and output files of the Common Data Analysis Pipeline (CDAP) run at NIST for the Clinical Proteomics Tumor Analysis Consortium (CPTAC). The pipeline is meant to produce peptide-level and protein-level reports from which differential analysis can be performed. It was designed to be compatible with both 'label-free' and 4plex iTRAQ™ workflows. The pipeline components include programs to extract spectra from RAW data files, interpret the MS2 spectra by database searching, localize phosphosites, and report peptides and their quantification linked to their scan numbers. Protein and gene assignment programs are in development. Production of a common analysis is intended to reduce the variability inherent in comparing result sets from different data analysis pipelines.

Authors

Paul A. Rudnick, Yuri A. Mirokhin, Sanford P. Markey, and Stephen E. Stein*
NIST
Mass Spectrometry Data Center
Biomolecular Measurement Division
Material Measurement Laboratory

* Contact: steve.stein@nist.gov

Table of Contents

Conventions used in this document

`This font is used to highlight a software program, command-line options, or to display the contents of a file.`

Bold is used to highlight a program name, file type or data file format.

Major output files

.psm – Peptide spectrum match reports in tab-delimited text

.mzid – MzIdentML files produced from the .psm files at the DCC

Step 1 – Data file QC

Files are retrieved from the CPTAC Data Coordinating Center (DCC) and checksums verified.

Step 2 – Spectrum Extractions

`ReAdW4Mascot2.exe` **(ftp://chemdata.nist.gov/download/peptide_library/software/current_releases/ReAdw4Mascot2/)** was used to convert Thermo Scientific™ mass spectrometry files to **MGF** and **mzXML** formats for MS/MS searching and MS1 data analysis, respectively. It makes use of XCalibur's software libraries if they are installed, otherwise it will attempt to use **MSFileReader** from Thermo Scientific™ (http://sjsupport.thermofinnigan.com/public/detail.asp?id=703).

Command-line Options for `ReAdW4Mascot2.exe`

For LTQ data, the following command-line is used:

```
"-sep1 -NoPeaks1 -MaxPI -metadata -PIvsRT -c -sepZC -xcal -xpw 10 -xpm
32 -XmlOrbiMs1Profile -c <file>.RAW <out directory>"
```

For Orbitrap and QExactive data files, the following options are added:

```
"-ChargeMgfOrbi -MonoisoMgfOrbi -FixPepmass"
```

For iTRAQ4plex, add:

```
"-iTRAQ "
```

For iTRAQ8plex, add:

```
"-iTRAQ8 "
```

Both **MGF** and **mzXML** files are for internal use only but are available by request.

`ReAdW4Mascot2.exe` also extracts iTRAQ4plex or iTRAQ8plex reporter ion values and reports them in the TITLE line of the **MGF** files. A value for variability of each iTRAQ channel is also given in these field as the dMZ/HWHM where dMZ = (m/z) / ExpectedResolution(at m/z). A value >1 typically indicates isolation window contamination. These values can be used to impose penalties on identified spectra with abundant impurities. AbFract is also calculated; this is the fraction of the MS2 TIC accounted for by the reporter ions. All iTRAQ values are copied in the PSM reports.

Step 3 – Identification of Peptide MS/MS Spectra by Database Search

MSGF+ (https://bix-lab.ucsd.edu/pages/viewpage.action?pageId=13533355) is used to identify peptides from protein sequences. This program was formerly named **MS-GFDB** and was developed at the University of California San Diego (http://www.ncbi.nlm.nih.gov/pubmed/20829449) (Kim S. *et al*, Mol Cell Proteomics, 2010, 12: 2840-52). Its development continues at PNNL by Sangtae Kim.

Command-line Options for MS-GF+

For Orbi/HCD and QExactive data, the options are the following:

```
"java –Xmx3500M –jar MSGFPlus.jar -d <file>.fasta -t 20ppm -e 1 -m 3 -
inst 1 -ntt 1 -thread 2 -tda 1 -ti 0,1 -n 1 -maxLength 50 -mod
<file>.txt"
```

For phospho, add:

```
"-protocol 1"
```

For iTRAQ, add:

```
"-protocol 2"
```

For iTRAQ and phospho, add:

```
"-protocol 3"
```

For Orbi/CID, change:

```
"-m 3" to "-m 1"
```

For Q-Exactive , change:

```
"-inst 1" to "-inst 3"
```

The contents of the **mods.txt** files are the following:

```
"NumMods=3
C2H3N1O1,C,fix,any,Carbamidomethyl
O1,M,opt,any,Oxidation
H-1N-1O1,NQ,opt,any,Deamidated
H-2O-1,E,opt,N-term,Pyro_glu
H-3N-1,Q,opt,N-term,Pyro-glu"
```

For phospho, add:

```
"HO3P,STY,opt,any,Phospho"
```

For iTRAQ4plex, replace:

```
"H-2O-1,E,opt,N-term,Pyro_glu
H-3N-1,Q,opt,N-term,Pyro-glu"
```

With:

```
"144.102063,*,fix,N-term,iTRAQ4plex"
```

And add:

```
144.102063,K,fix,any,iTRAQ4plex"
```

Reference FASTA Files

The protein FASTA file used for CompRef analysis is concatenated **RefSeq** *H. sapiens* (build 37), *M. musculus* (build 37), and the sequence for *S. scrofa* (porcine) trypsinogen. The FASTA file used for analysis of the TCGA human samples lacks the *M. musculus* sequences.

MS-GF+ Results Files

MS-GF+ results are produced in **(mzIdentML) mzid** format and are converted to tab-delimited text (**TSV**) using the following command-line options:

```
"java -Xmx3500M -cp  MSGFPlus.jar edu.ucsd.msjava.ui.MzIDToTsv -i
<file>.mzid -o <file>.tsv -showDecoy 1"
```

These **.mzid** and **.tsv** files are for internal use only. A new **.mzid** file is produced from the final **.psm** (see below) files at the DCC.

Step4 – MS1 Data Analysis Using ProMS

ProMS is a peak processing program developed at NIST used to calculate precursor peak areas from extracted ion chromatogram, peak widths, and other features of the MS1 data. **ProMS** reads **mzXML**

files, the previous standard for peak lists developed at the Institute for Systems Biology (ISB), which has been replaced by **mzML**.

ProMS expects to find a **proms.ini** file in the directory containing the **mzXML** files to be processed. A proms.ini file should contain the following:

```
"<mzXML file>.raw.mzXML
<search result file>.raw.FT.hcd.ch.MGF.mzid.tsv
<output file>.raw.txt
<search engine name: MSGF+, MSPepSearch, SpectraST, OMSSA)
<instrument: ORBI_HCD, ORBI, LTQ, QTOF>"
```

The program is run on each **mzXML** file separately to produce **txt** file reports. These reports are essential for calculating peptide ion abundances and for many QC calculations performed by **nistms_metrics.exe**.

Step 5 – Calculation of QC Metrics

For the purposes of the CDAP, no QC metrics are included in the reports. However, performance metrics, as described in Rudnick et al MCP 2010 Feb;9(2):225-41, are used internally for the purposes of quality control.

Step 6 – Phosphosite Localization by PhosphoRS

PhosphoRS (http://www.ncbi.nlm.nih.gov/pubmed/22073976) (Taus T., *et al*, J Proteome Res, 2011 12:5354-62) is used to score **MSGF+** assignments for site localization. These scores are embedded into the peptide sequence in the 'PhosphoRS' field of the **.psm** files. PhosphoRS scores >0.99 are used to confidently assign a localization. If all phsophosites for a PSM are confidently assigned, 'FullyLocalized' in the .psm files is set to'Y.'

Psm2Xml.exe – produces an input file for ProsphoRS reading scan numbers and sequences from the .**psm** file and corresponding spectrum from the **mgf** file.

```
Psm2Xml <input psm file>.<input mgf file> <output xml file>
```

PhosphoRS command-line:

```
java –jar phosphoRS.jar <input xml produced by Psm2Xml> <output xml>
```

Add_phospho_psm.exe – reads site probabilities from PhosphoRS output xml file and produces an updated **psm** file.

```
Add_phospho_psm <input original psm file> <input xml produced by
PhosphoRS> <output updated psm file>
```

These programs are available by request.

Report Files

This section describes the contents of the reports uploaded to the CPTAC Data Coordinating Center (DCC).

Structure of Peptide Spectrum Match Files (.psm)

The .psm files are given in tab-delimited text with Windows-style line-returns. The data are pre-filtered to a q-value level (FDR) of <0.01. **Warning:** q-values in files with <500 identifications may not be accurate.

The fields in **purple** below are copied directly from the search MSGF+ result files.

The fields in **orange** are for iTRAQ reports only.

The fields in **green** are for phosphopeptide reports only.

`FileName`
: Raw file name

`ScanNum`
: Thermo MS2 scan number

`QueryPrecursorMz`
: Precursor m/z used at search time (PEPMASS value in **MGF** file). This may be different than that in the **RAW** file.

`OriginalPrecursorMz`
: Precursor m/z recorded in the RAW file for this MS/MS spectrum.

`PrecursorError(ppm)`
: This is calculated as the difference between the theoretical and measured m/z in parts per million.

`QueryCharge`
: Precursor charge as listed in the peak list file.

`OriginalCharge`
: Precursor charge as calculated by XCalibur. This may be different than QueryCharge if **ReadW4Mascot.exe** disagrees (rare).

`PrecursorScanNum`
: Scan number of the previous MS1 scan. A "?"in this field denotes inferred precursor scan number.

`PrecursorArea`

This is the precursor area in ion counts as calculated by **ProMS** from the extracted ion chromatogram of the precursor. These values can be used for relative quantitation.

`PrecursorRelAb`

This value is calculated by **ProMS** as the fraction of the total ion count (TIC) accounted for by this precursor ion. This may also be useful for label-free quantitation.

`RTAtPrecursorHalfElution`

This is the retention time at which half of the precursor ion has eluted according to **ProMS**.

`PeptideSequence`

Peptide sequence annotated with modifications (by mass shift) from **MSGF+**

`AmbiguousMatch`

This field is marked 'Y' if multiple top-ranked hits are present. In these cases, multiple matches are listed for a single MS2 scan.

`Protein`

Accession number(s) for all protein matches attributable to the corresponding peptide sequence given in the PeptideSequence field

`DeNovoScore`

See MSGF+ documentation and publications

`MSGFScore`

See MSGF+ documentation and publications

`Evalue`

See MSGF+ documentation and publications

`Qvalue`

See MSGF+ documentation and publications. This value is used to filter the results in the .psm files to hits with QValues < 0.01 (or 1% FDR) per file.

`PepQvalue`

See MSGF+ documentation and publications.

`PrecursorPurity`

This field reports the precursor purity values for the isolation window in the previous and next MS1 scans. These values are useful for monitoring contamination in iTRAQ experiments.

`FractionDecomposition`

1 – the fraction of remaining precursor intensity in the MS2 spectrum. Useful for assessing how well the precursor was fragmented at this collision energy level (HCD only).

`HCDEnergy`

The collision energy applied to fragment the precursor to derive this scan. This is not the normalized values set by the instrument operator by the actual energy applied at fragmentation time.

`iTRAQ114`

Actual abundance of the 114 channel followed by a quality score dMZ/HWHM = dMZ / (MZ / Re) where dMZ is the deviation, MZ is the theoretical MZ for the reporter ion and Re is the expected resolution at 400 (reported for the scan). Values above 0.5 usually indicate a major problem with peak finding.

`iTRAQ115`

Actual abundance of the 115 channel followed by a quality score dMZ/HWHM = dMZ / (MZ / Re) where dMZ is the deviation, MZ is the theoretical MZ for the reporter ion and Re is the expected resolution at 400 (reported for the scan). Values above 0.5 usually indicate a major problem with peak finding.

`iTRAQ116`

Actual abundance of the 116 channel followed by a quality score dMZ/HWHM = dMZ / (MZ / Re) where dMZ is the deviation, MZ is the theoretical MZ for the reporter ion and Re is the expected

resolution at 400 (reported for the scan). Values above 0.5 usually indicate a major problem with peak finding.

iTRAQ117

Actual abundance of the 117 channel followed by a quality score dMZ/HWHM = dMZ / (MZ / Re) where dMZ is the deviation, MZ is the theoretical MZ for the reporter ion and Re is the expected resolution at 400 (reported for the scan). Values above 0.5 usually indicate a major problem with peak finding.

iTRAQFlags

'I' if the geometric mean of the two 'PrecursorPurity' values is <90%. (It is known that this threshold is too strict. Empirical data suggest that only PrecursorPurity values <80% may be subject to "compression"

'M' if one or more iTRAQ channels has a zero value

'D' if the quality score for one or more of the iTRAQ channels is > 1

iTRAQTotalAb

Sum of the ion current for the iTRAQ peaks

iTRAQFractionOfTotalAb

iTRAQTotalAb divided by the TIC for the MS2 spectrum

PhosphoRSPeptide

Peptide sequence annotated with PhosphoRS probability scores. Scores for any site > 99.0 should be considered localized. It is possible to only localize one of several sites. This peptide annotation will frequently be different than that given in PeptideSequence by **MSGF+.**

nPhospho

The number of phophosites expected by the precursor m/z

FullyLocalized

'Y' if all phosphosites score >99.0, otherwise 'N'