Portable Proteomics Pipeline (P3) MSGF Benchmark

Daniel Kristiyanto (daniel.kristiyanto@pnnl.gov)

May 26, 2016

1. Files:

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML"
```

```
## [1] "ID_003632_9011437E.fasta"
```

2. Identification:

`MSGFPlus.jar` was downloaded from https://omics.pnl.gov/software/ms-gf. Identification was run multiple times (within or without Docker), and resulted the same results.

Identification with different switches resulted different results.

```
# Command Line (Without Switch)
java -Xmx8000M -jar P3/MSGFPlus.jar -s \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML -o \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid \
-d ID_003632_9011437E.fasta

# Command Line (2)
java -Xmx8000M -jar P3/MSGFPlus.jar -s \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML -o \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzid \
-d ID_003632_9011437E.fasta  \
-t 10ppm -m 0 -inst 1 -e 1 -ti -1,2 -ntt 2 -tda 1 -minLength 6 \
-maxLength 50 -minCharge 2 -maxCharge 5 -n 1 -thread 7 \
-mod MSGFDB_Mods.txt -minNumPeaks 5 -addFeatures 1
```

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid"
## [2] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzid"
```

3. Result from PNNL

```
## [1] 12695
```

```
##    Scan MSGFDB_SpecEValue                              Peptide
## 1 11612      1.481722e-36 IQDLTGLDVSTAEELQVANYGVGGQYEPHFDFAR
## 2 14441      8.834771e-36  LTPFIIQENLNLALNSASAIGCHVVNIGAEDLR
##       Scan MSGFDB_SpecEValue Peptide
## 12947 3823        0.04856372  SPAGGG
## 12948 4099        0.05022644  PGGAGM
```

4. Result from MSGF+ without any switches (sorted by SpecEValue)

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue
## X9740.1           10815               277     4.434865e-30
## X8340.1            9275               234     2.157325e-27
## X10711.1          11883               157     1.386474e-25
##                                  pepseq
## X9740.1  SHSTEPGLVLTLGQGDVGQLGLGENVMER
## X8340.1          DLYANTVLSGGTTMYPGIADR
## X10711.1      ILGGVISAISEAAAQYNPEPPPPR
```

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue
## X13036.1          14441               381     4.861374e-08
## X10464.1          11612               279     5.246882e-08
##                                        pepseq
## X13036.1     KETDLKQIQTLIQGTQTRLKYSQNELEMIKK
## X10464.1 AGISEAQLTDAETSKLIYDFIEDQGGLEAVRQEMR
```

Reading the MZID file manually

```
java -Xmx2000M -XX:+UseConcMarkSweepGC -cp \
../P3/MSGFPlus.jar edu.ucsd.msjava.ui.MzIDToTsv -i \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid \
-showQValue 1 -showDecoy 1 -unroll 1
```

```
##   ScanNum                           Peptide            Protein SpecEValue
## 1   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041660.1   4.43e-30
## 2   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041659.1   4.43e-30
## 3   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041664.1   4.43e-30
```

5. Results with switches

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue
## X11197.1          12418               321     6.104450e-34
## X9050.1           10056               258     4.553317e-33
## X9060.1           10067               288     9.340615e-33
##                                 pepseq
## X11197.1 VYLASLETLDNGKPFQESYALDLDEVIK
## X9050.1     CLCLPSYVGALCEQDTETCDYGWHK
## X9060.1     CLCLPSYVGALCEQDTETCDYGWHK
```

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue       accession
## X13036.1          14441               345     2.196462e-06 ref|NP_060250.2
## X10464.1          11612               279     2.280634e-06 ref|NP_000487.1
##                                    pepseq
## X13036.1 FILPNVSTPVSDAFKTQMELLQAGLSRTPTR
## X10464.1  GDYKDSSDFGAPHPQVQSVRRIRDMQWHQR
```

Reading the MZID file manually

7. Filtering (PNNL -> Docker)

a. Spec-Evalue 10^{-10}

From PNNL

There are 4633 rows remained.

From Docker

There are 1665 rows remained.

```
## [1] "==Head=="
```

```
##       Scan MSGFDB_SpecEValue                      Peptide ms-gf:specevalue
## 3641 12418      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK     6.104450e-34
## 2717 10056      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK     4.553317e-33
##                            pepseq
## 3641 VYLASLETLDNGKPFQESYALDLDEVIK
## 2717    CLCLPSYVGALCEQDTETCDYGWHK
```

```
## [1] "==Tail=="
```

```
##      Scan MSGFDB_SpecEValue           Peptide ms-gf:specevalue
## 2508 9513      2.871564e-13   NIIGSSPVADFSAIK     8.381110e-05
## 223  1927      5.243567e-25 CCYDGACVNNDETCEQR     9.379988e-05
##                    pepseq
## 2508 LVDETEGQCGETDPNSEMPR
## 223       HECCFRYHCTGCCFR
```

Differences

Data from Docker is matched with from PNNL using scan number as ID. MSGFDB_SpecEValue, Peptide are values from PNNL. ms-gf:specevalue, pepseq are the corresponding values from Docker Container. There are 4621 of rows in total with 1227 exact match.

325 with the same peptide identification and different spec-evalue, 3069 of different peptide identification as well different spec-evalues (12 of them are unmatched values).

```
## [1] "==Unmatched=="
```

```
##      Scan MSGFDB_SpecEValue       Peptide ms-gf:specevalue pepseq
## 2291 8856      1.119096e-15 CSEGVFLLTTTPR               NA   <NA>
## 1416 6555      6.590530e-15  CQSLQEELDFRK               NA   <NA>
## [1] "==Different Spec eValues=="
```

```
##       Scan MSGFDB_SpecEValue                          Peptide
## 3127 11078      5.495878e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 2548  9602      2.682754e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
##      ms-gf:specevalue                           pepseq
## 3127     7.059070e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 2548     1.788561e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
```

```
## [1] "==Different Peptides identification and different Spec Evalues=="
```

```
##       Scan MSGFDB_SpecEValue                 Peptide ms-gf:specevalue
## 2595  9730      6.005831e-22     FFDHSGTLVMDAYEPEISR     2.231272e-20
## 4204 13789      4.811341e-14 NAQMAQSPILLLGGAASTLLQNR     6.118720e-19
##                      pepseq
## 2595 DLYANTVLSGGTTMYPGIADR
## 4204  IQTQLNLIHPDIFPLLTSFR
```

b. Spec-Evalue 10^{-15}

From PNNL

There are 2511 rows remained.

From Docker

There are 767 rows remained.

```
## [1] "==Head=="
```

```
##       Scan MSGFDB_SpecEValue                      Peptide ms-gf:specevalue
## 1763 12418      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK     6.104450e-34
## 1170 10056      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK     4.553317e-33
##                            pepseq
## 1763 VYLASLETLDNGKPFQESYALDLDEVIK
## 1170    CLCLPSYVGALCEQDTETCDYGWHK
```

```
## [1] "==Tail=="
```

```
##    Scan MSGFDB_SpecEValue           Peptide ms-gf:specevalue
## 49 1876      1.449447e-17    QNEAAVGNCAEHMR     3.858078e-05
## 50 1927      5.243567e-25 CCYDGACVNNDETCEQR     9.379988e-05
##             pepseq
## 49    DSFMRHRGHCER
## 50 HECCFRYHCTGCCFR
```

Differences

Data from Docker is matched with from PNNL using scan number as ID. MSGFDB_SpecEValue, Peptide are values from PNNL. ms-gf:specevalue, pepseq are the corresponding values from Docker Container. There are 2511 of rows in total with 572 exact match.

170 with the same peptide identification and different spec-evalue, 1769 of different peptide identification as well different spec-evalues (0 of them are unmatched values).

```
## [1] "==Unmatched=="
```

```
## [1] Scan              MSGFDB_SpecEValue Peptide           ms-gf:specevalue
## [5] pepseq
## <0 rows> (or 0-length row.names)
```

```
## [1] "==Different Spec eValues=="
```

```
##       Scan MSGFDB_SpecEValue                          Peptide
## 1437 11078      5.495878e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 1057  9602      2.682754e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
##      ms-gf:specevalue                           pepseq
## 1437     7.059070e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 1057     1.788561e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
```

```
## [1] "==Different Peptides identification and different Spec Evalues=="
```

```
##      Scan MSGFDB_SpecEValue             Peptide ms-gf:specevalue
## 1091 9730      6.005831e-22 FFDHSGTLVMDAYEPEISR     2.231272e-20
## 123  3378      2.191136e-20   ASGNYATVISHNPETKK     2.189761e-14
##                      pepseq
## 1091 DLYANTVLSGGTTMYPGIADR
## 123     NKPGPYSSVPPPSAPPPK
```

c. Spec-Evalue 10^{-20}

From PNNL

There are 936 rows remained.

From Docker

There are 223 rows remained.

```
## [1] "==Head=="
```

```
##      Scan MSGFDB_SpecEValue                      Peptide ms-gf:specevalue
## 571 12418      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK     6.104450e-34
## 327 10056      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK     4.553317e-33
##                            pepseq
## 571 VYLASLETLDNGKPFQESYALDLDEVIK
## 327    CLCLPSYVGALCEQDTETCDYGWHK
## [1] "==Tail=="
```

```
##    Scan MSGFDB_SpecEValue             Peptide ms-gf:specevalue
## 19 3273      1.342643e-21 ESCNPQDCVPEVDENCKDK     2.671039e-05
## 12 1927      5.243567e-25   CCYDGACVNNDETCEQR     9.379988e-05
##                      pepseq
## 19 AEGAYYDNHSVVYLDHGENGEK
## 12        HECCFRYHCTGCCFR
```

Differences

Data from Docker is matched with from PNNL using scan number as ID. MSGFDB_SpecEValue, Peptide are values from PNNL. ms-gf:specevalue, pepseq are the corresponding values from Docker Container. There are 936 of rows in total with 165 exact match.

54 with the same peptide identification and different spec-evalue, 717 of different peptide identification as well different spec-evalues (0 of them are unmatched values).

```
## [1] "==Unmatched=="
```

```
## [1] Scan              MSGFDB_SpecEValue Peptide           ms-gf:specevalue
## [5] pepseq
## <0 rows> (or 0-length row.names)
```

```
## [1] "==Different Spec eValues=="
```

```
##      Scan MSGFDB_SpecEValue                          Peptide
## 450 11078      5.495878e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 284  9602      2.682754e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
##     ms-gf:specevalue                           pepseq
## 450     7.059070e-31        MGLDPACQVDIICGDHLLEQYQTLR
## 284     1.788561e-30 WSGPLSLQEVDEQPQHPLHVTYAGAAVDELGK
```

```
## [1] "==Different Peptides identification and different Spec Evalues=="
```

```
##      Scan MSGFDB_SpecEValue              Peptide ms-gf:specevalue
## 299  9730      6.005831e-22  FFDHSGTLVMDAYEPEISR     2.231272e-20
## 401 10627      4.148874e-23 GMSLNLEPDNVGVVVFGNDK     7.877905e-10
##                      pepseq
## 299 DLYANTVLSGGTTMYPGIADR
## 401   GNAEPDPPVRLFQIHGNDK
```