

Introduction to *pepXMLTab*

Xiaojing Wang

April 7, 2020

Contents

1	Introduction	1
2	Convert pepXML to a tabular format	1
3	PSMs Filtering	4
4	Session Information	5

1 Introduction

Mass spectrometry (MS)-based proteomics technology is widely used in biological researches. MS/MS spectra generated by this technology are usually searched and assigned to peptides, and such assignments are typically described by a data format named pepXML developed at the SPC/Institute for systems biology. More detailed information about pepXML can be found at <http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML>. Recently a community standard format mzIdentML [1] has been defined by HUPO <http://www.psidev.info/mzidentml> and is set to replace pepXML. There is an existing R package *mzID*, which is designed to parse the mzIdentML file format.

As the first widely accepted data format and supported by many search engines, pepXML is still commonly used. Although this XML based format features a highly organized structure, it is less intuitive to human interpretation, thus converting it to human readable format is often desired. To this end, we developed this R package, *pepXMLTab*, which import the Peptide-Spectrum-Matches (PSMs) and related information from pepXML files and filter them based on user specified FDR threshold.

pepXMLTab has been tested using sample pepXML files generated from multiple search engines, MyriMatch [2], Mascot [3], X!Tandam [4] and SEQUEST [5]

2 Convert pepXML to a tabular format

In order to calculate FDR at the peptide level, *pepXMLTab* uses the function `pepXML2tab` to convert the 'spectrum_query' section of a pepXML file to a data frame. The structure of the output data frame is dependent on the input pepXML, with each column representing a section

of the information defined by the search engine. Different search engines use their own scoring method for PSMs. For instance, MyriMatch uses a sophisticated statistical scoring system. For each experiment spectrum, MyriMatch examines every m/z location and computes two probabilistic scores: an intensity-based MVH score and a mass error-based mz Fidelity score. In SEQUEST, a cross correlation score (XCrr) is used to represent an average of the differences between the m/z values in the observed and virtual spectrum. Please check the documents of each search engine for more details.

```
> tttt <- pepXML2tab(pepxml)
> tttt[1:2,]
```

	spectrum	start_scan	end_scan		
1	511_c402.2474_u402.2695_r41.13	0	0		
2	186_c660.3871_u662.3625_r58.85	0	0		
	precursor_neutral_mass	assumed_charge	index	hit_rank	peptide
1	401.240124	1	2	1	VAIGR
2	659.379824	1	3	1	VAIGR
	peptide_prev_aa	peptide_next_aa	protein	num_tot_proteins	
1	R	A	ECA0851	4	
2	R	A	ECA0851	4	
	num_matched_ions	calc_neutral_pep_mass	massdiff		
1	4	658.424789	+0.9550		
2	4	658.424789	+0.9550		
	num_missed_cleavages	is_rejected			
1	0	0			
2	0	0			
	protein_descr	protein_mw			
1	putative sugar ABC transporter ATP-binding protein	43111			
2	putative sugar ABC transporter ATP-binding protein	43111			
	ionscore	identityscore	homologyscore	expect	modification
1	4.73	28	15	12	NA
2	2.22	28	15	12	NA

```
> #SEQUEST example
> pepxml <- system.file("extdata/pepxml", "SEQUEST.pepXML", package="pepXMLTab")
> tttt <- pepXML2tab(pepxml)
> tttt[1:2,]
```

	spectrum	start_scan	end_scan		
1	mam_012808n_SW480_200ug_B05.02170.02170.2	2170	2170		
2	mam_012808n_SW480_200ug_B05.02170.02170.2	2170	2170		
	precursor_neutral_mass	assumed_charge	index	hit_rank	peptide
1	960.4736	2	1	1	LEELSDQK
2	960.4736	2	1	2	QNEVSEKK
	peptide_prev_aa	peptide_next_aa	protein	num_tot_proteins	
1	R	N	ENSP00000363435	1	
2	K	E	ENSP00000387188	2	
	num_matched_ions	tot_num_ions	calc_neutral_pep_mass	massdiff	
1	10	14	960.4764	-0.002790	
2	8	14	960.4876	-0.014020	
	num_tol_term	num_missed_cleavages	is_rejected	xcorr	deltacn
1	2	0	0	1.628	0.115
2	2	0	0	1.440	0.120
	deltacnstar	spscore	sprank	modification	

```

1      0.000 416.0000      7      NA
2      0.000 289.5000     52      NA

> #XTandem example
> pepxml <- system.file("extdata/pepxml", "XTandem.pepXML", package="pepXMLTab")
> tttt <- pepXML2tab(pepxml)
> tttt[1:2,]

      spectrum start_scan end_scan
1 mam_012808n_SW480_200ug_C10.00163.00163.3      163      163
2 mam_012808n_SW480_200ug_C10.00177.00177.3      177      177
  precursor_neutral_mass assumed_charge index retention_time_sec
1          1236.7882             3      1          702.929
2          1389.5482             3      2          710.847
  hit_rank      peptide peptide_prev_aa peptide_next_aa
1         1    LKSQPEPLVVK             E             G
2         1  NAEGEPVCNACGL             R             Y
  protein num_tot_proteins num_matched_ions tot_num_ions
1 ENSP00000357861             1             5          40
2 ENSP00000259090             4             4          48
  calc_neutral_pep_mass massdiff num_tol_term num_missed_cleavages
1          1236.7437     0.044             1             1
2          1389.5657    -0.018             1             0
  is_rejected hyperscore nextscore bscore yscore expect
1           0        12.6        12.6     8.7     9.3     19
2           0        15.2        14.5      0     12.8     12
  modification
1              NA
2 8;160.0306;11;160.0306

>

```

3 PSMs Filtering

After loading from the pepXML files, function `PSMfilter` was used to filter the PSMs based on score (defined by search engines), hit rank and peptide length. By default, `PSMfilter` selects the top ranking peptide hit with a minimum amino acid length of 6. The FDR estimation is based on decoy database matches. The calculation method is similar to what has been used in IDPicker2 [6]. All the peptides are separated into different peptide classes based on tryptic status and charge status. For each peptide class, PSMs were filtered based on user-specified FDR (Default is 0.01). PSMs that passed the FDR threshold in each class were then pooled together as output [6]. For example, considering the combination of three tryptic type (fully tryptic, semi tryptic and nontryptic) and three charge status (1+, 2+, 3+), all PSMs can be divided into 9 groups. In each group, we may keep the PSMs with FDR less than 0.01. The passed PSMs in each group are then pooled together as output.

```
> ## MyriMatch example
> pepxml <- system.file("extdata/pepxml", "Myrimatch.pepXML",
+   package="pepXMLTab")
> tttt <- pepXML2tab(pepxml)
> passed <- PSMfilter(tttt, pepFDR=0.01, scorecolumn='mvh', hitrank=1,
+   minpeplen=6, decoyprefix='rev_')
> passed[1, ]
```

```

                                spectrum
1 mam_121007n_RK0_200ug_IEF_A01.494.494.1
                                spectrumNativeID start_scan end_scan
1 controllerType=0 controllerNumber=1 scan=494          494      494
  precursor_neutral_mass assumed_charge index hit_rank peptide
1      622.22405861752          1    89      1  AMGNCA
  peptide_prev_aa peptide_next_aa      protein num_tot_proteins
1              K              - NP_004355.2              1
  calc_neutral_pep_mass      massdiff num_tol_term
1      622.685464 0.46140538248              2
  num_missed_cleavages num_matched_ions tot_num_ions
1              0              3              9
  number of matched peaks number of unmatched peaks      mvh
1              3              6 9.128796086702
                                mzFidelity      xcorr      modification NTT
1 12.780763189712 0.63469683320820336 5;160.0306484778      2

```

4 Session Information

```
R version 4.0.0 alpha (2020-04-05 r78150)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Mojave 10.14.6
```

```
Matrix products: default
```

```
BLAS:      /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK:    /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
```

```
locale:
```

```
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
attached base packages:
```

```
[1] stats      graphics grDevices utils      datasets  methods
[7] base
```

```
other attached packages:
```

```
[1] pepXMLTab_1.21.0
```

loaded via a namespace (and not attached):
[1] compiler_4.0.0 tools_4.0.0 XML_3.99-0.3

References

- [1] A. R. Jones, M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. J. Hubbard, J. N. Selley, B. C. Searle, J. Shofstahl, S. L. Seymour, R. Julian, P. A. Binz, E. W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J. A. Vizcaino, M. Chambers, A. Pizarro, and D. Creasy. The mzidentml data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*, 11(7): M111 014381, 2012.
- [2] D. L. Tabb, C. G. Fernando, and M. C. Chambers. Myrimatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*, 6(2): 654–61, 2007.
- [3] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, 1999.
- [4] R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, 2004.
- [5] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*, 5 (11):976–89, 1994.
- [6] Z. Q. Ma, S. Dasari, M. C. Chambers, M. D. Litton, S. M. Sobecki, L. J. Zimmerman, P. J. Halvey, B. Schilling, P. M. Drake, B. W. Gibson, and D. L. Tabb. Idpicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res*, 8 (8):3872–81, 2009.