My XTandem parser
Automated filtering and export of X!Tandem MS/MS results

Benoit Valot
valot@moulon.inra.fr
PAPPSO - http://pappso.inra.fr/

14 Avril 2010

Abstract

X!Tandem is an open-source software allowing peptide/protein identification from MS/MS mass spectra. X!Tandem is fast and accurate, but the Global Proteome Machine (GPM) is relatively limited regarding the processing of identification results.

My XTandem parser permit to filtered data according to statistical value at peptide and protein levels. The results are accessible to tabulated files (excel). Moreover, redundancy of protein database are fully filtered as :

- Proteins identified without specifics peptides compared to others are eliminated.
- Proteins identified with the same pool of peptides are assembled.
- Proteins are grouped by function (identified with at least one common peptide), and the specific peptides for each of the sub-group of proteins are indicated.

Contents

1 Installation

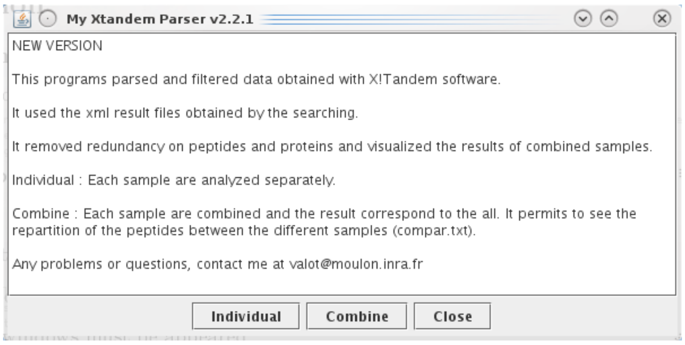
1.1 Requirements

My XTandem parser works on all plateforms (linux, windows). Java 1.6 must be installed : link

1.2 Start My XTandem parser

To run My XTandem parser, simply :

- Open My XTandem parser by using this link
- Allow the program to be executed
- The principal window will appear (Fig 1.2)



NEW VERSION

This programs parsed and filtered data obtained with X!Tandem software.

It used the xml result files obtained by the searching.

It removed redundancy on peptides and proteins and visualized the results of combined samples.

Individual : Each sample are analyzed separately.

Combine : Each sample are combined and the result correspond to the all. It permits to see the repartition of the peptides between the different samples (compar.txt).

Any problems or questions, contact me at valot@moulon.inra.fr

Figure 1: Principal window

1.3 License

2 Utilisation

Warning To perform analysis, My X!Tandem parser needs the have X!Tandem result files (.xml). The names of the files are used as **sample names**. To perform local X!Tandem identification in automated process, you can use our xtandem batch processing script available here.

2.1 Two modes

You can filter the MS/MS identifications and export the results in two different modes :

Individual mode
Each MS/MS result file is processed individually.
You cannot perform comparison by using this process.

Combined mode
The MS/MS result files are combined in one result file, and this file is filtered / exported.
This mode is usefull to compare different results.

In both modes, you have to :

1. Select the xml result files
2. Define the filter parameters (2.2)
3. Define the name of the result file to export
4. Define the export parameters (2.3)

2.2 Filter parameters

The filter window (Fig 2) defines the automated filter process :

Peptide E-value
Defines the E-value above which a peptide is considered as valid

Peptide number
Defines the number of valid unique[1] peptides necessary to validate a protein

Protein E-value
Defined the E-value above which a protein is considered as valid

- The protein E-values are re-calculated by the product of the valid unique peptides E-values, and are different from the protein E-values determined by X!Tandem.
- The values are expressed in log(E-value)

Sum to all
Defines how protein filter is performed when MS/MS results are combined :

No To validate a protein, the 2 parameters (peptide number and protein E-value) must be valid in at least one result. Interesting to compare 2DLC-MS/MS results, where peptides from a protein are in the same LC-MS/MS run

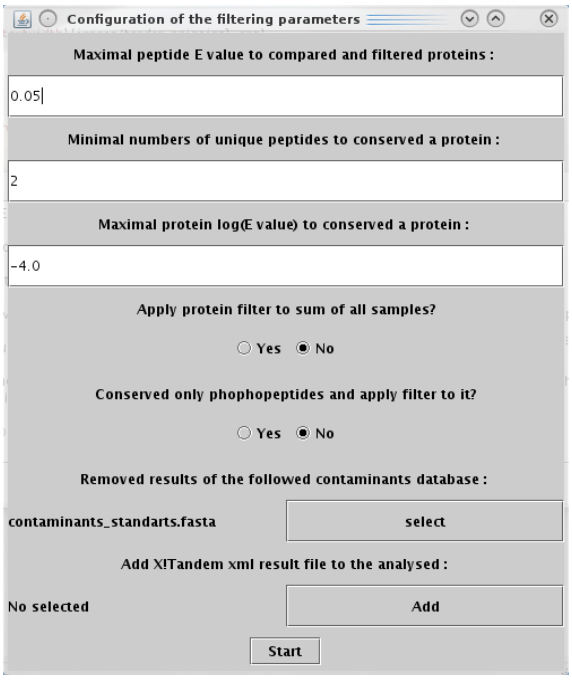
Yes To validate a protein, the 2 parameters (peptide number and protein E-value) must be valid in the sum of all results. Interesting to compare SDS-PAGE-LC-MS/MS results, where peptides from a protein are split in different LC-MS/MS runs

Phosphopeptide
Conserves only peptides containing phosphorylated residue modifications. All other peptides are invalided.

Contaminants
When you perform an analysis using different fasta databases, you can remove the result from one database by selecting this database. Interesting because it allows to always include the same contaminant proteins during the database search, and because it removes the contaminant proteins from the results.

Add results
At this stage, you can add other MS/MS result files to the analysis. If two files have the same name, they are combined in one result file. Interesting to combine X!Tandem results of the same LC-MS/MS run using different modification parameters or protein databases.

[1] Unique peptides are defined as peptides with different sequences. This excludes peptides with different modifications.



Configuration of the filtering parameters

Maximal peptide E value to compared and filtered proteins :

0.05

Minimal numbers of unique peptides to conserved a protein :

2

Maximal protein log(E value) to conserved a protein :

-4.0

Apply protein filter to sum of all samples?

○ Yes ● No

Conserved only phophopeptides and apply filter to it?

○ Yes ● No

Removed results of the followed contaminants database :

contaminants_standarts.fasta

Add X!Tandem xml result file to the analysed :

No selected

Figure 2: Filter window

2.3 Export parameters

The export window (Fig 3) shows the different types of available exports:

Default
Creates tabulated files containing identification results for proteins (*protein.txt) and peptides (*peptide.txt). When you perform a combined analysis, a *compar.txt file is created that contains the results of comparison between samples.

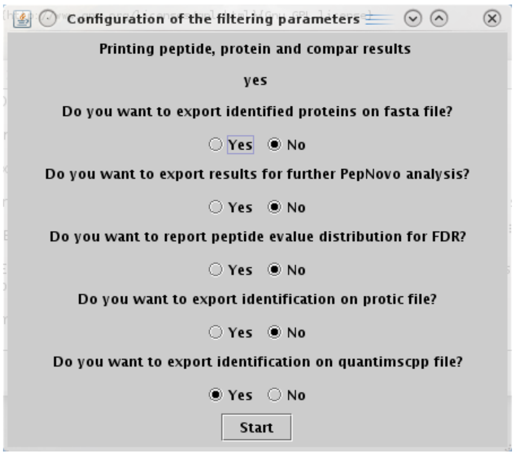
Fasta
Creates a fasta file for valid proteins.

PepNovo
Creates an xml file containing the peptide results to be removed for an automated *De Novo* interpretation in sequence using our pipeline.

FDR
Creates a tabulated file containing the number of valid peptides for the different peptide E-values in each database. Interesting to determine the E-value above which FDR value is acceptable.
Warning : Use very low parameters in peptide (0.1) and protein (-1) evalues, and set the number of unique peptides to validate a protein to 1.

Protic
Internal export in xml, defined to store results in a proteomic database called PROTICdb.

Quantimscpp
Internal export in xml, used as input to perform quantitative analysis using a home-made software 'Quantimscpp' currently in development.



Configuration of the filtering parameters

Printing peptide, protein and compar results

yes

Do you want to export identified proteins on fasta file?

○ Yes ● No

Do you want to export results for further PepNovo analysis?

○ Yes ● No

Do you want to report peptide evalue distribution for FDR?

○ Yes ● No

Do you want to export identification on protic file?

○ Yes ● No

Do you want to export identification on quantimscpp file?

● Yes ○ No

Figure 3: Export window

3 Export results

3.1 *protein.txt

All proteins identified are presented by sample by sample (MS/MS file, Fig 4). Proteins are grouped by function.

Group Group to which the protein belongs. All the proteins in a group have at least one peptide in common.

Sub-group Sub-group to which the protein belongs. All the proteins in a sub-group are identified with the same valid peptides.

Description Protein description as it is in the header of the fasta file.

log(E value) Protein E-value expressed in log.

- Statistical value representing the number of times this protein would be identified just by random.
- Calculated as the product of unique peptide E-values in the sample.

Coverage % of protein coverage.

MW Molecular weight of the protein expressed in KDa.

Spectra Total number of MS/MS spectra identified for the protein

Specifics Number of MS/MS spectra that are specific to the protein, compared to the other proteins of the same group.

Uniques Number of unique peptide sequences identified for the protein.

PAI Protein Abundance Index

- PAI estimates the relative abundance of the protein.
- PAI is calculated as the number of identified spectra divided by the number of theoretical peptides[2] of the protein.

Redundancy Number of proteins identified with the same pool of spectra. When there is redundancy, the above described parameters are shown only for the first protein of the subgroup (arbitrary chosen).Only the description of the other members of the subgroup is shown.

3.2 *peptide.txt

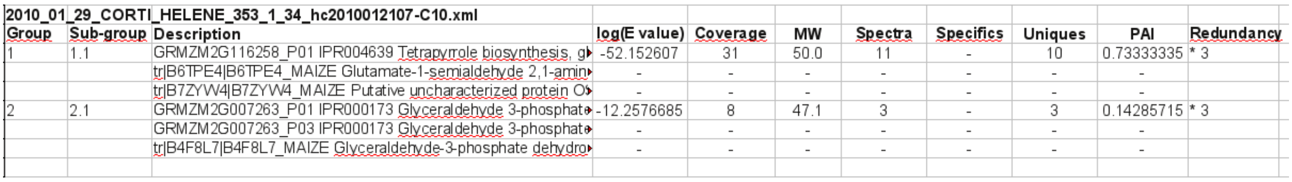
Identified peptides are grouped by group (Fig 5). One line corresponds to 1 MS/MS spectrum identifying one peptide, that can be present in one or more proteins.

FDR False discovery rate at peptide level, calculated by the peptide E-value of the complete analysis.

Group Group of the proteins containing this peptide.

Description Protein description if the peptide is specific to this protein.

[2] Theoretical peptides correspond to the peptides resulting from the theoretical digestion of the protein sequence by trypsin and that are visible in mass spectrometry ($800 < MH < 2500$)



2010_01_29_CORTI_HELENE_353_1_34_hc2010012107-C10.xml										
Group	Sub-group	Description	log(E value)	Coverage	MW	Spectra	Specifics	Uniques	PAI	Redundancy
1	1.1	GRMZM2G116258_P01 IPR004639 Tetrapyrrole biosynthesis, gl	-52.152607	31	50.0	11	-	10	0.73333335	* 3
		tr|B6TPE4|B6TPE4_MAIZE Glutamate-1-semialdehyde 2,1-amin	-	-	-	-	-	-	-	
		tr|B7ZYW4|B7ZYW4_MAIZE Putative uncharacterized protein O	-	-	-	-	-	-	-	
2	2.1	GRMZM2G007263_P01 IPR000173 Glyceraldehyde 3-phosphate	-12.2576685	8	47.1	3	-	3	0.14285715	* 3
		GRMZM2G007263_P03 IPR000173 Glyceraldehyde 3-phosphate	-	-	-	-	-	-	-	
		tr|B4F8L7|B4F8L7_MAIZE Glyceraldehyde-3-phosphate dehydro	-	-	-	-	-	-	-	

Figure 4: Protein results

Sample Name of MS/MS run file.

Scan Scan number of the MS/MS run analysis.

Rt Retention time of the peptide.

Sequence Sequence of the peptide.

Modifs Modifications on the peptide. [3]

Valid Indicates whether the peptide was validated by the filter parameters or not.

Used Number of protein sub-groups in which the peptide is present.

on a total of Total number of protein sub-groups in the group.
Rq : If the peptide is specific, there is only ′−′.

Sub-groups Protein sub-groups where the peptide is present.

E-value Peptide E-value.

- Statistical value representing the number of times this peptide would be identified at random.
- Calculated by the X!Tandem with an empiric model.

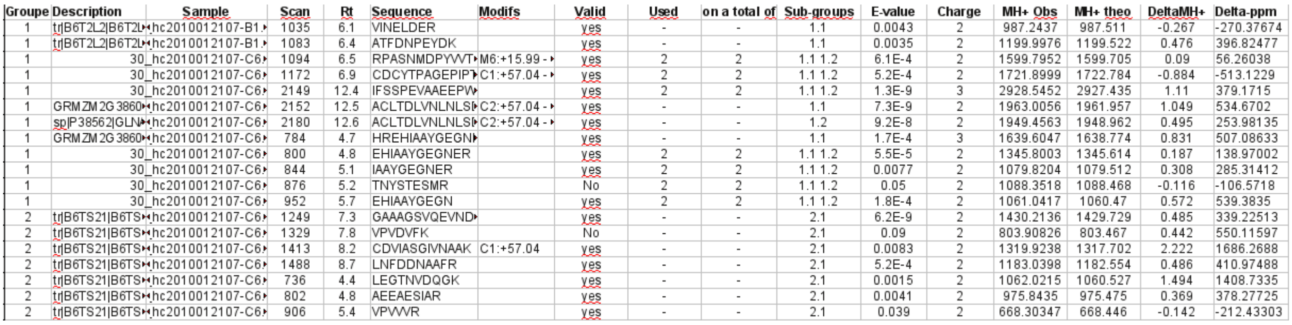
Charge Charge level of the precursor.

MH+ Obs Monoisotopic observed mass for the peptide + one proton (MH^+)

MH+ Theo Monoisotopic calculated mass for the peptide + one proton (MH^+)

DeltaMH+ Error in the precursor mass between observed and theoretical data (Da)

Delta-ppm Error in the precursor mass between observed and theoretical data (ppm)



Groupe	Description	Sample	Scan	Rt	Sequence	Modifs	Valid	Used	on a total of	Sub-groups	E-value	Charge	MH+ Obs	MH+ theo	DeltaMH+	Delta-ppm
1	tr|B6T2L2|B6T2L	hc2010012107-B1.	1035	6.1	VINELDER		yes	-	-	1.1	0.0043	2	987.2437	987.511	-0.267	-270.37674
1	tr|B6T2L2|B6T2L	hc2010012107-B1.	1083	6.4	ATFDNPEYDK		yes	-	-	1.1	0.0035	2	1199.9976	1199.522	0.476	396.82477
1		30_hc2010012107-C6	1094	6.5	RPASNMDPYWT	M6:+15.99 -	yes	2	2	1.1 1.2	6.1E-4	2	1599.7952	1599.705	0.09	56.26038
1		30_hc2010012107-C6	1172	6.9	CDCYTPAGEPIP	C1:+57.04 -	yes	2	2	1.1 1.2	5.2E-4	2	1721.8999	1722.784	-0.884	-513.1229
1		30_hc2010012107-C6	2149	12.4	IFSSPEVAAEEPW		yes	2	2	1.1 1.2	1.3E-9	3	2928.5452	2927.435	1.11	379.1715
1	GRMZM2G3860	hc2010012107-C6	2152	12.5	ACLTDLVNLNLSI	C2:+57.04 -	yes	-	-	1.1	7.3E-9	2	1963.0056	1961.957	1.049	534.6702
1	sp|P38562|GLN	hc2010012107-C6	2180	12.6	ACLTDLVNLNLSI	C2:+57.04 -	yes	-	-	1.2	9.2E-8	2	1949.4563	1948.962	0.495	253.98135
1	GRMZM2G3860	hc2010012107-C6	784	4.7	HREHIAAYGEGN		yes	-	-	1.1	1.7E-4	3	1639.6047	1638.774	0.831	507.08633
1		30_hc2010012107-C6	800	4.8	EHIAAYGEGNER		yes	2	2	1.1 1.2	5.5E-5	2	1345.8003	1345.614	0.187	138.97002
1		30_hc2010012107-C6	844	5.1	IAAYGEGNER		yes	2	2	1.1 1.2	0.0077	2	1079.8204	1079.512	0.308	285.31412
1		30_hc2010012107-C6	876	5.2	TNYSTESMR		No	2	2	1.1 1.2	0.05	2	1088.3518	1088.468	-0.116	-106.5718
1		30_hc2010012107-C6	952	5.7	EHIAAYGEGN		yes	2	2	1.1 1.2	1.8E-4	2	1061.0417	1060.47	0.572	539.3835
2	tr|B6TS21|B6TS	hc2010012107-C6	1249	7.3	GAAAGSVQEVND		yes	-	-	2.1	6.2E-9	2	1430.2136	1429.729	0.485	339.22513
2	tr|B6TS21|B6TS	hc2010012107-C6	1329	7.8	VPVDVFK		No	-	-	2.1	0.09	2	803.90826	803.467	0.442	550.11597
2	tr|B6TS21|B6TS	hc2010012107-C6	1413	8.2	CDVIASGIVNAAK	C1:+57.04	yes	-	-	2.1	0.0083	2	1319.9238	1317.702	2.222	1686.2688
2	tr|B6TS21|B6TS	hc2010012107-C6	1488	8.7	LNFDDNAAFR		yes	-	-	2.1	5.2E-4	2	1183.0398	1182.554	0.486	410.97488
2	tr|B6TS21|B6TS	hc2010012107-C6	736	4.4	LEGTNVDQGK		yes	-	-	2.1	0.0015	2	1062.0215	1060.527	1.494	1408.7335
2	tr|B6TS21|B6TS	hc2010012107-C6	802	4.8	AEEAESIAR		yes	-	-	2.1	0.0041	2	975.8435	975.475	0.369	378.27725
2	tr|B6TS21|B6TS	hc2010012107-C6	906	5.4	VPVWR		yes	-	-	2.1	0.039	2	668.30347	668.446	-0.142	-212.43303

Figure 5: Peptide results

3.3 *compar.txt

All identified proteins are presented in a list: one protein per protein, and one column per sample, i.e. per MS/MS run file (Fig 6). The list of proteins is repeated 4 times, corresponding to the 4 parameters that are used to compare samples (see Type for details).

Group Protein group. Groups roughly correspond to the different functions.

Sub-group Protein sub-group. All the proteins of a sub-group are identified with the same valid peptides.

[3] For example, M2:+15.99 means that the mass of the second amino acid, which is a methionine, is increased by 15.99. This mass increase indicates that the peptide is oxidized.

Description Protein description extracted from the fasta file.

MW Molecular weight of the protein (KDa).

log(E value) log of protein's E-value.

- Statistical value representing the number of times this protein would be identified just by random.
- Calculated by the product of unique peptide E-value in all sample.

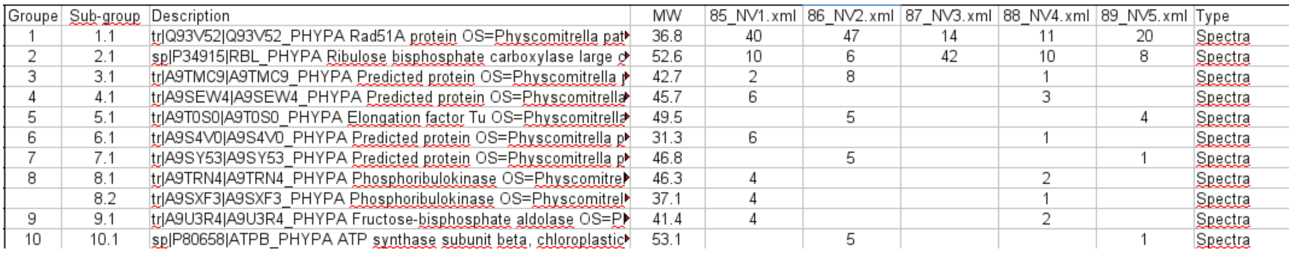
Type Defines the item that is compared between samples

Spectra Number of MS/MS spectra identified for the protein

Specifics Number of specific MS/MS spectra identified for the protein compared to the other proteins belonging to the same group.

Uniques Number of unique peptide sequences identified for this protein.

PAI Protein Abundance Index (3.1)



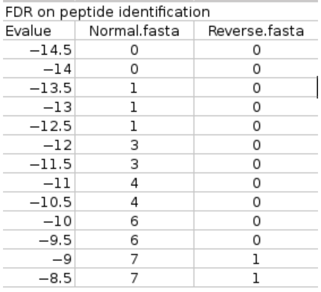
Groupe	Sub-group	Description	MW	85_NV1.xml	86_NV2.xml	87_NV3.xml	88_NV4.xml	89_NV5.xml	Type
1	1.1	tr|Q93V52|Q93V52_PHYPA Rad51A protein OS=Physcomitrella pat	36.8	40	47	14	11	20	Spectra
2	2.1	sp|P34915|RBL_PHYPA Ribulose bisphosphate carboxylase large c	52.6	10	6	42	10	8	Spectra
3	3.1	tr|A9TMC9|A9TMC9_PHYPA Predicted protein OS=Physcomitrella p	42.7	2	8		1		Spectra
4	4.1	tr|A9SEW4|A9SEW4_PHYPA Predicted protein OS=Physcomitrella	45.7	6			3		Spectra
5	5.1	tr|A9T0S0|A9T0S0_PHYPA Elongation factor Tu OS=Physcomitrella	49.5		5			4	Spectra
6	6.1	tr|A9S4V0|A9S4V0_PHYPA Predicted protein OS=Physcomitrella p	31.3	6			1		Spectra
7	7.1	tr|A9SY53|A9SY53_PHYPA Predicted protein OS=Physcomitrella p	46.8		5			1	Spectra
8	8.1	tr|A9TRN4|A9TRN4_PHYPA Phosphoribulokinase OS=Physcomitre	46.3	4			2		Spectra
	8.2	tr|A9SXF3|A9SXF3_PHYPA Phosphoribulokinase OS=Physcomitrel	37.1	4			1		Spectra
9	9.1	tr|A9U3R4|A9U3R4_PHYPA Fructose-bisphosphate aldolase OS=P	41.4	4			2		Spectra
10	10.1	sp|P80658|ATPB_PHYPA ATP synthase subunit beta, chloroplastic	53.1		5			1	Spectra

Figure 6: Comparison results

3.4 *fdr.txt

The result file indicates the number of peptides with an E-value less than the E-value indicated in the fist column (Fig 7). You just have to divide the number of peptides in the reverse or decoy database by the number of peptides in the normal database to obtain the false discovery rate at each E-value level.
This methods has 2 limitations :

- Normal and reverse databases must be saved in different fasta files.
- The filter parameters must be low (2.3)



FDR on peptide identification

Evalue	Normal.fasta	Reverse.fasta
−14.5	0	0
−14	0	0
−13.5	1	0
−13	1	0
−12.5	1	0
−12	3	0
−11.5	3	0
−11	4	0
−10.5	4	0
−10	6	0
−9.5	6	0
−9	7	1
−8.5	7	1

Figure 7: FDR results