

X!Tandem Batch Analysis

Automated performed database searching using X!Tandem in local version

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



14 Avril 2010

Abstract

X!Tandem software is an open-source software to performed peptides/proteins identifications from MS/MS mass spectra. To easily performed automated search using predefined parameters in a local version, we developed this perl script. This system is an alternative to install an local server of the Global Proteome Machine (GPM)

X!Tandem Batch Analysis script applied database searching with defined parameters in a list of MS/MS analysis with a list of protein databases.

Contents

1	Installation	2
1.1	Windows	2
1.2	Linux	2
1.3	License	2
2	Utilization	3
2.1	Running Script	3
2.2	Peak lists	3
2.3	Databases	3
2.4	Models	4
3	Results	5
3.1	Database searching	5
3.2	Viewing result	5
3.3	Automated export result	5

1 Installation

1.1 Windows

1. If not installed, download and installed Active-perl (Version 5.8.XXX) using this [link](#).
Warning : The script need version 5.8, and does not work on 5.10 or 5.12!!!!
2. Download the [archive](#) and unzip it.
3. Move the complete "Benperl/" folder directly at the C:/
4. Inside, go to the folder Benperl/installation/
5. Install perl library using *Perl-library-installation.pl* script¹

1.2 Linux

Requirements

- Installed *libwx-perl* package

Ubuntu

- Add this software [repository](#)
- Installed the *xtandem-tornado* package

Other distribution

- Download [source](#) of the X!Tandem and followed the instruction of compilation.

Script adaptation

1. Download the [archive](#) and unzip it.
2. Move the complete "benperl/" folder in your /home/user/ folder
3. Inside, go to the folder Benperl/
4. Open on a text editor the perl script *xtandem-batch-analysis.pl* and modified the path to the X!Tandem executable and to the folder containing the model (xml files).

1.3 License

Copyright (C) 2010 Valot Benoit

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the [GNU General Public License](#) for more details.

¹Simply double-clic it

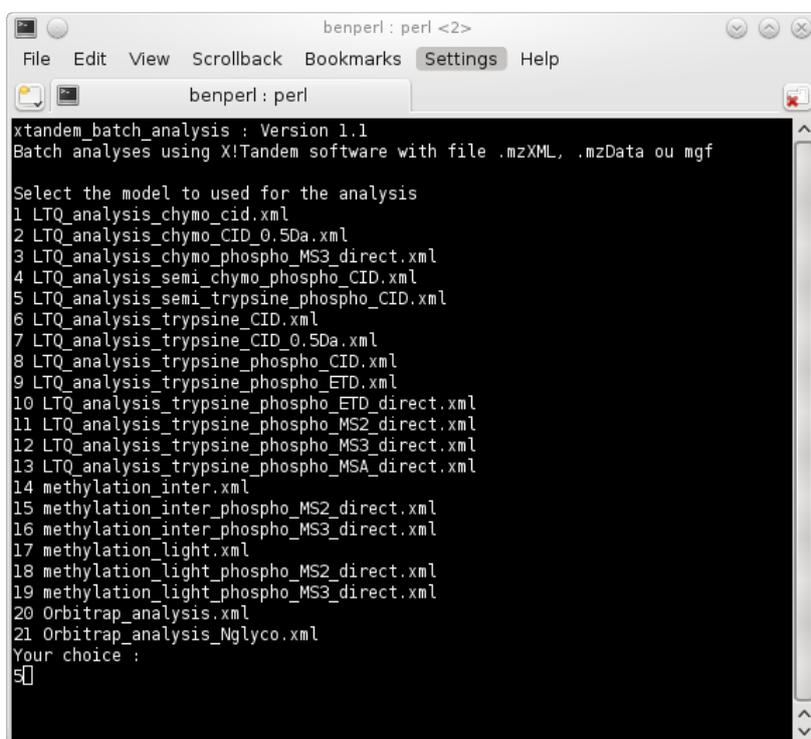
2 Utilization

It permits to analysis a list of peak list files on a list of protein database using the X!Tandem software. Three successive box permit to select mzXML file or other peaklist, to select protein databases and finally the folder were results are stored. The databases must be proteins, X!Tandem doesn't work on DNA database.

2.1 Running Script

To performed analysis, start the perl script *xtandem-batch-analysis.pl*²

1. Select peak lists files to be analyzed. (See 2.2)
2. Select databases files to be searched. (See 2.3)
3. Select the folder to save the results files.
4. Select in the model containing searching parameters. (See 2.4 and Fig 1)



```

benperl : perl <2>
File Edit View Scrollback Bookmarks Settings Help
benperl : perl
xtandem_batch_analysis : Version 1.1
Batch analyses using X!Tandem software with file .mzXML, .mzData ou mgf

Select the model to used for the analysis
1 LTQ_analysis_chymo_cid.xml
2 LTQ_analysis_chymo_CID_0.5Da.xml
3 LTQ_analysis_chymo_phospho_MS3_direct.xml
4 LTQ_analysis_semi_chymo_phospho_CID.xml
5 LTQ_analysis_semi_trypsin_phospho_CID.xml
6 LTQ_analysis_trypsin_CID.xml
7 LTQ_analysis_trypsin_CID_0.5Da.xml
8 LTQ_analysis_trypsin_phospho_CID.xml
9 LTQ_analysis_trypsin_phospho_ETD.xml
10 LTQ_analysis_trypsin_phospho_ETD_direct.xml
11 LTQ_analysis_trypsin_phospho_MS2_direct.xml
12 LTQ_analysis_trypsin_phospho_MS3_direct.xml
13 LTQ_analysis_trypsin_phospho_MSA_direct.xml
14 methylation_inter.xml
15 methylation_inter_phospho_MS2_direct.xml
16 methylation_inter_phospho_MS3_direct.xml
17 methylation_light.xml
18 methylation_light_phospho_MS2_direct.xml
19 methylation_light_phospho_MS3_direct.xml
20 Orbitrap_analysis.xml
21 Orbitrap_analysis_Nglyco.xml
Your choice :
5

```

Figure 1: Model selection

2.2 Peak lists

X!Tandem works with open peak list files like mzXML, mgf, mzData, mzML or pkl files.

2.3 Databases

X!Tandem software uses only protein database in fasta format. It doesn't work with EST³ sequences. You can transform your database using our application *Protein database manager*, available [here](#), or for direct [running](#).

²Simply double clic it.

³Expressed Sequenced tag

2.4 Models

To performed database searching, you must created or edited a model xml file. Some example are presents in the "/benperl/Xtandem" folder. Each xml file in this folder could be select in the terminal during analysis (Fig 1).

To modified the parameter xml file, open it in a text editor (Fig 2) and then follow the [documentation](#) at the X!Tandem web site.

```
<?xml version="1.0"?>
<bioml label="example api document">
<note type="heading">Paths</note>

<note type="heading">Spectrum general</note>
  <note type="input" label="spectrum, fragment mass type">monoisotopic</note>
  <note type="input" label="spectrum, fragment monoisotopic mass error">0.8</note>
  <note type="input" label="spectrum, fragment monoisotopic mass error units">Daltons</note>
  <note>The value for this parameter may be 'Daltons' or 'ppm': all other values are ignored</note>
  <note type="input" label="spectrum, parent monoisotopic mass error plus">2.5</note>
  <note type="input" label="spectrum, parent monoisotopic mass error minus">1.0</note>
  <note type="input" label="spectrum, parent monoisotopic mass error units">Daltons</note>
  <note type="input" label="spectrum, parent monoisotopic mass isotope error">no</note>

<note type="heading">Residue modification</note>
  <note type="input" label="residue, modification mass">57.04@C</note>
  <note type="input" label="residue, potential modification mass">15.99@M</note>

<note type="heading">Protein general</note>
  <note type="input" label="protein, cleavage site">[RK][P]</note>
  <note type="input" label="protein, cleavage C-terminal mass change">+17.00305</note>
  <note type="input" label="protein, cleavage N-terminal mass change">+1.00794</note>
  <note type="input" label="protein, N-terminal residue modification mass">0.0</note>
  <note type="input" label="protein, C-terminal residue modification mass">0.0</note>

<note type="heading">Scoring</note>
  <note type="input" label="scoring, minimum ion count">4</note>
  <note type="input" label="scoring, maximum missed cleavage sites">2</note>
  <note type="input" label="scoring, cyclic permutation">no</note>
  <note type="input" label="scoring, y ions">yes</note>
  <note type="input" label="scoring, b ions">yes</note>
```

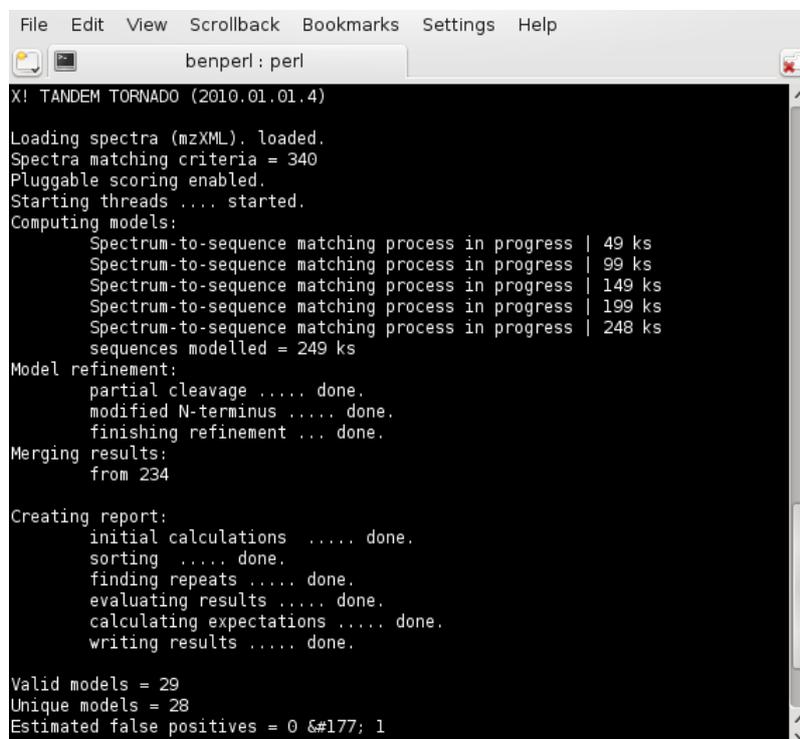
Figure 2: Model edition

To used complete performance of our computer, specify the number of cpu in the model at the line :
type="input" label="spectrum, threads".

3 Results

3.1 Database searching

Each peak list files are analyzed in series. You can follow advance of the processing in the terminal (Fig 3). To stop processing, simply close the terminal windows.



```
File Edit View Scrollback Bookmarks Settings Help
benperl : perl
X! TANDEM TORNAO (2010.01.01.4)
Loading spectra (mzXML), loaded.
Spectra matching criteria = 340
Pluggable scoring enabled.
Starting threads .... started.
Computing models:
  Spectrum-to-sequence matching process in progress | 49 ks
  Spectrum-to-sequence matching process in progress | 99 ks
  Spectrum-to-sequence matching process in progress | 149 ks
  Spectrum-to-sequence matching process in progress | 199 ks
  Spectrum-to-sequence matching process in progress | 248 ks
  sequences modelled = 249 ks
Model refinement:
  partial cleavage ..... done.
  modified N-terminus ..... done.
  finishing refinement ... done.
Merging results:
  from 234
Creating report:
  initial calculations ..... done.
  sorting ..... done.
  finding repeats ..... done.
  evaluating results ..... done.
  calculating expectations ..... done.
  writing results ..... done.
Valid models = 29
Unique models = 28
Estimated false positives = 0 &#177; 1
```

Figure 3: Database searching progress

3.2 Viewing result

The result are in xml format, you can view results on the [GPM web site](#) (Fig 4).

3.3 Automated export result

To automated processing the result files in excel view, you can used our application *My Xtandem parser* available [here](#).



The Global Proteome Machine

advanced [page](#)
view saved [xml data](#)

what is the [gpm](#)
powered by [tandem](#)
send us [email](#)

Eukaryote proteomes
[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

Boutique proteomes
[human](#) [mouse](#) [frog](#)
[cow](#) [bacteria](#) [plant](#)
[fish](#) [rat](#)

Algorithms
[X! P3](#) | [X! Hunter](#)

Information
[gpmDB](#) [wiki](#)

saving All of the data associated with a GPM search is stored in a single **XML** XML file, using an XML called **BIOML**. On the top of an results **data:** page, there is a link that allows you to view that file. To **Save** that data to a local file, simply click on that link and then use the "File->Save As" menu item on your browser. This creates a local copy of the data file on your computer.

using Once you have saved a GPM results file, using this facility you can **saved** upload the file to the GPM and view the data using the viewing **data:** tools. In this way, you can share data with colleagues: by sending them a saved file, they can easily view the information themselves.

data GPM BIOML files only (gzipped files with .xmz, .gz, .z or .zip
file: extensions automatically decompressed)

Figure 4: Viewing results on GPM site