

De Novo Pipeline : Automated identification by De Novo interpretation of MS/MS spectra

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



29 October 2010

Abstract

The classical method for protein identification of LC-MS/MS data uses database searching and matching, as it is the case in Mascot, Sequest or X!Tandem softwares. However, these identifications are possible only if the protein being searched is present in the database. To address this problem the *de novo* interpretation strategy can be used instead. There exist softwares that use this strategy, but they are often difficult to use in a direct, automated way.

The De Novo Pipeline uses the *de novo* technique to perform identification on data collected from ion trap mass spectrometers. It performs automated analysis by connection of two applications :

1. [PepNovo](#) : automated interpretation of MS/MS spectra in a possible peptide sequence,
2. [Fasts](#) : homology search in an iterative mode, to identify proteins from peptides sequences.

The De Novo Pipeline is complementary to X!Tandem : it allows you to remove spectra previously identified by X!Tandem and perform analysis on the remaining ones. The results can be graphically viewed and/or exported into tabulated files.

Contents

1	Installation	3
1.1	Requirements	3
1.2	Third party softwares for Windows	3
1.3	Third party softwares for Linux	3
1.4	Run De Novo Pipeline	3
1.5	License	4
2	Processing	5
2.1	DeNovo sequencing	5
2.1.1	Utilization	5
2.1.2	Search parameters	5
2.1.3	PepNovo calculation	5
2.2	Homology search	6
2.2.1	Utilization	6
2.2.2	Search parameters	6
2.2.3	Fasts calculation	7
3	Results	8
3.1	Tabulated files	8
3.1.1	File *protein.txt	8
3.1.2	File *peptide.txt	8
3.1.3	File *alignment.txt	9
3.2	Graphical view of the results	10
3.3	Export	11
3.3.1	PepNovo sequences	11

1 Installation

1.1 Requirements

The De Novo Pipeline works both on Linux and Windows platforms. Java 1.6 is required and can be found here : [link](#). Also, the PepNovo and Fasta softwares must be installed on the system as described below.

1.2 Third party softwares for Windows

1. Download the [De Novo pipeline archive](#) and unzip it.
2. Create a folder "Benperl/" directly in the C:/ directory.
3. Move the folders "Fasta" and "PepNovo_bin", from the archive to the new folder "C:/Benperl/".

1.3 Third party softwares for Linux

Ubuntu

- Add this software [repository](#) to your system.
- Install the *pepnovo* and *fasta* packages.

Other distributions

- Download the sources of PepNovo and Fasts included in this [archive](#) and followed the instruction of compilation.

1.4 Run De Novo Pipeline

To run De Novo Pipeline :

- open De Novo Pipeline by using this [link](#);
- allow the program to be executed;
- the main window will appear. (Fig 1)

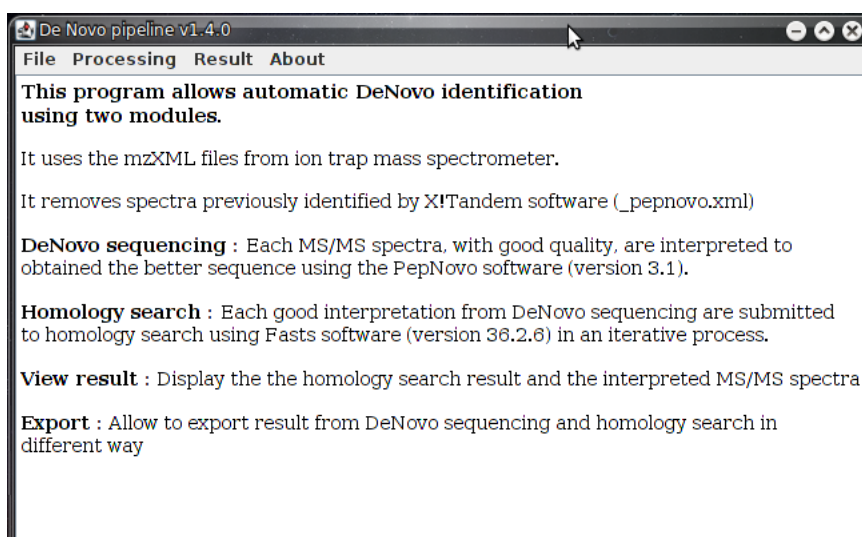


Figure 1: Principal window

1.5 License

Copyright (C) 2010 Valot Benoît

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.

2 Processing

2.1 DeNovo sequencing

The PepNovo software (version 3.1, [site](#)) interprets every MS/MS spectrum of good quality in order to obtain the best sequence possible. For each LC-MS/MS file :

1. the spectra are transformed from mzXML to mgf format;
2. a quality score is performed by PepNovo in order to remove poor spectra;
3. the spectra, previously identified by X!Tandem and filtered using our [X!Tandem pipeline](#), are removed;
4. for the remaining spectra, a possible sequence was determined by PepNovo.

2.1.1 Utilization

The processing is split as follows :

1. define the search parameters (Fig 2);
2. select the mzXML files to analyze (no other file format than mzXML is supported);
3. select the folder where to put the PepNovo result files (.pepnovo files);
4. you can watch the processing in progress (Fig 3).

2.1.2 Search parameters

Number of CPUs

Defines how many mzXML files are processed in parallel. For the best performance, you must use less than the maximum number of processors of your PC.

Quality score

This score represents the quality of the spectra. It lies between 0 (poorest quality) and 1 (best quality). This score is used for the filtering of the spectra : a quality score between 0.01 to 0.05 is sufficient to filter more than half of all spectra without losing valid peptides.

X!Tandem filter

If you have performed previous identifications using X!Tandem software, you can remove identified peptides using our [X!Tandem pipeline](#) (_pepnovo.xml file).

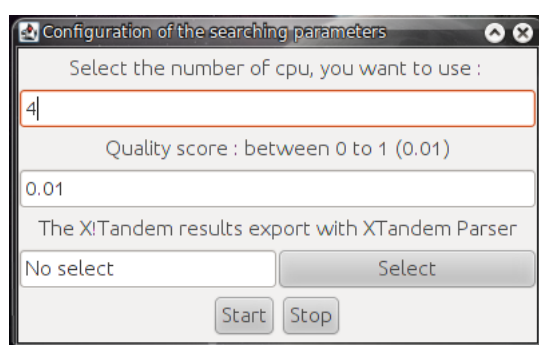


Figure 2: Parameter window

2.1.3 PepNovo calculation

During the PepNovo calculation, you can overview the progression state of the analysis.

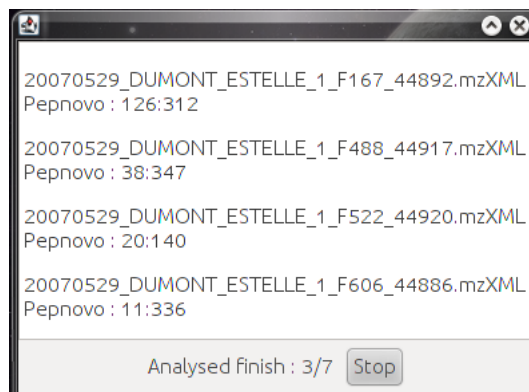


Figure 3: Progress window

2.2 Homology search

Every good interpretation obtained from DeNovo sequencing, is submitted to homology search using Fasts software (version 36.2.6, [site](#)), in an iterative process as follows :

1. sequences determined by PepNovo are filtered according to a filter score;
2. an homology search is run against 2 or 3 databases sequentially :
 - the contaminants database;
 - the proteins previously identified by X!Tandem;
 - the selected database.
3. After the homology search :
 - if a result with an Evalue less than 0.0001 is found, the proteins are conserved. Peptides use for the alignment are removed from the peptide sequence list. A new search is performed in the **same** database.
 - If no valid result is found, a new search is performed in the **next** database.
4. When all databases are searched, a complete export in tabulated files (see 3.1) is produced, containing the total result.

2.2.1 Utilization

The processing are split as follows :

1. define the search parameters (Fig 4);
2. select the pepnovi files to analyze (these files are created by the previous process);
3. select the protein database (.fasta) against whom homology search will be performed;
4. define the name of the tabulated result files (*.txt format). Fasts result files will be saved in the same folder.
5. You can view the processing in progress (Fig 5).

2.2.2 Search parameters

Number of CPUs

Defines how many .pepnovo files are processed in parallel. For the best performance, you must use less than the maximum number of processors of your PC.

PepNovo score

This score represents the confidence of *de novo* interpretation of the spectrum in a peptide sequence performed by PepNovo. Classical values lie from 50 to 100.

Contaminants database

You can select a contaminants database to remove peptides from keratins, trypsin, ... This way, they are not reported into the results.

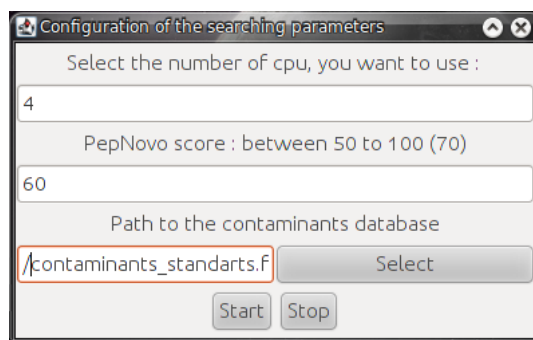


Figure 4: Parameter window

2.2.3 Fasts calculation

During the Fasts calculation, you can overview the progression state of the analysis.

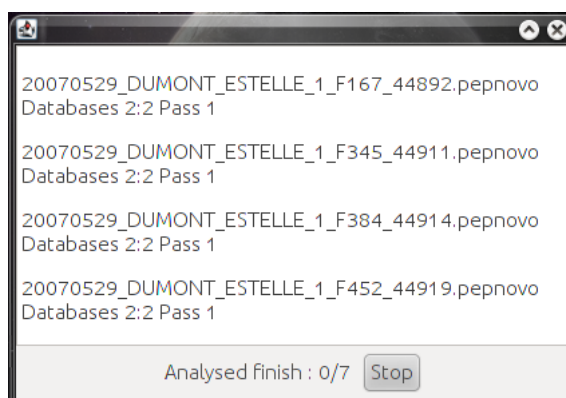


Figure 5: Progress window

3 Results

3.1 Tabulated files

Warning The results of homology search contain **redundancy** which must be filtered manually. In fact, a protein or homologous proteins may be reported more than once with different peptides and thus, appear in the results with different protein numbers.

The tabulated files contain sequentially every sample result (MS/MS file).

3.1.1 File *protein.txt

This file contains the proteins identified for each sample (Fig 6).

Number

Protein number representing the N^{st} homology search. In this example, a protein is identified twice (P08926 and P14226). In the second result, the protein number begins at 2, because one contaminant protein identified has not been reported.

Description

Description of the identified protein.

Evalue

Evalue of the homology result.

Peptides

Number of spectra used for the homology result.

Databases

Database used for this identification. If the database is X!Tandem :

Number

Protein number indicated in the X!Tandem result.

Peptides

Number of spectra identified for this protein in the previous X!Tandem search.

20070529_DUPOND_DUPONT_1_F167_44892.mzXML

Number	Description	Evalue	Peptides	Databases	Number	Peptides
1	sp P08926 RUBA_PEA RuBisCO large subunit-binding protein su	0	9	X!Tandem	1.1	15
2	sp P08926 RUBA_PEA RuBisCO large subunit-binding protein su	3.4e-1	5	X!Tandem	1.1	15
3	tr Q2UZ90 Q2UZ90_ARATH Mg chelatase subunit I OS=Arabidops	9.9e-3	12	Uniprot_Viridiplantae	-	0
4	tr B9N0E2 B9N0E2_POPTA Predicted protein OS=Populus trichor	8.8e-1	3	Uniprot_Viridiplantae	-	0

20070529_DUPOND_DUPONT_1_F606_44886.mzXML

Number	Description	Evalue	Peptides	Databases	Number	Peptides
2	sp P14226 PSBO_PEA Oxygen-evolving enhancer protein 1, chl	0	10	Uniprot_Viridiplantae	-	0
3	sp P14226 PSBO_PEA Oxygen-evolving enhancer protein 1, chl	1.7e-1	11	Uniprot_Viridiplantae	-	0

Figure 6: Protein result

3.1.2 File *peptide.txt

This file contains the details of the peptides identified in each sample (Fig 7).

Number

Protein number representing the N^{st} homology search. In this example, the protein number begins at 2, because one contaminant protein identified has not been reported.

Description

Description of the identified protein.

Scan

Scan number of the MS/MS spectrum.

Sequence

Sequence of the peptide determined by PepNovo. N- or C-ter of the peptide may have not been determined, in that case they are indicated by a mass in N- or C-gap.



Charge

Charge of the MS/MS spectrum

MH+theo

Monoisotopic calculated mass for the peptide + one proton (MH^+). If N or C-gap are present, this value is estimated.

MH+obs

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

DeltaMH+

Error in the precursor mass between observed and theoretical data (Da). If N or C-gap, this value is not determined (ND).

N-gap

Monoisotopic mass not interpreted in the N-ter of the peptide.

C-gap

Monoisotopic mass not interpreted in the C-ter of the peptide.

Sequence score

Sequence score determined by PepNovo.

Filter score

Quality score of the spectrum, lies between 0 and 1.

20070529_DUPONT_DUPOND_1_F522.mzXML

Number	Description	Scan	Sequence	Charge	MH+theo	MH+obs	DeltaMH+	N-gap	C-gap	Sequence score	Filter score
2	sp Q01517 ALFC2_PEA Fructose-bisphosphate aldolase 2, chlor										
		172	YLGDWSEEAQK	2	1325.6011	1325.7328	-0.131713	0.0	0.0	150.432	0.965
		201	SAAYEQQR	2	1115.5122	1115.1581	0.354125	0.0	0.0	75.139	0.791
		340	SLAKLGK	2	917.32324	918.1846	ND	200.856	0.0	123.382	0.509
		639	LGLETEANR	2	1432.631	1434.1942	ND	316.066	0.0	88.24	0.947
		675	LAMDSENA	2	1465.9894	1466.7211	ND	169.643	363.956	97.537	0.754
		770	TLNLLHR	3	1484.5223	1485.9032	ND	618.001	0.0	88.683	0.66
		829	QEALLFR	2	1018.2915	1018.7406	ND	141.797	0.0	108.978	0.908
		869	EYTLNLLHR	2	1328.4332	1329.2635	ND	169.806	0.0	145.998	0.887
		899	VSLPNDYFGLK	2	1452.3898	1452.8031	ND	199.732	0.0	99.358	0.223
		1170	LVDVLVLEELL	3	1992.569	1993.3284	ND	0.0	756.832	105.177	0.73
		1251	GSNNESWCQGL	2	2401.9492	2401.568	ND	550.15	618.328	81.011	0.717
3	tr A9RX76 A9RX76_PHYPA Fructose-bisphosphate aldolase OS=Ph										
		524	RLDSLGLTEANR	2	1587.8088	1587.6835	0.125366	0.0	0.0	134.653	0.921
		639	LGLETEANR	2	1432.631	1434.1942	ND	316.066	0.0	88.24	0.947
		701	VAEYTLNTYQKR	3	1485.77	1486.0815	-0.311523	0.0	0.0	75.876	0.552
		1177	LVEELL	2	1777.3734	1778.7916	ND	553.887	413.997	91.616	0.29

Figure 7: Peptide result

3.1.3 File *alignment.txt

For each homology result, an *alignment.txt file containing the alignment is created (Fig 8).

```

20070529_DUMONT_ESTELLE_1_F167_44892.mzXML
>>sp|P08926|RUBA_PEA RuBisCO large subunit-binding protein su...
Sample                               10      20
                               DLAFDKVA-----EAADAVGLTLGPR
                               :.:.:.:      :.:.:.:.:
sp|P08 QTSLSKVKVQHGRVNFQKPNRFVVKAAAKDIAFDQHSRSAMQAGIDKLADAVGLTLGPR
                               30      40      50      60      70      80

Sample --TLVLDEFGSPKVVNDGVT-----40DAGAALLR-----50DSAGDGTITASLL
                               :.:.:.:.:      :.:.:.:      :.:.:.:.:
sp|P08 GRNVVLDEFGSPKVVNDGVTIARAIELPDPMENAGAALIREVASKTND5AGDGTITASIL
                               90      100     110     120     130     140

Sample AR-----70LGLLNVTSANGLLKK-----80AALVEELEK-----90LSAGND
:: :.:.:.:.: :.:.: :.:.:.:.:
sp|P08 AREIIKLGLLNVTSANGANPVSIKKGIDKTVAAALVEELEKLARPVKGGDDIKAVATISAGND
                               150     160     170     180     190     200

Sample ELL-----100GYLSPQFVTN---110SLVEM
:: :.:.:.:.:
sp|P08 ELIGKMIAEAIDKVGPDGVL5IESSNSFETTVEVEEGMEIDRGYISPQFVTNPEKSIVEF
                               210     220     230     240     250     260

Sample 120ENARVLLTDQK-----130GLLNVAA
:: :.:.:.:.:
sp|P08 ENARVLITDQKISAIKDIIPLEKTTQLRAPLLIISEDITGEALATLVVNKLRGILNVAA
                               270     280     290     300     310     320

Sample LK-----TLNAD
:: :.:.:.:
sp|P08 IKAPGFGERRKALLQDIAILTGAEFQASDLGLLVENTTIEQLGLARKVTISKDSTIIAD
                               330     340     350     360     370     380

Sample 140AASK-----150SETDSLYDSTR-----160AATETELEDK---
:: :.:.:.: :.:.:.:.: :.:.:.:.:
sp|P08 AASKDELQSRVAQLKKELSETDSIYDSEKLAERIAKLSGGVAVIKVGAATETELEDKRLR
                               390     400     410     420     430     440

Sample -----170LEEGLVP-----LGADLVQK-----
:: :.:.:.: :.:.:.:.:
sp|P08 IEDAKNATFAAIEEGIVPGGGTALVHLSGYVPAIKEKLEDADERLGADIVQKALVAPAAL
                               450     460     470     480     490     500

Sample 180----AGLEADVVEK
:: :.:.:.:
sp|P08 IAQNAGIEGEVVVEKIKNGEWEVGYNAMTDYENLVESGVIDPAKVTRCALQNAASVAGM
                               510     520     530     540     550     560

```

Figure 8: Alignment result

3.2 Graphical view of the results

For each sample, you can watch the Fasts results. The application's graphical window is divided in three (Fig 9) :

- the list of the identified proteins and the expandable list of the peptides used for the homology;
- the alignment of the homology (click on a protein or a peptide to view it);
- the annotated MS/MS spectrum with b/y ions (click on a peptide to view it).

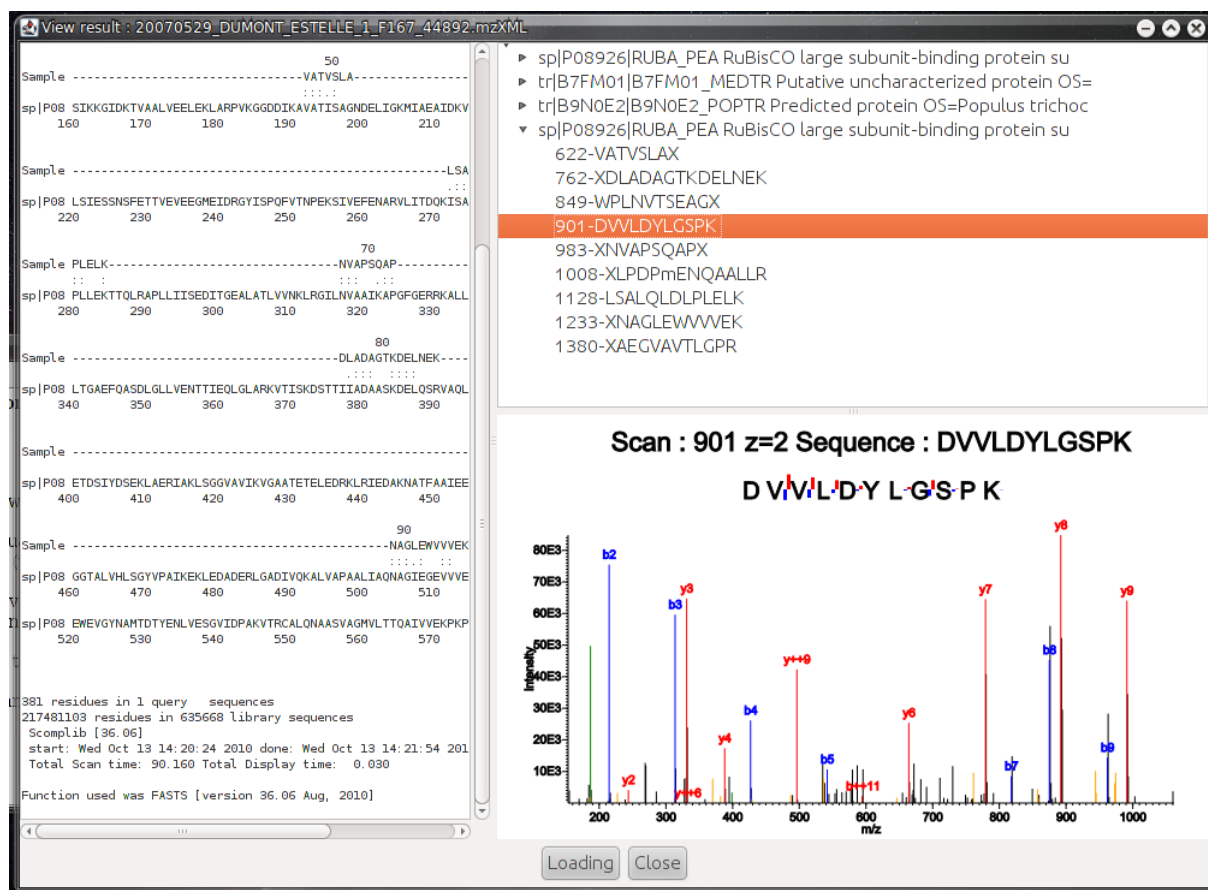


Figure 9: View result

3.3 Export

3.3.1 PepNovo sequences

You can export the sequence determined by PepNovo for each spectrum.