
PTMiner User Guide

(Last updated: February 21st, 2022)

Contents

1. Introduction.....	2
2. How to run PTMiner.....	2
2.1 Setting parameters using the interface	3
2.2 Loading an existing parameter file	9
2.3 Running PTMiner	10
3. Best Practice.....	10
3.1 Format of search results	10
3.2 Typical use of PTMiner.....	12
4. Output Files.....	13
5. Interpretation of Output	14
5.1 <i>filtered_result.txt</i>	14
5.2 <i>loc_result.txt</i>	15
5.3 <i>anno_result.txt</i>	15
5.4 <i>filtered_summary.txt</i>	16

1. Introduction

PTMiner is an efficient software tool for accurate filtering, localization and annotation of protein modifications identified by open database search in proteomics. It provides a user-friendly interface for setting parameters and running. The core of PTMiner was written in standard C++ and the interface was implemented in C# on the platform of Microsoft Visual Studio ultimate 2013 in Windows System. PTMiner is freely available at <http://fugroup.amss.ac.cn/software/ptminer/ptminer.html>.

2. How to run PTMiner

Now PTMiner can be run on Windows system smoothly. If your computer has not installed Visual Studio ultimate 2013, please install it first by double clicking the 'vcredist_x64.exe' in the program folder. Then, double click the icon  of *PTMiner.exe* to start up PTMiner. The graphic user interface (GUI) of PTMiner is as shown in **Figure 1**.

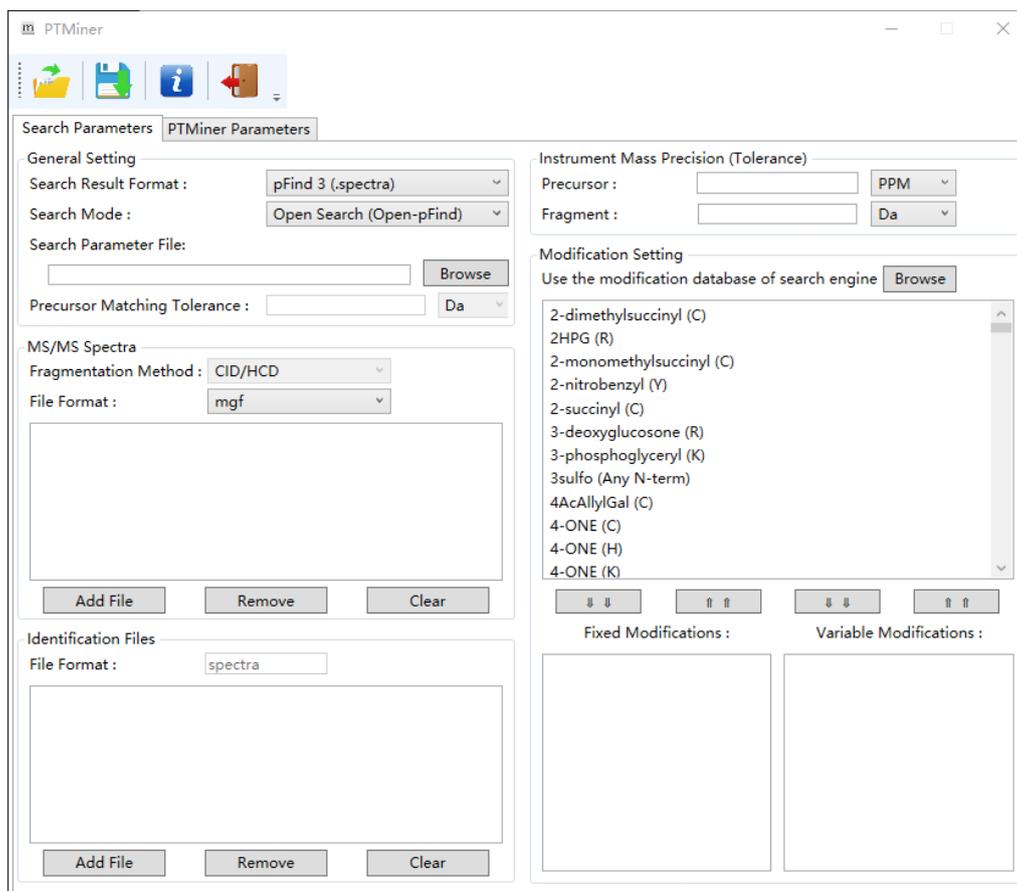


Figure 1. The GUI of PTMiner

2.1 Setting parameters using the interface

The parameters are separated into two groups, including the parameters used for database search (**Figure 2**) and the parameters used by PTMiner (**Figure 3-5**).

2.1.1. Setting search parameters

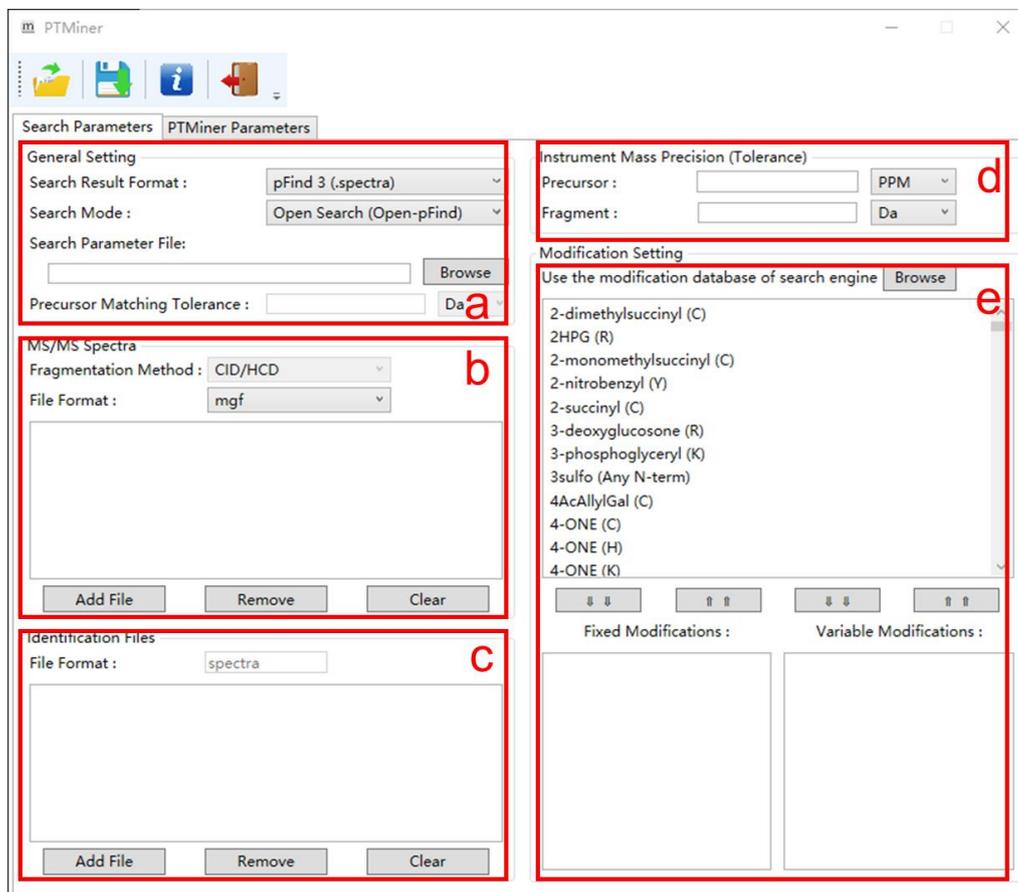


Figure 2. Search parameters setting

2.1.1.1 General Setting (**Figure 2a**)

A) Search Result Format

The following search result formats are supported by PTMiner,

- pFind 3 (.spectra)
- pFind 2.8 (.txt)
- MSFragger (.pepXML)
- Comet (.pep.xml)
- Sequest (.txt)
- PTMiner (.txt)

PTMiner (.txt) is our own format with headers as described in *filtered_results.txt* in section 4.1, and adjacent attributes delimited by tabulator.

B) Search Mode

There are two search modes, i.e., close search or open search. Close search means traditional database search using a tight precursor tolerance, e.g. 10 ppm, and open search means that a large precursor tolerance (500 Da for example) has been used. For pFind 3, there is an additional search mode option *open search (Open-pFind)* corresponding to the default open search (Open-pFind) in pFind 3.

C) Precursor Matching Tolerance

The precursor mass tolerance used in open search mode.

D) Fragmentation method

The fragmentation method in the experiment, CID/HCD or ETD are available.

2.1.1.2 MS/MS Spectra (**Figure 2b**)

A) File Format

PTMiner only supports *mgf* format of MS/MS Spectra at present.

B) Path

Click the “Add File” button to add MS/MS Spectra, the “Remove” button to delete one file selected, and the “Clear” button to delete all files.

2.1.1.3 Identification Files (**Figure 2c**)

A) File Format

It is specified based on the “Search Result Format” specified in search parameter table.

B) Path

Click the “Add File” button to add identification results, the “Remove” button to delete files selected, and the “Clear” button to delete all files.

2.1.1.4 Instrument mass precision (tolerance) (**Figure 2d**)

A) Precursor

This parameter is used to specify the instrument mass precision/tolerance of peptide precursors.

B) Fragment

This parameter is used to specify the instrument mass precision/tolerance of fragment ions.

2.1.1.5 Modification Setting (Figure 2e)

A) Use the modification database of search engine

The *browse* button after label *Use the modification database of search engine* make it possible to use the modification database of search engines for analysis instead of our original database. It is currently only available when the “Search Result Format” is pFind 3 or PTMiner.

B) Fixed Modifications

The types of fixed modifications used in database search. PTMiner lists all modifications in Unimod database, and the same modification names must be used for database search. One can use the  or  button to select or delete modifications.

C) Variable Modifications

The types of variable modifications used in database search.

2.1.2 Setting PTMiner parameters

The screenshot shows the PTMiner software interface with the following parameters and settings:

- FDR Filter (a):**
 - Enable FDR Filter:
 - Decoy Tag: REV_
 - FDR Threshold: 1 %
 - FDR Method: Transfer
 - FDR Formula: (decoy+1)/target
- Localization (b):**
 - Enable Localization:
 - Minimum PSM Number: 5
 - Using Prior Probability:
 - Filter Method: Probability
 - Filter Threshold: 0.5
 - Target Modifications: (empty list)
 - Modification Specificities: (empty list)
 - Buttons: Edit (for both lists)
 - Load the top [] modification(s) from... Browse
- Annotation (c):**
 - Enable Annotator:
 - Protein Database: (empty field) Browse
- Output (d):**
 - Output: (empty field) Browse
- Save & Run (e):** A button at the bottom center of the interface.

Figure 3. PTMiner parameters setting

2.1.2.1 FDR filter (Figure 3a)

A) Enable FDR Filter

Enable it to do FDR control. It is checked by default.

B) Decoy Tag

Decoy protein tag in the beginning of protein names. Default is "REV_" for pFind 3, "rev_" for Comet and "REVERSE_" for others.

C) FDR Threshold

Set the threshold of the false discovery rate. Default is 1%.

D) FDR Method

PTMiner provides three FDR estimation method, i.e. Global, Separate and Transfer FDRs. We recommend the Transfer FDR for the open search mode.

D) FDR Formula

PTMiner provides two estimation formulas for FDR, namely decoy/target and (decoy+1)/target

2.1.2.2 Localization (Figure 3b)

A) Enable Localization

Check this option to do modification localization. It is checked by default.

B) Minimum PSM Number

Only when the number of PSMs in one group of modification is more than this threshold can it be regarded as one type of modification. Default value is 5.

C) Using Prior Probability

Check this to use iteratively updated prior probability to localize modification sites. By default, it is checked when “Search mode” is “Open Search” (excluding “Open Search (Open-pFind)” of pFind 3), otherwise it is not check.

D) Filter Method

PTMiner provides two methods to filter localization results, i.e. Probability and FLR. By default, it is is “FLR” when “Search mode” is “Open Search” (excluding “Open Search (Open-pFind)” of pFind 3), otherwise it is “Probability”.

E) Filter Threshold

The threshold used to filter localization results. By default, it is 0.01 when “Filter Method” is “FLR” and 0.5 when “Filter Method” is “Probability”.

F) Target Modifications

This is for close search only. By clicking the “Edit” button, the window shown in **Figure 4** will be opened. Select the modifications that are going to be localized by PTMiner.

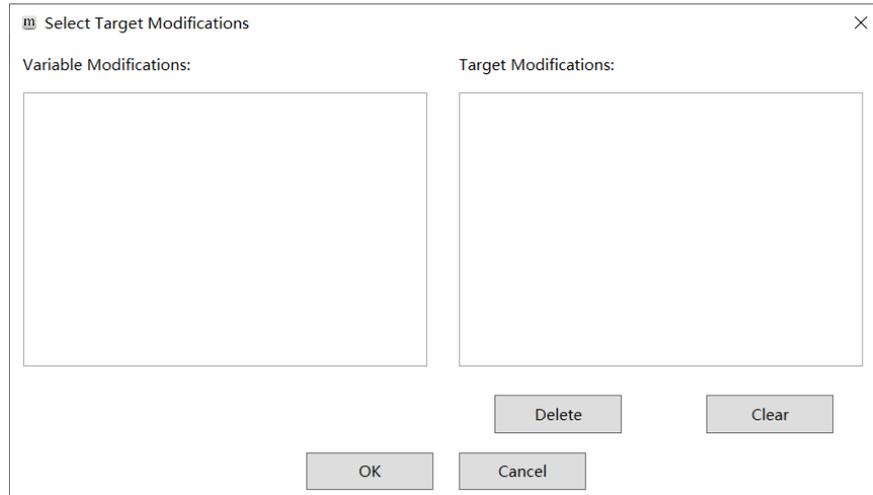


Figure 4. Select Target Modifications

G) Modification Specificities

This is for close search only. By clicking the “Edit” button, the window shown in **Figure 5** will be opened. Select the modification specificities for each target modification.

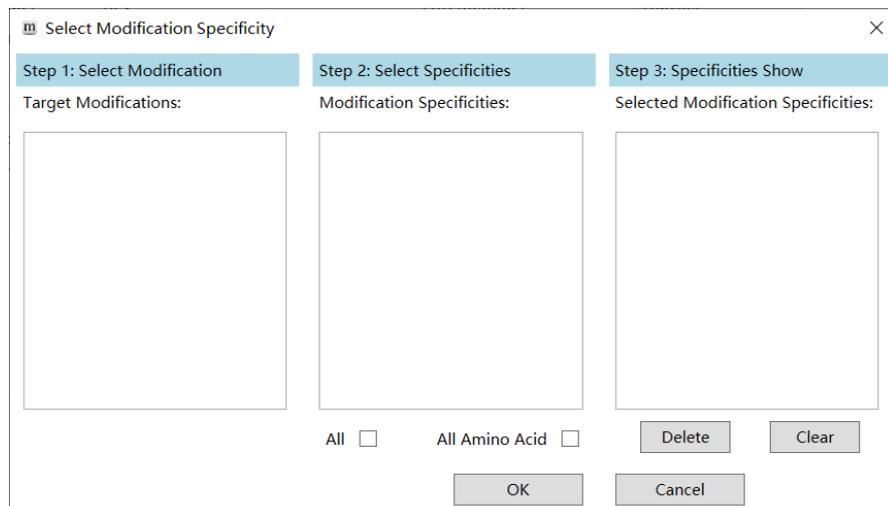


Figure 5. Select Modification Specificity

H) Load the top N modification(s) from file

Click the *browse* button after the “Load the top N modification(s) from file” label to load modifications with top N (specified in the text box) frequency from the summary file of search engine. All modifications will be loaded when the text box is not populated. This is only available when “Search Result Format” is “pFind 3”.

2.1.2.3 Annotation (Figure 3c)

A) Enable Annotation

Select this to do modification annotation. It is checked by default.

B) Protein Database

The protein database (*.fasta) used in the search engine.

2.1.2.4 Output (Figure 3d)

Specify the folder to save the results.

2.2 Loading an existing parameter file

If an existing parameter file (*.param) is available, a convenient way to set parameters is to load the file using the tool  in the toolbar (Figure 6)

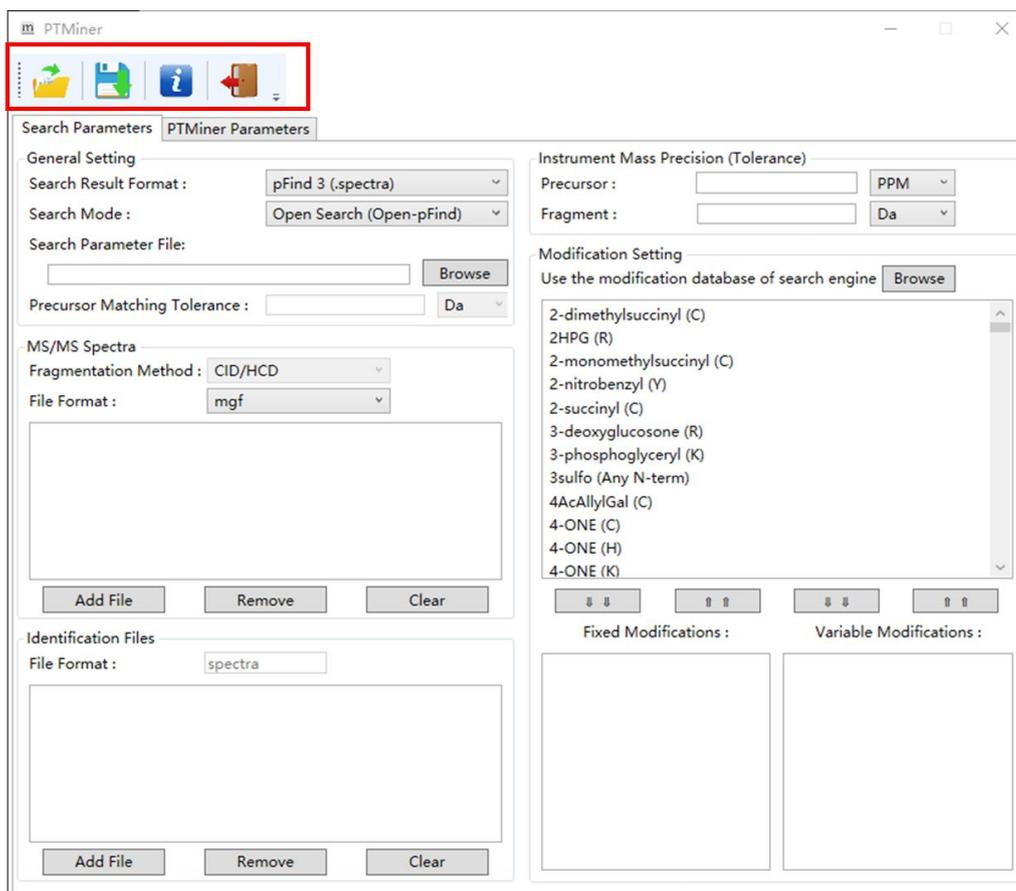


Figure 6. Load parameter file manually

2.3 Running PTMiner

Make sure that all parameters are set properly, and then click the “Run” button to start PTMiner (**Figure 3e**). The command console will appear to show the program progress.

3. Best Practice

3.1 Format of search results

PTMiner provides four specific formats of search engine (pFind 3, pFind 2.8.5, MSFragger, Comet and SequestHT in Proteome Discoverer 2.0 or above) results and a general format (PTMiner). If other search engine is used in your study, you can transfer your results into the general format to use PTMiner. The PSMConvert.exe tool provided in the PTMiner directory can convert MaxQuant and MS-GF+ result files to PTMiner format.

A) pFind 3. Only one .mgf file is supported in MS/MS Spectra. If you have multiple .mgf files, please use the FilesMerger tool to combine them into one.

B) pFind 2.8. Please note that PTMiner needs the result of pFind, not pBuild (post-processing program of pFind).

C) MSFragger. PTMiner supports its .pepXML format.

D) Comet. PTMiner supports its .pep.xml format.

E) SequestHT. If you want to use this search engine, we provide three search tips and result exportation methods.

(1) Search tips

- Search *raw* format files, and use *mgf* format files converted by *msconvert* (<http://proteowizard.sourceforge.net/>) as PTMiner input files.
- For the parameters of “Sequest HT”, the protein database should include decoy protein sequences.
- In the modular of “PSM Validation”, we recommend “Target Decoy PSM Validator”, and set both “Target FDR (Strict)” and “Target FDR (Relaxed)” to

be 1.0.

(2) Exportation method

The Sequest database search results should be exported through Proteome Discoverer in *.txt* format. We recommend the following operating steps,

- Select the table of PSMs
- Make sure that these 12 fields in the table are selected: “Sequence”, “Modifications”, “Protein Accessions”, “Charge”, “ ΔCn ”, “Rank”, “MH+ [Da]”, “Theo. MH+ [Da]”, “ $\Delta m/z$ [Da]”, “First Scan”, “Spectrum File”, and “XCorr”.
- Select “File” -> “Export”-> “To Text (tab delimited)...”, and then a window will appear. Choose the “PSMs” item to be exported, and double click the button “Export”.

E) General format PTMiner (.txt)

Thirteen attributes are needed as follows (see ‘*filtered_result.txt*’ in section five for details),

- (1) Dataset Name: Spectral file name;
- (2) Spectrum Name: Spectral name or scan number;
- (3) Sequence: Identified peptide sequence;
- (4) Charge: Precursor charge of the spectrum;
- (5) ObsMH: Observed precursor molecular weight with one proton;
- (6) Mass Shift: Identified mass shift (ObsMH-TheoMH, TheoMH is theoretical mass of identified peptide sequence with identified modifications);
- (7) Main Score: Identification Score;
- (8) High Score Better: Is greater ‘Main Score’ better than smaller? If yes, set it to 1 and otherwise set it to 0;
- (9) Identified Mod Name: Identified modification name;
- (10) Identified Mod Position: Identified modification position;
- (11) Protein Access: The protein access from which the ‘Sequence’ comes from;
- (12) Before AA: The adjacent amino acid before peptide in protein;

-
- (13) After AA: The adjacent amino acid after peptide in protein;
 - (14) Protein Start Position: The starting position of the peptide in the protein sequence.

3.2 Typical use of PTMiner

PTMiner can be run through a user-friendly graphic user interface (GUI), only need the user to provide spectral files, identification results and some basic parameters. The meaning of parameters please see section two. After specifying parameters, click 'Save & Run' button.

4. Output Files

<i>filtered_result.txt</i>	If “Enable FDR Filter” is checked, this file will appear in the folder specified in “Output”. PTMiner will load all identification results, and do FDR estimation for all of them. The results with scores greater than the threshold that fulfills the specified FDR level will be aggregated into this file.
<i>filtered_summary.txt</i>	If “Enable FDR Filter” is checked, this file will appear in the folder specified in “Output”. This file summarizes the modifications that appear in the <i>filtered_result.txt</i> .
<i>loc_result.txt</i>	If “Enable Localization” is checked, this file will appear in the folder specified in “Output”.
<i>loc_summary.txt</i>	If “Enable Localization” is checked, this file will appear in the folder specified in “Output”. This file summarizes the modifications that appear in the <i>loc_result.txt</i> .
<i>anno_result.txt</i>	If “Enable Annotation” is checked, this file will appear in the folder specified in “Output”. This file will NOT appear for close search results.
<i>anno_summary.txt</i>	If “Enable Annotation” is checked, this file will appear in the folder specified in “Output”. This file summarizes the modifications that appear in the <i>anno_result.txt</i> .

5. Interpretation of Output

5.1 *filtered_result.txt*

Dataset Name	Spectrum file name without path, e.g. “test.mgf”
Spectrum Name	Spectrum name or scan number
Sequence	Identified peptide sequence
Charge	Precursor charge
ObsMH	The experimentally observed peptide MH+
Mass Shift	Mass difference between observed and identified peptide masses
Main Score	<i>Final score</i> for pFind 3, <i>E-value</i> for pFind 2.8, <i>expect search score</i> for MSFragger and Comet, <i>XCorr</i> for SEQUEST, <i>SpecEValue</i> for MS-GF+, <i>Score</i> for MaxQuant
High Score Better	Whether higher main score is better. 1 for yes and 0 for no.
Identified Mod Name	Identified modification names including both fixed and variable modifications. If the PSM contains more than one modifications, they are separated by “;”, e.g., “Carbamidomethyl;Carbamidomethyl”. If the PSM contains no modifications, it is empty.
Identified Mod Position	Identified modification positions corresponding to the identified modifications. If the PSM contains more than one modifications, they are separated by “;”, e.g., “6;12”. If the PSM contains no modifications, it is empty. “0” represents the N-terminal of the peptide and “length of peptide+1” represents the C-terminal of the peptide, 1 to “length of peptide” indicate the corresponding position of the peptide sequence.
Protein Access	The Access of protein from which peptide comes. If the PSM contains more than one proteins, they are separated by “;”, e.g., “sp P05141 ADT2_HUMAN;sp P12236 ADT3_HUMAN”.
Before AA	The adjacent amino acid before peptide in protein. If the PSM

	contains more than one proteins, they are separated by “;”, e.g., “K;K”. Otherwise it is set to “-”.
After AA	The adjacent amino acid after peptide in protein. If the PSM contains more than one proteins, they are separated by “;”, e.g., “R;H”. Otherwise it is set to “-”.
Protein Start Position	The starting position of the peptide in the protein sequence. It is only available when the search engine is pFind 3 or pFind 2.8. Otherwise, it is set to -1.

5.2 *loc_result.txt*

The first 13 attributes are the same as in *filter_result.txt*. In addition, there are 3 more as follows:

Posterior Probability	The maximum posterior probability of the position localized in the peptide sequence
Position	The position with the maximal posterior probability
AA	The amino acid with the maximum posterior probability

5.3 *anno_result.txt*

In this type of file, there are two headers. The first is for basic information of localization result, and the second is for annotation result.

The first add attributes are as follows,

SDP Score	Similarity score between spectra of modified and corresponding unmodified peptides (with the same sequence and non-target modifications)
Annotation Type	Including three types, i.e. “Fully”, “Partially” and “None”. “Fully” indicates that both mass shift and site specificity are matched to existing modification(s) in Unimod. “Partially” indicates that only the mass is matched. “None” represents that the mass shift cannot be found in Unimod.

New Sequence	Updated new sequence after deleting or adding some amino acids on peptide termini. If a new peptide sequence (with or without modification) explains the mass shift better, this column will show the sequence. Otherwise, it is empty.
New Mod	The possible explanation of the new mass shift given the new sequence above
New Mod Position	The position of the new modification above

The second header includes attributes as follows,

*	This sign indicates that this row is annotation result.
# Mass, Mod	The order of the annotated mass (Mass) and the order of the annotated modification (Mod)
Annotated Mass	Annotated mass
Annotated Mod	Annotated modification
Annotated Mod Site	The same as the “Site” in the Unimod database
Annotated Mod Term Spec	The same as the “Position” in the Unimod database
Annotated Mod Classification	The same as the “Classification” in the Unimod database.

5.4 *filtered_summary.txt*

Name	The names of all modifications that appear in the <i>filtered_result.txt</i> . For open search mode, a mass shift will also be treated as a modification for statistics, and will be displayed in the form of “MassShift:X”, where X is an integer obtained by rounding the mass shift.
#Spectra	The number of spectra with the above modification
#Peptides	The number of peptides with the above modification. Peptides are considered as the same only when their peptide sequences and modification(s) (including site(s)) are the same.
#Sites	The number of sites with the above modification. Sites are considered as the same, only when the set of modified sites (each

	site is represented by the protein accession and amino acid position on the protein sequence) in the identification result are the same, they are considered the same site. This column will only appear when the search engine is pFind 3 or pFind 2.8. Since no location information can be acquired in the FDR filtering step, “MassShift:X” does not have a site-level count in the results of open search.
--	---

The *loc_summary.txt* and *anno_summary.txt* files have the same meaning as the above files, except that the corresponding *loc_result.txt* and *anno_result.txt* files are summarized respectively.