


pCluster User's Guide

version 2010.05.14

<http://pfind.ict.ac.cn/pcluster/>

Overview

pCluster is a software tool for detecting protein modifications independently of sequence databases by tandem mass spectral clustering. It has two algorithms for modification detection: One uses the peptide precursor information for spectral clustering, and the other uses the fragment information. The precursor-based algorithm named DeltAMT can detect abundant modifications in an extremely fast speed, while the fragment-based algorithm is able to detect lower-abundance modifications. Both algorithms are based on the assumption that the modified and unmodified versions of a peptide are simultaneously present.

pCluster was implemented using C++ language and can be used in Windows OS. Click the icon  in the pCluster directory to start pCluster. The following figure shows the user interface of pCluster.

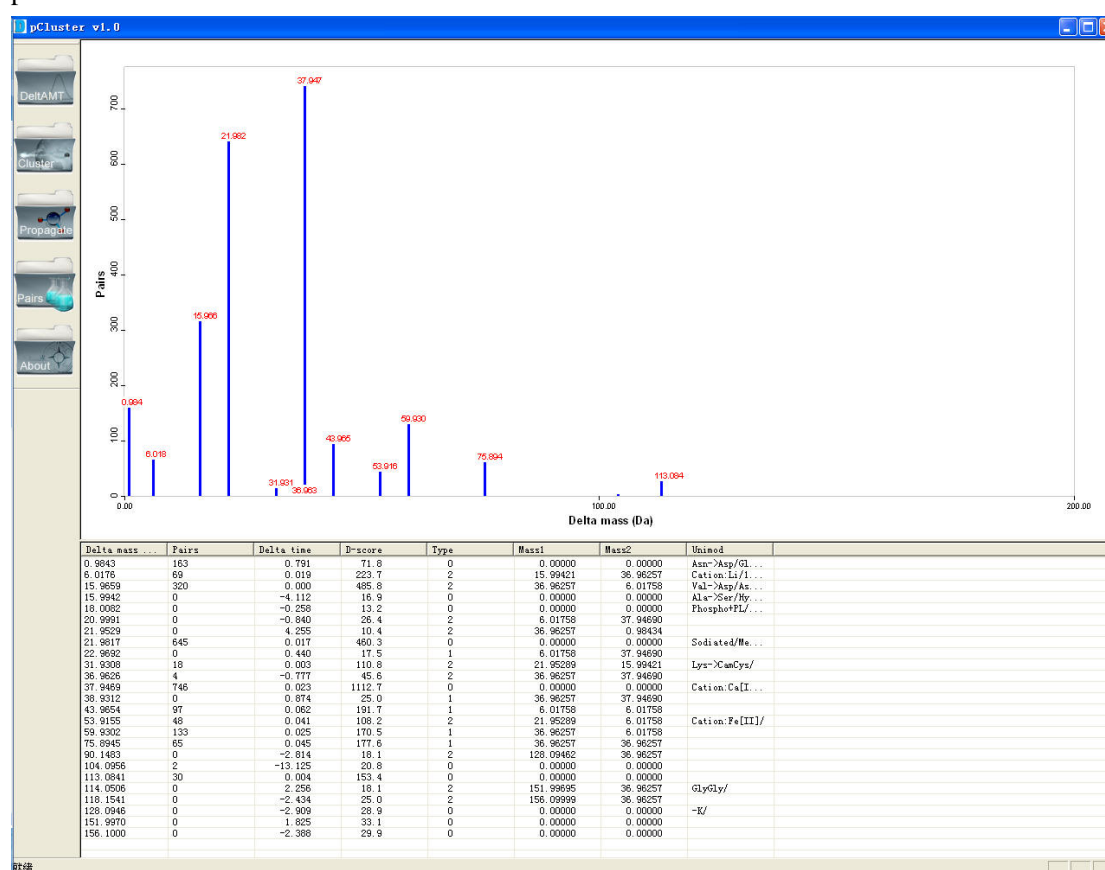


Fig 1. The user interface of pCluster.

The left tool bar provides the function buttons of pCluster, mainly including DeltAMT, Clustering and Propagation. Detected modifications are shown in the view zone and the list zone on the right of the window. In the view zone, mass shifts of detected modifications are displayed as a histogram. In the list zone, the details of detected modifications are listed.



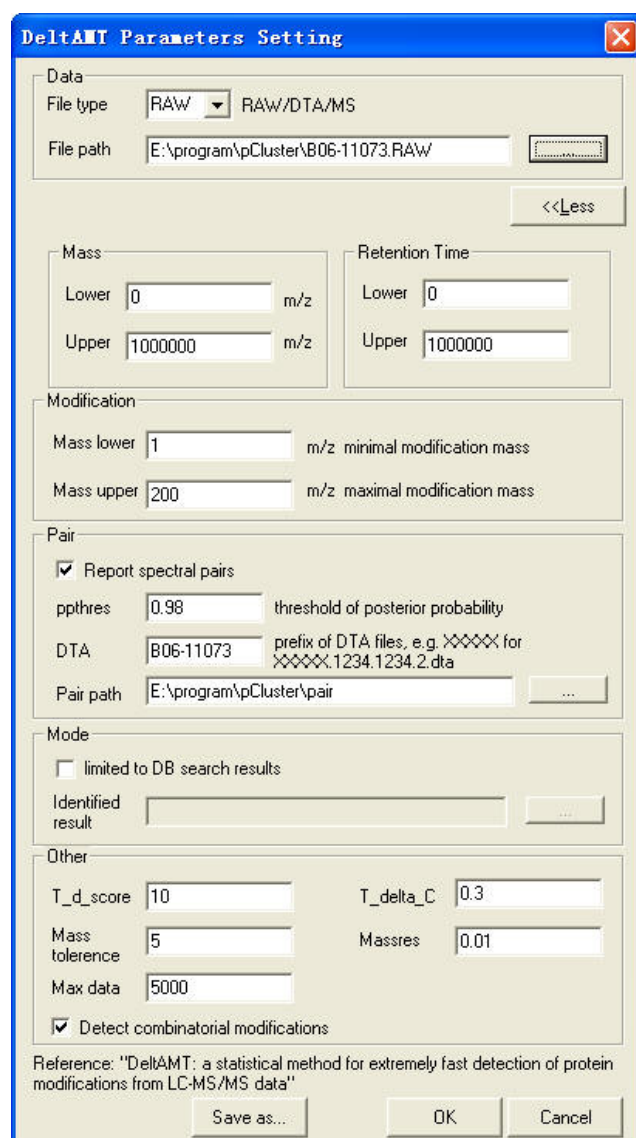
Precursor-based clustering

Click the DeltAMT icon to start the DeltAMT algorithm for modification detection. Figure 2 shows the DeltAMT parameter setting dialog. DeltAMT supports three types of data formats: RAW, DTA, MS2. To read data in the Thermo RAW format, Bioworks must be installed. Click the “...” button to select the data file/directory and then click the “OK” button to start analysis.



Fig 2. DeltAMT parameter setting dialog.

Click the “More >>” button to set more parameters, as shown in Figure 3.



Data

File type: data format

File path: data file path

Mass

Lower: lower bound of masses to be used

Upper: upper bound of masses to be used

Retention Time

Lower: lower bound of times to be used

Upper: lower bound of times to be used

Modification

Mass lower: lower bound of mass shifts

Mass upper: upper bound of mass shifts

Pair

Report spectral pairs: report or not

ppthres: threshold for posterior probability

DTA: DTA file name

Pair path: path to save pair data

Mode

Limited to DB search results: enable or not

Identified result: DB search results

Other

T_d_score: threshold for D-score

T_delta_C: threshold for delta C

Mass tolerance: used to remove redundancy

MassRes: mass resolution

Detect pseudo-modifications: enable or not

Fig 3. The whole parameters setting dialog of DeltAMT.



Fragment-based Clustering

Click this icon for PTM detection using fragment-based clustering algorithm. A parameter setting dialog shows after clicking the button. Figure 4 displays the spectral clustering parameter setting dialog.

The image shows a 'Cluster Dialog' window with the following fields and controls:

- Experiment name: 2010_Apr_29
- With Constraint:
- Input section:
 - Data type: dta
 - Browse as directory:
 - Data path: [empty] with a Browse button
- Output section:
 - Representative Spectra Path: [empty] with a Browse button
 - Clustering Result: e:\program\pCluster\pCluster\cluster.csv with a Browse button
- Similarity section:
 - Similarity Method: Dot Product
- Cluster section:
 - Consensus Spectra Method: Consensus
 - Threshold: 0.60
- Buttons: Advance>>, Save as, Start

Fig 4. Spectral clustering algorithm parameters setting dialog.

Click “Advance>>” button to set more parameters. See Figure 5.

All parameters are explained below.

“**Experiment name**”: the name for the parameter file with date default.

“**With constraint**”: with this button check means that what you do is a restrictive spectral clustering. Spectra from the same peptide are clustered together. And each cluster generates a representative spectrum. With this button uncheck means that what you do is an unrestrictive spectral clustering. Spectra from modified and unmodified peptide are clustered together.

“**Data type**”: here we support seven kinds of data formats: DTA, DTAS, MGF, PKL, MZML, MS2, RAW.

“**Browse as directory**”: check this button browse data files as a directory, so all data in this directory are included.

“**Data path**”: the path of data.

“**Representative spectra path**”: the path of representative spectra when doing restrictive clustering.

“**Clustering result**”: clustering result file.

“**Similar method**”: the similar method when comparing two spectra.

“**Consensus spectra method**”: the method of generating representative spectra.

“**Threshold**”: the threshold of similar score.

“**Spectral filtering**”: spectra with fragment peak number less than this number will be filtered from data.

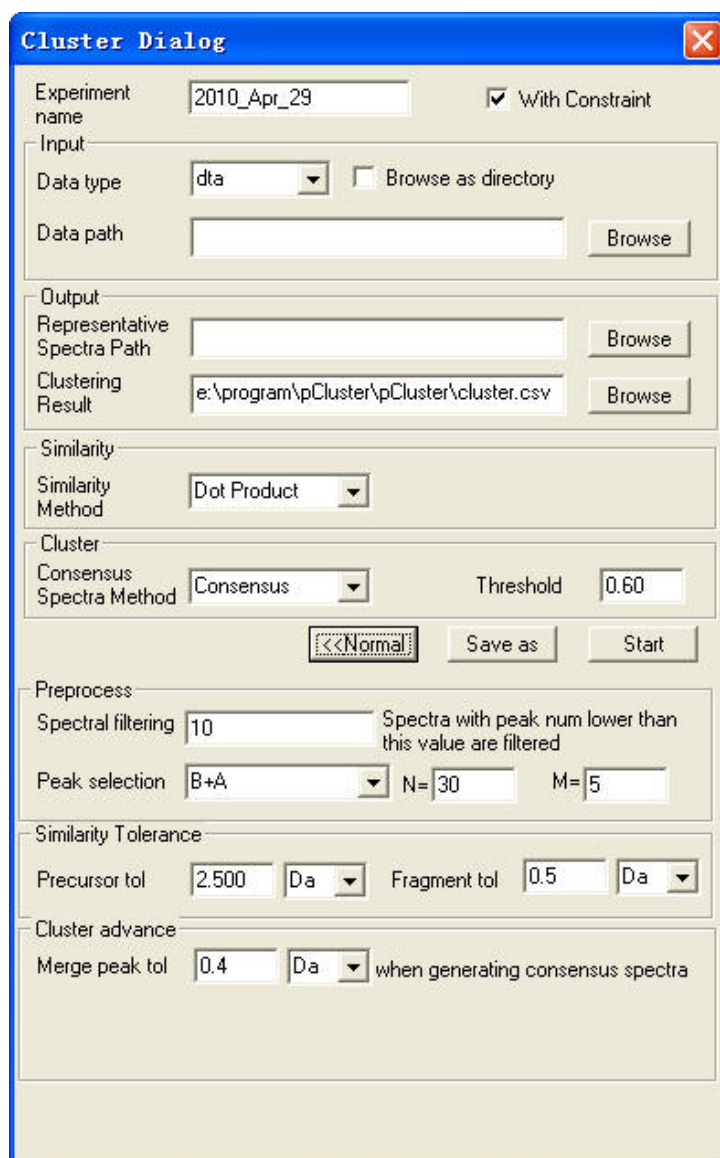


Fig 5. The whole parameters setting dialog of the spectral clustering algorithm.

“Peak selection”: method of how to select signal peaks.

“N=”: the top N highest intensity peaks of all peaks.

“M=”: the top M highest intensity peaks per 100peaks.

“Precursor tol”: precursor tolerance.

“Fragment tol”: fragment tolerance.

“Merge peak tol”: tolerance of merging neighbor peaks when using consensus spectra as representative spectra.

Output of spectral clustering algorithm:

1. When doing restrictive spectral clustering, it will generate a clustering result file and a series of representative spectra.
2. And when doing unrestrictive spectral clustering, it will generate a clustering result file and a series detected modifications and some relative information.
3. As the number of representative spectra is much less than original spectra and if using

consensus spectra as representative spectra, the consensus spectra are higher quality than original ones. So before doing unrestrictive spectral clustering, it is recommended to do the restrictive spectral clustering and using representative spectra as the input data of unrestrictive spectral clustering,

Display of detected modifications

In the view zone, detected modifications are displayed as a histogram, in which the horizontal axis is the modification mass and the vertical axis is the number of spectral pairs for each modification. In the list zone, the details of detected modifications are listed, including the mass shifts, retention time shifts, scores, types and inferred modification entries in the UniMod database.

Delta mass: estimated mass shift for each modification

Pairs: Number of spectral pairs detected for each modification

Delta time (for DeltAMT only): estimated retention time shift

D-score (for DeltAMT only): D-score of each modification

Type (for DeltAMT only): modification type, 0 for mono, 1 for combinatorial, 2 for differential

Mass1 (for DeltAMT only): involved modification masses for pseudo-modifications

Mass2 (for DeltAMT only): involved modification masses for pseudo-modifications

Unimod: inferred modification entries in the Unimod database

Click any item in the list, if the mass shift matches any entries in the Unimod database, the details will be shown in a new dialog.

Propagation

After modification-related spectral pairs are detected, peptide identifications obtained from sequence database searches can be propagated among these spectra pairs. At present, peptide identification files that pCluster can read are those exported by the pBuild tool (<http://pfind.ict.ac.cn/pBuild.htm>). The peptide propagation dialog (shown in Figure 6) can be opened by clicking on the 'Propagation' button.

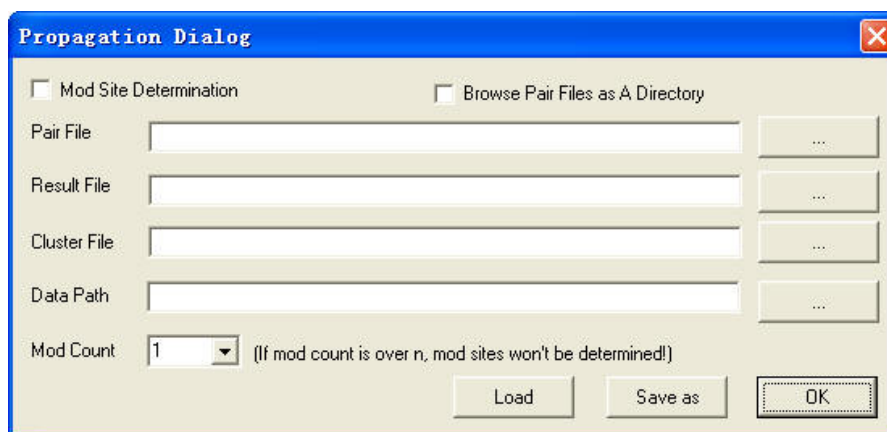


Fig 6. Propagation dialog.

“Mod site determination”: check this box if you want to do modification site localization.

“Browse pair files as a directory”: check this box if you want to do propagation in a batch mode.

“Pair file”: spectral pair files after DeltAMT or unrestrictive spectral clustering.

“Result file”: peptide identifications obtained by sequence database searches.

“Cluster file”(not required by DeltAMT): if you have done a restrictive spectral clustering, an inner-cluster propagation can be carried out by specifying the clustering result file.

“Data path”: spectral data path.

“Mod count”: maximal number of modifications on a peptide allowed for modification site localization.