# User manual for FTDR 2.0(not finished)

## 1.main GUI, including these items,

(1) Database search results selection panel. Supports formats: Mascot \*.dat, \*.html (version 1.9, 2.1, 2.2), Sequest \*.out, pepXML (with PeptideProphet validation, needs error tables to select the modeling data).

(2) Raw file selection panel. Only support Thermo raw files( \*.raw) because the calibration model may use the FT analyzer status data and the tailer extra data, which has not been included in the mzXML or mzML files currently.

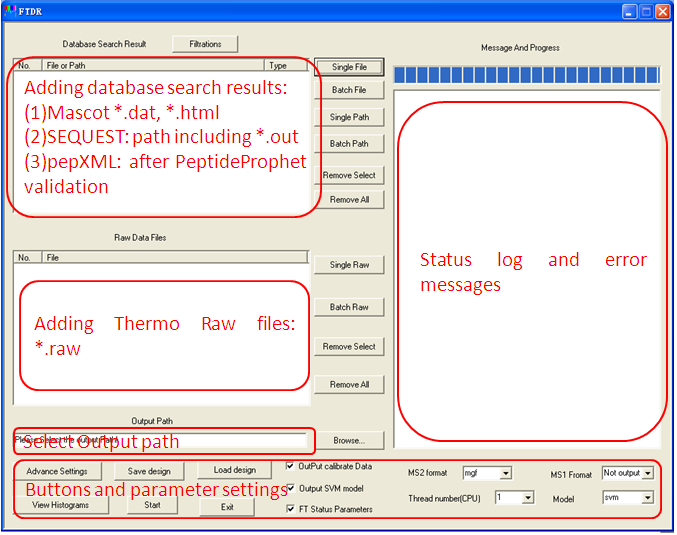


Figure 1. The main interface of FTDR 2.0.

### (3) The basic setting include:

a. output ms2 format selection, non for no output, mgf means general mgf files, extend mgf files will include 2 more header items: STOL and MSNUM. STOL means specific mz error tolerance for this spectrum, the unit is ppm. MSNUM means the number of parent ion observations in MS spectra.

b. ms1 format, non for no output, mzXML and mzML means the corresponding format.

c. Thread num: work thread number, if the computer has more than one cpu and the memory is enough, you can select up to 10 threads to proceed the calibration task. Do not make the thread number larger than the number of raw files because FTDR only take the calibration for a raw file as one unit, and do not distribute the calculations into different threads.

d. model selection. FTDR 2.0 supports linear model as used in FTDR 1.0, but with more parameters. A non-linear model using support vector machine (SVM) is more prefer to get a more accurate calibration. Also, you can select no model calibration, in which FTDR will only weighted average all the observed m/z in the XIC range. Generally, the svm model will consume more process times and linear model will be very fast.

e. check the “FT status parameters” checkbox if you wan to a more accurate calibration.

f. check the “Output SVM model” if you want to know the details of the SVM model.

g. check the “Output calibrate Data”. If you want to inspect the modeling data.

### (4) Advance settings.

The advace parameters used by FTDR 2.0 can be changed by push the button “Advance Settings” in the main GUI. A new dialog will appear as Figure 2.

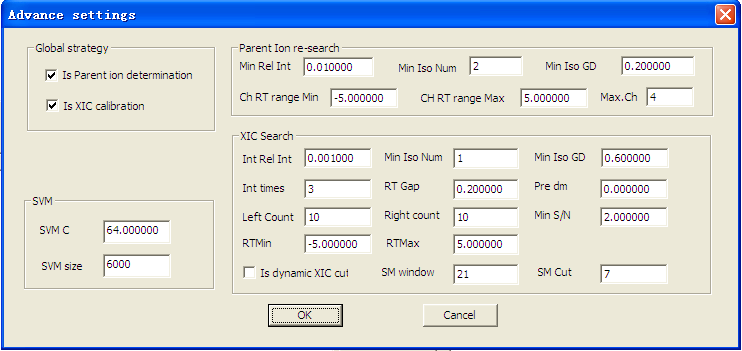


Figure 2. Adcanced parameters.

The parameters including:

a. global strategy: if the checkbox “Is Parement ion determination” is seected, FTDR will re-determine the parent m/z in the isolution window (generally 2.0 T), find all possible isotopic profile in this range by the group isotopic fitting (GIF) algorithm (see the detail in “method” section of the paper). GIF wil determine the monoisotopic m/z and charge state automatically. Multiple parent ions may be appear for one MS/MS spectrum.

If the checkbox “Is XIC calibration” is unchecked, FTDR will only using the parent ion information located in the nearby MS (before) spectrum to calibrate the m/z, do not perform the XIC search. If it is checked, FTDR will search a XIC for the parent ion, and then calculate all the calibration m/zs, a square intensity weighed m/z will be output as the final results.

b. SVM training parameters, including the sample size and the penalty factor (C). The larger value for C (not great than 256 generally in this problem) will generate more fitness model and taking more training time. More taining samples will generate more generalized models and consuming more training time. Note, the training samples will be randomly selected from the raw observations collected from the database search result and raw file pairs.

c. Parent ion re-search parameters. These parameters are used to filter the GIF results, including the mininal relative intensity(Min Rel Int), mininal isotopic peaks (Min Iso Num), minimal isotopic fittness (Min Iso GD), charge search retention time left range (charge RT Range Min), charge search retention time right range (charge RT Range Max), and Maximal charge (Max.ch).

Note:

The relative intensity is normalized to the base peak in the MS spectrum. The default value is 0.01.

The mininal isotopic peaks is required larger than 2, otherwise, the charge state can not be determined for a isolated peak.

The goodness of the isotopic match is calculated as the cosine of the experiment and theoretical distribution vector (Formula 1). The default value is 0.2, a relative low cutoff to let almost all the possible isotopic clusters pass. If this value is set too high, you will miss some low quality isotopic profiles in the MS2 nearby full scan, but may be very good in other MS1 spectra.

 (1)

Where , are the experiment and predicted isotopic instensities respectively. *N* is the number of isotopic peaks taken into account.

The distribution of the GD calculated by formula 1 on dataset ISB FT Mix3.

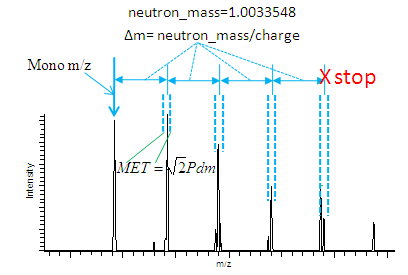


Figure 3. Experiment isotopic profile detection rules. Begin with the given Mono-m/z, searching the signals in a m/z error tolerance (MET) determinated bins with the center distancing the mono-m/z with the integral number of neutron m/z (=neutron\_mass/charge). If the next bin gives no signals, the search will be aborted. In this procedure, MET is , where *Pdm* is the pre-determined parent ion m/z error tolerance, which is used to match the experiment and theoretical values. MET is the error tolerance for the 2 experiment values. If we assume the experiment obxervations is norml distributed variables, the standard deviation of the difference of 2 random variables is  times of that of the random variables.

The experiment isotopic peaks is dectected using the rules shown in figure 3. The theoretical isotopic distribution is predicted from the natural mass by the formula (Formula 2) proposed by Valkenborg De et al[1]. We slightly modified the parameters(Table 1), by regressing this model on the accurate distribution data[2] form the trypisn digested peptides (up to 2 miss-cleavage) in IPI Human.3.49 (<ftp://ftp.ebi.ac.uk/pub/databases/IPI/old/HUMAN/ipi.HUMAN.v3.49.fasta.gz> ).

(2)

Table 1. The parameters for the iostopic distribution prediction.

|  |  |
| --- | --- |
| ***ai*** | ***bi*** |
| 1.007 | 0.0005792 |
| 0.0006321 | -0.09212 |
| 0.0005683 | 0.02292 |
| 0.0005526 | 0.09675 |
| 0.000568 | 0.1138 |
| 0.0005795 | 0.1215 |

The parent ion re-search will deterimine the charge of a peak by merging the MS signals in the [MS2RT+charge\_RT\_Range\_Min, MS2RT+charge\_RT\_Range\_Max] if the charge state can not be determined in the nearby pre-full MS scan of the MS/MS spectrum. The m/z values of the merged sub-spectrum is calculated by the intensity weighted average method[ ], and the signal intensities is the summary of all the combined signals. MET mentioned above is used to determine which peaks can be merged. The default value of charge\_RT\_Range\_Min, is -5, and +5 for charge\_RT\_Range\_Max.

#### d. XIC search parameters.

Background:

In LC-MS analysis, a peptide will be eluted in a time range ( from less than 1 to dozens of mins) and its parent ions will be observed many times. The observed signals can be used to reconstruct the chromatographic peaks, which is called extracted ion chromatogram (XIC[] or EIC[]). The measured m/zs can be viewed as the observations to the actual m/z with normal distributed random errors. According to the law of large numbers, we can reduce the random errors by averaging all the m/z observations. Another finding is that the random m/z error distributed more centralized with the increase of the signal intensity. That means, peaks with larger intenisty have a more accurate m/z observation[].

By taking into account the two above understandings, a more accurate m/z can be expected as fourmula 3.

 (3)

Where *Inti* is the observed parent ion intensity, *mzi* is the observed m/z.

As a case, Figure 4 gives the XIC of peptide “IFLNK”, and the idenfitied information is listed in Table 2. We extracted the observed m/z errors (in ppm) and intensities (Figure 5), the data proved the above 2 observations.

In MaxQuant[4], the 3D peaks feature detection step also ultilzed this rule, the calibration of MaxQuant is also proved to be effective.

As a more generalized investigation, a scatter plot of m/z measurement errors and signal intensities in gived in Figure 6 for dataset ISB FT Mix 3.

FTDR 2.0 implemented the XIC based global calibration with the parameter based local calibration. This section will introduce the XIC search parameters used in FTDR 2.0.

“Int Rel Int” means the relative intensity (normlized to the base peak) cutoff for parent ions. Generally, it is set to 0.001 according to the signal dyamic range of MS1 full scan.

“Min Iso Num” means the minimal retrived isotopic peaks, it is 1 at least which means that the monoisotopic m/z must be found.

“Min Iso GD” means the cutoff of goodness of isotopic profile match( to prediction ones). The default value is 0.2.

“Int times” (default 3) means interupt times tolerated by the searching. If FTDR can not find the parent ions in MS1 spectra with interupt times, the search will be trunked.

“RT gap” (default 0.2 min) has the same effect as “Int times” but it uses the absent RT time bins as a index.

“Min S/N”(default 2), the cutoff of signal to noise. The noise data is read out by Xcalibur COM interface.

“Left count” (default 10) and “Right count”(default 10). As shown in Figure 7, these 2 parameters (named as Lc and Rc) are used to limited the XIC time range by counting the observed parent ions. FTDR will extend finding the XIC signals from the MS2 present times, which can be convert from the scan number of the MS2 spectrum. In the training data collection step, it will be collapsed because of memory overflow if there a too many cases for high abundance peptides. Thus, FTDR need to trunk the collections by these 2 parameters.

“RTMin” (default -5, must be negative values) and “RTMax”(default +5, must be postitive values). These 2 parameters defined a time range based on the MS2 spectrum present time to trunk the XIC in the calibration step. If the “Is dynamic XIC cut” check box is not checked, these 2 parameters will be effective.

Dynamic XIC cut is implmented by a online Savitzky–Golay (SG) smoothing (adding the data one by one when searching the XIC). If a local minimum is found, the XIC extending procedure will be stoped. The local minimum is defined with *n* decreasing neibours in the left and *n* increasing neibours in the right, where *n*=(SM\_Cut-1)/2, and SM\_Cut is user input in the edit box “SM Cut”(default 7, this number must be odd). The window of SG smoothing is user defined by inputting it in edit box “SM window” (default 21, you may use 7 for short XIC. This number should be odd.) and the order of the local fitting polymerization is 2. The SG smoothing is implemented using GNU Scientific Library (GSL) according the algorithm introduced in ref 5.

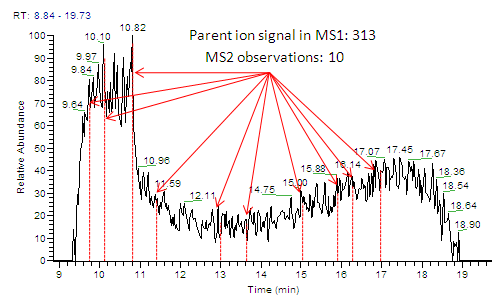


Figure 4. XIC of peptide “IFLNK”. 313 parent ion signals are found in the MS1 spectra, only 10 MS2 spectra were collected for it. The case from the 18 mixtrure standard proteins sample of Institute of System Biology (ISB)[3]. The Mascot database search results are list in Table 2. The m/z after calibration is with a error to the value ppm.

Table 2. Mascot database search records for peptide “IFLNK” in the ISB FT Mix 3 dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Scan | Score | Delt mass (Da) | Delt mass (ppm) |
| 581 | 30.37 | 0.001527 | 2.403208 |
| 610 | 24.03 | 0.001263 | 1.987722 |
| 654 | 24.02 | 0.001479 | 2.327665 |
| 683 | 27.39 | 0.001579 | 2.485046 |
| 774 | 24.00 | 0.001729 | 2.721118 |
| 802 | 25.87 | 0.001389 | 2.186022 |
| 870 | 29.41 | 0.001559 | 2.45357 |
| 918 | 24.20 | 0.001393 | 2.192318 |
| 943 | 26.40 | 0.001371 | 2.157694 |
| 985 | 17.20 | 0.001033 | 1.625746 |

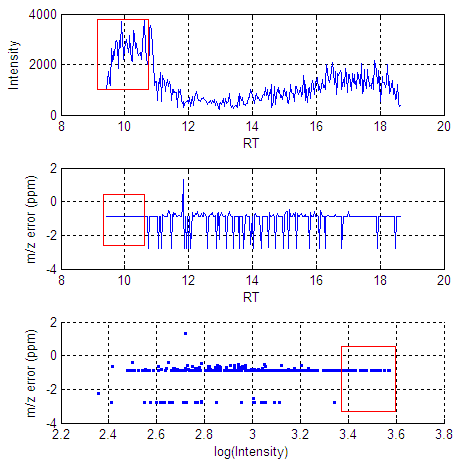
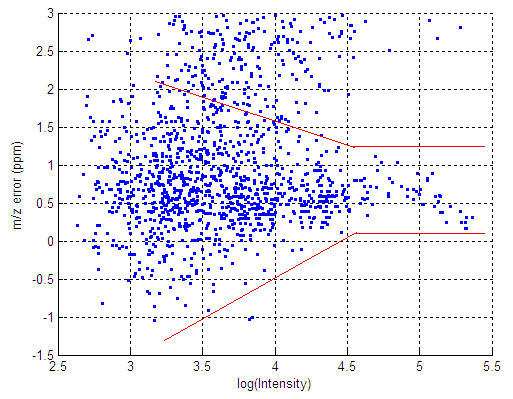


Figure 5. The XIC and m/z measurement errors. In the middle and bottom view, we can clearly see that the m/z error variance is little for high signals (red box).



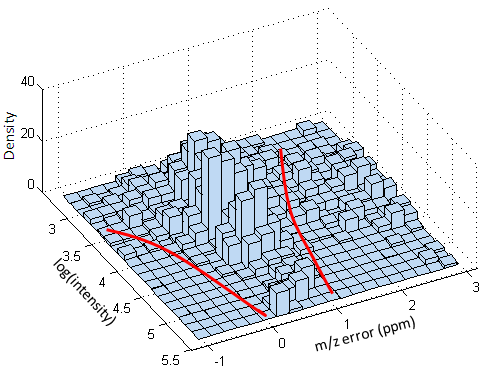


Figure 6. m/z measurement errors decrease with the increasing of intensities. Top view is the scatter plot, and bottom view is the 3d histogram.

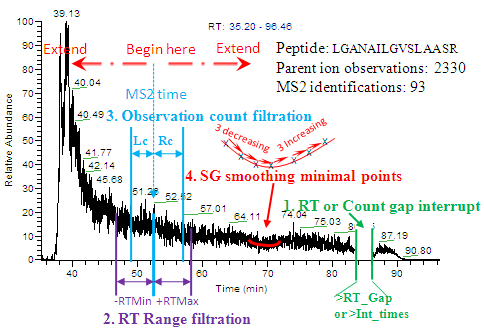


Figure 7. XIC search filtrations in FTDR 2.0. 4 strategies were implemented to trunk the XIC for different usage. The 1st on is using the RT or Count gap to give a continuity restraint, which was used in any XIC searching step in FTDR. Thw 2nd is a RT range filtration, only used in the calibration step and will be automatically disabled when the 4th rules was used. The 3rd rule is used to limited the volume of training datasets by counting the observations. It is only used in the model training step. The 4th strategy want to give a more reasonalbe cutoff by ultilzing the shape characteristic of XIC curve. But it will be failure for the split chromatographic peaks[6] as shown in Figure 4. This case is from the LTQ/FT dataset published in ref 7.

### (5) functional buttons. FTDR is operated by clicking the buttons on the main GUI.

a. click the “Filtrations” to set the database search result validation rules (Figure 8). Only Mascot and Sequest search results need these rules, pepXML files will be filtered by the built-in error table with a estimated FDR of 0.01.

b. click “Save design” to store the current setting and file pathes. All the parameters will be stored in the given file, and you can load them by click “Load design” button. These 2 operations can facilitate the try and error parameter tuning.

c. The m/z error distributions can be viewed by clicking the “view histograms” button after the calibration tasks are finished. The popup dialog will gives 2 normalized histograms to caompare the m/z error distributions before and after calibration. Parameters and curves of normal distribution fittings will be given also (Figure 9).

d. “start” button can start the tasks, and will be disabled until all the tasks have been finished.

e. “Exit” button can withdraw from FTDR, even the tasks are runing. All the tasks will be lost if they are not finished.

f. File or path adding button, you can operate with a single or batch file or path.

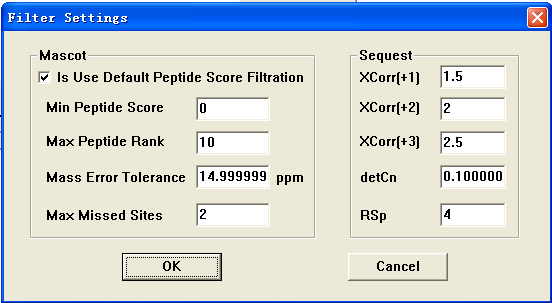


Figure 8. Database search results validation for Mascot and Sequest. These parameters is obvious and has been introduced in ref 8.

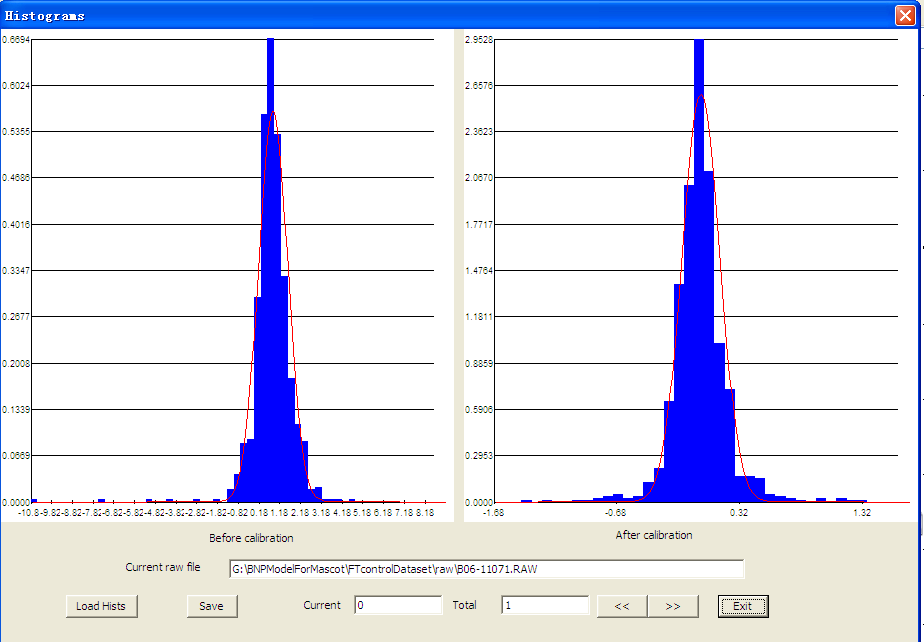


Figure 9. m/z error histograms before and after calibration. The red curves are normal distribution fittings.

## Reference

1. Valkenborg De et al. Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach.Rapid Commun Mass Spectrom. 2007;21(20):3387-91.

2. Yergey J. A general approach to calculating isotopic distributions for mass spectrometry. Int. J. Mass Spectrom. Ion Phys. 1983, 52, 337-349.

3. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng JK, Aebersold R, Martin DB. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. J Proteome Res. 2008 Jan;7(1):96-103.

4. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008 Dec;26(12):1367-72.

5. S.J. Orfanidis.Introduction to Signal Processing.Prentice-Hall, Englewood Cliffs, NJ (1996) (Chapter 8).

6. Sadroddin Golshan-Shirazi, Georges Guiochon. Combined effects of finite axial dispersion and slow adsorption desorption kinetics on band profiles in nonlinear chromatography. J. Phys. Chem., 1991, 95 (16):6390–6395.

7. Liu K, Zhang J, Wang J, Zhao L, Peng X, Jia W, Ying W, Zhu Y, Xie H, He F, Qian X. Relationship between Sample Loading Amount and Peptides Identification and Its Effect on Quantitative Proteomics. Anal Chem. 2009;81(4):1307-14.

8. Zhang J, Ma J, Dou L, Wu S, Qian X, Xie H, Zhu Y, He F. Mass Measurement Errors of Fourier-Transform Mass Spectrometry (FTMS): Distribution, Recalibration, and Application. J. Proteome Res. 2009; 8 (2), 849–859.