IPeak user guide

version 1.0

1 Contents

1 Overview

IPeak is a peptide identification pipeline that is designed to combine a post-processing algorithm and multi-search strategy to enhance the sensitivity of peptide identifications without compromising accuracy. Currently, it can combine results from OMSSA, X!Tandem and MS-GF+. It can be used either in command line mode or in GUI (visualized) mode. It was implemented into the mzidLibrary and is available to download from the download page (https://code.google.com/p/mzidentml-lib/). Some example files and a user guide file are contained in IPeak software package.

2 Software requirements

The following software must be installed before you run IPeak.

Java: Java 1.7 or newer is required. If your operating system don't install Java, please download and install it from the website http://java.com/download.

Percolator: After you download IPeak software and unzip it, please run "check_percolator.bat" (Windows) or "check_percolator.sh" (Linux) to make sure Percolator work well in your computer. If it doesn't work, please download and install the suitable Percolator package according to your operating system from the website http://www.nada.kth.se/~lukask/percolator_dist/. And then

please copy the executable percolator file to folder “percolator” in IPeak software and rename it to percolator_win.exe (for Windows) (percolator_linux.exe for Linux, percolator_mac.exe for OS X). And finally, add the installed Percolator path into the list of the current environment variable settings. Currently, only version 2.01~2.04 are supported. I recommend that you use the version 2.04. We have done much test based on this version. Currently, IPeak can’t work with the latest version of Percolator 2.05, but it will be supported in the future.

3 Run IPeak

IPeak can be used either in command line mode or in GUI (visualized) mode.

a. Run in GUI mode

Please run IPeak with GUI mode through double click the file “run_IPeakGUI_Windows.bat” (Windows system) or run “sh run_IPeakGUI_Linux.sh” (Linux system and OS X). After you double click this file, then you will see a GUI like below:

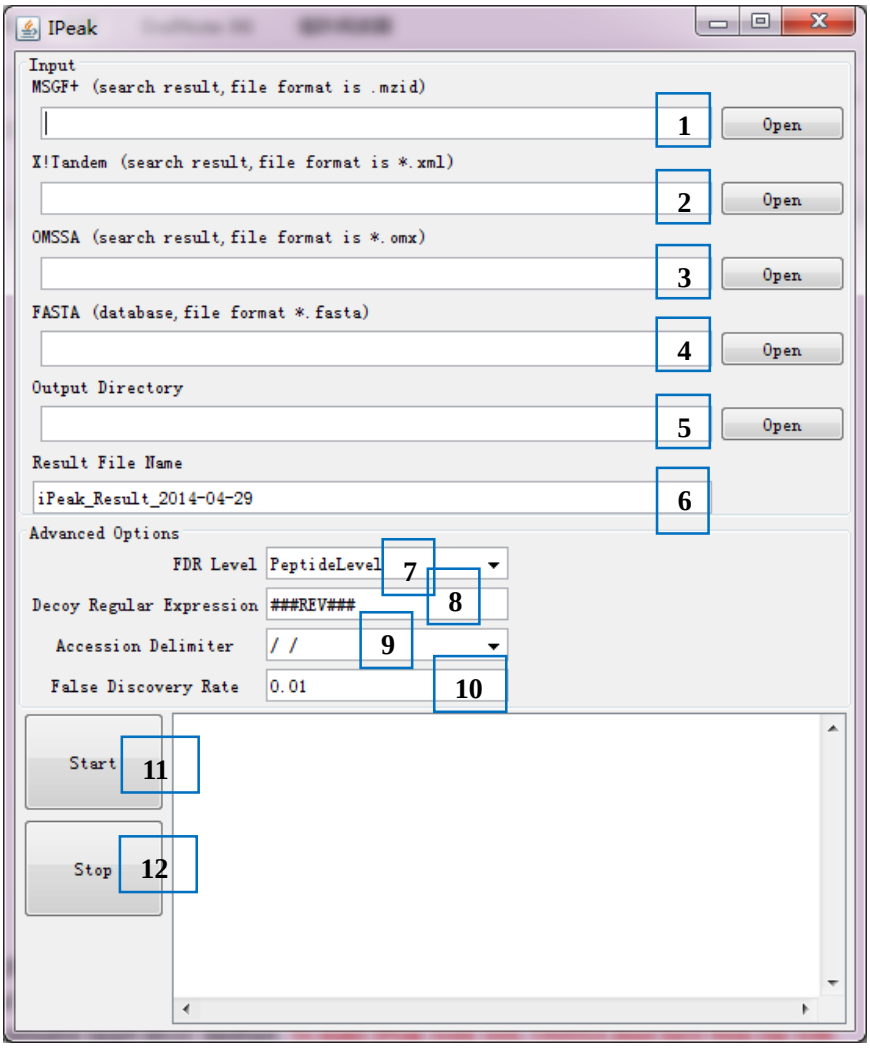

Figure 1. The interface of IPeak

IPeak takes the native format search files (MS-GF+ “.mzid”, OMSSA “.omx”, X!Tandem “.xml”) and fasta format protein database as inputs. The search engines must have been run using a

concatenated target-decoy database. To make IPeak work well, OMSSA must have been run with the option "-w" include spectra and search parameters in search results; X!Tandem must have been run with option "output, parameters = yes" when performing the search (http://thegpm.org/tandem/api/index.html); MS-GF+ must have been run with option "-addFeatures 1" to output additional features used for Percolator.

An example setting for the input parameters is shown below:

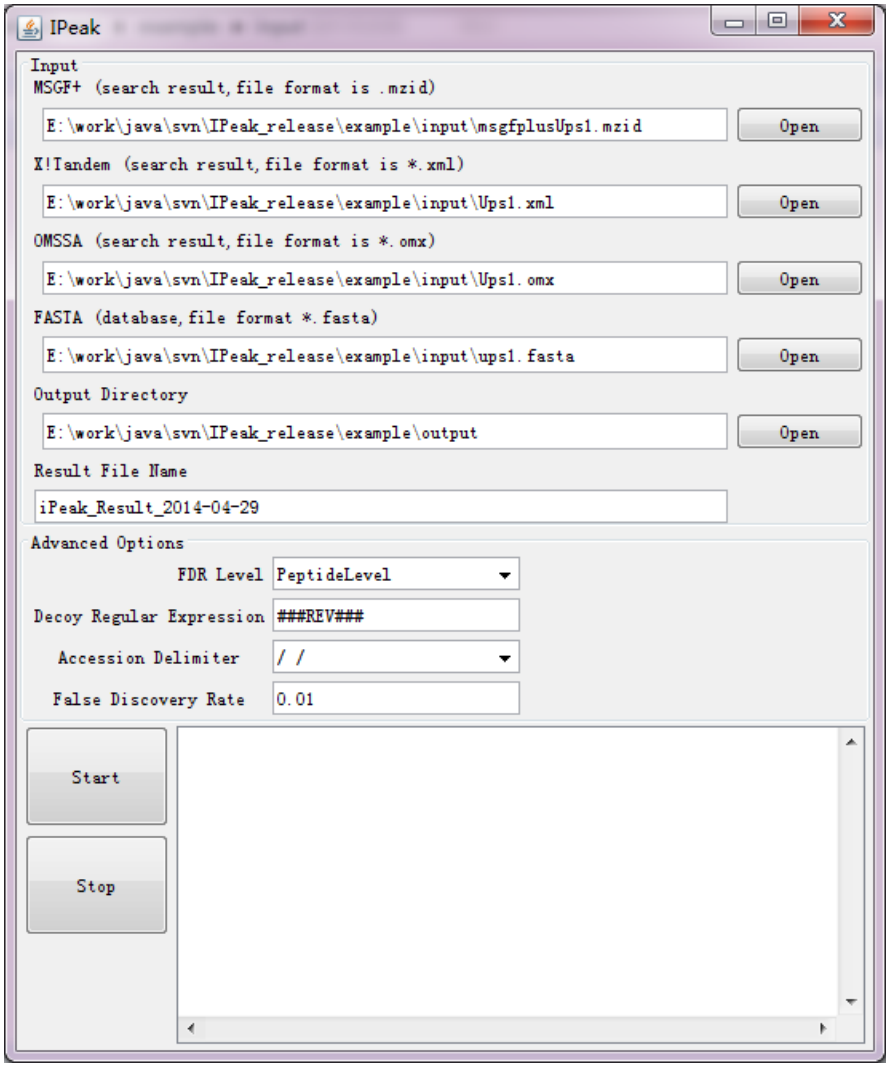

Figure 2. An example setting for IPeak

When the parameters are set correctly, you can click the "Start" button to run the program. You can use the example files in "example/input" directory to do a test. A detailed explanation about the parameters in GUI is shown below (the number of following are consistent with that on figure 1):

⑴. The result file of MS-GF+, mzid format;

⑵. The result file of X!Tandem, xml format;

⑶. The result file of OMSSA, omx format;

⑷. The concatenated target-decoy database for searching. You can use the Perl script decoy.pl to generate the decoy database. This script can be downloaded from website http://www.matrixscience.com/help/decoy_help.html;

⑸. The directory of output files;

⑹. The prefix of the output files;

⑺. Filter result in peptide-level or PSM-level;

⑻. The symbol of decoy sequences;

⑼. The char of used to separate the protein ID and protein description in database fasta file, surrounded by forward slashes, e.g. "/ /". Usually it is space, you can use the default value;

⑽. The FDR threshold value to filter the result;

⑾. Start to run the program;

⑿. Stop the program when it is runing.

When the program is end, you can find the result files in the output directory. Usually, three files are useful for the user: *combined_percolatorProteinSummary.txt (containing the protein group information), *combined_percolatorPSMSummary.txt (containing the peptide identification result), and *combined_percolator.mzid (a mzid file about the result).

b. Run in command line mode

To run IPeak from the command line, usually, you can use like below:

```
java -Xmx2G -cp mzidentml-lib.jar bgi.ipeak.IPeak -xtandem example/input/Ups1.xml -omssa example/input/Ups1.omx -msgf example/input/msgfplusUps1.mzid -database example/input/ups1.fasta -ofr test -mod mods.xml -pp /path/percolator.exe -outdir example/output/
```

You can find some example files in directory “example/input” and you can double click the file “run_example_CommandLine_Windows.bat” (in Windows system) or run “run_example_CommandLine_Linux.sh” (in Linux system) to see how it work. Please change the percolator path to that in your system before you run the example.

A complete list of arguments can be found below:

Options	**Required?**	**Description**
-msgf	Required	MS-GF+ result file (*.mzid format): a single file or a directory containing multiple files
-omssa	Required	OMSSA result file (*.omx format): a single file or a directory containing multiple files
-xtandem	Required	X!Tandem result file (*.xml format): a single file or a directory containing multiple files
-database	Required	The concatenated target-decoy database for searching. You can use the Perl script decoy.pl to generate the decoy database. This script can be downloaded from website http://www.matrixscience.com/help/decoy_help.html
-decoyregrex	Optional	The symbol of decoy sequences, default is ###REV###
-outdir	Required	The directory of output files
-prefix	Required	The prefix of the output files
-maxfdr	Optional	The FDR threshold value to filter the result
-u	Optional	Filter result in peptide-level FDR (true or false), default is true
-accessionSplitRegex	Optional	The char of used to separate the protein ID and protein description in database fasta file, surrounded by forward slashes, e.g. "/ /". Usually it is space, you can use the default value;

-delete	Optional	Remove PSMs that don't pass the threshold
-debug	Optional	Just for testing

When the program is end, you can find the result files in the output directory.

4 Call Percolator on individual search engine result file

a. Call Percolator on X!Tandem result

INPUT_FILE	[input].xml
OUTPUT_FILE	Output directory
PARAMS	-decoyRegex decoyregex -compress true|false
Description	Running XtandemPercolator

Example command line:

java -Xmx2G -jar mzidentml-lib.jar XtandemPercolator example/input/Ups1.xml example/output -decoyRegex "###REV###" -compress false

b. Call Percolator on OMSSA result

INPUT_FILE	[input].omx
OUTPUT_FILE	Output directory
PARAMS	-database database -decoyRegex decoyregex -compress true|false
Description	Running OmssaPercolator

Example command line:

java -Xmx2G -jar mzidentml-lib.jar OmssaPercolator example/input/Ups1.omx example/output -database example/input/ups1.fasta -decoyRegex "###REV###" -compress false

c. Call Percolator on MS-GF+ result

INPUT_FILE	[input].mzid
OUTPUT_FILE	Output directory
PARAMS	-decoyRegex decoyregex -compress true|false
Description	Running MsgfPercolator

Example command line:

java -Xmx2G -jar mzidentml-lib.jar MsgfPercolator example/input/msgfplusUps1.mzid example/output -decoyRegex "###REV###" -compress false

5 Common errors when using IPeak

The most common problems maybe result from the following:

- The input file is not in the correct format. You must do the database searching according to the instruction in section 3.a.
- The parameters are incorrectly specified, please check these carefully against the instructions before submitting a bug report.
- Out of memory errors. You may need to increase the memory available to Java by editing the command for the routine or the graphical interface. Example providing a large amount of memory (4G): java -cp mzidentml-lib.jar bgi.ipeak.IPeak

If you have any suggestion or question about IPeak, please don't hesitate to contact me: wenbo@genomics.cn.