Mayu Software Manual

General

Mayu is a software package to determine protein identification false discovery rates (protFDR) and peptide identification false discovery rates (pepFDR) additionally to the peptide-spectrum match false discovery rate (mFDR).
This software is licensed under the CC-GNU GPL version 2.0 or later. This software and any associated documentation is provided “as is” and there is no warranty for this software.

Installation

Mayu can be run directly after the unpacking of the zip file if a perl interpreter is installed. To run Mayu open a command line change to the directory where Mayu.pl is located (use the command cd) and type 'perl Mayu.pl' on the command line for a help.

> Install a perl interpreter (download e.g. the ActivePerl language distribution from www.activestate.com).

For graphical output the R statistical package needs to be installed and R recognized as a command on the command line.

> Download the R package from http://www.r-project.org/ and install it on your system. Add the path to the R binaries (e.g. C:\R-2.4.1\bin) to your path environment variable.

The program will run without additional perl module installation, however an xml parser for proper pepXML parsing can be installed (-xmlparser option after installation of the parser).

> If you wish to use an xml parser for pepXML parsing, install the libxml parser on your system (required modules are XML::Parser::PerlSAX and XML::SAX::Base). Use the ppm program (Programmer's Package Manager) of the ActivePerl distribution or go to http://search.cpan.org to download the required package(s).

Prerequisites

- data has to be searched against one target decoy database (reversing recommended)
- search results formatted as pepXML, mascot .csv or comma separated Mayu table files
- target decoy fasta database that was used for the database search (use the script reverse_fasta.pl that is located in the folder var/ to create a concatenated target decoy database for your database of choice)

Recommendations

- keep sequence redundancy of the protein database as low as possible
- all data should be searched with similar options
- for proper peptide identification false discovery rates the data should be searched in fully tryptic mode prior to Mayu analysis

Options

Type 'perl Mayu.pl -manual' for a detailed description of the options

Run

Unzip the program to the directory of your choice and run the program in this directory from the command line with the command 'perl Mayu.pl'. This will print a help how to run the program with the proper input.

Examples:

1. standard analysis, main analysis table printed out
   'perl Mayu -B example.csv -C tardecdb.fa -v -s'
2. plot graphics using the R statistical package
   'perl Mayu -B example.csv -C tardecdb.fa -v -s -runR'
3. remove peptides smaller than 10 amino acids from target and decoy PSM
   'perl Mayu -B example.csv -C tardecdb.fa -D 10 -v -s'
4. do calculations of error rates in 51 steps between 0 and 5% PSM FDR
   'perl Mayu -B example.csv -C tardecdb.fa -G 0.05 -H 51 -v -s'
5. print out more result tables in separate files
   'perl Mayu -B example.csv -C tardecdb.fa -PmFDR -PbinProt -PprotFeat'
6. start a long run on a unix system and log the standard output
   'nohup perl Mayu -B example.csv -C tardecdb.fa ... -v > log.txt &'
7. print out target and decoy PSM, target PSM with a PSM FDR of 0.01
   'perl Mayu -B example.csv -C tardecdb.fa -P mFDR=0.01:td'
8. print out target PSM whose protein ids correspond to a protFDR of 0.05
   'perl Mayu -B example.csv -C tardecdb.fa -P protFDR=0.05:t'
9. use pepxml as input
   'perl Mayu -A sequest_pepxml.xml -C tardecdb.fa'
10. pepxml as input, print out a .csv file of input for faster reanalysis
    'perl Mayu -A sequest_pepxml.xml -C tardecdb.fa -Pio'
11. sort the LC-MS/MS runs by orthogonality (run is recognized by its scan base) and perform the analysis on cumulative data sets in 11 steps
    'perl Mayu -B example.csv -C tardecdb.fa -N 5 -O 11'

Input File Formats

Search results can be passed in three formats

1. pepXML (-A, .xml): This is an open format that was developed as part of the TPP. The format is described here http://sashimi.sourceforge.net/software_tpp.html
2. Mayu format (-B, .csv)
   a comma separated file with the following columns:
   1. scan (run.scannr.scannr.charge)
   2. raw peptide sequence
   3. protein identifier (decoy ids must have a prefix)
   4. modifications (pos1=mass1:pos2=mass2)
      position: position starting with 1, 0 and L+1 for N and C-terminal modifications respectively
      mass: amino acid mass minus water plus modification in dalton
   5. discriminant score (e.g. PeptideProphet probability score)

   example line representing a PSM:
   run1.2208.2208.2,KLAHDTKMLK,F02H6.4,8=147.192:10=147.192,0.6824
3. Mascot table format (-B, .csv)

The **search database** has to be provided in fasta format as a concatenated target decoy database. Decoy entries have to be marked with a prefix (-E <prefix>)

Output File Formats

- **_mFDR...txt**
  1. PPs: PeptideProphet probability score or any other discriminant score
  2. mFDR: peptide spectrum match (PSM) false discovery rate estimated using the target decoy strategy
  3. FP: false positive target PSM
  4. TP: true positive target PSM
  5. TD_mFDR: PSM false discovery rate (mFDR) for target and decoy PSM
  6. TD_FP: false positive PSM for target and decoy
  7. TD_TP: true positive PSM for target and decoy
  8. target_PSM: target PSM
  9. decoy_PSM: decoy PSM

- **_prot_size_local_FDR...txt**
  1. nr_runs: number of LC-MS/MS runs
  2. nr_files: number of input files
  3. mFDR: mFDR cutoff
  4. protein_size_bin: index of the protein size bin
  5. bin_desc: description of the boundaries of that protein size bin
  6. target_prot: number of target protein from the total database in that protein size bin
  7. target_protID: number of target protein identifications in that protein size bin
  8. decoy_protID: number of decoy protein identifications in that protein size bin
  9. FP_protID: number of false positive protein identifications in that protein size bin
  10. FP_protID_stdev: standard deviation of false positive protein identifications in that protein size bin (derived from the hypergeometric model)
  11. TP_protID: number of true positive protein identifications in that protein size bin
  12. protFDR: protein identification false discovery rate in that protein size bin

- **_feat_prot...txt**
  1. id: protein id
  2. mFDR: mFDR cutoff
  3. nr_files: number of input files
  4. nr_runs: number of LC-MS/MS runs
  5. NP: number of distinct peptide identifications mapping to this protein at this mFDR cutoff
  6. NS: number of PSM mapping to this protein at this mFDR cutoff
  7. PAT: PSM alignment type at this mFDR cutoff:
     0: one PSM
     1: two PSM on same peptide identification
     2: two PSM on distinct peptide identification
     3: three PSM on same peptide identification
     4: three PSM on two peptide identifications
     5: three PSM on three peptide identifications

6: more than three PSM
8. PSL: protein sequence length
9. acNTP: corrected number of tryptic peptides for this protein
10. decoy: 0 decoy id, 1 target id

- **_main...txt**
  1. nr_runs: number of LC-MS/MS runs
  2. nr_files: number of input files
  3. mFDR: mFDR cutoff
  4. target_PSM: target PSM
  5. decoy_PSM: decoy PSM
  6. FP_PSM: false positive target PSM
  7. TP_PSM: true positive target PSM
  8. target_pepID: number of target peptide identifications
  9. decoy_pepID: number of decoy peptide identifications
  10. FP_pepID: false positive peptide identifications
  11. FP_pepID_stdev: standard deviation of false positive peptide identifications (derived from the hypergeometric model)
  12. TP_pepID: true positive peptide identifications
  13. pepFDR: peptide identification false discovery rate
  14. target_protID: target protein identifications
  15. decoy_protID: decoy protein identifications
  16. FP_protID: false positive protein identifications
  17. FP_protID_stdev: standard deviation of false positive protein identifications (derived from the hypergeometric model)
  18. TP_protID: true positive protein identifications
  19. protFDR: protein identification false discovery rate
  20. target_protIDs: target single PSM protein identifications
  21. decoy_protIDs: decoy single PSM protein identifications
  22. FP_protIDs: false positive single PSM protein identifications
  23. TP_protIDs: true positive single PSM protein identifications
  24. protFDRs: single PSM protein identifications false discovery rate
  25. target_protIDns: target all but single PSM protein identifications
  26. decoy_protIDns: decoy all but single PSM protein identifications
  27. FP_protIDns: false positive all but single PSM protein identifications
  28. TP_protIDns: true positive all but single PSM protein identifications
  29. protFDRns: all but single PSM protein identifications false discovery rate

- **_psm...csv**
  1. scan: scan id
  2. pep: peptide sequence
  3. prot: protein id
  4. mod: modification info
  5. score: discriminant score
  6. decoy: decoy or target (decoy = 1, target = 0)
  7. mFDR: corresponding mFDR