NAME

DTarray_pro - Combine DTASelect-filter files.

SYNOPSIS

DTarray [**-i** <*mode*>] [**-o** <*output_files*>] [**-p** <*output_files*>] [**-lr** <*output_files*>] [**-g** <*group_method*>] [**-n** <*0/1*>] [**-modG** <*group_method*>] [**-modS**] [**-u**] [**-c**] [**-seqC**] [**-s** <*output_format*>] [**-saint** <*bait_file*>] [**-rev** <*0/1*>] [**-loc**] [**-fxn**] [**-mw** [<*fasta_file*>]] [**-mact**] [**-act** <*file*>] [**--unicode** <*0/1*>] [**-seq** [<*fasta_file*>]] [**-f** [<*prefix*>]] [**-d** <*dir*>] [**-flist** <*file*>] [**-rw** <*arg*>] [**--purge**] [**-h**, **--help**]

DESCRIPTION

DTarray_pro extracts Uniprot ID numbers, molecular weights, and spectral counts from .dtafilter files stored in the working directory. Protein data is combined into one dataset and written to the working directory as a tab delimitated text file (.tsv). At runtime, a file list named *dtarray_pro_flist.txt*, containing all the valid DTAfilter files found in the working directory, is generated. If a file list already exists, the existing file list is used. Various options are available to, modify how files are treated in the working directory, extract additional data from the DTAfilter files, and to change the output file format.

OPTIONS

Command line options are processed from left to right. Options can be specified more than once. If conflicting options are specified, later specifications override earlier ones.

INPUT/OUTPUT OPTIONS

-i, --in <*mode*>

Specify which input mode is used. **std** is the default.

std

Read in DTASelect-filter files stored in a common directory with the name <*sample_name*>.dtafilter

subdir

Read in DTASelect-filter files stored in subdirrectories where the sample name is the dirrectory name and the filter file is named *DTASelect-filter.txt*. In subdir mode, DTarray will search all dirrectories one level below the current working directory for valid DTASelect-filter files.

-o, --out <*output_files*>

Specify the output file format. **1** is the default.

0

Do not include protein output files.

1

Use wide output format.

2

Use long output format.

3

Generate both wide and long output files.

-p, --peptides <*output_files*>

Generate output file containing spectral counts for each peptide. **0** is the default.

0

Do not include peptide output files.

1

Use wide output format.

2

Use long output format.

3
Generate both wide and long output files.

-lr, --locReport <*output_files*>
Generate table of subcellular locations identified in each sample. Columns are included for count, sum of spectral counts, and total peptides identified. **0** is the default. **-s** and **-u** are supported in loc_report if specified. **-loc** option is also set automatically when loc_report is included.

0
Do not include loc_report.

1
Use wide output format.

2
Use long output format.

3
Generate both wide and long output files.

-saint <*bait_file*>
Generate input files for saintExpress. <*bait_file*> should be a text file with three columns separated by tabs: IP name (or sample name), bait name, and T if sample is test or C if sample is control. Two files are generated in working dirrectory: interaction_file.txt and prey_file.txt.

-d, --dir <*dir*>
Specify parent directory from which to run program. By default, the current working directory is used.

-flist <*file*>
Specify file list. By default the automatically generated file list is used.

RUN SPECIFIC SUPPLEMENTARY INFORMATION

-u, --unique
Include spectral counts for unique peptides in output file.

-c, --coverage
Include percent sequence coverage for proteins in output file.

-seqC Include sequence count for proteins in output file.

-s Specify how to group sup info columns in output file. **0** is the default. One or more run specific sup info options must be specified to use this option.

0
Group columns by sample.

1
Group columns by observation.

PROTEIN SPECIFIC SUPPLEMENTARY INFORMATION

-loc Use *~/scripts/DTarray_pro/db/humanLoc.tsv* to lookup subcelluar localization information for proteins in output file. humanLoc.tsv contains Uniprot annotations for subcelluar localization by Uniprot ID, updated as of Jan 18 2017. Currently, sub cell location information is available for human proteins only.

-fxn Use *~/scripts/DTarray_pro/db/humanFxn.tsv* to include panther category for proteins in output file.

-seq Use sequence information in *~/scripts/DTarray_pro/db/humanProteome.fasta*, to include protein sequences in output file.

<fafsa_file>

The user can optionally specify a fafsa formatted file to lookup protein sequences. The *<fasta_file>* used to calculate protein molecular weights does not have to be the same as the *<fasta_file>* used to search for protein sequences.

PEPTIDE FILE OPTIONS

-g, --group Specify how peptides are grouped in peptide output files. **1** is the default.

0

Do not group peptides. In this format, each peptide will be output on a separate line in a long formatted peptide output file. Columns with information specific to each scan, i.e. obsMH, parent file, and scan, are included in output file. Only long output format is supported for this group method.

1

Group peptides by parent protein. A separate entry for each charge state of a given peptide will be included in peptide output files.

2

Group peptides by parent protein and charge. Peptides found in multiple charge states will be grouped in output files.

-modG Specify how to group modified peptides in peptide output files. **0** is the default.

0

Peptides with the same sequence, but different modification status will not be grouped. A separate entry will be included for each modification status found for a peptide.

1

Ignore modification status when grouping peptides.

-n, --nullp

Specify whether to include peptides with 0 spectral counts in long peptide output file. **0** is the default.

0

Do not include peptides with 0 spectral counts in output file.

1

Include peptides with 0 spectral counts in output file.

PROTEIN AND PEPTIDE COMPATABLE OPTIONS

-modS Include information about number of modified peptides. Seperate columns for number of spectral counts for modified peptides and total spectral conts will be included in protein and peptide output files.

-mw Calculate protein/peptide molecular weights and molecular formulas. Columns will be included for average mass, monoisotopic mass and molecular formula. Peptide/protein masses and formulas are calculated from *atomCountTable.txt* which contains the number and types of atoms found in each amino acid and a table located at *~/scripts/DTarray_pro/db/atomCountTable.txt* containing the masses of each atom. By default the atom count table at *~/scripts/DTarray_pro/db/atomCountTable.txt* is used. The user can also supply a custom *atomCountTable.txt* file with the **-act** option. Protein sequence information is stored in a fasta formatted file. The default sequence file is *~/scripts/DTarray_pro/db/humanProteome.fasta*.

<fafsa_file>

The user can optionally specify a fafsa formatted file to lookup protein sequences.

-mact, -makeAtomCountTable

Copy default atom count file to working directory and exit program.

-act, -atomCountTable *<file>*

Use user specified atom count table. If the **-mw** option is not also specified, this option will be ignored.

--unicode *<0/1>*

Specify whether to use UTF-8 encoding to write molecular formulas with subscripts in output files. If the **-mw** option is not also specified, this option will be ignored. **0** is the default.

0

Do not write molecular formulas with subscripts.

1

Write molecular formulas with subscripts. Output files must be imported as UTF-8 text to see subscripts in Excel.

-f Include columns for sample name and replicate number in long protein and peptide output files. If the sample name is in the format <sample name>_<number> all text after the last underscore (with the exception of the extension) in the sample name is used as the replicate number.

<prefix>

Remove *<prefix>* from all sample names. If *<prefix>* is not found in sample name, name is unchanged. In long output format, columns will be included for long sample name, short sample name and replicate number.

-rev Choose whether to include reverse matches in protein and peptide output files. **1** is the default.

0

Do not include reverse matches.

1

Include reverse matches.

OTHER

-rw *<arg>*

Rewrite existing param files in working directory.

flist

Rewrite input file list.

smod

Rewrite static modifications file. See **-mw** for details on smod file.

--purge

Remove file list, static modifications file, and all DTarray output files from current working directory and exit program. Only files with default names will be removed.

-v, --version

Print binary version number and exit program.

-h, --help

Display this help file.

EXAMPLES

DTarray

Run DTarray using default parameters.

DTarray -p 1

Run **DTarray**, generating wide formatted protein and peptide (**-p 1**) output files.

DTarray -p 1 -g 2

Run **DTarray**, generating wide formatted protein and peptide (**-p 1**) output files. Group peptides with the same sequence but different charge state onto the same line (**-g 2**).

DTarray -u -s 1

Run **DTarray** and include spectral counts for unique peptides in output file (**-u**), grouping columns by spectral counts then unique peptide spectral counts (**-s 1**).

AUTHOR

DTarray_pro was written by Aaron Maurais. Email questions or bugs to: aaron.maurais@bc.edu