DTarray_pro-1.8.0 User Guide

Aaron Maurais

14 June 2019

Contents

1 Introduction

DTarray_pro extracts Uniprot ID numbers, molecular weights, and spectral counts from DTASelect-filter files. Protein data is combined into one dataset and written to the working directory as a tab delimitated text file (.tsv). This document will describe how to install and use the latest version of DTarray_pro step by step. Some experience using a unix shell is assumed.

2 Installation

2.1 Downloading DTarray_pro source code.

- DTarray_pro expects to be installed in $HOME/local. The program needs files stored in $HOME/local/DTarray_pro/db for some features to work. First, make $HOME/local if it doesn't exist.

```
$ mkdir -p ~/local
```

- If you do not wish to build the program in $HOME/local, you can run the configure script from the desired location with the --program-directory <path> option.

- DTarray_pro is hosted at GitHub, which is a free hosting service for distributed version control in software development. To clone the latest version of the stable branch from the GitHub repo into your current working directory, run:

```
$ git clone --recurse-submodules --single-branch --
   branch stable https://github.com/ajmaurais/
   DTarray_pro
```

2.2 Compiling DTarray_pro

- DTarray_pro is written in vanilla c++ 11 and can be compiled on linux or macOS systems with gcc. Other compilers may work but have not been tested.

- To configure and build DTarray_pro, run:

```
$ cd ~/local/DTarray_pro
$ ./configure
$ make
```

- After you have run `make`, there should be several new files in the `DTarray_pro` directory. If everything worked, two executable files should be located at `DTarray_pro/bin/`

```
$ ls bin
DTarray          DTsetup
```

- `DTarray` is the executable for `DTarray_pro`
- `DTsetup` is the executable for a script which will automatically setup DTASelect-filter to be read into `DTarray_pro`.

3 Usage

3.1 Setting up filter files with DTsetup

`DTsetup` searches all directories one level below where it is run for a file named `DTASelect-filter.txt`. If a matching file is found, it is copied into a folder named `<date>_<time>_dtarraySetup` with the name `<sample_name>.dtafilter`. Once `DTsetup` is run, `DTarray_pro` can be run from the newly created directory to compare spectral counts for proteins found in multiple runs.

To run `DTsetup`, first navigate to a directory containing subdirectories with `DTASelect-filter.txt` files and run the command:

```
$ DTsetup
Creating dtarraySetup folder in parent dirrectory...
Copying DTA filter files to 18-09-28_121625_dtarraySetup
        Adding ./sample_1...
        Adding ./sample_2...
        Adding ./sample_3...
        Adding ./sample_4...
Done
```

3.2 DTarray_pro Input format options

Users have two options for how the input DTASelect-filter files read in by `DTarray_pro` are structured.

3.2.1 std mode

In `std` mode, DTASelect-filter are stored in a common directory and the files are named with the format `<sample_name>.dtafilter` (This is the default input mode for `DTarray_pro`.)

```
$ ls
sample_1.dtafilter                sample_2.dtafilter
sample_3.dtafilter                sample_4.dtafilter
```

The user can either manually setup the folder, or use `DTsetup` to automatically generate the folder. To run `DTarray_pro` in `std` mode, run the commands:

```
$ cd <path_to_directory_with_filterfiles>
$ DTarray

DTarray_pro v1.7

Adding sample_1...
Adding sample_2...
Adding sample_3...
Adding sample_4...

4 files combined.

Writing protein data... done!
Protein data written in wide format to: DTarray_pro.tsv
```

If `DTarray_pro` was successful, two files should have been generated. `DTarray_pro.tsv` is a tab delimitated text file (.tsv) which contains a row for each protein and a column for the spectral counts of that protein in each sample. `dtarray_pro_flist.txt` is a list of valid `.dtafilter` files found in the directory. If a file list already exists in the folder, the existing file list is used. The user can edit the file list to change the order of the columns in the output file and add additional files to the list.

3.2.2 subidr mode

In `subdir` mode, DTASelect-filter files are stored in a separate directory, the directory name is the sample name, and the filter files are named `DTASelect-filter.txt` (This is the default input mode for `dtarray.pl`.)

```
$ find */*.txt
sample_1/DTASelect-filter.txt
sample_2/DTASelect-filter.txt
sample_3/DTASelect-filter.txt
sample_4/DTASelect-filter.txt
```

To run `DTarray_pro` in `subdir` mode, the user has to include the `-i subdir` option, because `subdir` is not the default behavior.

```
$ cd <path_to_parent_directory>
$ DTarray -i subdir
```

Upon completion, a files named `DTarray_pro.tsv` and `dtarray_pro_flist.txt` are generated formatted the same as in `std` mode.

3.3 DTarray_pro Output file options

This document will not provide an exhaustive list of options for `DTarray_pro`. Instead, this section will provide examples of options will likely find most useful. For the full list of optional arguments see `DTarray_pro/helpFile.pdf` or use `DTarray -h` to see the help file from the terminal.

3.3.1 Get spectral counts for unique peptides by protein

the `-u` option is used to include a column for the total counts for unique peptides in the `DTarray_pro.tsv` output file.

```
$ DTarray -u
```

By default, the columns for spectral counts and unique peptide spectral counts are grouped by sample. See section 3.3.2 to change this behavior.

3.3.2 Specify how to group supplementary information columns in output file

The examples in this section all use the `-u` option (section 3.3.1), but the `-s` option is also compatible with other `DTarray_pro` options including the `-lr`

option (see section 3.3.5)

By default, the columns for sample specific supplementary information are grouped by sample as shown in table 1.

	Sample_1		Sample_2	
Protein	SC	Unique_SC	SC	Unique_SC
ALBU	2149	2149	3092	3092
TRFE	661	661	698	698
IGHG1	573	152	382	52

Table 1: Default column arrangement

The `-s` option can be used to control this behavior. To group the columns by sample , then observation (table 2) add the `-s 1` option.

```
$ DTarray -u -s 1
```

	SC		Unique_SC	
Protein	Sample_1	Sample_2	Sample_1	Sample_2
ALBU	2149	3092	2149	3092
TRFE	661	698	661	698
IGHG1	573	382	152	52

Table 2: Grouping columns by sample

3.3.3 Get spectral counts for peptides

By default no peptide file is generated. To also generate a file containing spectral counts for peptides in each sample, use the `-p 1` option.

```
$ DTarray -p 1
```

An additional file will be generated named `peptideList.tsv` containing peptide data.

3.3.4 Modify how peptides are grouped in output file

By default, peptides are grouped by sequence and parent protein. A separate entry for each charge state of a given peptide will be included in peptide output files. If the `-g 2` option is set, peptides will also be grouped by charge; i.e., the spectral counts for each peptide will be the sum of all charge states identified for that peptide.

```
$ DTarray -p 1 -g 2
```

The `-modG <group_method>` specifies how to group modified peptides in `peptideList.tsv`. By default 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. To ignore modification status when grouping peptides, use the `-modG 1` option.

```
$ DTarray -p 1 -modG 1
```

If the `-p 1` option is not set, the `-g` and `-modG` will be ignored. The `-g` and `-modG` options can also be combined as desired.

3.3.5 Get subcelluar location data for proteins

`DTarray_pro` can use `DTarray_pro/db/humanLoc.tsv` to lookup subcelluar localization information for proteins. `humanLoc.tsv` contains Uniprot annotations for subcelluar localization by Uniprot ID, updated as of Apr 19, 2019. Currently, sub cell location information is available for human proteins only.

There are two ways in which subcelluar location information can be compiled.

1. The `-loc` option will add a column for the location of each protein in `DTarray_pro.tsv`. To include the subcelluar location column in `DTarray_pro.tsv` run the command:

   ```
   $ DTarray -loc
   ```

2. The `-lr 1` option will create a file named `loc_summary.tsv` with the sum of spectral counts, sequences, and proteins identified for each subcelluar location. To generate `loc_summary.tsv` run the command:

   ```
   $ DTarray -lr 1
   ```

By default, columns are arranged by sample. The `-s 1` option can be used to arrange the columns by observation. See section 3.3.2 for an explanation of the `-s` option.

3.3.6 Calculate molecular weights for peptides and proteins

`DTarray_pro` will calculate protein/peptide molecular weights and molecular formulas when the `-mw` option is provided. Columns will be included in output files for average mass, monoisotopic mass and molecular formula.

Three files are required for the calculation:

1. An atom count table, named `atomCountTable.txt` which contains the number and types of atoms found in each amino acid (similar to `Cimage` table).

2. An atom mass table, located at `DTarray_pro/db/atomMasses.txt`, containing the masses of each atom.

3. A `.fasta` file to lookup protein sequences located at `DTarray_pro/db/humanProteome.fasta` (required for proteins only) Currently, the `-mw` option is supported for human proteins only.

By default the atom count table located at `DTarray_pro/db/atomCountTable.txt` is used. The default atom count table includes a static modification on cysteine for iodoacetamide alkylation. To calculate peptide and protein masses with default residue masses, run:

```
$ DTarray -p 1 -mw
```

The user can also supply a custom `atomCountTable.txt` file with the `-act <file_name>` option. A copy of the default atom count table can be generated in the working directory with the `-mact` option.

```
# make default atomCountTable.txt in working directory
$ DTarray -mact
# run DTarray with custom atomCountTable.txt
$ DTarray -act ./atomCountTable.txt -p 1 -mw
```

The user can also edit the default atom count table as at `DTarray_pro/db/atomCountTable.txt` as desired, but editing `DTarray_pro/db/atomMasses.txt` is not recommended.