group3\_task4 (1) Peptide statistics group3\_task4(1)

NAME group3\_task4.py - basic statistics for digested peptides using mass-to-charge values

SYNOPSIS group3\_task4.py [--inputfile][--analysis][--minimum] [--maximum][--enzyme] [--sequence][--binss]

DESCRIPTION

group3\_task4.py calculates basic statistics for mass-to-charge, under user defined parameters.

User can define four filters for mass-to-charge values: analysis, minimum, maximum, enzyme and sequence; filters are used in listed order. To visualise for histogram user can define: bins.

Pre-analysis: Task 3 four files "*.masses*" were merged into *task3.masses*. This file is transformed to data-frame and named *task3\_pd*. This file is filtered for values of mass-to-charge between 1000Da - 1500Da to make it comparable with experimental data from spectrometer. Resulting file is *task4dat* and is the file used for further analysis.

The first parameter is -n range or full (default). For -n range, a second filter is required, -i minimum and -x maximum values, if not default values of -i 1000 and -x 1500 are used. These -i and -x values are the conditions cond1 and cond2 respectively which are used to filter *task4dat* and create *data1*. If neither -z nor -s are defined, then *data1* is analysed. If -z is not given but -s is defined, then a condition cond3 for -s is used to create *data11* from *data1*. If -z is given, condition cond4 for -z is used to create *data2* from *data1*. If -z is given but not -s, *data2* is analysed. If -z and -s are given, then *data2* is filtered for condition cond5 for -s and *data21* is created and analysed. For -n range, any combination of -i, -x, -z, -s filters data and a summary of results are in *group3\_task4.tab* file. This file contains for each -b bins: lower limit (bin\_left), upper limit (bin\_right), number of peptides within the bin limits (number\_peps), mean value for mass-to-charge (mean\_mz).

If the first filter -a is not defined (full by default), a complete *task4dat* is used. If neither -z nor -s are given, all *task4dat* is analysed. If -z is not defined but -s is given, a condition cond6 for -s is used to create *data3* from *task4dat*. If -z is given, condition cond7 for -z is used to create *data4* from *task4dat*. When user defines -z but not –s, then *data4* is analysed. Finally having -z and -s, condition cond8 for -s is used to filter *data4* and create *data41*. Mean value of mass, number of peptides, table of filtered data and a histogram are created in *statistics.tab* file. For -n full (default), -i 1000 (default) and -x 1500 (default); any combination of -z, -s filters data and a summary of results are in *group3\_task4.tab* file. This file contains for each -b bins: lower limit (bin\_left), upper limit (bin\_right), number of peptides within the bin limits (number\_peps), mean value for mass-to-charge (mean\_mz).

OPTIONS

-f --inputfile

Input file or files to be processed, it must have .masses format.

-n --analysis

Type of analysis "range" or "full". The first option is for mass-to-charge analysis in shorter range between a 1000Da-1500Da, the second option uses all data between 1000Da - 1500Da.

-i --minimum

If -N is defined as "range" then this is a Boolean value greater than 1000, but not greater than 1500; default value is 1000.

-x --maximum

If -N is defined as "range" then this is a Boolean value lower than 1500, but greater than 1000; default value is 1500. It is preferable a value greater than the minimum.

-z --enzyme

Is an str value deifined by user, it can be t for Trypsine,l for Endoproteinase Lys-C, a for Endoproteinase Arg-C, e for V8 proteinase. Only one letter must be write.

-s --sequence

Is a str sequence of amino - acids user selects within data. Any combination of the following uppercase letters can be searched: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y.

-b --bins

Number of bins for histogram. Default value is 30.

FILES

/task31.masses File from four merged files; it contains peptide name, number of peptide from peptide name, mass-to-charge value in Daltons, z value, p value, enzyme used for digestion, sequence of amino-acids.

/group3\_task4.tab File in tab format for quantitative analysis; it contains a mass-to-charge mean value, number of peptides, a filtered table for parameters given by user (peptide name, number of peptide from peptide name, mass-to-charge value in Daltons, z value, p value, enzyme used for digestion, sequence of amino-acids) and a histogram.

ENVIRONMENT group3\_task4.py It runs in python 3 for user defined -n, -i, -x, -z and -s.

DIAGNOSTICS Diagnostics may be issued on command line:

Input file type.

The input file must have digested\_x.masses format, where x is t, l, a, e; if not it will show error.

Bin number.

It is recommended to use high number of bins, but lower than 2500 because of 0.2Da accuracy of comparing data.

BUGS If none -i or -x are outside limits of 1000-1500 could get error. Also if -i is higher -x, not comparable results can be obtained. For high number of bins, some bins could count zero and mean\_mz could give an ZeroDivisionError.

EXAMPLE

For the parameters below, lines under it are the results in *group3\_task4.tab* file

python3 group3\_task4.py -f task3.masses -n range -i 1200 -x 1300 -s M -b 10

Mass to charge within the range and for defined sequence is: 1249.8503209380174 Da

Total number of peptides within the range and for defined sequence is: 4776

bin\_left bin\_right number\_peps mean\_mz

1200.0 1210.0 476 1205.047

1210.0 1220.0 437 1214.677

1220.0 1230.0 516 1224.767

1230.0 1240.0 478 1235.149

1240.0 1250.0 464 1244.791

1250.0 1260.0 514 1254.793

1260.0 1270.0 473 1264.889

1270.0 1280.0 492 1275.007

1280.0 1290.0 481 1284.95

1290.0 1300.0 445 1295.021

AUTHOR

Moises Gualapuro

SEE ALSO

pandas(1), matplot.hist(2)

Python Last change: NOVEMBER 2018 1