

|  |
| --- |
| **Listing of Analytically Relevant Sequences LARS 1.17** |
| An information and selection tool for improving the selection of sequences for local deuterium uptake analysis using the waters HDX package |

# Table of contents

[Table of contents 1](#_Toc499025506)

[Manual 1](#_Toc499025507)

[File loading 1](#_Toc499025508)

[Output files 1](#_Toc499025509)

[Filtering and scoring 2](#_Toc499025510)

[The sorting specific values: 2](#_Toc499025511)

[Sorting and score values 3](#_Toc499025512)

[Analysis settings 3](#_Toc499025513)

[The minimum sequence score (7-0) 3](#_Toc499025514)

[Maximum allowed RT SD 3](#_Toc499025515)

[Pre-process and process 3](#_Toc499025516)

[Save and load 4](#_Toc499025517)

[Mode of operation 4](#_Toc499025518)

[RT\_User\_SD\_guiding\_determination 4](#_Toc499025519)

[Calculation of sequence coverage 4](#_Toc499025520)

[Calculation of redundancy 5](#_Toc499025521)

# Manual

In order to use this software you must have the 3.6.1 version of python installed. The version can be downloaded here <https://www.python.org/downloads/>. LARS will be developed along with newer versions of python and new releases can be found here https://github.com/LarsSoerensen/LARS

## File loading

The software works exclusively on the \_IA\_final\_peptide.csv file generated from a PLGS MSE data treatment procedure.

The files are added by using the Add an Ion Accounting File



If you want to remove any of the added PLGS files select it with the mouse cursor and use the remove an Ion Accounting File



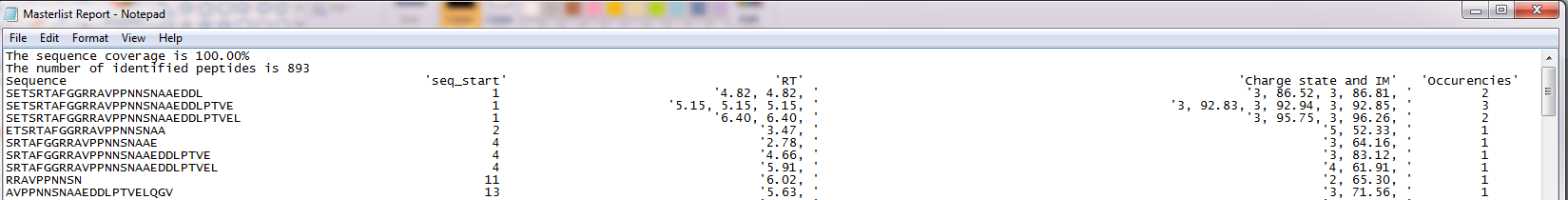
## Output files

LARS produces 3 files. When the user presses process “Run the LARS processing” the user will be asked to specify the location that the files will be written to. The files are PROTEIN\_IDENTIFIER\_IA\_final\_peptide.csv, PROTEIN\_IDENTIFIER\_Helper.txt, PROTEIN\_IDENTIFIER\_Score\_overview.txt

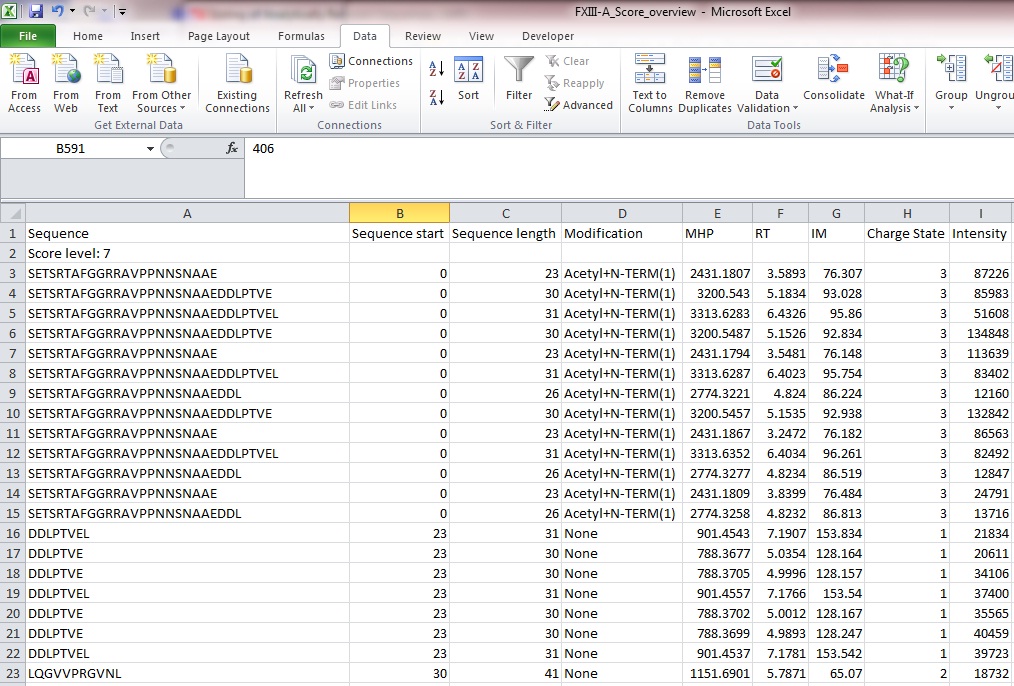
Each file will be named using the user specified protein identifier, see the next section.

PROTEIN\_IDENTIFIER\_IA\_final\_peptide.csv is used to transfer the results of the LARS filtering directly into DynamX.

PROTEIN\_IDENTIFIER\_Helper.txt, contains all the necessary information for validating a DynamX assignment, and provide information regarding the charge state for which a given sequence were identified



PROTEIN\_IDENTIFIER\_Score\_overview.txt



Contains all sequence identifications sorted according to the achieved LARS score and the placement within the sequence, so the user can quickly identify any sequence that fit into a sequence gap and peptides that may improve the sublocalization of the deuterium uptake.

## Filtering and scoring

The filtering is split into two. This is done in order to facilitate a more intuitive graphic user interface (GUI). All Boxes needs to contain an appropriate number or characters or the program will not work. The requirements will be listed below.

### The sorting specific values:

The protein identifier can be found by looking in the FASTA/text files that are used as the database in the PLGS (HD)MSE workflow. The protein identifier is the part of the header marked in red

FXIII-A | P00488 | F13A\_HUMAN Coagulation factor XIII A chain OS=Homo sapiens GN=F13A1 PE=1 SV=4

Only one protein identifier can be used.

In addition to manually typing in the identifier, the user can select one from the list in box below the Ion accounting files.

Maximum MH+ Error (ppm) needs to be a whole number or the software will not work. The value in the entry box will be the highest included.

The maximum and minimum sequence length needs to be a whole number or the software will not work.

### Sorting and score values

All entry boxes are inclusive, meaning that the value in the entry box will be the lowest value included depending on the criteria.

Minimum intensity needs to be a whole number or the software will not work.

Minimum products needs to be a whole number or the software will not work.

Minimum products per amino acids can be a whole number or a floating point number.

Minimum consecutive products needs to be a whole number or the software will not work.

Minimum sum for identified products needs to be a whole number or the software will not work.

Minimum peptide score needs to be a whole number or the software will not work.

Minimum Number of replicates needs to be a whole number or the software will not work.

## Analysis settings

### The minimum sequence score (7-0)

The value in the entry box defines the lower score limit. The limit tells the software when to stop adding new sequences. A preview of the effect of different score cut-offs can be seen by pressing the pre-process button. Values that are not whole number and between 7-0 will break the software.

### Maximum allowed RT SD

When the standard deviation between the RT values exceeds the user determined value, the user is prompted with all the identifications and can then choose the identification that has the RT they believe to be the correct one. A detailed explanation of this function can be found in the Mode of operation, RT\_User\_SD\_guiding\_determination section.

## Pre-process and process

The Update button allows the user to see the results of the analysis without the writing of the different files.

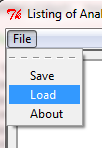
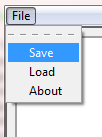


The process button uses the user determined settings and produces the described files



## Save and load

If you want to save the file list and current settings, locate file in the top menu bar and press the save button. You will be prompted to determine the location of the produced \_log.txt file. The \_log.txt file can be loaded using the load the saved file list and settings.



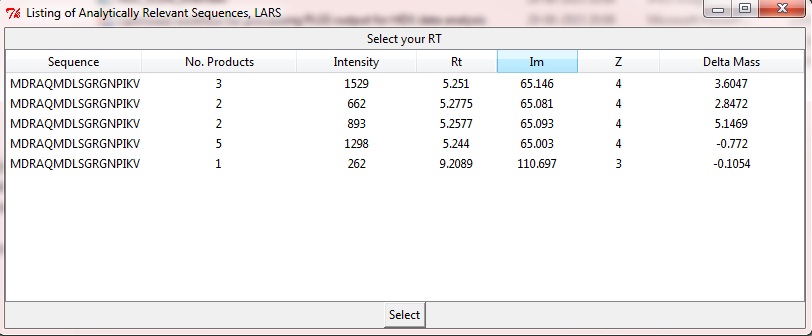
Save

Load

# Mode of operation

### RT\_User\_SD\_guiding\_determination

This is name of the function that prompts the user when an average standard deviation is higher than the user determined maximum, a pop up window will appear. This popup window contains all identifications for a given sequence. The user can then select which identification they believe contains the true RT value. This can be done by marking it with the cursor and pressing Select or double clicking the identification.



### Calculation of sequence coverage

The starting and end point for the sequences forms the basis of the sequence coverage calculation. Lists of all PLGS identified sequence start and end points are created. The lowest numerical sequence start point and the highest numerical end point are used to assess the sequences length. For the individual sequence the start and end point are used to create segments of amino acid positions. Depending on the score cut-off these segments are added to a set list. All redundant amino acid positions are removed by changing the list to a set list.

The limitation of this method is if you don’t have peptides covering the N or C terminal you might end up with a false representation of the sequence length and thereby the wrong sequence coverage will be reported.

### Calculation of redundancy

From the segments calculated in the in the sequence coverage the number of times an amino acid location is identified is added to a list. All these numbers are added together and dived with the number of amino acids found, not the sequence coverage.