

LARS, Listing of Analytically Relevant Sequences

File

2014-06-26_FXIII_B_01
2014-06-26_FXIII_B_02
2014-06-26_FXIII_B_03
2014-06-26_FXIII_B_04
2014-06-26_FXIII_B_05
2014-06-26_FXIII_B_06

Add an Ion Accounting File Remove a Ion Accounting File Update

Pep_A
FXIII-B
FXIII-A

Select Protein Identifier

The minimum sequence score (0-7) 4

Maximum allowed RT SD 0.5

Run the LARS processing

Insert your filter values

Protein Identifier FXIII-B

Maximum MH+ Error (ppm) 10

Minimum Sequence Length 7

Maximum Sequence Length 55

Minimum Intensity 100

Minimum Products 1

Minimum Products Per Amino Acid 0.01

Minimum Consecutive Products 0

Minimum Sum for Identified Products 50

Minimum Peptide Score 0

Minimum Number of Replicates 0

Cut-off value

Minimum scoring value

Sequence Passes

Score cut-off	Sequence coverage	Number of peptides	Redundancy
7	90.6%	146	4.16
6	99.4%	329	9.59
5	100.0%	694	18.39
4	100.0%	1078	28.47
3	100.0%	1445	37.04
2	100.0%	1722	42.77
1	100.0%	1806	44.26
0	100.0%	1806	44.26

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
Manual	1
File loading	1
Output files.....	1
Filtering and scoring.....	2
The sorting specific values:	2
Sorting and score values.....	3
Analysis settings.....	3
The minimum sequence score (7-0)	3
Maximum allowed RT SD.....	3
Pre-process and process	3
Save and load	4
Mode of operation.....	4
RT_User_SD_guiding_determination	4
Calculation of sequence coverage.....	4
Calculation of redundancy	5

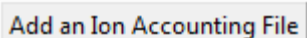
Manual

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>

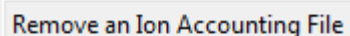
File loading

The software works exclusively on the _IA_final_peptide.csv file generated from a PLGS MS^E data treatment procedure.

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

A rectangular button with a light gray border and a subtle gradient. The text "Add an Ion Accounting File" is centered in a blue, sans-serif font.

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

A rectangular button with a light gray border and a subtle gradient. The text "Remove an Ion Accounting File" is centered in a blue, sans-serif font.

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

Sequence	'seq_start'	'RT'	'charge state and IM'	'occurencies'
SETSRFAFGGRRVPPNNSNAE	1	'4.82, 4.82, '	'3, 86.52, 3, 86.81, '	2
SETSRFAFGGRRVPPNNSNAEEDLPTVE	1	'5.15, 5.15, 5.15, '	'3, 92.83, 3, 92.94, 3, 92.85, '	3
SETSRFAFGGRRVPPNNSNAEEDLPTVEL	1	'6.40, 6.40, '	'3, 95.73, 3, 96.26, '	2
ETSRFAFGGRRVPPNNSNAE	2	'3.47, '	'5, 52.33, '	1
SRFAFGGRRVPPNNSNAE	4	'2.78, '	'3, 64.16, '	1
SRFAFGGRRVPPNNSNAEEDLPTVE	4	'4.66, '	'3, 83.12, '	1
SRFAFGGRRVPPNNSNAEEDLPTVEL	4	'5.91, '	'4, 61.91, '	1
RRVPPNNS	11	'6.02, '	'2, 65.30, '	1
AVPPNNSNAEEDLPTVELQGV	13	'5.63, '	'3, 71.56, '	1

PROTEIN_IDENTIFIER_Score_overview.txt

Sequence	Sequence start	Sequence length	Modification	MHP	RT	IM	Charge State	Intensity
Score level: 7								
SETSRFAFGGRRVPPNNSNAE	0	23	Acetyl+N-TERM(1)	2431.1807	3.5893	76.307	3	87226
SETSRFAFGGRRVPPNNSNAEEDLPTVE	0	30	Acetyl+N-TERM(1)	3200.543	5.1834	93.028	3	85983
SETSRFAFGGRRVPPNNSNAEEDLPTVEL	0	31	Acetyl+N-TERM(1)	3313.6283	6.4326	95.86	3	51608
SETSRFAFGGRRVPPNNSNAEEDLPTVE	0	30	Acetyl+N-TERM(1)	3200.5487	5.1526	92.834	3	134848
SETSRFAFGGRRVPPNNSNAE	0	23	Acetyl+N-TERM(1)	2431.1794	3.5481	76.148	3	113639
SETSRFAFGGRRVPPNNSNAEEDLPTVEL	0	31	Acetyl+N-TERM(1)	3313.6287	6.4023	95.754	3	83402
SETSRFAFGGRRVPPNNSNAEEDDL	0	26	Acetyl+N-TERM(1)	2774.3221	4.824	86.224	3	12160
SETSRFAFGGRRVPPNNSNAEEDLPTVE	0	30	Acetyl+N-TERM(1)	3200.5457	5.1535	92.938	3	132842
SETSRFAFGGRRVPPNNSNAE	0	23	Acetyl+N-TERM(1)	2431.1867	3.2472	76.182	3	86563
SETSRFAFGGRRVPPNNSNAEEDLPTVEL	0	31	Acetyl+N-TERM(1)	3313.6352	6.4034	96.261	3	82492
SETSRFAFGGRRVPPNNSNAEEDDL	0	26	Acetyl+N-TERM(1)	2774.3277	4.8234	86.519	3	12847
SETSRFAFGGRRVPPNNSNAE	0	23	Acetyl+N-TERM(1)	2431.1809	3.8399	76.484	3	24791
SETSRFAFGGRRVPPNNSNAEEDDL	0	26	Acetyl+N-TERM(1)	2774.3258	4.8232	86.813	3	13716
DDLPTVEL	23	31	None	901.4543	7.1907	153.834	1	21834
DDLPTVE	23	30	None	788.3677	5.0354	128.164	1	20611
DDLPTVE	23	30	None	788.3705	4.9996	128.157	1	34106
DDLPTVEL	23	31	None	901.4557	7.1766	153.54	1	37400
DDLPTVE	23	30	None	788.3702	5.0012	128.167	1	35565
DDLPTVE	23	30	None	788.3699	4.9893	128.247	1	40459
DDLPTVEL	23	31	None	901.4537	7.1781	153.542	1	39723
LQGVVPRGVNL	30	41	None	1151.6901	5.7871	65.07	2	18732

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)MS^E 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

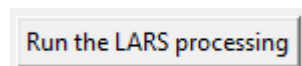
When the standard deviation between the RT values exceeds the user determined value, the user is prompted with all the identifications and can then choose the identification that has the RT they believe to be the correct one. A detailed explanation of this function can be found in the Mode of operation,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



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.

Listing of Analytically Relevant Sequences, LARS

Select your RT						
Sequence	No. Products	Intensity	Rt	Im	Z	Delta Mass
MDRAQMDLSGRGNPIKV	3	1529	5.251	65.146	4	3.6047
MDRAQMDLSGRGNPIKV	2	662	5.2775	65.081	4	2.8472
MDRAQMDLSGRGNPIKV	2	893	5.2577	65.093	4	5.1469
MDRAQMDLSGRGNPIKV	5	1298	5.244	65.003	4	-0.772
MDRAQMDLSGRGNPIKV	1	262	9.2089	110.697	3	-0.1054

Select

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.

$$\text{Sequence coverage} = \frac{\text{Sum of the unique amino acids positions}}{\text{Estimated sequence length}} * 100$$

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.

$$\textit{Redundancy} = \frac{\textit{Number of amino acids found}}{\textit{Unique amino acids positions}}$$