

LandmarkMatch Documentation

Module name:	LandmarkMatch
Description:	Increase the number of identified peptides in an LC-MS experiment.
Author:	Jake Jaffe, D. R. Mani and Vincent Fusaro (Broad Institute) jjaffe@broad.mit.edu
Reference:	Jaffe et al., <i>Mol. Cell. Proteomics.</i> , 5:1927 (2006)

Summary: Landmark matching is a method to propagate identified peptides over time onto accurate mass LC-MS features in such a way as to maximize total identified peptides from disparate data acquisition methods. Using a combination of accurate mass and local retention time information it is possible to determine the likely identification of an unknown peak based on its relative location to known peaks.

1. Peptides sequenced during one experiment are mapped onto features identified by a feature selection program, such as MapQuant, using a loose m/z tolerance (+/- 25 ppm) and an absolute retention time radius of 0.3 min.
2. A more stringent m/z recalibration is calculated using a least squares quadratic fit based on preliminary matches between the peptides sequenced and the features identified by MapQuant. The new tolerance is typically less than 5 ppm.
3. Peptides are mapped onto features from the experiment using the new m/z tolerance. These become the landmarks for the single experiment.
4. Any peptides observed in any related experiment are mapped onto features in the experiment under consideration using a high accuracy m/z data and a relative time heuristic termed the landmark score.

Landmark matching is part of the Platform for Experimental Proteomic Pattern Recognition (PEPPER) pipeline.

The table below summarizes the different options available and which parameters are required.

Parameter	Input
peakList.filename	Required
retentionTime.filename	Required
sampleInfo.filename	Required
globalID.filename	Required
accurateMass.filename	Optional
prefitCoef.filename	Optional
Bootstrap	Required (default = No)

GenePattern

Parameters

Name	Description
peakList.filename	A zip archive containing a set of peak list files, each of which is generated according to the specified format below. Each file name stub must be specified in the sampleInfo.filename. See below for important naming convention information regarding the zip archive structure.
retentionTime.filename	A zip archive containing a list of scan number-to-retention time relations according to the specified format below. Each file name must be specified in the sampleInfo.filename.
sampleInfo.filename	A comma delimited file with a header that specifies the following: experiment ID, sample, class. The peak list and retention time files must be named according to the experiment ID. See the file format below.
globalID.filename	A global list of all the identified peptides. The file format must be specified according to the format below.
accurateMass.filename	Optional: Specify an accurate mass table according to the software used to identify the peptides. The default is from SpectrumMill.
prefitCoef.filename	Optional: Specify different coefficients for calibration.
Bootstrap	Optional: This option defaults to "no" because it involves less computational time. If selected the analysis will take longer and a *.enhanced file will be created containing various match statistics.

File naming convention for files in zip archives:

Each file in the zip archives specified above (as supplied for peakList.filename and retentionTime.filename) must use the following naming convention:

<raw_data_stub>.pgr (for peak list files)
<raw_data_stub>.rts (for retention time files)

Where the <raw_data_stub> portions of the files are the first column entries in the sample information file (see below). There must be one file for each experiment listed in the sample information file in each archive.

Example: given raw data VARMIX_A_01.raw, you would have:

VARMIX_A_01.pgr (peaklist)
VARMIX_A_01.fts (retention times)

as members of the zip archive submitted at run time.

See the associated example data files for further treatment on this topic. The names of zip archives themselves do not follow any convention other than ending in '.zip'

GenePattern

File Formats:

Peak list: The peak list file is a tab delimited generic peak list format. The number of columns must remain the same (you can leave the columns blank if you don't have all the data to fill in the columns). The number of rows is unlimited. The header must be included. If the **mz_err**, **rt_err**, and **carbons** data are unknown, these columns can be zero-filled.

feature	m/z	rt	z	abund	mz_err	rt_err	carbons
1	373.1936	40.82	2	70.8841	0.00501	0.1157	34
2	373.7314	23.61	2	269.1172	0.00443	0.0668	33
4	374.9436	19.98	4	126.9451	0.00673	0.0361	59
5	375.4211	21.21	4	156.4705	0.00733	0.0368	55

Retention time list: The retention time file is a tab delimited list of scan numbers in the first column and time in the second column. There is no header in this file.

1	0.00
2	0.02
4	0.03
5	0.04

Sample information file: this is a comma delimited file that specifies the details of the experiment. The header must be included. The experiment names must match the file names for peak list and retention time. For example, there should be a peak list file called "VARMIX_A_01.pgr" and there should also be a retention time file called "VARMIX_A_01.scan2rt."

LandmarkMatch is designed to work with another module called PeakMatch. This file format supports both modules. It is possible to run LandmarkMatch using only the Experiment column however it would require creating another file in order to use the PeakMatch module.

experiment	sample	class
VARMIX_A_01	A_01	Mix1
VARMIX_A_02	A_02	Mix2

GenePattern

Global ID file: The global ID file is a tab delimited list of the identified peptides. The experiment name must be included in the first column and can be part of the file name as in this example. The global list contains all the identified peptides from all the experiments. There is no header. It follows these conventions:

- Each line has a minimum of 9 columns with no set maximum of columns
- Different lines may have different numbers of columns
- Columns 3 through 7 are not currently used by the landmark matching program but could conceivably be used for filtering by score in a later implementation. For now, you can put anything you want in those columns.

Column 1: spectrum filename (following extract_msn convention, see text)

Column 2: peptide sequence in the form:
[preceding residue].[peptide sequence].[next residue]
Example: K.SDRPELTGAK.V

Column 3: primary score (from your MS/MS search engine)
i.e. XCorr, etc.
-or any arbitrary float-

Column 4: secondary score (from your MS/MS search engine)
i.e. DelCN, etc.
-or any arbitrary float-

Column 5: tertiary score (from your MS/MS search engine)
i.e. SPI, Rank, etc.
-or any arbitrary float-

Column 6: any arbitrary float (some search useful, i.e. ppm error, etc.)

Column 7: any arbitrary float (some search useful, i.e. ppm error, etc.)

Column 8: primary database accession number
i.e. 'gi' number from GenBank database

Columns 9 to (n-1):
additional accession numbers

Last column: number of additional accessions beyond the primary accession
if the last column is '0', then there are only 9 columns for that entry

Example:

VARMIX_A_01.0002.0063.4.pkl	K.KKEEAPSLRPV.A	18.24	90.2	-0.0015	0	0	gi 71826	0
VARMIX_A_01.0010.0135.2.pkl	K.YKELGFQG.-	10.01	72.3	0.0002	0	0	gi 70561	0
VARMIX_A_01.0011.0119.2.pkl	K.YNGVFQEccQAEDK.G	20.41	97.7	-0.0003	0	0	gi 1351907	0
VARMIX_A_02.0179.0238.4.pkl	K.KKEEAPS.A	16.95	86.1	-0.0063	0	0	gi 71826	0
VARMIX_A_02.0219.0330.2.pkl	K.YNGVFQEccQAEDK.G	19.86	96.5	-0.0031	0	0	gi 1351907	0

GenePattern

Accurate mass file: The accurate mass file is a tab delimited table of symbols and accurate mass values. This file may change depending on the program used to perform peptide identification or the instrument mass accuracy. The following is an example (not all amino acids or symbols are listed). There is no header. The default included table will work for SpectrumMill and possibly SEQUEST depending on how your search is defined.

A	71.03711
C	160.0307
c	160.0307
D	115.0269
W	186.0793
Y	163.0633
#	15.99491
@	57.02146
*	43.00581
PROTON	1.007276
WATER	18.01057

Prefit coefficients: The prefit coefficients is a table that initially defines the values for the first pass m/z tolerances. It should be specified according to the following format. The first line contains quadratic coefficients a , b , and c (set to 0 before any fitting is done). The second line contains the r^2 value (set to 1 before any fitting is done). The third line contains the ppm tolerance (set to 25 ppm before recalibration is done).

[0 0 0]
1
25

References:

- Jaffe J.D., Mani D.R., et al. (2006) "PEPPER, a Platform for Experimental Proteomic Pattern Recognition," Mol Cell Proteomics, 5(10) 1927-1941.
- Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., and Church, G. M. (2006) MapQuant: open-source software for large-scale protein quantification. Proteomics 6, 1770.

GenePattern

Return Value:

There are many outputs from this module. Due to the GenePattern format they are all displayed. Only a few files are actually important.

1. LMOutput.zip – is a directory containing all the processed data files.
2. stderr.txt or lsf_log.txt – should be inspected for any obvious errors during processing.

Platform dependencies:

Task type:	Proteomics
CPU type:	any
OS:	Windows, Linux
Java JVM level:	1.4
Language:	Perl