Methodology for MALDI-TOF Data Analysis

Once the peptide mass data of individual protein is obtained from MALDI-TOF, the next step is for identification of protein through database search. The database matching for query sequence depends on lot of parameters, which need to be optimized to come up with the best fit.

Learning Objectives:

After interacting with this learning object, the learner will be able to:

- Prepare in handling the database search tools.
- Operate on setting up the parameters.
- Analyze the result output from the database.
- Assess the troubleshooting steps involved in the experiments.

Note: The current IDD exists in two modes- interactive and automatic. Students taking lab course should select interactive (set as default), while the automatic mode may be selected for general users.
Data Input

In MS analysis, sample firing data that is generated can be saved in the excel format by making a note of m/z of the peaks or by saving the spectrum in ASCII format.
Parameter Settings

Open the matrixscience browser window to carry out online data analysis. In case of peptide analyses, please select peptide mass fingerprint option. More information can be obtained by clicking each of the highlighted buttons.
Parameter Settings

The details for each of the parameters must be entered depending on the background of sample type and its source. Output data for the best hit depends a lot on the input parameter details.
Parameter Settings

Provide details like the user id, which may be required in case of any network problem, so that the output data can be mailed to the user.
Parameter Settings

The following parameters should be selected

Databases: The primary sequence protein databases, including NCBI and SwissProt against which the query will run.

Enzyme: Used during sample preparation before its mass spectrometric analysis.

Missed Cleavage Allowed: Occurrence of partial digests during trypsinolysis of sample protein at one or two Arginine and Lysine sites.
Parameter Settings

In case of Taxonomy: the search query must be limited to a particular species or a group of species for which the sample belongs.
Parameter Settings

For Fixed Modifications: These are modifications that are applied collectively across the database to account for change in mass of specific residue/s.

Variable Modifications: These are mass changes suspected to occur during sample handling and accounted for by increasing the number of primary sequences compared against experimental masses.
Parameter Settings

The parameters can be changed depending upon user needs.

Protein Mass: Is the mass of intact protein in the form of a contiguous stretch including all matched peptides.

Mass Values: To specify the type of charge of the analyte being examined.

Monoisotopic Mass Vs Average Mass Value: Depending upon the mass accuracy of a spectrometer, the experimental masses calculated for identification of analyte by Peptide mass fingerprinting either by choosing monoisotopic mass or the average mass of its isotopic elements.
Parameter Settings

A Query can be uploaded by feeding data or by browsing the file from the database. The result output in the form of the hits can be selected and decoy helps to search with same parameter across database.
Parameter Settings

Do make sure all the parameters are set and mass data is uploaded, user can click the start search button. Depending on the set parameters the search in the database begins.
Data Output

The result output is folded into three sections:

In Section 1: the summary of set parameters defined by user.

Section 2: Mascot Score Histogram. The number of protein hits with score is plotted along the graph.

Section 3: Summary report in which matched proteins from the database, with the details of important parameters are displayed either in concise format, protein format and the data can be exported too.
Data Analysis

If the search parameters are not the best fit, the software generates an error message. Depending on the error message user needs to change the parameter setting and do the search again.
Data Analysis

In Section 2: Mascot Score Histogram, the number of protein hits and their score is displayed along the graph.
Data Analysis

The section: 3 concise protein summary report gives out more details about the protein identification. Matched proteins from the database, with name, mass, score and other details of important parameters are given. For individual protein information, the details can be obtained by clicking on the blue link for each protein.
Data Analysis

In protein view section, displays matching of the query peptide to the protein sequence in the database. The sequence type, the matched region, what is the expected and calculated values of the query peptide and sequence details.
Data Analysis

In protein summary report, index displays the very concise details followed by the details of each of the hits separately. The details contain the ID, Mass, score, expected value and number of matches.
Data Analysis

In export search result report, user has options for data information to be stored with respect to parameters selected. The result can be exported and saved, the data can be later taken for pathway analysis and literature study. In case user has done MS/MS of a particular protein spot, the search parameters are all the same with inclusion of few more parameters for the best fit.
Data Analysis

For better protein identification and to increase protein score, CID of each peak generated is carried out to generate MS/MS data. For such data analysis, MS-fit option is selected from the matrix science browser window. In MS/MS search tool, more input parameters like Quantitation, MS/MS tolerance, peptide charge, instrument etc. in addition to the fields for PMF and rest other parameters are similar to that of Peptide mass fingerprint.
Data Analysis

Depending upon the process carried out for data generation, a selection in the Quantitation must be made. Quantitation of the extracted protein, peptide, mixture of both, extracted from the different instruments before the MS analysis. For example: iTRAQ, Tandem mass Tags for fixed mass/charge values, ICAT, ICPL for precursors within a single data set, SILAC for ion fragments peaks, XICs for precursors in multiple data set etc. For more information on iTRAQ, ICAT, ICPL and SILAC follow the respective IDDs.
Data Analysis

Define the instrument that has been used to generate the raw data.
Data Analysis

All the parameters added must be relevant to the background of sample details to get the best hit from the database search.
Data Analysis

The result output generated is almost similar to mascot output. The Tandem MS protein analysis is used to obtain protein identities from each of the sequenced peptides. The results page begins with a list of probable protein identities and their respective sources. The score histogram provides details similar to the PMF analysis, with the probability distribution being displayed graphically. The green shaded region is indicative of a match that has greater than 5% chance of being random while the red peak indicates that the chances of a random match is less than 5%. 
Data Analysis

The summary report lists all the protein matches obtained from the database search with their respective molecular weight, protein score, source organism and details regarding each of its fragmented peptides. Further information about any of the protein sequences can be obtained by clicking on the corresponding protein link. Data regarding each of the peptide fragmentation patterns can also be obtained by clicking on the peptide link indicated by the query number.
Data Analysis

Regarding the protein score, molecular weight, isoelectric point, the sequence coverage of the protein etc. Protein scores above 67 are considered significant and greater the percentage sequence coverage more is the number of matching peptides for that particular protein. All sequences are displayed with the matching sequences being indicated in red.
Data Analysis

The protein view obtained on selecting a particular protein link, is very similar to the protein view observed in PMF. It provides details Information about each of the matched peptides is also displayed. The start and end amino acid positions, calculated and experimental molecular weights, number of missed tryptic cleavages, sequence of each peptide fragment and their corresponding ion scores are shown. The highest ion scores are used for computing the final protein score.
Data Analysis

Each peptide in Tandem MS/MS undergoes a second round of fragmentation when it passes through the second mass analyzer before it reaches the detector. This provides significantly larger amount of information regarding each peptide fragment which can be viewed by clicking on the peptide links provided in the summary report. The fragmentation pattern is displayed graphically, which can be zoomed into as per the requirement by adjusting the x-axis plot values.
Data Analysis

At low collision energy, each peptide fragment is cleaved at the amide bond which can result in the formation of two types of ions - the y ion & b ion. In y-ions, the positive charge is retained on the C-terminus of the peptide ion while in b-ions; charge is retained on the N-terminal. These ion masses can be used to compute the amino acid sequence by calculating the mass difference between consecutive ions. Each mass difference value corresponds to a particular amino acid, which can be obtained from a standard information table. The y-ion series & the b-ion series run opposite to each other as indicated in the example above. Please go through the future IDDs for more information and better understanding of the concept.
Definitions of the components/Keywords

Methodology for MALDI-TOF Data Analysis

1. Peptide Mass Fingerprinting: Probable protein identification method, which compares peptide mass values of protein analyte to a database of known proteins to arrive at its probable identity in the form of the "best fit".

2. Spectrum from MALDI Analysis: The peptide fragments generated after proteolytic digestion are analyzed by MALDI-TOF and the data generated is represented in the form of spectrum. The spectrum data can be used for online sequence databases search.

3. Online Search: Several open source databases are available online, the protein data is updated on regular basis which allow analysis of the MS spectrum generated.

4. Open Shareware for PMF: Database search algorithms used for comparing experimental peptide masses against theoretically calculated peptide masses derived by applying "cleavage rules" to large primary sequence protein databases. The open shareware consists of the following fields which need to entered by the user during the search:
   a) Name and Email: Used for identification of search entry and also for e-mailing results page in case of loss of connection without requiring re-entry of data.
   b) Search Title: Used to identify and label search entry and typically includes the name of the protein whose information is required.
   c) Databases: The primary sequence protein databases, including NCBI and SwissProt against whom the query is run. A contaminants database is also recommended to eliminate contaminants such as keratin, trypsin and BSA.
   d) Taxonomy: The search query to be limited to a particular species or a group of species.
Definitions of the components/Keywords

Methodology for MALDI-TOF Data Analysis

e) Enzyme: Used during sample preparation of analyte before its mass spectrometric analysis. The popular one is trypsin but if any other enzyme is used it’s site specificity is expected to be equal to or better than that of trypsin.

f) Missed Cleavage Allowed: Occurrence of partial digests during trypsinolysis of analyte protein at one or two Arginine and Lysine sites is a common phenomenon and needs to be accounted for during search against calculated peptide masses.

g) Modifications: During sample prep for Mass Spec Analysis of proteins, some changes in the mass of specific residues might occur, such as oxidation of methionine, carboxymethyl and cysteine etc. To account for these mass changes, the algorithm allows two types of modifications to be pre-selected- Fixed and Variable.

i) Fixed Modifications: Modifications that need to be applied collectively across the database to account for change in mass of specific residue/s. Most common fixed modification is the selection of the mass of carboxymethyl over cysteine replacing its mass as 161 Da.

II) Variable Modifications: These are mass changes suspected to occur during sample handling and accounted for by increasing the number of primary sequences compared against experimental masses. Most common variable modification is the oxidation of methionine residue in the analyte protein.

h) Protein Mass: Mass of intact protein in the form of a contiguous stretch including all matched peptides. If mass is unknown, this parameter can be left empty and the mass will remain unrestricted.

i) Peptide Tolerance: This is a parameter associated with accuracy and resolution of the mass spectrometer and is used to account for shifts in isotope spacings.
Definitions of the components/Keywords

Methodology for MALD-TOF Data Analysis

j) Mass Values: To specify the type of charge of the analyte being examined by Peptide Mass Fingerprinting, i.e. MH+, M-H- or if the masses correspond to neutral values like M0.

k) Monoisotopic Mass vs Average Mass Value: Depending upon the mass accuracy of a spectrometer, the experimental masses calculated for identification of analyte by Peptide mass fingerprinting is either chosen to be monoisotopic mass or the average mass of its isotopic elements. The selection of monoisotopic mass rests upon the ability of the instrument to resolve isotopes, and accurately determine peak mass. Average mass is the sum of abundance-weighted masses of all isotopes while the monoisotopic mass is the sum of masses of the most abundant isotope of each element. If the instrument has insufficient mass resolution capabilities combined with poor signal to noise ratio, the peptide mass of experimental values must be selected as being average to provide better identification.

5. Best fit - Score histogram: The “best fit” is defined as the primary identification of the analyte protein made by the database search algorithm representing either the exact protein being analyzed or the protein with the closest primary sequence homology, unusually with equivalent function in a related species. The score histogram depicts the distribution of protein scores for all the hits obtained by the query.
Which one of these is common across all Mass Spec based proteomics experiments carried out?

- Liquid Chromatography
- Proteolysis
- 2-D Gel Electrophoresis
- Isoelectric Focusing

Congratulations, you have chosen the correct answer.
Peptide Mass Fingerprinting or PMF is defined as?

- Finding the best fit for peptides identified by fragmentation.
- Finding the best fit for protein by sequencing in a Triple Quadrupole Analyzer.
- Finding fingerprints of proteins on 2-DE Gels.
- Finding the best fit for masses of peptides identified by MALDI-TOF.

Congratulations, you have chosen the correct answer.
Which one of these mass values represents a protein/peptide ion?

- M-H-
- M-H+
- MH+
- MH-

I am sorry, the correct answer is MH+. 
The average mass of which of the following amino acids corresponds to 87.0782?

- Serine ✔
- Glycine ❌
- Alanine
- Glutamine

I am sorry, the correct answer is Serine.
The wavelength of laser used for ionization?

- 337nm (Correct)
- 437nm
- 537nm (Incorrect)
- 637nm

I am sorry, the correct answer is 337nm.
References

Papers:


Books:


2) Biochemistry by Stryer et al., 5th edition

3) Biochemistry by A.L. Lehninger et al., 3rd edition

4) Biochemistry by Voet & Voet, 3rd edition