LECTURE-13

Peptide Mass Fingerprinting

HANDOUT

PREAMBLE

Mass spectrometry is an indispensable tool for qualitative and quantitative analysis of proteins, drugs and many biological moieties to elucidate the structure, elemental composition and identification. The basic principle involved in MS is to ionize the molecule and separate them based on mass/charge ratio in electric or magnetic field. Though mass spectrometry has long history, significant advancement has happened during the last twenty years. The initial mass spectrometry has EI and chemical ionization as a source for ionization of the molecule, which were hard ionization methods not suitable for proteomic application. Therefore, soft ionization techniques such as matrix-assisted laser desorption ionization (MALDI) and electron spray ionization (ESI) has been introduced for large biomolecule studies. ESI and MALDI are now routinely used for the analysis of biological macromolecules and become indispensable part of proteomics applications.

OUTLINE

1. Introduction to MALDI TOF-MS
2. Sample preparation for PMF
3. Peptide Mass Fingerprinting (PMF)
4. Data analysis
5. Conclusions

BOX-1: TERMINOLOGY

1. Ion source: One of the major components of any MS instrumentation, which fragments the sample into an ionic form for further detection. MALDI and ESI are most commonly used for proteins samples.

2. Matrix Assisted Laser Desorption Ionization (MALDI): MALDI is an efficient ionization source for generating gas-phase ions of peptides and proteins for mass spectrometric detection. Target analyte embedded in dried matrix-sample is exposed to short, intense pulses from a UV laser.

3. Mass analyzer: The mass analyzer resolves the ions produced by the ionization source on the basis of their mass-to-charge ratios. Various characteristics such as resolving power, accuracy, mass range and speed determine the efficiency of these analyzers. Commonly used mass analyzers include Time of Flight (TOF), Quadrupole (Q) and ion trap.

4. Time-of-Flight (TOF): This is a mass analyzer in which the flight time of the ion from the source to the detector is correlated to the m/z of the ion.

5. Flight tube: Connecting tube between the ion source and detector within which the ions of different size and charge migrate to reach the detector.

6. Reflectron: The reflectron acts as an ion mirror, and extends the flight length without increasing the instrument size. The reflectron compensates for the initial energy spread of ions having the same mass.

7. Detector: The ion detector determines the mass of ions that are resolved by the mass analyzer and generates data, which is then analyzed. The electron multiplier is commonly used detection technique.

1. INTRODUCTION TO MALDI-TOF MS

Mass spectrometry has mainly three constituents, an ion source to generate ions, mass analyzer to separate the generated ions in vacuum and detector to capture the separated ions and generate spectrum according to the m/z values. MALDI is an efficient process for generating gas-phase ions of peptides and proteins for mass spectrometric detection. It is widely used in proteomics research as a high-throughput technique to identify proteins and their post-translational modifications. In case of MALDI, matrix that is mostly aromatic acids absorbs laser energy and transmits the photon to the adjacent peptide fragments to ionize. The ionized fragments travel in vacuum according to the mass-to-charge ratio and give the corresponding spectra.

Illustration: Fundamentals of MALDI-TOF MS

The time-of-flight analyzer resolves ions produced by the ionization source on the basis of their mass-to-charge ratio. The TOF tube can be operated in the linear mode or the reflectron mode depending on the sample to be detected. In case of small molecules, this mode usually provides sufficient resolution. The generated ions are accelerated towards the detector with the lighter ions travelling through the TOF tube faster than the heavier ions. The flight time of the ions is correlated with the m/z ratio. The TOF analyzer can also be operated in the reflectron mode, which is more commonly used for proteomics studies. A reflectron, which acts as an ion mirror, is incorporated at one end of the TOF tube. This helps in extending the path length and in turn the flight time of the ion without having to increase the actual size of the instrument. This helps to even out any kinetic energy differences between ions having the same mass and thereby improves the resolution. The flight time is directly proportional to the square root of mass of ion.

2. SAMPLE PREPARATION FOR PMF

Two dimensional electrophoresis or SDS-PAGE or liquid chromatography has ability to separate hundreds of proteins based on different properties. However, further information such as identification of protein, co- and post modifications, N-terminal sequencing, and quantification cannot be done without MS. PMF or peptide mass fingerprinting (PMF) or peptide mass mapping is required to get more information after separation of the complex proteome. Prior to PMF, proteins are digested with protease to get the peptide fragments and then run on mass spectrometry to generate the spectrum (m/z values), which is the input file for PMF identification. The widely used proteolytic enzymes for mass spectrometry analysis and their role in proteolysis are provided in Fig. 1. Mass spectrometry analysis requires good sample preparation. Sample processing should be very pure and desalting steps needs to be performed to get rid of salts and obtain good spectrum. Once peak list of m/z values are obtained, it can be imported to databases for protein identification.

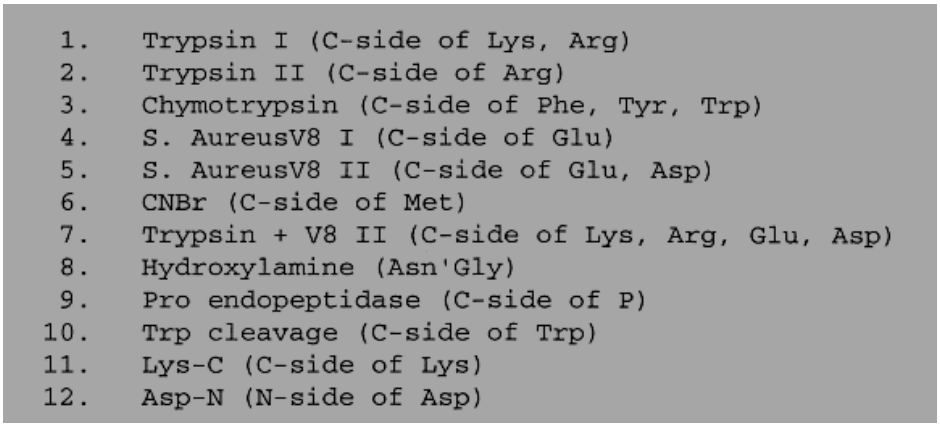


Fig 1. Widely used proteases for mass spectrometry analysis of the proteins (Ref: Thiede B et al., Methods 35, 2005, 237–247)

Illustration: Sample preparation and spotting

The protein sample must be prepared suitably before MS can analyze it. The purified protein of interest is excised from the gel on which it has been electrophoresed and dissolved in a suitable buffer. Trypsin is then added to this in order to carry out digestion of the protein. This enzyme cleaves the protein at the C-terminal of its Arginine & Lysine residues unless there is a Proline present immediately after. The protein is thus digested into smaller fragments of manageable size. Once the protein sample has been digested, all the salt, buffers and any detergents must be removed from this sample. This can be efficiently done with the help of filters (e.g. ZipTip). It offers several advantages such as quick purification, sample enrichment and ensuring there is no contamination. However, it can purify only limited volume of the sample and also adsorbs some amount of the protein sample thereby leading to losses. The purified protein sample is then mixed with an aromatic matrix compound like a-cyano-4-hydroxycinnamic acid, sinapinic acid etc. in the presence of an organic solvent. The components are then mixed thoroughly. The solution containing the organic matrix with embedded analyte is then spotted onto a metallic MALDI sample plate.

Illustration: Ionization and detection

The target plate containing the spotted matrix and analyte is placed in a vacuum chamber with high voltage and short laser pulses are applied. The laser energy gets absorbed by the matrix and is transferred to the analyte molecules, which undergo rapid sublimation resulting in gas phase ions. The gas phase ions generated are accelerated and travel through the flight tube at different rates. The lighter ions move rapidly and reach the detector first while the heavier ions migrate slowly. The ions are resolved and detected on the basis of their m/z ratios and a mass spectrum is generated. Parameters such as geometric design, power supply quality, calibration method, sample morphology, ion beam velocity etc. all affect the accuracy of mass detection.

Illustration: Methodology for the matrix preparation

Matrix is selected depending on the sample used for the MS analysis. The reagent is prepared and matrix is dissolved into the reagent solution. It is necessary to prepare fresh matrix solution for effective MS analysis. MALDI plate provides the surface to load the sample along with matrix. There are circles made on the plate, which define the spotting area for the sample. Used MALDI plate needs to be cleaned before use. Do not use alcohol or sonication processes, these will alter the surface of the plate.

In most of the cases sandwich method is used for sample and matrix deposition. The idea behind this is to make matrix excited even before the laser starts firing the analyte. As soon as the analytes are exposed to the laser they must accept the photon released from the matrix with no much delay. Sample is not in direct contact with laser, so less sample can be used in this method. The dried-droplet method involves mixing the matrix solution and sample, and spotting it on plate. In case if the protein load is more this step can be used. The droplet method involves spotting sample on top of the matrix solution, so that the mixing happens within the drop. To avoid sample loss like in dried-droplet, this method is used. In this method, mixing is uneven and liquid may come out of the circle boundary. Robotic spotting is fast, but in most of the labs manual spotting is carried out. Soon after drying the plate it can be loaded into the MALDI instrument and it is ready for MS analysis.

3. PEPTIDE MASS FINGERPRINT (PMF)

Like our fingerprints, all peptide ions generated in mass spectrometer are unique and hence have a separate fingerprint. PMF refers to the use of exact mass of the peptide in question to derive the identity of a particular protein. In PMF, the peptide sequence is generated, which is used to match the existing sequences in the database. A particular peptide sequence identity will match with its exact homologous sequence with higher confidence.

Illustration: Methodology for MALDI instrumentation

MALDI plate is spotted with the matrix-sample solution and dried before inserting it in MALDI instrument. The plate loading into the system is controlled by software. Before operating the instrument, the plate must be loaded, by giving command over acquisition mode. User must check the instrument status before inserting the plate. If the pressure is high, the vent in chamber gets jam to prevent the door opening. If it is high, wait for few minutes for pressure to come down. Once the door opens, place the plate in the groves provided. The plate will be placed on the surface only in one direction. Once the plate is loaded inside the instrument, the plate parameters need to be specified.

For experimental parameters, in the beginning linear or reflector mode is selected. In reflector mode the sensitivity is better. For data collection the firing parameters need to be set. For some parameters the default works, in some case user can set the parameters depending on the need. First select the spot, which need to be fired. Initially user needs to align the laser with the target spot before setting other parameters. The parameters need to set by user are default. Do fine adjustment for small fine alignment, and to find the sweet region within the spot. If the parameters are optimum, user can save it and reload the same parameters for future use. In CID the peptide of our interest is further broken down to precursor/daughter ions. The PMF data of each peak can later be combined together to perform a MASCOT search.

Initially fire the standard mix spot. The standard peaks help to calibrate the instrument. Calibrate the system by providing least tolerance value. After calibration the system is ready for firing the targets. Once the PMF data is generated, data in the excel format can be exported, and saved. The peptide spectrum can be saved directly.

Prior to the peak detection of the original sample, calibration of the mass spectrometry is required to ensure that there is no mass drift during analysis. Pep mix having peptides in the range of 700-3500 Da can be used for this purpose to make sure that instrument is working fine. The calibration data is shown in Fig 2. Now test sample can be run on MALDI and spectra obtained can be used for the protein identification. The peak-list and spectrum are provided in Fig 3. The peak list can be saved in ASCII files or directly paste in .XLS for analysis as shown in TABLE 1.

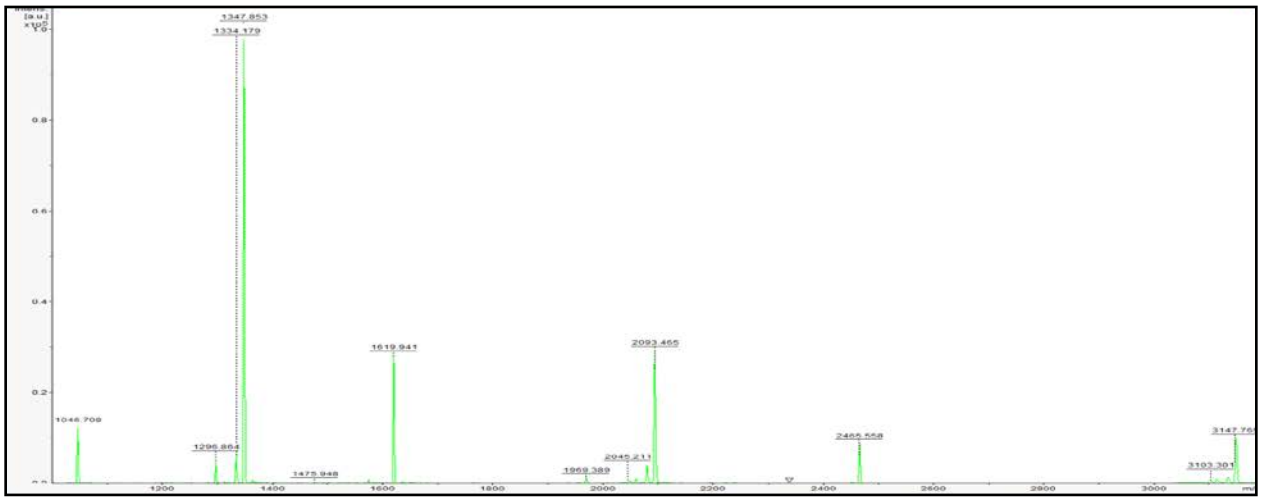


Fig 2. Pep mix standard spectra obtained from the MALDI-TOF. The green peaks indicate the peptide peak of the pep mix. Total seven intense peaks are identified.

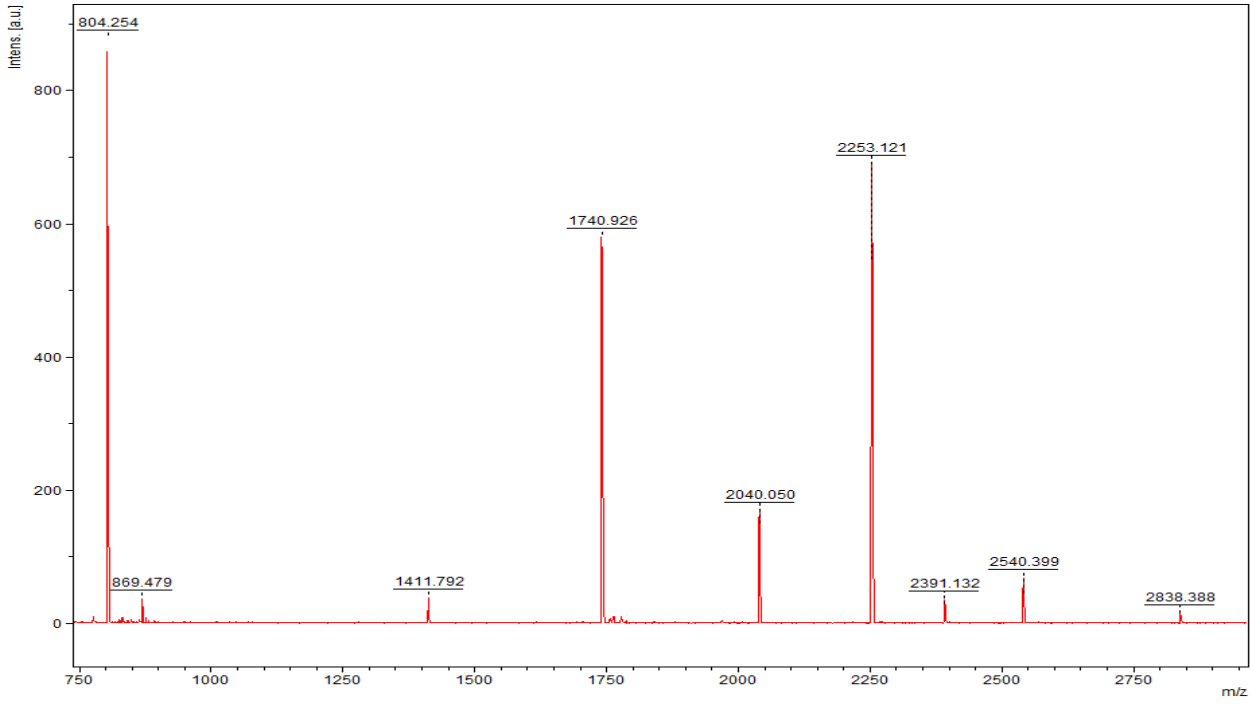
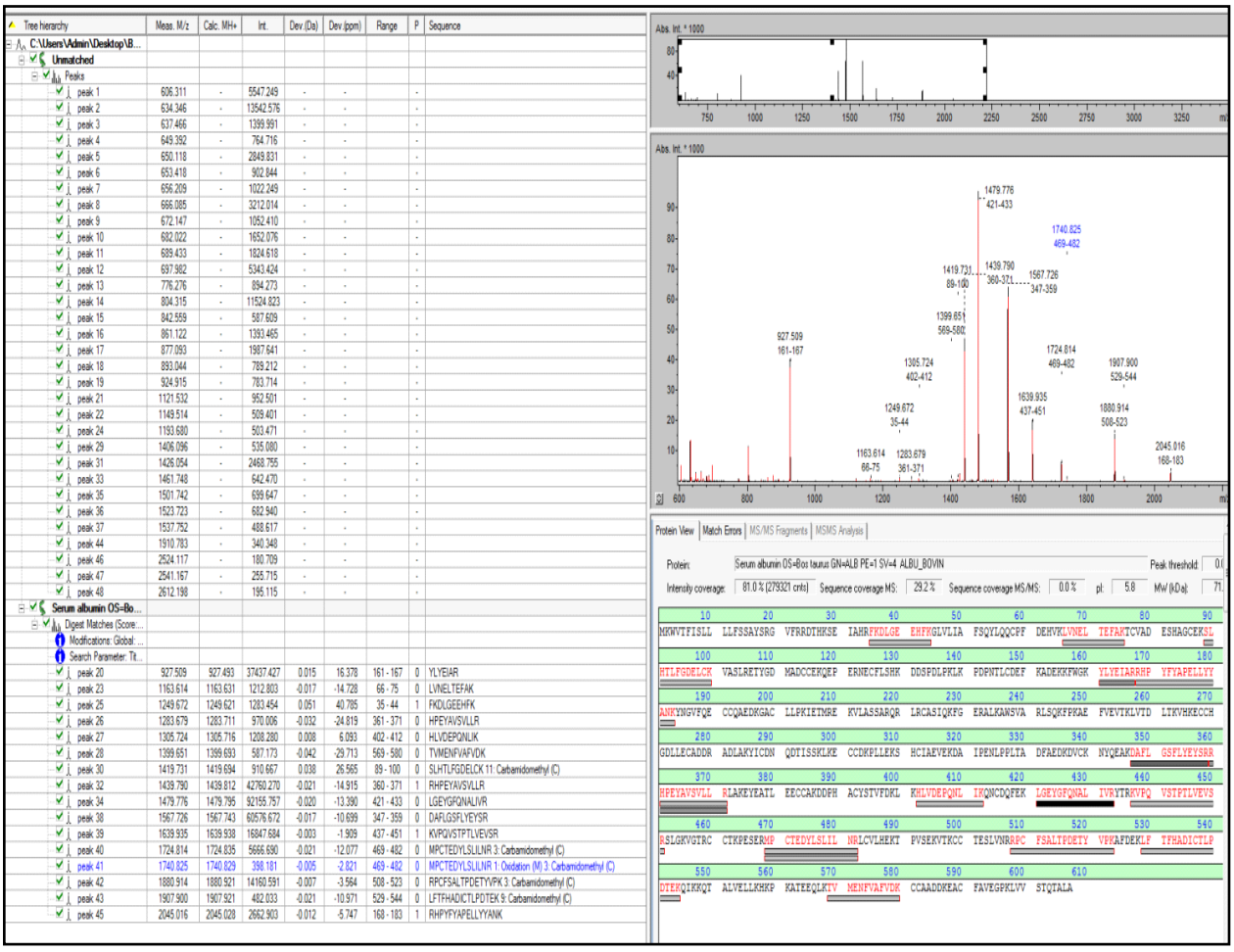


Fig 3. The spectra obtained for the test sample, peaks indicate m/z values for PMF analysis.

Table 1. Peak list generated by the MALDI-TOF. The m/z data is required for the PMF search.

m/z	S/N	Quality Fac.	Res.	Intensity	Area
2838.37	7	401	22401	19.8	10.5
2540.46	15	408	14048	50.4	31.2
2540.27	15	322	31357	51.1	14.2
2391.15	13	562	24689	47.1	14.9
2253.06	184	3691	13781	728	466
2040.19	19	282	39894	91.4	15
2040.01	54	5397	20190	251	93
1740.93	125	909	8335	629	381
1411.79	9	873	9621	47.4	16.4
869.471	10	693	10108	56.2	8.88
804.245	208	52628	7831	1115	203



Tree hierarchy	Meas. M/z	Calc. MH+	Int.	Dev.(Da)	Dev.(ppm)	Range	P	Sequence
C:\Users\Admin\Desktop\B...								
Unmatched								
Peaks								
peak 1	606.311	-	5547.249	-	-		-	
peak 2	634.346	-	13542.576	-	-		-	
peak 3	637.466	-	1399.991	-	-		-	
peak 4	649.392	-	764.716	-	-		-	
peak 5	650.118	-	2849.831	-	-		-	
peak 6	653.418	-	902.844	-	-		-	
peak 7	656.209	-	1022.249	-	-		-	
peak 8	666.085	-	3212.014	-	-		-	
peak 9	672.147	-	1052.410	-	-		-	
peak 10	682.022	-	1652.076	-	-		-	
peak 11	689.433	-	1824.618	-	-		-	
peak 12	697.982	-	5343.424	-	-		-	
peak 13	776.276	-	894.273	-	-		-	
peak 14	804.315	-	11524.823	-	-		-	
peak 15	842.559	-	587.609	-	-		-	
peak 16	861.122	-	1393.465	-	-		-	
peak 17	877.093	-	1987.641	-	-		-	
peak 18	893.044	-	789.212	-	-		-	
peak 19	924.915	-	783.714	-	-		-	
peak 21	1121.532	-	952.501	-	-		-	
peak 22	1149.514	-	509.401	-	-		-	
peak 24	1193.680	-	503.471	-	-		-	
peak 29	1406.096	-	535.080	-	-		-	
peak 31	1426.054	-	2468.755	-	-		-	
peak 33	1461.748	-	642.470	-	-		-	
peak 35	1501.742	-	699.647	-	-		-	
peak 36	1523.723	-	682.940	-	-		-	
peak 37	1537.752	-	488.617	-	-		-	
peak 44	1910.783	-	340.348	-	-		-	
peak 46	2524.117	-	180.709	-	-		-	
peak 47	2541.167	-	255.715	-	-		-	
peak 48	2612.198	-	195.115	-	-		-	
Serum albumin OS=Bo...								
Digest Matches (Score:...								
Modifications: Global: ...								
Search Parameter: Tit...								
peak 20	927.509	927.493	37437.427	0.015	16.378	161 - 167	0	YLYEIAR
peak 23	1163.614	1163.631	1212.803	-0.017	-14.728	66 - 75	0	LVNELTEFAK
peak 25	1249.672	1249.621	1283.454	0.051	40.785	35 - 44	1	FKDLGEEHFK
peak 26	1283.679	1283.711	970.006	-0.032	-24.819	361 - 371	0	HPEYAVSVLLR
peak 27	1305.724	1305.716	1208.280	0.008	6.093	402 - 412	0	HLVDEPQNLIK
peak 28	1399.651	1399.693	587.173	-0.042	-29.713	569 - 580	0	TVMENFVAFVDK
peak 30	1419.731	1419.694	910.667	0.038	26.565	89 - 100	0	SLHTLFGDELCK 11: Carbamidomethyl (C)
peak 32	1439.790	1439.812	42760.270	-0.021	-14.915	360 - 371	1	RHPEYAVSVLLR
peak 34	1479.776	1479.795	92155.757	-0.020	-13.390	421 - 433	0	LGEYGFQNALIVR
peak 38	1567.726	1567.743	60576.672	-0.017	-10.699	347 - 359	0	DAFLGSFLYEYSR
peak 39	1639.935	1639.938	16847.684	-0.003	-1.909	437 - 451	1	KVPQVSTPTLVEVSR
peak 40	1724.814	1724.835	5666.690	-0.021	-12.077	469 - 482	0	MPCTEDYLSLILNR 3: Carbamidomethyl (C)
peak 41	1740.825	1740.829	398.181	-0.005	-2.821	469 - 482	0	MPCTEDYLSLILNR 1: Oxidation (M) 3: Carbamidomethyl (C)
peak 42	1880.914	1880.921	14160.591	-0.007	-3.564	508 - 523	0	RPCFSALTPDETYVPK 3: Carbamidomethyl (C)
peak 43	1907.900	1907.921	482.033	-0.021	-10.971	529 - 544	0	LFTFHADICTLPDTEK 9: Carbamidomethyl (C)
peak 45	2045.016	2045.028	2662.903	-0.012	-5.747	168 - 183	1	RHPYFYAPELLYYANK

4. DATA ANALYSIS

MASCOT is shown here as a representative search engine for PMF data analysis. The PMF data can be imported or peak-list can be manually entered (Fig 4). Before analysis starts, few parameters need to be selected for protein identification. These parameters include taxonomy, type of proteases used for digestion, database need to search, possible missed cleavage by the enzyme, the manual modifications during sample processing, the variable modifications which depends on sample type, peptide tolerance, type of instrument used for data acquisition and type of ions generated etc. During the search, we need to set the parameters otherwise search will take long time and may not provide specific information. Additionally, contaminant peaks such as Keratin, Trypsin etc should be removed.

Once the parameters are set, the search engine will provide the probable hits based on the input m/z values. The preliminary results display accession number, molecular weight, MASCOT score, number of matched peptides, expected p-value, name of the protein and origin of species. The Mascot score is very important for the identification of the protein. The Mascot score is -10*log (P), where P is the probability of the random hit. Therefore, lesser the P value greater the Mascot score. In mascot results, the hits that are indicated outside the green portion with red bar indicate the $P<0.05$ where as the hits in green box shows only 95% accuracy for identification. As red hit goes away from the green box, its precession and accuracy will be enhanced (Fig 5). Further information can be obtained by clicking on accession number. It will show all the details such as type of database used for search, score, nominal mass (Mr), calculate pI, taxonomy, enzyme used for digestion, m/z values searched and

matched, the protein fragments matched will be highlighted in red with sequence coverage value, elaborated view of matched peptide sequence with observed, calculated and expected mass values (Fig 6).

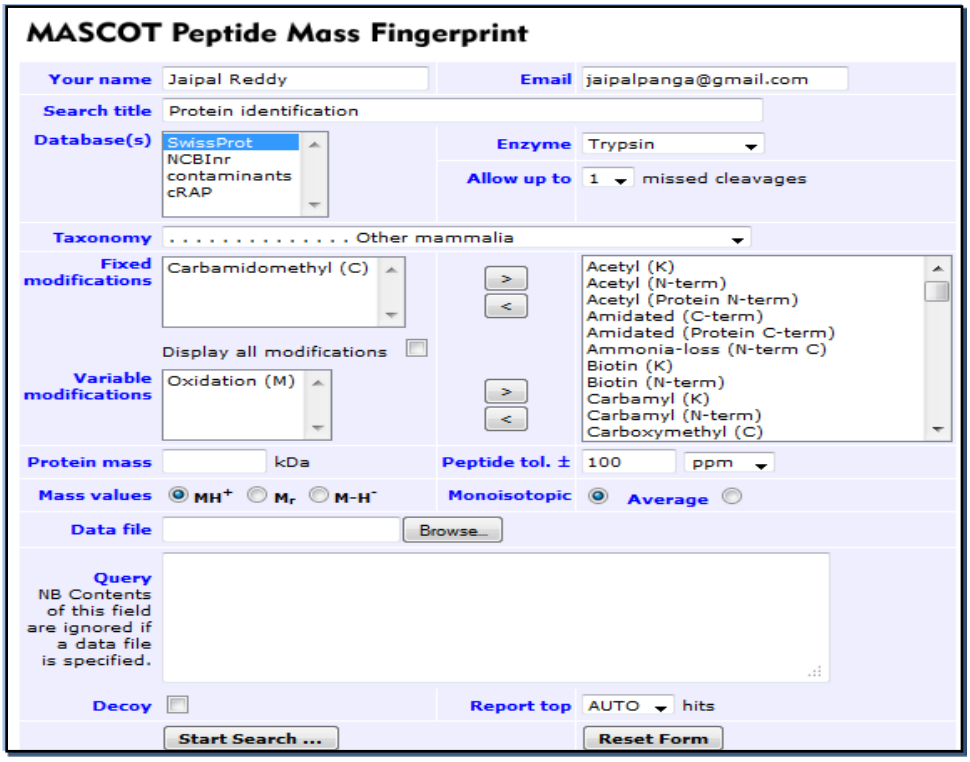
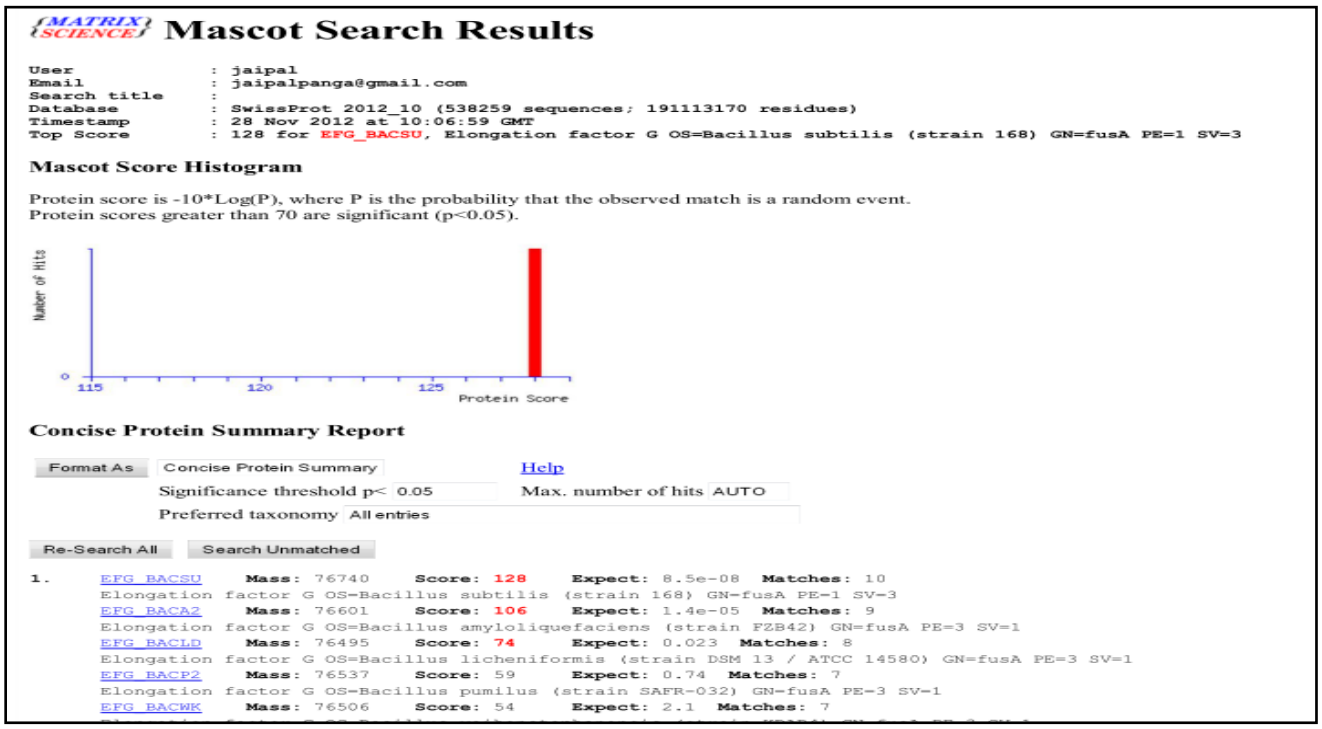


Fig 4. Mascot database for PMF analysis using peak-list.



MATRIX SCIENCE **Mascot Search Results**

User : jaipal
Email : jaipalpanga@gmail.com
Search title :
Database : SwissProt 2012_10 (538259 sequences; 191113170 residues)
Timestamp : 28 Nov 2012 at 10:06:59 GMT
Top Score : 128 for EFG_BACSU, Elongation factor G OS=Bacillus subtilis (strain 168) GN=fusA PE=1 SV=3

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 70 are significant (p<0.05).

Concise Protein Summary Report

Format As | Concise Protein Summary | Help
Significance threshold p< 0.05 | Max. number of hits AUTO
Preferred taxonomy All entries

Re-Search All | Search Unmatched

1. EFG_BACSU Mass: 76740 Score: 128 Expect: 8.5e-08 Matches: 10
Elongation factor G OS=Bacillus subtilis (strain 168) GN=fusA PE=1 SV=3
EFG_BACA2 Mass: 76601 Score: 106 Expect: 1.4e-05 Matches: 9
Elongation factor G OS=Bacillus amyloliquefaciens (strain FZB42) GN=fusA PE=3 SV=1
EFG_BACLD Mass: 76495 Score: 74 Expect: 0.023 Matches: 8
Elongation factor G OS=Bacillus licheniformis (strain DSM 13 / ATCC 14580) GN=fusA PE=3 SV=1
EFG_BACP2 Mass: 76537 Score: 59 Expect: 0.74 Matches: 7
Elongation factor G OS=Bacillus pumilus (strain SAFR-032) GN=fusA PE=3 SV=1
EFG_BACWK Mass: 76506 Score: 54 Expect: 2.1 Matches: 7

Fig 5. Results obtained from Mascot PMF search engine displays score, protein identify, matched peptides and expected probability values.

MASCOT Search Results

Protein View: EFG_BACSU

Elongation factor G OS=Bacillus subtilis (strain 168) GN=fusA PE=1 SV=3

Database:	SwissProt
Score:	128
Expect:	8.5e-08
Nominal mass (M_r):	76740
Calculated pI:	4.80
Taxonomy:	Bacillus subtilis subsp. subtilis str. 168

Sequence similarity is available as an NCBI BLAST search of EFG_BACSU against nr.

Search parameters

Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications:	Carbamidomethyl (C)
Variable modifications:	Oxidation (M)
Mass values searched:	11
Mass values matched:	10

Protein sequence coverage: 23%

Matched peptides shown in ***bold red***.

```
   1 MAREFSLEKT RNIGIMAHID AGKTTTTERI LFYTGRIHKI GETHEGASQM
  51 DWMEQEQERG ITITSAATTA QWKGYRVNII DTPGHVDFTV EVERSLRVLD
 101 GAVAVLDAQS GVEPQTETVW RQATTYGVPR IVFVNKMDKI GADFLYSVGT
 151 LRDRLQANAH AIQLPIGAED NFEGIIDLVE NVAYFYEDDL GTRSDAKEIP
 201 EEYKEQAEEL RNSLIEAVCE LDEELMDKYL EGEEITIDEL KAGIRKGTLN
 251 VEFYPVLVGS AFKNKGVQLV LDAVLDYLPA PTDVAAIKGT RPDTNEEIER
 301 HSSDEEPFSA LAFKVMTDPY VGKLTFFRVY SGTLDSGSYV KNSTKGKRER
 351 VGRILQMHAN SREEISTVYA GDIAAAVGLK DTTTGDTLCD EKDLVILESM
 401 EFPEPVIDVA IEPKSKADQD KMGIALAKLA EEDPTFRTQT NPETGQTIIS
 451 GMGELHLDII VDRMKREFKV EANVGAPQVA YRETFRTGAK VEGKFVRQSG
 501 GRGQFGHVWI EFEPNEEGAG FEFENAIVGG VVPREYIPAV QAGLEDALEN
 551 GVLAGFPLID IKAKLFDGSY HDVDSNEMAF KVAASMALKN AVSKCNPVLL
 601 EPIMKVEVVI PEEYMGDIMG DITSRRGRVE GMEARGNAQV VRAMVPLAEM
 651 FGYATALRSN TQGRGTFTMH MDHYEEVPKS VAEEIIKKNK GE
```

Unformatted sequence string: 692 residues (for pasting into other applications).

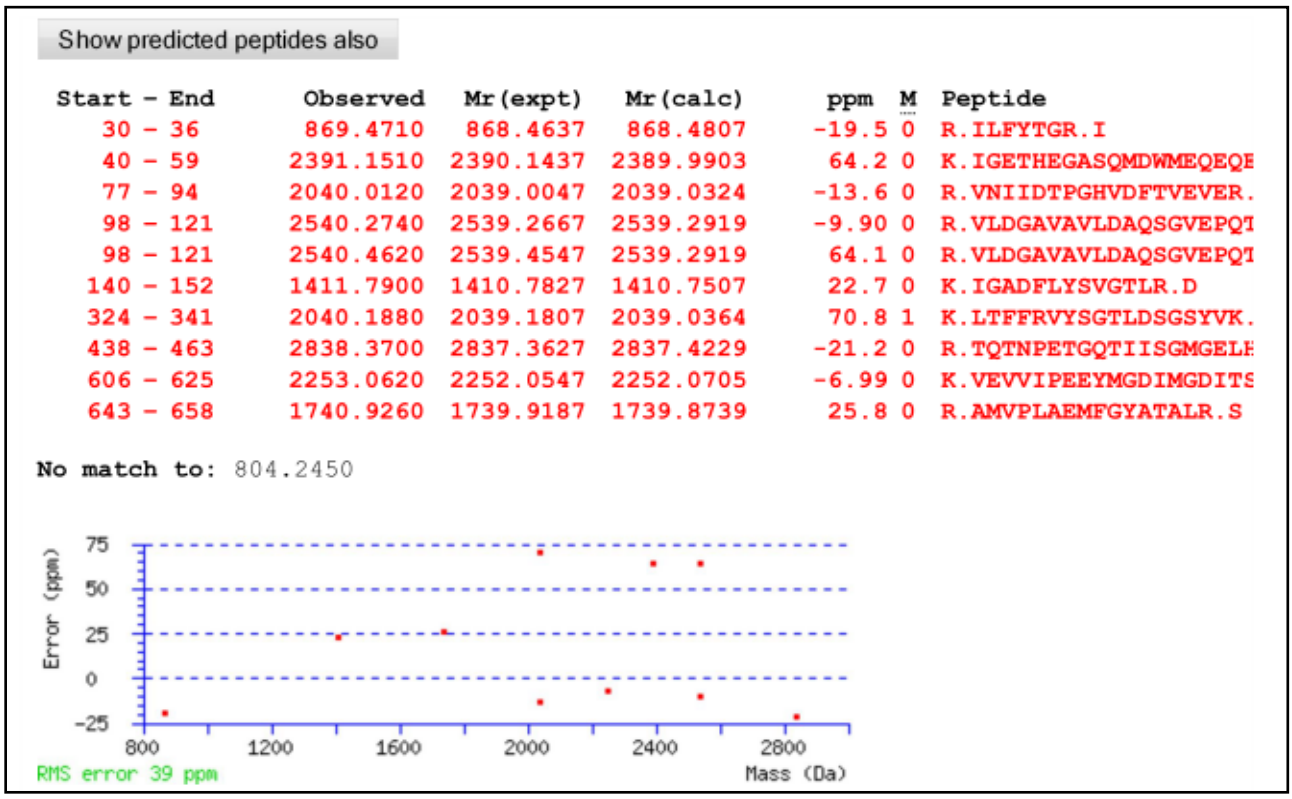


Fig 6. Mascot search results showing details of protein view, including fragments matched with database, sequence coverage, error rate, unmatched peptides and parameters set for protein identifications.

Illustration: Peptide Mass Fingerprinting (PMF) Data Analysis

There are many MS analysis software available online, which allows data generated from MALDI. It requires inputs from the user regarding the experimental parameters used such as enzyme cleavage, protein name, fixed modifications etc. and search criteria like taxonomy, peptide tolerance, taxonomy etc. Commonly used protein databases against which the MS information is processed to retrieve sequence data include NCBI, MSDB and SwissProt. The data file generated from MS is uploaded and the search carried out. The data analysis is demonstrated using Mascot (www.matrixscience.com).

The final results of the search are depicted in a concise report, starting with a protein score histogram. The protein score is a measure of the statistical significance of the protein hit. The histogram seen here displays the distribution of protein scores. Random matches made during database comparison are generally found in the green shaded

region, where the probability of finding a random hit is greater than 5%. The single red peak at the end of the histogram is the protein that has less than 5% chance of being a random hit, making it a statistically significant identity of the unknown protein analyte.

The concise summary report provides details of the peptide matches made by the algorithm which deduces the most probable protein match. The first hit is usually the "best fit" to the experimental masses that were entered in the search query. A protein score higher than threshold is considered to be a significant score. And a lower E-value indicates that the probability of the hit being a random event is extremely low. Significant amount of information about the protein can be obtained from the report by clicking on the corresponding protein link.

On selecting a particular protein link, the protein view provides details regarding the protein score, molecular weight, isoelectric point, the sequence coverage of the protein etc. The greater the percentage sequence coverage, more the number of matching peptides for that particular protein. All sequences are displayed with the matching sequences being indicated in red. Sequence of each peptide fragment processed in database is displayed along with information regarding its molecular weight, starting and ending amino acid number and the number of missed cleavages during tryptic cleavage.

Illustration: Methodology for molecular weight determination by MALDI-TOF MS

For demonstration purpose let us take apomyoglobin sample for the analysis. In MALDI experiment first one need to find sweet spot in spot region, where peaks are more in number with high intensity. After 100 profiles user can save the data. During the firing step user have option for abort, resume, suspend and clear data. User can select these options depending on the profile data obtained. In most cases the default parameters for peak processing are best suited. If user is not fine with data collected, user can make peak processing with the required settings.

Once the PMF data is ready, data in the excel format can be exported, and saved. The mass can be calculated from any two peaks by taking the difference and applying the formulas. The molecular mass of apomyoglobin is around 16951.49u. The same spectrum can be analyzed through MASCOT engine.

5. CONCLUSIONS

Mass spectrometry has revolutionized high-throughput proteomic applications**.** Peptide mass fingerprinting provides tool for protein identifications based on m/z values obtained from the MALDI-TOF MS. In addition to protein identification, PTM analysis and quantitative proteomics etc. are some of the applications of mass spectrometry.