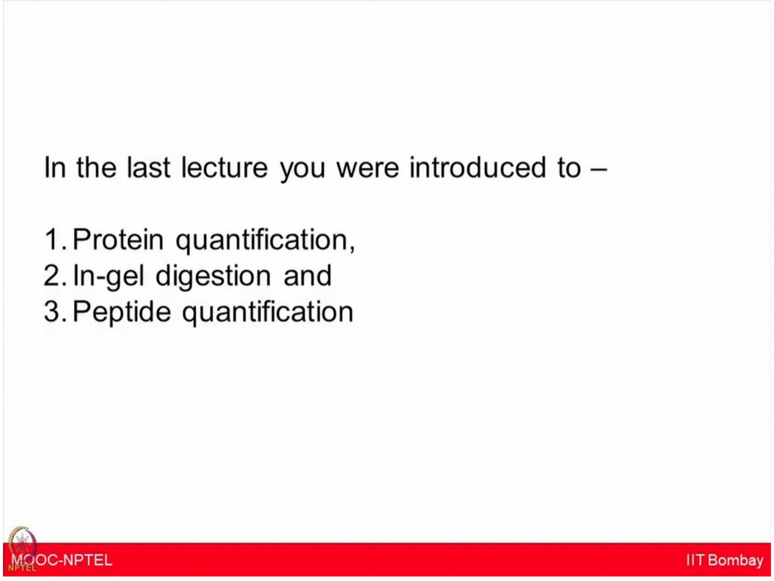
An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Science, Bombay

Supplementary Lecture- 10
Mass spectrometry – Sample preparation and analysis – Part II

(Refer Slide Time 00:21)



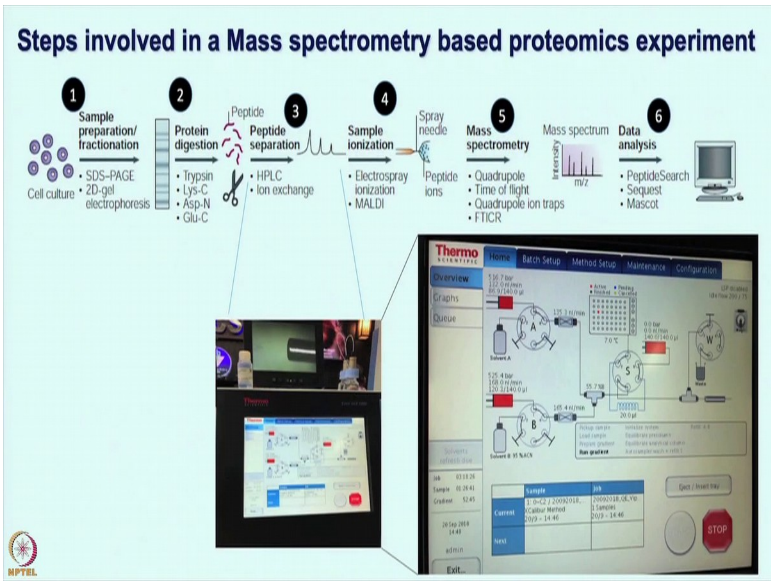
In the last lecture you were introduced to –

1. Protein quantification,
2. In-gel digestion and
3. Peptide quantification

MOOC-NPTEL
IIT Bombay

Welcome to MOOC course on Introduction to Proteogenomics. So, we are proceeding from the protein quantification, running on the SDS page gel, doing the sample clean up, peptide quantification, now you are ready to inject the sample in the LC, liquid chromatography followed by the MS analysis. So, now, it come the liquid chromatography. Here we you are using reverse space column C18 material.

(Refer Slide Time 00:47)



Steps involved in a Mass spectrometry based proteomics experiment

1 Sample preparation/ fractionation
- SDS-PAGE
- 2D-gel electrophoresis

2 Protein digestion
- Trypsin
- Lys-C
- Asp-N
- Glu-C

3 Peptide separation
- HPLC
- Ion exchange

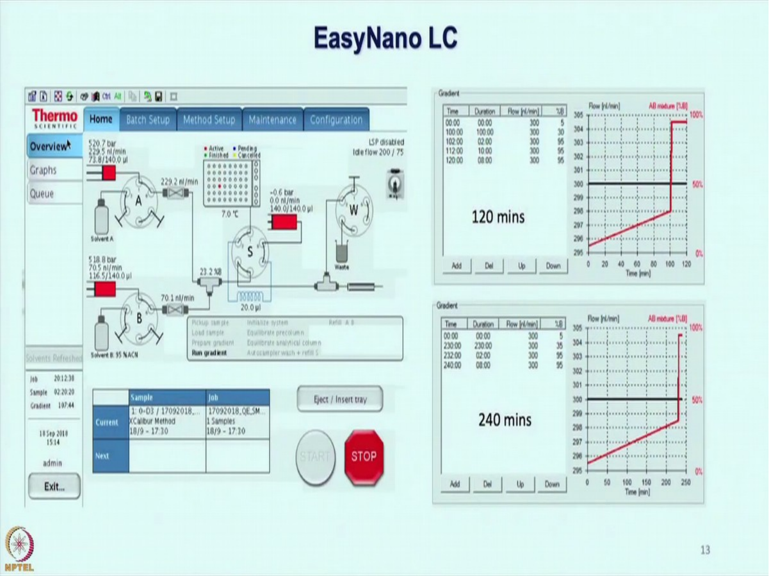
4 Sample ionization
- Electrospray ionization
- MALDI

5 Mass spectrometry
- Quadrupole
- Time of flight
- Quadrupole ion traps
- FTICR

6 Data analysis
- PeptideSearch
- Sequest
- Mascot

Peptides are going to buy into the column and then you want elute the peptide based on the hydrophobic hydrophilic properties, so you will use a different gradient using acetonitrile, 5 percent to 80 percent or you can go even up to 90 percent with 0.1 percent formic acid. And, thes concepts I have already you know briefed in the beginning in the previous lectures, so I am not repeating again. But you need to pay attention to the parameters for what should be the best gradient for doing a liquid chromatography.

(Refer Slide Time 01:21)



EasyNano LC

Gradient

Time	Duration	Flow [nl/min]	%B
00:00	00:00	300	5
100:00	100:00	300	30
102:00	02:00	300	95
112:00	10:00	300	95
120:00	08:00	300	95

120 mins

Gradient

Time	Duration	Flow [nl/min]	%B
00:00	00:00	300	5
230:00	230:00	300	35
232:00	02:00	300	95
240:00	08:00	300	95

240 mins

What is shown on the screen here is the again a refresher we talked earlier as well, that you will use different parameter depending on the kind of sample you have. And, you would like to achieve a good Gaussian distribution of the peptides which are eluted from the column. You would like to see that you know very soon may be after 5 to 10 minutes time peptides start eluting out of the column, then after as you increase the gradient of acetonitrile more and more peptides are coming out of the column.

And, eventually when you have reached to the saturation level, then finally, all the peptides are out of the column and you are then washing and requilibrating it being ready for the next injection. So, that ideal set up should give you a Gaussian distribution of good intensity of the all the peptides. One need to play with this parameters, but again shown here that you need to work on each of these, and now, let us we will talk more about these parameters in the lab session.

Hello. So, here I am going to explain about mass spectrometry. So, basically there are main two component, one is liquid chromatography, another is mass spectrometry. So, I am going to explain first about liquid chromatography.

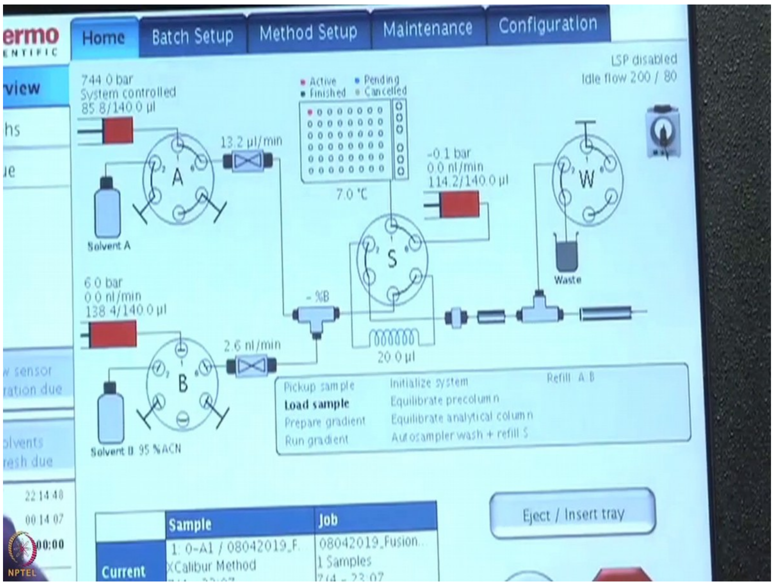
(Refer Slide Time 02:21)

So, here you can see there are main two solvent.

(Refer Slide Time 02:33)

This is one solvent A which having of 0.1 percent formic acid. Another is solvent B, which is 80 percent ACN in 0.1 percent formic acid. Now, here if you see this screen; now, here you can see these are pump A and this is pump B, basically which regulate the flow of solvent A and solvent B.

(Refer Slide Time 02:53)



Pickup sample
Load sample
Prepare gradient
Run gradient

Initialize system
Equilibrate precolumn
Equilibrate analytical column
Autosampler wash + refill S

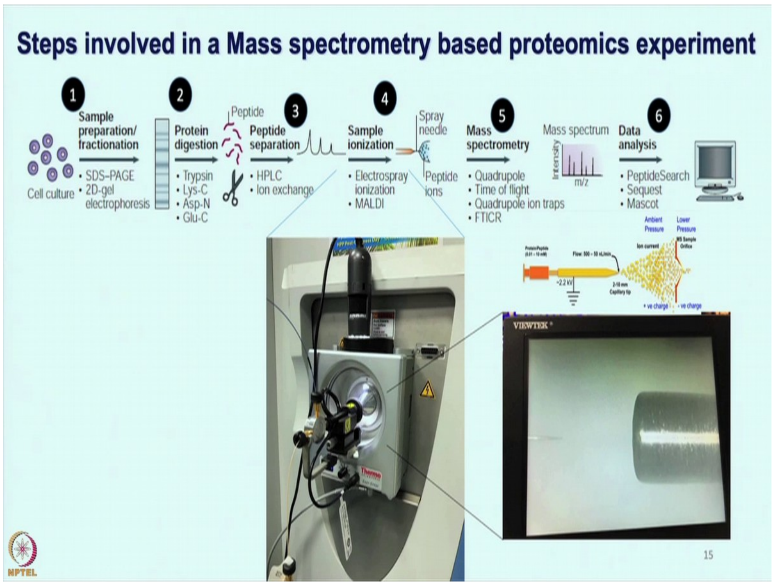
Refill A B

	Sample	Job
Current	1: 0-A1 / 08042019_F. XCalibur Method	08042019_Fusion.. 1 Samples 7/4 - 23:07

This is another pump which is pump S it control the taking how to sample that how much amount of the sample has to be injected that will be controlled by pump S. So, here I am ejecting the tray, where you can keep your sample. Now, here we can see this plate, and this is vial where you can keep your sample. So, generally here we are loading around 1 microgram, accordingly you can calculate the amount of volume how much you have to inject in mass spectrometry.

Now, you are familiar with the liquid chromatography and how nano LC can be used, how different parameters, different pumps are important, keeping watch on the pressure is very crucial. Now, you are ready for injection with the electro spray ionization.

(Refer Slide Time 03:43)



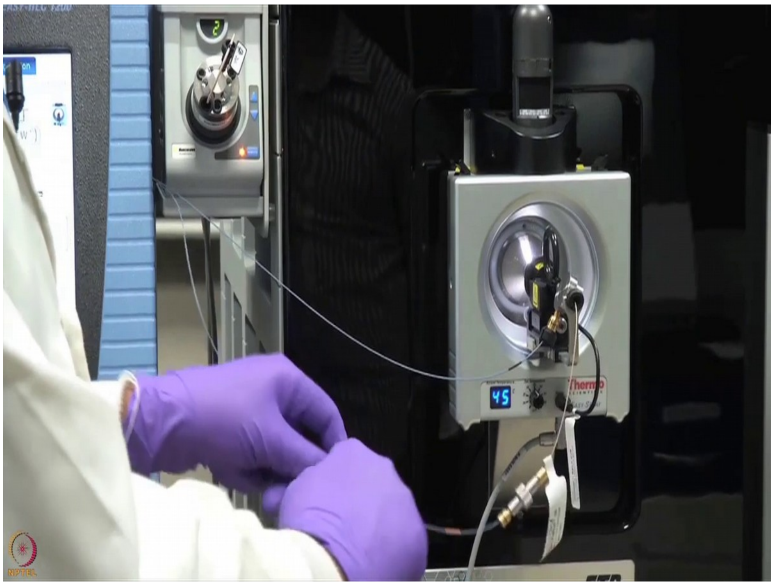
Steps involved in a Mass spectrometry based proteomics experiment

This is the very crucial part because now all your peptides have converted to the gaseous ionized forms and they have to move it inside the mass analyser for further analysis. So, various settings are and again the voltage these criteria are important and you are going to learn more about them in the lab session. But briefly refreshing you here again that while all these ions are comings your major effort is most of them can you move them inside the mass analyser, and that is where the pressure and the charge, these parameters are very crucial and you need to make sure that most of the ions are actually going inside the mass analyser otherwise the proteome coverage will be effected.

So, next let us assume that you know you have done a good electro spray ionization. Then you are ready to separate these ions inside the mass analyser. There could be different type of mass analysers and different type of mass spec configuration. For example, we can have you know ESI Q-ToF a popular configuration or orbitrap. In this case we are going to talk about orbitrap fusion technology which is having a tribrid, mass spectrometer and let us have a laboratory session about using orbitrap fusion technology and different parameters for MS and MS-MS analysis.

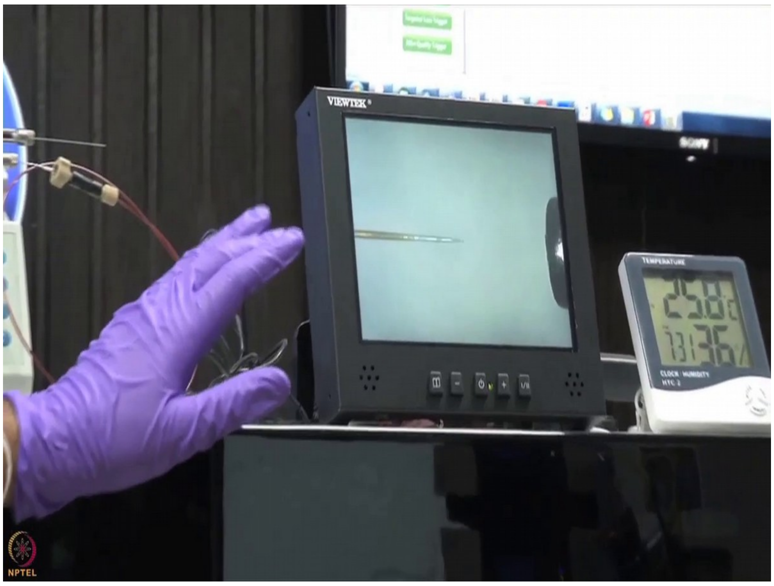
Once you have kept your sample in nano LC, and then you will monitor through the software.

(Refer Slide Time 05:07)



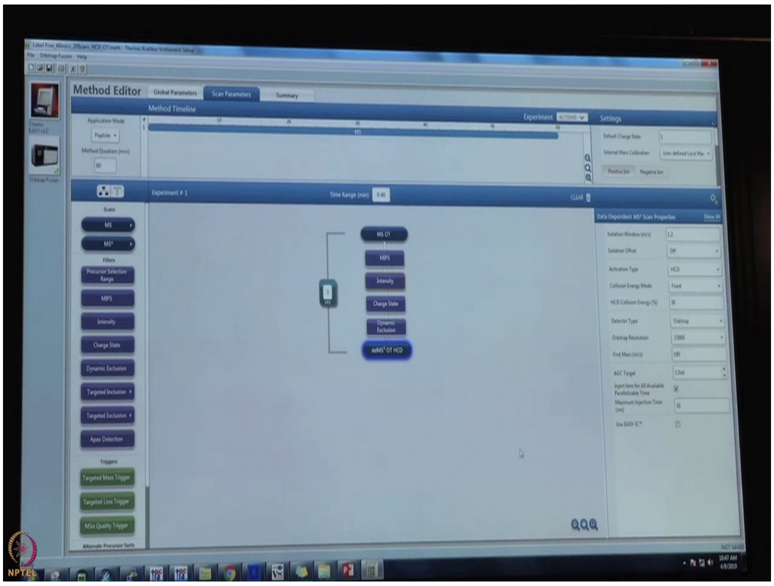
So, your sample will go through this tube and this is first column which is pre-column, where your sample start to clean up, basically desalting. And your peptide will start to bind in this column and then waste will go through this tube into a waste beaker . Now, once your sample will be cleaned in the pre-column then second column which is analytical column, then your sample will start pass through this analytical column and they will start to fractionate, and slowly the first hydrophilic peptide will start elute and then hydrophobic peptide will start to come. So, once your sample is start to elute from this tip of column you can see here, and then it will start get ionized.

(Refer Slide Time 05:45)



Now, here we are applying voltage around 2.2 kV and your sample like high highly charged and they will start to elute.

(Refer Slide Time 05:55)



Settings

Default Charge State: 1

Internal Mass Calibration: User-defined Lock Ma

Positive Ion | Negative Ion

Data Dependent MS² Scan Properties

Isolation Window (m/z): 1.2

Isolation Offset: Off

Activation Type: HCD

Collision Energy Mode: Fixed

HCD Collision Energy (%): 30

Detector Type: Orbitrap

Orbitrap Resolution: 15000

First Mass (m/z): 100

AGC Target: 1.0e4

Inject Ions for All Available Parallelizable Time: ☑

Maximum Injection Time (ms): 30

Use EASY-IC™: □

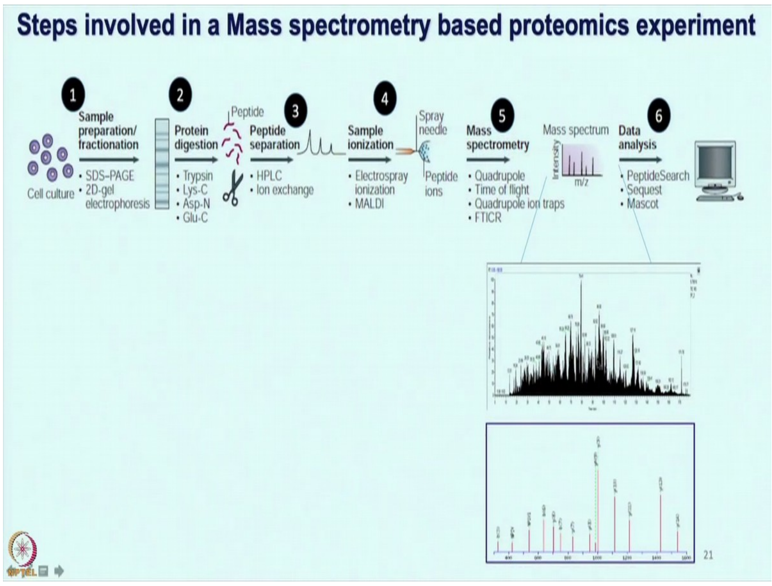
Now, here I am going to explain parameter for MS. So, this is MS OT. So, these are parameter for the MS, where I keeping orbitrap resolution around 60,000, scan range 375 to 1700 M/Z. So, generally for peptide it is in a that is optimised. RF lens I am keeping 60. AGC target is like automatic gain control, like how many ions you want to accumulate, so that you have to define, which I am taking here around 4.0 e to power 5. Maximum injection time is 50 millisecond and polarity is positive, because our peptides ionise and they are positively charged.

In intensity you have to define the intensity how much intensity threshold should be there which is 5.0 e to the power 3. Now, generally peptide is charged. So, I have to define how much charge should be there the range of charges, ok. Now, here you can see 2 to 6 I am keeping, it will not take singly charged and then more than 6, it will consider only 2 to 6 charged peptide. And, then I am defining dynamic exclusion which is 40 second here. Mass tolerance should be in ppm and this is the parameter for the MS-MS now, where I am keeping isolation window is 1.2.

Now, because the MS is happening on the peptide, now those peptide have to be fragmented and for that I am using the HCD, cell high collision dissociation and energy collision energy more said when we fixed and this is the energy which I am applying for the fragmentation for our peptide. And, a detector is here isorbitrap and a resolution for the MS-MS is 15,000, and first mass which has to be detected is 100. So, these are the like parameter, which I have shown you these are the like already optimised for the label free quantification. If you are using like different-different type of technique, if you are using a labelled base in case of iTRAQ and TMT accordingly those parameter will be changed. So, accordingly you can do it change so, this was the parameter for the LC and then MS parameter.

So now, you are familiar that how you can use different parameters are settings for the MS and MS-MS analysis. After doing the a good run from the experiment from the same sample, now you will see these chromatograms which is shown here on the screen which shows that you know the time versus the intensity of these peptides, as I mentioned you would like to see a good Gaussian distribution of the peptides coming out of you know from your sample.

(Refer Slide Time 08:13)



Steps involved in a Mass spectrometry based proteomics experiment

1 Sample preparation/ fractionation
- SDS-PAGE
- 2D-gel electrophoresis

2 Protein digestion
- Trypsin
- Lys-C
- Asp-N
- Glu-C

3 Peptide separation
- HPLC
- Ion exchange

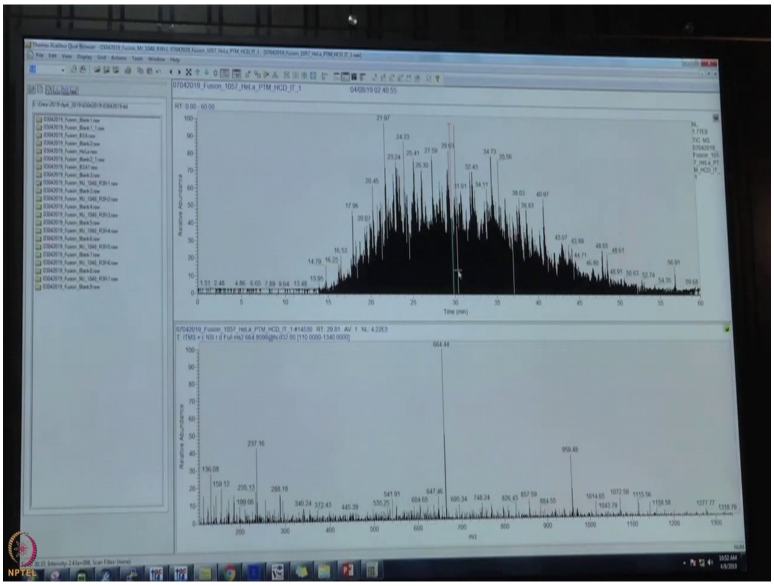
4 Sample ionization
- Electrospray ionization
- MALDI

5 Mass spectrometry
- Quadrupole
- Time of flight
- Quadrupole ion traps
- FTICR

6 Data analysis
- PeptideSearch
- Sequest
- Mascot

Now, from this sample how to make cells of this information that you know what these proteins are. So, if you remember that we talked briefly about you know looking at b and y ions. So, again if you keep looking at walking through the entire chromatograms, you will see the pattern of the spectra which will give you good idea for you know the b ion, y ion various ions generated, and now, you can use this information for the database search. So, let us have another lab session, to discuss more about these chromatograms and looking at these data.

(Refer Slide Time 08:59)

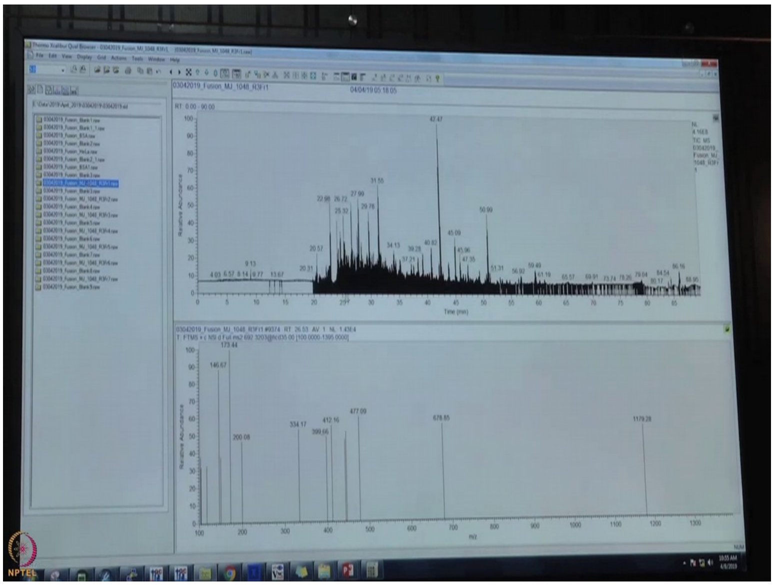


So, till now we have seen the different liquid chromatography parameters and the mass spectrometry parameters required for the success of a LC-MS experiment. So, as you now know that the charge species enter into the mass spectrometer and get fragmented. These fragments are then detected and by use of suitable software, the identification of the peptide is revealed. However, just by looking at the chromatogram one can easily deduce whether the sample run was good enough for it to be taken to further analysis.

Let us now take a look at the example of a very good chromatogram. On the screen, we see a very Gaussian distribution of the peptides, this is the MS 1 chromatogram; that means, all the peptides which have entered into the mass spectrometer as charge species have got detected at the MS 1 level. Further, based on the abundance of each of these peptides they are fragmented at the MS 2 level and detected. It is to be noted that most of the peptides which are less abundant are likely to be ignored and only the high abundant peptides get fragmented.

So, the bottom panel shows the MS 2 fragmentation pattern for the selected peak. So, if you can see here, this is the MS 2 pattern for the selected peak. So, you see different fragments which have been generated and also the signal is relatively less noisy. This helps in better interpretation of the data for the software. So, we now move to another chromatogram which is not that great.

(Refer Slide Time 11:11)

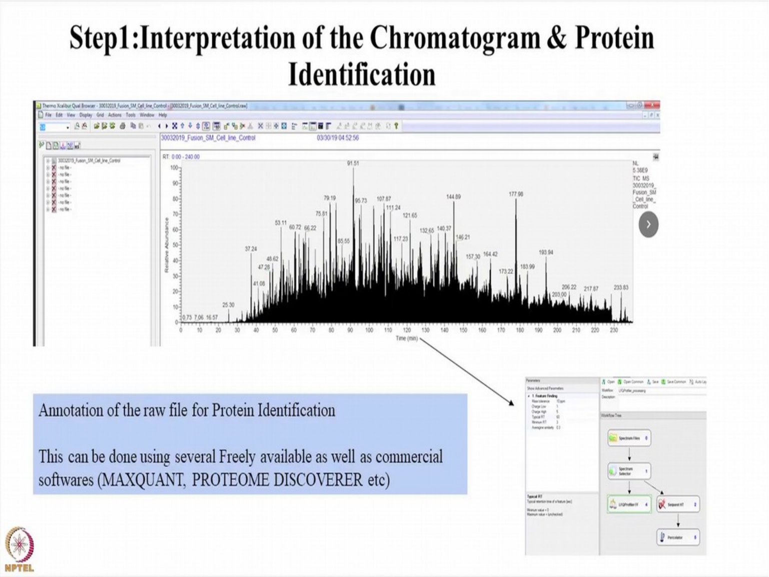


So, on the screen if you try to correlate, if you try to compare this chromatogram to the chromatogram which was previously shown, you can very clearly see that the distribution of the peptides is not a Gaussian distribution. Also, you see that there are gaps in the chromatogram which are indicative of issues with the sample and with the electrospray ionization. So, there are certain segments of the chromatogram where probably nothing entered into the mass spectrometer or the peptides did not get ionized properly for them to be detected by the mass spectrometer. If you now look at the MS 2 pattern, the MS 2 here has significantly less number of fragments which is a feature of a very bad chromatogram resulting from a very bad sample.

This issue could have been due to improper handling of the sample or due to issues with the column, but this is the basic information that one can deduce from merely looking at the chromatogram at the MS 1 level and the MS 2 level. The raw data that is generated is then further analyzed using specific softwares which can deduce the information in the chromatogram and reveal the identity of the peptides subsequently leading to the identification of the proteins.

I hope you got a very good glimpse of doing the mass based proteomics work flow. But how one could use the same set of information these you know work flow for a you know any case study, for any biological problem. In this slide I have invited one of my PhD student Shuvolina Mukherjee to talk about how she has used this mass spectrometry based proteomics work flow in her own research. Very briefly she will walk you through these steps and these strategies for data analysis and in a nut shell that how it can give you some biologically meaningful insight from a clinical problem.

(Refer Slide Time 13:25)



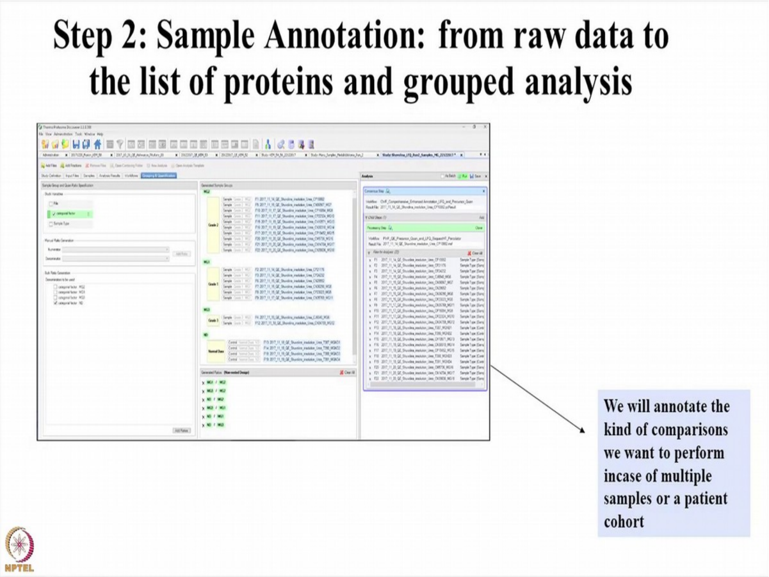
Step1:Interpretation of the Chromatogram & Protein Identification

Annotation of the raw file for Protein Identification

This can be done using several Freely available as well as commercial softwares (MAXQUANT, PROTEOME DISCOVERER etc)

So, now that you have an idea about the chromatogram that comes out after your mass spectrometry experiment, we will talk about how to interpret it and how to do the protein identification. So, you can interpret the raw data coming out using several freely available software as well as commercial software. So, one of the software is MAXQUANT and another software that I will be giving you a little glimpse of is PROTEOME DISCOVERER. So, what the software does it? It takes the raw data into account and then it does the spectral matching and counting and also it uses a database, that is a background database. So, if you sample is from human origin, you give a Homo sapiens database and then it does the annotation of the peaks.

(Refer Slide Time 14:11)



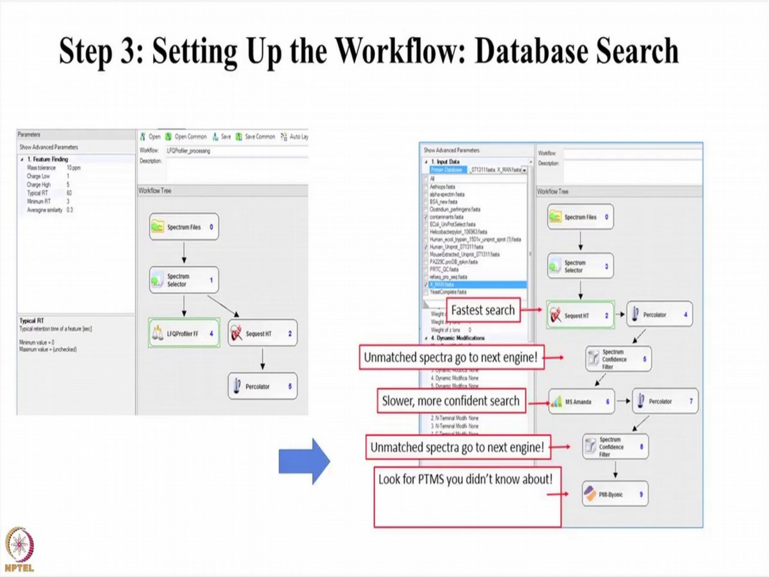
Step 2: Sample Annotation: from raw data to the list of proteins and grouped analysis

We will annotate the kind of comparisons we want to perform incase of multiple samples or a patient cohort

Also, in the process of sample annotation from raw data you can also do a grouped analysis. For example, in this case if you are looking at a cancer sample, wherein you are comparing the normal or the non-tumour samples with the tumour samples, you can give different grades of the tumour. For example, this is grade I, this is grade II and this is grade III of a tumour and these are the normals.

So, when you annotate the raw files in such a way that you already specify that which group it belongs to, the software then takes into account these considerations and then the it gives you the details of how much of the protein abundance is present across these groups; so, that you can know whether you have some dysregulated proteins that can be subsequently used for identification of bio markers.

(Refer Slide Time 14:59)



Step 3: Setting Up the Workflow: Database Search

Parameters
Show Advanced Parameters
1. Feature Finding
Mass tolerance 10 ppm
Charge Low 1
Charge High 5
Typical RT 60
Minimum RT 3
Averagne similarity 0.3

Typical RT
Typical retention time of a feature [sec]
Minimum value = 0
Maximum value = (unchecked)

Workflow: LFQProfiler_processing

Fastest search

Unmatched spectra go to next engine!

Slower, more confident search

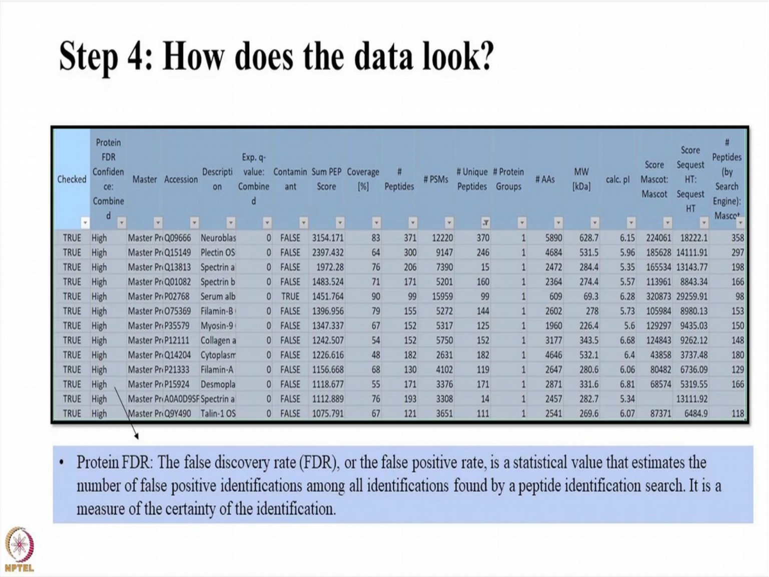
Unmatched spectra go to next engine!

Look for PTMS you didn't know about!

So, for the setting up of the work flow first of all we do the database search. As you can see here, here we have used SEQUEST HT and also at the these are the steps that are followed that is you can annotate the parameters, that is for example, you have put a parameter of mass tolerance which is 10 ppm, then you have gives the charge states and then the retention time. Also here, you can see this spectrum files will be taken and then it will go to the spectrum selector, then it will go to the database and then your there is another work flow which is called the percolator.

Other than these things you can also use other search engines like you can use MASCOT in parallel with SEQUEST HT, and then if there are unmatched spectra it can go to the next search engine also. For example, in this case it is the spectrum confidence filter. And then furthermore you can also use other softwares for knowing into the other modifications present in the peptide. For example, if you want to know whether your peptide is glycosylated or phosphorylated, then you can use these kind of, these kind of filtering to annotate those changes in your peptide.

(Refer Slide Time 16:11)



Step 4: How does the data look?

Checked	Protein FDR Confidence: Combined	Master	Accession	Description	Exp. q-value: Combined	Contaminant	Sum PEP Score	Coverage [%]	# Peptides	# PSMs	# Unique Peptides	# Protein Groups	# AAs	MW [kDa]	calc. pI	Score Mascot: Mascot	Score Sequest HT: Sequest HT	# Peptides (by Search Engine): Mascot
TRUE	High	Master Pr	Q09666	Neuroblas	0	FALSE	3154.171	83	371	12220	370	1	5890	628.7	6.15	224061	18222.1	358
TRUE	High	Master Pr	Q15149	Plectin OS	0	FALSE	2397.432	64	300	9147	246	1	4684	531.5	5.96	185628	14111.91	297
TRUE	High	Master Pr	Q13813	Spectrin a	0	FALSE	1972.28	76	206	7390	15	1	2472	284.4	5.35	165534	13143.77	198
TRUE	High	Master Pr	Q01082	Spectrin b	0	FALSE	1483.524	71	171	5201	160	1	2364	274.4	5.57	113961	8843.34	166
TRUE	High	Master Pr	P02768	Serum alb	0	TRUE	1451.764	90	99	15959	99	1	609	69.3	6.28	320873	29259.91	98
TRUE	High	Master Pr	O75369	Filamin-B	0	FALSE	1396.956	79	155	5272	144	1	2602	278	5.73	105984	8980.13	153
TRUE	High	Master Pr	P35579	Myosin-9	0	FALSE	1347.337	67	152	5317	125	1	1960	226.4	5.6	129297	9435.03	150
TRUE	High	Master Pr	P12111	Collagen a	0	FALSE	1242.507	54	152	5750	152	1	3177	343.5	6.68	124843	9262.12	148
TRUE	High	Master Pr	Q14204	Cytoplasm	0	FALSE	1226.616	48	182	2631	182	1	4646	532.1	6.4	43858	3737.48	180
TRUE	High	Master Pr	P21333	Filamin-A	0	FALSE	1156.668	68	130	4102	119	1	2647	280.6	6.06	80482	6736.09	129
TRUE	High	Master Pr	P15924	Desmopla	0	FALSE	1118.677	55	171	3376	171	1	2871	331.6	6.81	68574	5319.55	166
TRUE	High	Master Pr	A0A0D9SF	Spectrin a	0	FALSE	1112.889	76	193	3308	14	1	2457	282.7	5.34		13111.92	
TRUE	High	Master Pr	Q9Y490	Talin-1 OS	0	FALSE	1075.791	67	121	3651	111	1	2541	269.6	6.07	87371	6484.9	118

- Protein FDR: The false discovery rate (FDR), or the false positive rate, is a statistical value that estimates the number of false positive identifications among all identifications found by a peptide identification search. It is a measure of the certainty of the identification.

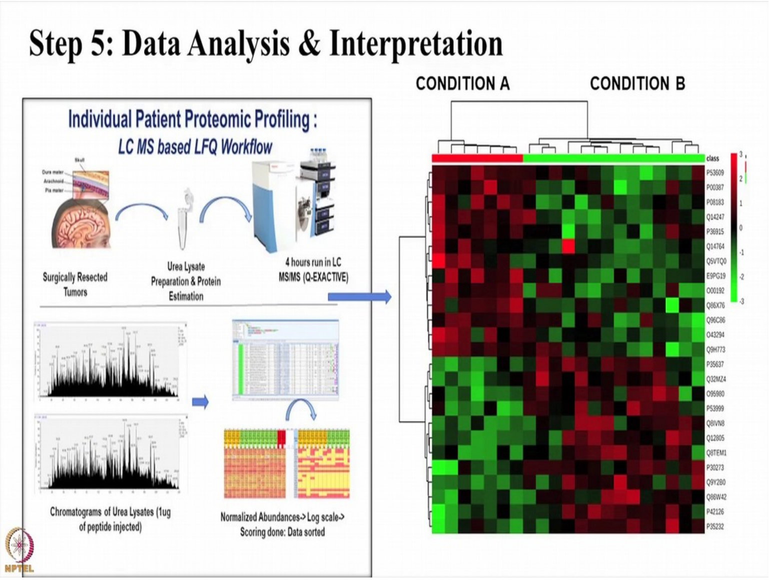
So, how does the data look? So, as you can see after get going through a lot of filtering criteria the data that comes out is of high confidence because you have put on stringent filters, so that whatever hits you get are the true data and not false positives. So, here you can see that the there are different tabs associated here. For example, this is the, this is where you have annotated whether the protein FDR confidence how much is the level. So, here you can see all of these proteins that have been identified, have high FDR. So, false discovery rate is the statistical value that estimates the number of false positive, identifications among the peptide, and it is also measure a certainty for the identification as in how much you are confident that the protein you have identified is a true match.

Then we also have a contaminant database which we can plug in with the work flow. The contaminant database is will actually indicate presence of keratin or serum albumin which are high abundant protein and often are responsible for giving false positives or masking your actual protein of interest. So, as you can see here also, in our data we have got serum albumin and the software has marked here as true. So, while all the other proteins there is false in case of serum albumin, it is marked as true, so you can remove this kind of protein in case of your subsequent analysis.

Other than that, you will also have a plethora of information like the unique proteins. So, the unique proteins again gives you an idea as to how confident you are of identifying the particular protein that has been ascribed to. For example, the first hit which is a neuroblast differentiation protein has 370 unique peptides; that means, the mass spectrometer is encountered the peptide 370 times or it has actually annotated the protein, the software is annotated the protein with very high confidence. Then you can also know about the score which is the score that is given to a by the search engine and many other details.

So, thus you will have all the information available after you have run the protein through, after you have run your raw file through a software and the software has different tabs that you can customize and you can set your parameters that you are looking for.

(Refer Slide Time 18:31)



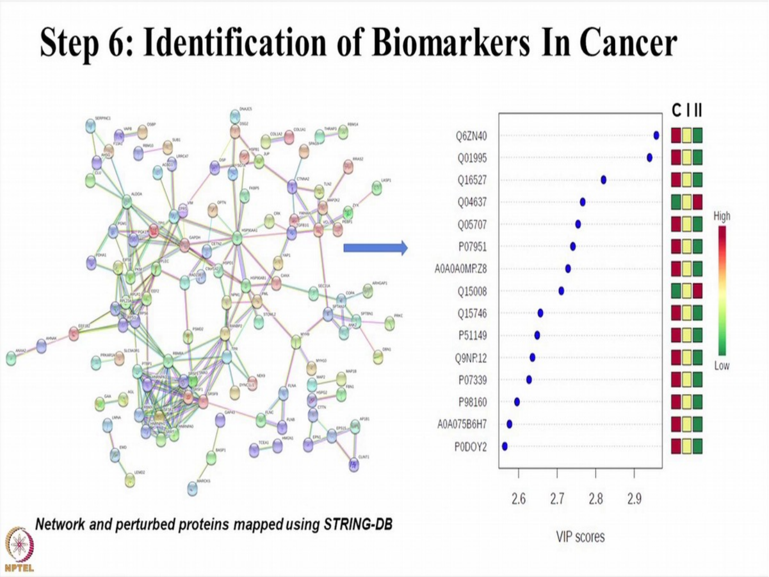
Step 5: Data Analysis & Interpretation

CONDITION A
CONDITION B
Individual Patient Proteomic Profiling :
LC MS based LFQ Workflow
Skull
Dura mater
Arachnoid
Pia mater
Surgically Resected Tumors
Urea Lysate Preparation & Protein Estimation
4 hours run in LC MS/MS (Q-EXACTIVE)
Chromatograms of Urea Lysates (1ug of peptide injected)
Normalized Abundances-> Log scale-> Scoring done: Data sorted
class
P53609
P00387
P08183
Q14247
P36915
Q14764
Q5VTQ0
E9PG19
O00192
Q86X76
Q96C86
O43294
Q9H773
P35637
Q32MZ4
O95980
P53999
Q8IVN8
Q12805
Q8TEM1
P30273
Q9Y2B0
Q86W42
P42126
P35232
3
2
1
0
-1
-2
-3
NPTEL

So, basically we have started from here, that is we have taken a cancer sample and normal samples we have used a patient cohort, then we have extracted the protein and then we have run the samples in the mass spectrometer. So, after doing all these exercise, you have got raw values, you have got annote the different peaks that you now you need to know that what are they. So, you have used the software which I told you that you can either use a freely available software like MaxQuant or a license software whichever resources available to you and then you can use the software and the various parameters to now do the data mining.

So, after the data mining what you are actually looking for is something like this wherein you can see a clear difference between condition A and condition B. So, as you can see there are signature list of proteins which are highly abundant in condition A and there are a signature list of proteins which are highly abundant in condition B. So, these are indicative of the actual biological changes that are happening in the patient sample.

(Refer Slide Time 19:35)



Step 6: Identification of Biomarkers In Cancer

Network and perturbed proteins mapped using STRING-DB

Furthermore, you can do the data curation and network analysis using again doing several bio informatics software; like string-DB, metaboanalyst etcetera. And now, you can see that after doing a the whole exercise of using mass spectrometer and the software for annotating the data, we get the, we can map the proteins in various networks like this. And, then we can also see which are the ones which are classifying the different grades of tumour or the or are different between the control and the cancer samples.

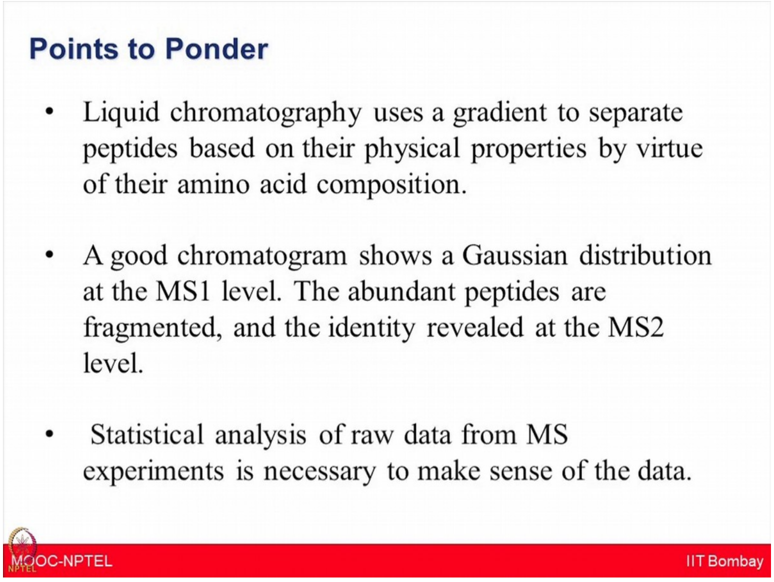
For example, these are a set of markers as you can see. So, you can see in this one, this is very high in the C sample which is the control sample and relatively low in the grade II samples. Similarly, there is a reverse trend for this protein you can see Q04637, so this is again showing a sequential increase that is it is low in the control samples and going high as the tumour progresses. So, thus using these kind of tools you can answer a lot of biological questions.

Alright so, we will started with work flow of mass spectrometry based proteomics. We talked about how to do the protein quantification, protein digestion, again we have talked about peptide quantification, then you are ready for doing the LC-MS/MS based analysis. So, liquid chromatography and MS parameters, what we generate the chromatograms, how to interpret chromatograms and how to review the whole data set, make more meaningful insight from this data for the clinical case studies.

I hope this gives you very basic. Of course, it is not so detailed, but you know a good glimpse of the work flow involved in doing MS based proteomics. A lot can be done using mass spectrometry based proteomics. We have just talked about protein identification and label free quantification work flows. We can also think about quantitative proteomics which I talked in the theory classes earlier, about using iTRAQ or TMTs and those work flows can be very useful as well. But, as long as you have done this work flow of sample preparation and their separation very well, then you are ready for the quantitative proteomics based work flow as well.

Thank you.

(Refer Slide Time 21:49)



Points to Ponder

- Liquid chromatography uses a gradient to separate peptides based on their physical properties by virtue of their amino acid composition.
- A good chromatogram shows a Gaussian distribution at the MS1 level. The abundant peptides are fragmented, and the identity revealed at the MS2 level.
- Statistical analysis of raw data from MS experiments is necessary to make sense of the data.

MOOC-NPTEL

IIT Bombay