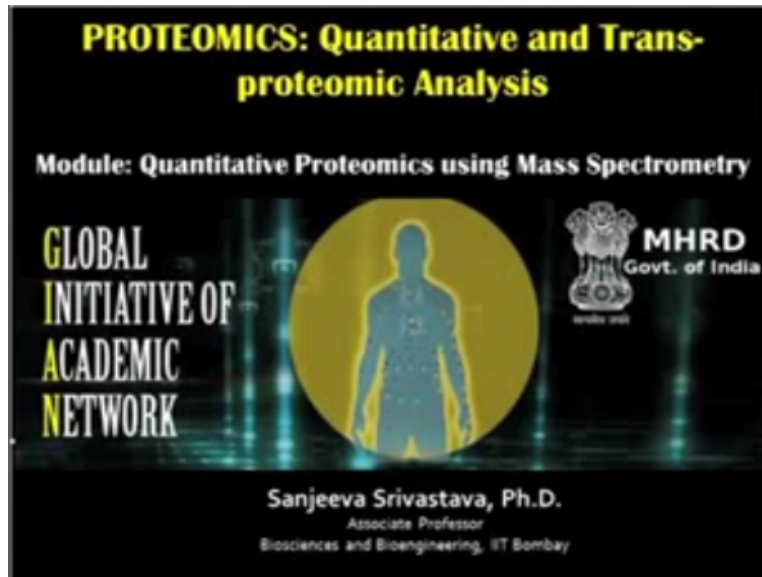


Introduction to Proteomics
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology – Bombay

Lecture–35
Quantitative Proteomics Data Analysis

(Refer Slide Time: 00:32)

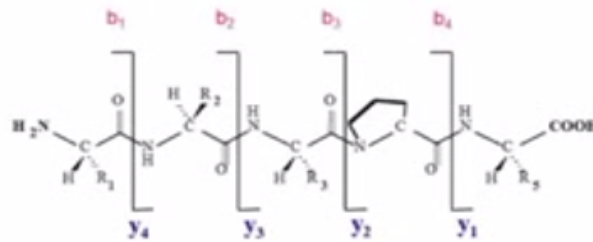


Okay. so we are talking about quantitative proteomics using mass spectrometry and we have discussed variety of instrument platform, variety of approaches which one can use to do the quantitative proteomics, right.

(Refer Slide Time: 00:47)

Peptide Ion Fragmentation: b and y ions

- At low collision energy peptides fragment typically at amide bond
- bond between carbonyl oxygen and amide nitrogen cleaves to form "y-ion" and "b-ion"



y-ions: positive charge retained on C-terminus of peptide ion
b-ions: charge is retained on the N-terminal

QAN

El Bombay

3

So, as you can know that during the collision in (()) (00:51) association when collision energy is subjected at that point, these peptide bonds they are broken and then you can see B ion and Y ion getting generated. So, Y ions are these positive charge which is retained on the C-terminus of the peptide ions, like on this side and that is why the Y ion series starts from the C-terminus and then B ions are those which are charge retained on the Normal-terminal side, so that you have B ions series starting from B1 to B4 from the N-terminal. So, these things you have already been aware now.

(Refer Slide Time: 01:26)

"y-ion" Series More Intense than "b-ion" Series

- Tryptic peptides generate doubly charged ions
- due to Lys/Arg at C-terminus
- Basic side chains in Lys/Arg retains +ve charge at C-terminus

QAN

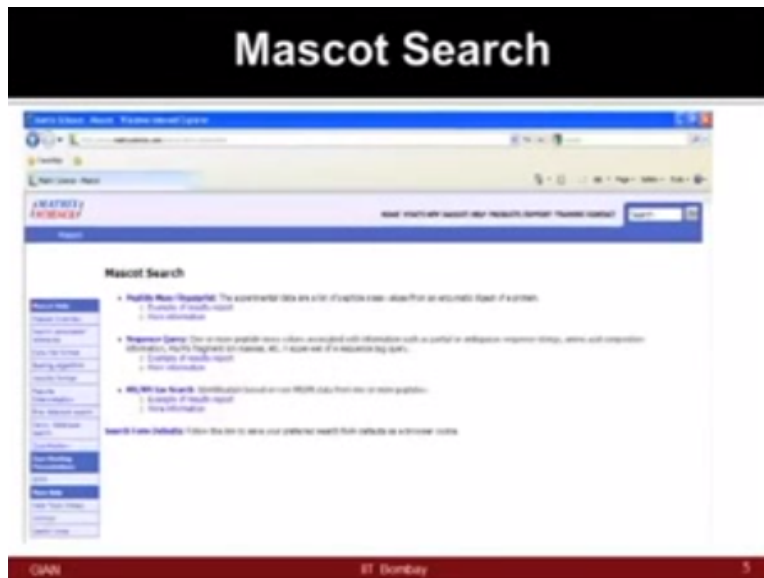
El Bombay

4

Question is why Y ion series is more intense than the B ions because if you look at the actual data you will find that many Y ions are red whereas very few B ions then are actually the red,

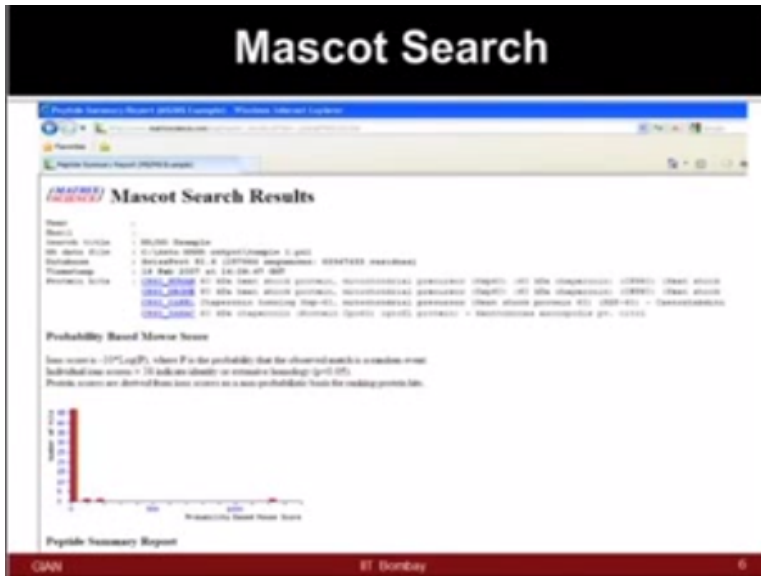
right. The main reason for that is basic side chains especially when you are doing a trypsinization for the cleavage which happens at the lysine and arginine residues, they are generating the positive charge at the C-terminus which is a part of your Y ion series and therefore because of the (()) (01:57) cleavage which happens at arginine and lysine residue, you will have much more intense Y ion series as compared to the B ion, right.

(Refer Slide Time: 02:08)



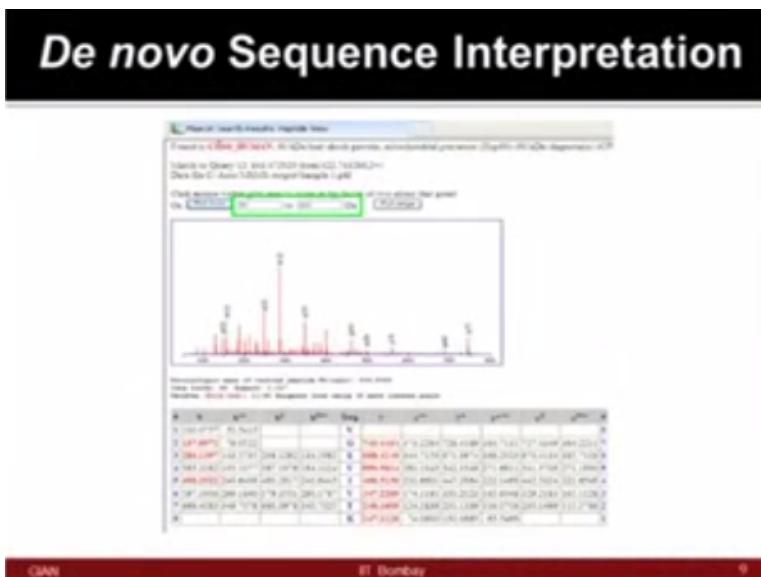
Okay, so we had then talk to you about Mascot Search, especially one of the open access database software available to do data analysis and you can do variety of analysis over there, peptide mass fingerprinting, especially if you have data coming from the MALDI TOF/TOF platform. Sequence query, you can do MS/MS ion search. There are example file. There are description for each parameter, I think those you should read in much more detail.

(Refer Slide Time: 02:37)



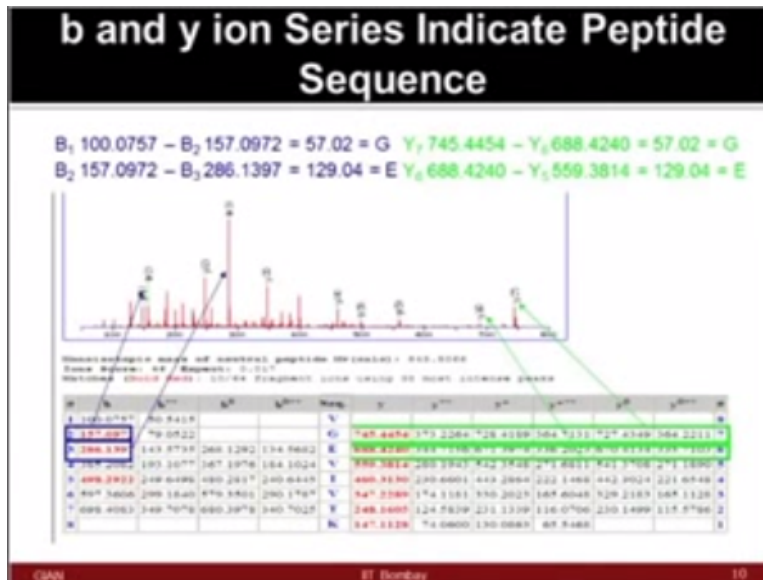
As you have seen in this example that how if your entire sample preparation (02:44) everything has worked out well and then your database search is able to pick up the right protein, then you can see a good score for a given protein and then you can establish the identity of the protein with good confidence. So, if you look into each of these peptides, you can then start seeing the B ion and Y ion series.

(Refer Slide Time: 03:03)



What is important here as I mentioned to you that if you zoom into one of these region you will see almost all of the Y ions and only very few of the B ions. This is what is expected. Now, how to cross check this sequence in formation, if this sequence is correct or not.

(Refer Slide Time: 03:25)



In other way, in which way can you establish the sequence assignment to yourself. Can you establish the sequence identity of these amino acids and peptide sequence yourself? So, as you can see here like I have, you know, kept it hiding for the time being. If you start looking at Y ion series, Y7-Y6 for example, that gives the difference of 57 which corresponds to amino acid glycine. Now, if you come to Y6-Y5, you are finding 129 which is you got another amino acid.

Now, if you come back here on B ion series, important thing here is that you are cross checking the data, okay. You are cross checking the data and if now from the B ion also you are getting the same difference, then only you can see glycine is correct. So, it means both the side both ways. You have done Y7-Y6 and you have done B1-B2. So, what you have to keep in mind that you are doing from Y7-Y6, not Y6-Y7 not in that manner.

Y7-Y6 which corresponded to 57 correspond to glycine and if you have crosschecked that for B1-B2 you are getting the same difference of 57 and then you can be confident that this is going to give rise to the same amino acid and in this manner if you derive, now you can derive this entire amino acid sequence and peptide sequence can be derived from this.

Not every B ion is available to you for calculation and crosscheck but as long as you can crosscheck two or three which are red in colour the intense B ions, then you are very confident that these amino acid sequences are correct. So, despite that you may not able to read all the

exact differences in your table and sheet given to you, but as long as you can cross check from both the sides C-terminus and N-terminal that these difference are almost similar, then you can be confident about that amino acid being correct.

So, this way one could actually manually derive these peptide sequences and although from the software you can immediately get this information because these values are already fit there and by now I think you can do these things yourself.

(Refer Slide Time: 05:32)

Factors Affecting Generation of Good y and b ions

- Certain amino acids exhibit unique fragmentation
 - loss of H₂O from side chains of Ser, Thr
 - loss of H₂S from Cys
 - loss of ammonia from Glu, Asp
- Effect of Proline on fragmentation
 - Pro peptide bonds are resistant to fragmentation
 - tryptic cleavage is prevented when Pro is located on C-terminal side of Lys/Arg
 - effect due to unique cyclic structure of Pro
- Different peptide bonds fragment differently
 - E.g. cleavage near Asp/Glu produces intense ions

QMS

IT Bombay

11

There are many factors which also generates how good these fragmentation ions can be, right. For instance, if you have the loss of water from the side chains of serine and threonine that is going to affect the fragmentation pattern and then your fragmentation may not be as good as you would expect. Same, if you have the hydrogen sulphide released from cysteine that may affect the fragmentation pattern or loss of ammonia from the glutamine and aspartic acid, those are also going to affect your fragmentation pattern.

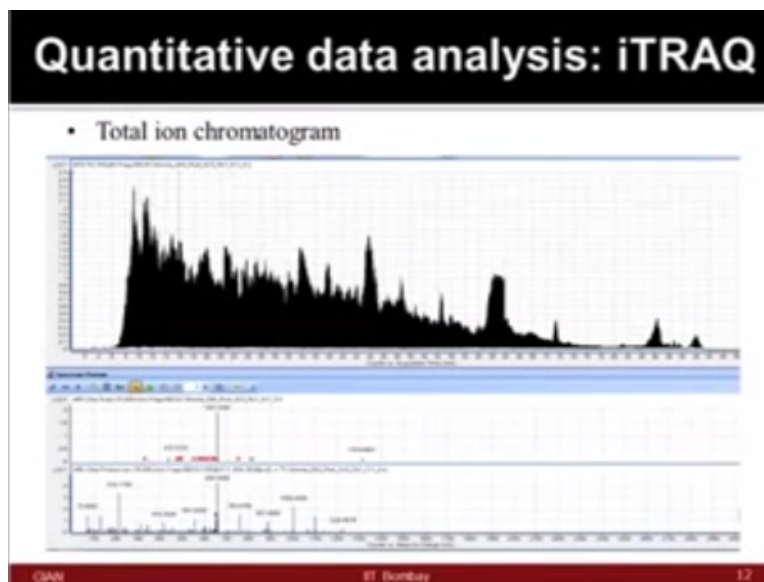
Proline being unique in nature having the cyclic structure is always, you know, creates some sort of challenges for the analysis part. So, it is resistant to the fragmentation. So, if the proline is just before or after the lysine and arginine residues, you will see the problem in the cleavage and therefore again your fragmentation and efficiency of cleavage will be affected. So, therefore there are different enzymes being also used if you think your protein is rich in proline amino

acid, okay.

So, then you will in addition to the trypsin, you should use chymotrypsin, you should use other peptide as well. So, different peptide bonds the fragment differently because there are certain amino acids near those there are very intense B ions being produced, so therefore aspartate glutamate they are going to create some anomaly in terms of the reporter or intensity. So, what it means that it is not always that your collision energy and your sample preparation are the only factors.

The kind of abundance of amino acid present in your given proteins is also going to govern how good your fragmentation is going to happen and what kind of intense you are going to see that.

(Refer Slide Time: 07:20)



I am just briefly covering about the qualitative data analysis especially the iTRAQ-based workflow. As you are aware that you have to do the pre-fractionation strategies to simplify the complex proteome, right. One strategy which you have seen is OFFGEL fractionation based on liquid isoelectric focusing. So, if you have done IEF and you have collected 20 to 24 fractions, each one of those fractions you are going to run on the mass spec, right.

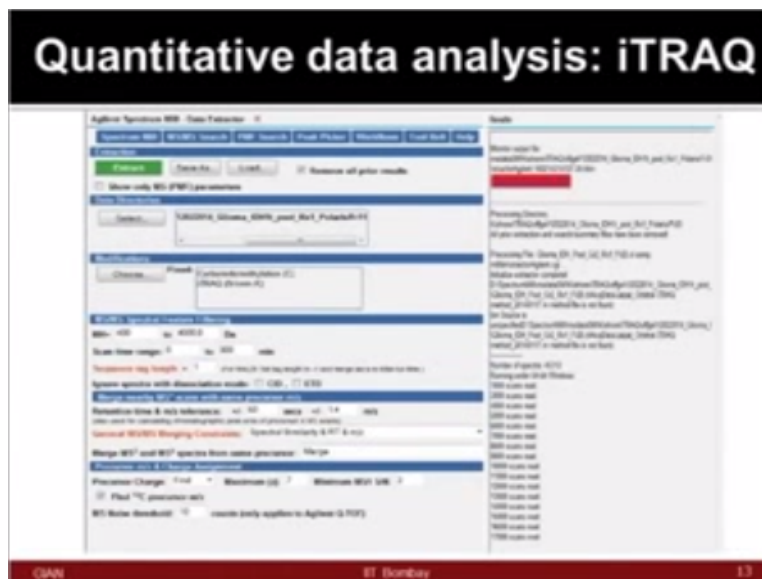
When you are running in the mass spec, you going to look for how good the peptide patterns are coming. The chromatogram is coming for each one of those fractions because if you would have

just run 20 minutes run for the entire complex proteome, still you will get, you know, this kind of pattern and if you have now run 20 fractions each around 20 to 30 minutes time, now you are seeing this kind of pattern. So, it means much more number of peptides you can now read.

Because the scan rates are fixed, you cannot change those, right. They are very far but still the way your ions are passing in the mass spec and the way the MS scan speed is there; these two things are you know you have no control. So, one way you can control that is the pre-fractionation, so that less number of peptides and less number of ions are coming each time and then later on you can pool all of this information to generate the entire proteome coverage.

Because proteome when we say, we are interested to know about let us say all the 5000 or 10,000 protein in a given sample. So, in that way coverage becomes very important and to do that people follow this strategy in the iTRAQ based workflow. You can see the MS ion, like if you just pass through you will certain MS fragmentation pattern. If you do MS/MS, now you can generate the reporter ion which could be used for the quantification.

(Refer Slide Time: 08:59)



The image shows a screenshot of a software interface for quantitative data analysis, specifically iTRAQ. The title bar reads "Quantitative data analysis: iTRAQ". The interface is divided into several sections. On the left, there are various input fields and checkboxes for parameters such as "Scan rate", "Scan time", "Scan length", "Scan delay", "Scan start", "Scan end", "Scan delay", "Scan start", "Scan end", "Scan delay", "Scan start", "Scan end", "Scan delay", "Scan start", "Scan end". On the right, there is a "Results" section displaying a list of data points, including "Scan rate", "Scan time", "Scan length", "Scan delay", "Scan start", "Scan end", "Scan delay", "Scan start", "Scan end". The interface is designed for detailed configuration and data review.

While I will not talk in much detail right now about the parameters, there are various platforms available for doing data analysis for iTRAQ. One platform which you will see today is based on the commercial sector what we have here available but you can still you use the Mascot if you do not have the specialised software available, you can still use Mascot for doing data analysis.

Only thing is if you have access to the Mascot complete database, probably your search is very much more refined.

Nevertheless, in the case, one need to define the range, one need to look at the scan time range, one need to look at the retention time by the tolerance what is the precursor charge you are expecting, the MS (()) (09:37) all of these parameters were need to define depending on the instrument being used and your knowledge about your sample and the runtime. All of these will be crucial factor for deciding these parameters.

(Refer Slide Time: 09:48)

The screenshot displays the 'Agilent Spectroscopy MS/MS Search' software interface. At the top, a black banner contains the text 'Quantitative data analysis: iTRAQ'. Below this, a bullet point indicates 'MS/MS search'. The interface is divided into several sections: 'Search Parameters' on the left, which includes fields for 'Sample Name', 'Database', 'Scan Range', 'Retention Time Range', 'Molecular Weight Range', and 'Charge Range'; and 'Search Results' on the right, which shows a table of search results. The table has columns for 'Scan', 'Retention Time', 'Molecular Weight', and 'Charge'. The interface also includes a 'Results' section at the top right and a 'Search' button at the bottom right. The bottom of the screenshot shows a red bar with the text 'QAN' and 'ET Dumbay'.

So, then you will go back further for doing MS/MS search from the same samples and now you have to look for various parameter for instant databases, what protein isoelectric point you are expecting because you would have use object fractionation for example, so you know which we PI value you have used, right 4 to 7 or 3 to 10.

You are looking at what kind of fixed modification, variable modification you may have, what instrument platform is used, monoisotopic mass, what is the precursor and the product mass tolerance which one need to define and you have to also mention that how many peptides your minimum want to see that, is it only one or more than one. So, all of these parameters are very critical and I am sure when you will do analysis today you will see by changing one by one how much impact that will make in terms of the overall data availability for that.

(Refer Slide Time: 10:40)

Quantitative data analysis: iTRAQ

• Protein summary

Protein	Peptide	Log2 Ratio	Peptide Count
1	1	1.14	28
2	2	1.15	19
3	3	1.17	18

At the end from doing this you will generate the protein summary data, although these things may not be so clear to you to read but what it generates that various spectra which has come from each of these samples, you will see how many peptide sequences you are able to derive. So, those which means are good hits. For instance, the ones in the red are 28, 19 and 18 peptides. It means of course as you can understand, these are much more confident hits you will have.

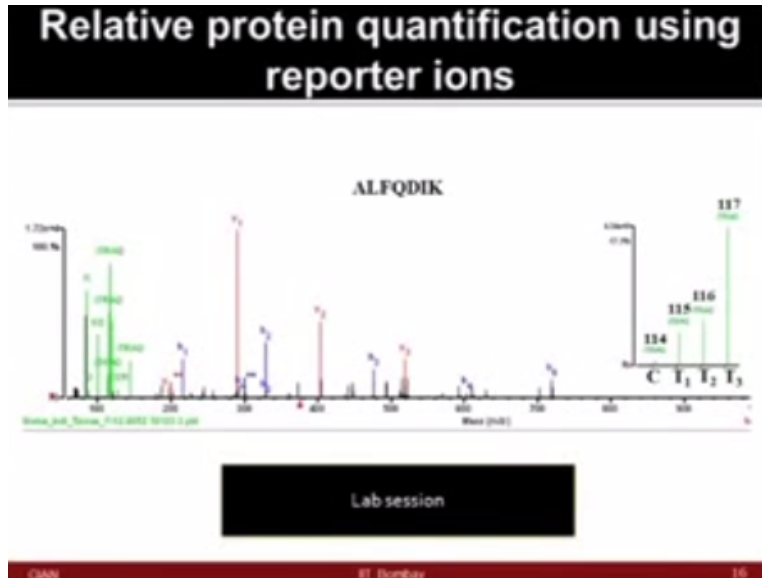
You have many peptides for that given protein and then you are looking at log ratios of iTRAQ labels. So, let us say it was a four place experiment of your control and three treatments. So, now you are looking at how 114, let us say that is controlled as compared to one of the treatment condition 115, what is the log for change ratio available, then you are looking at 116 x 114, you are looking at 117 x 114 accordingly.

So, depending on what reporter ions you have used, you can look into the four changes now. So, all of this information you can generate in the protein summary data. So, this entire analysis is pretty automated. What is important here is your stringency which is very important because while you will always think about how good are my coverage is, how many more protein I can get for differentially expressed ones.

At the same time what you need to really make sure that your thresholds are very stringent, so

that even you have less number of protein but those should be biologically relevant and they should pass the essential parameters and the filters, okay.

(Refer Slide Time: 12:06)



At the end of doing the entire work, you will derive these kind of sequences and you will look into for each one of these peptide sequences what is the reporter ions for MS/MS which will help you to do the quantification. So, these are the way of doing data presentation for this kind of iTRAQ workflow. Not every single data you can show in your papers and in your reports, but at least certain peptides you need to show that your quantification looks accurate, your claims what you are telling for the up or down regulation.

They are actually making sense from various peptides. So, that you need to show. At the end, you have to report all your data in the supplementary data and also now you have to upload all the data files into the PRIDE and some other open access databases where we are need to upload all the data file what you are generating from the mass spec because every single data you cannot put in the papers, but community should have access to your data.

So therefore it is important for you to upload the data in the generic raw data files.

(Refer Slide Time: 13:02)

**NOTE: This video is a part of GIAN
course conducted at IIT Bombay**