

**An Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. Karl Clauser**  
**Department of Biosciences and Bioengineering**  
**Indian Institute of Technology, Bombay**  
**Broad Institute of MIT and Harvard**

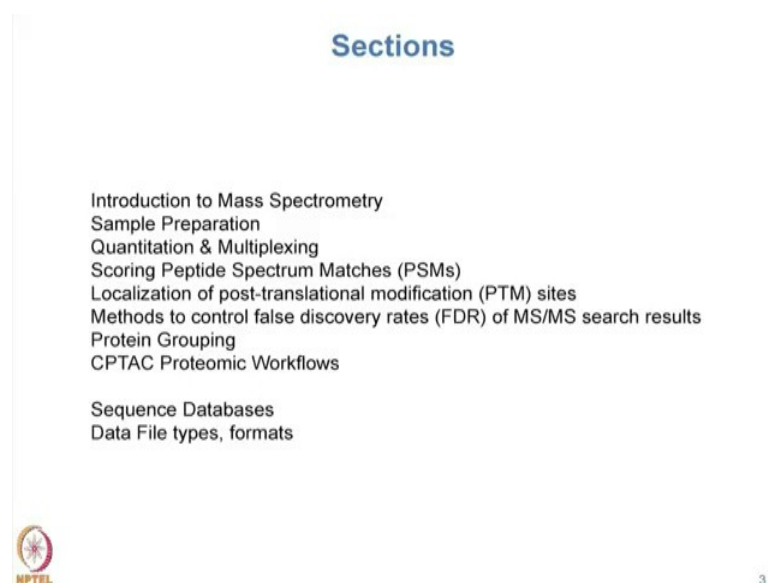
**Lecture – 14**  
**Introduction to Mass Spectrometry based Proteomics – I**

Welcome to MOOC course on Introduction to Proteogenomics. In the first module we learnt about genomics, in the second module we started discussing about proteomic tools, and in this light, in the last 3 lectures I try to give you the basics and foundation of proteomics.

Today lecture is going to be conducted by Dr. Karl Clauser, who is a principal scientist at the proteomics platform at the Broad Institute of MIT in Harvard. Dr. Clauser will focus on the basics of mass spectrometry based proteomics with emphasis on the electrospray ionization, the factors which influence good ionization efficiency and the architecture of a mass spectrometer.

So, let us welcome Dr. Karl Clauser for his lecture.

(Refer Slide Time: 01:22)




If you have seen the description of what the session was supposed to cover, it looks like this. As I put all the slides together I decided not to have them in quite that order. So, the

general flow of sections that I am going to go through is illustrated here. And so, let us get started, all right.

(Refer Slide Time: 01:39)

**Modern Mass Spectrometer (MS) Systems**



Fusion Lumos      Q-Exactive Plus & HFX      Quantiva

Discovery/Global Experiments      Targeted MS

**MS systems used for proteomics have 4 tasks:**

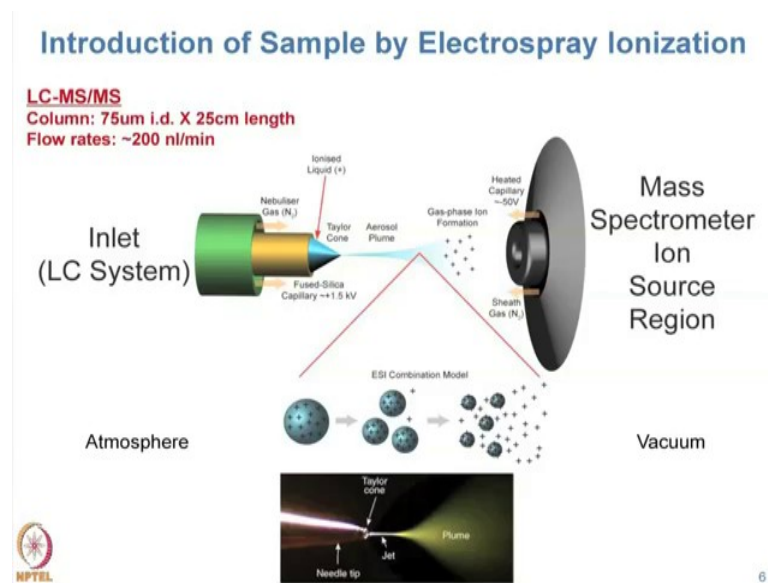
- Create ions from peptides
- Separate the ions based on charge and mass
- MS: detect precursor ions and determine their mass-to-charge
- MS/MS: select and fragment precursor ions to determine peptide sequence

NPTTEL 5

So, the first section I am going to teach you some of the basics of mass spectrometry. Today, in our lab in Boston we do a range of types of experiments in proteomics, including discovery based experiments and targeted experiments. These are examples of all the different kinds of instruments that are available from the manufacturer thermo. We have at least one of each in our lab and all of them have to do some of the same basic functions, ok. They have to create ions from peptides, separate based on charge and mass, ok. Although, we do mass spectrometry we do not actually measure mass, we always measure mass to charge ratio, ok.

The instruments basically do two types; generally two types of spectra, an MS spectrum or an MS-MS, right. The MS spectrum measures the complete mass of a peptide and the an MS-MS spectrum is after there is fragmentation and you measure the masses of the fragments that are produced from that spectrum or from that peptide, all right.

(Refer Slide Time: 02:52)



The, as of today in proteomics there is essentially one major ionization technique predominates the field and that is electrospray ionization. If this was 10 or 15 years ago there would have also been some MALDI or matrix assisted laser desorption ionization, but today most of what we do is really driven by electrospray, ok. So, this is a depiction of what is on the outside of the mass spectrometer. Here at atmospheric pressure you end up with a liquid that flows out the end of an LC column ah. There is a voltage applied, that causes droplets to be formed that exists as ions, and then those ions are transmitted into the mass spectrometer and along the way they have to become desolvated, ok.

This is a cartoon representation this is an actual picture of what that electrospray looks like. And the liquid is flowing, is most when doing proteomics most often it is through a liquid chromatography interface flowing at the most common scale to do this is 200 nanoliters a minute. Can't be flowing salt or detergent, peptides have to be in water a acetonitrile and a dilute acid most often that is formic acid at about pH 3, ok. If you have salt or detergent still in your sample you end up gunking up the front of the mass spectrometer and it does not work as well, all right, ok.

(Refer Slide Time: 04:32)

**Stable isotopes of most abundant elements in peptides**


**Most elements have more than one stable isotope.**  
1.1% of C atoms have an extra neutron, making their mass 13 Da.

**How does that help?**

- High resolution mass spectrometers resolve the isotope peaks.
- High resolution yields better mass accuracy.
- Isotope spacing reveals charge state.

Element	Mass	Abundance (%)
H	1.0078	99.985%
	2.0141	0.015
C	12.0000	98.89
	13.0034	1.11
N	14.0031	99.64
	15.0001	0.36
O	15.9949	99.76
	16.9991	0.04
	17.9992	0.20

$\Delta = 1.0034$   
 $\Delta = 0.996$



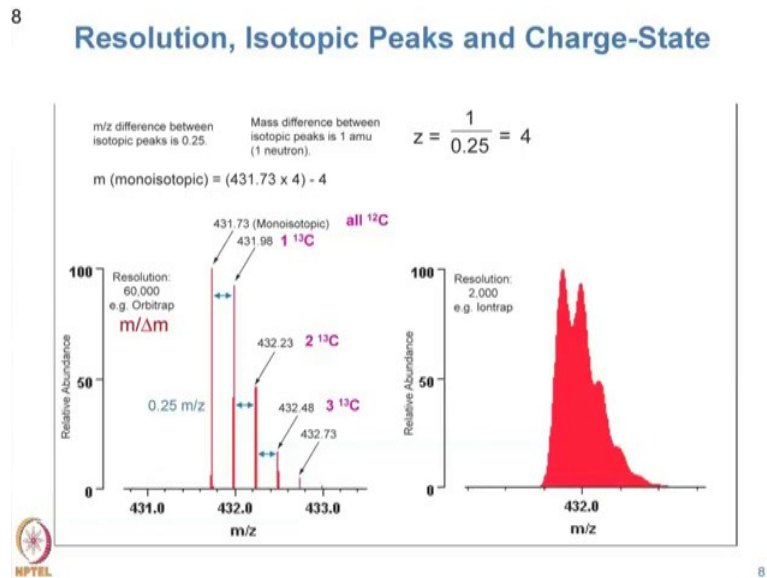
7

We measure mass, ok. Well, what is mass? Ok. So, the elements they contribute to proteins are dominated by hydrogen, carbon, nitrogen and oxygen. The masses are shown here. I want you to it is important to understand that each of those atoms or elements does not exist as a single species, ok, they have an isotope form. Those isotopes are not very abundant. So, you can see the one with the greatest abundance is carbon, and so 1 percent of carbon exists in the C 13 form, ok.

We are also going to be able to take advantage of that, it is crucial for doing quantitation these days for there to be the existence of carbon 13 and nitrogen 15. So, one other aspect that I want you to keep in mind is that the difference between carbon 12 and 13 is not the same as the difference between nitrogen 14 and nitrogen 15 that is going to be very valuable to us when we talk about multiplexing later, ok.

So, it was actually I have been doing this for about 25 years and it was only a few years ago that I was looking at something like this and I said wait a minute, a neutron it has a different mass depending upon whether it is attached to nitrogen or to a carbon. And the answer is yes of course, ok, because we do not measure the actual mass of something we measure the mass based on energy, ok. And the binding energy of the protons and the neutrons in a nucleus is different depending upon which element you are talking about, ok. And that is the source of the difference between adding a neutron to a carbon or adding a neutron to a nitrogen, all right.

(Refer Slide Time: 06:32)



So, once you get some data out of a mass spectrometer and these are some of the basic characteristics of it. I told you that there was isotopes, ok. So, instead of getting a single peak we get multiple, ok. This first one the leftmost peak or lowest mass peak is referred to the mono isotopic peak; that means, it was measured of a molecule that contained entirely the first isotope of the elements that are contributed to, ok. Because carbon is the most abundant isotope the major contributor to the other isotopic peaks is carbon 13, ok. So, here you have 1 carbon 13, 2 carbon 13s, 3 carbons, 4 carbons, ok. But I told you is only 1 percent, right. This is huge; it is almost as big right. Why? Well, once you add up something that has an intact mass of about 2000, there is going to be 100s of carbons, ok. So, the chance now of being at least one of them being carbon 13 goes way up, right.

So, once you have these isotope patterns there is also some extra valuable information encoded in there, right. Approximately the mass of a neutron is 1, but we do not measure mass or measure mass to charge ratio, ok. So, it is m divided by z. Say, in this case you can see that those isotopes spacing is 0.25; that means, the charge now is 4, and we can determine that because of that isotope spacing, all right.

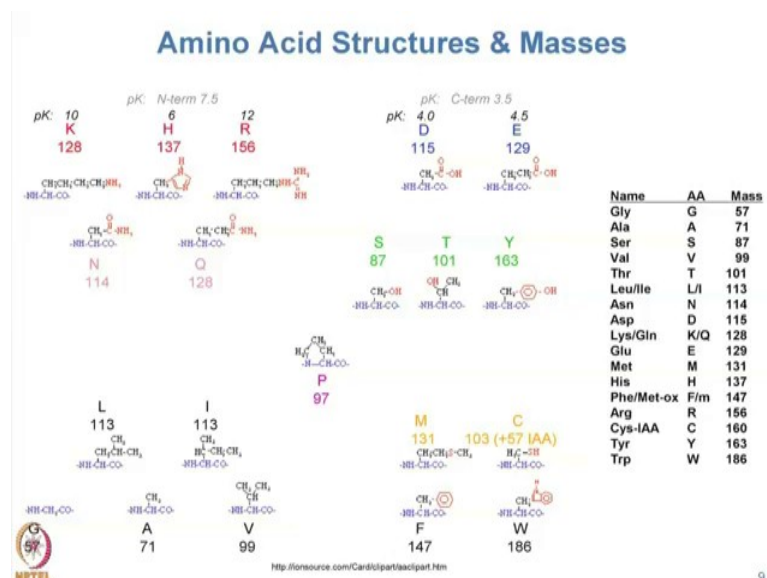
On a high resolution instrument this is what an isotope cluster will look like the numeric value assigned resolution is something like 60,000 that is the typical resolution that one will run an MS 1 spec scan on when using an orbit trap instruments, ok. Where does the 60,000 come from? The typical measure of resolution is m over delta m, so that means,

you would use in the numerator you would use 431, but the delta m part comes from the width of the one of the isotopic peaks, ok. So, that is very narrow. When you have a low resolution instrument 2000 would be the resolving power of this, right.

Now, it is also kind of complicated when you compare one instrument from one manufacturer to one instrument for another manufacturer because although they both use them; they will all use the measure of resolution, the resolution is not uniform across the mass range of an instrument, ok. And the particular manufacturers do not necessarily quote their resolution at the same mass, and furthermore do you make it even more complicated thermo uses different measure depending on which instrument they are talking about, ok. So, they will quote you a resolution number at mass 400 for their quadrupole type instruments and for the hybrid instruments they are quote mass 200, ok. So, when you start to say oh I want this instrument has got higher resolution. It does not, it is just they have changed where they give it to you.

But what you really need to know is how narrow are those peaks going to be and are they narrow enough to do things that are useful, ok. TMT 10 resolution is something we are going to talk about later and the ability to resolve the N and C isotopes of TMT reporter ions is important, ok, all right.

(Refer Slide Time: 09:53)



So, I have already said the gap there was 0.25 right, we measure amino acids, ok. Kelly showed you earlier when you were working with DNA or RNA, there is only four things,

right, and they are not glycine, threonine, alanine cysteine, which is what they would be if they were talking about amino acids, all right. There are 20 amino acids that are depicted here. This organization of the amino acids has them showing different properties of the amino acids and I was told that there might be a quiz conducted throughout the course of the workshop here and it just might be useful to know that my personal favorite amino acid is proline, ok, all right. And pick you. The mass of tyrosine is 163, ok. A couple slides from now you are going to need to know that, ok, all right.

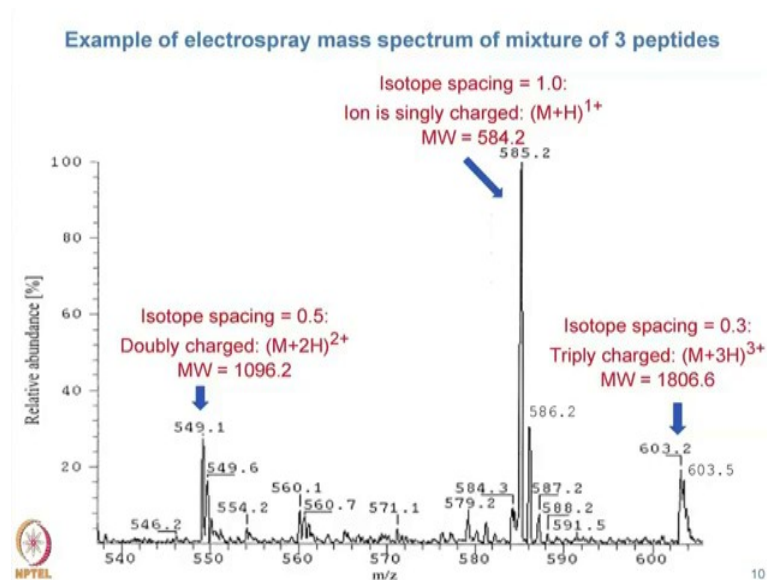
So, some other important; these particular amino acids over here leucine, you cannot tell the difference between leucine and isoleucine with a mass spectrometer because they have an exact same elemental composition, ok. But from the standpoint of measuring ions these ones are kind of boring, they do not too much. The reason proline is my favorite is because it causes all kinds of havoc. The side chain is bound to the backbone and it likes to put retain charge there, and so you get a nice huge peak in an MS-MS spectrum when there is a proline and you get an ion cleaving on the N-terminal side, ok.

We could not do proteomics if it was not for these 3 amino acids, in particular lysine and arginine, ok. They are basic, they bear charge, there are the basis for making the ions have charge. We do positive ion mass spectrometry, it is possible to do negative ion a spectrometry. I have doing this for 25 years, I never do negative ion, ok. If to me if it was negative I would be totally boring because you cannot do peptides work very well that way, all right.

But what I give you some property information here the pK of the basic group on the side chain is quite different and the arginine is much more basic, ok. It really wants to hold on to that charge much more strongly than do lysine or histidine. Do you have a question or a thought? No? Ok, all right.

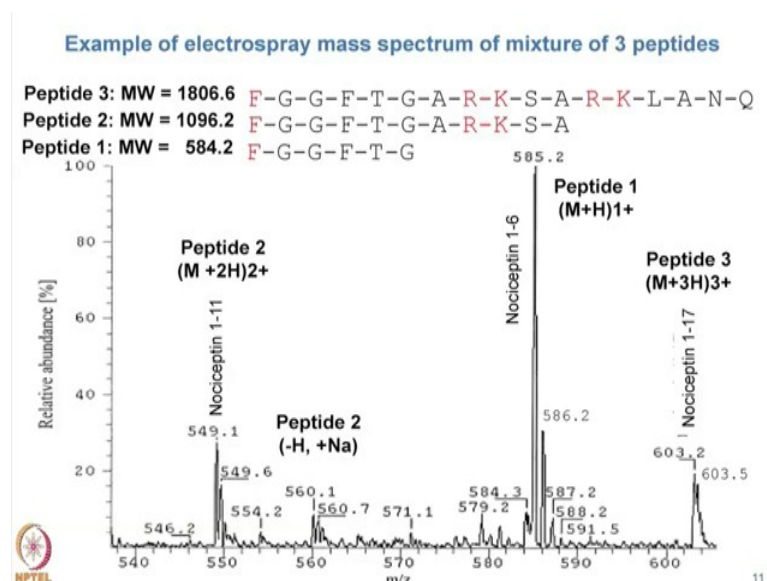
Please feel free to into interrupt me when, ok. So, later when we interpret spectra by hand, if you know all the masses you would be able to keep up with me. I have a big advantage over you that I just know all know my amino acid masses, ok. So, after we do some math, I can just tell you right away what the amino acid is, ok.

(Refer Slide Time: 12:52)



You are at least going to know tyrosine is 163, ok. Here is an old slide that illustrates an MS 1 spectrum. There is a few things I want to point out here. This is measuring multiple peptides at one time and you can see that the, this one is a singly charged peptide its isotope spacing is 1.0. This one over here is the doubly charged, and then I think one is triply charged here, right, so the spacing is 0.3. You can tell for them resolution or the peak widths there, so that this is an old instrument, ok. So, those things are wide, ok.

(Refer Slide Time: 13:26)

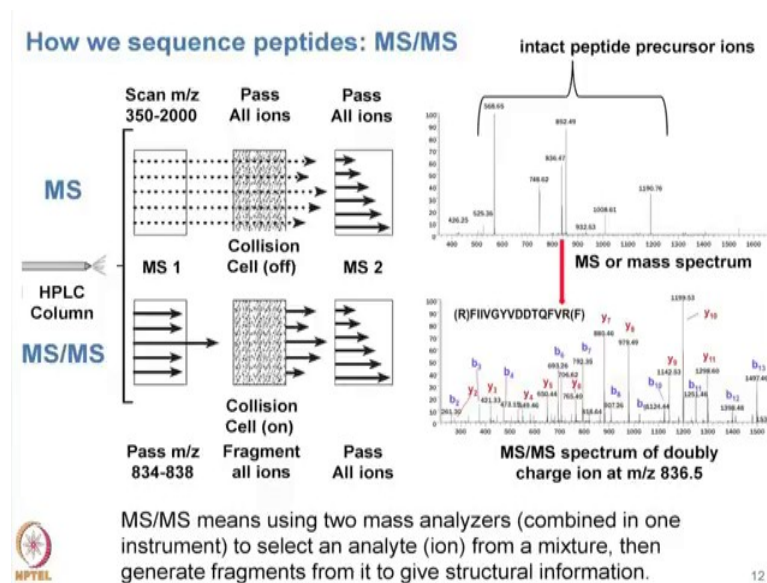




If we do label it a little differently one of the things you can see is that the masses of those peptides after you account for the charge converted back to singly charged, this peptide has no basic residue which is why it is easily singly charged. This peptide can hold two charges and you can typically will have charged on arginine and lysine and then on the interim terminus as well, ok.

It is also possible to hold charge on asparagines and glutamines, ok. But, here we have easily two basic residues, when they are right next to each other it is hard for them both to be charged, ok, all right. But here this one has got enough basic residues peptide 3 is triply charged. If the mass range was expanded wider some of these peptides would produce multiple charge states, ok. So, you would see 2 and 3, 3 and 4 something like that, all right.

(Refer Slide Time: 14:35)



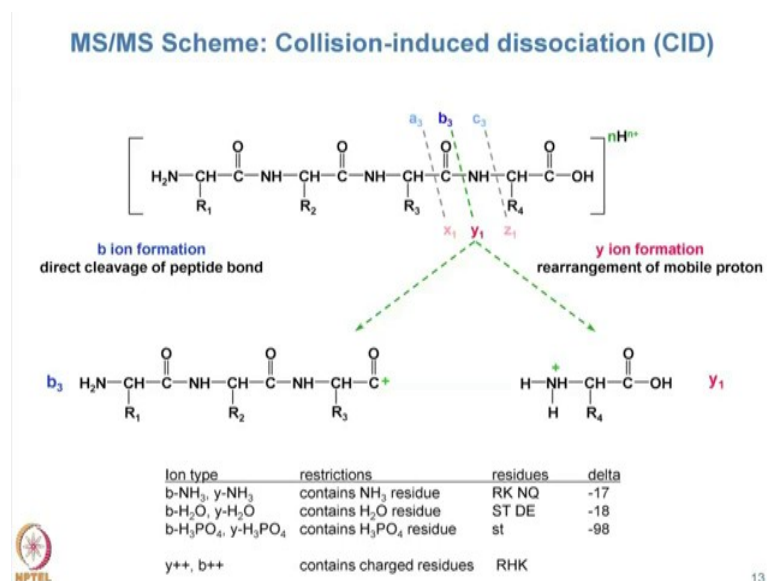
So, how do we do? How do we do MS and MS-MS? Ok. We use one instrument, ok, but it is going to do two things in. This is the simplest instrument, where there is one mass analyzer in a second mass analyzer with the collision cell in between. Graphically, this is equivalent to a triple quad instrument, where to do an MS measurement you let the first quad just pass everything through and then we turn off the collision cell and we measure every ion that went into the instrument and that gives you intact precursor ion information you may hear me switch between the terms precursor ion and parent ion, ok. When I am doing that, when I say parent ion I and I usually mean the singly charged

version of the ion and when I say precursor I mean the multiply charged that is not universal it is just my personal way of distinguishing what I am talking about, all right, right.

To do an MS-MS spectrum, we set the first analyzer to only pass ions of a particular mass, and then the width that we allow might be only about two Daltons, ok. So, only things that have, ok; in this figure it is about it is an old figure that, that width was 4, ok. So, we are going to allow anything with a mass of 834 to 838 to pass through collision cell it is going to turn on fragments into pieces and you get an MS-MS spectrum, ok. You can put 3 mass spectrometers together and you would get MS to the third, that works a little bit differently, but the principle is the same, ok, all right.

The mass spectrum right now is labeled b and y and we will skip ahead to talk about that, right. After this slide, we are almost never going to see elements again, we are always going to talk in terms of amino acids and we will use letter codes for amino acids.

(Refer Slide Time: 16:20)

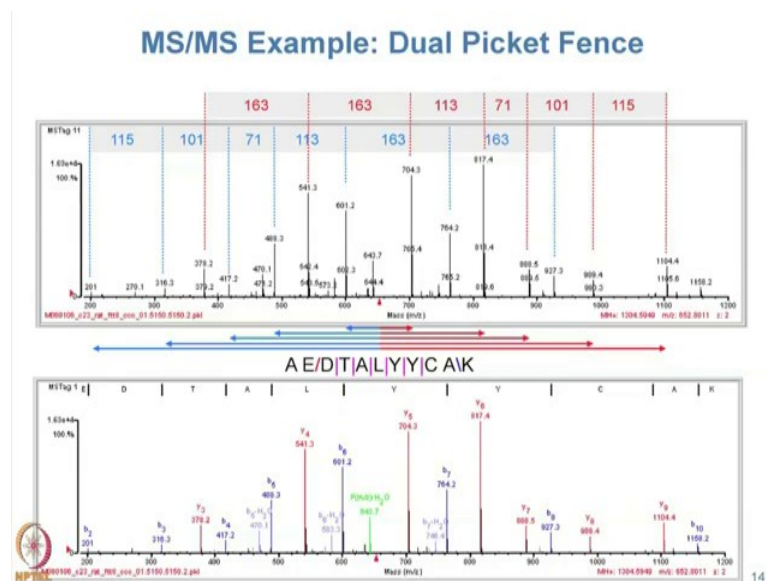


But I want you to keep in mind that a peptide is actually a molecule, and it has bonds between carbons and nitrogens and oxygens. This is a peptide backbone and if you look close you will see that there are essentially 3 different types of bonds here. The most common one that will fragment in during MS-MS is this one, and that will lead to a b ion. If the charge is retained on this side of the peptide and it will lead to a y ion if charge is on that side of the peptide, ok.

When we have a mass spectrum it is not the measurement of one ion, it is the measurement of many ions. Some of them are going to fragment here, others are going to fragment there, others are going to fragment different places and so the spectrum is the sum of all those events being measured together, ok. Why would you call this b and y? Right; well, that's because there was three things, ok. So, we call them a b and c, or x y and z. The most common dominating ion types though are being y, ok.

But it is not that quite that simple. You can also lose break bonds that are in the side chains of the amino acids, and in particular you can lose water and ammonia dependent upon having these particular amino acids in those side chains. If you have a phosphorylated residue you can lose 98 from the side chain of serine and threonine, ok. Again you can also have if you start out with triply charged or doubly charged peptide the fragments can have multiple charges on, ok, all right.

(Refer Slide Time: 18:01)



are going to whenever I show you these things, there are several things you should know. This is the file name which means nothing to you, but it tells me what project the spectrum came from. This little red carrot is the position of the precursor mass in the MS-MS spectrum. Over here is the parent mass which is singly charged. Here is the precursor mass  $m/z$  and then the precursor charge state there is listed as  $z$ , ok.

So, if you look at this spectrum, you can see there are spacing of peaks they looks pleasant, ok. There is also some apparent symmetry to the left and to the right, ok. That is it is helpful to look at that symmetry, ok. I have colored those blue and red because they are going to turn out to be b ions and y ions, we do not yet know which what they are, but a b ion plus a y ion that the fragments at the same place adds up to the precursor mass and that is the basis of the symmetry, ok. So, you once you have two ions that are on the opposite side of the precursor mass is symmetric they cannot be the same ion type, they must be complementary, ok.

So, if you then start to do math or subtraction to figure out the mass gaps between them you do not need to do math on ones that are symmetric to each other, ok, all right. So, then we look at some of the mass differences from peaks and if you had memorized all of them you would immediately recognize that those are masses that correspond to amino acids, ok, all right. Mass 163 corresponds to.

Student: It is Tyrosine.

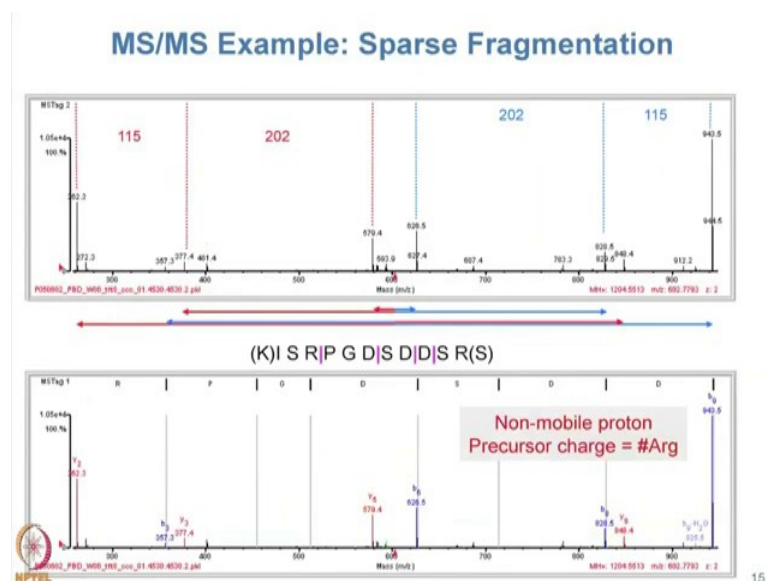
Come on say. Tyrosine, all right tyrosine. So, we can tell from this set of mass gaps this is the sequence. This part of this peptide is going to be tyrosine and tyrosine, leucine or isoleucine cannot tell the difference, alanine, threonine 115 is aspartic acid.

And then, this is a symmetric spectrum and you can find those same masses going the other way, ok. I have already colored them blue and red, but there is no obvious way to say that one is a b set and one is a y set, ok. And so, how do you decide? Ok. Well, if you can get to the end that helps you, ok. If you can get to, and in this case we cannot readily get to the end in either case, ok. This 201, there is going to be probably 2 amino acids that you have to add together to get to that one that is probably that could be a b<sub>2</sub> ion, ok. This comes from an older instrument where we do not have low mass available to us, ok. This is a tryptic peptide, ok, so that means, it is very likely to have a c terminal amino acid that is arginine or lysine, ok.

y 1 mass for arginine is 175 the y 1 mass for lysine is 147, ok. We cannot readily get from 201 to either 147 or 175 in a distance that is consistent with a amino acid mass. So, that suggests that is more likely to be a b ion than it is a y ion, ok. We could also figure out if we could get from here to the top which is that says 1364 is alright, ok. So, what I am trying to show you here is that there is a lot of information, ok. If we have a complete interpretation of the sequence it would be shown here, ok. The way I have slashes between the amino acids is to indicate that there is fragmentation. This red slash indicates that there was only a y ion there, ok. This says that there was only a b ion in that position and pink means that there is both b and y, ok.

So, this peptide is fragmented nearly completely, but it has not given us information about the order of the first two residues, ok. Together they add up to 201, but we do not know whether it is A E or E A, ok. And similarly over here we do not know the order of cysteine or alanine, ok, alright. But as with mass spectrum or driving down the street in India not all two wheelers are Royal Enfields , ok, sadly, alright.

(Refer Slide Time: 23:38)



So, some you get peptides that will fall apart to not give anywhere near as complete sequence information, And it is actually much easier to so manually interpret these things because you cannot get very far, ok. So, the best we can do is we can tell that there is some symmetry there. We have at least one mass gap that is consistent with an amino acid and then there are a couple of combinations of amino acids that will add up

202, ok. When you search a database this is the only peptide that will be consistent with all of that information, ok, alright.

(Refer Slide Time: 24:20)


**Factors Effecting Fragmentation and Interpretation**

**Range of Fragmentation Completeness**

- Too many or too few basic residues (R,K, H)
- Variability of fragment ion types and intensity
  - Instrument type dependent
  - Amino acid dependent
  - Altered by chemical labels, post-translational modifications

**Interpretation**

- Complementary ion types (b/y)
- Redundant information
- Possible direction uncertainty
- Unanticipated chemical or post-translational modifications
- Missing ions lead to composition with sequence order ambiguity
- Isobaric AA's yield sequence ambiguity
  - I = L (C6 H11 N1 O) = 113.08406
  - K - Q (C6 H12 N2 O, C5 H8 N2 O2) 128.09496 - 128.05858  $\Delta = 0.03638$
  - F - m (C9 H9 N O, C5 H9 N O S) 147.06841 - 147.0354  $\Delta = 0.0330$
  - GG = N (C4 H6 N2 O2, C4 H6 N2 O2) 114.04293
  - GA = Q - K (C5 H8 N2 O2, C5 H8 N2 O2, C6 H12 N2 O) 128.09496 - 128.05858  $\Delta = 0.03638$
  - DA - W - VS (C7 H10 N2 O4, C11 H11 N2 O, C8 H14 N2 O3) 186.06405 - 186.07931 - 186.10044  $\Delta = 0.01526 \Delta = 0.02113$

 16

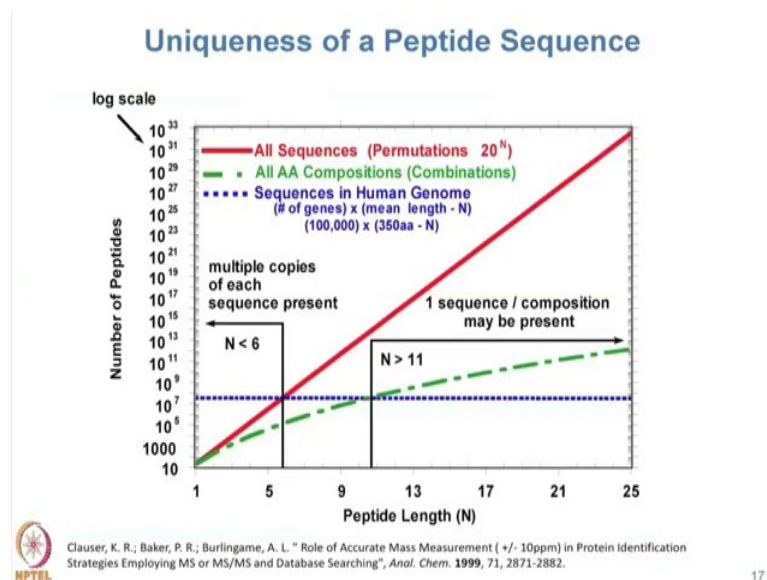
So, what I have tried to illustrate for you is that unlike DNA sequencing, where its typically the case that you get a clear complete answer after every step of the process. With an MS-MS a spectrum you have incomplete information, ok. And what I have tried to go through and tell you is some of the factors that contribute to that information being incomplete, and here is a list of those things here, ok.

The fragment ion types and the tendency of for them to occur depends a little bit on what kind of fragmentation mode you are using, there are different ones out there and I will come to that in a later slide, ok. You have complementary information in both directions, ok, which but that can lead to some uncertainty as to which direction you are in, ok. And I told you there are 20 amino acids and if you were looking for mass gaps you would hope to find a one of those 20, but there may be other modifications that happen that you have to be aware of and that is going to alter the mass, ok.

Then there are some amino acids, like I told you a leucine, isoleucine where you cannot tell the difference because they have the same mass. There happen, to also be some combinations of things that have the same mass, ok, all right and or very similar masses, ok. So, lysine and glutamine for example, both have 128 and if you have low resolution, low accuracy, you cannot tell the difference, but today with an orbitrap type of

instrument you do have the mass accuracy to tell the difference between those things, ok. Two glycines together is identical to one asparagine, ok. The glycine and alanine is identical to a glutamine. So, these things go together to make your life a little bit difficult.

(Refer Slide Time: 26:09)



This is a figure that I made in graduate school because I was faced at that time with many spectra there were more like the poor quality one, ok. And I wanted to know that this was before genomes were done, ok. And I wanted to know once the genome is done are those poor quality spectra still going to be good enough with a determine a sequence with a database of a complete proteome. The good news is I am still standing here today doing mass spectrometry. So, yes the answer is its good enough, ok.

But this is the nature of the figure that I made at that time it was still unknown how many genes were in the human genome and one of the highest estimates was that it was a 100,000 genes, ok. So, I used that as an upper limit and then I took the mean length of a gene or of a protein in the swissprot database at the time and that was 350. And then I made these calculations.

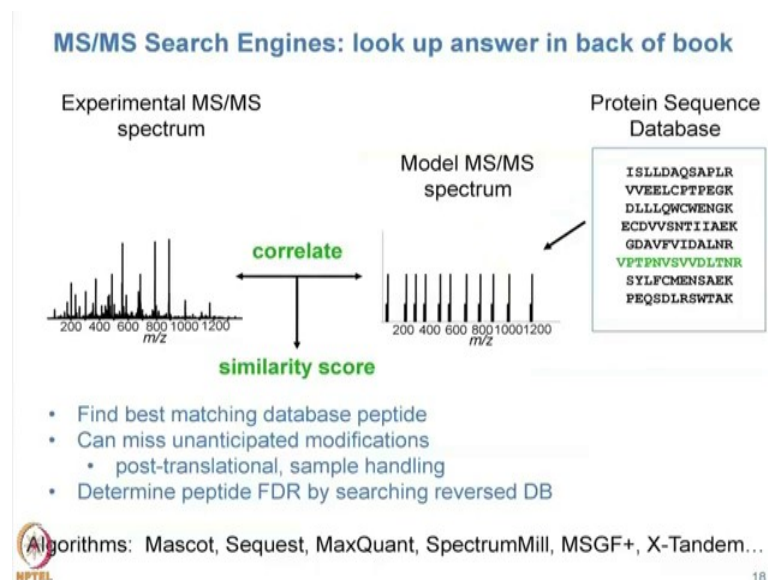
Given the peptide length here I wrote a program one of my first programs to go through and count how many peptides there were going to be at that length given these assumptions, ok. And so the red line is the simplest one which is to say if you could make all possible sequences and count them this is what that is how many there would



be, ok. Then I the blue line is how many peptides there would be in the human genome, ok. And then the green; once I found out that this was nowhere near close to the red line, I said well how little information could you actually do this with ok, and that is where I calculated the green line. So, the green line means we do not actually know the order of any of the amino acids, all we know is the amino acids that are present in the peptide there, amino acid composition. And to my surprise at the time you could start to get out to a point where if you had a long enough length the amino acid composition was going to be unique, ok.

So, you can still do a lot of good with these partial sequences, ok. Life of course is a lot easier when you get complete sequences, but life is not always giving us that, ok, alright.

(Refer Slide Time: 28:29)



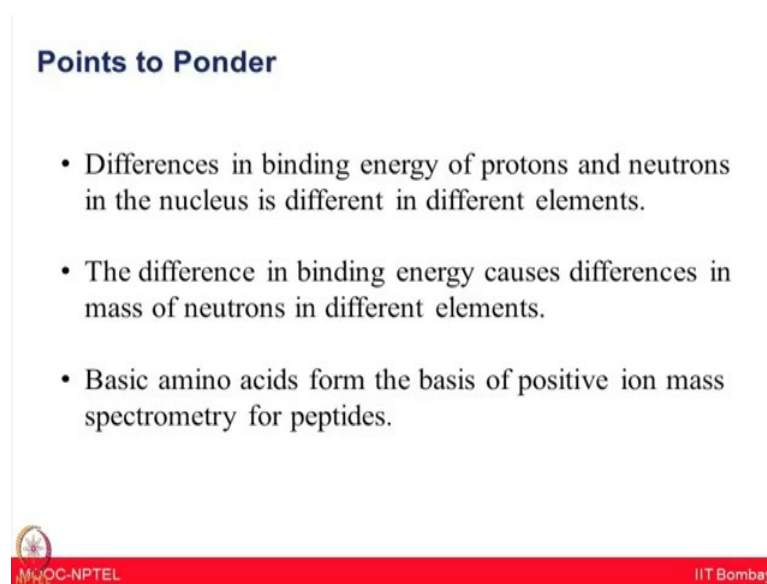
So, programs then have been written to do this kind of thing, and I have written one of them. The software package that I am responsible for developing is something called spectrum mill, alright. And, but nonetheless all of the software packages out there do this basic type of thing, ok. You will start with an experimental spectrum, you start with a database of sequences, ok, you have to take those sequences and digest them into peptides theoretically and then you make a model spectrum from the sequence and you match those two up. If you get a match, however you score it, then what you are looking for is to find the peptide in the database that gives you the best match, ok, alright.



So, when you do this if the database is incomplete then you cannot get the right answer, ok. And that actually causes some trouble because the programs are always going to give you the best answer and what you really like to know is can you only give me the right answer and if it is not there could you please just tell me that.

But they do not do that, ok, alright. So, that is can be frustrating, ok, alright. So, as a consequence though now what we do is since we cannot be sure we have got the right answer, let us at least estimate what our error rate is. And, so I am going to talk to you about calculating a false discovery rate from this type of thing.

(Refer Slide Time: 30:05)



**Points to Ponder**

- Differences in binding energy of protons and neutrons in the nucleus is different in different elements.
- The difference in binding energy causes differences in mass of neutrons in different elements.
- Basic amino acids form the basis of positive ion mass spectrometry for peptides.

MOC-NPTTEL IIT Bombay

So, building further from where I started about the basics of proteomics workflows and focus on the mass spectrometry, Dr. Karl has further given you the detailed concepts of the principle of electrospray ionization and why it is important in proteomics. Then he talked about importance of high resolution of isotopes for proper identification of peptide sequences, how collision induced ionization results in the formation of different types of ions which is useful for their identification was also covered. Then he showed you examples of different spectra, and how those could be interpreted manually. Something similar to what I talked in the previous lecture, but now you have seen much more detail about how to derive this peptide spectra manually.

Additionally, you are also informed about the factors which could affect the fragmentation of peptide. And lastly an effort were also made to help you understand the architecture of two very advanced mass spectrometers used in proteomics.

So, I hope from my previous 3 lectures and Dr. Karl Clauser's today's lecture, now you got a very good understanding about how to use mass spectrometry and different you know important considerations for doing proteomics using these kind of instruments. In the next lecture, Dr. Clauser will help you appreciate the importance of sample preparation and then moving on to the quantitative proteomics with the use of iTRAQ and TMT labels for quantitative proteomic analysis. Further he will talk to you about use of various search engines in protein identification.

Thank you.