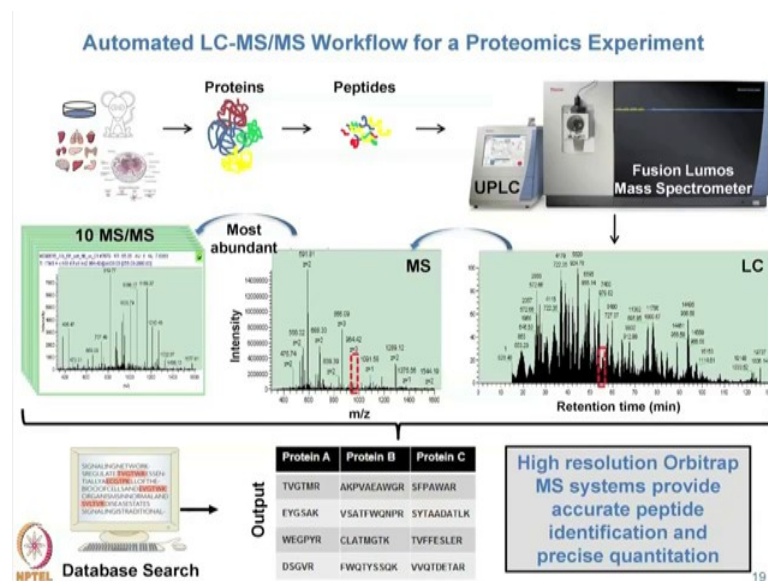


An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Karl Clauser
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
Broad Institute of MIT and Harvard

Lecture - 15
Introduction to Mass Spectrometry based Proteomics - II

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture, Dr. Karl Clauser introduced you to the basics of mass spectrometry based proteomics. Today's lecture we focus on the crucial steps in sample preparation for mass spectrometry based proteomics and also to provide a glimpse of label based quantitative proteomic approaches. Further the concepts of peptide spectrum match or PSMs and a spectrum library matching will be covered. So, let us welcome Dr. Clauser for his second lecture.

(Refer Slide Time: 01:21)



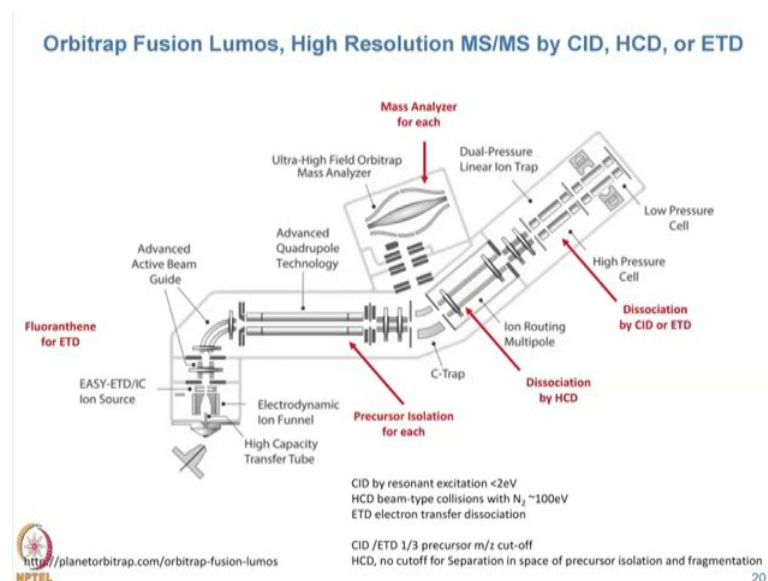
So, you start out with, to make that automated workflow happen you are going to start out with a source of material which could be tissue, could be cell lines. You are going to extract that into proteins, then we most often will digest the proteins into peptides, the peptides then go into a mass spectrometer and then this automated system does 3 basic steps, ok. It is going to separate the peptides chromatographically, eluting them over time based on their hydrophobicity and so, this in this description here that that runtime takes

about a 120 minutes, ok. At some given point in time a scan is going to happen this takes about 10, 10 to 100 milliseconds now.

The first thing you would do in a cycle has taken MS scan, you measure the masses of all of the peptides that are present, and then you will very quickly collect some number of MS-MS spectrum a common number to do now is 10. So, and it will take the 10 biggest things that it was observed in this MS scan and do MS-MS on it, ok. Today this cycle like this can happen in a second, ok, all right.

And then on this huge collection of spectra that are generated automatically. Then, will get put into a software program and it will try to match up assign peptide sequences to each of the mass spectra and then you will have some additional software that will try to take the peptides that belong to the same protein and you get out of list of proteins that were observed and all of the peptides that you have observed with that, ok.

(Refer Slide Time: 03:05)



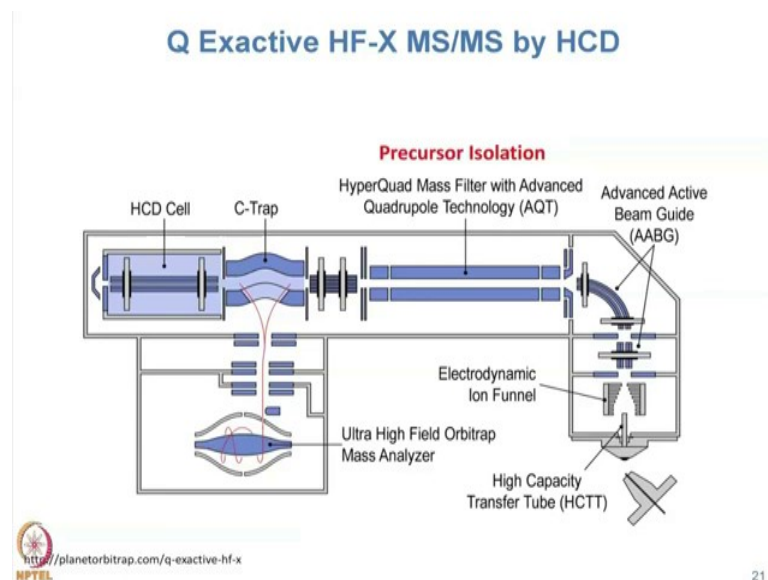
This is one of the once a most desirable instrument but that it is most desirable only if you can afford it. It is one of the most expensive instruments and it can do a whole lot of things, ok.

The CPTAC program is currently generating almost all of its data on this kind of an instrument but we do not use the entire capability of the instrument, ok. So, this instruments of Fusion Lumos from thermo, you put ions in here, you then have a way of

isolating precursor ions here and then you can do MS-MS by 3 different techniques. You can do something called high energy higher and higher energy collision dissociation, you can do collision induced association or you can do electron transfer dissociation, ok and then you measure things the spectra in the orbitrap.

It is also possible to measure them in the ion trap out here at lower resolution. You can go faster with lower resolution if you go out here, ok. In practice and the CPTAC program for generating proteogenomic data, we generate only HCD spectra and we collect mass spectra only in the orbitrap.

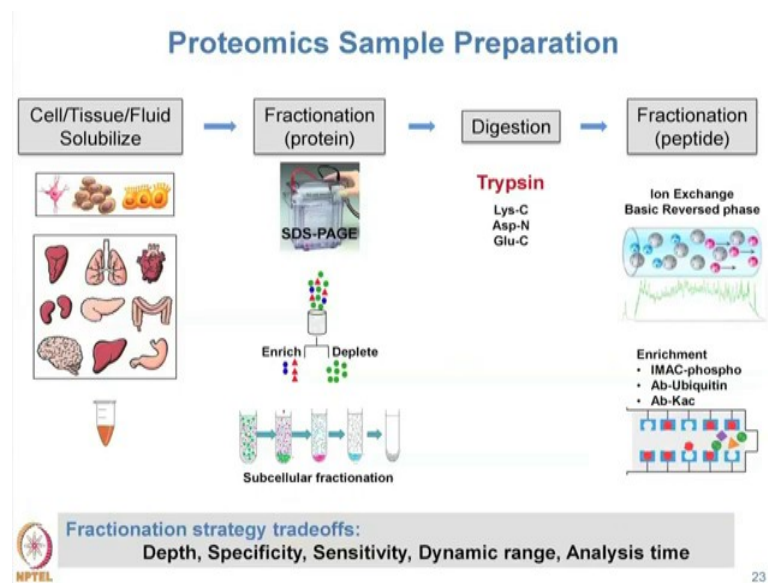
(Refer Slide Time: 04:29)



So, we are not taking full advantage of the instrument. The reason we are using those instruments is because this instrument which does only the things that we really want to did not become available till just earlier in this year and we started the grants that we were doing a year before that, ok, all right.

So, this I think of as today the workhorse type of instrument that if one was setting up a lab to do, proteogenomics in the way that we are going to describe having done it in the CPTAC program, this instrument would be the one would get today.

(Refer Slide Time: 05:03)



Sample preparation, ok; so, in proteomics these are some of the basic considerations that you have to do in designing your experiment, ok. We are quite often going to start with either cells tissue, fluid, fluid might be blood for example, and then there is going to be some set of separations that we are going to choose to do, ok. You have the choice of maybe you want to do some fractionation at the protein level, you might want to do some enrichment or depletion at the protein level, if you are working on cells and you care about mitochondria you might do a preparation that gives you an enrichment of that subcellular fraction that you are interested in, ok. From the standpoint of proteogenomics we do not do any fractionation at the protein level, ok.

The first thing we do is digest in peptides and then it is all about separation of peptides after that. If you are going to do, so fractionation of the protein level it is usually because you are after some particular subset of things or let us say you are doing a plasma, plasma; the most abundant protein in plasma is.

Student: Albumin.

Albumin, right; and it is the least interesting protein. But it is. So, what is the first thing you want to do is get rid of it, ok. So, you use a depletion step to get rid of albumin before you go to peptides, right. But for the purpose of doing cancer proteogenomics, we take our tissue grind it up, go to peptides and then we are going to fractionate peptides. Typically, if you are going to do it offline before you go to the instrument what you want

to do is choose a methodology this going to give you a different kind of separation than the one that is going into the instrument.

So, two common ways of doing that or either ion exchange or what we most commonly do now which is basic reverse phase, ok. So, it means we are running a reverse phase separation, but we run it at pH 10, ok. The separation that goes into the instrument goes at pH 3, ok.

The another thing that you want to want to do is enrichment, ok. So, if you after Phospho peptides you do not have to sequence everything else to get to your phospho peptides. So, you use something to pull them out, we use immobilized metal affinity chromatography. If you are interested in lysine acetylated peptides you can isolate them with anti-acetyl lysine antibody, ok, all right. So, in choosing what you want to do you are looking to make a trade off among these criteria, ok, all right and most proteomics today is done in a way where we are there is going to be a digestion step to peptides, ok.

(Refer Slide Time: 07:48)

Protein fragments are easier to analyze by MS

- Proteins are enzymatically digested into **peptides**
- **Trypsin** most commonly used enzyme in proteomics
- Cleaves C-terminal of **Arg** and **Lys** (positive charge)
- High specificity (cleavage rate >90%)
- Range of peptide length is **7-35** residues
- Tryptic peptides are at **least 2+** and easily ionized (N-term and C-term)
- Cysteines typically reduced/alkylated before digestion

Alternative enzymes can enhance sequence coverage

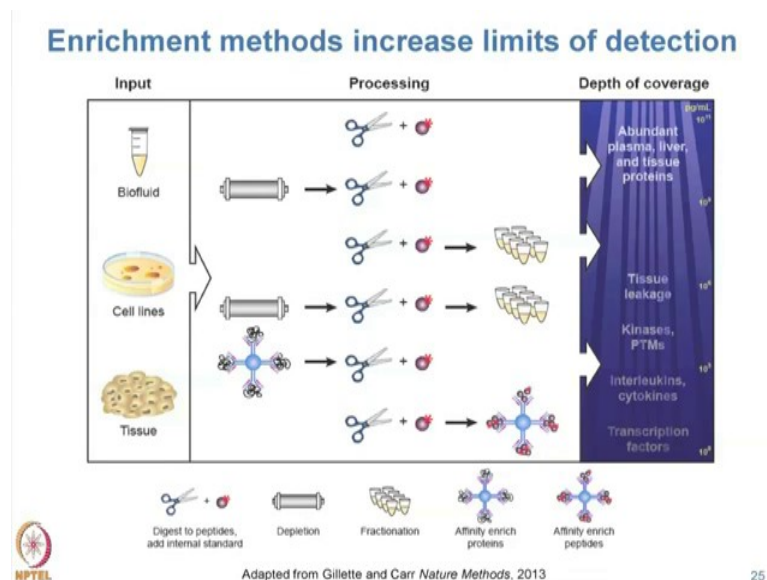
Trypsin cleavage sites (C-terminal)

24

Trypsin is by far the most common enzyme to use, and it gives you convenient lengths of peptides that are generally tend to work well on a mass spectrometer. They have the property that they have a basic amino acid at the C-terminus which is going to give you somewhat better fragmentation then if it is not at the C-terminus, ok.

Cysteines can be disulfide linked when they are in a protein. If you just reduce them they are very hard to chemically maintain throughout your process. So, what we typically do is reduce the; break the disulfide bonds and then alkylate them with some agent like iodoacetamide that then makes them readily detectable, ok, all right.

(Refer Slide Time: 08:37)



So, when you are doing enrichment, here you want to think about whether you are doing enrichment or depletion at the protein level. Here then there would be a digestion step, and then consider fractionation or affinity enrichment. The reason that you would make all of these kinds of strategic decisions is probably with some goal of increasing your depth of coverage.

So, if you want to start out with a complex sample and you are only interested in these things that are low abundance. There is it going to be typically some form of affinity enrichment involved or depletion of more abundant components, all right.

(Refer Slide Time: 09:18)

PTMs currently amenable to large-scale LC-MS/MS

▪ PTMs are substoichiometric and require enrichment!

PTM	Mass shift (Δm ;Da)	Amino acids	Frequency (tryptic peptides)	Enrichment methods
Phosphorylation	79.9663	Ser, Thr, Tyr (Asp, His)	3.1%	IMAC, TiO2 and antibodies
Ubiquitination (diGly tag)	114.0429	Lys	0.08%	Anti-diGlycine tag antibodies
Acetylation	42.0106	Lys	0.07%	Anti-acetyl lysine antibodies

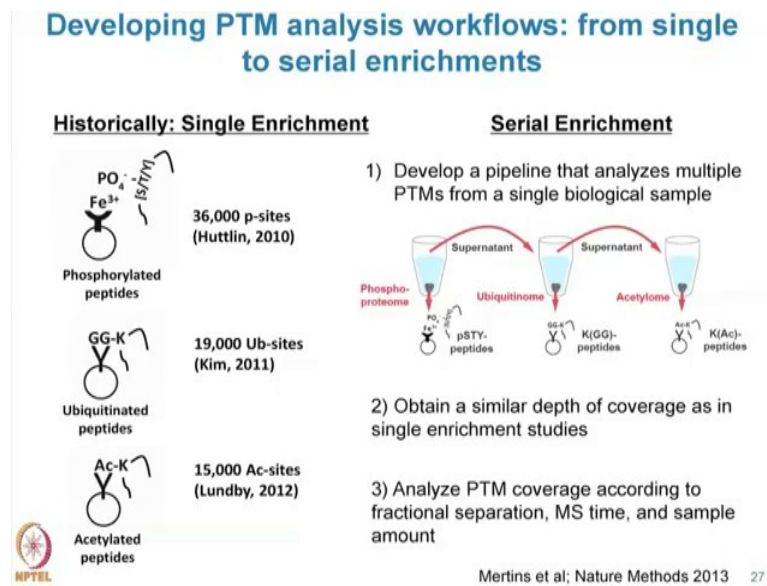
The slide displays five chemical structures representing different PTMs. From left to right: 1. p-Serine: A serine residue with a phosphate group attached to the hydroxyl oxygen. 2. p-Threonine: A threonine residue with a phosphate group attached to the hydroxyl oxygen. 3. p-Tyrosine: A tyrosine residue with a phosphate group attached to the hydroxyl oxygen of the phenolic ring. 4. diGlycine-Lysine: A lysine residue with a di-glycine tag (two glycine residues) attached to the epsilon-amino group of the lysine side chain. 5. Acetyl-Lysine: A lysine residue with an acetyl group attached to the epsilon-amino group of the lysine side chain. The structures are labeled below as p-Serine, p-Threonine, p-Tyrosine, diGlycine-Lysine, and Acetyl-Lysine. The NPTEL logo is in the bottom left, and the number 26 is in the bottom right.

So, the ones that are the most common post translational modifications that are that people can work on that are do by large scale methods phosphorylation of course is the most significant one. In our lab we also do a lot of ubiquitination work; this happens, this is done by having a glycine-glycine which is the starts out on the ubiquitin. So, the ubiquitin is covalently bonded to a lysine in a protein. When you treat it with trypsin it cuts off the ubiquitin but leaves two glycines that were the C-terminals of the ubiquitin, ok.

Acetylated lysines are something else that we also now are doing routinely in the CPTAC program, ok. So, and you can do these by using an anti-acetyl lysine antibody, all right.

So, we also have done this in a way where we do not have to split up the sample and loose require more sample and they dedicate only some of it to one modification only some of it to another, some of it to a third. Instead, we do the enrichments one after the other, ok.

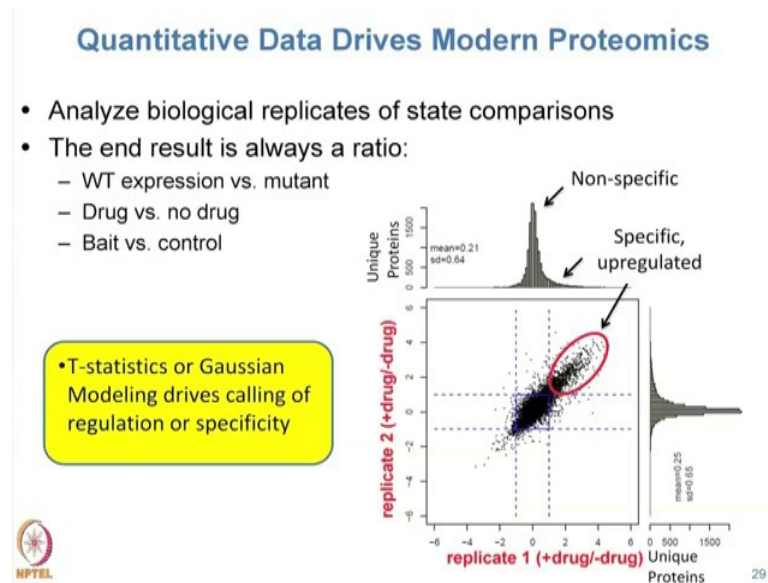
(Refer Slide Time: 10:37)



So, the supernatant of what comes through a IMAC column can then be used to, so the things that do not bind to the IMAC column come through you can then do the next step of enrichments for something else. And in our case, we can do we; published work on doing those 3 things in serial, ok. So, you start with less total sample in and achieve all each of those items, all right, ok.

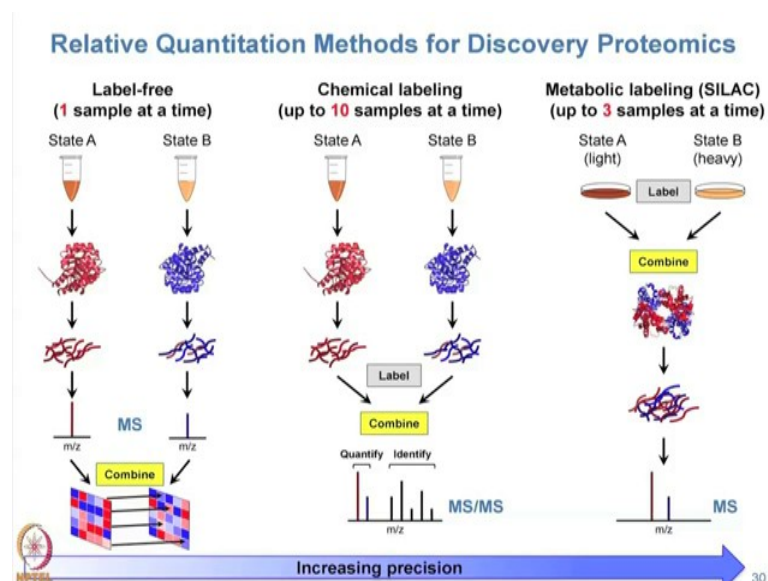
Ok Quantitation and multiplexing; the almost anything that we do in our lab today in proteomics and quite a lot of labs are trying to do things that are quantitative, ok. And the basics, basis of doing something quantitative and having some statistical power requires that you have replicates, ok.

(Refer Slide Time: 11:35)



So, not only do you want to have replicates, but you typically want to compare two, at least two conditions and examples of this might include wild type versus mutant expression, treatment with a drug or without a drug or capturing something with a bait or not. And then most of what you detect is probably going to be unchanged between the conditions and you are looking to do statistics to recognize some subset of things which change between the conditions that you do, ok, all right.

(Refer Slide Time: 12:07)



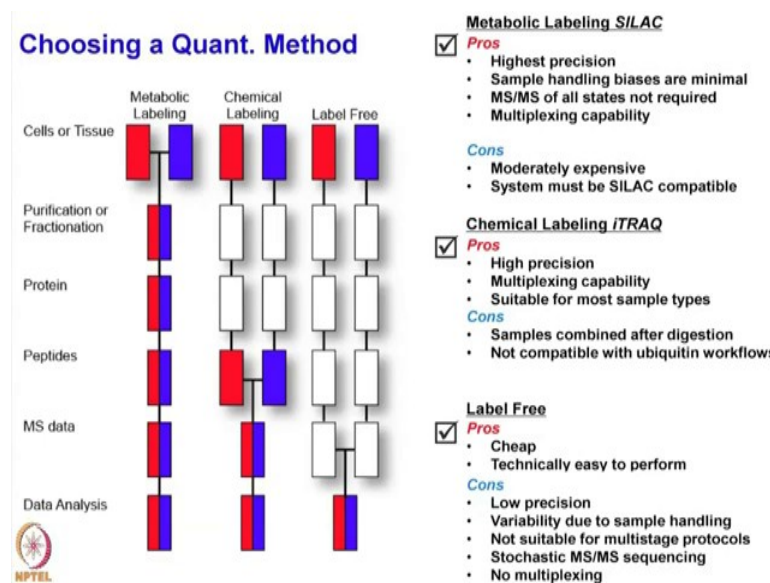
What sort of experimental design considerations that you put into this; ok. I am going to show you 3 different techniques, one here is called label-free where you basically combine the samples at the end; right then I am going to show you two labeled techniques, one is SILAC; stable isotope labeling of amino acids in cell culture and then the third one is something called a TMT or iTRAQ where you are using a chemical labeling agent.

You then are going to combine the samples and then put them into a mass spectrometer. You do MS-MS the quantitation comes at the level of the MS-MS spectrum, right. In this technique here you combine earlier in the process and the quantitation comes at the MS level, ok. Multiplexing wise you can do 3 things, 3 different 3 samples at a time; picture there are only 2. You can do light and heavy; the third one would be medium, ok, all right.

With a TMT 10 reagent you can put together 10 samples. If you have an iTRAQ agent, there is actually two iTRAQ agents, one is called iTRAQ 4 and the other one is called the iTRAQ 8, ok. So, that tells you how many samples you can put together there is also something called TMT 6, ok; and I will get into some of the differences in what you have to have to be able to do those kinds of experiments, ok. So, here are some of the features about this; takes a lot more time to do an experiment this way, a lot more instrument time.

Here there is some loss of accuracy in the quantitation due to compression that I will talk about the reagents are can be expensive, ok, all right. Here you have a less potential to duplexing and in order to get a heavy label you have to be able to add that to the to the cell culture that that is going on, so that means, you cannot label humans, ok. So, you can this is really suited to working with cells and cell culture, ok, all right.

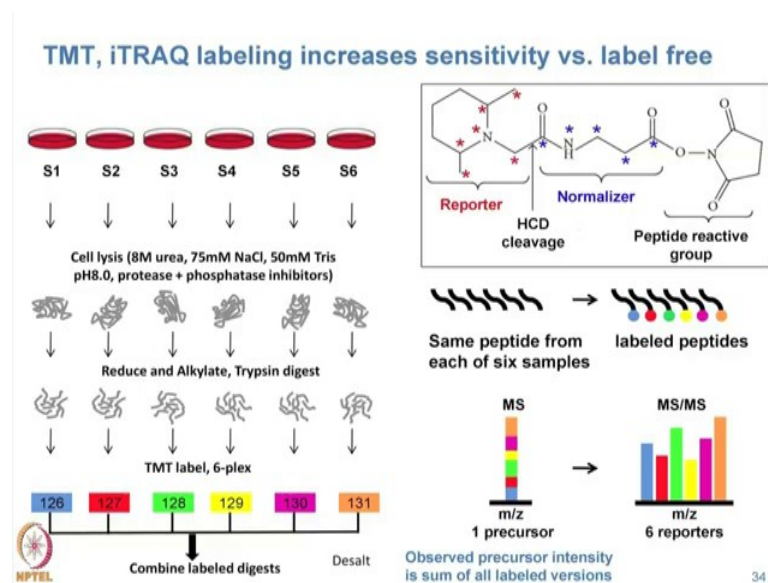
(Refer Slide Time: 14:36)



And the quality of the quantitation is shown here and would be highest over here, ok. Why is it highest over there; ok, ideally, when you are going to mix things together you would like to mix them as early in the process as possible, so there any of the experimental variable variability that happens to all the samples together, ok.

But because of the way you do the experiment, you cannot necessarily mix things until a later stage, ok. So, in the case of chemical labeling you have to mix after you have done digestion but if you do it in cell culture you get it and do the combination way back when they are just after the cells have grown, right; and so, I think I have already said some of the pros and cons about this.

(Refer Slide Time: 15:26)



So, let us move on, all right. Let us go straight to what happens when you do a chemical labeling approach, ok.

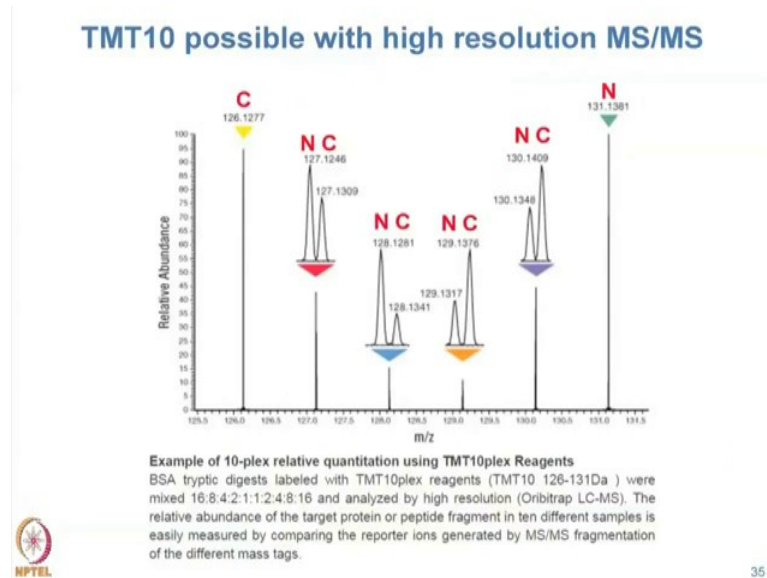
So, the idea here is this you might here I am illustrating TMT 6, ok. So, you would have 6 samples. You lyse each of those samples that gives you proteins, then you reduce and alkylate and trypsin digest the peptides. After you have peptides you use the TMT labeling agent. These are amine chemistry based reagents, so they are going to put a label on the side chain of lysine and on the N terminus of the peptide, ok; and so, the reagent normally comes in 6 colors, that these are actually masses and the mass is shown here are the reporter ion masses that are present in the MS-MS spectrum, ok.

So, then after you do the labeling you mix the samples and then you have 6 different things labeled. The purpose of doing it this way is you the labeling reagent causes all of the samples to have the same mass and the label is going to have a different mass, but only after you do MS-MS, ok.

So, the signal that you see in an MS 1 scan is the sum of all of the 6 samples, which is good, right. It means; you get more signal when you combine the samples, ok and then after your fragment you are going to have the reporter ions that allow you to get the peak height that is shown here is going to enable you to do the quantitation back to the samples that they came from, all right.

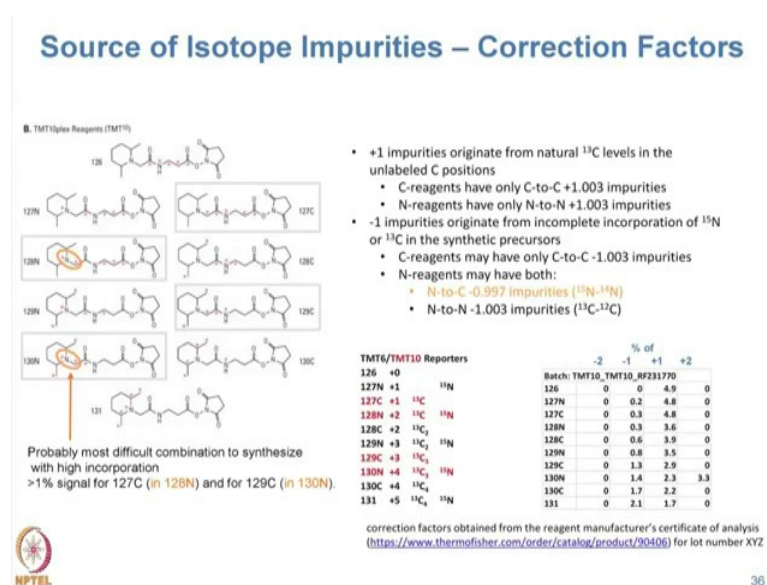
This is the chemical structure of the label all right and shown where the asterisks are is where you would put C 13 or N 15, all right. And in order to do the labeling you are going to put in 5 labels, but depending upon where you put them you can end up with a 126 ion. If you put, I have another slide that will show you where, but the idea is you are going to put them in different places, so that you have different labeling capability, ok.

(Refer Slide Time: 17:43)



Now, because I told you the very beginning you could tell the difference between N 15 and C 13, right. If you have high enough resolution you can separate and you can get 10 different things; ok, and that is all going to go back to whether there was a label on this nitrogen or the C-13 in this position, ok.

(Refer Slide Time: 18:07)



So, now this slide it is harder to see but the dots show us where the labels are for each of the different reagents, ok, all right and then the things here colored in black corresponds to the reagents that are for TMT 6. The additional things in red are the extra channels that you use for TMT 10, ok.

Unfortunately, it is a little bit complicated, ok. So, you have some impurities that were, that you have to deal with in this thing. And there are two types of impurities, ok, one sort of an impurity comes from how pure is the C 13 that you start to put in a label, ok. You can get over 99 percent pure C 13 to incorporate these days, but if there is some level of impurity the same is true of nitrogen 15, ok. But there is a second set of impurity which is this is in the unlabeled positions, ok.

So, this is over here, there is, these carbons over here that are naturally occurring levels of C 13, ok, and so, if you end up with a C 13 in one of these positions it is going to be one carbon higher in mass than it would be, ok.

So, if you when you obtain the reagents they also give you a set of correction factors, ok. That software will apply to correct the intensities to account for the impurities present in the labels, ok. If you obtain data from some public repository and you want to reprocess it all from scratch make sure you get the correction factors that are provided by the people who generated the data.

Unfortunately, they do not always remember to give you the correction factors is when they deposit the data somewhere and you might have to send email asking for them and hopefully someone came right back and give them to you, ok. One of the things that we try to do from our lab is always provide these but I often have to chase down the people who did the experiment and say you need to provide these before we can put their data in a public repository, ok.

(Refer Slide Time: 20:23)

Reporter Ion Intensity Correction Method

Shadforth IP, Dunkley TPJ, Lilley KS, and Bessant C
 BMC Genomics 2005, 6:145 doi:10.1186/1471-2164-6-145

BMC Genomics

Software
i-Tracker: For quantitative proteomics using iTRAQ™
 Ian P Shadforth*, Tom PJ Dunkley*, Kathryn S Lilley* and Conrad Bessant†

Address: *Department of Analytical Science and Informatics, Cardiff University, Cardiff, Wales, United Kingdom; †Oxford Proteomics Centre for Proteomics, Biotechnology Department, Cambridge University, Cambridge, United Kingdom

Email: Ian P Shadforth - ip.shadforth@cardiff.ac.uk; Tom PJ Dunkley - tpdunkley@cam.ac.uk; Kathryn S Lilley - k.s.lilley@proteomics.cam.ac.uk
 Corresponding author: † Conrad Bessant - c.bessant@proteomics.cam.ac.uk

Background
 Each batch of iTRAQ reagents supplied by ABI is labelled with various priority values indicating the percentages of each reporter ion that have masses differing by +2, -1, +1 and -2 Da from the nominal reporter ion mass due to isotopic variation. This information can be used to correct the peak areas calculated for each reporter ion to account for the losses w₊ and gains from, other reporter ions. Losses to ion peaks due to the reporter ion range are also accounted for in this method.

Results
 The simultaneous equations needed to solve this problem are fully completed, but can be formal such that Cramer's rule may be applied. A detailed explanation of the method is available at www.bmc.com/submit/article/145/145.html and the various priority correction values (as percentages) in the order:
 114.1 - 2 Da, 115.1 - 2 Da, 116.1 - 2 Da, 117.1 - 2 Da, 114.1 - 1 Da, etc...

Methods
 w₊ represents the percentage of each peak expected to be present in the mass of the reporter ion associated with that peak. Here, w is for 114.1, x for 115.1 etc.
 $w = [100 - (a + e + i + m)]$
 $x = [100 - (b + f + j + n)]$
 $y = [100 - (c + g + k + o)]$
 $z = [100 - (d + h + l + p)]$

The area (A_j) of each reporter ion peak (j), is calculated above, can now be written in terms of the true areas of peaks (T_j):

$$A_{114.1} = (w^+ T_{114.1}) - (e^+ T_{114.1}) - (i^+ T_{114.1})$$

$$A_{115.1} = (x^+ T_{115.1}) - (f^+ T_{115.1}) - (j^+ T_{115.1}) - (n^+ T_{115.1})$$

$$A_{116.1} = (y^+ T_{116.1}) - (g^+ T_{116.1}) - (k^+ T_{116.1}) - (o^+ T_{116.1})$$

$$A_{117.1} = (z^+ T_{117.1}) - (h^+ T_{117.1}) - (l^+ T_{117.1}) - (p^+ T_{117.1})$$

4 equations and 4 unknown
 → Solve for unknowns using Cramer's rule

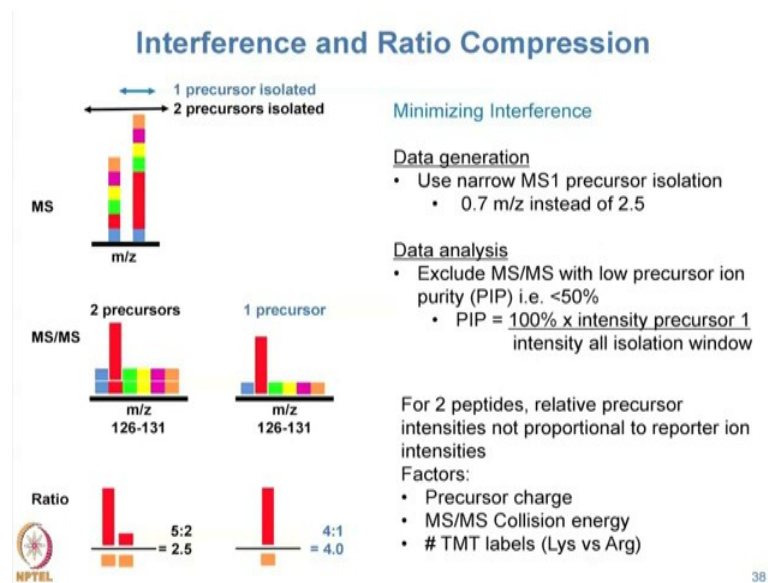
The task is now to calculate each T_j according to these equations.

The determinant of the matrix of coefficients can be found:

$$|K| = \begin{vmatrix} w^+ & 0 & 0 & 0 \\ -e^+ & x^+ & 0 & 0 \\ -i^+ & -f^+ & y^+ & 0 \\ 0 & -j^+ & -k^+ & z^+ \end{vmatrix}$$

Those correction factors are then used with an algorithm to apply them and correct the intensities. This is a publication that is about 10 years old. This is, we use the same method that they describe. We do not we do not use the exact same software because this publication is old enough that it only applied to iTRAQ 4, we have modified it to be able to work with TMT 6 and TMT 10, ok, all right.

(Refer Slide Time: 20:50)



There is another sort of part of complication and working with TMT quantitation and that comes down to interference. There which goes back to if your fragment more than one thing at a time, ok; and so, what I am trying to do is draw a cartoon here to illustrate how this works, ok.

If you had two peptides of very similar precursor mass that were present at the same time and let us say one of them is uninteresting, there is no; no one of the 6 samples that has either up or down regulated levels of protein. But in the sample that is right next to it has up regulated levels of the red, ok. So, there is way more red in this one than there is in the other one.

If you are using an older instrument you might have to set the precursor mass window to be 2 Daltons wide which would cause both of these things to be transmitted at the same time, ok. The labels produce the same reporter ion masses and you cannot tell which one they came from, all right.

So, if you were to transmit this whole thing then you would have a reporter ion set that look like this. Only, when the data came off the instrument it would not have this white line through it that allowed you to tell which one was you would just have the sum of these things, ok.

If you were able to use an instrument that had a narrow precursor window then that what information you would get will be just derived from this one peptide, ok, all right. So, if you put the quantitation together and you combine these things, the ratio that you would calculate if you calculated the red divided by the pink, I am sorry; let us call that orange you would get a ratio of 2.5. If you had only the one together you would get a ratio of 4, ok. So, the ratio of 4 is what you wanted to observe, but it is compressed to 2.5 because of this effect, ok, all right. So, if you, so this is just an example of what might happen, ok.

And so, that there is a couple of things you can do to deal with this, right. The first is you could do a better experiment, right. If you have an instrument that allows you to do better transmission just all of the CPTAC work that is going to be presented later in the week and is already published is all iTRAQ data run on a Q-Exactive instruments that at the time had a window of two daltons per precursor transmission. What you can now routinely do on a lumos instrument or a Q-Exactive HF-X is run a 0.7 m/z tolerance or window width and so, you would be able to in this kind of case transmit only the one thing, ok.

The second thing you could do is you could have data analysis that would go back and look at all your MS scans, and say, if we have got this thing let us throw away that data point, ok; and because we are expecting most of our proteins to be detected by multiple peptides, we have some ways of taking and recognizing that some data points are better than others, and so, we can exclude those, ok.

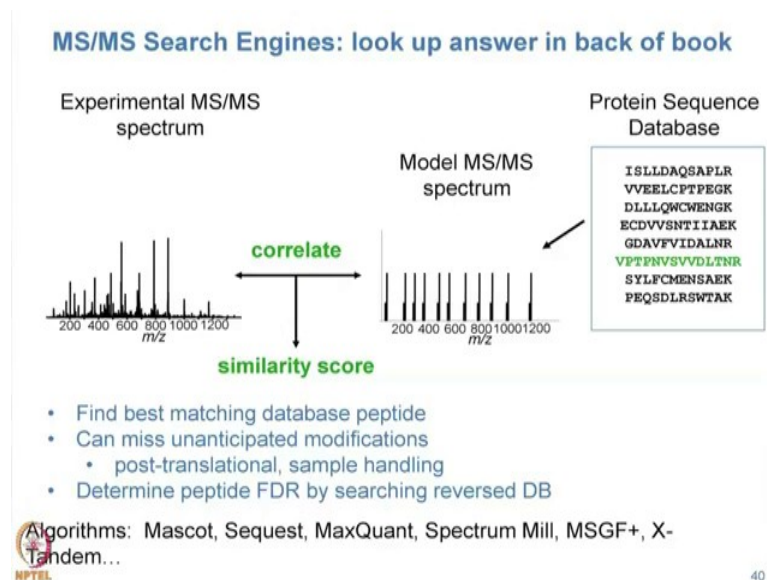
So, a common thing to do it different people do this, but they do not all call it the same thing which is to take some measure of whether how many things are here and what is the relative abundance of those things that are there and when the relative abundance of those things is high then you throw away the data point, ok.

Now, that is an approximation, because although I have shown you in this cartoon example that the ratios of the MS 1 peaks is the same as the relative ratios of the MS or the of the reporter ion that is not always what actually happens, ok. When an individual peptide fragments you are going to get some reporter ion signal and some sequence ion signal but sometimes the balance is like this, sometimes it is like that, ok; and so, even if

this peak right here in the MS 1 scan is taller it does not necessarily mean it is going to contribute more reporter ion signal, right.

So, that is some of the uncertainty that is present in this type of data. And getting better at this is there is room for improving our data analysis, ok, all right, all right. Scoring peptide spectrum matches, all right. So, this slide I already showed you once, it was several equipment failures earlier, all right.

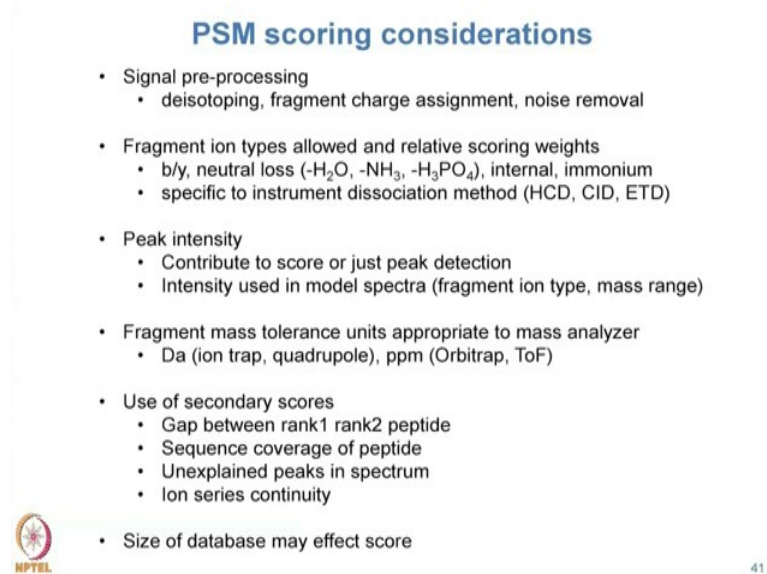
(Refer Slide Time: 25:41)



But the idea is that you are going to take a sequence database and in your experimental spectrum, you have programs that are going to approximate what the spectrum is expected to look like and then score them. These are some examples of some names of programs that do this, ok.


If you are going to design an algorithm to do this; these are the kinds of things that you would have to think about, ok. And when you when you start to just look at other programs, these are some of the things that you could you could think about in terms of evaluating or reading about what they do, right. So, but they are all going to have one way or another have to deal with these kinds of things, ok.

(Refer Slide Time: 26:23)



PSM scoring considerations

- Signal pre-processing
 - deisotoping, fragment charge assignment, noise removal
- Fragment ion types allowed and relative scoring weights
 - b/y, neutral loss (-H₂O, -NH₃, -H₃PO₄), internal, immonium
 - specific to instrument dissociation method (HCD, CID, ETD)
- Peak intensity
 - Contribute to score or just peak detection
 - Intensity used in model spectra (fragment ion type, mass range)
- Fragment mass tolerance units appropriate to mass analyzer
 - Da (ion trap, quadrupole), ppm (Orbitrap, ToF)
- Use of secondary scores
 - Gap between rank1 rank2 peptide
 - Sequence coverage of peptide
 - Unexplained peaks in spectrum
 - Ion series continuity
- Size of database may effect score

 41

So, there is going to have to be some step maybe it is not within the search program itself, might be a program that you can run ahead of time, they will do peak detection, ok. And it is going to do these kinds of things. It is going to do de isotoping. It is it could assign fragment charge and do some sort of signal to noise a processing, so that you are hopefully trying to only use a signal peaks, all right

You have to have when you design the algorithm know what fragment ion types are possible, ok; and when you start to use a program you often have to choose what instrument type it is that you use to generate that spectrum and when you have done that it is going to behind the scenes be consulting a configuration file that is got appropriate things like what ion types are possible for that instrument, and some part and potentially some different scoring values for the different ion types, ok, all right. Then when your algorithm also not only does it have mass information; it has intensity information.

Today search programs generally make not very much use of intensity other than to say present and not present, ok, all right. With some of the machine learning approaches that are starting to be imposed one of the goals of those is to make better use of intensity information, ok, right. You are going to have to choose some fragment tolerance units, ok. I told you, resolution was different across the mass range in certain instruments that is particularly true in orbitraps in time of flight instruments.

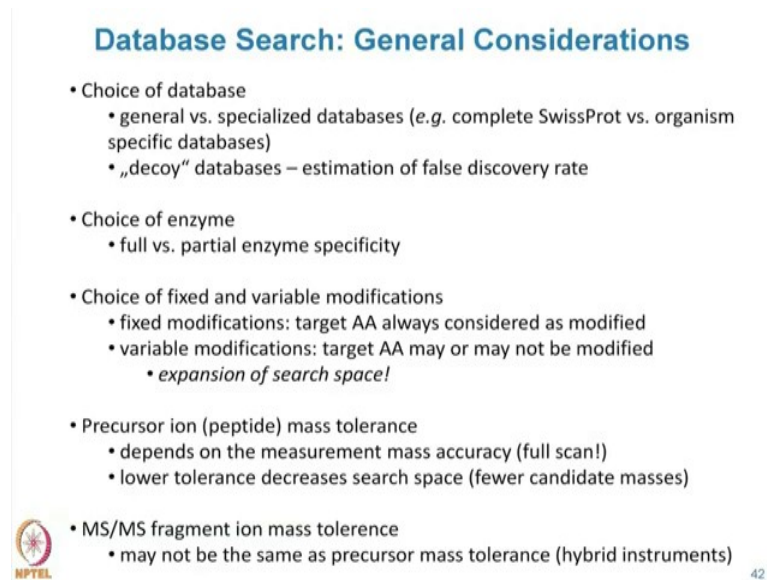
The mass accuracy is also different across the measure, across the mass range, and so, we use different units. If you use a parts per million unit a typical value of a good mass accuracy on a high results may not be plus or minus 5 parts per million, ok; and you would say that across the entire mass range. But when you convert that parts per million into Daltons it means it is a wider, I am sorry wider mass window at high mass and a narrower mass window in Daltons at low mass, ok.

So, if your instrument data has your mass accuracy specifically in units of ppm, ideally you would like to use a search program they could also support mass accuracy in ppm units, ok. But it is actually quite common to use the program where it only has Dalton mass accuracy and so, what you have to do is compromise and set the tolerance to only use the high mass one when you should be able to in principle use it at lower mass and have a narrower tolerance, ok, all right.

Most search engines produce a score that is the primary score that is used to make most decisions, but along the way they might calculate extra things and that might be possible to use in reducing your list to the confidently assigned peptides, ok. Some scoring systems are going to be dependent upon the size of the database, others are going to be only dependent upon the scoring of the ions and a particular sequence, and if you take that sequence and put it in a big database or a little database the scores going to be the same, ok.


Some search engines will however, take the size of the database into account, all right. So; so, that is what you have to do if you are designing an algorithm you consider all those things. If you are going to use one you have to consider these kinds of things, ok. You have to choose a database, ok.

(Refer Slide Time: 30:13)



Database Search: General Considerations

- Choice of database
 - general vs. specialized databases (e.g. complete SwissProt vs. organism specific databases)
 - „decoy“ databases – estimation of false discovery rate
- Choice of enzyme
 - full vs. partial enzyme specificity
- Choice of fixed and variable modifications
 - fixed modifications: target AA always considered as modified
 - variable modifications: target AA may or may not be modified
 - *expansion of search space!*
- Precursor ion (peptide) mass tolerance
 - depends on the measurement mass accuracy (full scan!)
 - lower tolerance decreases search space (fewer candidate masses)
- MS/MS fragment ion mass tolerance
 - may not be the same as precursor mass tolerance (hybrid instruments)

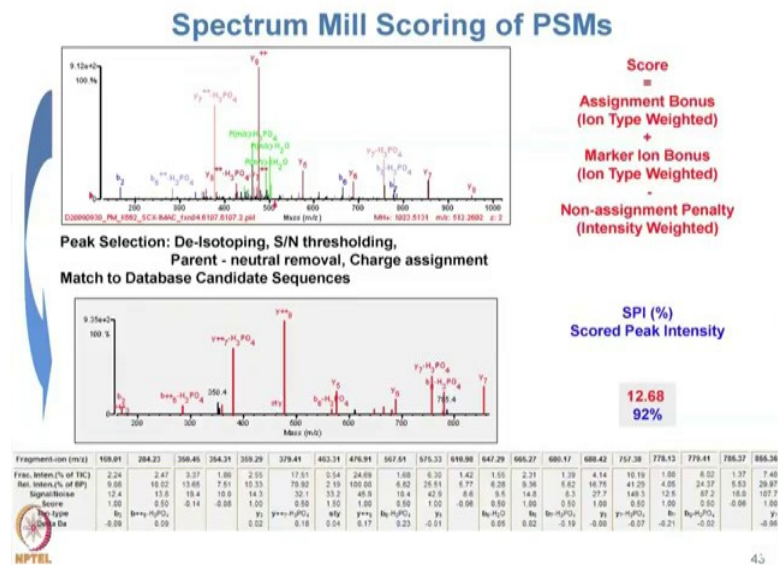
 42

Most of the time what it today we are there is also an opportunity to choose to somehow do a decoy database that is used to calculate false discovery rate, ok. As you read literature you will find that there are certain groups that always you allow for partial enzyme specificity, ok. Well, other people will may require that fully specific. So, the trypsin had to cleave on both ends of the peptide, ok.

If you are using a partial enzyme specificity that increases the search space that the spectra are going to be matched against, the program is going to run slower and you usually have to have higher score thresholds to meet your FDR, ok.

When you are going to choose a fixed and variable modifications you want to choose things that you can expect to find in your sample, and if you are interested in these things that are rare especially if you choose many of them; it is going to slow down the search and I have a slide a little bit we will talk about expansion or search, ok; then you have to choose like I said precursor ion tolerance and fragment ion tolerance, ok, all right. This is; this is how this spectrum is scored in my software called Spectrum Mill right.

(Refer Slide Time: 31:39)



And this up here is shown with the all of the peaks that are present in the spectrum as it is generated from the instrument. The instrument does not have a colored blue, red and green; well it is all black, ok, all right. There is a pre-processing step that does peak detection, that does these three and these several things de-isotoping, signal noise thresholding, removes the parent ion's neutral peaks so. These are the only peaks that are left that are subject to the scoring, ok.

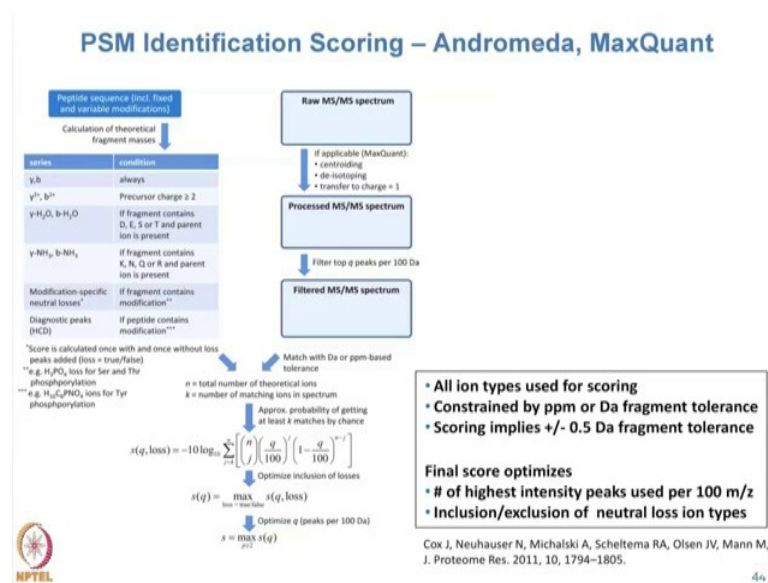
The scoring has 3 components to it. There is a positive component; that means, you the mass matches a fragment ion type that match the score of that is independent of the intensity but it is weighted by what ion type it is, ok. There is a bonus for having composition information like immonium ions and then there is a negative portion of the score that is for peaks that are not assigned, ok; and so, basically a tall peak in a spectrum that is unassigned that is bad, right, that suggests that you have a incorrect interpretation or you have got multiple things that are being fragmented at the same time, ok.

The different ion types have different scores b and y has the highest score they have scored one things that are b minus water, y minus water, a ions those give you less information about the sequence, that because you have already got information from the presence of b and y ion so, that a ions behind b minus water, b minus ammonia they score less, they score a half, ok, all right. So, you do all of those things and you end up

with a score in this particular case the score is 12, the peak detection will produce no more than 25 peaks, maximum score is 25, ok, all right.

Now, something that is quite a bit more different and less intuitive is something like a one of these scores that is use a probability based approach and this is the binomial probability equation.

(Refer Slide Time: 33:49)



It is the basis for scoring in the Andromeda search engine as part of maxquant. This roughly the same approach is used in MASCOT. And the way this works is that all ion types are given the same weight, ok; and in order to calculate the probability, you have to account for the chance of there being a randomly matched peak, ok.

And the way that this is put into the binomial probability essentially comes down to breaking up the mass range into a 100, I mean a 100 Dalton chunks then you say if we are going to look for say 6 peaks then the chance of 6 of randomly matching would be 6 out of a 100, ok. It may not be immediately obvious but that also suggests that the mass tolerance you were allowing was plus or minus a half a Dalton, ok, all right.

Now, in practice MaxQuant has allowed you to specify a fragment ion tolerance, and, but that is not used as part of the scoring, ok. And for up until about 1 or 2 years ago MaxQuant did not allow you to use parts per million as a fragment tolerance you had to use Daltons, and it is because of the way that the scoring is built into the probability, ok.

(Refer Slide Time: 35:22)


True Probability or Just Effective Scores?

Peak selection assumptions

- All regions of spectrum equally likely
 - multiply charged fragments below precursor
 - some 100-300 m/z values not possible, di-peptide AA combinations
 - tolerance in Da, not ppm
- Tall and short peak intensities equally diagnostic

Fragment ion type assumptions

- All ion types equally probable
- Neutral losses possibly ignored, γ -H₃PO₄, γ -H₂O, γ -NH₃



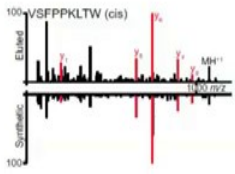
45

So, from my point of view; the probability is not true probability but the scores are still effective, ok; and the reason that the probability is not true or for the reasons that I have listed here and I have already talked through, ok, all right.

Now, what I want to do to show you here is a contrast and this is what you would do if you knew what that peptide fragmentation was going to look like, ok. And you would know what the spectrum is going to look like because you already had a spectrum that you trust and you is used as the reference presumably because you knew you had the peptide maybe you made it synthetically generate the spectrum, the spectrum comes put in a library and then all the experimental spectrum you generate you just match to the library, ok.

(Refer Slide Time: 36:04)

Spectral Library Matching



Match an experimental MS/MS spectrum against a library of previously annotated experimental spectra

spectral dot-product score (SDP_Score)

$$\text{SDP_Score} = \frac{\sum_{\text{peak_hits}} I_Q \times I_L}{\sqrt{\sum_{\text{query_peaks}} I_Q^2} \times \sqrt{\sum_{\text{library_peaks}} I_L^2}}$$

Score range: 0 - 1.0
Perfect match: 1.0
"Reliable" match: > 0.7

- Makes explicit use of intensity
- Requires prior observation/annotation that was trusted
- Instrument characteristics should be very similar (collision energy)
- Chemical labeling should be same
- FDR estimation based on decoy libraries is a "work-in-progress" (less universally practiced than for DB search)

Tools: SpectraST, X!Hunter, Bibliospec

46

The particular case that I am showing actually is one of these things where somebody is trying to demonstrate that the thing that they observed in a complicated experiment, they made the synthetic peptide, the spectrum looks almost the same, you can calculate a spectral similarity metric and it passes the threshold and they can say; see, we this is what we said it was, ok, all right.

So, the equation that that is gets used here is a dot-product score. There are a few different variations on this, and I am not going to go through the math, but the point is that you are really taking advantage of the intensities. And you are not allowing for all possible fragment ion types that could occur to a peptide, you are only allowing for the ones that actually occurred to generate the reference spectrum, ok. Some software programs that do this kind of spectral library searcher are listed right here, ok.

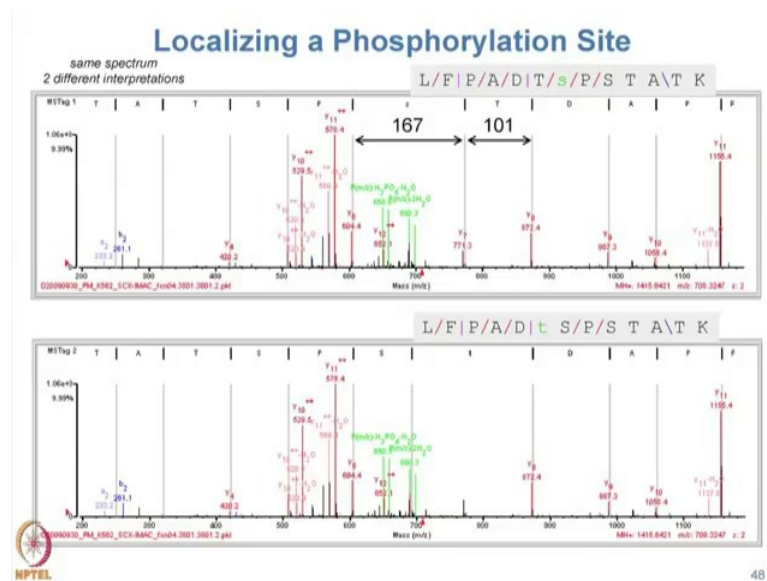
The FDR method that is calculated is sort of today not thought I was being as statistically regular rigorous as what is used for database searching and as I would characterize that as a work in progress to be able to do good false discovery rate calculations, ok. Now, it is also the case that in order to do this effectively, you have to have a good set of reference spectra to match to, ok.

And one of the things we found in our lab is that once we have got just a good reference library somebody came up with a new chemical labeling agent and we switch it over it and they all they fragment all differently and now we have to start over, ok, all right. But

after we have done a lot of work somebody can collect all of our spectra and then use that as the basis for creating a library, ok, all right.

Now, let us talk a little bit about localizing in post translational modification site, all right. So, what I have got here is a MS-MS spectrum of a phosphopeptide, this is not two spectra; this is one spectrum, it is just labeled two different ways, ok.

(Refer Slide Time: 38:20)



The say you can see it is the same peptide sequence that, all right. And the only difference is whether the phosphosite is on this serine or the phosphosite is on this threonine . I want you to raise your hand if you think it is on threonine . Now, I want you to raise your hand if it is on the serine. You have to pick one come on, [Laughter] ok. Serine.

Student: Yes.

Who else wants to go serine? Anybody does not vote, does not get the lunch coupon tomorrow, [Laughter] all right. So, the answer is yes there is a serine, and you should have been able to vote, ok. Because you do not have to know anything to see that when you look at the labeling there is something that is not assigned here and it is assigned here, ok, all right.

So, let us talk about what is assigned and why, ok. So, the fundamental premise is of being able to pick where the thing is you have to have fragmentation between the

possibilities, all right. So, in this case, you have this single ion right here in the spectrum which can be interpreted as the y 7 ion for cleavage right there, where 101 would be the mass of threonine in its unmodified form, 167 is the mass of serine which is 87 plus 80 which is the phosphate, ok, all right.

So, if you were instead to allow 87 for the threonine or for the serine that would stick this in here in a sort of messy part of the spectrum, and then if the gap out here would be shown as this ion to this ion and then that would leave that unexplained, ok, all right. But because those two residues are right next to each other you are not going to get much information to work with in order to make your decision, ok.

So, in cases like this and in all it is going to be often the case that if you have to determine two choices that are right next to each other, you are going to have to make that decision on maybe one or two peaks in the spectrum, ok, all right.

(Refer Slide Time: 40:48)

PTM Site Localization		
Test all Locations, Examine Score Gaps		
	Locations Tested	Conclusion
No possible ambiguity	AV ^s EEQQPALK	AV ^{S(1.0)} EEQQPALK # PO ₄ sites = # S, T, or Y
Single Site	AP ^s LTDLVK *	APS(0.99)LT(0.0)DLVK
	APSL ^t DLVK -	
	sSAGPEGPQLDVPR * SsSAGPEGPQLDVPR * SSsAGPEGPQLDVPR -	
Multiple Sites	VTNDI ^s PE ^s SPGVGR *	VT(0.0)NDIS(0.99)PES(0.50)S(0.50)PGVGR
	VTNDI ^s PE ^s PGVGR *	
	VTNDISPE ^{ss} PGVGR -	
	V ^t NDI ^s PES ^{ss} PGVGR -	
	V ^t NDISPE ^s SPGVGR -	
	V ^t NDISPE ^s PGVGR -	

Let us talk about the range of possibilities now that could happen here, ok. If you have if you are looking for phosphorylation sites the precursor mass is 80 Daltons higher, so, so you know you have got phosphate. And then you look at the sequence candidate. What is going on what is here, ok, all right. If you look at that sequence there is only one serine threonine or tyrosine in it. So, you do not even need to look at the mass spectrum to figure out which one is labeled or which is phosphorylated it is going to be that, all right, ok.

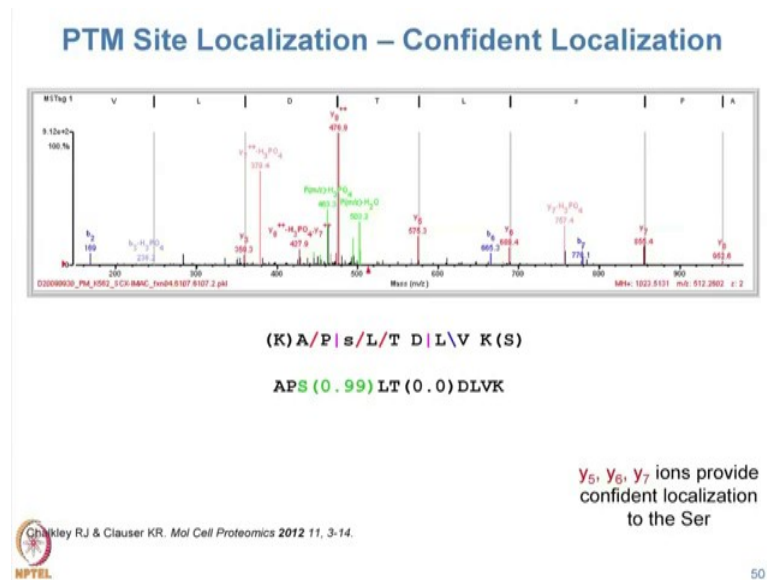
I am having, I am going to switch to the pointer here, ok, It takes a lot of time to figure that back to get in the red spot on that thing. So, I am just going to switch, ok. So, in this case you will have a peptide sequence where you have a serine or threonine and so, you could if you have enough information you could confidently say that the phosphate is on the serine and we would call that a 99 percent chance of being correct, ok.

Let us suppose here you have one phosphate and it could be on any of these 3 serines out here if you have fragmentation between them, you can tell the difference, ok. And I am going to show you the spectrum where there is fragmentation between serines 2 and 3, so we can say it is not on serine 3, but we cannot tell the difference between first and second serine, ok.

When you get multiple phosphosites in the same peptide that gets a bit trickier and this is illustrating all the possible places the combinations that you could put them and then I am going to show you a spectrum that gives you the ability to tell that there has to be a one on this serine not on that threonine but then the second one we cannot tell where it is, ok.

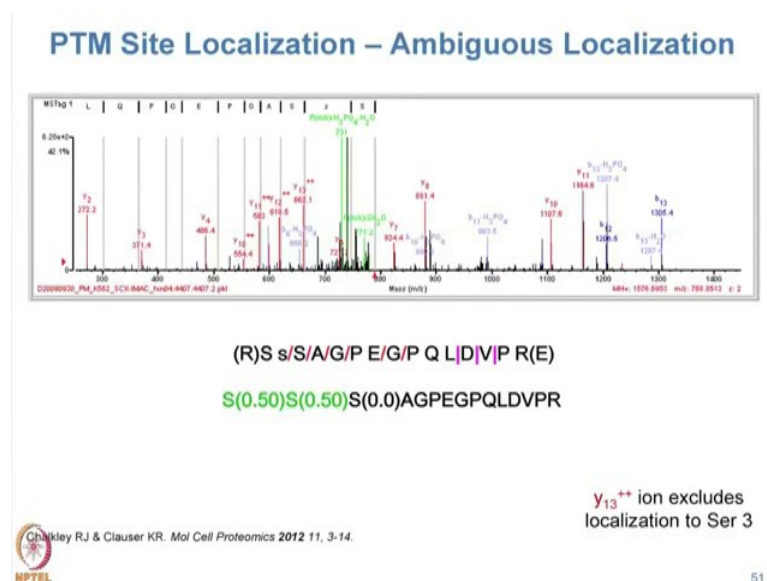
So, this is how complicated this kind of stuff gets and when you are doing proteogenomic work and you want to look at the phosphodata set, and you look at the list you know like, there is all these things in this list that do not have clear assignments of the serine threonine well that is a feature of the data that you got to deal with. One of the ways you might deal with this throw out everything that is not confidently indicated to a particular position, ok.

(Refer Slide Time: 43:01)



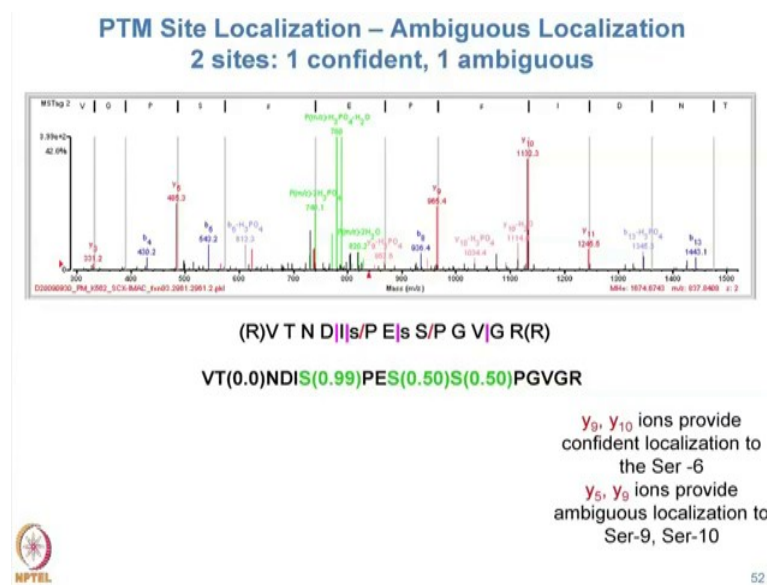
Here are the spectra that give you the cases that I just described, ok. So, here is a spectrum where we can confidently put the phosphate on the serine and these ions in the spectrum y 5, 6 and 7 are separated by the right masses, they should have been labeled, ok. This is going to be a 113 gap, this is going to be a 167 gap and then 97, ok. So, that can place the phosphate on that serine, not on that threonine. This gap over here is going to be 101, ok.

(Refer Slide Time: 43:40)



Here is the example where the y 13 ion, y 13 doubly charged right here allows us to fragment between the second and third serine. So, we know that the third serine now is not phosphorylated, but there is no fragmentation between the first two, ok. So, we cannot tell where that is, all right. Here is the complicated one where there is two phosphorylation sites, the precursor mass is 160, greater than the unmodified version for this sequence.

(Refer Slide Time: 44:10)




We have a fragment ion of y 9 and 10 that gap there is 187 which is going to say that that is phosphoserine and then y 5 and 9, here there is not very good fragmentation between those, and so, we cannot tell where the localization is, ok, all right.

(Refer Slide Time: 44:34)

Key Aspects of Scoring Localizations

- ➔ • Select peaks in spectrum to be used for identification/localization
- Test all sequence/location possibilities
- Assign fragment ion types to peaks
 - Allow for peaks to have different ion type assignments for conflicting localization possibilities
- Use score differences to make decision on localization certainty/ambiguity
 - Decide upon conservative/aggressive thresholds.
- ➔ • Provide a clear representation of the certainty/ambiguity in localization of each site
 - Allow for multiple sites with mix of certainty and ambiguity in localization
 - Distinguish between:
 - Ambiguity – no distinguishing evidence, i.e. either possibility
 - Ambiguity – conflicting evidence, multiple co-eluting isoforms present



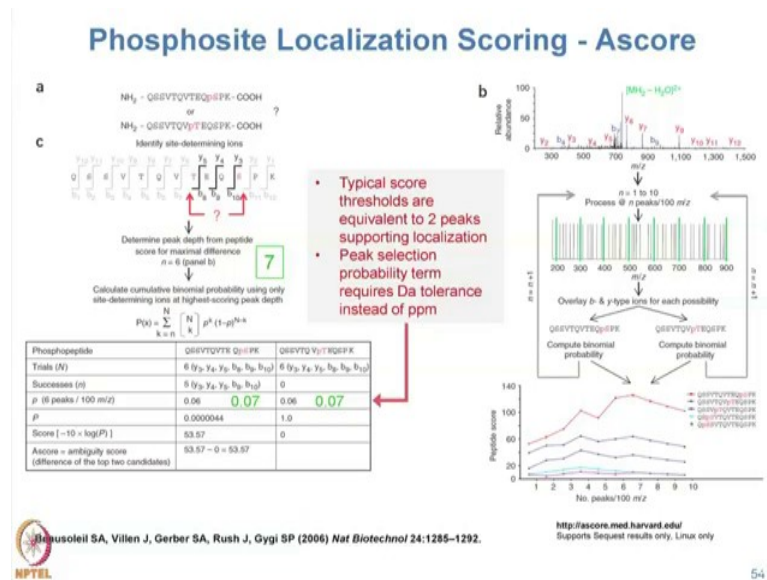
53

So, when you are going to write some, so I tried to show you graphically this is when you look at the spectrum. Can you have the information? If you are going to write a program to do this, all right, these are some of the things that you got to put into the design of your experiment.

You are going to think about all of these things, ok. I think the most important of those things are shown here the choice about how you decide what peaks are going to be used to make your decision and then how do you clearly represent the certainty and ambiguity in the localization decisions that the program is made, ok.

There will be different choices made by people that who write the programs about how to deal with the rest of these issues, ok. And then today there is not a universally applied way of determining a false localization rate from these scoring things, whereas, the target decoy calculation for identification is practiced throughout the field, ok, all right.

(Refer Slide Time: 45:42)



This is a one of the first automated scoring approaches and it is again using this binomial probability theorem but instead of using the calculation based on all the possible fragmentation of the peptide, it is limited to just the fragmentation between the sites that you are trying to distinguish which have the localization, ok. But otherwise it uses the same framework the same mass accuracy assumptions and when you get down to the what your score threshold you are going to use, it comes down to essentially saying that we are going there has to be two good peaks that meet the scoring threshold, ok.

At the time that this was published the authors used a particular score threshold, I forget exactly what the value was and then like a year or two later they decided they could say they have more identifications if they made the threshold lower, ok. And it was essentially by saying instead of two peaks you would allow one peak to make the decision, ok, all right. But you have this nice descriptive way of using a mathematical calculation, ok.


(Refer Slide Time: 46:58)

Spectrum Mill Variable Modification Localization Score

VML score = Difference in Score of same identified sequences with different variable modification localizations

VML score > 1.1 indicates confident localization

Why a threshold value of 1.1?
1 implies that there is a distinguishing ion of b or y ion type
0.1 means that when unassigned, the peak is 10% the intensity of the base peak



55


When I wrote the calculation, I have tried to think of it more intuitively and I calculated the score difference in the identification scores, given various possible places and the decisions I made were on the quality of the information that will that gave those score distinctions, ok. I said that I want the ion type that you allow to make the decision has to be one of the highest information ion types, it is got to be a b or y ion.

You are not going to make the decision based on one ion that is a b minus water ion. You are also not going to make the decision based on tiny little peak that could be mistaken for noise, ok. So, what I sought to do is say that it is going to be a b or y ion and the relative intense that it has to be at least 10 percent of the base peak, so it is a solid peak, it is not noise. That works out to giving you a score threshold that is 1.1.

(Refer Slide Time: 47:54)

Points to Ponder

- Depletion of plasma sample during preparation for MS can result in better coverage of low abundant proteins.
- Enrichment of sample helps in better identification of low abundant proteoforms.
- In a label based MS experiment, reporter ion signal at the MS2 level may not always be similar to the sequence ion signal seen at MS1 level.



NPTEL IIT Bombay

In conclusion, we hope that this lecture and the series of 5 lectures so far has helped you to appreciate the importance of sample preparation for mass spectrometry based identification of peptides, the need for enrichment of post translationally modified forms of peptide prior to MS analysis.

Direction has also provided you the glimpse of how impurities in this sample can lead to the errors during the identification of peptides and additionally you were introduced to the concepts of PSMs and how a specific software like a Spectrum Mill uses PSM to score the hits. Lastly, you were explained the concepts of phosphosite localization and a scoring using suitable examples.

In the next lecture, Dr. Karl Clauser will conduct hands on sessions to help you interpret the MS-MS spectrum manually.

Thank you.