1. An Introduction to Proteogenomics

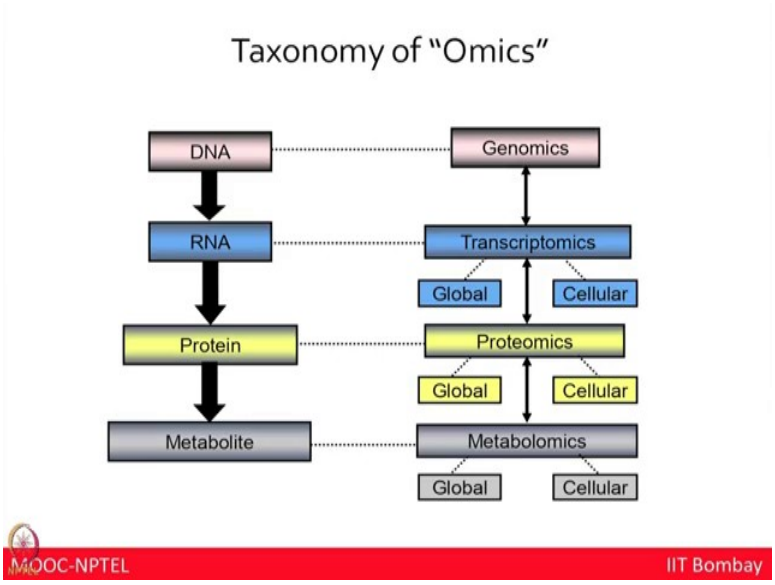
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 11
Introduction to Proteomics

Welcome to MOOC course on Introduction to Proteogenomics. In the first module, you were introduced about latest developments in the area of genomics. Our eventual goal of this course is to introduce you the concepts of proteogenomics, but you would like to develop the concepts of genomics first, proteomics and then try to integrate both genomics and proteomics. So, this module, we are going to talk to you about proteomic technologies, some basics, more advancement in the areas of mass spectrometry, and how one could analyze the data using various latest tools available.

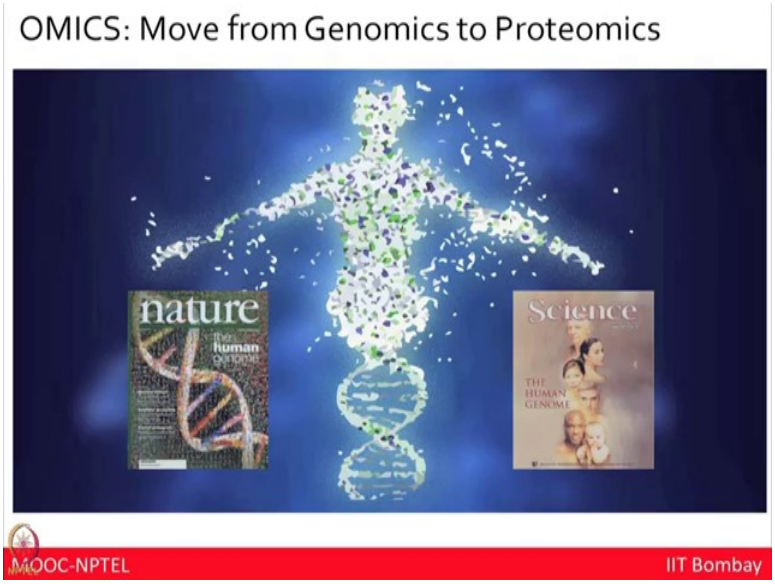
So, let us first start about some of the basic concepts of proteomics especially an overview of proteomic technologies from discovery to function. So, let us first look at the taxonomy of OMICS. I am sure you have been listening variety of terminologies linked to OMICS field. OMICS field aims to look at a given system in its totality. For example, all the genes or DNA of a given system can be studied under genomics, all the RNA or transcripts under transcriptomics, all the proteins under proteomics, all metabolites as metabolomics.

(Refer Slide Time: 01:41)



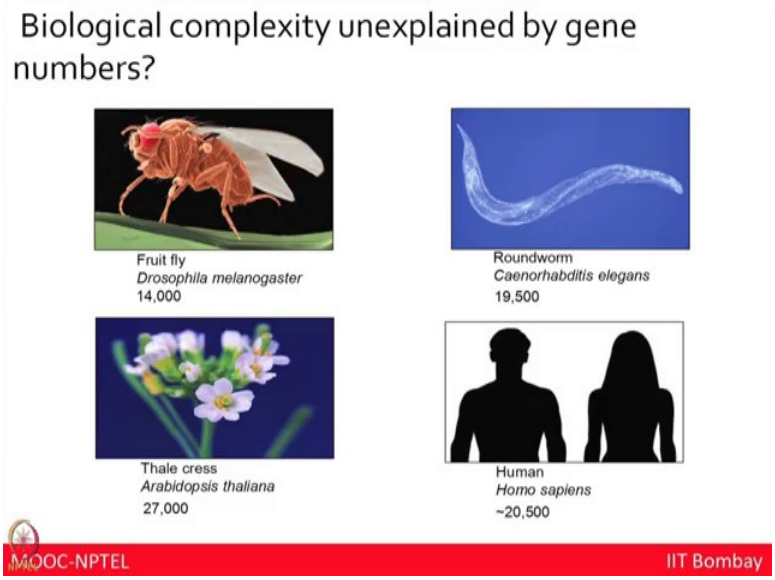
Let us say if the context was to look at all the possible transcripts genes or proteins, then you say that we are going to study the global proteome or global genome analysis. But if the concept is linked to a specific cell or an organ, then that cellular or organelle proteomics. So, in this way, I hope that you know now you are clear about layers of information and different type of OMICS terminologies. All the started from the success of the human genome projects and the success of genomic technologies, especially in the 1990s, that time various genome projects were in progress.

(Refer Slide Time: 02:25)



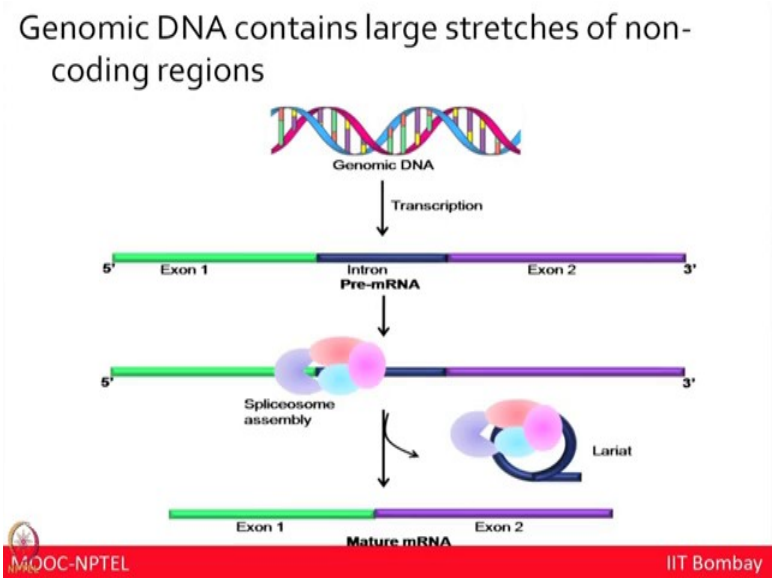
And especially 2001 and 2003 during that time the draft human genome maps were published, which was one of the major accomplishments of first time knowing about all the possible human genes. While this was you know one of the major breakthroughs which will happened in front of many of us, a major milestones in biology.

(Refer Slide Time: 02:55)



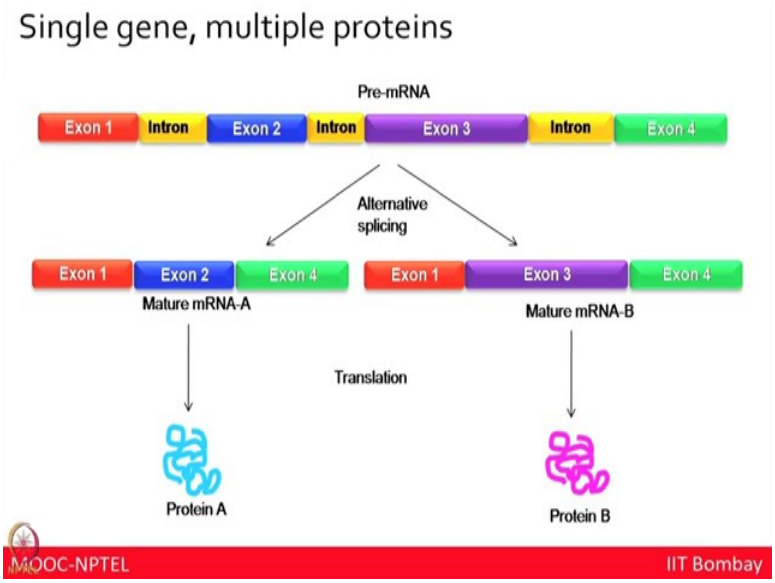
But when people looked at variety of model organisms especially fruit fly which is *Drosophila melanogaster*, Roundworm, *C elegans*, Thale cress *Arabidopsis thaliana*, human *Homo sapiens*, variety of model organism showed the number of genes are not very different in different systems. The numbers may vary around 17,000 to 20,000. So, what makes different systems, show different morphologically and physiologically? So, eventually the focus shifted from the genome moving onto the protein level investigation or proteome.

(Refer Slide Time: 03:31)



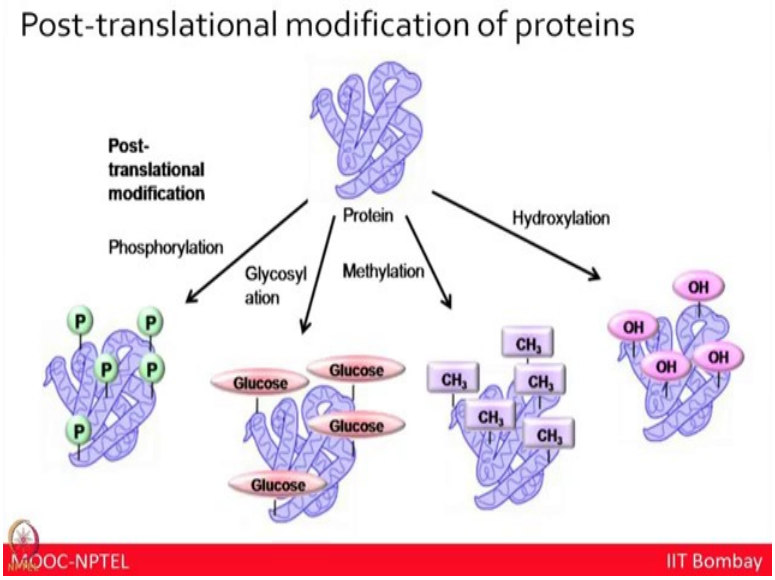
Let us look at some of the basic information basic biology. From the genomic DNA, the after transcription, the RNA's are formed. And in the process of it splicing, you can see that you know the exon and the introns are there. And when introns are out, the functional mature mRNA being formed with combination of various exons.

(Refer Slide Time: 03:51)



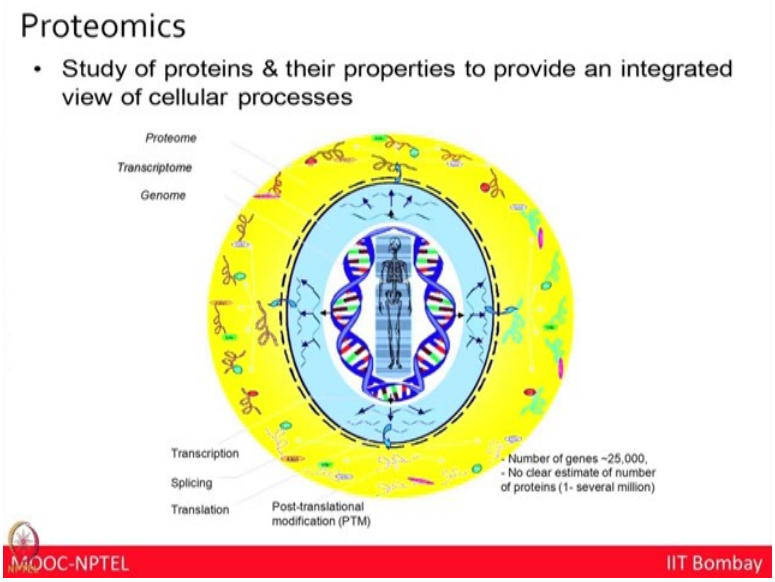
Now, during this process from pre mRNA to mRNA or the mature m RNA formation, we know variety of combinations can happen different exons can come in two different combinations to give rise to different type of mature mRNA, which may eventually result in to different protein forms. From same gene different transcripts and different proteins could be produced.

(Refer Slide Time: 04:25)



And further now this information could be much more complicated because each protein may also gets modified based on the attachment of phosphate residue, methyl residue, glucose at the sugar molecules, hydroxyl residues, many of these modifications known as post-translational modifications are very relevant for the functional information of the protein.

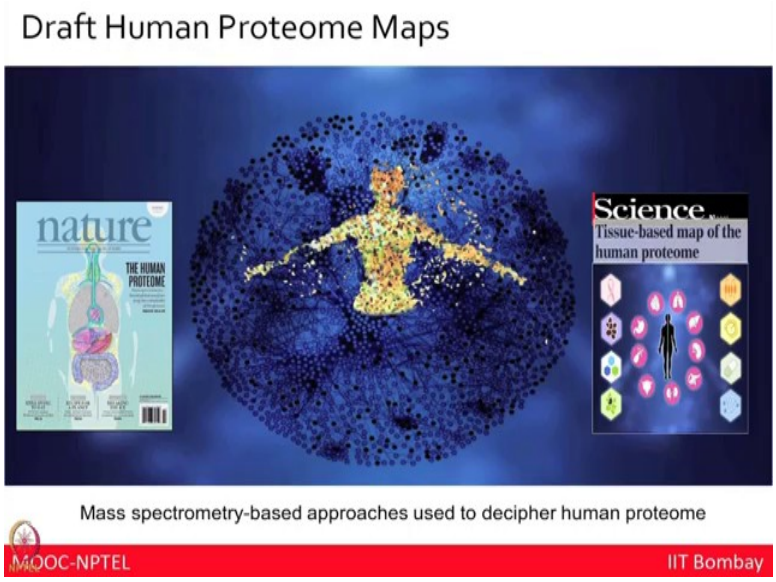
(Refer Slide Time: 04:45)



So, looking at this information, now let us look at this cartoon where we can see the inner sphere that is the genome. Then the transcripts are being formed; the blue circle

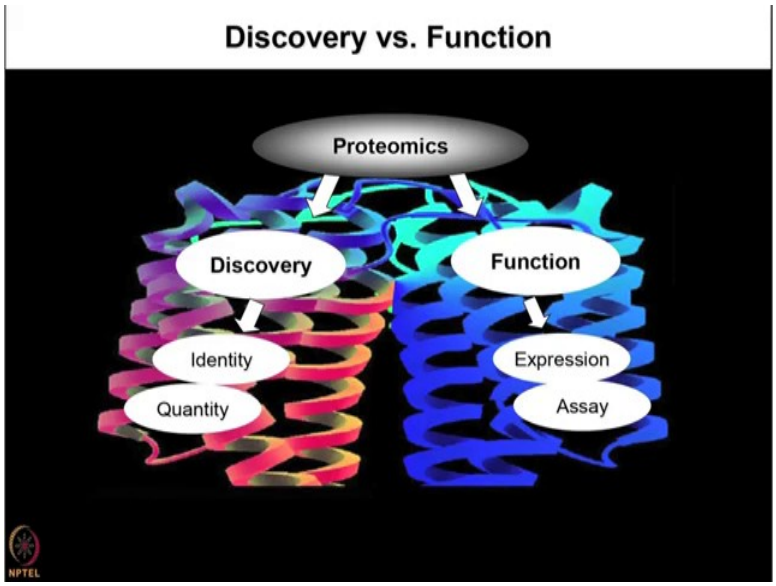
shows the transcriptome, and then eventually the proteins are formed and gets modified that is the proteome. The study of proteins and their properties to provide a broad integrated view of cellular processes is known as proteome, and the field which aims to study that is known as proteomics. I would like to highlight that after genome reference maps which are published, the eventual goal of the entire community was to look at possibility of all the proteins and trying to obtain evidence that these proteins do exist in human with various type of experimental approaches like mass spectrometry and tissue based microarrays.

(Refer Slide Time: 05:29)



All those efforts took a lot of time, but eventually 2014 and 15, the first draft human proteome map was published, which tried to capture and provide evidence that proteins do exist for all the genes which were already decoded. And then you know almost 17,500 proteins where you know the evidence for those were provided. In two different seminal papers, you know based on the mass spectrometry in 2014, and then eventually based on the tissue based map of the proteome in 2015. So, these are again the major milestone accomplishment. After the draft human genome maps, we now have the draft human proteome maps available.

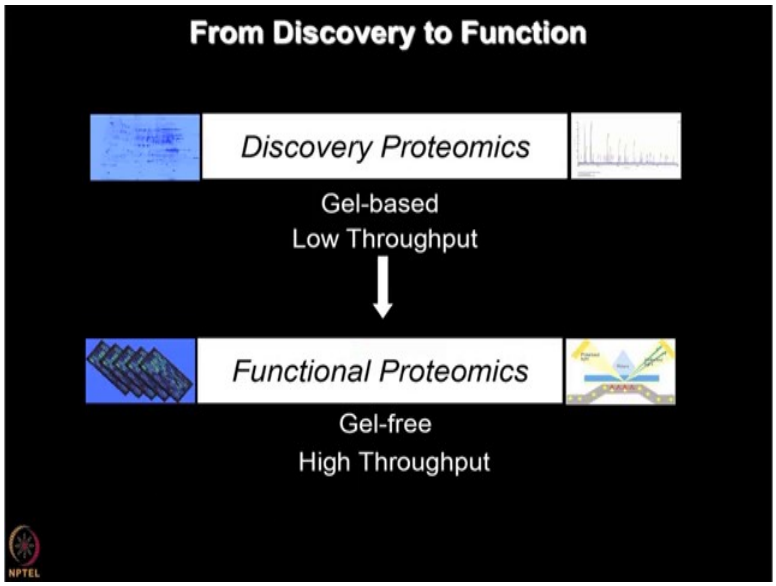
(Refer Slide Time: 06:31)



So, broadly if we want to study the proteomics or the field of proteomics, you can think about there are two major objectives; one is to look at the new discoveries or new kind of targets in context of a given disease or given stress or given comparison which you want to make, or you want to understand the function of the proteins. So, one could term as a discovery or abundance-based proteomics, another can be known as function-based proteomics.

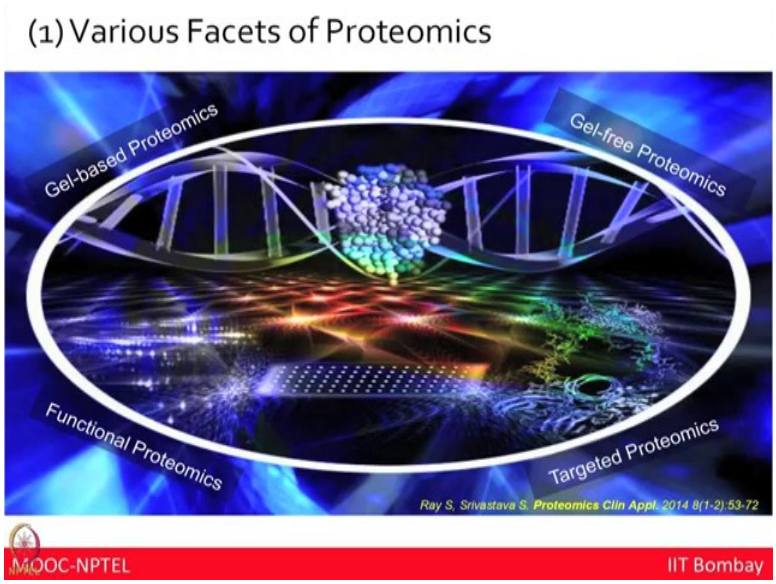
So, discovery based proteomics aims to look at the new protein molecules which could be showing change in the abundance because of the disease or a stress condition or drug treatment, and what are their identification and what is their quantification. So, these information could be obtained under discovery or abundance-based proteomics. Then one could look at function-based proteomics, where aim is to identify their function of unknown proteins or the new proteins which you have discovered from the discovery set. So, where the proteins are expressed the assays to try understand the function of these proteins are the major goal of functional proteomics.

(Refer Slide Time: 07:35)



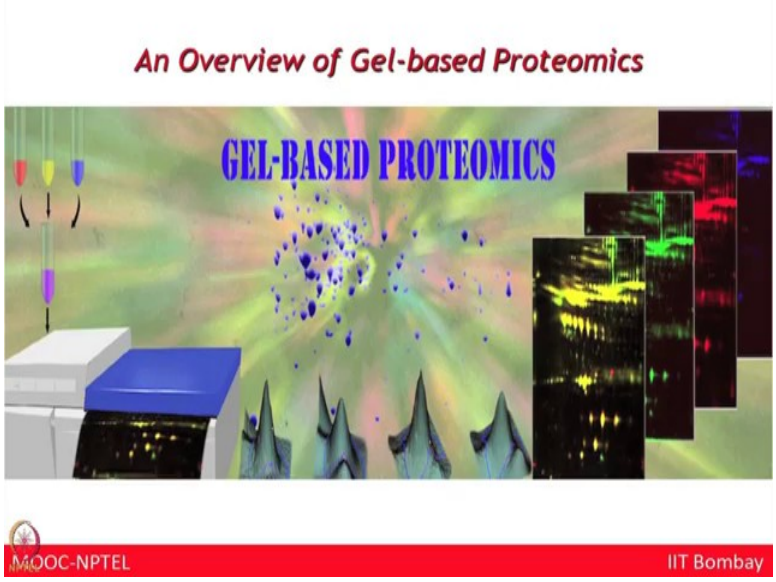
So, if you look at the technology which could be used for discovery proteomics essentially various type of gel-based technology like 2-DE gels and DIGE. And mass spectrometry where quite handy to use the discovery based workflows. And then how to look at the functional information was based on the protein microarrays and SPR - Surface Plasmon Resonance based technologies which provide more information at the functional level. So, much more gel-free kind of approaches, more high throughput approaches were used for the functional studies.

(Refer Slide Time: 08:09)



So, one could say that you know there are various facets of proteomics, starting from the beginning of the proteomics field which is started with gel-based proteomics or the gel-free proteomics which is predominantly build on the mass spectrometry, and various label free biosensors or functional proteomics where you want to study the function of the given protein or targeted proteomics where you want to target a specific protein a peptide sequence, and do their validation in the high throughput manner. I will give a very brief overview of you know each of these domains. Of course, there are separate courses which I have offered in the past where you can get much more detail about each one of these technologies than in the full modules on those.

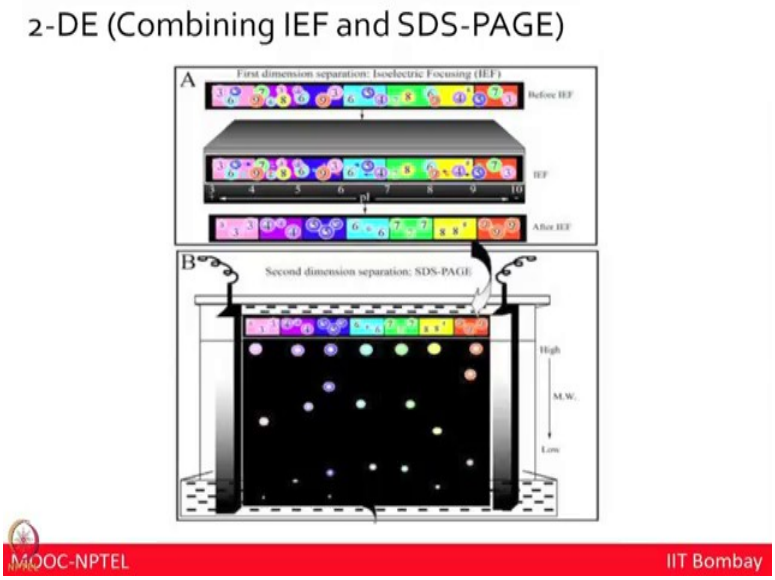
(Refer Slide Time: 08:39)



But let kind of capture very briefly then you know various technology the available for studying proteome. So, in this slide, gel-based proteomics comes very first, because all of you have studied at least some level of gels SDS-PAGE gel or even two-dimensional gel electrophoresis or the advanced form of gels which is known as DIGE or difference in gel electrophoresis.

So, the study of you know looking at the proteome or separating proteins on the gel have been from very long time electrophoresis field has in size from 1950s. And especially the 2-DE gel came into the light in 1975 when Klose and Farrell, they first time showed that one could separate proteins based on their two properties of molecular weight and isoelectric point.

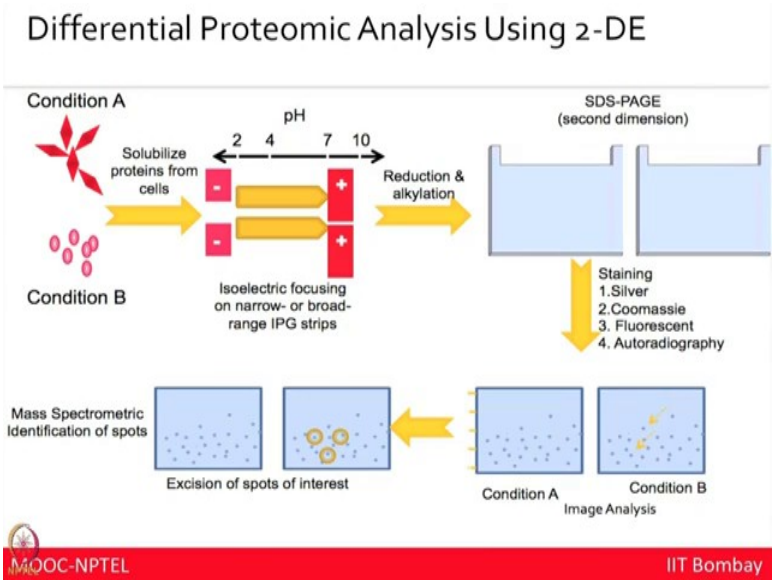
(Refer Slide Time: 09:27)



So, this is the you know a cartoon diagram showing you the concept of two-dimensional electrophoresis, but the first dimension protein separation can be done on the IPG strip or immobilized, pH gradient a strip in the isoelectric focusing field based on their PI values. When pH equals PI, there is no net charge, then protein stops migrating in the electric field, and this is how you can separate the proteins based on their PI values based on either 3 to 10 IPG strip or 4 to 7 IPG strip, you can choose as per the relevance.

Once you have done the first dimension separation then put the same strip on the SDS-PAGE gel unit, and now you can separate proteins on the second dimension based on their molecular weight. So, then eventually you generate an image which is shown in this particular spots, where you can obtain information for each protein at their molecular weight as well as isoelectric point level.

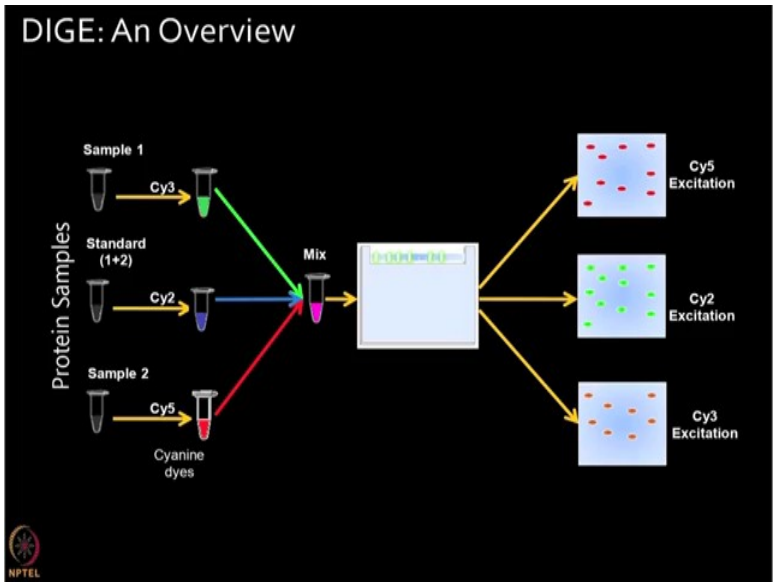
(Refer Slide Time: 10:25)



Now, let us say that your goal was to do the differential proteomics or abundance-based proteomics to look at the new targets in context of a given you know condition A and condition B comparison. So, you have done the protein extraction solubilization of the proteins, and then separated them based on their isoelectric point in the first dimension. After reduction alkylation, proteins are separated on the SDS-PAGE-gel in the second dimension. You can now separate those 2-DE gels you can now stain those for visualization purpose using either coomassie stain or silver stain. Now, these two gel patterns that emerge from condition A and condition B.

After analysis we can see that you know some spots are differentially expressed. And if those are reproducible, you can excise or cut the spots, and then do the protein identification using mass spectrometry. While this is all you know very straightforward, very nice technology, but because of the reproducibility issue, because of the running artifacts, and because of the you know the low throughput the number of protein coverage issues, a new technology you know was much more easy to accept that was DIGE technology which is difference in gel electrophoresis especially for the quantitative proteomics.

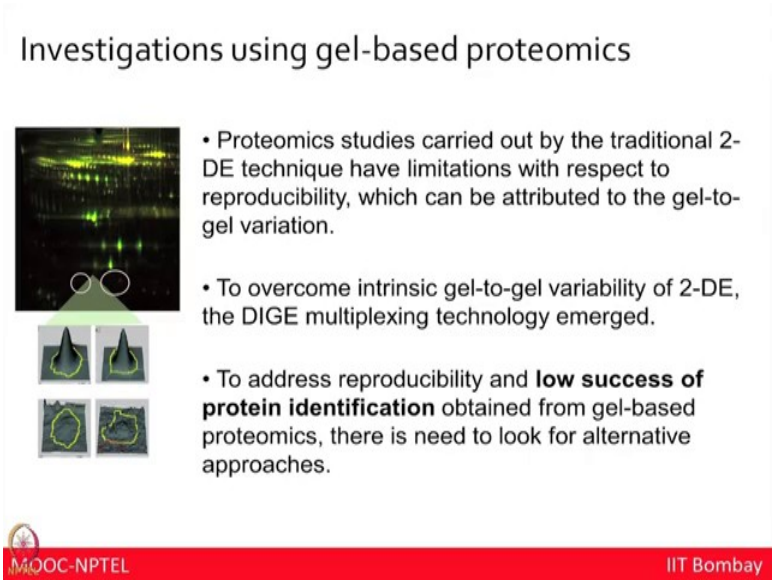
(Refer Slide Time: 11:43)



So, in this particular technology, let us say you have two samples for condition A and condition B which you want to compare, you can now label them from you know Cy 3 and Cy 5 dyes; these are fluorescent dyes from the cyanine. And then in addition to these two dyes you are also adding making a mixture of the both the samples from sample 1 and you know sample 2. As a reference pool and then use that reference pool to label with Cy 2 dyes mix all the three samples together, then on one gel and then you are going to obtain the three images after scanning for Cy 2, Cy 3 and Cy 5.

So, in this light, now you got control and treatment Cy 3 and Cy 5 images, and Cy 2 as the reference pool. Let us say if you have to run three biological replicates in all the three biological replicate, Cy 2 image will remain the same, because you have already made the pool of that and the control and treatment images will change. So, this became much more reproducible much more quantitative.

(Refer Slide Time: 12:43)



And if we now want to summarize, I will have just shown your data from the lab where now you can see after comparison these are the 3D views of the spots of from the DIGE gel bottom panel is for the control spots and the top l for the DIGE spots. So, looks like something is very interesting three dimensional view like showing you know higher expression of this protein in a given disease context.

So, if you want to summarize the overall what we learned from the gel-based proteomics that gel-based proteomics are you know very handy, very simple easy root technology, but because of the reproducibility issue the limitations of gel to gel variability, the focus shifted more towards the quantitative proteomics based on the DIGE. DIGE is the multiplexing technology which offers much more sensitive and much more reproducible way of doing gel-based proteomics.

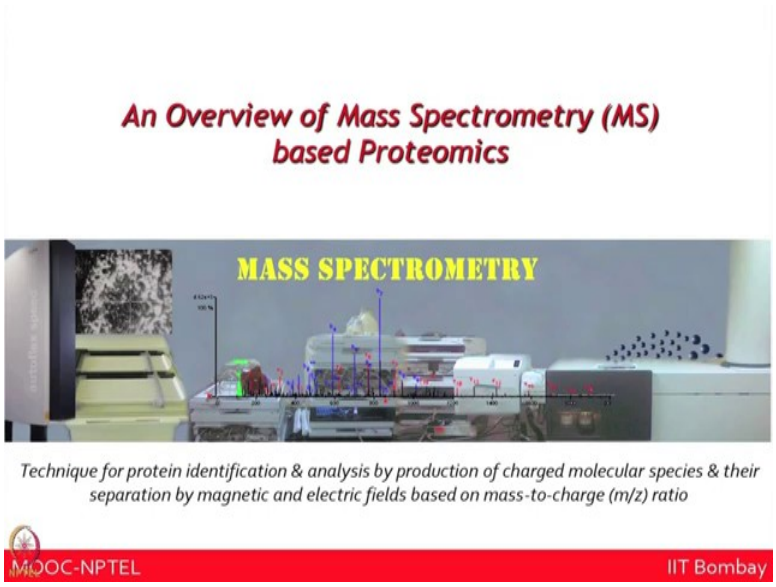
However, the major challenge of both the 2-DE gel or DIGE gel remains that after you know looking at the significant protein spots, you want to identify them using mass spectrometer. But from the gel when you excite a given spot the protein amount present in that a spot may be too little, too low for you to you know reproducibly capture that information and then get any prepared identification.

So, even now you see this particular 3D view spot looks pretty nice and you know biologically very relevant, but what this protein is, we have no idea no information. So, can we now you know identify the protein using mass spectrometer, and if we can do

that then only we will have any sort of biological information. But if mass spectrometer fails to identify this information, because there is not sufficient peptides available, then I think our all effort have failed.

So, therefore, gel-based proteomics is you know great it can do many things, but it is still the success of identifying the protein of relevance could be only 50 or 60 percent because you cannot identify all the protein spots. So, a lot of focus shifted in the overall proteomics field from the gels and moving to the gel-free proteomics especially based on the mass spectrometry. So, let us now shift to the next module especially looking at mass spectrometry based proteomics.

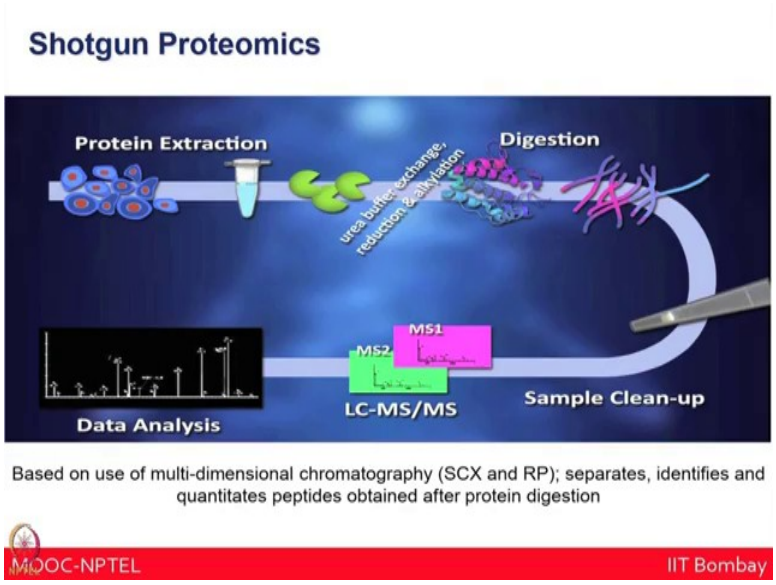
(Refer Slide Time: 14:51)



I must say this particular technology has almost you know influenced very positively the entire field of proteomics and even when metabolomics. And as a result, you will see a lot of developments happening in this area. I will probably give a separate lecture in my next lecture just focused on the MS based proteomics. But let us kind of you know very get very brief overview of mass spectrometry based proteomics, the mass spectrometers, they are aiming to provide you information based on the master charge ratios of these particular you know you are peptides to give you identification and quantification. And to achieve that you are trying to separate these ions in the electric and magnetic field.

So, diffusion is mentioned here that is a technique for protein identification, and analysis by the production of charged molecular species. And their separation by magnetic and electric field based on their mass and charge ratio.

(Refer Slide Time: 15:51)



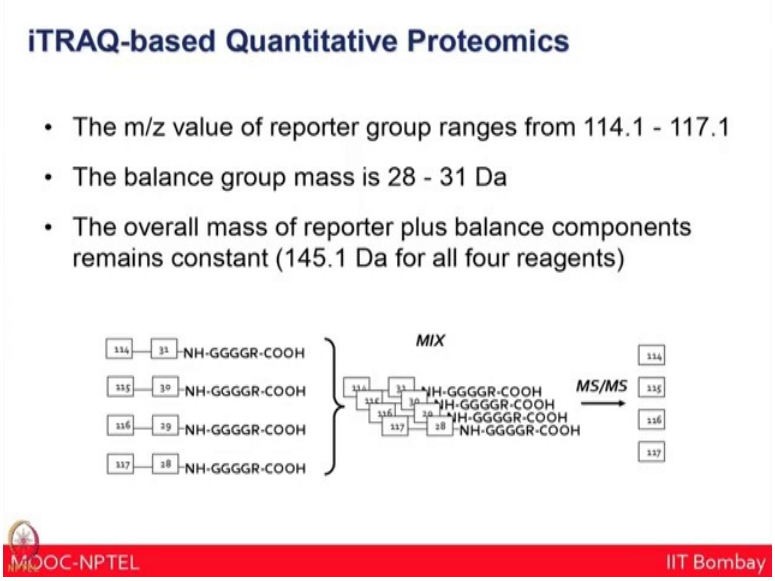
This image is shown here is for shotgun proteomics. You do not want to use the gels directly solubilize the protein do the know reduction alkylation followed by treat with enzyme trypsin for digestion. Now, all the proteins will convert to the peptide form. These peptides can now be you know cleaned up for the salt and other debris. You remove those artifacts and now the clean peptides could be directly subjected and injected to the inside the mass spectrometer for doing MS analysis or MS MS analysis.

In this area, you will also use the chromatography like a specially strong cation exchange or reverse phase for peptide separation. And then you will use ionization source like electrospray ionization or MALDI to ionize is these peptides into the gaseous ionized forms, then separate them based on their various m by z properties in the mass analyzers. The different type of mass analyzers available could be you know quadrupole, time of flight, ion trap, orbitrap, many type of configurations are available and then one could use MS/MS to generate the peptide spectrum. And then use databases for peptide search and eventually you will get the hit for the protein identification.

In just nutshell of course, you know we will touch upon some of this concept in much more detail in the next lecture. But I just trying to give you the overall you know field of

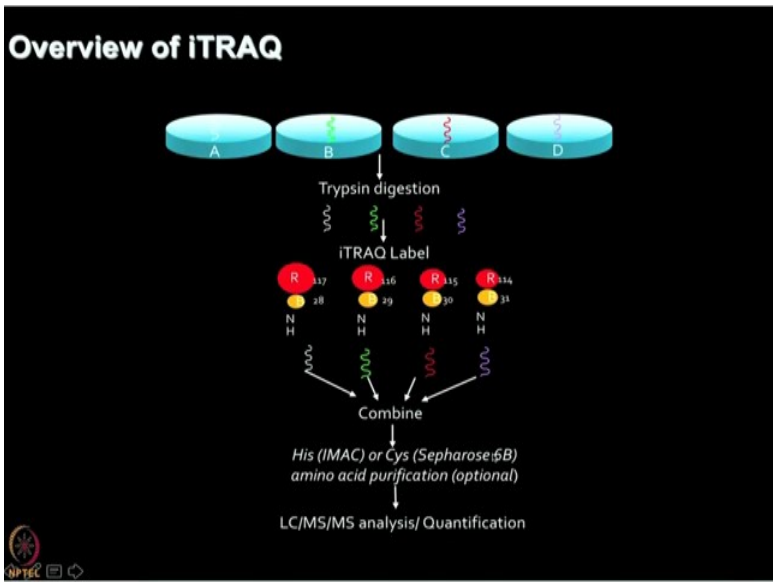
proteomics especially gel-based we covered and now I am trying to give you the feel for the mass spec based proteomics specially shotgun proteomics. So, if you want to do quantitative proteomics using mass spectrometry the way we talk for gel-based where we have used then you know concept of DIGE technology. Here one of the successful example is iTRAQ based quantity proteomics, iTRAQ in isobaric tagging for the relative and absolute quantification.

(Refer Slide Time: 17:39)



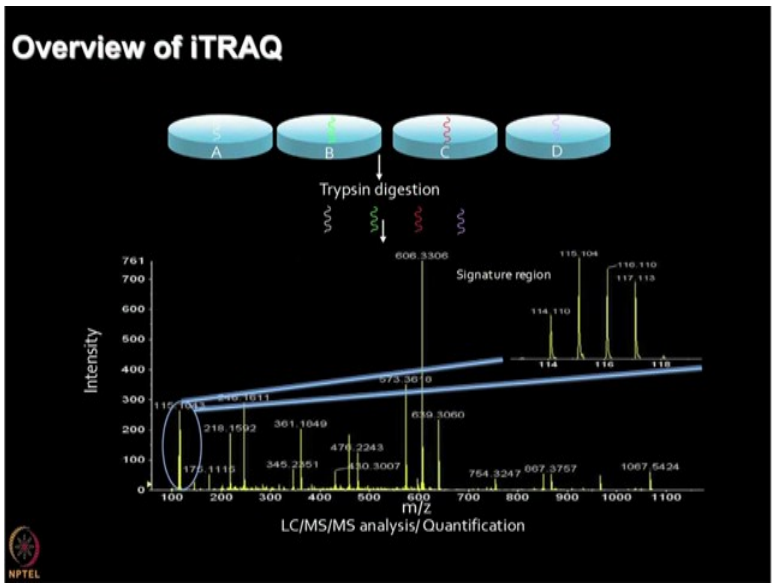
And here again you can take now 4 samples or 8 samples or even you know a similar technology like TMT or Tandem Mass Tag have come forward where you can study even 10-plex studies. These are isobaric tag. So, it means you are not changing over addition of the mass on any of the 4 condition or 8 condition. So, in this way you know, we have this iTRAQ label 114, 115, 116, 117, they put the balancer region is 31, 30, 29, 28. Overall mass added is 145, which is all same in all the four conditions. You have labeled the peptide of four different conditions mix them all together and do MS MS analysis.

(Refer Slide Time: 18:21)



This cartoon shows you that you have four conditions. You have done protein extraction followed by digestion from those, label the samples using iTRAQ labels combine them, do the you know some sort of chromatographic purification if required, and then followed by LC MS analysis.

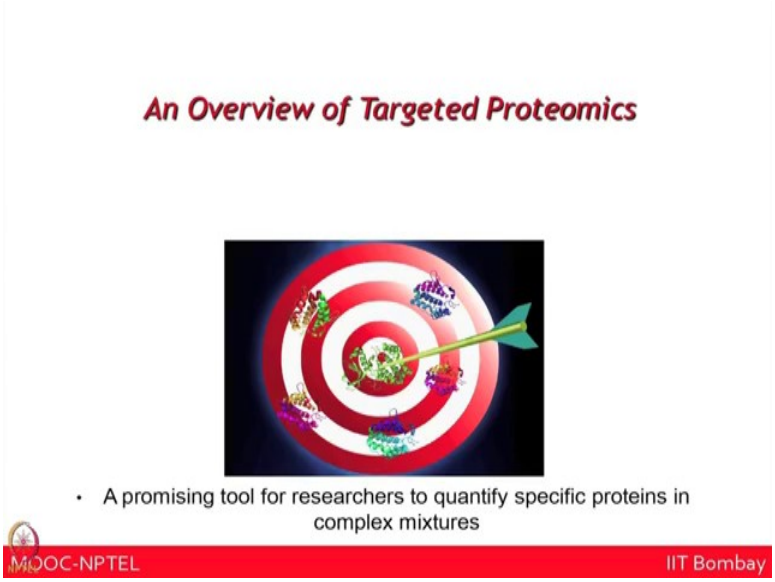
(Refer Slide Time: 18:39)



At the MS/MS level you can see these reporter ions could be visualized 114, 115, 116, 117, and this information could be used for the quantification of peptides. So, this is the

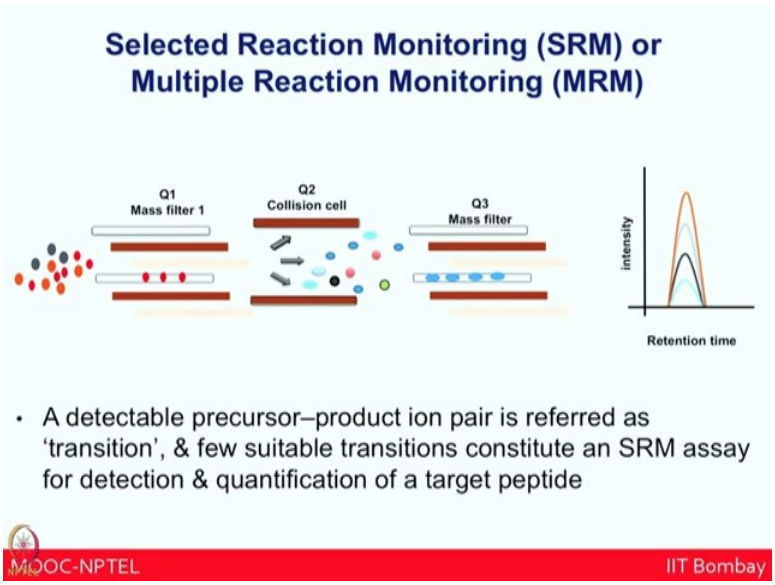
in a very nutshell a very brief about using mass spectrometer directly for the connotation. You are not using any gels in this approach.

(Refer Slide Time: 19:05)



Next technology which is coming forward is targeted proteomics, which is also based on the mass spectrometry revolution where aim is to look at peptides directly using triple quadrupole based mass spectrometers, and do their validation.

(Refer Slide Time: 19:15)



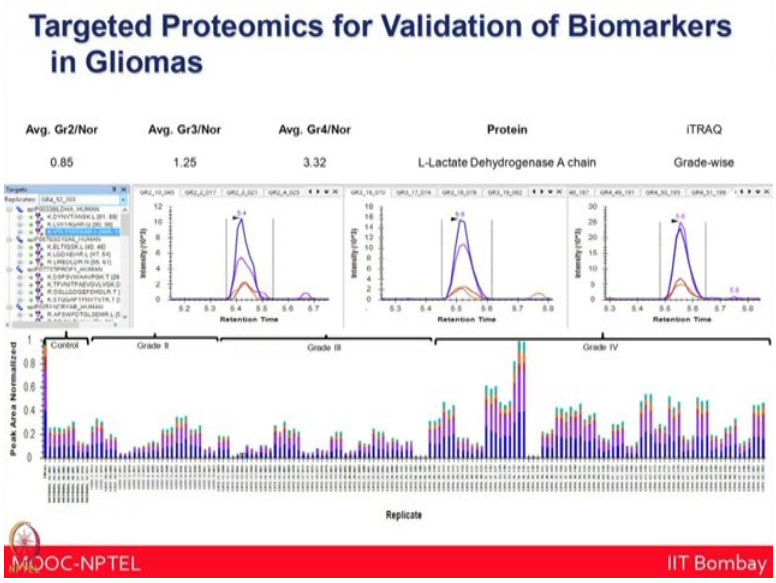
So, one approach which has come forward very promising is known as a SRM - Selected Reaction Monitoring, or MRM - Multiple Reaction Monitoring. And in these you know

technology approaches, these are says now you will see there are triple quadrupole mass spec being used. For example, Q1 which is a mass filter here you are going to select a given peptide, so you are already identified the peptide sequences of interest from discovery workflow

Now, you want to validate a specific peptide. So, you are selecting those you know specific peptide now. Let us say shown in different colors of red, blue and black. Further red one now you are doing the fragmentation in the collision cell, and then further analyzing, its you know product ions. So, you are looking at the precursor and product ion pair which is referred as the transition. And then you are monitoring at least three transition of a given peptide, something shown in these cartoons. And you are looking at least 3 peptides for a given protein to be very confident about your quantification.

So, now using this information that for a given peptide you have measured at least three transitions, and then for a given protein you have taken at least three peptides. Now, you can validate peptides you directly using mass spectrometer. And of course, you will have much more detail eventually in the you know in the lectures as well as some of the demo sessions where you will be introduced to the concepts of you know promising software like a skyline, how to analyze the target data and how to conduct these experiments using targeted proteomics triple quadrupole mass spec in the lab setting, we will also have a demo session on that.

(Refer Slide Time: 21:01)



One could utilize these information for the you know actual clinical work, for example, now we identified some specific targets for one of the studies in the brain tumor. And now we are only looking at that target across large number of patients. These are the patients suffering from you know various grades of gliomas, grade 2, grade 3 and grade 4. This particular protein lactate dehydrogenase, we are trying to measure only that peptide across all of these patients, and we see there are much more abundance of this peptide in Grade IV, and something similar result we were also seen in the discovery workflow when we use the iTRAQ based work.

So, you know you can see that there are different technologies which have complementary nature which could be utilized to complement the information and rather than now generating the antibodies separately for doing validation, you can just use the mass spectrometer even for the validation. And of course, once the validation looks very promising on large number of samples, then you can make synthetic peptide which are heavy labeled, and then try to do the more accurate quantification.

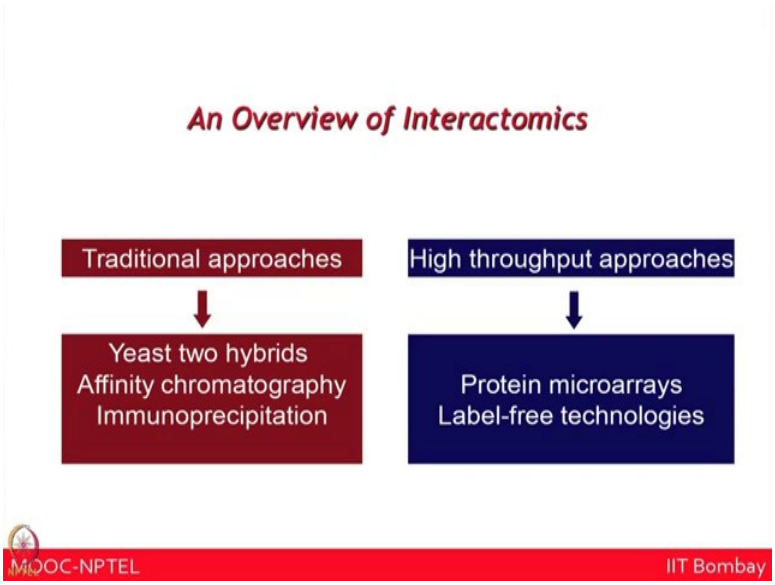
Once even you have passed that kind of you know road block, and you will see that now there is you know detectable amount very reproducible, accurate quantification of given peptide from the clinical samples, then you can take it forward next level, you probably now want to develop a clinical assay directly using mass spectrometer or you can raise antibodies or some other kind of biochemical assays which could be utilized for doing the much simpler based assay like ELISA or Western Blot for the clinical utility. So, this can be you know a path in the workflow for taking your, you know the project of interest from discovery to the you know validation, and followed by taking to the clinical translation work.

So, I know I am kind of you know going fast I am trying to cover variety of fields in a very nutshell, each one of these require lot more detail for you know studying about the details of each technology. But let us kind of capture the at least the breadth of the entire field of proteomics and especially we started from gel-based move to mass spectrometer. And now let us shift slightly toward the interactomics and function proteomics technologies.

So, variety of you know technologies are being developed in this areas of interactomics, where aim is to look at how a protein interacts is another protein, another biomolecule or

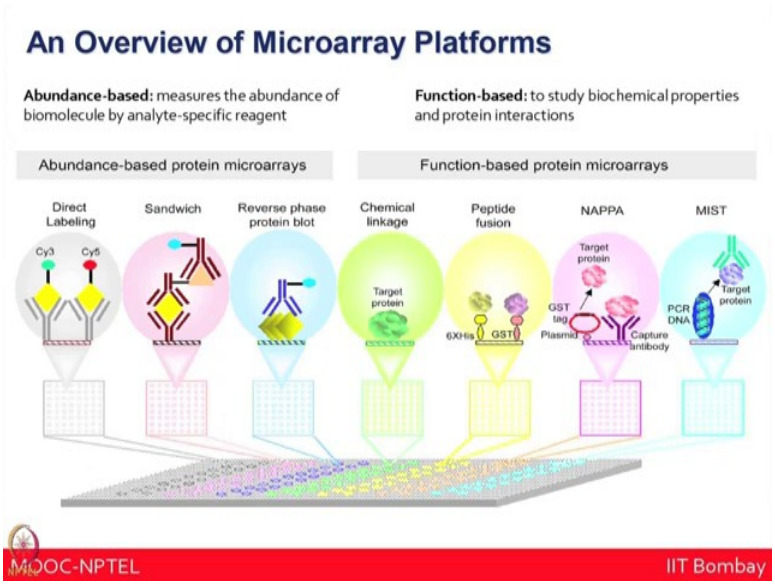
a drug molecule, and how to identify their function or at least get a glimpse of what could be possible function of an unknown protein.

(Refer Slide Time: 23:17)



And this slide traditional approaches like yeast two hybrid or even affinity chromatography Immunoprecipitation (IP) were heavily used. More recently we have started using protein microarrays and label free technologies.

(Refer Slide Time: 23:35)



This complex slide which shows you variety of platforms available for doing protein microarray based work. Let us start you know understanding it you know in the small

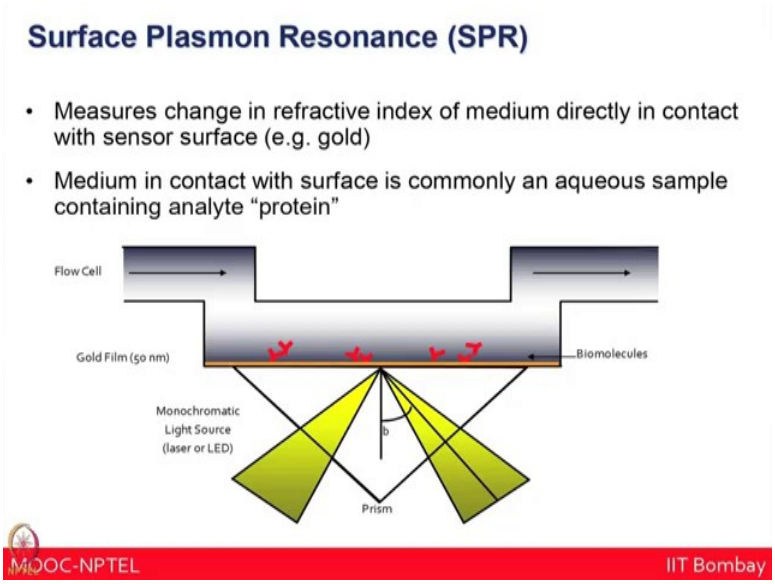
pieces. If you have printed antibodies on the glass slides that is let us say you know these three cartoons which is known as abundance-based proteomics. So, you are trying to measure the abundance of a given protein for which you already have an antibody. And then the way you can develop your assay, it can be direct labeling, it could be sandwich assay, or it can be reverse phase protein arrays. In reverse phase protein arrays, you would like to you know take the tissue lysate or cell lysate and then probing for a given antibody to measure its abundance.

So, this is known as abundance-based proteomics, where aim is to measure the abundance of biomolecules by analyte specific reagents. Or you can also look at the function-based protein microarray approaches, where aim is to study the biochemical properties or the protein interactions for which either you have purified a protein and printed that on the chip. Or you know the specific peptide sequences and you have tried to use those information for the you know printing on the chip, or you can just take the cDNA directly for your not purify the protein, but directly you are taking the DNA or the cDNA print them on the chip, and use in vitro transcription translation mix on top of that, and then try to synthesize protein on the chip. This field is known as cell free expression microarrays or cell free based proteomics.

So, the technologies like you know nucleic acid programmable protein arrays or NAPPA or multiple spotting techniques are promising approaches to do these cell-free expression with microarrays. But all of these four which I have showed are a possible way to look at the function of the proteins. So, in some way that you know either a variety of platforms available for doing the proteomics especially even the interactomics using microarray based workflows.

Additionally another technology if you are now aiming to look at more quantitative information for biomolecular interactions, then you do not want to label the proteins with any given fluorescent or fluorophore. So, here your aim is to look at the biomolecules in its native state and you are looking at in the level free manner.

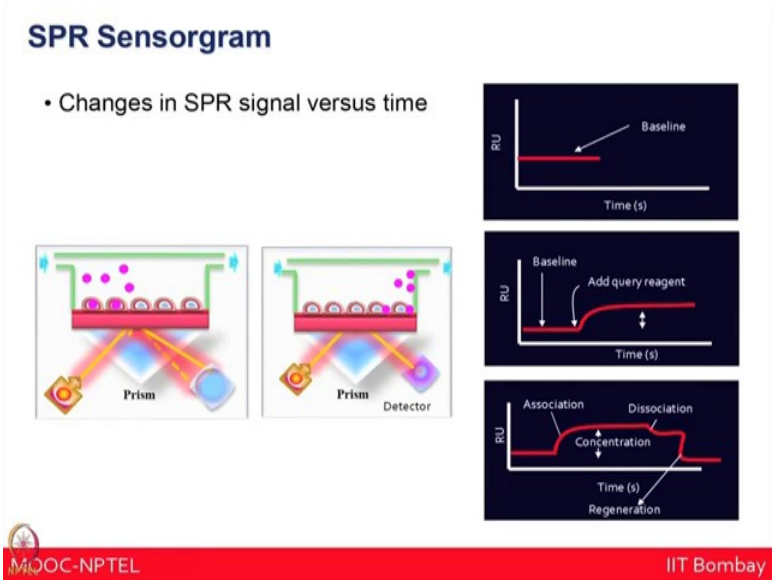
(Refer Slide Time: 25:45)



Surface Plasmon Resonance or SPR is one of the promising technology in this manner, which measures the change in the percentage reflectivity as a result of you know the two biomolecules comes into the binding, you will see the change in their percentage reflectivity or in the refractive index of the medium will change. And therefore, you can see the binding is happening or you know still very weak binding is happening, and that is you know measure under the sensorgram which I will show in the next slide.

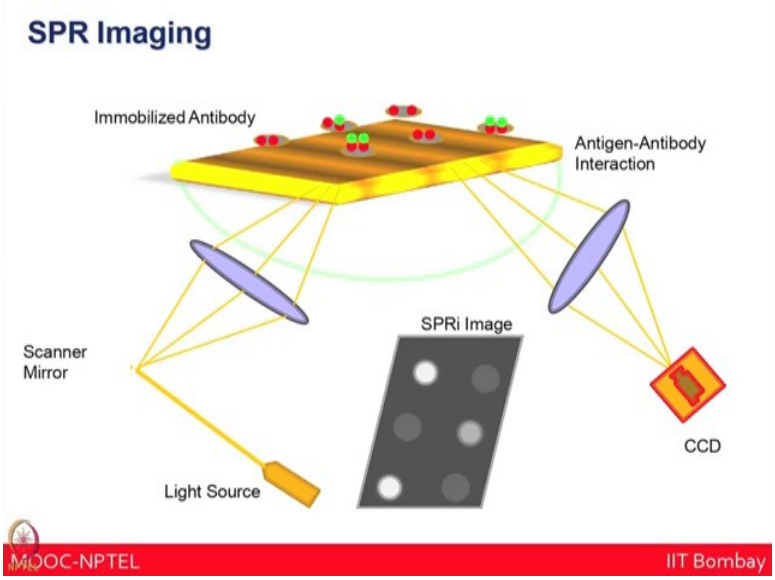
But here this cartoon what it shows that you have a gold slide, you have a light source, and you have immobilized some antibodies of interest, for example, you are you know passing your proteins of interest to measure that there is there is no binding happening for this protein or not. And then if there is a no binding, then you will see the change in the percentage refractivity and followed by you will monitor the sensorgram.

(Refer Slide Time: 26:35)



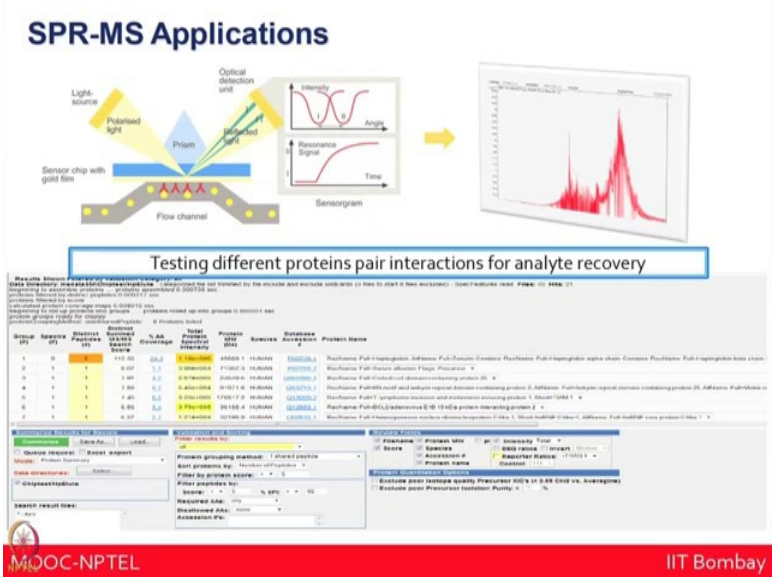
So, if you now look at what you obtain from this kind of experiment. Initially you have you know interest straight baseline, you are now measuring the sensorgram which is a response unit versus timescale. The baseline is very straight. And after you know a binding it starts happening on the chip, you will see the on rate and then you keep flowing the buffer so that you know the molecules cannot dissociate, and that will come under the dissociation phase or the offer rate. And then you want to utilize the same chip for further experiment and that could be studied, after regeneration when now you can use some of the mild acids to strip off your binding. So, this is you know the change in the SPR signal versus time is known as SPR sensorgram_shows you the binding activity and provides information for the on rate off rate or the KD values of a given interaction.

(Refer Slide Time: 27:27)



A new advancement in this area is SPR imaging where now the aim is to look at you know something like you know combining microarray based concept to the SPR based concept. You want to study the molecular interaction in high throughput manner like microarray, but also want to get their KD values and the, you know quantity of information in the high throughput manner and that is the SPR imaging is one of the promising approach.

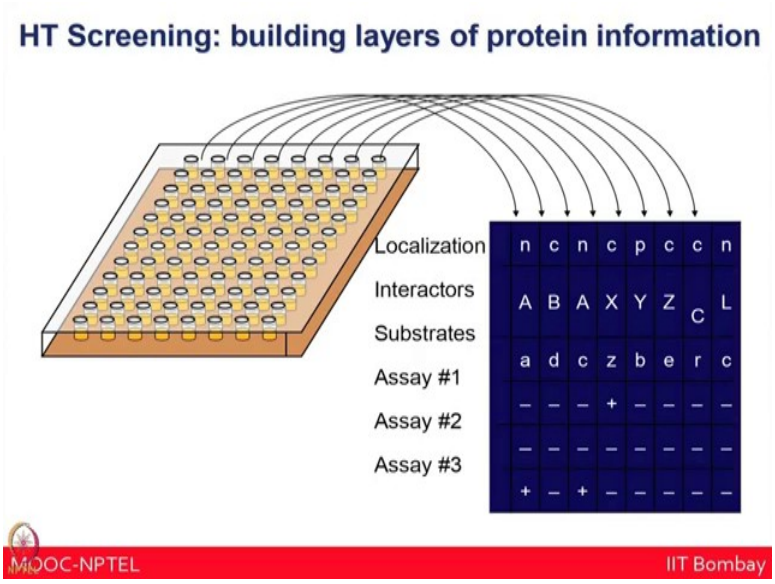
(Refer Slide Time: 27:49)



Additionally one could look at identification of new interactors, let us say you have a you know an antibody which you have immobilized on the gold slide, you are passing a cell lysate or tissue lysate. And you are you know expecting to you know bind some of the unknown proteins on the antibody. So, therefore, you know you want to now identify which are the potential interactors which are binding to this you know antibody of interest.

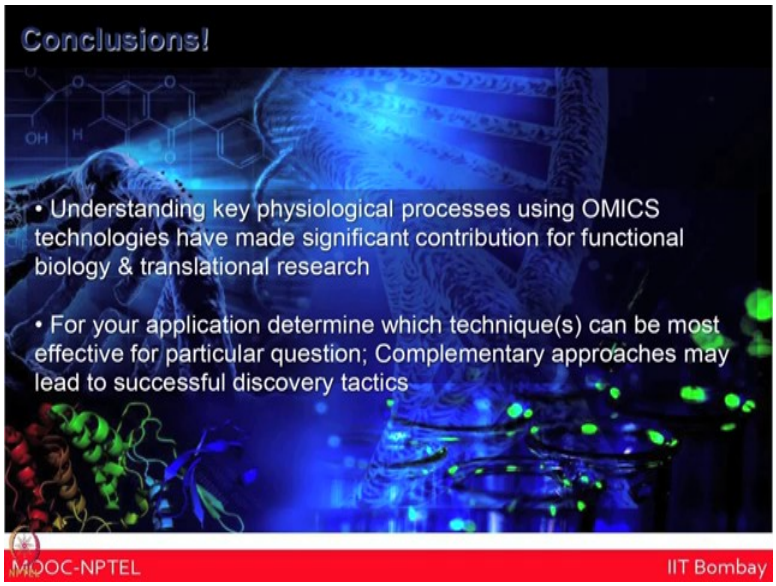
So, can you now identify those interactors and that is where you need to bring in the SPR MS, a new technology approach. So, can you now you know strip of this particular you know bound molecule or the analyte, which is bound on the gold chip, and then generate sufficient of those peptides, so that now you can run on the mass spectrometer, we have tried to optimize some of these technologies. And after you know multiple run cycle, you can know generate sufficient amount of peptide which has given a detectable amount of peptide in the mass spectrometer, and you can identify these in tractors or the proteins of interest.

(Refer Slide Time: 28:47)



So, finally, what we are doing, we are trying to build the layers of protein information using variety of proteomic technologies where proteins are localized, how they interacts which are the substrate for these assays, and variety of ways we are trying to understand the protein function which is very very complex.

(Refer Slide Time: 29:03)



So, in this slide, I would like to conclude that there are many promising proteomic technologies are in development. You need to be very cautious and clear, which technology platforms you are going to use for your addressing a biological question. All of these proteomic technologies have made you know huge revolution of understanding various welfare system, but depending on your need you can choose some time a gel-based proteomic technology or a directly mass spectrometers or use micro arrays or SPR of any other technologies are available to address there a biological question.

So, depending on what is your question, what you want to obtain the answer, you can choose the right type of complimentary proteomic approach. I hope it kind of gives you the now the good foundation and at least the overview of the field of how various type of proteomic technologies could be used for we studying any biological system. As I go along and as you know we are going to cover this part in the workshop, we will talk more about mass spectrometry based workflows and different type of data analysis tools available which will help you to now get the depth of this field, especially how to utilize in your own research.

Thank you.

(Refer Slide Time: 30:15)