Eukaryotic Gene Expression

Prof. P N Rangarajan

Department of Biochemistry

Indian institute of Science, Bangalore

Lecture No. # 39

Genomics & Proteomics

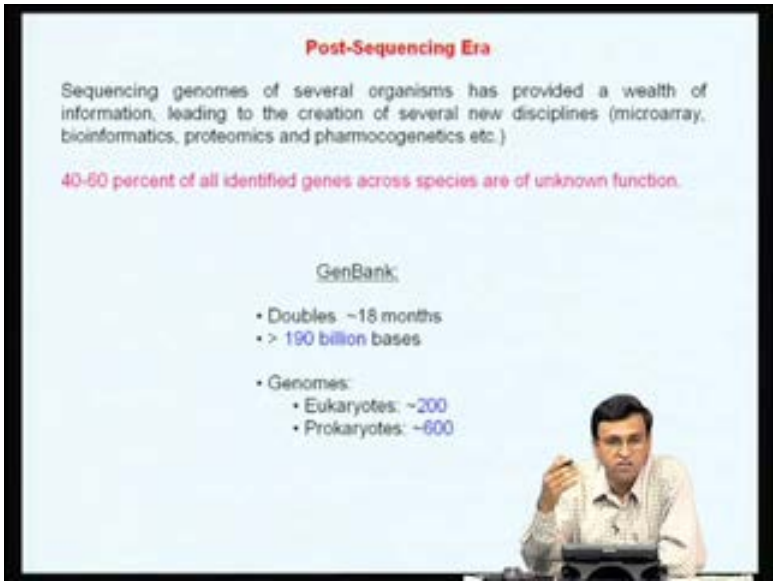
Welcome to this lecture series on eukaryotic gene expression basics and benefits. Today, we are in the lecture number 39. We are primarily going to discuss about a very interesting and very marvel topic, genomics and proteomics.

(Refer Slide Time: 00:36)

Now, the purpose of giving this lecture is, to make you understand, how we can study global changes in gene expression; how we can understand global changes that take place in gene expression; what are the new technologies that are available to us, and what are the new challenges ahead of us. Now, the year 2000-2001 is a kind of a landmark in the molecular biology as well as gene expression and eukaryotic genome biology, because the complete genome sequence of the humans was published by two independent groups, one that appeared in

say science and another appeared in nature. So, this sequencing of the human genome told us about the number of genes that are organized in the human genome and paved way for sequencing a number of other genomes as well.

(Refer Slide Time: 01:23)



So, we are basically in a post-sequencing era, where sequencing genomes of several organisms has provided a wealth of information, leading to the creation of several new disciplines, be it microarray based disciplines, bio informatics (( )) genetics to just name a few. Now what is become evident, once the genome sequence became available not only for humans, but also for many mammalian species, which has mouse, rat, and so on and so forth. About 40-60percent of all the genes that were identified were classified as unannotated genes, because we do not know what the function genes is, we do not know what kind of proteins they code for, and we do not know anything about the function of these proteins. So, one of the important things that came out of this human genome sequencing or sequencing of mammalian genomes is that or even microbial genomes that we still do not understand the function of many of these genes.

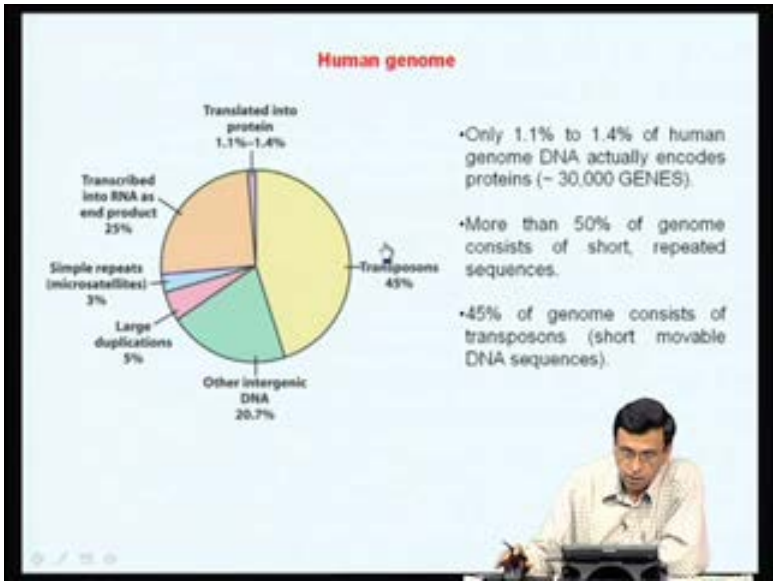
So, with more and more sequencing technologies becoming available and more, more refinement of this technologies, more and more genome sequence are being made available, and it is actually estimated that in the GenBank, which is actually a depository for all these such genome sequences, the sequence data almost doubles every 18 months. And about 190 billion base

sequences, bases of sequence data is being added to this GenBank, almost every year. And there are approximately about 200 eukaryotic genomes and about 600 prokaryotic genomes either being sequenced or closed (( )) relation.

(Refer Slide Time: 03:11)

So, what is happening is that a huge volume of data is being generated and to actually make sense what of all these data, it became impossible for a human to handle all this huge volume of data. So, a new era of computational biology known as bio informatics came into prominence. So, you required software's to be developed, you required computations aids, you required computers to handle this huge volume of data, and organize so that one can actually go and retrieve information that you are interested at.

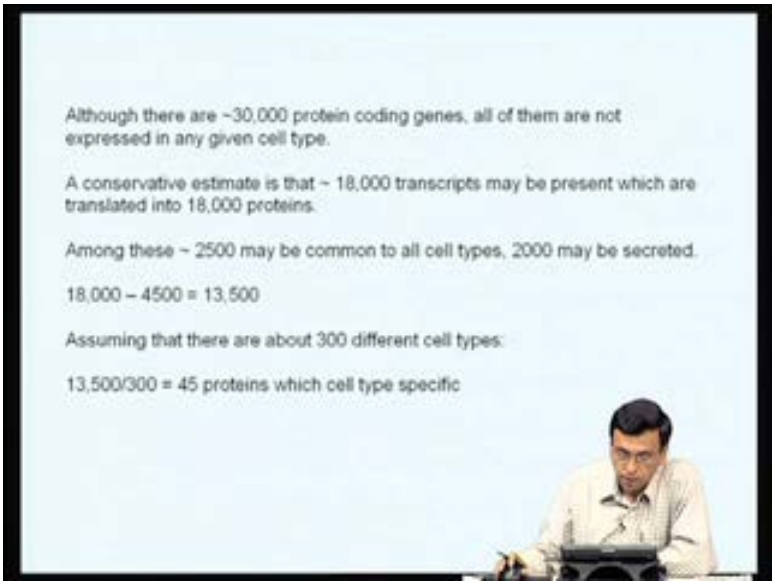
(Refer Slide Time: 03:38)



So, let us now just take a look at what the human genomes sequence data told us. The human genome data sequence actually told us that only 1.1-1.4 percent of the human genome D N A actually codes for the proteins, that means, only a very small fraction of the human genome actually is involved in coding for proteins. There contains a protein coding genes as I can shown here. So, only 1.1-1.4 percent of this entire genome is actually transferred into proteins and very surprisingly almost 25 percent of this human genome actually codes for R N A , which never gets translated into proteins. So, the end product of this genome is actually R N A and we have studied in our previous classes that how R N A can have very important regulatory functions like the micro R N A s, si R N As, and so on so forth. In addition to the small R N As, there are also number of other R N As, which are broadly known as the nc R N As because they do not code for any proteins and the R N A itself is the final functional entity. So, a lot of information is coming out about the structural or functional or regulatory roles of these various R N As, which are the ultimate end products of a huge part of the human genomes and it turned out about 3 percent of the human genome consists of what are called as microsatellites, which became very-very important for identification of disease markers and for what is called as a functional genomics and also some are known to contain large duplications and many other regions of D N A, which is broadly classified as intergenic D N A and very interestingly about 45 percent of the human genome sequence consists of what are called as moving elements or transposes, which

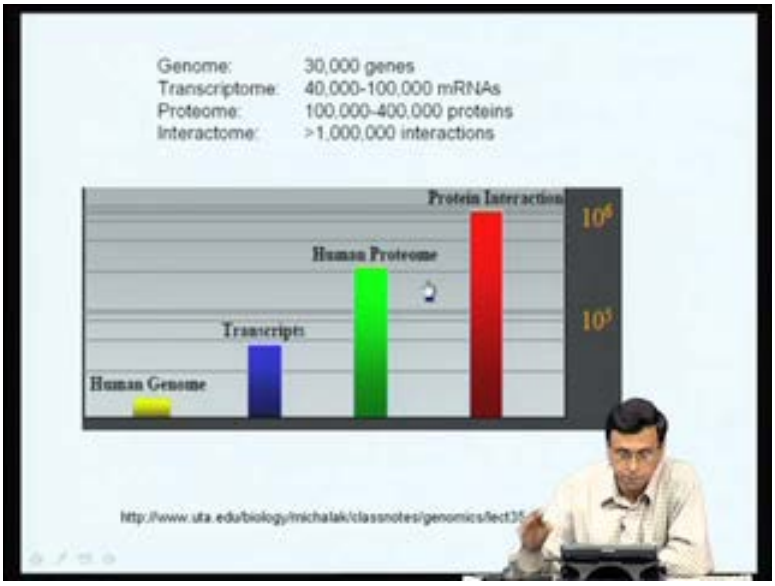
can jump from one place to another. So, as you can see what the human genome actually told is that, people who have been focusing their attention all the time on protein coding genes and how R N A polymer is to transcribe genes, how messenger R N A synthesized, they all became (( )) with this kind of information saying that what they are actually studying is only a very small fraction of the human genome and the rest of them probably we do not know anything about.

(Refer Slide Time: 05:53)



So, if we now look at these about 30,000 protein coding genes, that is what they estimate in case of the human genome. All these 30,000 protein coding genes are not expressed in any given cell type and it is estimated that about 18,000 transcript may be presents and therefore, these 18,000 transcript may be translated into about 18,000 proteins. This is the approximate estimate in about an average number of transcript that may be present in a given cell type. Now, among this 18,000 proteins, about 2500 may be common to all cell types. These are what called as the house keeping proteins, which are required for they are involved in energy metabolism or structural component and so on and so forth and about 2000 may be secreted out. So, if you now deduct this from 18,000 approximately 13,000 genes may be cell specific and assuming there are about 300 different cell types. In various mammalian systems, there are about 45-50 proteins that may actually be highly cell type specific.

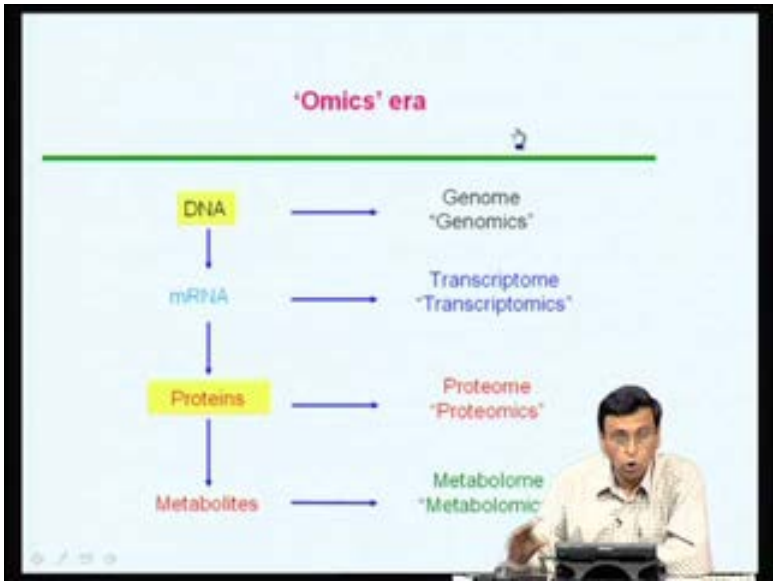
(Refer Slide Time: 07:02)



So, these are all the average estimates that one gets out of the information that is available in the uh human genome databases. So, you have a human genome, which approximately codes for about 30,000 gene protein coding genes and these 30,000 genes by a number of other processes such as differential transcription, differential splicing, or differential polyadenylation, and so on and so forth can actually generate multiple transcripts because we have already studied that how more than one messenger R N A species can emanate from a single gene. So, it is estimated that about the human transcriptome that is the number of R N A molecule that can be synthesized from the human genome is anyway varies from 40,000 to100,000. So, this many messenger R N As can be synthesized or the human genome has the potential to synthesize this many number of messenger R N As. Again when this messenger R N A gets translated into proteins, it can undergo splicing or they can undergo various post translation modifications and each post translation modification can give a different function to the protein. So, the proteome that is the protein component of the human genome actually consists a newer protein 100,000 to 400,000 proteins, which can have different functions and these proteins again they are not in the cell. These proteins interact with each other and form from key components of either metabolic reactions in the form of enzyme complexes or major signal transducers in major signal transduction pathways and these protein-protein interactions may almost be around one million.

So, you can see the ultimate physiology of an organism of a mammal ultimately depends on these one million protein-protein interactions that ultimately govern the various functions of the cell or the organism. So, you have genome and you have protein transcriptome, you have proteome, and these proteins interact to generate what is called as (( )) interactome. So, as you can see from the human genome sequence it became very clear as to functional side of the mammal the number increases logarithmically.

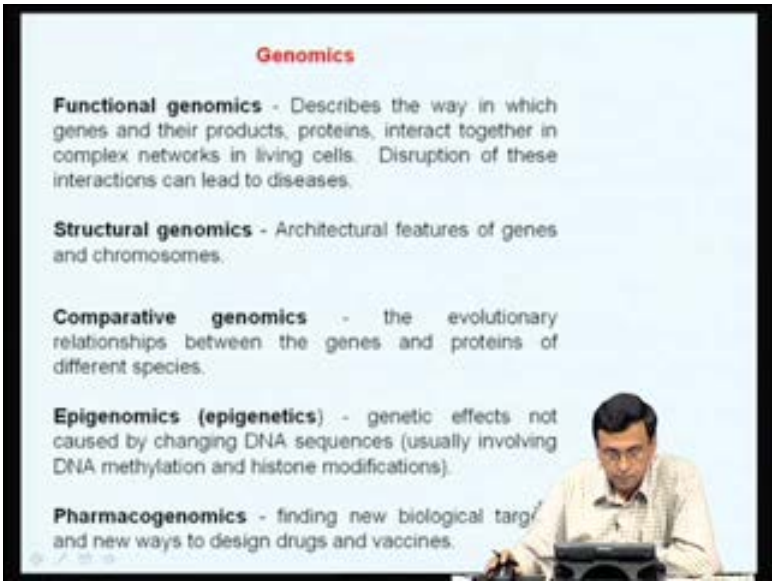
(Refer Slide Time: 09:07)



So, we have now basically entered what is known as the omics era, which is (( )) just now explained the analysis of the sequence information of the genome gives rise to what is called as genomics and when you start looking at how these genes are transcribed into messenger R N A and then look at the diversity of this messenger R N As you get what is called sub-transcriptome or a transcriptomics and then when messenger R N A gets translated into proteins, do you want to look at the various variations in proteins and how these proteins get modified and how they contribute the function, you get what is called as proteome or proteomics and when these proteins interact with each other and form interacting networks and then contribute to specific metabolic pathways or signal transduction pathways you get what is called as a metabolome or metabolomics. So, as we go from genome to metabolome, the complexity increases several folds. Within genomics, again several sub disciplines came into being, I have just listed a few here, we have what is called as a functional genomics, which describes the way in which genes and their

products, that is the proteins interact together in complex networks in living cells. And if we disrupt these interactions, it can lead to number of diseases.

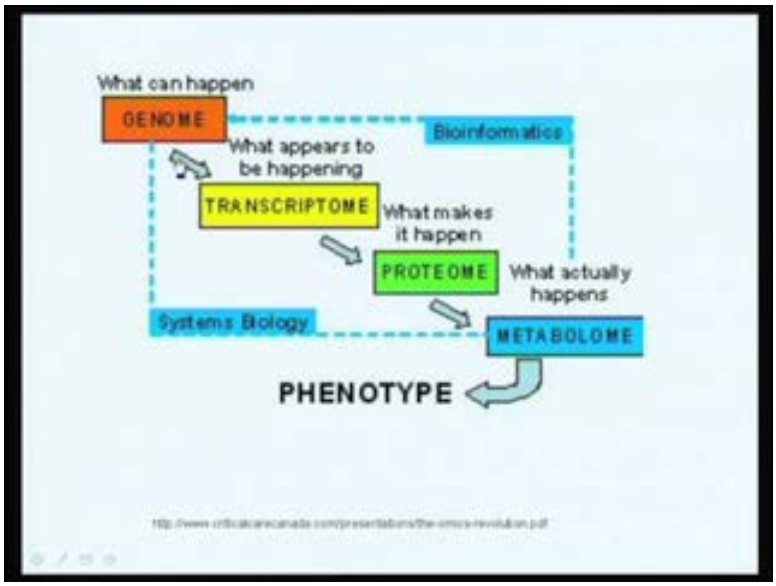
(Refer Slide Time: 10:03)



So, functional interactions between genes and gene products leading to normal physiology disruption on, how these disruptions or interactions can lead to disease phenotypes, what is known as the functional genomics. And you have structural genomics, which primarily focus on architectural features of genes and chromosomes. How these genes are organized? Euchromatin, heterochromatin, the clusters intra and extra organization and so and so forth. The repeat sequences and all these things then you have what is called as comparative genomics, because as more and more genome sequences became available from a number of organisms, people started comparing, for example, if you compare a human genome with chimpanzee genome, it is more than 99.5 or more than 99 percent identical. So, what distinguishes between chimpanzee genome and human genome is only one percent. So, you can really see, these variation that we see between organisms or between species may not been entirely accounted for the sequences diversity. So, we need to come down further towards proteins and their functions and the R N A and their functions to really understand how this genome ultimately contributes to diversity among various organisms.

So, I have compared to genomics which basically looks like evolutionary relationships between genes and proteins of different species. And epigenomics, which was studied in great detail in the last few lectures, which basically talks about genetic effects not actually caused by changing the D N A sequences, epigenome that is above genome sequence. And usually these kinds of epigenomic changes involve either D N A methylation or histone modifications. We have studied extensively, how epigenetic plays a very important role in number of diseases as well as development. Then, we have, what is called as pharmacogenomics that is how you can make use of these genome information and the proteome information to develop, design new drugs or new vaccines. How you can develop new drug targets, how you can develop new drugs, and so on and so forth. So, you can see the sequencing of genomes has really opened up whole new areas of research in the area of biology and a number of new disciplines came into being to really understand, what is going on inside the cell, and how all these transcripts, proteins and the genes contribute to the diversity among organisms.

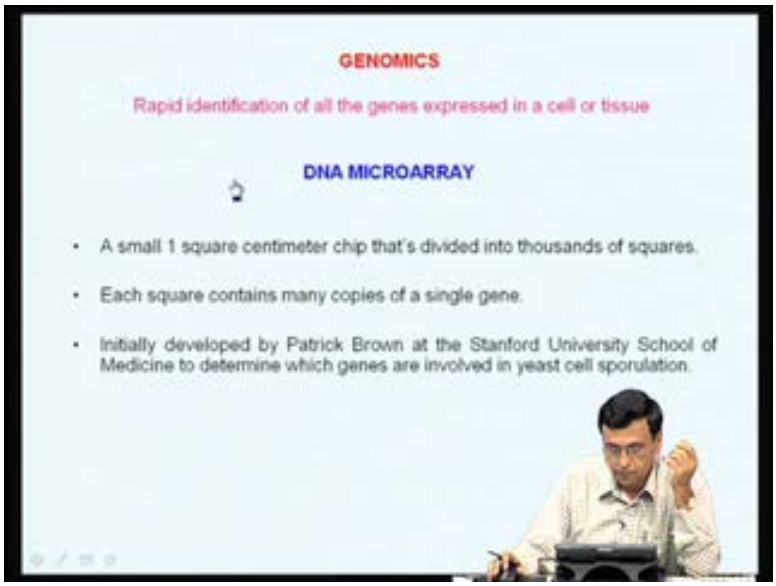
(Refer Slide Time: 12:43)



So, the genome sequence basically tells you about what can happen? Now once the R N A gets made or once the transcripts are generated from genome, the transcript on basically tells you what appears to be happening. Once this R N A gets translated into proteins and you start looking at the proteome, you understand because proteins are the workhorses of the cell, it actually tells you what makes it happen and then when ultimately these proteins work together or

interact with each other (( )) networks and you start look at metabolic interactions and metabolic networks, you actually (( )) what actually happens. So, in order to understand to connect this genome information to the metabolome, you require key players and one is called as the bioinformatics huge amount of computing power is required to put together all these data and make connections between what is happening at the gene level and what is happening at the metabolism level and also a very new area that actually does this uses the computing power or uses computational biology and this sequence information and the various experiments that are available is actually what is known as the systems biology. So, very complex areas of research have actually come into being with the analysis of the genomes, proteomes, and so on and so forth.

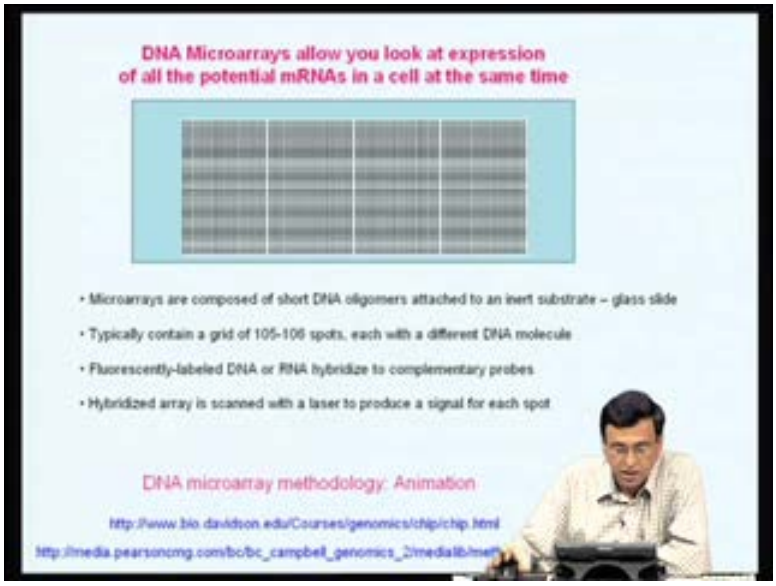
(Refer Slide Time: 14:07)



So, lets us spend time to now understand what actually is going on in the area of genomics, what are the exerting developments that are taking place in the area of genomics. Now what is genomics, a simple straight forward definition of genomics would be rapid identification of all the genes expressed in a cell or a tissue its the simplest definition of genomics. How will you quickly identify that what are the total number of genes that are expressed in a given cell. Now, so far in all our studies earlier, we have being studying only gene expression changes that are happening at the gene level, at the most we are looking at a cellular, at different stages of development, and how each transcription factor or how each gene contributes to certain

important changes, but today with technology has developed and technology has made it possible to understand all kinds of genes are expressed that is the total changes that takes place in a genome at the level of transcription and you can look at the entire gene expression changes that is taking place either at the cellular level or at tissue level. The key player for that enables us to understand this kind of global changes in gene expression that is taking place in a cell or in a tissue is what is called as a D N A microarray.

(Refer Slide Time: 15:42)



So, let us spend some time to understand how actually one can do genomics by using these microarrays. What is the microarray? A microarray is nothing, but a small one square centimeter chip that is divided into several thousands of small squares. Now each square contains many copies of a single gene and this microarray was initially developed by Patrick Brown at the Stanford University School of Medicine, who determine which genes are actually involved in these cells correlation. Let us spend little bit more time to understand this microarray because that is the key player for understanding for doing genomics. So, what happens with the D N A microarrays? D N A microarrays allow you to look at expression of all potential messenger R N As in a cell at the same time.

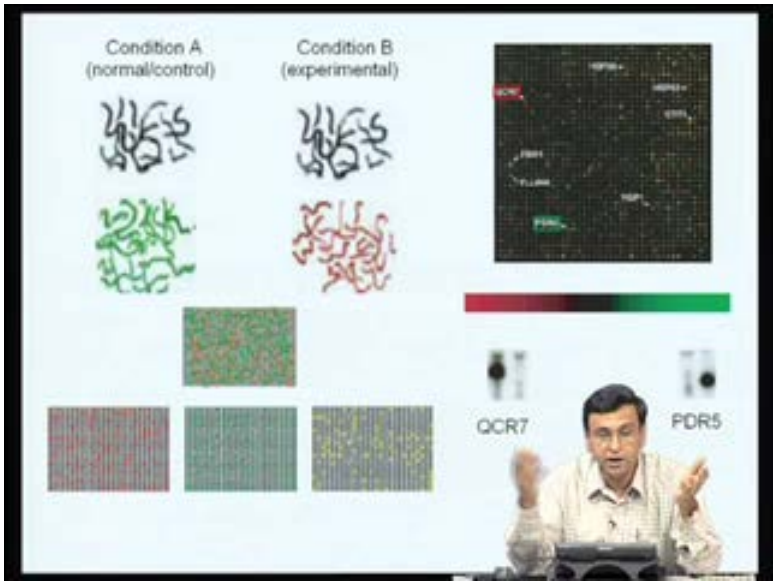
Now we have earlier studied techniques like RT-PCR, northern blot, and so on and so forth where at the most you can look at handful of genes maybe 10 genes, 20 genes. You can look at

the number of genes that are expressed in a particular tissue, but the number is a limited, you cannot go beyond 10 or 20 genes, but if you want to look at all the genes which are getting expressed in a given cell, you can only do it with technologies like microarray. So, what actually happens? A microarray as I said is nothing, but a chip which is actually present on a glass slide. They are composed of short D N A oligomers attached to inner substrate or a glass slide. The each dot that you see here represents one particular gene probe. So, let us say there are about 30,000 spots here; that means, 30,000 individual genes that have been laid on this microarray. Now typically each one of this grid consists of about 105 to 106 parts, each with a different D N A molecule. So, these D N As are actually spotted on to this glass slide by using a technique called as lithography. In earlier times, people usually actually synthesize this D N A, either in the form of oligonucleotides or cD N As and then this used to be spotted onto this microarrays by robotics, but today it is actually possible to do the entire gene synthesis right on the chip. So, we can actually synthesize different D N A molecules right on the chip and using what are called as oligonucleotides arrays. So, once you have all these genes, which have been spotted on this genome microarrays, you do not take this microarray for example, if you call it as a mouse microarray when you spot all the genes present in the mouse genome or oligonucleotides corresponding to different genes of the mouse or you call it as a human microarray, if human genes are spotted and so on and so forth. So, once we have these kinds of microarrays ready, you then hybridize these chips to fluorescently labeled D N A or R N A, so which are complimentary. So, I can isolate, for example, if you want to look at R N A, which is expressed in mouse liver under two different conditions, we extract this R N A and then label them differentially using two different dyes and hybridize to this particular chip and then you wash off the unhybridized probe and then scan this using a laser scanner. I would not like to go to too much details of this thing because it is a very fascinating technology. I would actually suggest you to actually look at the animation of D N A microarray methodology. How actually these microarrays are developed and how the microarray experiments are carried out. There is a very nice animation, you can go to this particular website either (( )) just cut and paste it and then just look at a beautiful video that actually tells you how actually this microarrays are made and how actually microarray experiment can be carried out. So, I am just going to very briefly tell you how exactly a micro experiment is done. The rest you can and watch that video. Let us assume for example, there are two different conditions, one is a normal condition and another is an

experimental condition. This can be anything. For example, you want to look at a normal cell, what happens to gene expression in normal cell versus not after gene expression a cancer cells or you can you can look at what is happening between a unstimulated cell and a stimulated cell. Stimulated cell, I mean you can add a hormone or you can add a growth factor any component you are interested in. So, any two different physiological situations one can take and then you basically isolate R N A from these two different conditions. So, I isolate R N A from the normal cells or you isolate R N A from whatever condition that you are interested in. Once you isolate this R N A, you then convert this R N A into cD N A's and then these cD N A's are actually labeled with fluorescent dye. The situation condition, A, you label the R N A with a fluorescent dye, which fluoresces green and the condition B, you label this R N A with a dye that fluoresces red. Then you mix these two R N A's together and then you hybridize the microarray, which we described with the previous slide. So, this microarray actually contains, let us say for example, this R N A is made from a mouse and you are hybridizing this R N A for a mouse microarray and both these R N A's now hybridized. So, let us now look at this microarray carefully, so you have all the genes, which are actually present in the mouse genome, let say about 30,000 genes and each pot corresponds to one particular gene and depending upon whether the gene is getting expressed in condition A, if this particular gene number one is expressed in condition A, then the green labeled R N A will hybridize to this ,red will not. On the other hand, if gene number two is actually expressed in condition number B, but not condition number A, then the red labeled R N A will hybridize to this, but if the gene number c is expressed in both conditions, then both R N A's will hybridize and you will ultimately get a yellow signal. So, depending upon which gene is getting expressed with corresponding R N As, complementary R N A's, will hybridize and accordingly you get either a green signal or a red signal or a yellow signal and that is what is shown here. Red actually indicates to all those genes, which are active in condition B. The green spots here indicate all those genes, which are active in condition A. The yellow spots here actually tell you genes, which are active under both conditions A and condition B. Now, sophisticated softwares are now available, you cannot do this analysis manually because they are all very-very highly microscopic regions. So, you take this hybridization pattern and then put it upon using appropriate software, you can actually understand which of these genes, which are getting expressed in condition A or condition B or in both. For example, I have just given a example here say for example, I know exactly the region on the microarray here actually

correspond to gene called QCR7 and this gene if it is red; that means, this QCR7 actually expressed in condition B. Similarly here is another gene for example, we have YGP1, which is spotted which is red here, which means this gene actually express in condition B, but not condition A and so on and so forth. So, using appropriate softwares and computers, you can actually identify, which are the genes which are getting expressed in condition A, which are the genes which are getting expressed in condition B, or which are getting expressed in both conditions.

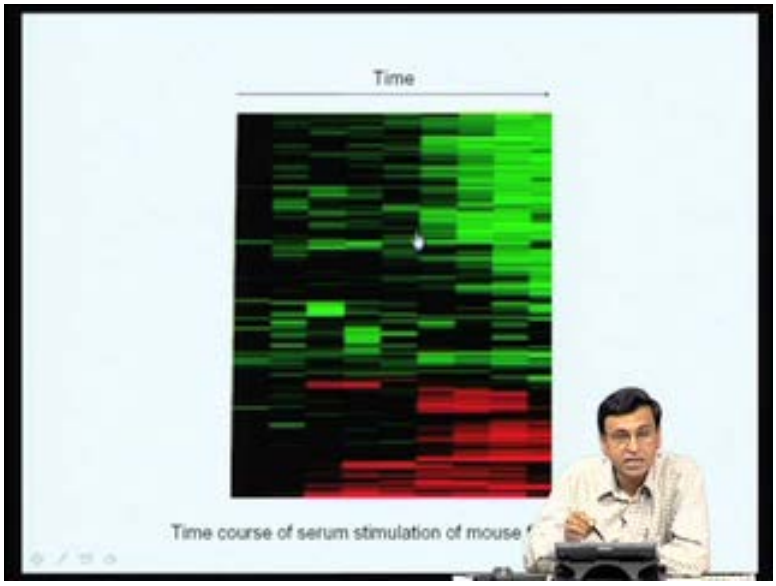
(Refer Slide Time: 19:16)



So accordingly, you can actually also get the information on the fold induction, how many fold the R N A is more in one particular condition compared to the other again using computational biology from biological techniques and once you have actually (( )) the data you have got actually is valid (( )) experiment for example, if I want to really see whether the QCR7 is actually getting expressed only in condition B, but not in condition A, you again take the R N A from condition A and condition B and make a QCR7 probe and then do a northern blot analysis and actually show that yes this data is correct. Similarly, I can do an experiment actually, take another proof for example, if I am expressed here, I should validate again doing a particular northern blot or you can also do what is called as RTPCR or real time PCR and so on and so forth. So, this is basically how a microarray experiment is carried out. So, you can see what this microarray actually given you the ability to analyze the expression of thousands of gene at any

given time. So, you can understand or you can identify all the genes that are getting expressed in that particular cell at a given time and you can actually compare two different conditions and make conclusions of what kind of expressed or what kind of genes may be responsible for that particular physiological response. (( )) For example, you had a hormone, if it is going to activate what are all the genes, which are activated by that particular hormone in that particular cell type. This can be found out by using these kinds of microarrays. So, microarray became a very-very powerful tool to understand global changes in gene expression. I have just going to give a few examples just to clear how actually a microarray data looks like.

(Refer Slide Time: 23:34)

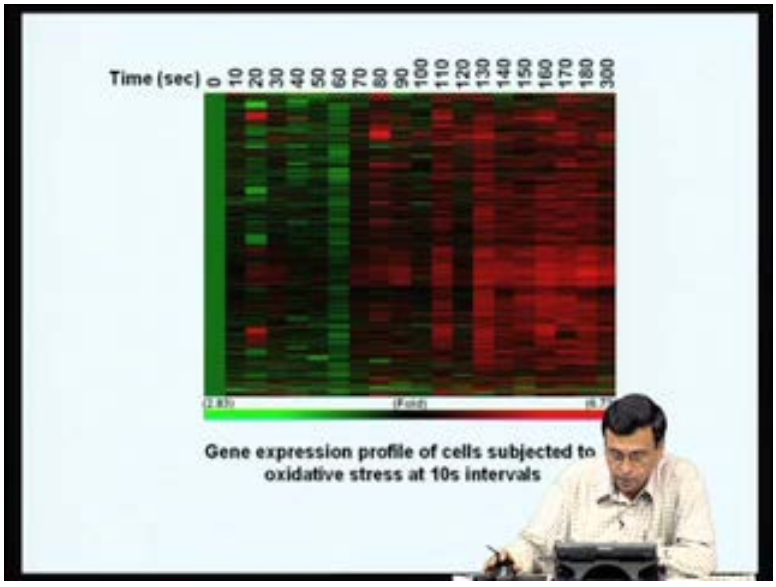


For example, I am interested in looking at the time course of serum stimulation of mouse fibroblasts; that means, mouse fibroblasts, which are normally quiescent and as you know if the mouse fibroblasts have to proliferate in cell culture, you require serum. Now as you know serum contains a number of growth factors, so when you add serum to the fibroblasts it stimulates cell proliferation. So, a number of genes that are involved in cell proliferation should go up and now I would like to know what kind of genes are activated.

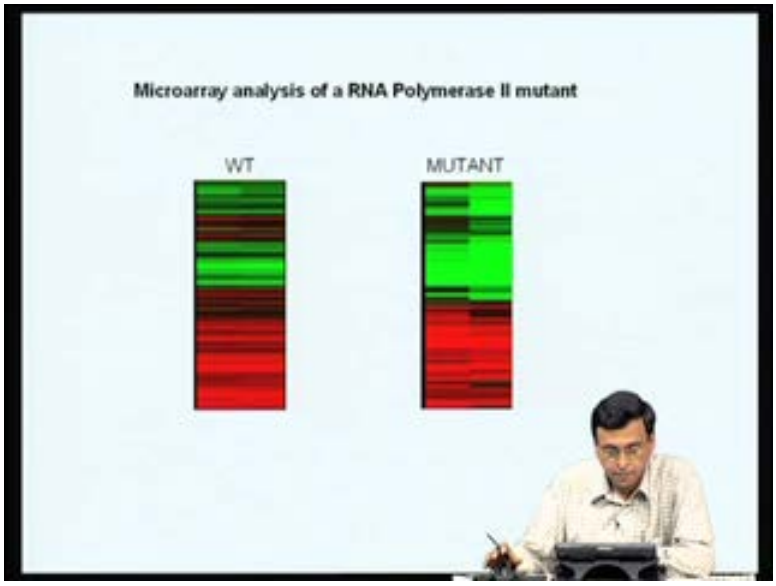
First what kind of gene expression genes are activated at the later time points, so I do a time course. So, starting from minutes to hours, I now isolate R N A from different time points and do this microarray and you can see clearly for very few genes are active when there is no serum and

once you add serum and then start doing time course, you can see a number of genes are getting activated at the same time a number of genes are getting repressed and I can now to the microarray and using those software data, I can actually analyze and see what kind of genes are getting repressed when I add serum and what kind of genes are getting activated when I add serum and then I can experimentally validate and then find out the physiological functions for these gene products. So, I can now correlate, this is the mechanism by which serum addition of serum stimulate cell proliferation. These are the gene products involved and how these products actually interact with each other and then contribute to this cell proliferation phenotype. Similarly, here is another example where let us say I am interested in gene expression profile of cell subject to oxidative stress at 10 second intervals. So, I can do the oxidative stress and isolate R N A at 10 second intervals and then look at the gene expression changes that takes place because of this oxidative stress at different time points and again you can see some genes are going up, some genes are going down, and one can look at these data and then make proper annotations and make certain conclusions.

(Refer Slide Time: 24:53)

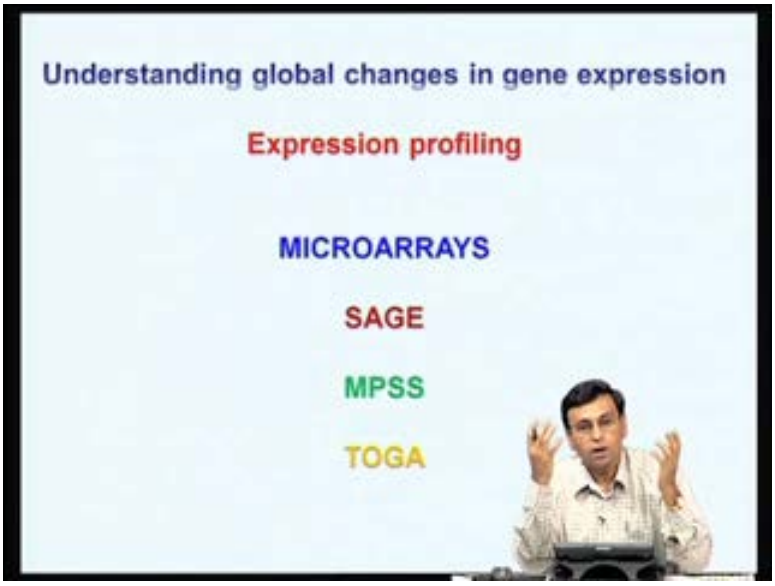


(Refer Slide Time: 25:21)



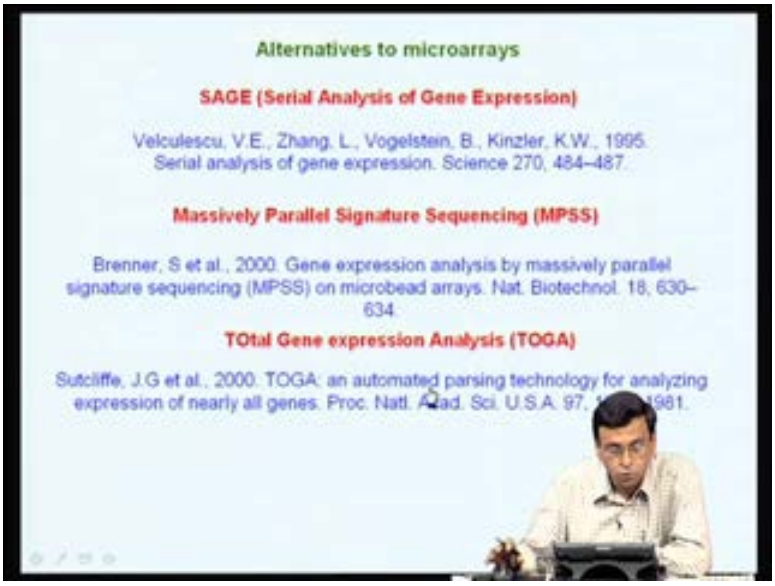
Here is another example, suppose I made a mutation in one of the R N A polymerase to subunits and I want to see how does this mutation or his R N A polymerase effect gene expression because as you know R N A polymer is very-very important for transcription. So, again I take R N A from a wild type and I take a R N A from mutant and then compare these two different conditions and then see how this particular mutation of this particular subunit of R N A polymerase, what kind of genes are getting affected and you can see these are the genes which are affected by this mutation in a R N A polymerase to subunit. So, I can.

(Refer Slide Time: 25:52)



So, I can go on giving a huge number of examples, but the purpose is not to load you with information, but actually tell you what is the experimental approach that one follows by if you want to understand global gene expression changes that are taking place the microarray be has become a wonderful tool to understand how gene expression changes can be measured at the genome level. So, this understanding gene expression change is at the global level is actually known as expression profiling. Now, what I told you is just one technique that is how microarrays can be used for doing this expression profiling, but there are number of other techniques also which are not as popular as microarrays, but they can actually compliment certain defects of microarrays. These are called as SAGE, MPSS, TOGA and so on and so forth,

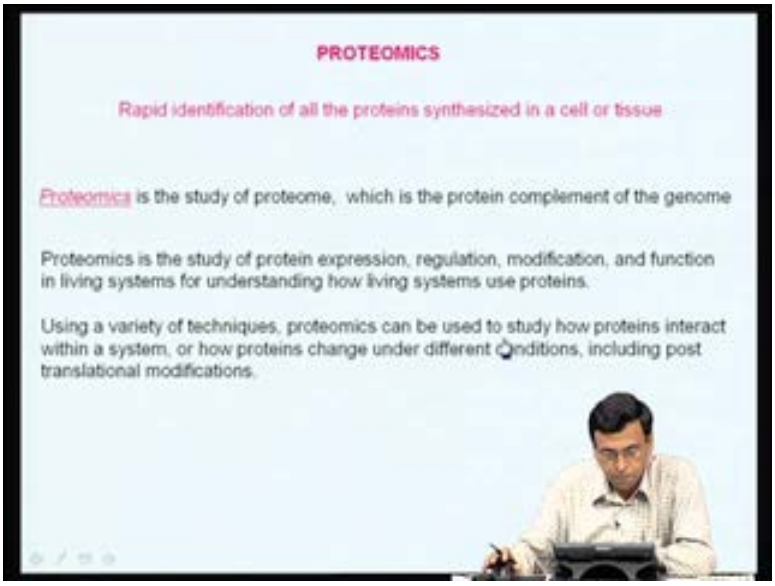
(Refer Slide Time: 26:39)



but to lack of time I will not go into this other alternate techniques, but I what I have done here is SAGE actually refers to what is called as Serial Analysis of Gene Expression. These are all the technologies, which are also used to understand global changes in gene expression and their alternative to microarrays. So, SAGE, Massively Parallel Signature Sequencing (MPSS), TOtal Gene expression Analysis or TOGA. I have given the references one whoever is interested can actually read this very nice articles and try to how these techniques also can complement the microarray technology for understanding global changes in gene expression and these are all very-very powerful tools and gives you a huge volume of data and huge volume of information about disease phenotypes and so on and so forth. Ok, So, what I told you is to give a very comprehensive view about the importance of genomics, how genomics can be carried out by doing microarray experiments, and how the microarray experiments, I gave you few examples of how information can be obtained on two different physiological situations and how one can use the microarray data to understand global changes that take place in the area in gene expression. Now, once the genomics and once people started doing this kind of a genomics and once people started mining what is called as a data mining using this microarrays, people thought they can understand the biology of the organism, but it soon became clear that just by doing gene expression profiling, we cannot understand everything because ultimately, we also, I need to understand how proteins function as I said once (( )) transient to protein, the protein can have

pleiotropic functions depending upon its own state, depending upon its primary structure, secondary structure, tertiary structure, its interaction with other proteins as well as post translation modifications of proteins, all these things play ultimate role in ultimately giving a phenotype per a particular cell, whereas, gene expression profiling does not tell you what is happening to the protein. It can only tell you that these messenger R N A ‘s are going up, these are going down, and then extra polite and say these are the protein levels are going up or going down, but what is the functional consequence of this. The proteins levels may go up, but unless it is phosphorylated, the protein may not be active. So, the phenotype of a phosphorylated protein may be difference on a phenotype of a dephosphorylated protein, similarly glycosylation or similarly unless the protein interacts another protein. So, it is not just enough that protein goes up another protein is required for a particular phenotype because unless these two protein interact, humanoid get a particular phenotype. So, it became very clear just by doing microarray analysis or just by doing expression profiling, you cannot explain all the things that are go happening the organism. So, you need to also develop tools to understand global changes that are taking place at the level of proteins and this analysis of global changes that are taking place they are all proteins came to be known as proteomics. So, genomics logically lead to proteomics because it became clear it is not enough to just do expression profiling, you also have to look at the changes that are taking place at the protein level and you need to have powerful tools just as the microarray, you also need to have powerful tools to understand changes that are taking place at the protein level. So, let us now try to understand, what is this proteomics and how proteomics as contributed to the understanding of various biological phenomena.

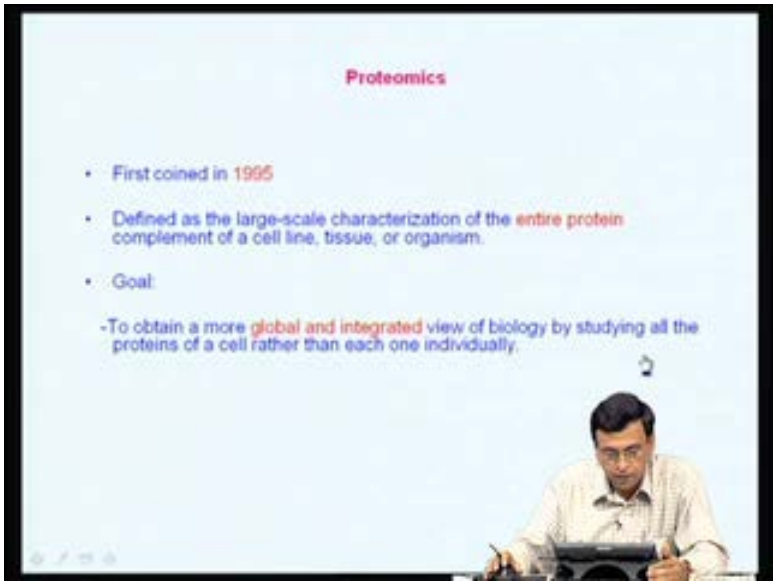
(Refer Slide Time: 30:02)



So, what is proteomics? As again just like we defined genomics, proteomics is rapid identification of all the proteins synthesized in a cell or a tissue. So, just as genomics is rapid identification of all the R N A that is inside in a cell, here proteomics actually is a rapid identification of all the protein synthesized in a cell or tissue. So, proteomics is actually the study of a proteome, which is the protein complement of the genome and proteomics is the study of protein expression, regulation, modification, and function in living systems for understanding how a living systems used proteins because they are the ultimate workhorses.

So, using a variety of technique, proteomics can be used to study, how proteins interact within a system or how proteins change in a different conditions including post translation modifications and protein-protein interactions and so on and so forth. So, let us now would discuss it briefly, what kinds of techniques have been developed to do proteomics, how does one go about doing proteomics, and what are the applications of proteomics.

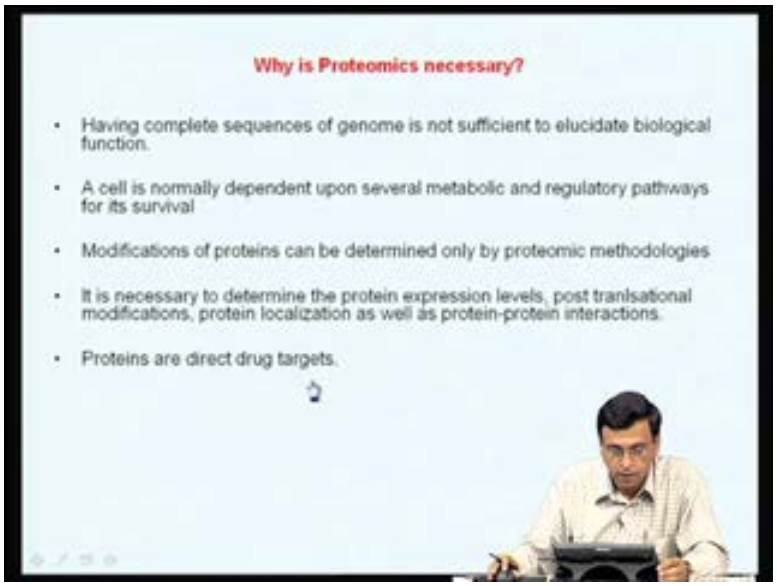
(Refer Slide Time: 31:04)



The term proteomics was first coined in the year 1995 and it is actually defined as the large scale characterization of the entire protein complement of a cell line tissue or organism.

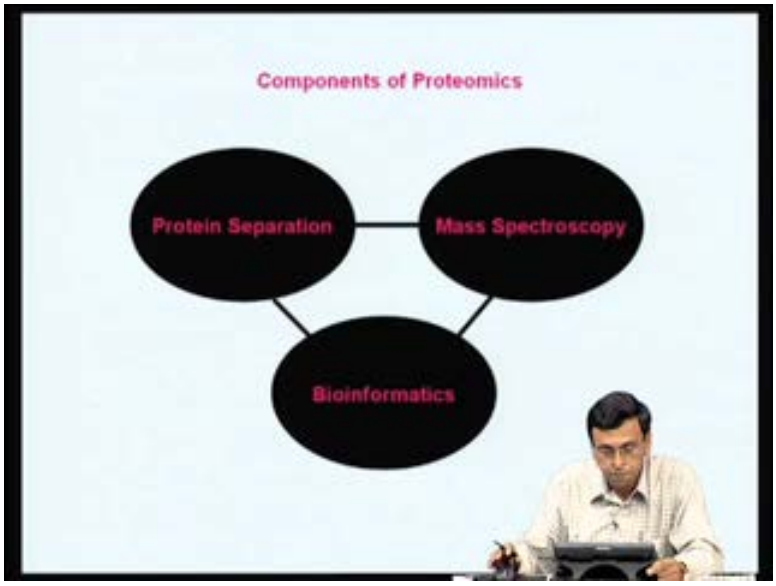
The goal of proteomics at that time is to obtain a more global or integrated view of biology by studying all the proteins of a cell rather than each one individually. This was the goal of proteomics.

(Refer Slide Time: 31:27)



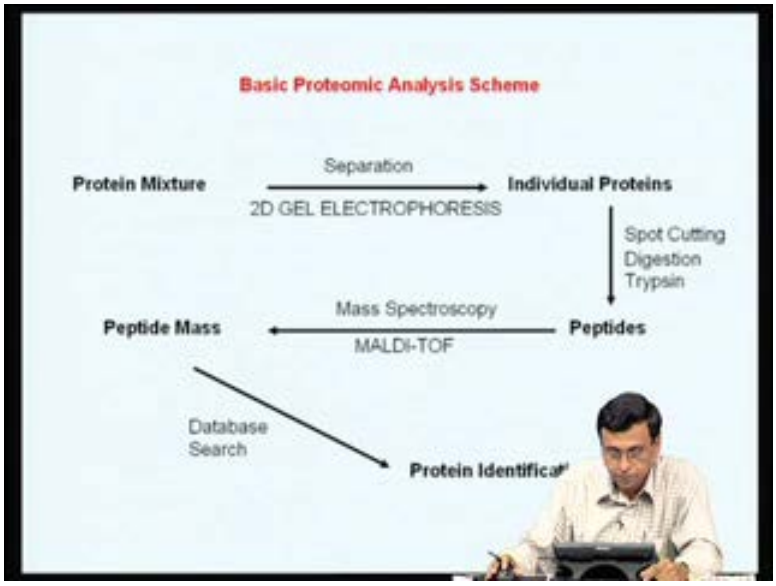
Why was proteomics necessary as I just explained to you, I have now given a few more reasons. Having complete sequences of genome is not sufficient to elucidate biological function. So, it became clear just by sequencing genome after genome at the end just looking at the genome organization, you cannot really understand what is going on inside the cell because ultimately you need to understand how these genes are getting transcribed and how the proteins are being made and how these proteins function at the cell. Now, a cell is normally dependent upon several metabolic and regulatory pathways for its survival and modifications of proteins can be determined only by (( )) proteomic methodologies. So, many at times how these metabolic or regulatory pathways are regulated, ultimately depends on proteins and in our previous classes, we have clearly understood how various signal transaction pathways function and how protein-protein interactions and post translation modifications of proteins play very-very important role in various metabolic reactions as well as in various signal transductions pathways. So, it is necessary to determine the protein expression levels, post translational modifications, protein localization as well as protein interactions, and only then we can have get a comprehensive view of how exactly the genome is functioning and how the proteins which are transcribed from the genome actually are able to function your cellular processes and remember mane times it is ultimately the proteins, which are the drug targets. So, unless you understand and exactly how the protein is functioning, you cannot really develop novel drugs or they cannot be drug targets. So, understanding this proteins and how global what kind of changes takes place in the protein level became very-very important.

(Refer Slide Time: 33:06)



So, what are the components of proteomics? So, if you want to look at very rapidly what kind of changes are taking place at the protein level, how do you go about. The key components involved in (( )) first you need to have a very good protein separation technology then you need to have what is called as a mass spectrophotometry and of course, we need to have a huge computing power to interrupt with the data bioinformatics. So, proteomics combines protein separation techniques with mass spectrophotometry and bioinformatics and together you can understand what kind of global changes are taking place at the proteome level.

(Refer Slide Time: 33:42)



So, I will just briefly schematically described here, although the whole (( )) complex, I tried to simplify the whole thing, how exactly go about doing proteomics. This is the basic proteomic analysis scheme that would be here. So, suppose we have a protein mixture, let us say I have proteins from a liver extract. I first have to separate these proteins, now normally we all do what is called as a sodium dodicil sulfate polyacrylamide gel electrophoresis, but unfortunately the S D S page is primarily based on the separation of proteins based on their mass because you neutralized the charge, all the proteins become as a negatively charged and they become reduced because you are adding S D S and beta-mercaptoethanol and many at times when you separate proteins on S D S page you cannot separate proteins based on their charge,. Suppose, there is a protein which is differentially phosphorylated or different acetylated, you cannot separate these two proteins because their molecule weight is still more or less or the same and the find difference between the molecular weight cannot be resolved on this one dimensional S D S page. So, if you want to do actually proteomics, you need to combine this S D S polyacrylamide gel electrophoresis with one another methodology that is actually based on separating proteins based on their charge. So, what you first you do is, when you have a mixture of proteins you first separate this proteins by what is called as a two dimensional gel electrophoresis where the individual proteins are first separated by what is called as isoelectric focusing where the

separating is based on the charge and once the protein got separated based on the charge, then you separate the proteins based on their mass or even what is called as a S D S page.

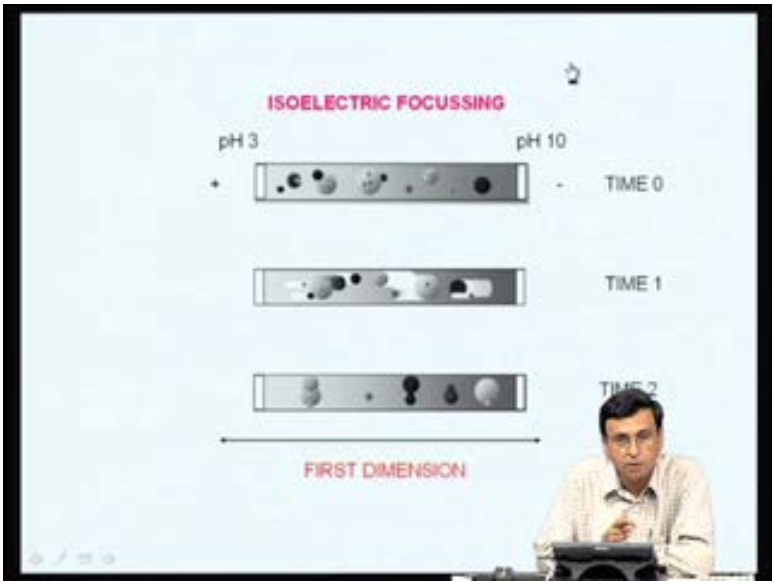
So, by using isoelectric focusing and polyacrylamide gel electrophoresis you can resolve a number of portions much better than either of them alone. So, this combination of these two techniques is actually called as two dimensional gel electrophoresis or two d g e. So, combination isoelectric focusing combine with polyacrylamide gel electrophoresis. So, once you have resolved the individual proteins based on the charge as well as was mass you can now take individual parts individual proteins and break them in to peptides using enzymes like trips in and then this peptides is fed in to an instrument called mass spectrophotometer mass spectrometer we will not go into the function of mass spectro meter because that is a total different thing, but basically in the mass spectrometer the protein gets ionized and the mass of each one of this peptides is analyzed.

And. So, what ultimately you have that data that comes out of mass spectrometer is the mass of various peptides that have been generated for a particular protein by the trypsin digestion. So, you have a series of a mass of peptides depending upon the number of trips in cleavage particular protein you get a whole mixture of peptides with different masses now there are actually databases what they very populated database is called as mascot where there predictor or a theoretical peptides masses of all known proteins are available.

So, you take the protein peptide mass that you have got from your experiment and then if you now feed into the database it will now compare the data generated by your experiment with the predicted peptide masses of all the known proteins and then match and then tell you this is the likely protein that is present in your sample. So, it will identify protein based on this kind of a comparison.

So, this is basically is the basic proteomic analysis that one carries out. So, i like to spend couple of minutes understand how exactly two dimensional electrophoresis because that is the key technique. So, as I told you the two dimension gel electrophorus basically is done first by Isoelectric focusing followed by S D S page.

(Refer Slide Time: 37:20)

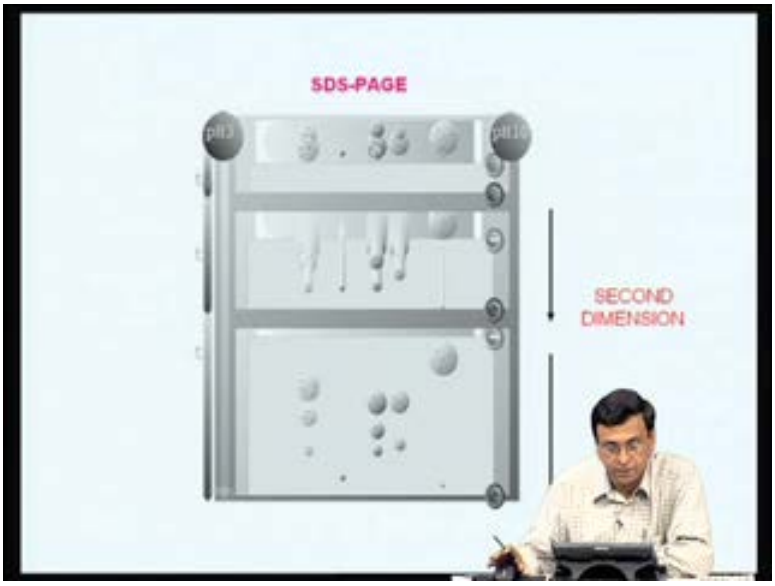


Now in the Isoelectric focusing basically you have a gradient from various page that can ph-3 to ph-10 or if you want to separate proteins at narrow range you have p h strips where you can start from p h seven to p h ten and. So, on. So, forth

So, depending upon the P I that you are interested in you can use different kinds of P H steps now one then you apply your protein samples and pass current from negative to positive charge now time zero before you apply the current the proteins will be all over because they are under various P I. Now, once the current is supplied the proteins start migrating under the electric field and they will stop migrating when they reach their the Isoelectric point.

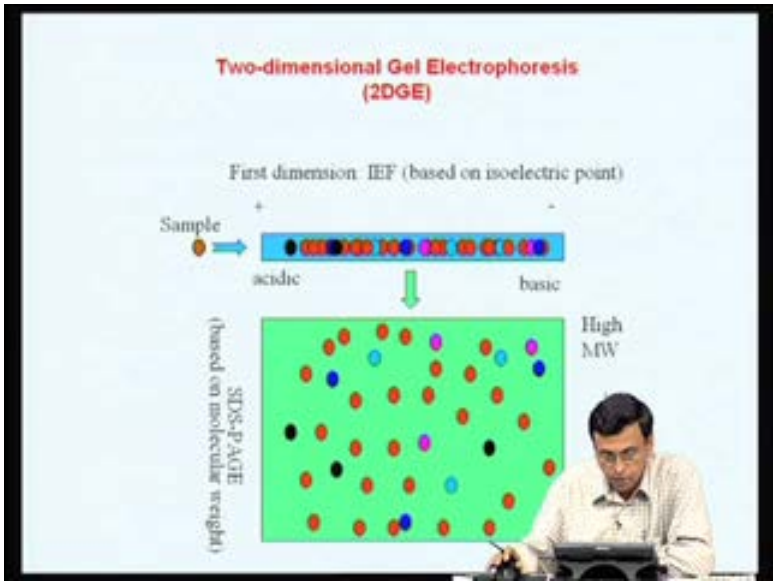
So, here the proteins get selected Isoelectric focusing depending up on their P I and you get various proteins that gets separated based on their P I. So, these actually as called as a first dimension electrophoreses

(Refer Slide Time: 38:20)



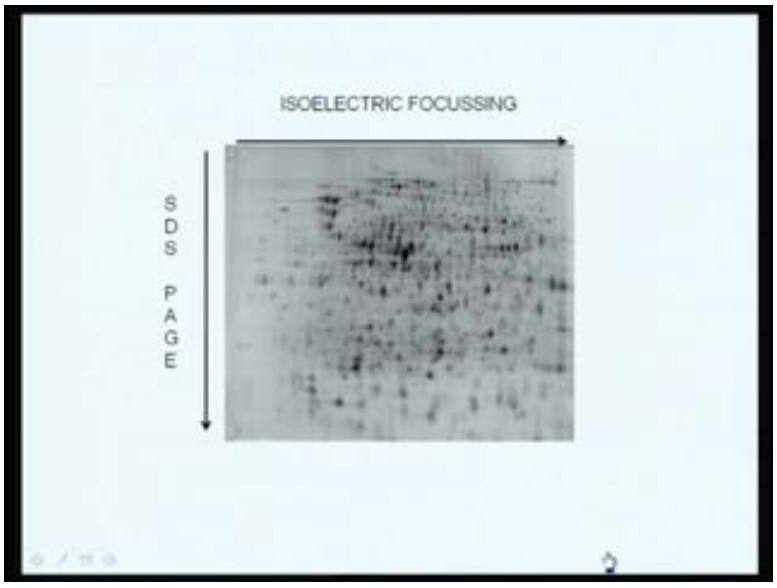
where proteins have been separated based on their p I now once you have done that you now take this strip and then now subject it to S D S electrophoreses where the now the proteins which have been resolved based on their P I now they move in their electric field and now they get a separated based on their mass high molecular proteins gets move less migrated less and low molecular proteins move further and so, ultimately what to you end up at the end of this two dimensional gel electrophoreses is proteins which have been resolved based on their charge as well as the mass.

(Refer Slide Time: 38:48)



So, this is the schematic of how exactly you get a mixture of proteins which have now been first separated on the I E F based on their charge and now when you subjected to the second dimension is this electrophoresis get resolved based on their mass with high molecular proteins at the top and low molecular proteins move in the farthest.

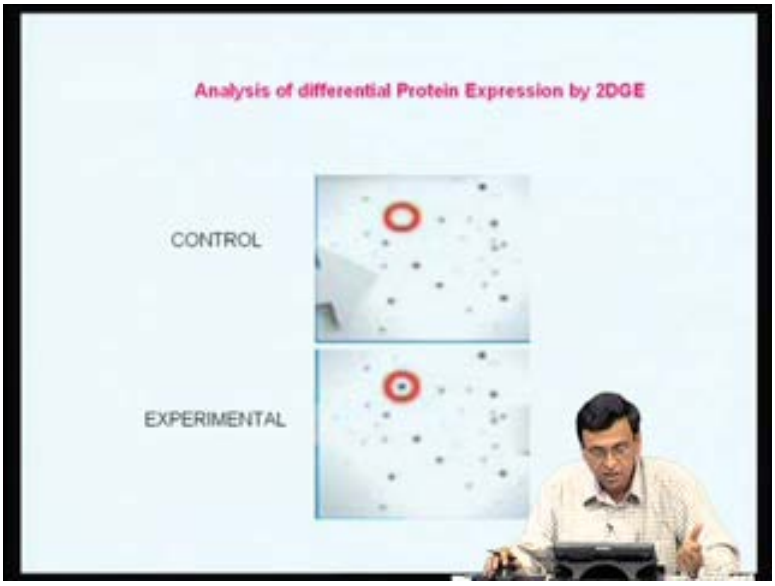
(Refer Slide Time: 39:09)



This is how a basically two d gel looks like you can either stain this gel by coomassie blue or you can do silver staining or if you have used the radioactive proteins you can do auto radical proteins and then pick up the radioactive parts.

So, you can see each dot here represents a particular protein and. So, once you have separated the proteins now how do you identify this proteins these proteins are actually identified, here is for example,

(Refer Slide Time: 39:32)



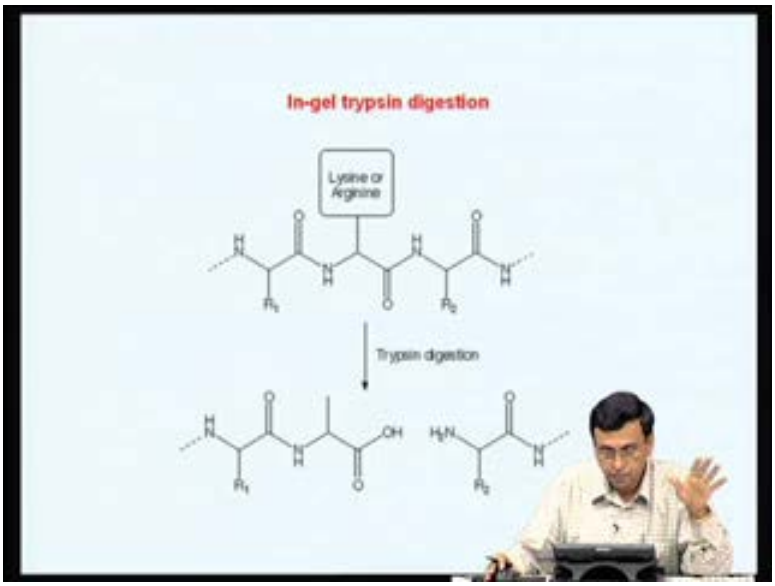
I have given you one example of what did you do by two dimensional electrophoreses let us say for example, this is the protein mixture from a control sample and is the protein mixture from a experimental sample and as a mention in the case of micro array it can be any a two different two physiological situations control means stimuli unstimulated or stimulated it can be quiescence cell verses poly plating cell without ligen with ligen any two different physiological conditions.

And you can see here all the proteins are more or less same, but you can see one particular protein which is now present in the experimental sample, but is absent in the controlling sample. So, I now make conclusion that here is particular protein which is getting induced which is present only in this experimental conditions, but not on the controlled conditions now you want to identify what is protein is, and you can see the power of this technique you have not purified

any protein you have not separated any proteins you have simply taken a whole extract or the entire protein complement of a cell or a tissue and under two different conditions you separated all this proteins and we have already see there are differences between the two.

There are protein profile which is different now let us see how you proceed further. So, once proteins are separated by two dimensional electrophoresesis now how do identify the protein. So, for example, in this case now I want to now see what is this protein that is actually appearing only in this particular condition, but is absent in under this condition what do I do, I now excise this gel P which contains this particular protein sample and then subject to what is called as a trips in digestion followed by mass spectrometry.

(Refer Slide Time: 41:06)

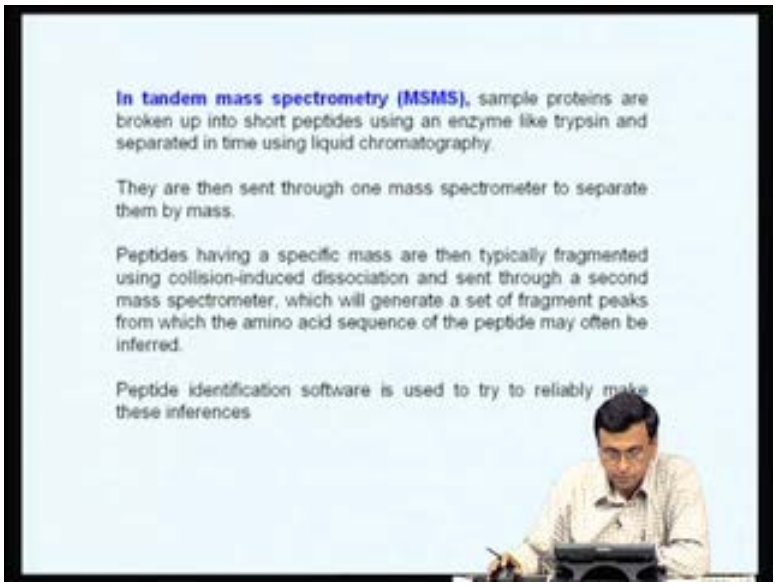


So, as you know strips in always cuts after a basic amino acid either lysine or arginine. So, it take this gel p containing tryps in and subjected to what is called as a in gel or a in gel tryps in digestion and then you take this peptides and then put it inside a mass spectrometer. So, these peptides when they are introduced a mass spectrophotometer as I said I explained earlier these peptides will be ionized and the mass of each one of these peptides is determine by the machine and the values that are coming out of the machine can now be fed into a database and this is what is called as a peptide finger print.

(Refer Slide Time: 41:42)

Now, So, depending upon the mass of the peptide that are generated and comparing the mass of this peptides with the known databases you can actually identify why this proteinates. Now there is one more advanced technique it is

(Refer Slide Time: 43:43)



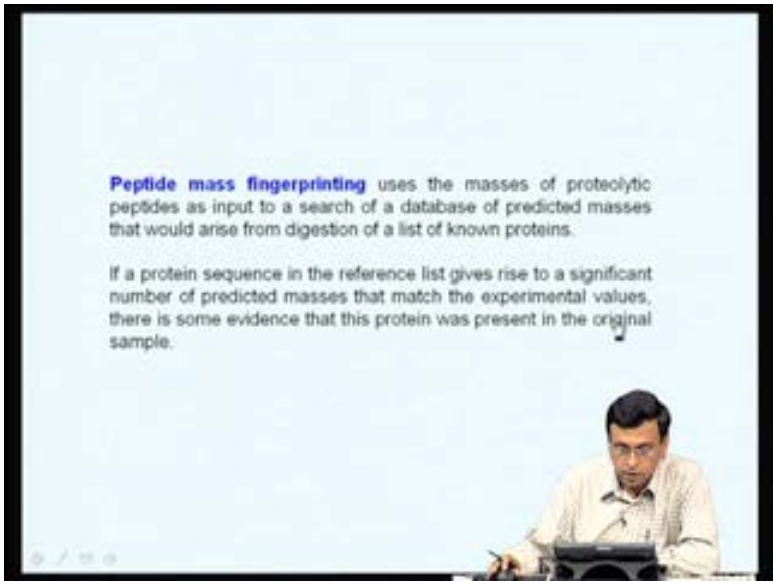
what is called as a tandem mass spectrometry or M S M S wherein once the peptides are have been generated and one peptides have been characterized and you can now ask the machine

again to take each any particular peptides that you are interested in and this peptides again in subjected to one more mass spectrometry wherein these amino acids are cleaved from the peptides and these amino acids again ionized and again based on the mass one can now identify exactly what is the amino acid sequence of that particular peptides.

So, at doing two mass spectrometer thrice twice therefore, it is called M S M S or tandem mass spectrometry. So, first you are looking at the mass of peptides then you can pick up peptide of your interest again subjected to the mass spectrometry and you can identify what is actually amino acid sequence of these particular peptides. So, it is very powerful tool mass spectrophotometer became a very powerful tool wherein you can actually identify a particular protein from a crude sample from a tissue or a cell extract without any purification.

So, this approach is actually called as a bottom approach proteomics where in proteins are identified at the peptide level.

(Refer Slide Time: 43:01)

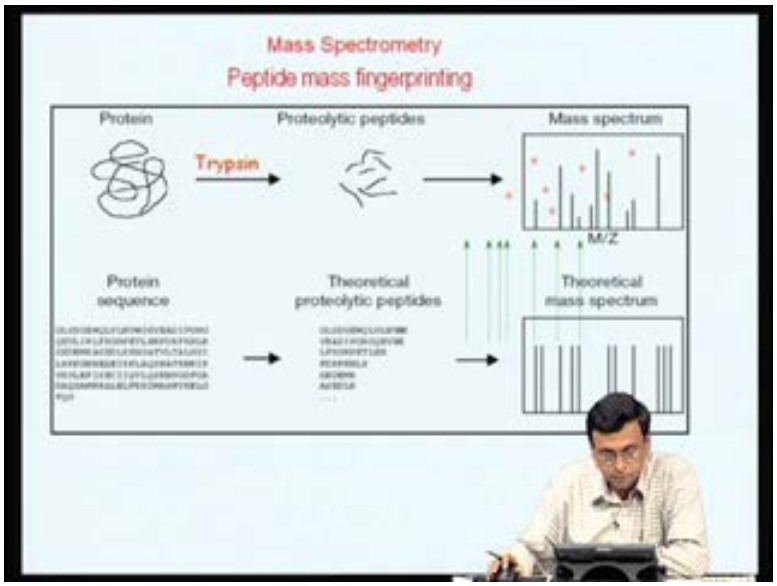


So, peptides mass finger printing actually uses the masses of the proteclytic peptides as input to search a database of predicted masses that would arise from digestion of a list of known proteins this is what I explained just now. So, if a protein sequence in the reference list gives rise to a significant number of predicted masses that match the experimental values there is some

evidences that this proteins was actually present in the original sample. So, the peptides which are generated the masses of the various peptides you have generated experimentally you now compare against the predicted peptide digestion of all known peptides and if these peptides fragmentation matches then you can reasonably confident that you are your protein is likely to be what is depicted from the database. So, in the tandem mass spectrometry or M S M S the sample proteins are first broken into short peptides using trypsin separated using chromatography and then they are sent through one mass spectrometer to separate them by mass and then peptides once they are separated by mass peptides having a specific mass are then again typical fragmented using collision induced dissociation and sent through a second mass spectrometer mass spectrometer which will generate a set of fragment peaks from which amino acid sequence of the peptide may be inferred. So, by doing mass spectrometer twice one first to generate the peptides mass then actually to generate the amino acid mass and by comparing these things you can actually identify using appropriate software you can actually identify the sequence of the peptides and then infer what the protein is.

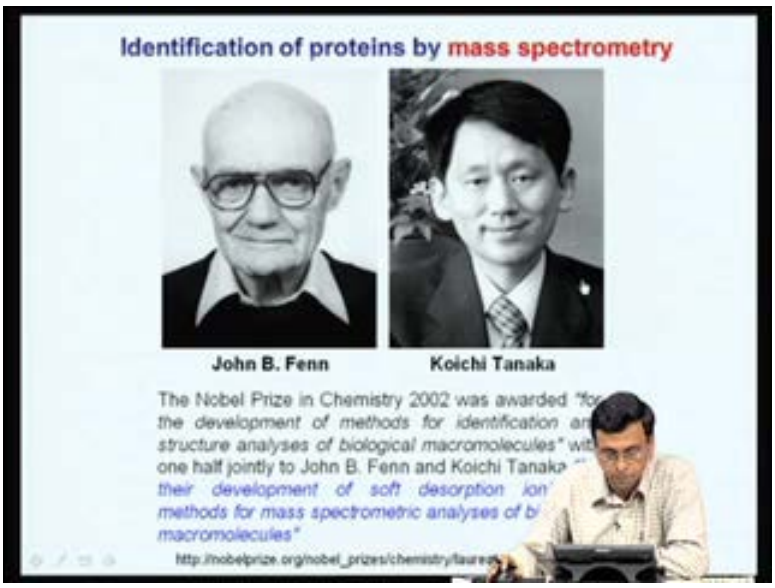
So, instead of doing this is say mass spectrometry became a very powerful tool to identify proteins from crude samples and became a very good alternative for the classical amino acid sequencing or protein sequencing using admin degradation. So, on and. So, forth.

(Refer Slide Time: 44:45)



So, this is the schematic I have shown here again I will not repeat basically take a proteins subject to trypsin digestion take the protein peptides put in the mass spectrometer you get various peptides and then you take this peptides and you have databases of a predicted protein sequences which are where the predicted tryptic peptide masses are given by comparing with the theocratic appropriable we will identify what is protein is likely to be.

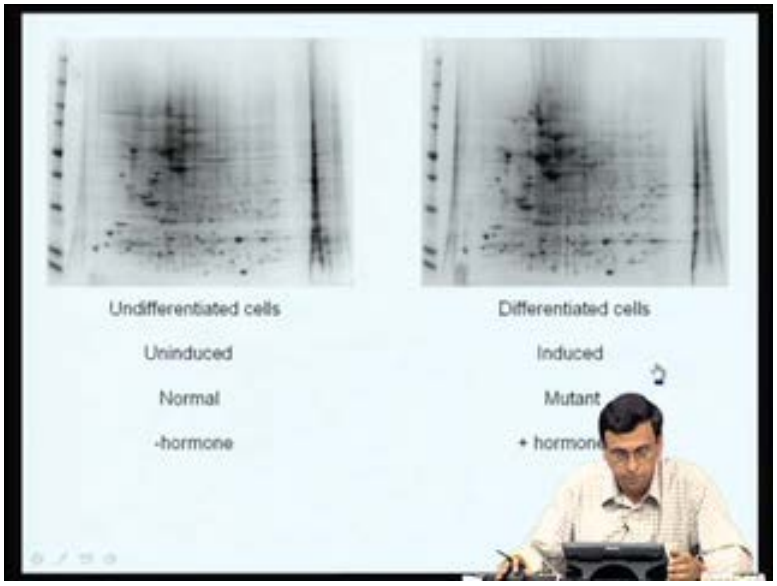
(Refer Slide Time: 45:08)



So, mass spectrometry became such a powerful tool that the people who actually developed these techniques john fen and koichi tanaka were actually awarded nobel prize in the 2002 the citation actually says that the nobel prize in chemistry two thousand two was awarded for the development of methods for identification and structural analyses of biological macromolecules with one half jointly to john fenn and koichi tanaka for their development of soft desorption ionization methods for mass spectrometric analyses of biological macromolecules.

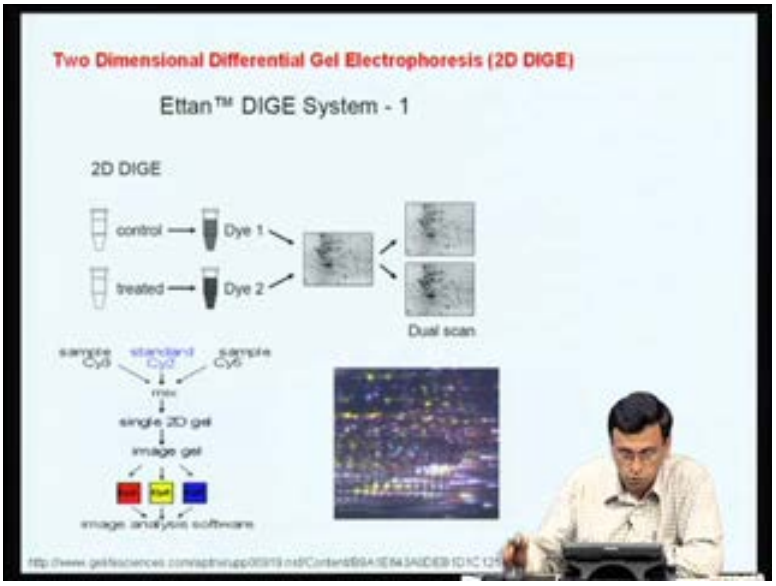
So, the mass spectrometry became a very powerful tool for identification characterization of protein molecules and it became a fantastic tool to understand how proteins vary from one particular situation to another situation become excellent to normal verses genome type or different external conditions to identify proteins by using this kind of a mass spectrometric techniques.

(Refer Slide Time: 46:08)



So, here is for shown you couple of examples you can actually compare the proteome of undifferentiated cell versus a differentiated cells or you can compare the two proteomes are of un-induced verses induced it can it can be hormone it can be growth factor and so on so forth, or it can be normal versus mutant wherein transcription factor or a significance specific component of significant transcription pathway has been mutated now I want to show you want to have find out what kind of proteins are getting by this particular mutations, or it can be minus hormone or plus hormone you can see what kind of proteins are getting synthesized when you had a particular hormone sky is the limit you can un you can compare any two physiological situations and get the variations in the gel.

(Refer Slide Time: 46:47)

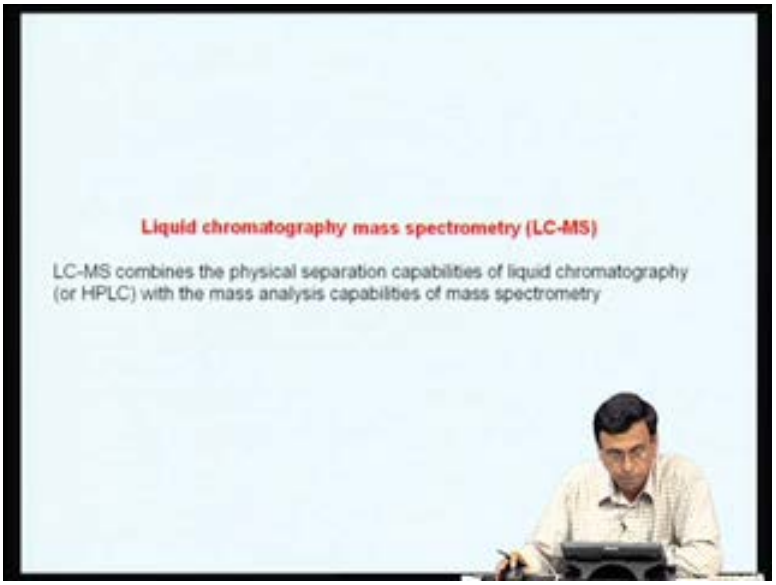


And. So, for I have talked about analyzing in one gel only one particular condition. So, if you want to compare two different conditions you have to run two gels now you have what is called as a two dimensional differential gel electrophoresis or two D I G E wherein here you do not have two types sub two gels you can mix both this you can take the protein from proteins from two condition and mix them together and you can get by running a single gel you can infer the data this is now similar to what we did in the case of micro array where you label the R N A from condition with green dye another condition from a red dye mix them together hybrids to micro wave.

Same way here you can now take the proteins and label them with a dye one which flows as differently and take the other condition two and label those proteins at the another dye which flows differently mix them together run the gel and then you can see identify these differentially layer clotting proteins. So, the two dimensional gel electrophoresis or two D I G E you do not have to mix run two different gels you can run by a running a single gel you can find out the difference in the proteomic factor and mangier times now you can also induce a standard which should not change basically tells you that your experiment is alright, so you have actually red colored red label proteins blue label proteins and when you mix them together you get a come down this is how a typical gel looks like you have raw blue dots, red dots and yellow dots so on so.

So, by running a single gel you can get information about differential expression of the proteins .

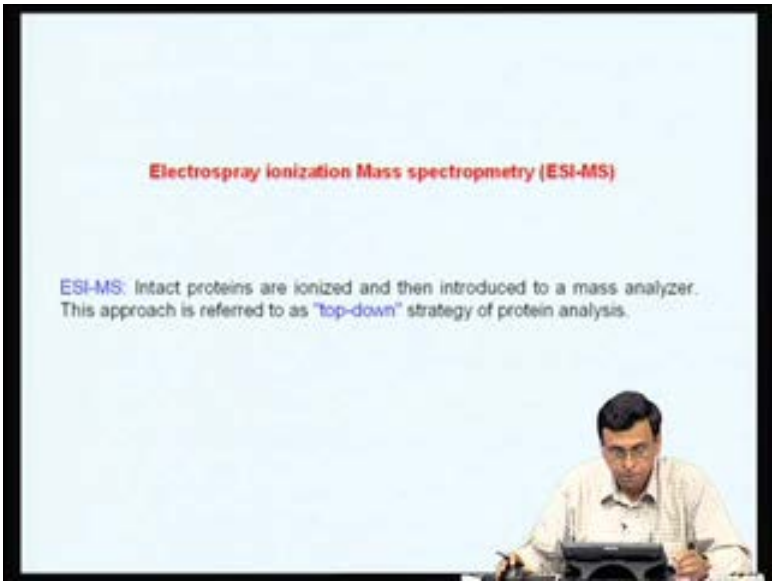
(Refer Slide Time: 48:18)



Now there are a number of variations of this proteomic techniques that are now available what i told you only about molding or mass spectrometry based on molding or m s m s you have what is called liquid chromatography mass spectrometry or L C M S where instead of separating the protein on a two dimensional gel by I E F and S D page you can separate the proteins based on H P L C by liquid chromatography system and then this proteins are fed into mass spectrometry.

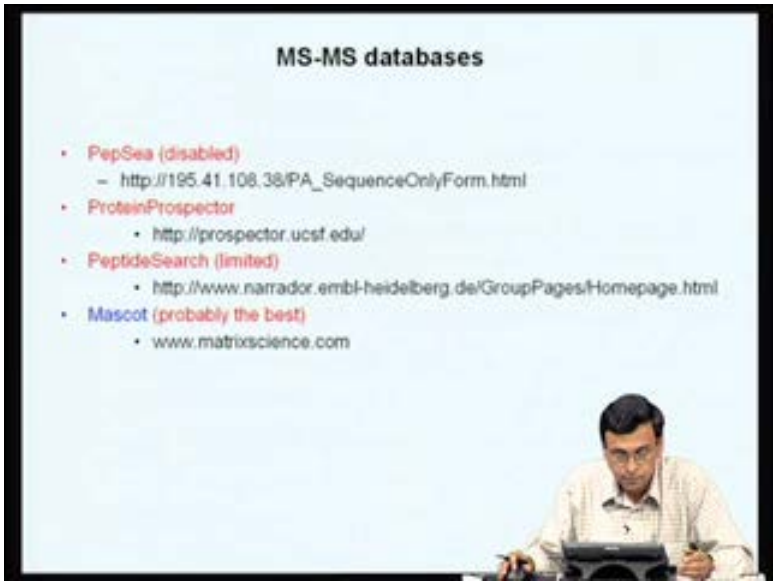
So, combining the mass spectrometry with the liquid chromatography technique is known as L C M S. So, either can separate protein based on S D S page or you can separate proteins based on liquid chromatography systems and ultimately you can combine them with mass spectrometry and get the data.

(Refer Slide Time: 49:04)



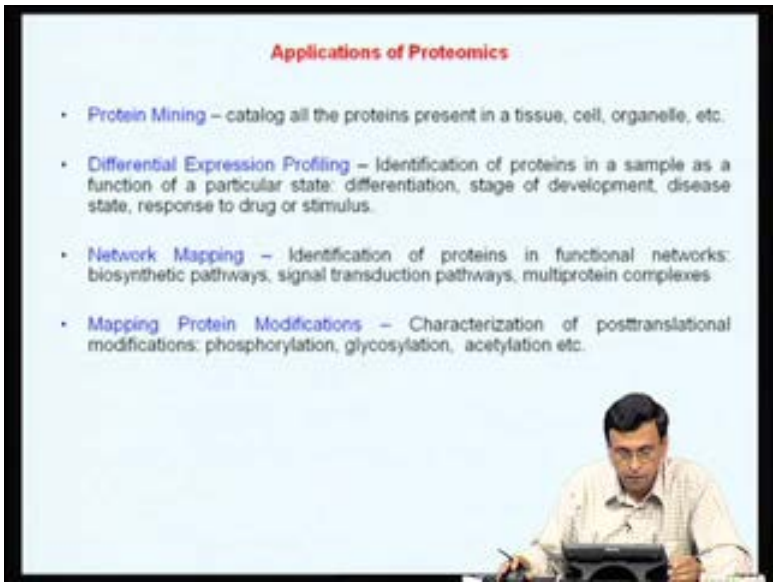
This is what called as a electro spray ionization mass spectrometry or E S I M S again where intact proteins are ionized and then introduced to a mass analyzer and this approach is referred to as top down approach for the strategy of protein analysis. There are number of databases now available, the most widely used database for that is once you generate the peptide finger print it means once you got masses ways peptides you take this and then go to this databases and feed this numbers into these databases and the software will now compare your mass values with the predicted peptic patterns of all known proteins and give you the identity of the protein. So, number of databases is now available.

(Refer Slide Time: 49:17)



So, what are the applications of proteomics I am just going to give few examples just I am I have mentioned about the case of genomics you can give proteomic for protein mining that is you can catalog all proteins which are present in tissues cell or organelle. So, just as you have a gene expression signature using D N A micro array you can get the protein of signature what all the proteins are given signature in given cell or a given tissue by using proteomics.

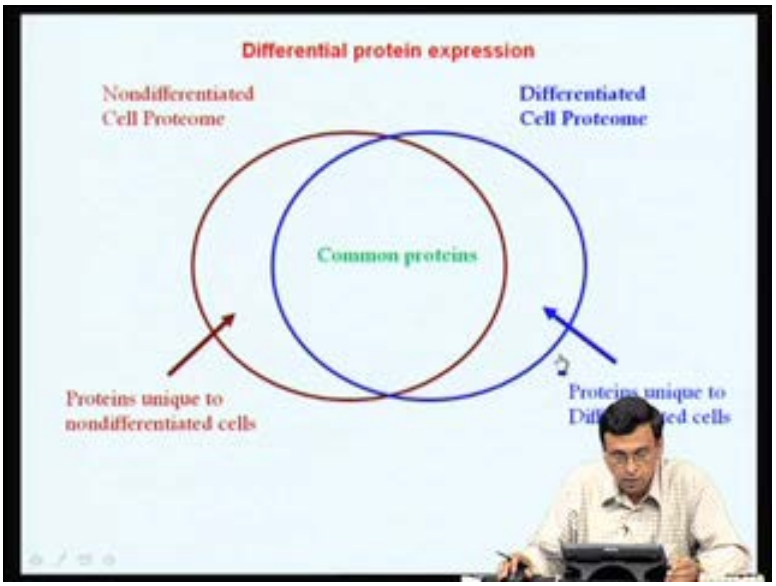
(Refer Slide Time: 49:46)



And as I mentioned you can do differential expression profiling identification of proteins in a sample as a function of particular state it can be differentiation different stages of development can be compared, disease state normal state versus, disease state response to a drug or stimulus you name an example it can be done by doing proteomics you can do network mapping identification of proteins in functional networks biosynthetic pathways, signal transduction pathways, and multi protein complex.

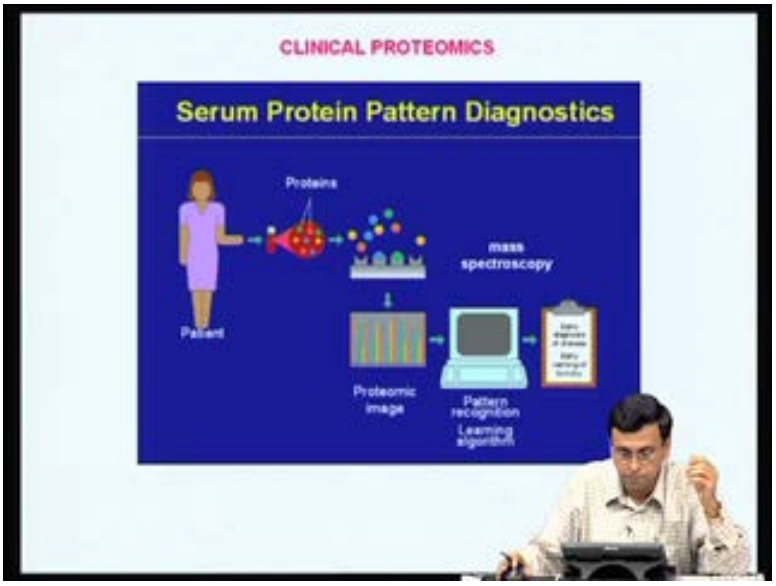
I will just give one example as we go along mapping protein modifications again, when you run a two dimensional electrophoresis you can actually separate the proteins based on their posttranslational modification because the charge will vary. So, you can characterize the post translation modification such as phosphorylation glycosylation acetylation etc.

(Refer Slide Time: 50:55)



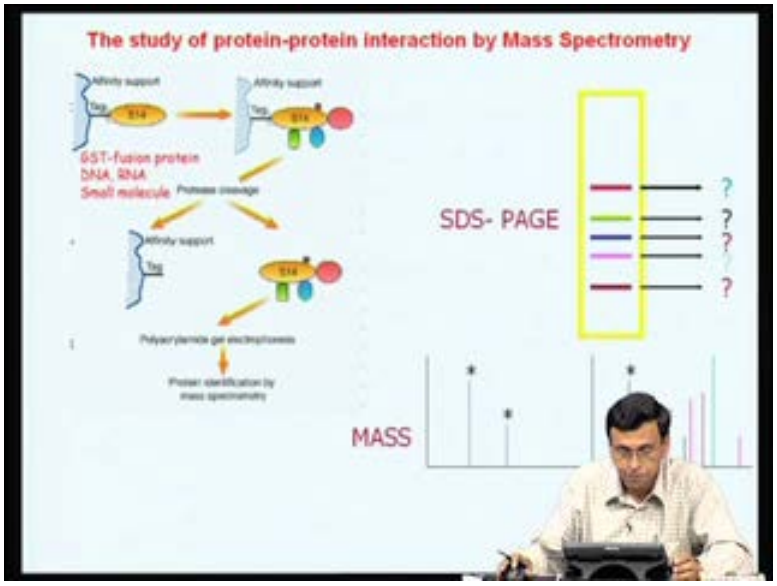
So, I just give an example of a differential protein expression for example, by I by identifying the proteins present in a undifferentiated verses differentiated you can actually put them together and find out which are the proteins which are unique to differentiated cells which are the proteins which are unique to indifferent cells and what are common proteins which are present in both differentiated verses undifferentiated states. So, lot of data can be obtained.

(Refer Slide Time: 51:19)



Clinical proteomics now serum is the most important source for or more easy way of identifying where versatile method of identifying a disease phenotype and suppose there are certain serum proteins which can serve as biomarkers. So, you simply collect the serum and you can feed these entire serum proteins by to a separate them by two dimensional electrophoresis or L C M S subject in the mass spectrometry and then see whether any protein is missing or whether a new protein is appearing and based on that by identifying the protein you can correlate whether it can be lead to any particular disease phenotype. So, clinical proteomic is another very major area that came out of the proteomics understanding this serum protein impact identifying protein mark biomarkers in various serum.

(Refer Slide Time: 52:54)

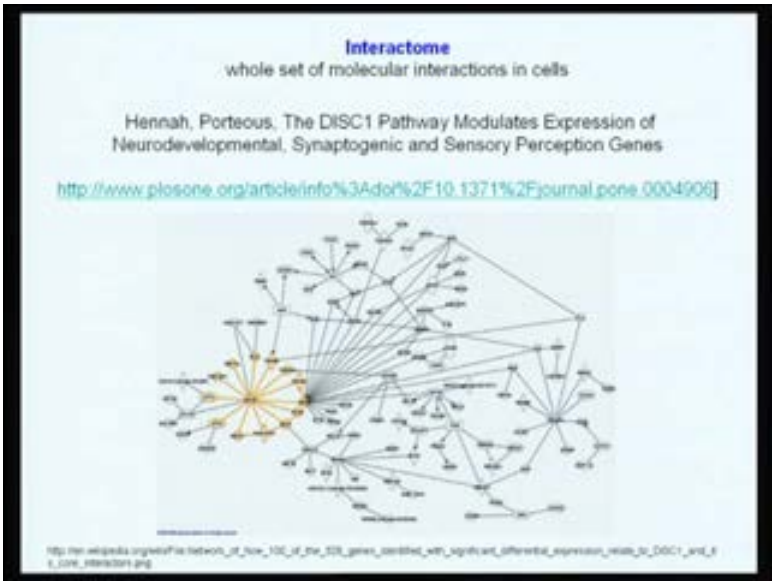


Study of protein interactions as I said ultimately either it be signal transaction pathways or metabolic networks it is the protein interactions ultimately are response.

So, suppose I want low let us a for example, a protein called sfourteen what are the various other proteins which are interact in the s fourteen. So, I can take this s fourteen express it as G S T tags protein couple it to the affinity column and then pass a cell extract through this column and then by using different trenching I can see what kind of proteins are interact in the s fourteen and then I can simply take this again do ah mass spectrometry and identify what are this various means which are interact s fourteen based on this I can develop what is called as a interact home interactive-map.

Interacting network what is called as protein interaction network I can develop and then make some conclusions on how s fourteen interaction with all this proteins can contribute a particular path way or a particular physic phenotype. I have just given you example here how for example,

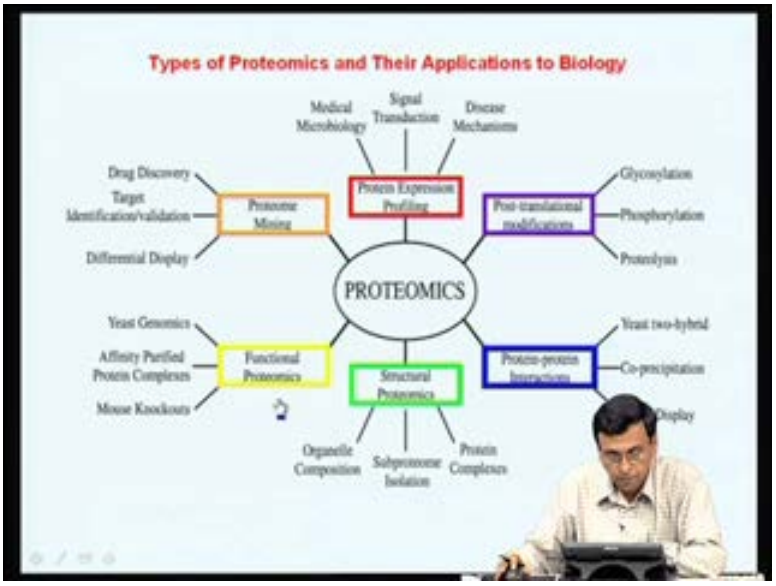
(Refer Slide Time: 53:00)



a protein called as disk one by doing this kind of proteomes what kind of protein proteins are actually intact in disk one and how this is this protein plays a very important role in Psychophrenia

So, by understanding what kind of proteins are actually intact in this one protein one can interfere and what is basis for the Psychophrenia and how this protein may be involved in Psychophrenia. so on so forth.

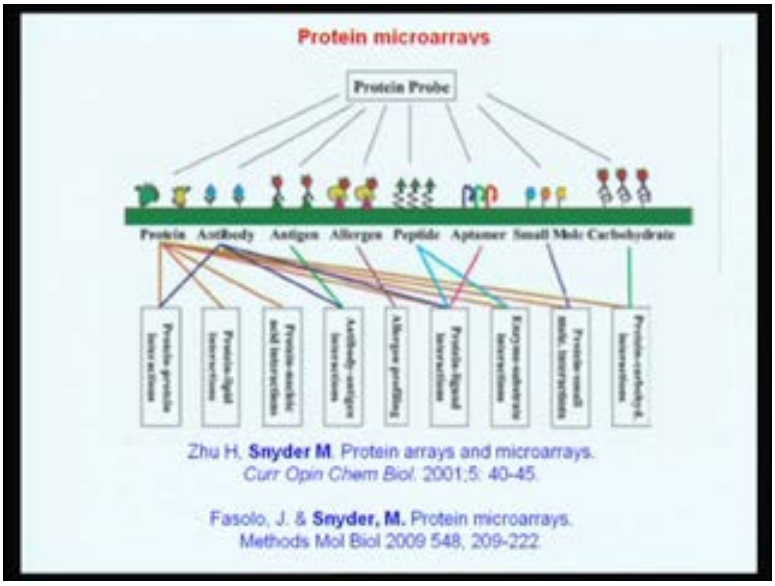
(Refer Slide Time: 53:28)



So, this is just a cartoon that actually tells you various applications of proteomics expression profiling can identify post translate modifications protein interactions structural proteomic functional proteomic proteome mining and so on so forth.

So, following genomics proteomics became a very versatile tool for all understanding how proteins function inside the cells and how protein interactions can ultimately contribute to metabolic transaction pathways. Now, so far you talked about D NA micro arrays now there are coming protein microarrays.

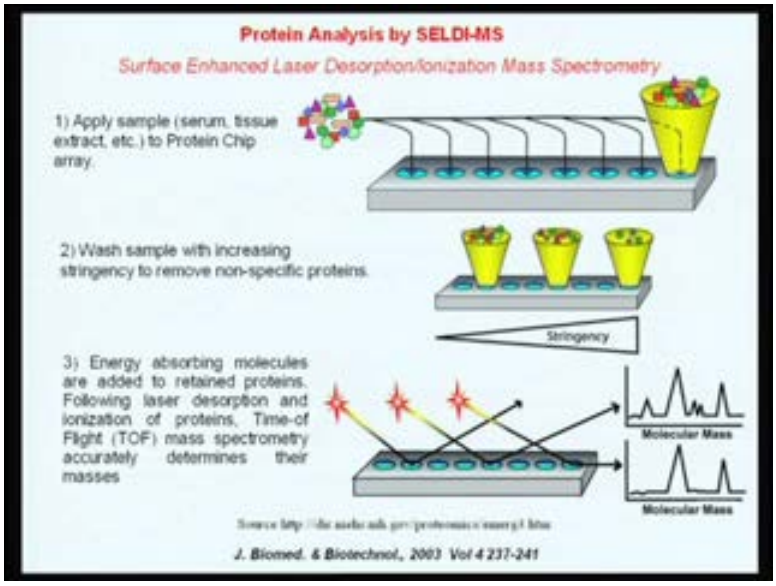
(Refer Slide Time: 54:04)



So, the proteins are actually being spotted on this glass level slides and using this kind of a protein micro arrays one can study antigen antibody interactions one can study protein protein interactions you name it. Small molecule protein interactions and so on so forth. Suppose you identified a protein as a drug target you can simply and suppose you have a chemical library of various compounds you want to see which kind of a drug molecule actually binds with this protein you can simply spot this protein and pass various drug candidate drug molecules and then see which drug molecules actually binds this molecule's highest affinity. So, protein micro is or antigen anti body interactions lien protein intersections so on so forth.

This is an excellent term review article to understand protein micro arrays and their comparison D N A micro arrays in current opinion chemical biology and methods and molecular biology I suggest one can by reading this you can understand more d potential of protein micro arrays. There are newer techniques coming up there is a very sophisticated technique called as surface enhanced laser desorption ionization mass spectrometry or SELDI mass.

(Refer Slide Time: 54:56)



Wherein for example, you can take the protein array and for example, I want to identify serum proteins which are interacting with particular protein. So, I can now apply a serum sample to this protein chip and then by using different wash buffers different stringency I can wash out. So, by increasing stringency as you can see the number of protein molecules interact in this protein chip minimized and you can see what are the molecules who are interacting in the higher stringency versus what are the molecules interact the lowest stringency.

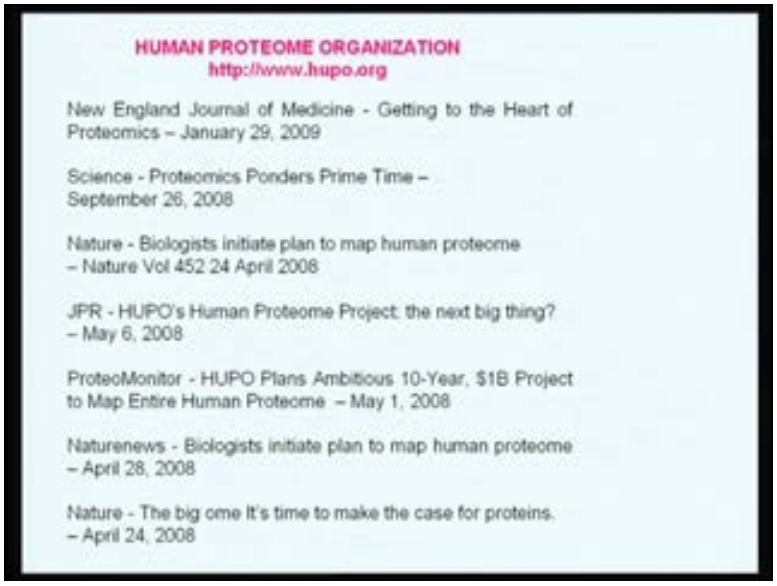
And again I can now take off by actually energy absorbing molecules are added to retained proteins following laser desorption and ionization of proteins at time of flight mass spectrometry you can actually identify you can actually remove this proteins and then generate various peptides of different mass and you can actually identify this proteins who are interacting with the identify what kind of serum proteins are interacting with the protein of your interest.

So, I think what I basically told you is about the potential of D N A micro array and potential of protein arrays how you can do genomics you can do proteomics and what is the potential of genome analysis and proteome analysis how it has contributed to the understanding of various biological phenomena and has really revolutionized the way biology medicine is research is carrying is out biological medicine.

(Refer Slide Time: 56:21)

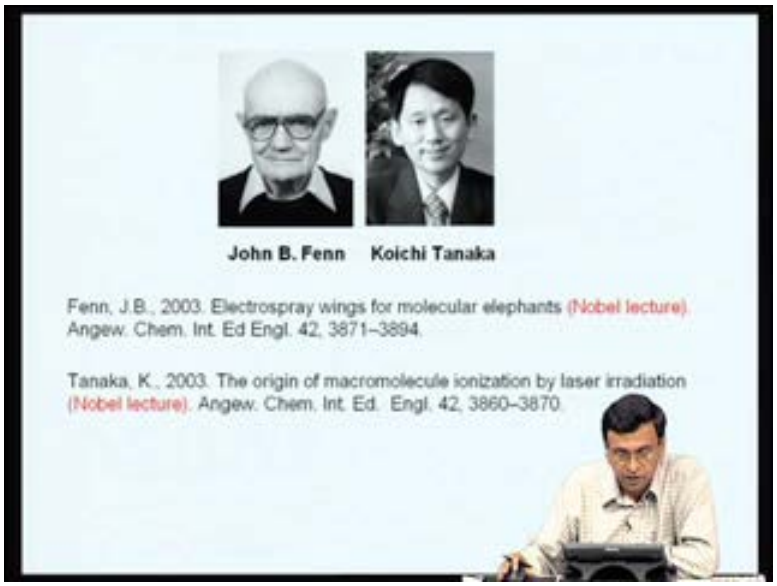
I have given a few references here one can actually go to this some of this websites and gain more information about how D N A micro arrays are being made D N A micros used, affymetrix Stanford university and number of other websites one can visit there is actually now a separate organization called human genome organization Hugo one can go to the web site and you can get fantastic amount of information of how human genome analysis sequencing and genome analysis has contributed to understanding biology and medicine.

(Refer Slide Time: 57:04)



There are very nice article review articles and scientific article proteomics one can actually download this as PDF article by going to this particular website it gives you how proteomics is expected to change the way biology in medicine is research is being done in these areas there are some very nice review articles have listed here just as we have a human organization now we have a human proteome organization and again you can go to the website

(Refer Slide Time: 57:12)



And read some of this articles gives excellent information and ultimately one must be read these to noble lecture by fen and tanaka who explained how the development of the mass spectrometric as revolutionized biology and medicine I think I will stop here

(Refer Slide Time: 57:25)

I think it is last slide where I actually this database called as Kyoto encyclopedia of genes and genomes which actually talks about various metabolic pathways all about two hundred and forty organisms twenty thousand organic specific pathways containing about more than seven hundred thousand genes have been complied together and networks there pathways have been built and very use full to go and then look at some of this pathways compares these various pathways across organisms and so on so forth. I think I will stop here.