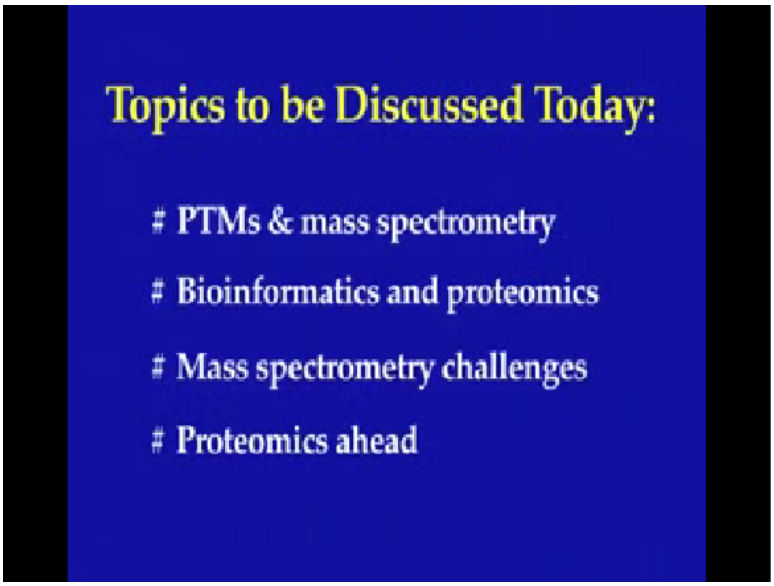
Introduction to Proteomics
Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology - Bombay

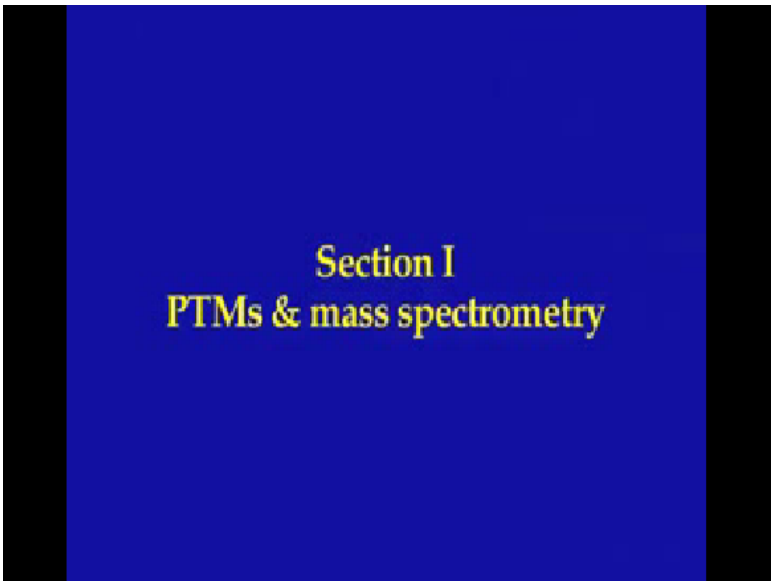
Lecture – 37
Challenges in proteomics

In today's lecture, we will talk about post-translational modifications, structural proteomics, role of bio-informatics challenges and future direction.

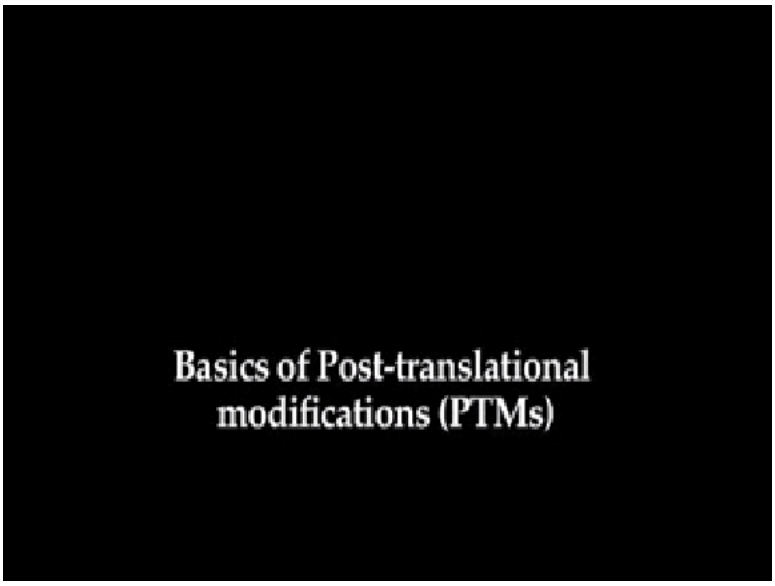
(Refer Slide Time: 00:20)



(Refer Slide Time: 00:33)

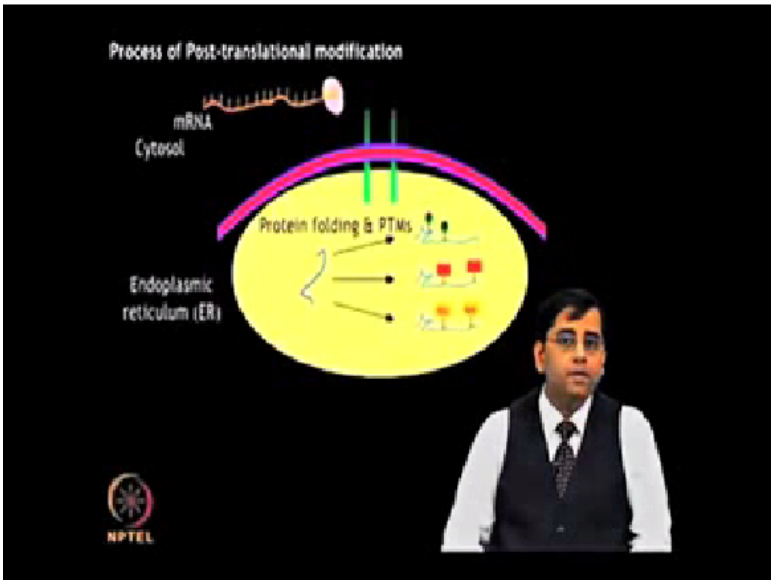


(Refer Slide Time: 00:38)



So let us start with PTMs. Post-translational modifications are vital cellular control mechanisms known as cellular switches that affect protein properties such as protein folding, conformation, activity and functions. As a result, they play very important role in various diseases.

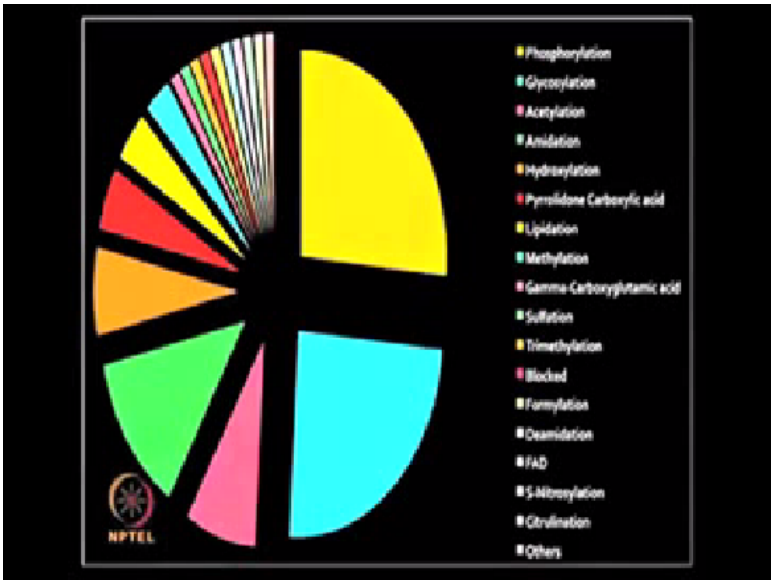
(Refer Slide Time: 01:07)



The protein complexity arises due to gene splicing and post-translational modification. Once the protein is synthesised by the ribosome from its corresponding mRNA in the cytosol, many proteins get directed towards the endoplasmic reticulum for further modification. Certain N and C terminal sequences are often cleaved in the endoplasmic reticulum after which they are modified by various enzymes at specific amino acid residues.

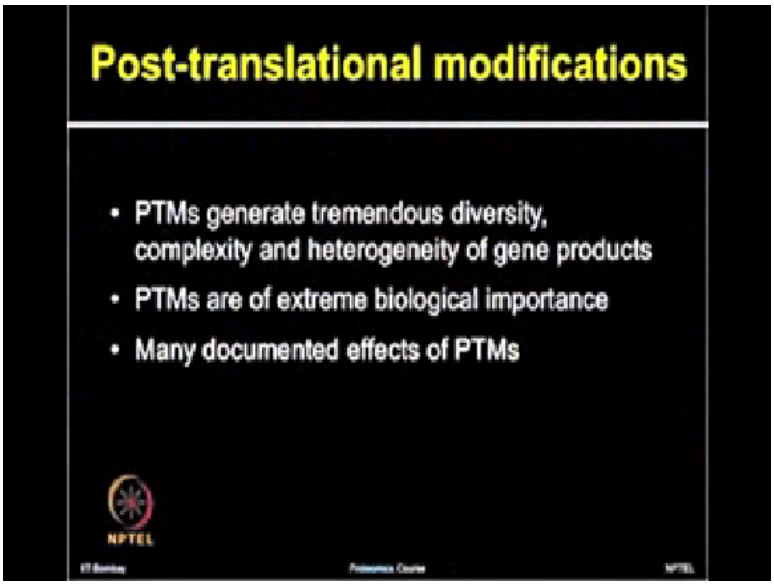
This modified protein then undergoes proper folding to give functional proteins. Due to these modifications, the number of proteins are 3 order of magnitude higher than the total number of genes encoded in genome. There are several types of post-translational modifications that can take place at different amino acid residues.

(Refer Slide Time: 02:22)



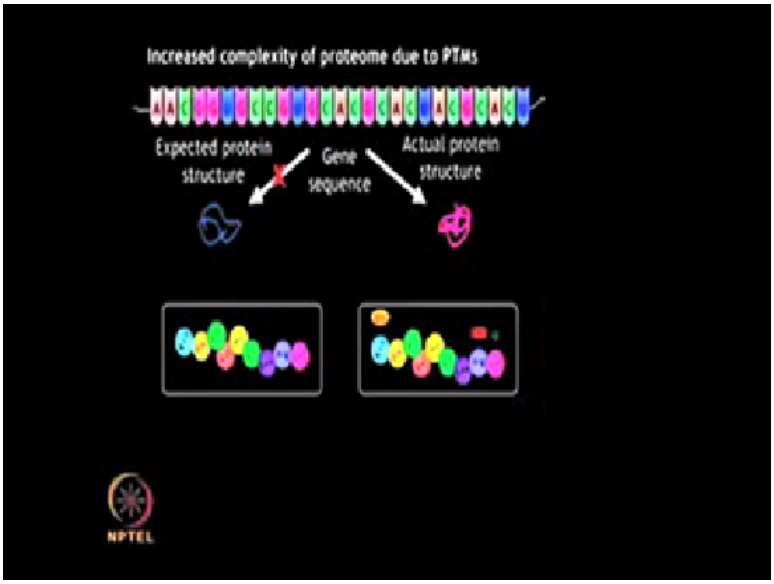
The most commonly observed PTMs include phosphorylation, glycosylation, methylation, as well as hydroxylation and acylation. Many of these modification, particularly phosphorylation serves as regular mechanism for the protein action. PTM generate tremendous diversity and are extremely important.

(Refer Slide Time: 02:54)



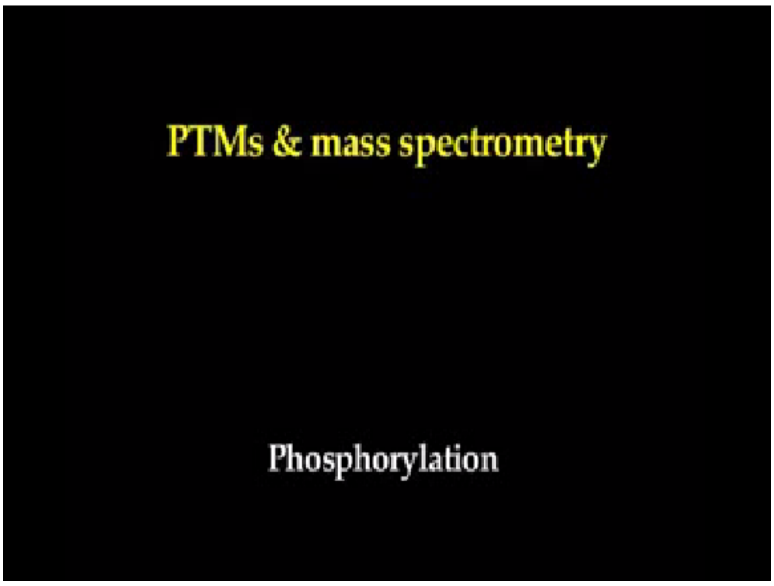
Many documented effects of post-translational modification include change in enzymatic activity, ability to interact with other proteins, sub-cellular localisation and targeted degradation.

(Refer Slide Time: 03:13)



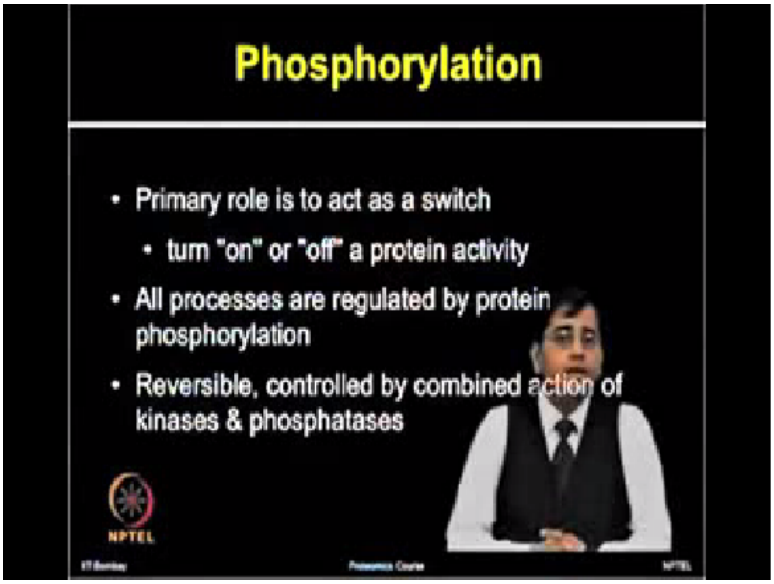
The final structure of functional proteins most often does not correlate directly with the corresponding gene sequence. This is because of PTMs that occur at various amino acid residues in the protein which cause changes interaction between the amino acid side chains thereby modifying the protein structure. It further increases the complexity of the proteome as compared to the genome.

(Refer Slide Time: 03:54)



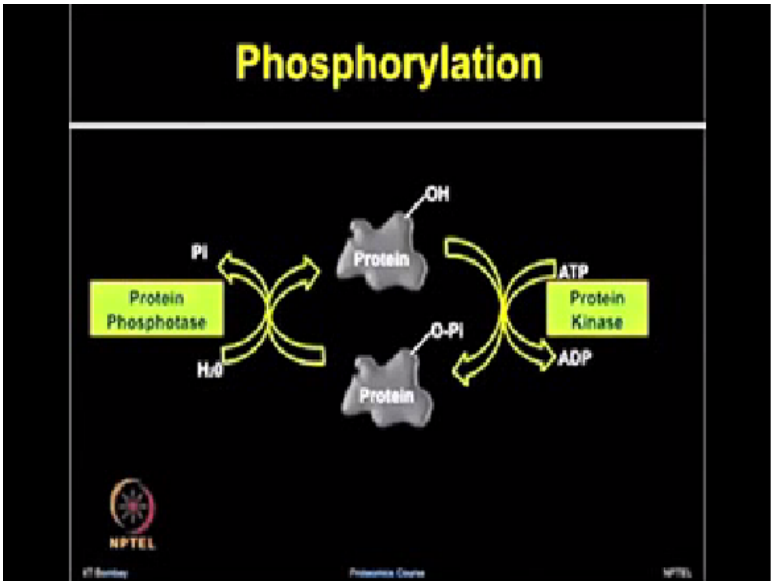
The protein phosphorylation acts as a switch to turn on or turn off the protein activity.

(Refer Slide Time: 04:08)



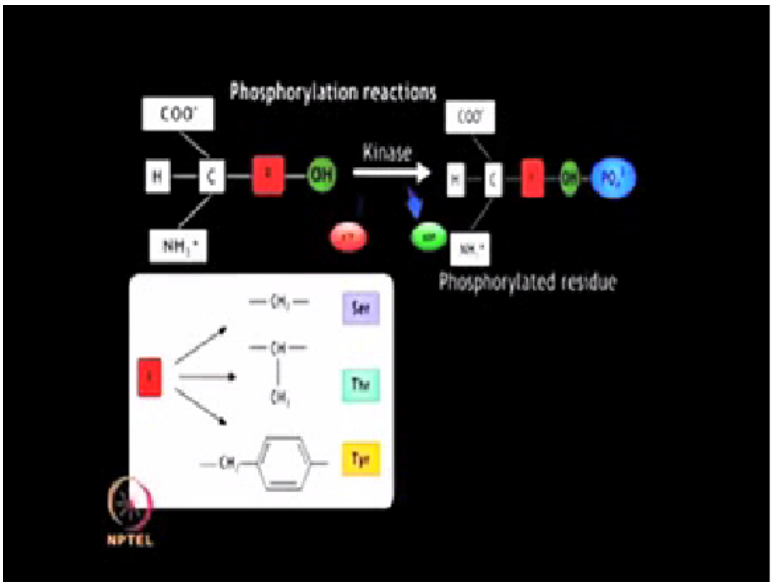
And governs wide range of polypeptides from transcription factors, enzymes, to cell surface receptors.

(Refer Slide Time: 04:19)



The reversible phosphorylation of proteins catalysed by kinesis and phosphatases regulates important cellular functions.

(Refer Slide Time: 04:30)

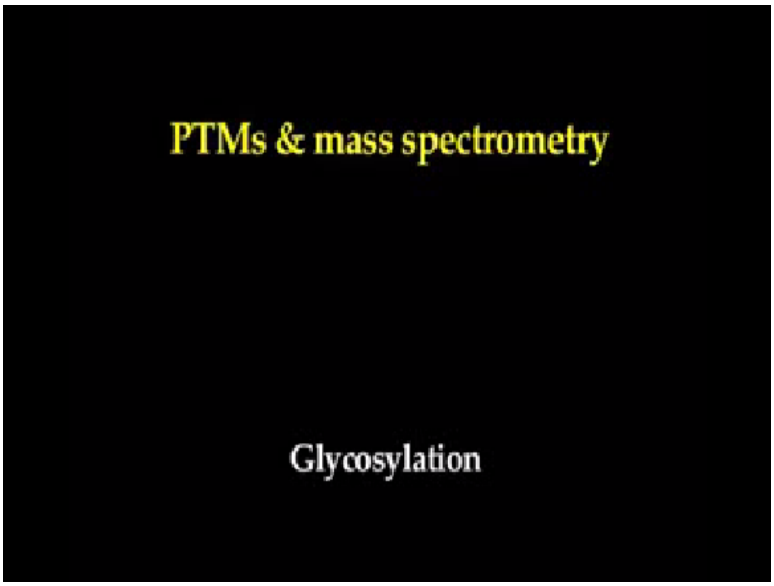


Phosphorylation of amino acid residues is carried out by a class of enzymes known as kinases that most commonly modify side chains of amino acids containing hydroxyl group. Phosphorylation requires the presence of a phosphate donor molecule such a ATP, GTP or other phosphorylated substrates. Serine is the most commonly phosphorylated residue followed by threonine and tyrosine.

The removal of phosphate groups is carried out by phosphatase enzyme and it forms one of the

most important mechanisms for protein regulation.

(Refer Slide Time: 05:25)

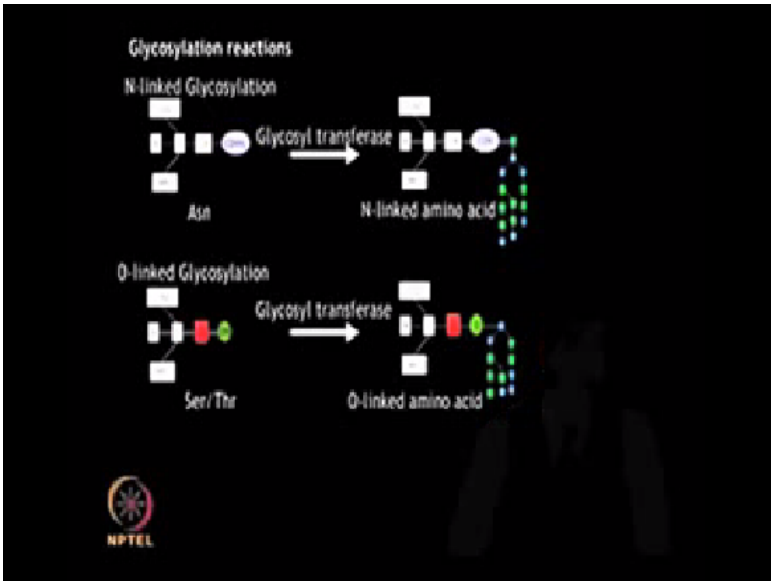


Glycosylation involves linking saccharides to proteins in presence of glycosyltransferases enzymes giving rise to a glycoprotein.

(Refer Slide Time: 05:44)

Glycosylation play vital role in various biological functions such as antigenicity of immunological molecules, cell division, protein targeting stability and interactions. The aberrant glycosylation forms research into various human congenital disorders.

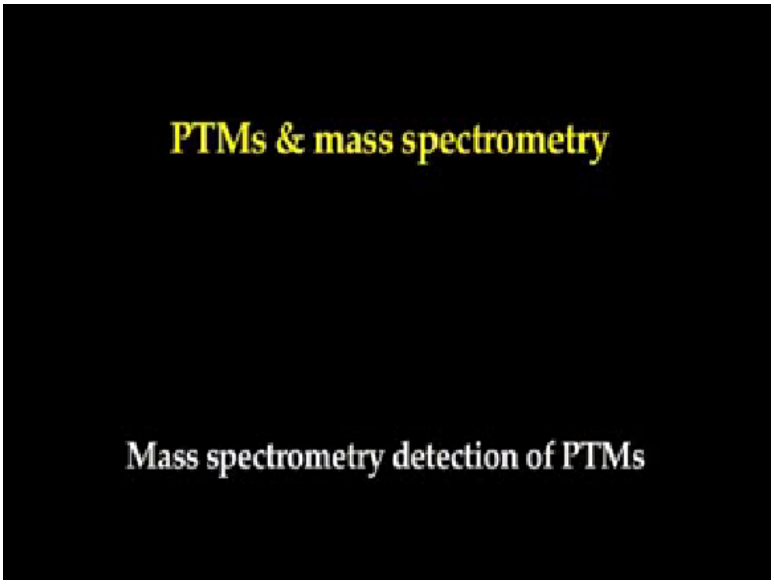
(Refer Slide Time: 06:12)



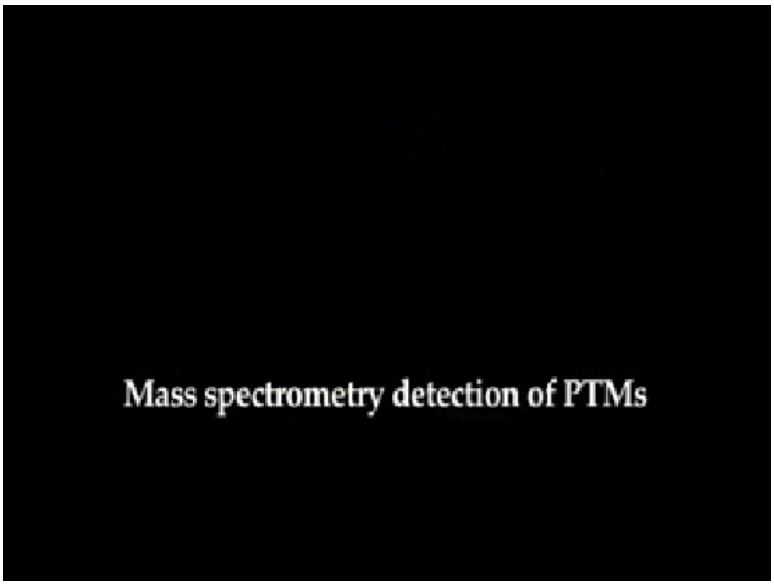
Depending on the linkage between the amino acid and the sugar moiety. There are 4 types of glycosylation, N-linked glycosylation, O-linked glycosylation, C-mannosylation and glycophosphatidylinositol anchored GPI attachments. Glycosylation involves the enzymatic addition of saccharide molecules to amino acid side chains.

This can be of 2 types, N-linked glycosylation which links sugar residues to the amide group of asparagine and O-linked glycosylation which links the sugar moieties to the hydroxyl group of serine or threonine. The glycosyltransferase enzymes catalyse these reactions. Sugar residues that are attached most commonly include galactose, mannose, glucose, N-acetylglucosamine, etc.

(Refer Slide Time: 07:28)



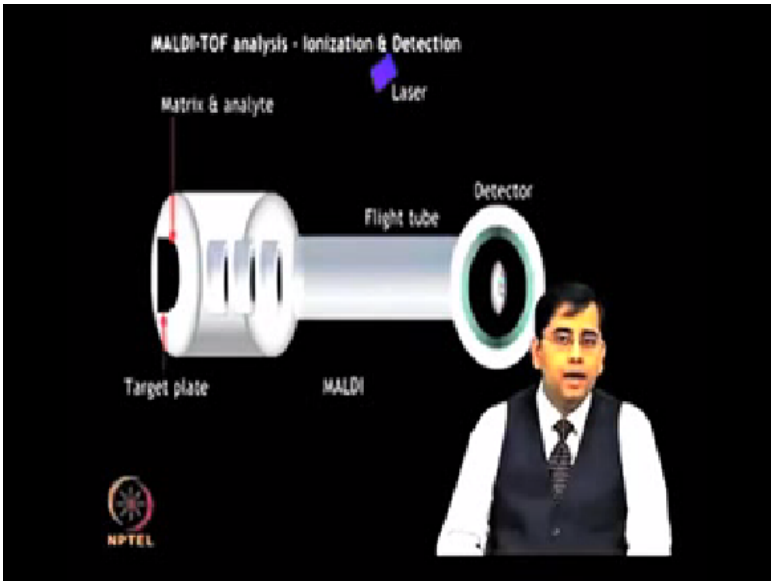
(Refer Slide Time: 07:33)



Protein mixture is digested with proteolytic enzyme such as trypsin and resultant peptides can be analysed by MALDI-TOF or LC-MS/MS. The top down mass spectrometry involves analysis of intact protein using high resolution mass spectrometry techniques. High resolution MS platforms such as FTICR-MS, Orbitrap-MS with PTM friendly dissociation techniques such as electron capture dissociation and electron transfer dissociation, ETD are commonly used.

Features can be detected by means of mass spectrometry due to the unique fragmentation patterns of phosphorylated serine and threonine residues.

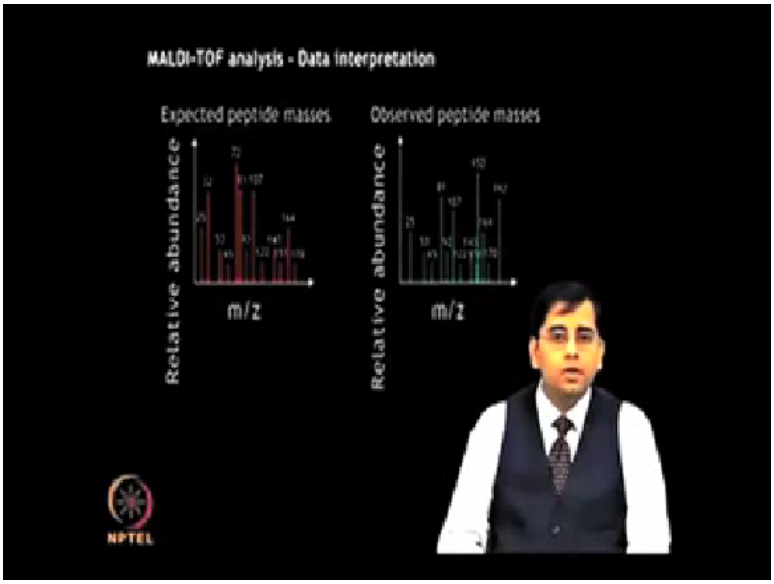
(Refer Slide Time: 08:39)



The modified protein of interest is digested into small peptide fragments using trypsin. This digest is then mixed with a suitable organic matrix such as alphacyano-4-hydroxycinnamic acid, sinapinic acid, etc. and then it is spotted on to MALDI plate. The target plate containing spotted matrix and analyte is placed in a vacuum chamber with high-voltage and short laser pulses are applied.

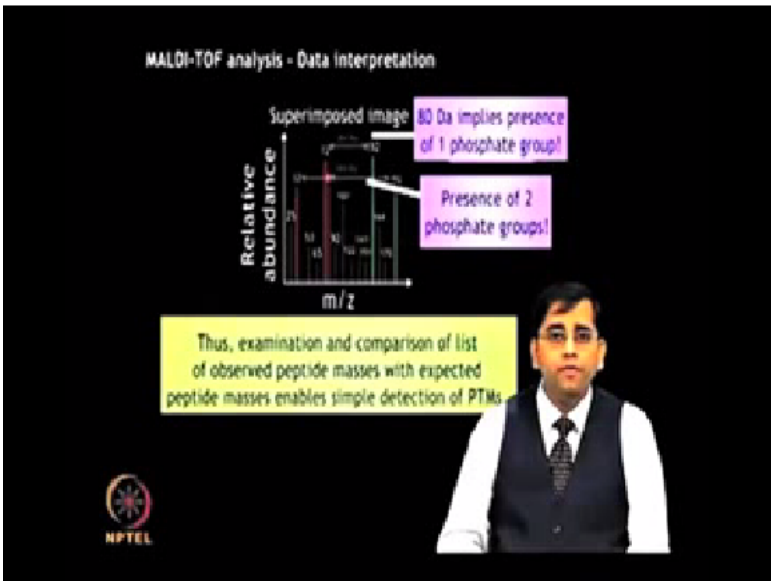
The laser energy gets absorbed by the matrix and is transferred to the analyte molecules which undergo rapid sublimation resulting in gas phase ions. These ions are accelerated and travel through the flight tube at different rates. The lighter ions move rapidly and reach the detector first while the heavier ions migrate slowly. The ions are resolved and detected on the basis of their m/z ratio.

(Refer Slide Time: 10:02)



And a mass spectrum is generated. Identification of post-translational modification by MS largely lies in interpretation of results.

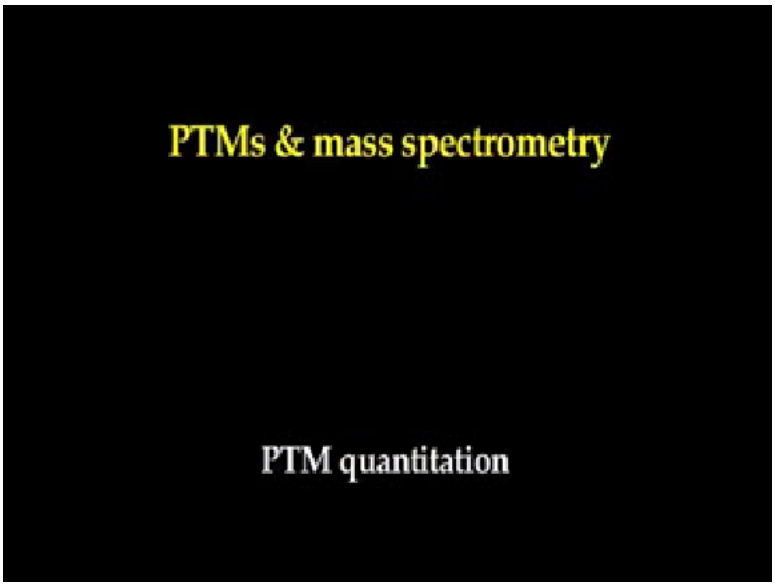
(Refer Slide Time: 10:17)



Comparison of list of the observed peptide masses from the spectrum generated with expected peptide masses enables the identification of those peptide fragments that contain any post-translational modification due to added mass of a modifying group. In this hypothetical example, 2 peptide fragments are found to have different m/z values differing by 80 daltons and 150 daltons.

It is known that the added mass of a phosphate group causes an increase in m/z of 80 daltons. Therefore, this principal of mass difference enables the detection of modified fragments.

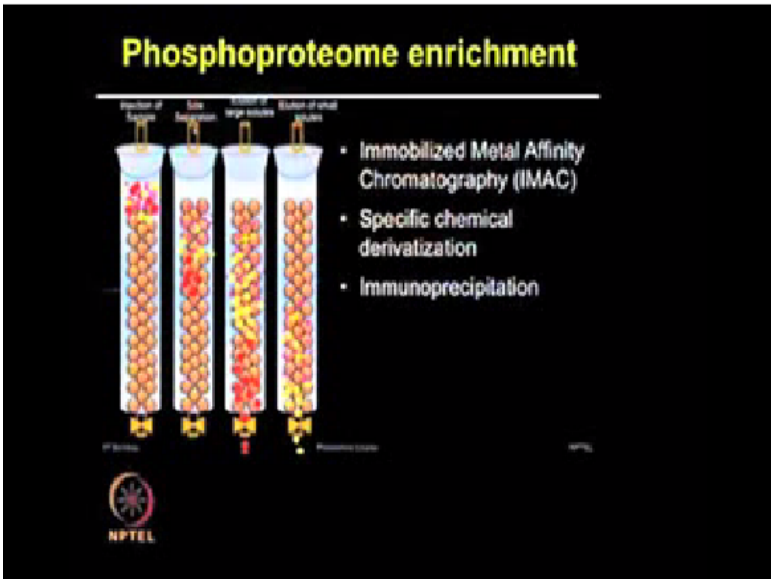
(Refer Slide Time: 11:15)



Affinity-based enrichment, immunopurification and metal-affinity chromatography are

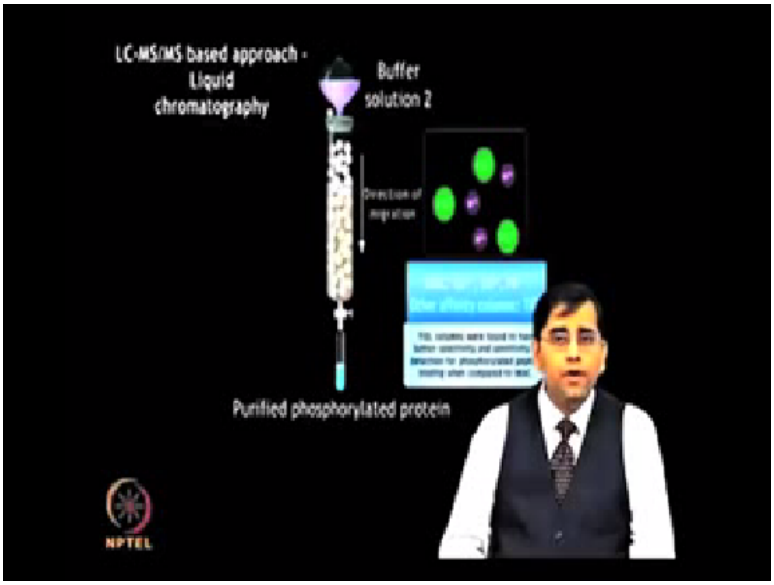
commonly employed for purification of proteins containing specific PTMs.

(Refer Slide Time: 11:38)



Immobilized metal affinity chromatography, IMAC and metal oxide affinity resin such as titanium dioxide, Fe3O4 are also commonly used for the enrichment of phosphoproteins.

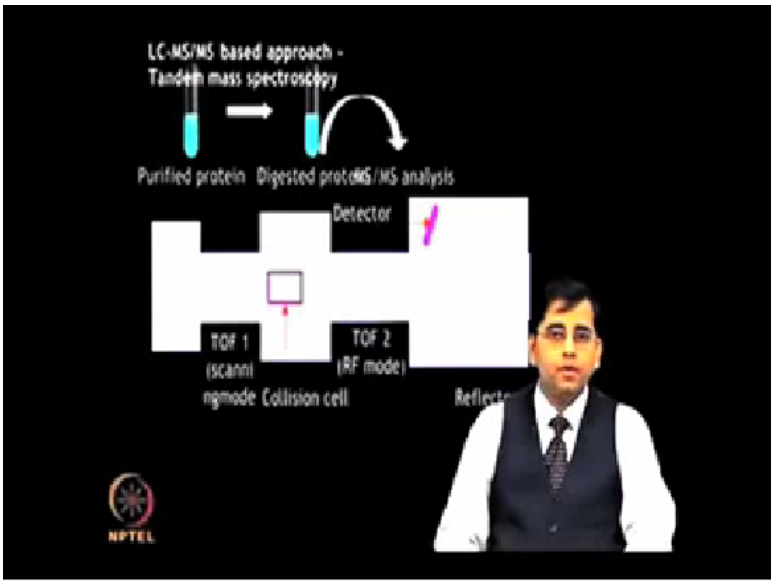
(Refer Slide Time: 11:57)



A protein phosphorylation experiment is shown here. The complex protein sample is loaded onto a miniaturised affinity column which interacts specifically with proteins having the post-translational modification of interest. The IMAC chromatographic columns containing ions such as zinc, iron, titanium dioxide is specifically chelate the phosphorylated proteins. The unwanted proteins are removed by washing the column with a suitable buffer solution after which the

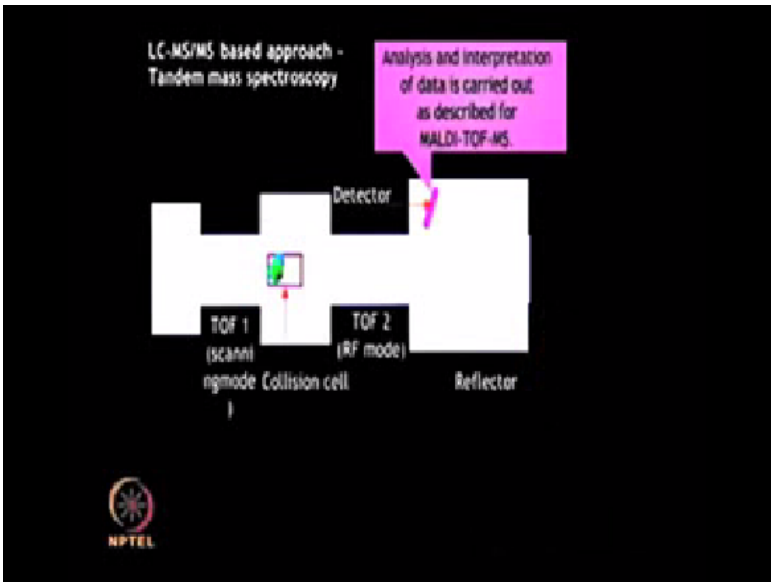
phosphorylated protein of interest can be eluded out by modifying the buffer solution.

(Refer Slide Time: 12:52)



The protein purified by liquid chromatography is then subjected to tryptic digestion.

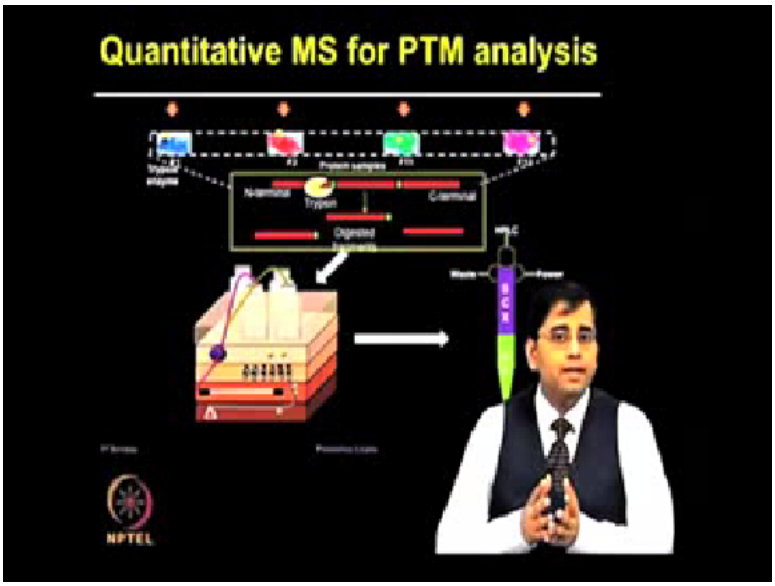
(Refer Slide Time: 13:00)



Followed by analysis using tandem mass spectrometry. Here I have demonstrated the use of MALDI-TOF-TOF MS for resolution of the generated ion fragments. Separation is based on the flight time of the ions and greater resolution is achieved due to the presence of 2 mass analyzers. The peptide ion spectrum generated is analyzed by comparing it with the expected spectrum thereby allowing determination of modified peptides having different m/z values.

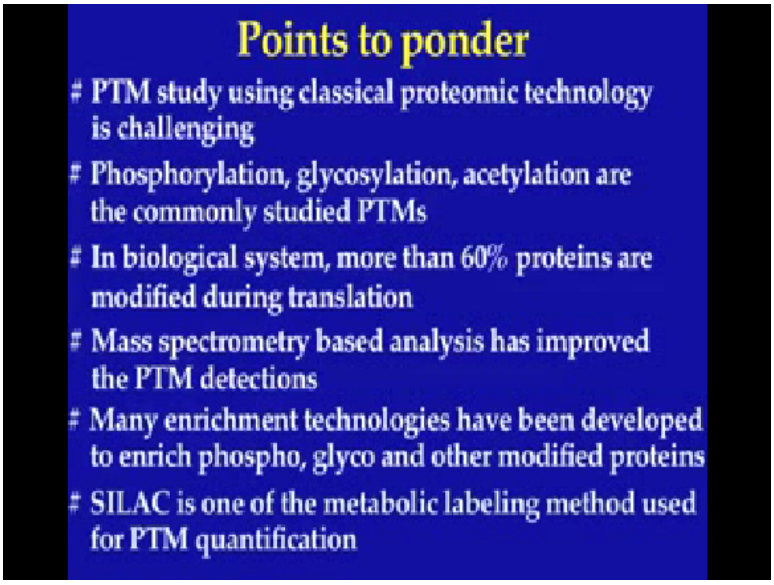
The metabolic labelling methods such as SILAC is used for label-based quantitation of PTMs.

(Refer Slide Time: 13:55)

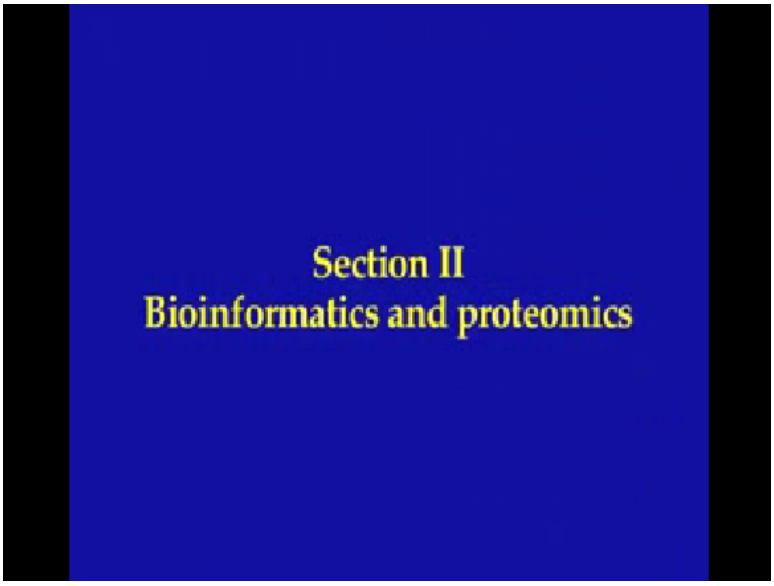


However, this strategy can only be used for the living cells. Other chemical labelling methods such as iTRAQ is also used for PTM analysis.

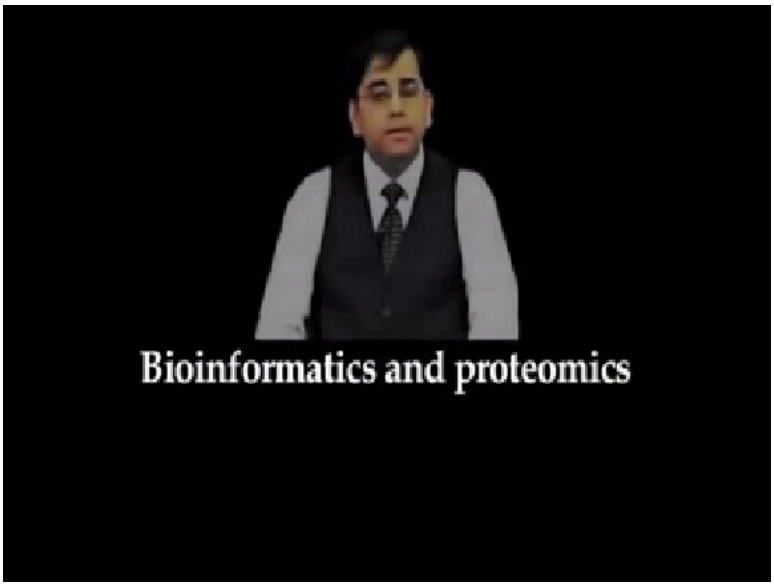
(Refer Slide Time: 14:09)



(Refer Slide Time: 14:28)



(Refer Slide Time: 14:31)

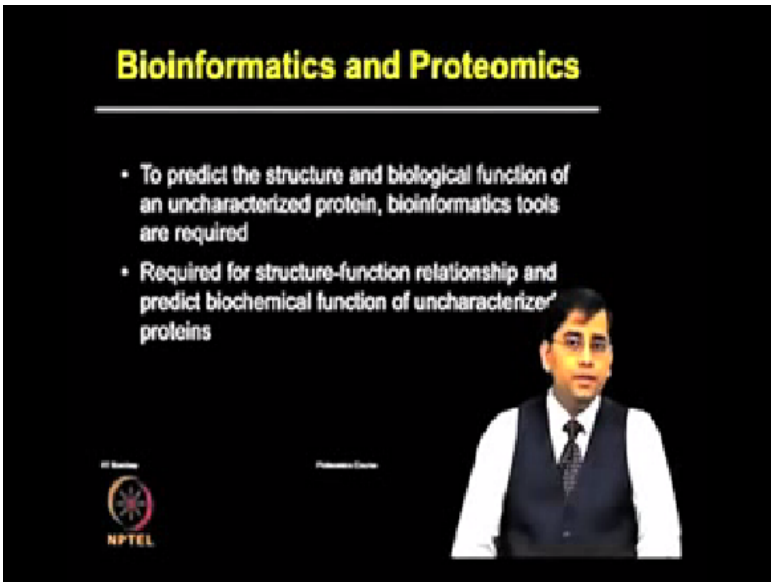


To predict the structure and biological function of an uncharacterised protein, computational methods rely on structural homology of unknown protein from proteins with known structure and biological function. By relying on such methods for structure function correlations, it is possible to predict biochemical function of uncharacterised proteins based on the structure homology to another protein with a known function.

Recent advancement in proteomics and other OMIC technologies allow large-scale analysis of biological samples and generate an unprecedented amount of digital data. In different modules, we have discussed different bioinformatic tools and software for analysing proteome and system-

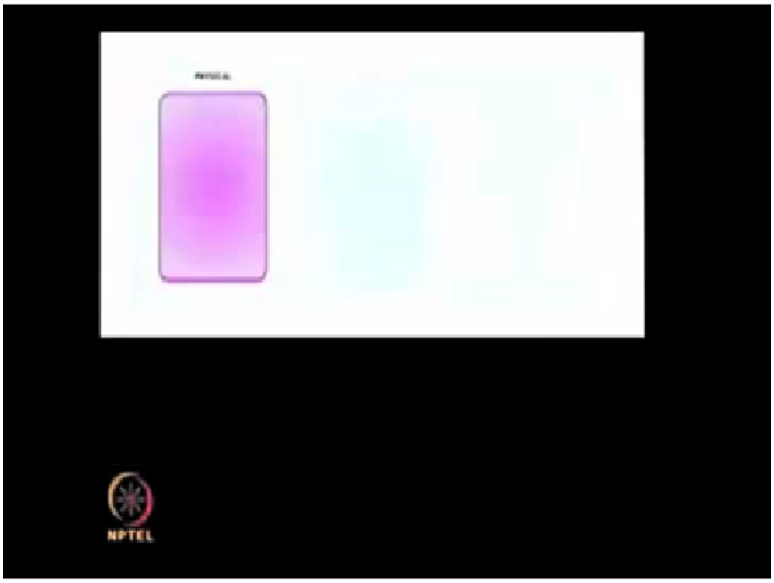
level investigation using 2-dimensional electrophoresis, mass spectrometry, micro-arrays and surface plasmon resonance.

(Refer Slide Time: 15:50)



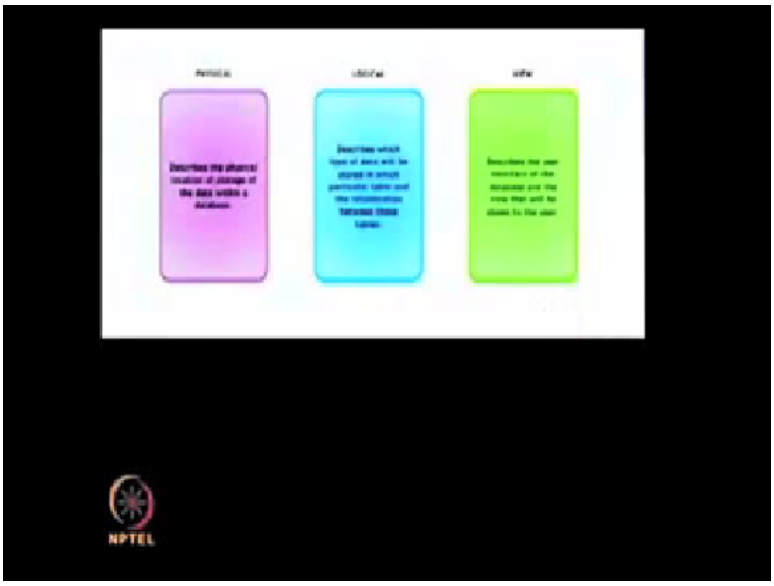
Computational challenges associated with proteomic studies have recently emerged as some of the most critical and limiting factors in this rapidly evolving discipline. Bioinformatic tools have been widely used for protein sequence analysis.

(Refer Slide Time: 16:17)



It is also used for protein motive detection and epitope prediction. Active site determination determining transmembrane domains as well as identification of DNA binding residues.

(Refer Slide Time: 16:33)

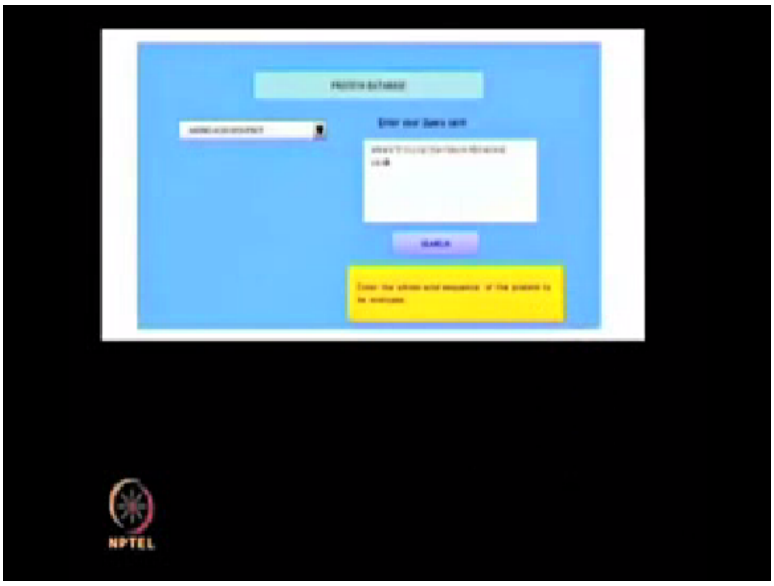


Data-base designing is done at various levels such as physical, logical and conceptual. At the physical level, the purpose of the database is defined which is in accordance with the proposed usage. At the logical level, the tables, attribute of the tables and data types are defined. At the view level, the views and appearance of the databases are defined.

(Refer Slide Time: 17:11)

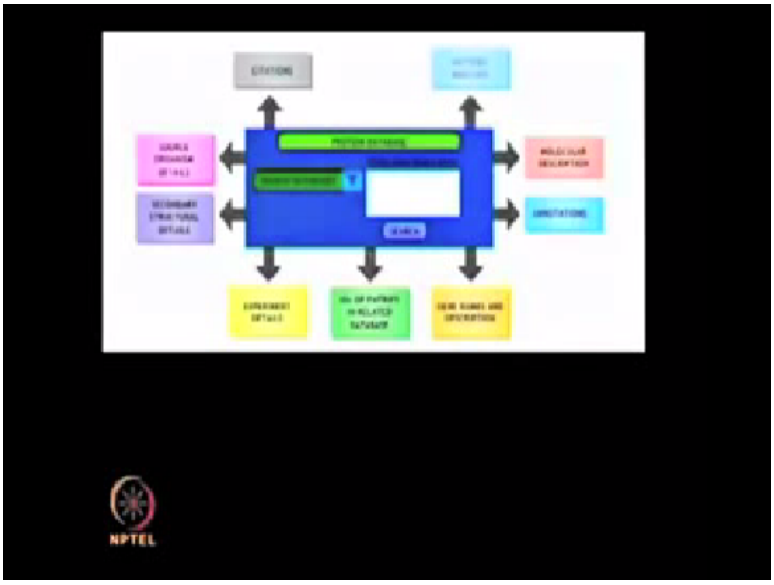
A typical biological database can be characterised by its type and its tool. The type defines the category of data that it includes such as sequence, domains, or structure. This implies that the particular database is most prominent feature include either sequences, domains, or structure and it is particularly used for their analysis. The analysis tool defines the platforms that the side will provide for gaining an insight into the protein data.

(Refer Slide Time: 17:59)



For extracting the protein information from a database, users can give a variety of input terms. This can be unique ID, molecular name, amino acid sequence, keyword, literature, gene, taxonomy, etc.

(Refer Slide Time: 18:22)



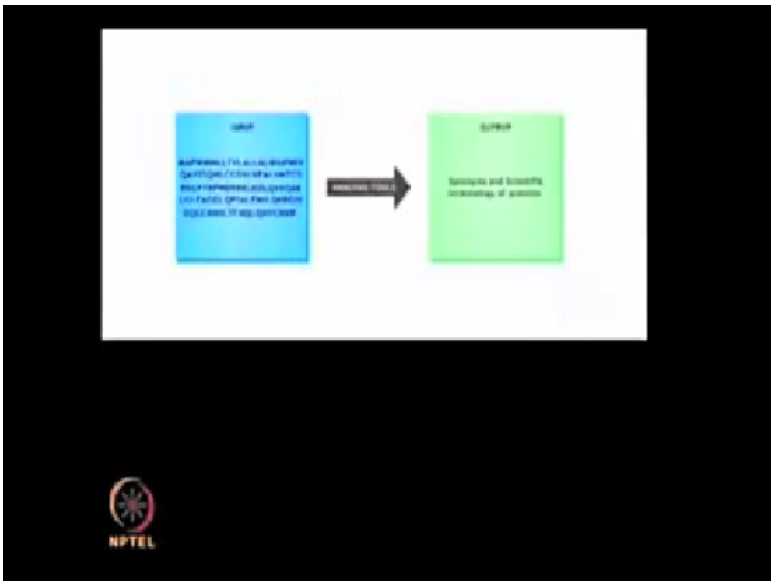
While the user submits a query, the output can be of multiple formats. The generalised information that users can obtain from protein reserve databases is general description of the protein molecule. The generalized information that users can obtain from protein databases is general description of the protein molecule, annotation of the protein, name and description of the gene that transcribes them.

ID of the same protein in other relevant databases, details of the experiment conducted for characterising proteins, details of proteins secondary structures, details of the organism which was used as a source for obtaining the protein and citations of research conducted.

(Refer Slide Time: 19:28)

Database analysis tools. The different kinds of analysis can be conducted on a given protein sequence. The query can be protein name, sequence, or any other identifier of the protein. Various kinds of research output can be obtained, identity of protein from sequence, identify physicochemical properties, molecular weight, isoelectric point, sequence tag information. Similarly search algorithms such as versions of BLAST, FASTA and multiple sequence alignment.

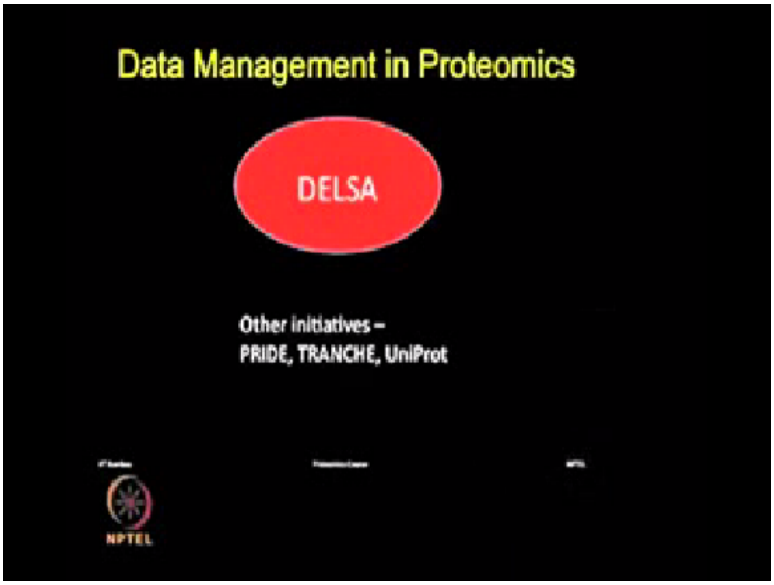
(Refer Slide Time: 20:18)



Finding conserved and variable domains in the protein to study its evolutionary relationship with other proteins, molecular modelling and visualisation tool, secondary and tertiary structure prediction and structural analysis, biological test analysis such as biomedical acronyms, gene protein synonyms, etc. Database mining in proteomics and visualisation tools.

Collective improvement in the search field can be accelerated by sharing scientific data among different research groups across the world. Seeing as it, allows other researchers to access, validate and re-analyse ones finding and correlate the results with their own observations.

(Refer Slide Time: 21:22)

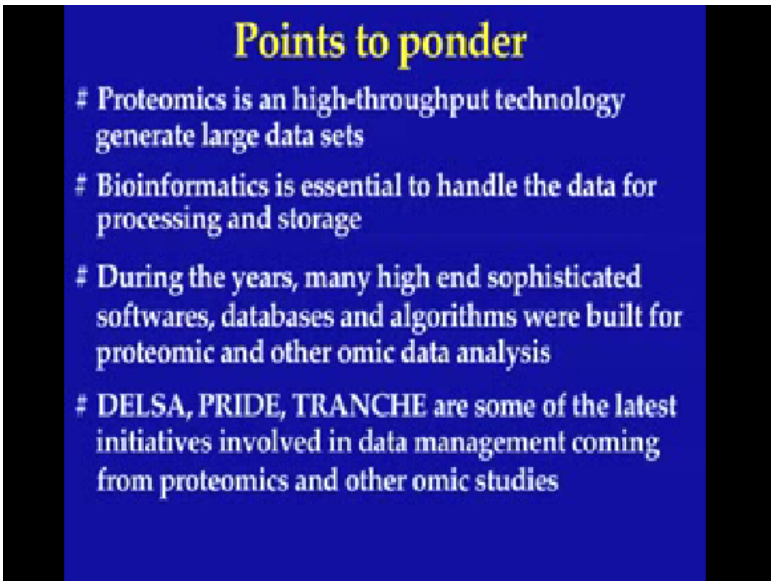


Data management is critical when using high throughput proteomic techniques. Several internet

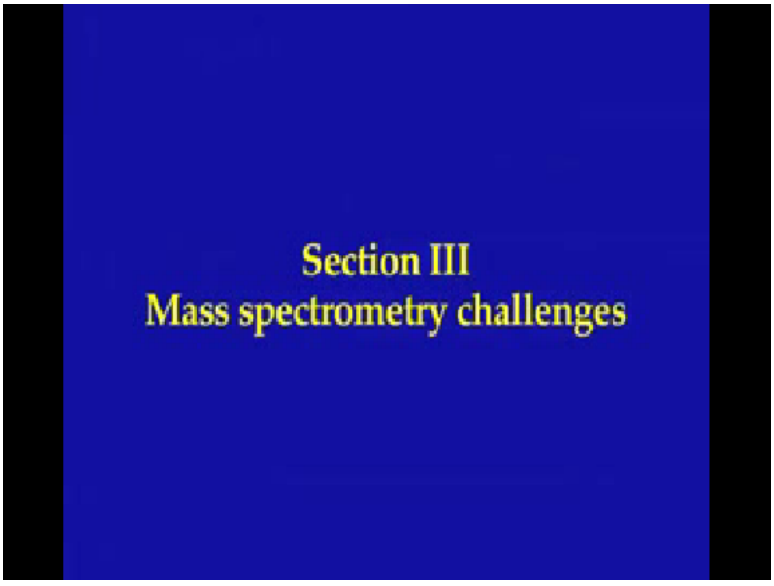
databases have been established to collect the proteomic datasets. Data-enabled life sciences Alliance DELSA is a timely and important initiative to create a common databank where on one hand, we can access the huge dataset generated by various research groups, on other hand.

We can also deposit our datasets which may be useful for a wide range of researchers working in similar fields. At present the broad field of DELSA encompasses biological sciences, ecology, environmental sciences, evolution, genomics and proteomics, computer sciences, cyber infrastructure, management, health sciences and policies for global distribution.

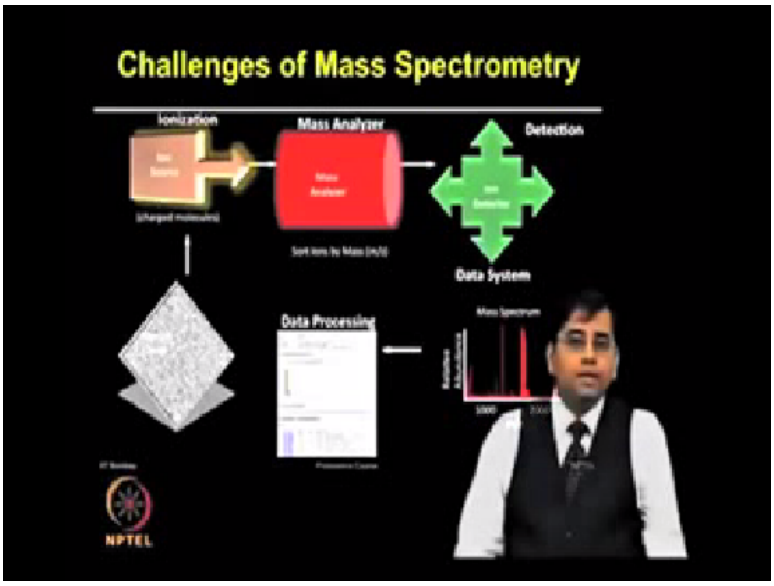
(Refer Slide Time: 22:34)



(Refer Slide Time: 22:50)



(Refer Slide Time: 22:55)



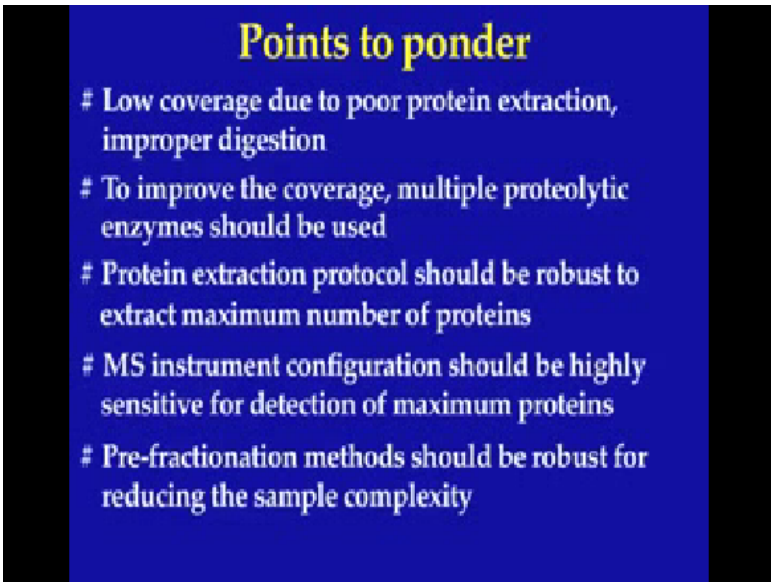
The MS-based proteomics encounters the following biological problems while analysis of huge number of proteins. Fragile nature of proteins, substantial losses occurring during the sample collection and processing steps, presence of multiple isoforms of single protein, the wide dynamic range of protein concentration in biological fluids, presence of high abundance proteins masking low abundant marker proteins.

Additionally, the technological limitations associated with most of the commonly used MS-based approaches include typical dynamic range of only 10 to the power 2 to 10 to the power 4. Inadequate coverage of whole proteome unless sample is fractionated extensively, low throughput and issues of robustness and cost over-fitting the data, machine fluctuation, instrument noise and contaminants in a spectrum and lack of standard procedure for analysis and interpretation of MS and MS/MS spectrum.

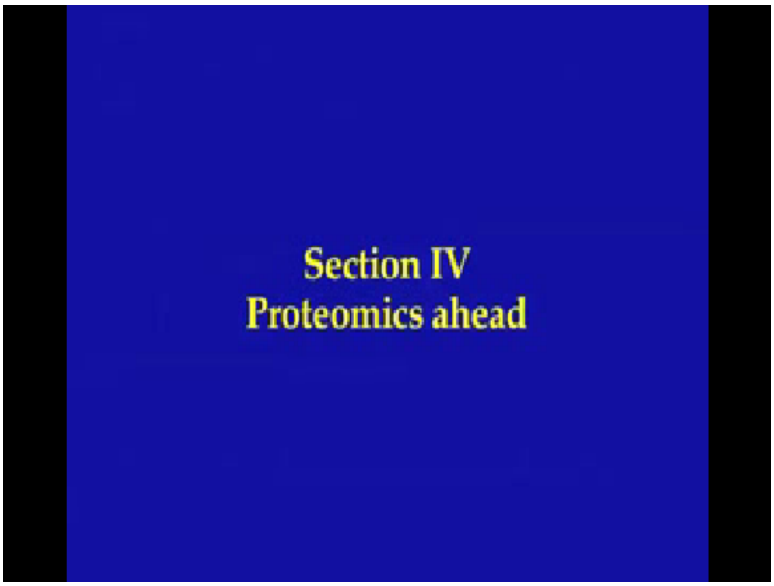
To overcome these technological challenges, different novel and amalgamated approaches have emerged in the last few years. The most promising advancements include large-scale quantitative proteomics, culture derived isotope tags and super SILAC-based technologies. Multiplexing tandem mass tags, TMTs and iTRAQ 8 plexin, quantitative accuracy, level free LC-MS/MS, low sample consumption and large-scale analysis, chip-based and nano LC-MS, sensitive quantitation of proteins within complex pictures, biomarker discovery, multiple reaction

monitoring MR-MS, large-scale biomarker discovery, etc.

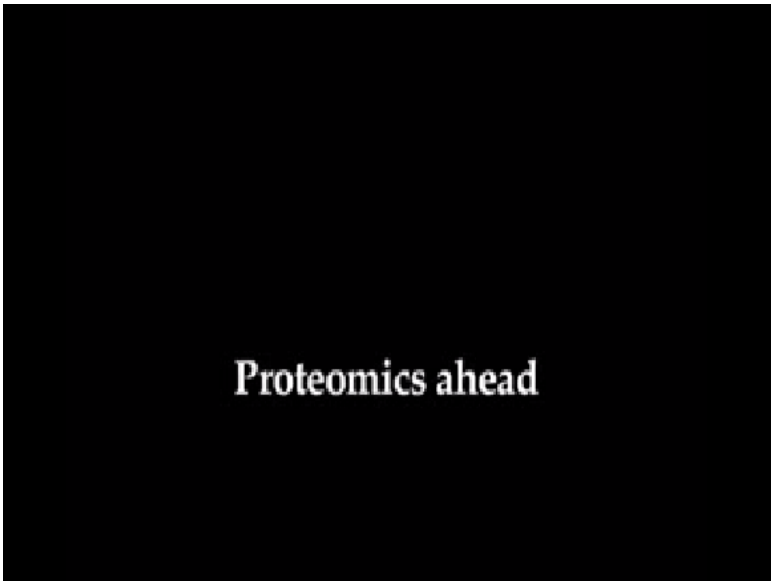
(Refer Slide Time: 25:37)



(Refer Slide Time: 25:53)



(Refer Slide Time: 25:58)



Impeding future of this promising research area will highly depend on the collaborative initiatives at global level and establishment of effective data repositories accessible to the proteomics researchers across the world.

(Refer Slide Time: 26:20)

In 2010, Human Proteome Organisation, HUPO have launched a global human proteome project, HPP. This project is designed to map the entire human proteins encoded by the genome.

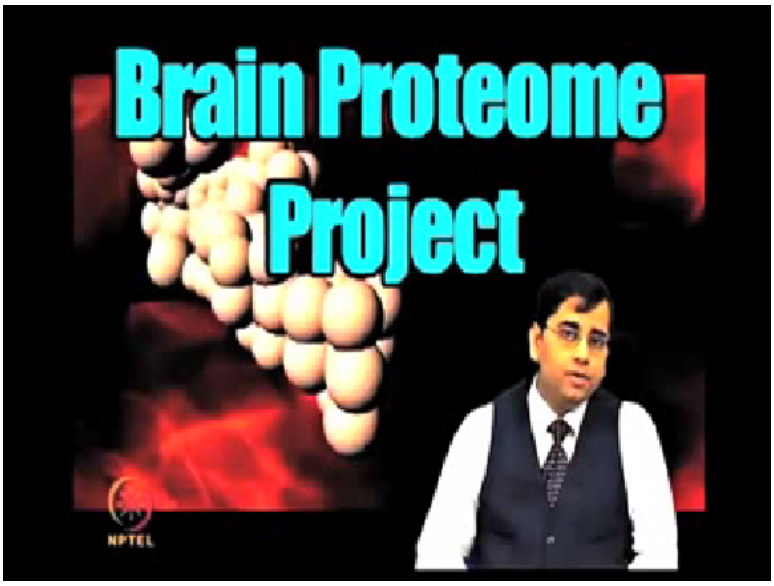
(Refer Slide Time: 26:39)

Let us now discuss some of the targeted focussed initiatives. The Human Liver Proteome Project. This is the first initiative of human proteome project for human organ, tissues with an intention of generation of comprehensive protein atlas of the liver and international liver tissue network, collection and distribution of normal liver sample and validation of new discoveries.

(Refer Slide Time: 27:18)

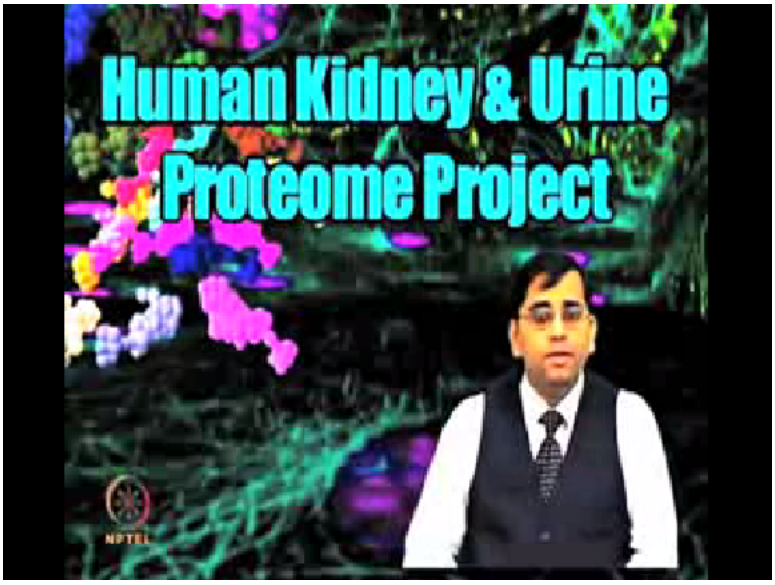
Human Plasma Proteome Project. Analysis of the protein constituents of human plasma and serum.

(Refer Slide Time: 27:27)



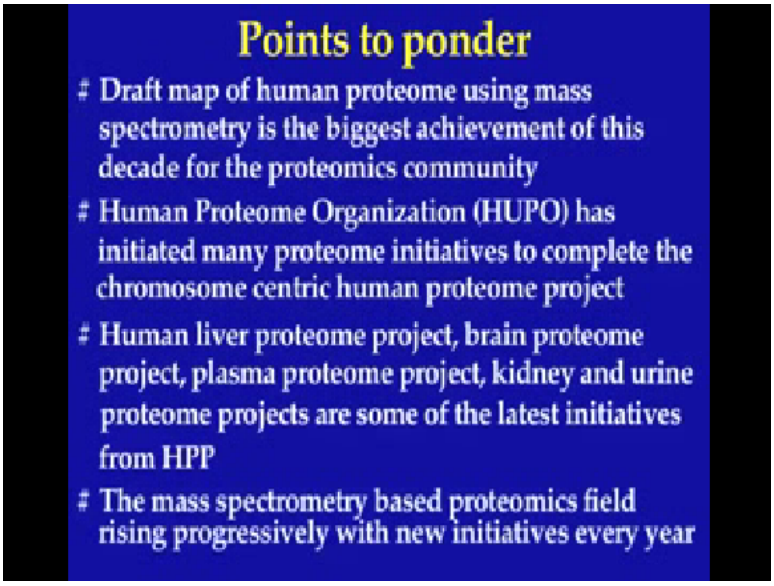
Human Brain Proteome Project, BPP focuses on the revolution of the brain related proteomics alteration focusing on understanding neurodegenerative diseases, ageing and identification of prognostic and diagnostic biomarkers.

(Refer Slide Time: 27:51)



Human Kidney and Urine Proteome Protein Project aims to understand kidney functions, mechanism of chronic kidney diseases at the protein level and discovered biomarkers and target molecules for due therapeutics of kidney diseases.

(Refer Slide Time: 28:13)

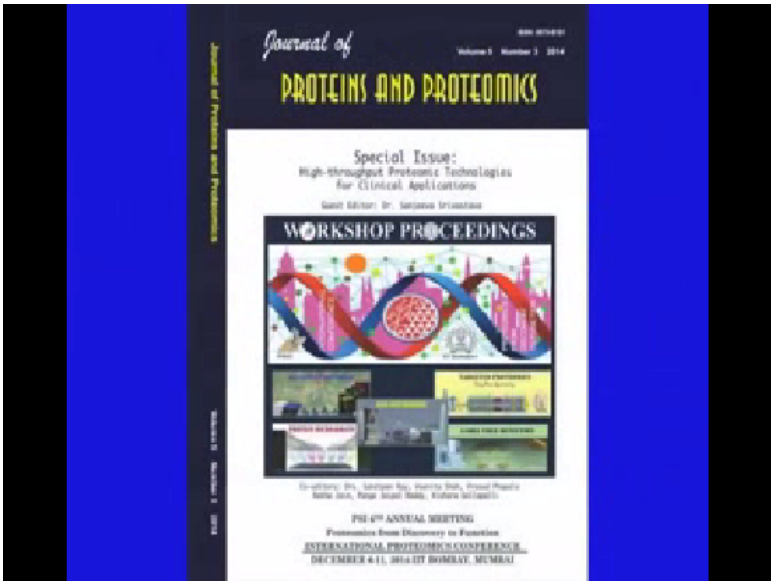


Over the last decade, proteomics research is progressing in different regions of India with a considerable interest.

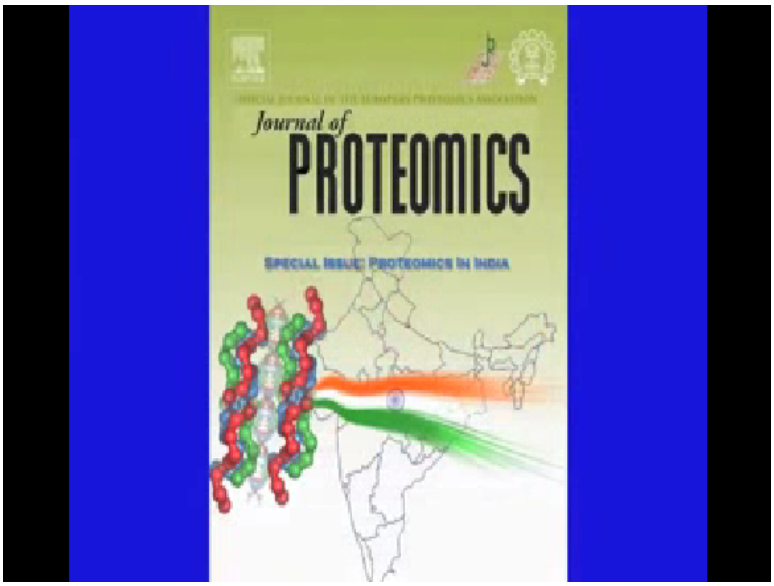
(Refer Slide Time: 28:36)

India is playing an increasingly significant role in global genomics and proteolysis research and development as it is evident from research publications and patents.

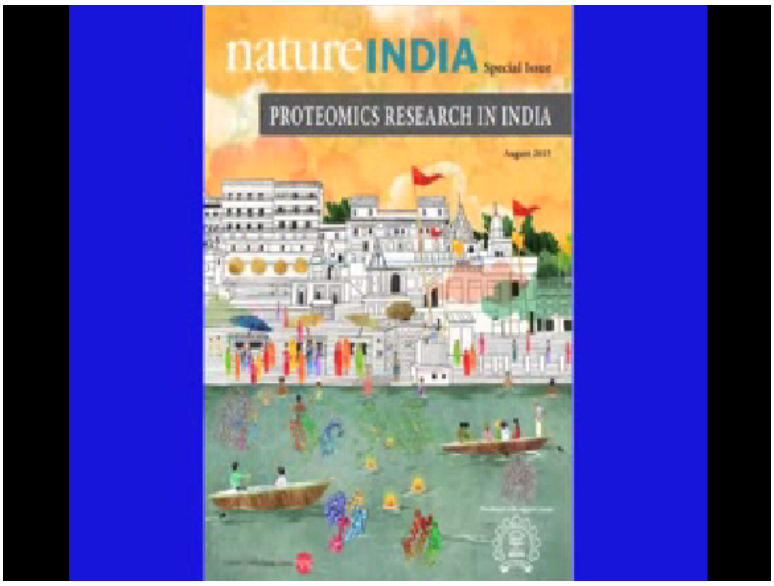
(Refer Slide Time: 28:57)



(Refer Slide Time: 28:59)

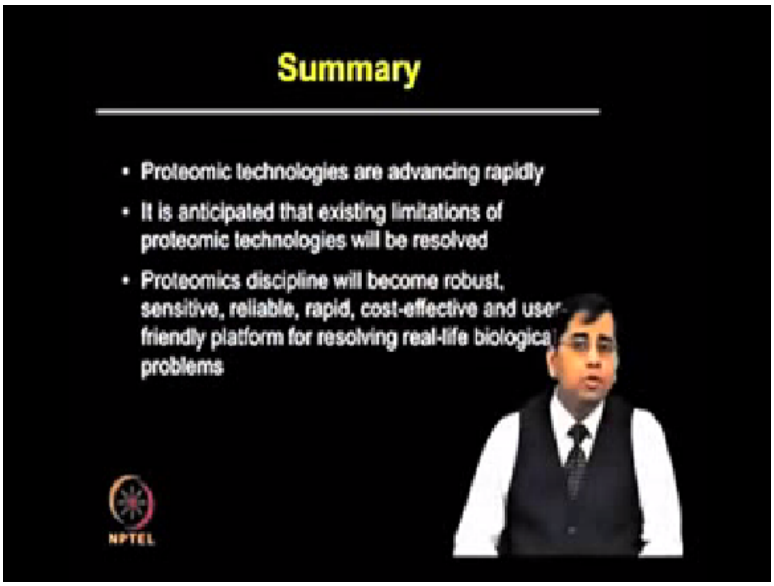


(Refer Slide Time: 29:02)



Indian government is also supporting the basic and applied proteomics as well as other OMICS-based research and multiple national and international funding agencies are providing investment on existing and new research projects.

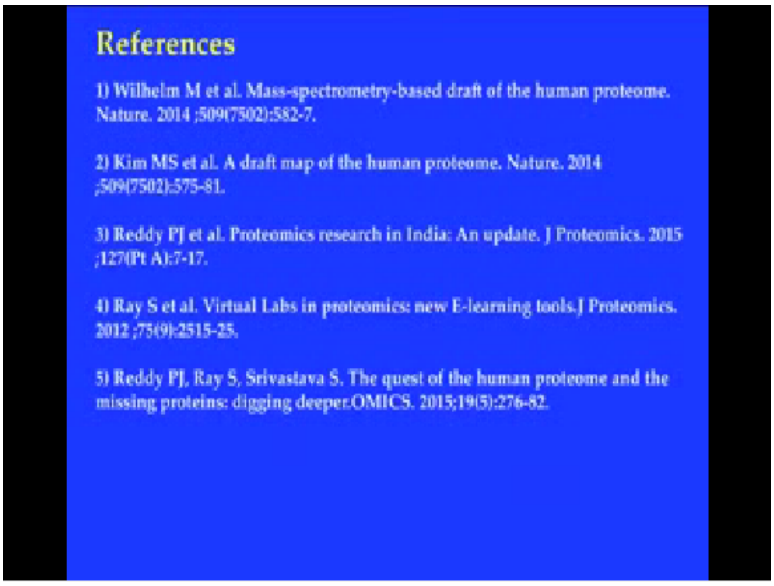
(Refer Slide Time: 29:40)



Concept in the space of emerging proteome-level research, it can be anticipated that in coming future, some amicable solutions for the existing limitations associated with the burgeonic field of proteomics will come forward through worldwide research initiatives and this discipline will become more robust, sensitive, reliable, rapid, cost-effective and user-friendly for resolving real-life biological problems.

Hope this course has given you foundation for proteomic concepts and enthused you for research in proteomics area. Thank you.

(Refer Slide Time: 30:18)



(Refer Slide Time: 30:22)