LECTURE-1

INTRODUCTION TO PROTEOMICS

TRANSCRIPT

Welcome to NPTEL course on Proteomics. My name is Sanjeeva Srivastava. I am in Department of Biosciences and Bioengineering of IIT Bombay. In this introductory lecture I will discuss about proteomics and provide an overview of entire course, which consists of 40 lectures.

Proteome describes the protein complement expressed by a genome, or more precisely, the protein complement of a given cell at a given time, including the set of all protein isoforms and modifications

Study of entire compendium of proteins encoded by a genome is known as “proteomics”

In this slide I have illustrated the complexity of human proteome as compared to genome or transcriptome. The extent of diversity and complexity due to alternative splicing and post-translational modification is tremendous, therefore studying proteins and proteome is very important.

Steps involved in proteome analysis: protein extraction followed by their separation, identification and characterization. Protein extraction from whole cells, tissue or organisms is first requirement for proteome analysis in majority of the proteomics experiments. Protein separation and quantification is achieved by gel-based (e.g. 2-DE) and gel-free techniques (e.g. iTRAQ) and identification by MS. The functional characterization of proteins using novel proteomic platforms opens new horizon for exploration in biology.

Abundance based proteomics aims to measure the abundance of protein expression, whereas functional proteomics aims to determine the role of proteins by assessing protein interaction and biochemical activities.

How did proteomics field emerge?

Advancement of various techniques such as 2DE and mass spectrometry, genome sequencing information and computational algorithms led to the emergence of this field.

Protein chemistry to Proteomics

Proteomics research originates from classical protein chemistry and it has embraced new HT techniques to analyze complex samples. Many of the techniques used under modern proteomics banner (e.g. 2-DE, MS) have actually originated several years ago. So what is new?

Technological advancements in protein analysis with increased sensitivity, resolution and capability to carry out high throughput studies has led to a transition from protein chemistry to the new field of proteomics.

Protein analysis by MS was challenging due to complete degradation of samples with available hard ionization techniques. This limitation was overcome by development of soft ionization techniques, MALDI and ESI. These techniques greatly improved proteomic studies as they facilitated MS analysis of protein samples.

Protein sequencing by Edman degradation is time-consuming and cumbersome. Several rounds of sequencing are required for analysis of long polypeptide chains. However, peptide sequencing by MS is much faster, and allows large number of samples to be analyzed in short time.

Development of IPG strips facilitated proteome analysis using 2DE: The pH gradient in tube gels are established by ampholytes gradients, which are not always very stable and tend to break down upon addition of concentrated samples. Analysis of protein mixture by 2-DE using tube gels often gives a lot of variation across gels.

The problem of reproducibility was overcome to a large extent by the development of Immobilized pH gradient strips. Minimal gel-to-gel variation was observed when samples were run by 2-DE using IPG strips, which made this technique suitable for large scale proteomic applications.

Completion of several genome sequence projects: Genome sequences of several organisms, including humans, have been successfully completed and these genome databases are extremely useful in correlating gene and protein sequences. Several databases are now readily available which can easily help in identifying gene sequence of a protein that has been sequenced by mass spectrometry.

Genomics to Proteomics

Genome represents an important starting point towards understanding complexity of biological functions. However, proteins, provide a much more meaningful insight into the mysteries of essential biological processes.

To obtain better understanding of cellular processes & regulation, there has been an increasing interest in studying proteomics.

There are several reasons why one need to study proteomics?

Genomic DNA contains large stretches of non-coding regions: Pre-mRNA is synthesized from genomic DNA by the process of transcription. mRNA contains both exons, the coding sequences, as well as introns which are intervening, non-coding sequences .

By involving series of steps, finally free 3' hydroxyl group of the first exon attacks the 5' end of the second exon such that they are joined to give the mature mRNA.

Single gene, multiple proteins: Alternative splicing is a process by which exons or coding sequences of pre-mRNA produced by transcription of a gene are combined in different ways during RNA splicing. Resulting mature mRNA give rise to different protein products by translation, most of which are isoforms of one another. The diversity of proteins encoded by a genome is greatly increased due to alternative splicing.

Post-translational modification of proteins: The proteins obtained by translation undergoes folding and various PTMs such as phosphorylation, alkylation, glycosylation, hydroxylation etc. to give the final functional protein. PTMs generate diversity, complexity and heterogeneity of gene products and its functional consequences can be modulation in protein dynamics and alteration of its functional activity.

Central Dogma, Omics and Systems Biology

During the last decade we have witnessed the revolution in biology, as this discipline has fully embraced "omics" tools. The emergence of genome-wide analyses to understand cellular DNA, RNA and Protein content by employing genomics, transcriptomics and proteomics at systems level has revolutionized our understanding of control networks that mediate cellular processes. These concepts will be discussed in first module.

Genes are the blue-print for life and proteins are the effector molecules. Due to this fact the central dogma has guided research at the systems level. After completion of human genome sequence number of genes ~25,000 are surpassed by an estimated number of proteins in millions.

Studying large scale study of protein structure and function, requires a thorough understanding of protein composition and their various structural levels by employing HT tools.

Proteins play an important role in essential characteristic of living systems, how they function and replicate themselves through intricate molecular interactions.

Amino acids constitute the basic monomeric units of proteins, which are joined together by peptide bonds.

The linear sequence of amino acids constitutes primary structure. Folding of polypeptide/protein chain into regular structures like a-helices, b-sheets, turns and loops gives rise to secondary structure.

Three dimensional compactly folded structure of proteins makes tertiary structure, which represents overall organization of secondary structural elements in 3-D space.

Quaternary structure - refers to interaction between individual protein subunits in a multi-subunit complex.

Sickle cell anemia is caused due to single nucleotide substitution which converts a glutamic acid residue to valine in beta chain of hemoglobin. Thalassemia is caused due to abnormalities in hemoglobin synthesis.

Protein folding is an elegant example of biological self-assembly. Understanding the mechanisms through which protein folding takes place remains challenging for scientific community.

Anfinsen tested the ability of reduced and unfolded Ribonuclease A protein to spontaneously fold into its native state. Protein folding is a cooperative process which arises from simultaneous formation of multiple interactions within a polypeptide chain. Protein folding is thermodynamically favorable and spontaneous process.

Folding efficiency could be limited by processes such as aggregation. The molecular chaperones are designed to promote protein refolding.

From complex proteome it is challenging to purify a protein in a single chromatographic step. Therefore, sequential pre-fractionation steps involving different modes of chromatography becomes necessary.

Gel filtration chromatography separates protein on basis of difference in size. When a protein sample is applied to column, small proteins passes from the pores of the beads while the large proteins are excluded, therefore this technique is also known as “molecular exclusion”.

Ion-exchange chromatography relies on differences between number of charges and distribution of charge groups in defined pH and solvent condition.

Affinity chromatography is based on affinity of protein to other molecules, where protein of interest binds through a specific interaction.

GENOMICS

Studying genome of an organism by employing sequencing and genome mapping is known as “genomics”.

Several genome sequencing projects that aim to elucidate the complete genome sequence of organisms have been undertaken by several research groups all over the world. From a genomic library clones were isolated and ordered into a detailed physical map. Further, individual clones were sequenced by shotgun sequencing to provide the complete genome sequence.

Recently next-generation sequencing strategies have dramatically increased the pace of sequencing by several order of magnitudes. NGS based on nanopore structures is known as nanopore sequencing. For NGS various commercial platforms such as Illumina, Pyro-sequencing, Helicos, Ion Torrent etc are available.

TRANSCRIPTOMICS

Study of all the mRNA molecules expressed by a particular cell type of an individual is known as transcriptomics. The transcriptomic analysis measures the genes that are being actively expressed at any given time and varies significantly with external environmental conditions. Various techniques such as microarrays, Q-PCR etc. have been widely used for transcriptomics analysis.

In microarray experiment, mRNA from control and test samples are extracted and reverse transcribed into its corresponding cDNA. The cDNA samples are labeled with Cy5 and Cy3 dyes and mixed cDNA sample is incubated on printed DNA microarray. This allows hybridization to occur between the probe oligonucleotides on the array surface and the labeled cDNA samples of interest. In this manner expression level of thousands of genes can be measured and analyzed simultaneously.

Different type of proteomic technologies such as 2DE, Mass Spectrometry, Microarrays and Label-free techniques will be discussed in detail.

SYSTEMS BIOLOGY

In omics era, technological advancements in genomics, proteomics, and metabolomics have generated large-scale datasets in all aspects of biology. These large data-sets has

motivated the computational biology and systems approaches with objective of understanding biological systems as a whole.

Systems biology and biological network modeling aims to understand biological processes as whole system rather than isolated part by synergistic application of experiment, theory, technology & modeling.

Systems level studies aim to develop computationally efficient and reliable models of underlying gene regulatory network. Quantitative analyses measures and aims to make models for precise kinetic parameters of a system network components. It also uses properties of network connectivity.

Gel-based proteomics

Several techniques used in proteomics typically aim to elucidate the expression, localization, interaction, and cellular function of proteins. SDS-PAGE, 2DE and DIGE are commonly used gel-based techniques.

Protein extraction is the first step for proteomic analysis. The protein extraction methods aim that most, if not all the proteins in a cell or its organelles are extracted by the procedure and the presence of interfering compounds are minimized.

Different biological samples pose different challenges. E.g. serum proteome analysis shown here, illustrates that proteins in biological systems such as serum may have difference of several order of magnitudes. Albumin and immunoglobulin are the most abundant proteins in serum, which mask other proteins at lower concentrations. It is therefore preferred to remove these high abundance proteins by using affinity chromatography based methods. Once the serum has been processed using a depletion columns, proteins can be extracted.

In bacterial protein sample preparation sonication is an important step to disrupt the bacterial membrane. Sonication breaks open the cellular membranes to release the intracellular contents. Protein extraction can be performed using different methods and protein pellets are reconstituted in lysis buffer for proteomic analysis.

Protein quantification is sensitive to detergents or certain ions therefore it is crucial to select the correct quantification method. In Bradford color reagent transfer of electrons converts the dye to its blue form thereby giving the solution a blue color. Absorbance of standard and unknown protein samples can be measured at 595 nm and protein concentration can be determined from the standard plot of the absorbance values.

In gel-based proteomics, proteins are commonly analysed using SDS-PAGE and two-dimensional (2D) gel electrophoresis. Separation in SDS-PAGE occurs almost exclusively on the basis of molecular weight and in 2DE the complex mixtures are resolved first by isoelectric point and then by size on a polyacrylamide gel. Some of the limitation of 2DE can be overcome by Difference gel electrophoresis (DIGE) technique. 2DE in combination with mass spectrometry has been the standard technique for proteome analysis.

Two-dimensional electrophoresis involves protein separation on a pH gradient based on their isoelectric point (pI) using isoelectric focusing (IEF) followed by separation in the second dimension using SDS-PAGE where the proteins are resolved according to their molecular weight.

To perform 2DE, add the reconstituted protein sample to the rehydration tray and place the IPG strip for rehydration. Isoelectric focusing involves the application of an electric field, which causes the proteins to migrate to the position on the pH gradient that matches the pI of a specific protein after which it does not move in the electric field owing to the lack of charge. The proteins migrate along the strip and come to rest at a point where their net charge becomes zero known as their isoelectric point.

Prior to second dimension separation, an equilibration step is required. In equilibration, Dithiothreitol brings about cleavage of the protein disulphide bonds while iodoacetamide prevents reformation of these bonds by binding to free sulphydryl groups.

On SDS-PAGE gel, proteins get separated on the basis of their molecular weight with the low molecular weight proteins having high mobility and migrating further through the gel and the high molecular weight proteins remaining close to the point of application.

Gels can be visualized by different dyes such as Coomassie blue staining, Silver staining, Cyanine dyes etc.

Gel data analysis will be discussed with application expert of GE healthcare. The gel analysis involves, images processing, detection of spots, making match-set, landmarking, viewing histograms etc.

Various information regarding the spots such as their area, volume, intensity and statistical parameters such as standard deviation, can also be calculated.

2DE has high resolving power but it has several limitations such as staining and reproducibility.

Fluorescence two-dimensional difference in-gel electrophoresis (2-D DIGE) is an advanced 2DE technique that allows for accurate quantification with statistical confidence, while controlling non-biological variation.

In DIGE, proteins extracted from different types of cells or tissue samples are labeled with different fluorescent reagents (Cy2, Cy3 and Cy5), mixed and then separated by 2 DE on a single gel. The proteins are detected separately using Cy2, Cy3 and Cy5.

The commercial software such as DeCyder facilitate the automated analysis of DIGE gels and provide differential expression analysis, principal component analysis, pattern and discriminant analysis.

2DE, DIGE followed by mass spectrometry technique has been applied for many applications. Some of these applications will be discussed in this module of gel-based proteomics.

Mass Spectrometry

Mass spectrometry is technique for protein identification & analysis by production of charged molecular species in vacuum, & separation by magnetic and electric fields based on m/z ratio.

MS has become the method of choice for analysis of complex protein samples in proteomics studies due to its ability to identify thousands of proteins.

The gel-based techniques typically resolve only products of a few hundred genes at best, had low throughput and low dynamic range.

To overcome such issues, Mass Spectrometry has become an important analytical tool in proteomics, and in biology in general. It offers high-throughput, sensitive and specific analysis for many applications. In this module we will discuss basic concepts of mass spectrometry, ionization sources, mass analyzers, hybrid MS configurations, and quantitative mass spectrometry techniques such as SILAC and iTRAQ.

The basic components of MS involve “Sample inlet” to transfer sample into the ion source. “Ionization source” which converts neutral sample molecules into the gas-phase ions, A “mass analyzer” to separate and analyze mass of ionic species. Detector, which measures and amplifies ion current of mass-resolved ions and data system to process and analyze data.

Soft ionization techniques such as matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI), are now widely used for proteomics applications.

In MALDI protein is mixed with matrix and laser beam ionizes matrix molecules. It is an efficient process for generating gas-phase ions of peptides and proteins for mass spectrometric detection.

ESI requires sample of interest to be in solution and produces gas-phase ions from solution. The distinguishing feature of ESI is its ability to produce multiply charged ions.

Mass analyzer disperses all ions based on their (m/z) ratio and focuses all mass-resolved ions at a single focal point and maximizes their transmission.

Time of flight - measures m/z ratios of ions based on time it takes for ions to fly in analyzer & strike the detector

Ion Trap - traps ions using electrical fields and measures mass by selectively ejecting them to a detector.

Quadrupole consists of four parallel metal rods and mass separation is accomplished by the stable vibratory motion of ions in high-frequency oscillating electric field.

Some of the latest hybrid MS technologies will be discussed with application experts. The Orbitrap technology will be discussed with application expert of Thermo Scientific. LTQ orbitrap technology shown in video consists of both Linear trap and Orbitrap. It operates by trapping ions radially about a central spindle electrode. Unlike iontrap, the orbitrap uses only electrostatic fields to focus and analyze ions.

Another latest MS technology, Q-TOF and TQ will be discussed with application expert of Agilent Technologies.

Many such advancement in MS during the last decade have provided new ways for protein analysis and facilitated proteomic analysis of various biological systems.

The Q-TOF LC/MS system performs MS/MS analysis using a quadrupole, hexacollision cell and time of flight (TOF) mass analyzer.

Quadrupole selects precursor ions, which are further fragmented in collision cell. The product ions move to detector and spectrum is generated.

Protein labeling with stable isotopes are effective methods for quantitative proteome profiling using MS. Stable isotope labeling by amino acids in cell culture SILAC, which is a metabolic-labeling strategy to encode whole cellular proteome, is widely used methods for quantitative proteomics.

In SILAC two groups of cells are cultured in media that are identical in all respects except that one contains a heavy, isotopic analog of an essential amino acid while the other contains the normal light amino acid. After a number of cell divisions, the grown cells are combined and digested using trypsin. The complex mixture is further separated by SDS-PAGE to simplify the analysis. Further purification is carried out by liquid chromatography and purified peptide fragments are analyzed by MS/MS.

iTRAQ, it is a MS based technique for relative and absolute quantification of proteins. iTRAQ reagents are a set of four isobaric amine-specific labeling reagents (114, 115, 116, or 117). An iTRAQ reagent consists of a reporter group, a balance group, and a peptide reactive group.

Pooled samples are purified on a strong cation exchange column to remove excess unbound reagent. These isobaric labels are detected upon fragmentation and release in MS.

The data obtained from mass spectrometry can be analyzed by using search engines such as Mascot. The analysis requires inputs regarding the experimental parameters such as enzyme cleavage, modifications, instrument used, peptide tolerance etc. The data file generated from MS is uploaded and the search carried out by employing databases such as NCBI, MSDB and SwissProt.

Interactomics

Biology has evolved several mechanisms that regulate interactions, including a variety of PTMs and the presence or absence of an activator or inhibitor molecule. A detailed understanding of protein interactions provides an opportunity to understand the protein functional behavior.

Inspired from the success of gene microarrays various protein microarray approaches have emerged. The development of protein microarrays, on which thousands of discrete proteins are printed at high spatial density, offers a novel tool to investigate protein function.

Some of the widely recognized technologies that have been used to map protein-protein interactions at large scale, such as yeast-two-hybrid, IP with mass spectrometry and different types of microarray platforms will be discussed in Interactomics module.

In Y2H, binding domain is fused with the bait protein while the activation domain is fused with the prey protein. Binding of either one of the fusion proteins to the promoter

is insufficient to bring about transcription of the gene. When the bait protein bound with the binding domain interacts with the prey protein fused with the activation domain, there will be expression of the reporter gene which can easily be detected.

The gene coding for the protein of interest is expressed in a suitable heterologous host system such as E. coli. Protein purification can be done by chromatographic procedures to obtain the pure target protein. Tags like His6 are often fused with the protein of interest to facilitate the purification process due to its specific affinity towards nickel. Protein purity is tested on SDS-PAGE gels.

The array surface can be functionalized with suitable chemical reagents such as Aldehyde and silane derivatizations that will react with groups present on the protein surface. Protein is printed on to the array surface in extremely small volumes by using robotic printers.

There are two types of protein arrays that are commonly used. In forward phase arrays, immobilized antibody is probed by the test lysate. In reverse phase arrays, cellular lysate is immobilized on the array surface and then probed using detection antibodies specific to the target of interest.

In protein detection using direct labeling, all the target proteins are labelled with a fluorescent tag. In sandwich assay, however, a fluorescent tagged secondary antibody that recognizes a different epitope on the target antigen binds to it and is detected by means of the fluorescence.

Protein purification is a laborious and time-consuming procedure which posses several technical challenges eg. protein purity, protein folding and functionality during the purification and immobilization steps.

Cell-free expression based microarrays

These limitations have motivated the advent of cell-free expression based microarrays, which carry out in situ transcription and translation, and eliminate the drawbacks of traditional cell-based methods.

Nucleic Acid Programmable Protein Array (NAPPA) replaces complex process of spotting purified proteins with simple process of spotting plasmid DNA. By using recombinational cloning, & cell-free expression system, proteins are produced in vitro, and captured on array.

In DNA Array to Protein Array (DAPA) slides bearing the DNA template and the protein tag-capturing agent are assembled face-to-face with a lysate containing permeable

membrane placed in between. The expressed protein slowly penetrates the membrane and gets immobilized on the slide surface through its capture agent.

Multiple Spotting Technique (MIST) involves addition of template DNA on to the solid array support. and second spotting step involves the addition of the cell-free lysate directly on top of the first spot. Transcription and translation can begin only after the second spotting step.

In HaloTag method slide is activated with the HaloTag ligand which captures the expressed protein through firm covalent interactions thereby preventing any material loss and ensuring oriented capture of the protein.

Microarrays have become integral part of clinical and drug discovery process.

The protein microarray experiment involves Blocking the slides with milk or SuperBlock. Application of primary AB, and washing with milk followed by incubation of secondary AB and signal detection.

Although microarray experiments are simple but data analysis is very challenging. Biological research has witnessed a paradigm shift from focused reductionist approaches to a greater dependence on data provided by large “Omics” techniques to provide insight into biological systems and organization of physiological networks.

The microarray scanning and data analysis will be discussed with application expert of Spinco for Molecular Devices.

Single or multiple slides can be scanned by using scanner. Using defined scanning parameters, robotic arms can select slides and position it for scanning. The laser power wheel can adjust the laser strength. Fluorescence signal is collected from the photomultiplier tubes. Each channel is scanned sequentially and tiff images are saved.

Data analysis becomes crucial to make sense out of massive amount of data. Software tools can help but they can’t answers all the questions related to functional genomics and proteomics.

It is more important to have a good understanding of both the biology involved and the analytical techniques rather than relying only on software. Challenges of microarray data analysis will be discussed with Prof. Sudesh from Tulane University USA.

Protein microarrays have found wide applications for discovery and functional proteomic studies. Microarrays are used for analyzing both antigens and antibodies in blood

samples and other biological fluids for biomarker discovery. Some of the representative applications will be discussed.

Label-free detection technique

Several conventional label-based detection approaches such as fluorescence, chemiluminscence and radioactive isotopes are in practice but researchers are exploring methods for label-free analysis to get rid of the interference due to the tagging molecules and reduce the complexity and assay time.

Label-free techniques rely on measurement of inherent properties of the query molecules such as mass and dielectric property, and allow direct, real-time detection of biomolecules in a HT manner eliminating the requirement of secondary reactants.

Many label-free techniques such as SPR, SPRi, Ellipsometry, Interference and nano-technique based approaches are emerging rapidly as a potential complement to labeling methods and it will be discussed in module of label-free detection techniques.

SPR-based biosensors provide label-free, real-time detection of interactions. SPR sensorgram describes the changes in SPR signal versus time.

Since the introduction of the BIAcorer SPR instrument the SPR spectroscopy has been widely used to characterize biological surfaces and monitor biomolecular-binding events. The BIAcore technology will be discussed with research product expert of GE Healthcare.

The SPR biosensors have played an important role in research into biomolecules and their interactions and now they are increasingly being used for detection and identification of chemical and biological substances.

Performing good SPR experiment and accurate interpretation of binding reactions from biosensors are always very challenging. Performing good SPR experiments, data collection and processing can eliminate artifacts and provide good quality data.

The success of SPR experiments depends on the kinetic measurement in real time, monitoring adsorption of unlabeled analyte molecules to the surface and, ability to monitor weakly bound interactions due to high surface sensitivity. Detailed guidelines for the SPR data analysis has been provided by the researchers and it will be discussed in this module.

Several nanotechniques such as carbon nanotubes, nanowires, quantum dots, gold nanoparticles are increasingly being used for proteomic applications and this field of "nanoproteomics" will be discussed in this module.

As demonstrated here the binding of target protein to functionalized carbon nanowires lead to detectable changes in electrical conductance. The phenomenon of Diffraction and Interference and use of diffraction-based biosensors will be discussed with Prof. Cynthia Goh from University of Toronto.

These techniques holds great promise to become a technically robust and user-friendly platform for clinical and diagnostic studies.

This century is considered as century of biology, in which life science research is undergoing a profound transformation by employing various omics technologies.

Unraveling structural and functional details of proteins at the proteome scale is very daunting task. However, Proteomics has come to mean virtually everything in protein research and it has quickly evolved to become an integral aspect of human biology and medicine.

Today I gave you an overview of modules and lectures, which will be discussed in this course. It was not possible to discuss all the modules and all the lectures but during the subsequent lectures I will take you to a journey of protein and proteomics research by providing basic concepts and details of proteomic techniques. I hope it will enthuse you to learn about proteomic techniques and proteomic concepts.

Thank you for your attention!