LECTURE-3

Protein Chemistry to proteomics

HANDOUT

PREAMBLE

Proteins are the most dynamic and versatile macromolecules in a living cell, which regulates essential activities of the cell. The classical protein chemistry studies aimed at isolation, identification and functional elucidation of protein. Protein chemistry dealt mainly with sequence retrieval of the amino acids, followed by generating X Ray crystallographic image of the same and finally elucidating the function of the protein. However, as the technology advanced, a new discipline proteomics came into the existence, which aimed to look into the protein properties from a global perspective, i.e., not undertaking one protein at a time, but an entire set of proteins in the milieu. Proteomics mainly deals with the interaction of proteins amongst each other, thereby attempting to create a modeling strategy to answer systems biology of the organism as a whole. The advancement in technologies like mass spectrometry and electrophoresis, and availability of genome sequences and bioinformatics tools was major contributor responsible for transition from protein chemistry to proteomics.

OUTLINE OF LECTURE

1. Introduction
2. Evolution of proteomics
    a) Advancement in Mass Spectrometry
    b) Electrophoresis

c) Completion of genome sequence projects and introduction to bioinformatics

d) Introduction of protein microarrays

3. Proteomics vs. Protein chemistry

BOX-1: TERMINOLOGY

1. **Difference gel electrophoresis (DIGE):** An electrophoretic technique that allows more than one protein sample to be run simultaneously on a single 2-DE gel by carrying out differential labelling of each sample. This helps in eliminating any gel-to-gel variations and simplifies the process of analyzing large number of samples. An internal standard consisting of equal amounts of all samples being run in the experiment is also run thereby further reducing any variations.

2. **IPG strip:** Commercially available immobilized pH gradient (IPG) gel strips have replaced tube gels and have considerably facilitated the process of isoelectric focusing by eliminating the tedious steps of gel preparation and pH gradient establishment using ampholyte solutions. These strips, available across the pH range, contain a preformed pH gradient immobilized on a precast polyacrylamide gel placed on a plastic support.

3. **Protein microarrays:** These are miniaturized arrays normally made of glass, polyacrylamide gel pads or microwells, onto which small quantities of many proteins are simultaneously immobilized and analyzed. Protein microarrays can be generated by either traditional cell-based methods or more recently developed cell-free methods

4. **Ionization source:** This is responsible for converting analyte molecules into gas phase ions in vacuum.

5. **Mass analyzer:** The mass analyzer resolves the ions produced by the ionization source on the basis of their mass-to-charge ratios.

1. INTRODUCTION

The central dogma of life places translation process and protein products at the final stage. Protein play most important role in determining the role of the cell in the system. The DNA sequence determines the sequence of amino acids in the protein, but it is actually the modifications in the proteins resulting from alternate splicing and posttranslational modification, which finally dictates the physiological function of the protein. The structure and the function of the proteins are closely associated. In fact, it is the 3D structure of the protein, which governs the function of the protein. Hence an approach should be available so that all information regarding proteins can be obtained in high throughput manner.

Proteomics differs from the conventional protein chemistry approach of identification, structural and functional elucidation of proteins, in many aspects. Proteomics aims to decipher the protein properties from a global perspective, not taking one protein at a time. This approach becomes a holistic model, while studying the system, because in a system the protein in question is not alone. It is associated and interacting with several other proteins and bio-molecules, which together determine its role in the system. These two approaches, protein chemistry and proteomics, also differ in the techniques that are employed for these approaches as well as scale of usability. While traditional protein chemistry also employed techniques like electrophoresis and mass spectrometry, it is the advancement of these techniques, which led to the divergence from protein chemistry to proteomics. The sensitivity, resolvability, robustness and the high throughput approaches made the transition possible. The completion of genome sequence projects and advancement in bioinformatics and

microarrays were also major catalysts for the advancement of proteomics. These advancements in area of protein chemistry and biology gave rise to this new discipline of proteomics, and with continuous evolution of several advanced technologies, this field is continuously advancing.

2. EVOLUTION OF PROTEOMICS

The number of protein coding genes in the human genome is approximately 5% of the entire genome. It is an astonishing fact that how these 5% of the genes account for the entire diversity of proteins in the cell. It is the dynamic properties of proteins, which lead to several diseases. For example, cell cycle regulating proteins or cyclin dependent kinases phosphorylate activating the cyclins, which promote the cell cycle by allowing DNA replication to pass through the check points. Proteomics emerged from protein chemistry solely because of the advancement in the existing technologies. In the following sections, a more detailed description of the advancements is explained.

2.1 ADVANCEMENT IN MASS SPECTROMETRY

Mass spectrometry has been into existence during last several decades, since the days of protein chemistry. Mass spectrometry ideally measures the mass of an analyte by producing charged molecular species in vacuum, and their separation by magnetic and electric fields based on mass to charge (m/z) ratio. The pre-requisite for a mass spectrometry analysis is the generation of ionized particles. However, proteins being large soluble polymers of amino acids could not be ionized by the conventional gas chromatography without fragmenting it into constituent amino acids. This for the time being limited the usage of mass spectrometry in protein chemistry. However, with major discoveries of soft ionization techniques like MALDI and ESI, the mass spectrometry became a robust analytical technique for protein study.

With advancement in ionization techniques, sophisticated mass analyzers and detectors, the mass analysis of as low as 1 ppm with excellent resolving power is possible. The sensitivity of the mass spectrometry also increased, making it possible to detect the attomole of substances. The tandem mass spectrometry involving two mass analyzers like Tof-Tof/ Q-Tof etc. helped in protein sequencing. The Edman degradation process of amino acid sequencing though was extremely useful, was not high throughput as it could sequence a stretch of maximum 40 amino acids, whereas mass spectrometry could do it for large proteins. The ability of mass spectrometry for relative and absolute quantitation of proteins by employing iTRAQ, ICAT and SILAC labels have definitely advanced the field of quantitative proteomics.

Illustration: Mass spectrometric analysis Vs Edman degradation

Protein analysis by MS was challenging due to complete degradation of samples due to the hard ionization techniques. This limitation was overcome by development of soft ionization techniques such as MALDI and ESI. The vaporized sample is ionized by means of an electron beam in ESI or by a laser beam in MALDI. This results in charged peptide fragments, which get accelerated towards the mass analyzer. These two techniques greatly impacted 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. Protein sequencing by MS, however, is much faster, which allows large number of samples to be analyzed in same amount of time.

2.2 Advancement in Electrophoresis

Electrophoresis refers to the process of separation of charged particles under the influence of an external electric field. Electrophoretic separation of proteins had been used widely for quite a long time; however, recent advances in electrophoretic techniques in the form of protein separation, staining and detection have advanced this technique for further usage in proteomics.

The first advancement came in the transition from tube gels to immobilized pH gradient strips. Earlier, isoelectric focusing was performed using tube gels, which had biggest disadvantage of their stability. The pH would often change and result in erroneous results. The tube gels suffered from extremes of variations, discontinuity in the entire gel, which led to its breakage when concentrated samples were added. Prof. Angelika Gorg has made a remarkable contribution in the field of electrophoresis by substituting the tube gels with immobilized pH gradient strips. These were Acrylamide coated plastic strips containing immobilins of various pH spread across them. The biggest advantage of these strips was in the stability of the pH ampholytes inside the gel. As a result gel to gel variations got reduced and the physical stability of the strips were enhanced, as much as, they can be stored after the first dimension for a week at -20^0C before the second dimension can be done.

Staining techniques also increased the usability of gel electrophoresis in proteomic studies. The conventional coomassie brilliant blue dye had the limitation of 40 μg protein requirement. Although silver staining increased the sensitivity but it had its own

limitation in terms of background staining and incompatibility with mass spectrometry for the protein identification. To overcome these problems, cyanine dyes were introduced for staining which have extreme sensitivity as well as the specificity. The cyanine dyes interact with the protein with their hydroxysuccinamide residues and fluoresce when subjected to light of appropriate wavelength. This property of cyanine dyes is utilized into developing a technique known as DIGE (Difference in gel electrophoresis). The biggest advantage of DIGE was in its ability to resolve gel-to-gel variation. Since an internal standard is included in the gel, there is no need to perform electrophoresis on many samples. The sensitivity of the dyes, coupled with the latest software could thus even check the smallest amount of differential expression level of a protein with reliability. Although electrophoretic techniques such as 2DE is now considered to be primitive in proteomics; however, quantitative approaches such as 2D-DIGE still hold good place for quantitative proteomic analysis.

Illustration: Immobilized pH gradient (IPG) strips Vs tube gels

The pH gradient in tube gels was established by means of ampholyte solutions, which consist of low molecular weight organic acids and bases that are subjected to an electric field. These gradients were not always very stable and tend to break down upon addition of concentrated samples. Analysis of the same protein mixture by 2-DE using tube gels resulted into lot of variation in results across gels. The problem of reproducibility has been overcome to a large extent by the development of IPG strips, which are commercially manufacture gel strips having a preformed pH gradient. These strips only need to be rehydrated before use for 2-DE. Minimal gel-to-gel variation is observed when the same sample is run by 2-DE using IPG strips, thereby making them extremely suitable for large-scale proteomic applications.

2.3. COMPLETION OF GENOME SEQUENCE PROJECTS AND INTRODUCTION TO BIOINFORMATICS

The Human genome Project was headed by 6 countries and was completed in 2003 by 20 laboratories throughout the globe. The human genome project estimated the amount of coding genes in our genome to be roughly around 25,000. If each gene has at least two splice variants and each protein has at least two post translational modifications, then the total protein pool will have 1,00,000 proteins. Imagine the vast data so generated from the human genome project. To handle such vast data many bioinformatics approaches should be used. Bioinformatics led to the development of huge databases, which were designed to store the information that could be accessible to any laboratory. Along with the human genome project many organism's genome also got sequenced under other genome projects. A combination of informaticians, biologists and chemists together led to the development of this entire branch of bioinformatics. The need for development of extensive databases, software and tools arose from the completion of the sequence projects of many organisms. The concept of reverse genetics and engineering also popularized with the availability of these genome sequence data. *De novo* sequencing of proteins using mass spectrometry based approach led to the field of reverse genetics, where by knowing the sequence of the protein, one can elucidate the sequence of the genome.

Illustration: Completion of several genome sequence projects

The genomic DNA is cleaved using a suitable restriction endonuclease and inserted into the bacterial artificial chromosome. The amplified sequences are sequenced using an automated sequencer and then mapped by aligning the overlapping fragments to obtain the original DNA sequence. 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.

2.4. INTRODUCTION TO PROTEIN MICROARRAY

The success of DNA microarray also led the development of a similar approach for protein microarrays. Like DNA microarray, the idea was to immobilize proteins in the chips or antibodies against the proteins on the chips and then detect them. However, this success proved to be short-lived and extremely costly. Unlike DNA microarray, where a piece of the DNA is immobilized into the chip, in protein microarray the entire proteins needed to be immobilized with its function still intact, or else, if its 3D structure is destroyed it would also destroy the binding site for antibodies for detection. The printing of so many proteins proved to be extremely cumbersome and hence led to the establishment of various forms of protein microarray where instead of proteins the DNA was printed and the protein was translated *in situ* for detection.

3. PROTEOMICS vs. PROTEIN CHEMISTRY

The debate still persists whether proteomics emerged from protein chemistry or is it a new field. Although proteomics employs the basic techniques and principles of protein chemistry its approach and ultimate goals are different. While protein chemistry has always dealt with single proteins, proteomics is more interested in studying the global proteome as a whole, making its usability for high-throughput studies. The approach of protein chemists has always been targeted: isolation of proteins followed by structural elucidation and functional analysis. On the other hand, proteomics has always taken up the challenge of studying large number of proteins together, and is more interested in the interactions between the proteins, which ultimately helps to understand the function of the protein. In a way, this approach leads to long-term goal of understanding the role of the protein in the cell or the organism. The protein chemists have emphasized on the structural aspects of the protein, while the proteomics researchers have aimed to look upon the proteins in terms of developing a mathematical model for the system as a whole, using a systems biology approach.

Protein chemistry began its sequence retrieval using Edman degradation technique, whereas proteomics does the same using mass spectrometry. Edman degradation method still holds good but only when there is limited number of proteins; however, for high throughput studies the mass spectrometry based sequencing has become the method of choice. The additional benefit of mass spectrometry based sequencing is that there is no need for sequencing the entire protein. Few peptide sequences and bioinformatic tools are sufficient to identify the protein. In summary, protein chemistry

and proteomics have unique outlook of the problem, and methodology employed to address the problems; however, both are complimentary to each other.

3. REFERENCE

1. Gorg, A. et al. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 2000, 21, 1037-1053.
2. Gorg, A. et al. Two-dimensional polyacrylamide gel electrophoresis with immobilized pH gradients in the first dimension (IPG-Dalt): the current state of the art and the controversy of vertical versus horizontal systems. *Electrophoresis* 1995, 16, 1079-1086.
3. Gorg, A et al. 2-DE with IPGs. *Electrophoresis* 2009, 30, S122-S132.
4. Discovering Genomics, Proteomics & Bioinformatics, 2nd edition, A.Malcolm Campbell & Laurie J. Heyer.