## LECTURE-08

## PROTEOMICS AND SYSTEMS BIOLOGY

### Transcript

Welcome to the proteomics course. Today we will talk about proteomics and systems biology.

The lecture outline, first I will start on proteomics and then we will move on to systems biology. Today we will continue on first proteomics and then how different type of omic data can be applied for systems approach analysis that will be discussed.

In this slide I have shown you how you can define different type of omic platforms. So omic is a suffix derived from the Greek work ome all or every. The omics is used as a suffix which has enabled the explosion of terms genomics, transcriptomics, proteomics and metabolomics and so on. The omics also implies an integration of biology with information science and conveys large scale biology using systems approach. As you can see in this slide if you are studying about DNA that will be in totality known as genomics, RNA study is transcriptomics, proteins in proteomics and metabolites in metabolomics. If you are looking at all of the similar contents of proteome that will be known as cellular proteomics or cellular genomics and similarly all of the proteome of an organism will be known as global proteomics or similarly at the gene level global genomics.

Let's first start with proteomics.

If you remember from the previous lecture proteome is set of all the proteins which are expressed by a genome, proteomics is study of full set of proteins and their properties to provide an integrated view of cellular processes. What are these different properties? These properties include the extent of protein expression, how different type of post-transcriptional modifications occurs in the cell, different type of enzyme regulation whether it is activation or inactivation and then different type of intermolecular protein-protein interaction. The current goals of proteomics are very broad, including the diverse properties of proteins which we have discussed in the earlier lectures looking at the, side chain chemistry of amino acids and different levels of protein structures so whether it sequence, quantity, the state of modification, activity, interactions of proteins with other proteins and other molecules and sub-cellular distributions and structure analysis all of these are broad goals of proteomics.

Central dogma concept was discussed earlier which is an orderly and unidirectional flow of information encoded in base sequences of cells passed on from DNA to RNA and then to proteins this is simplest definition of central dogma. The genome sequencing projects have provided researchers with unprecedented information of genome sequences; however, there are numerous proteins which can be encoded by the genome therefore the analysis of static genome by doing sequencing alone is not sufficient. A gene can code for several type of proteins because of its alternative splicing and post-transcriptional and post-translational modifications; therefore, it suggest that studying proteins is more challenging than genome or transcriptome therefore proteomics has great significance to understand biological systems.

The completion of genome sequencing projects of several organisms including human has been one of the most remarkable achievements of this century however these have not been sufficient to unravel the mystery of complex biological processes. The similar genes numbers of many diverse group of organisms has failed to explain their varying biological complexity. A more meaningful understanding of biological function can be obtained through the characterization of products of gene expression, the protein which serve as ultimate effectors molecule of biological systems.

The proteomics refer to the study of entire protein complement expressed by an organism at any given time, while the genome of an organism is mostly static the proteome is dynamic and it changes with environment and time; thereby alleviating its complexity level. The gene regulation is regulated by several post-transcriptional and post-translational modifications due to which the number of proteins expressed in a cell is much greater than its genomic counterparts. The proteomics aims to decipher structure and function of all the proteins in a given cell under specific conditions and to obtain a global view of cellular processes at the protein level, study at the DNA level is known as the genomics, RNA level – transcriptomics and protein level – proteomics. Analysis of the proteome involves protein extraction, separation, identification and finally characterization of various proteins.

There are various proteomic techniques which are currently employed in different applications. We will talk in detail about different proteomic technologies in the subsequent modules of the course but briefly very broadly you can group these technologies as gel-based methods, gel-free MS based methods, Mass spectrometry techniques, techniques to study protein interactions and structural proteomics.

As I have shown some abbreviations in this slide, broadly different type of gel based proteomics such as Sodium Dodecyl Sulfate Polyacyrlamide Gel Electrophoresis (SDS-PAGE), Two dimensional electrophoresis (2-DE), Difference in gel electrophoresis (DIGE), Blue native page (BN-PAGE) as well as different type of staining methods such as coomassie, silver, fluorescent dyes, phosphor-stains such as Pro-Q diamond and multiplex staining methods all these can be grouped under gel-based methods.

Gel-free MS based methods includes SILAC which is stable isotope labeling by amino acids and cell culture, CDIT- culture derived isotope tags, ICAT- isotope coded affinity tags, VICAT- visible isotope coded affinity tagging, MCAT- mass coded affinity tagging, QUEST is quantitation using enhanced signal tags, ITRAQ- isobaric tagging for relative and absolute tagging, GIST- global internal standard technology, ICPL- isotope coded protein labeling, AQUA- absolute quantitation, SISCAPA- stable isotope standards captured by antipeptide-antibodies, COFRADIC- combined fractional diagonal chromatography and MudPIT which is multi-dimensional protein identification technology. All of these are various new advancements in the gel-free methodologies.

The basic mass spectrometry which is central to the proteomic application includes different type of ionization sources such as matrix assisted laser desorption ionization-(MALDI), electrospray ionization (ESI), and different type of mass analyzers such as quadrupole, time of flight, ion trap and Fourier transform mass spectrometry then different type of tandem MS-based systems are also used. The surface enhanced laser desorption ionization-time of flight (SELDI-TOF) is also used for various clinical applications.

The protein interaction methodologies include immune-precipitation, yeast two hybrid method and different type of protein microarray platforms such as antibody arrays, nucleic acid programmable protein arrays, multiple spotting techniques and various other cell-based and cell-free expression based protein microarrays. The detection can be either based on labeled using fluorescence, chemiluminescence or radioactivity or it could be different type of label-free methods such as surface plasmon resonance, interferometry based methods or different types of conductance based methods employing nanotubes and nanowires.

The structural proteomics it involves X-ray crystallography, Nuclear magnetic resonance (NMR), transverse relaxation optimization spectroscopy-TROSY, circular dichroism-CD, and different type of microscopy methods including atomic force microscopy and electron microscopy.

So as we have seen there are large numbers of proteomic technologies, which are currently available for various applications. Many times to address just one biological question different type of methodologies come together and then provide solutions to that problem. For example looking at some clinical sample for identification of biomarkers of a disease one can employ the samples such as tissue or blood or different type of body fluid and then either directly extract the protein and subject to the mass spectrometry or first resolve on 2-DE followed by identification on mass spectrometry or take these samples and directly apply on the microarray based platforms and then detect using label-based or label-free methodologies. Eventually these types of results will enhance knowledge for monitoring the therapy response as well as identification of early disease diagnosis. This is just one example similarly multiple types of proteomic technologies can be used for different applications.

There are several proteomics techniques, which are employed for studying these proteins such as two dimensional gel electrophoresis, mass spectrometry, protein microarrays as well as some label-free detection techniques such as surface plasmon resonance.

After discussing proteomics, now let's talk about systems biology. So what is systems biology? An estoreic knowledge? A method to understand biological systems or a tool to solve the practical problems?

Systems biology is the examination of a biological entity as an integrated system rather than study of its individual characteristic reactions and components which is termed as systems biology.

Study of all the mechanisms underlying complex biological processes in the form of integrated system of many interacting components is studied under systems biology.

Systems level understanding require information from different levels- as you can see DNA to RNA to proteins form a system and then that information can be applied to understand a complex system for different organism. The biological information is represented by the networks of interacting elements and dynamic responses to the perturbations. These networks provide insights which cannot be analyzed from the isolated component of the system. The common elements of systems biology include networks, modeling, computation and dynamic properties.

The different type of biological networks such as protein-protein interaction networks, gene regulatory networks, protein-DNA interaction networks, protein lipid, protein-other molecules network and metabolic networks.

The various ingredients of systems biology, for example if you are studying about a cell and the systems behavior one need to look at the genome, its transcriptome profile, proteome profile, how protein-DNA and different type of transcriptional networks are altered, protein-protein signaling networks, multimeric complexes how they are formed, how protein is localized in the intracellular dynamics and metabolic networks. All of these are ingredients of studying about a system.

Now systems biology study can be done at different levels for example to study the complex physiology of human one can look at individual systems such as respiratory systems, nervous systems or other physiological systems. Studies can be done at the intercellular or intracellular level and finally at the molecular level involving genomics, transcriptomics and proteomics.

So why there is need for systems biology? The study of biology at the system or subsystem level for understanding the biological processes and network is very much required. As you can see, to understand even simpler system of a cell how it is regulated with the extracellular space and the cytoplasm and different other components, examination of structure and dynamics of cellular and organismal function is very much required for understanding of systems rather than characteristics of isolated parts of the cell or the organism.

So what is the aim of systems biology? To understand the biology in holistic approach rather than the reductionist approach. The systems biology aims to quantitate the qualitative biological data and provide some level of predication by applying different type of computational methods.

The systems biology approach involves first of all collection of large experimental data sets and then mathematical models to provide insight of some significant aspects of data sets.

The simple systems biology approach would involve experiments by adding new datasets which will be used for model construction and model analysis and the biological insight derived from these models can be used to propose new hypothesis. So the properties of systems are probably more than just the sum of all its individual properties or its components therefore it is possible that system may have its own property by applying all the components.

So what are different approaches have been taken to study systems biology?

The different approaches of systems biology include the model-based and data-based methods. The model-based approach involves some prior information which can be

implemented in the models whereas the data-based, the objective is to find the new phenomenon. The model-based relies on computational modeling and simulation tools whereas the data-based method relies on the omics data sets.

In model-based systems biology approach it is difficult to build detailed kinetic models but in data-based systems the complex relationship among various type of omics information, metabolic pathways and networks can be created.

Studying systems components is very challenging. Systems biology and biological network modeling aims to understand the systems structure and function for better understanding of system properties like its robustness as well as its use for prediction of systems behavior in response to the perturbations.

The reductionist approach involves disintegrating the systems into its components and studying the whereas, the integrative approach involves integrating the study of individual components to form conclusions about the system.

What is systems biology triangle? First of all the systems information is generated at various levels as we have discussed starting genes to mRNA to proteins to metabolites or identifying regulatory motifs, metabolic pathways, functional modules and different large scale organizations. This information has to be stored and processed and further executed to identify the system level information.

Even simpler systems such as cell can be linked with various properties its genome sequences of different molecules, intracellular signals, transcription factors, different type of cis binding activities the expression profiling of RNA and proteins and different type of cellular processes.

So what approach one can take to study about the systems? Extraction and mining of complex and quantitative biological data, Integration and analysis of these data sets for development of mechanistic, mathematical and computational models, Validation of these models by retesting and refining after proposing some hypothesis.

Different online databases and repositories are nowadays developed and available for sharing large datasets and various systems models. The systematic approach how molecules act together within the network of interaction that make up life is definitely going to be useful to understand the systems biology.

The systems biology triangle as you can see here involves the experimental data sets could be derived from different type of omics platforms, technologies how the computational analysis can be  performed, different type of bioinformatics softwares and tools and then computational modeling by obtaining some theoretical knowledge. The

synergistic application of the experiment, theory and technology with modeling to enhance the understanding of biological processes as whole system rather than the isolated part is termed as systems biology triangle.

In systems biology triangle the wet lab experiments or bioinformatics based data analysis can be used to propose a model. The model building as an aid to understand complex system and some hypothesis can be generated which could be used further to propose more quantitative models or predictive models and also it can be used for independent techniques for the model validation.

What does system study? First of all the difference between systems study and component study one need to understand and what we have tried to emphasize in previous slides, after generating the data set and creating the systems biology triangle then this information can be used for understanding the systems in the more complex and mechanistic level.

**System study and model building**

The system science include synthesis, modeling concepts, analysis; life sciences discipline provides quantitative measurements, genetic modifications and deriving some hypothesis; the information sciences enables the visualization, the modeling tools and different databases. This model building as an aid to understand the complex systems is very useful for systems level investigation.

System is an entity which maintains its existence through a mutual interaction of its constituent parts. Systems biology research consists of the: Identification of the Parts, Characterization of the components, Exclude the ones which are not a part of the system, Identify the interaction of the components with each other, Identify the interaction of the components with the environment which modulate the parts either directly or indirectly through modulation of internal interactions

The systems biology concepts can be understood with help of two approaches, namely, the "Reductionist" approach and the "Integrative" approach. "Reductionist Approach" focuses on disintegrating the system into its component parts and studying them, whereas, Integrative Approach focuses on integrating the study of individual components to form conclusions about the system.

Consider a cell with its component molecules. Let's say we want to study the "metabolic pathway" as a "biological system". When the environment of the cell is perturbed a little, the individual components undergo unique changes, such as increase in production rate, or decrease in their amount. At this stage, due to lack of knowledge

of the nature of interaction of proteins, we cannot interpret how the system gets affected.

But when we study the interaction of one component with the other, we can conclude, that the increase in rate if DNA binding protein leads to increase in the synthesized amount of DNA, which further changes the final amount of lipo-protein produced. Thus we can see that to study a system, we need to analyze not just the components, but their interactions. These biological systems can be: Protein-protein interaction networks, Gene regulatory networks, Protein-DNA networks, Protein lipid interactions, Metabolic Networks.

To study the systems we need to know about the components and its interactions. The data about the components comes from Genomic and Proteomic studies. The data about the molecular interactions comes from interactomics studies.

Here it is shown, that is a systems approach- experiment, technology and computational modeling, this triangle is very important. This has to be linked with the theory to form a systems triangle. A typical work flow for approaching a biological problem with a systems biology perspective is selection of biological significant problem, creation of a model for the problem which represents computable set of assumptions and hypothesis that will be subjected to experimental validation. Compatibility of the model with established experimental facts will reveal the adequacy of assumptions made. If incompatible the model will be modified or rejected. The systems analysis is connducted on the consistent models to make predictions for the system. From all the predictions a small is selected to validate them using wet lab experiments which are designed for this purpose. The data from the experiments is analyzed and inadequate models are discarded.

There are different technologies which have been employed to study the systems biology. Obviously you need high throughput data sets, which could be derived from microarray platforms, RNA deep sequencing, different configuration of MS, different type of proteomic tools and protein interaction data sets.

Some of the technologies which are commonly employed in systems biology can be classified broadly under the following techniques:

For genomics the high throughput-DNA sequencing methodologies, mutation detection using SNP method; for transcriptomics the transcript measurement can include serial analysis of gene expression-SAGE, gene chips, microarrays and RNA sequencing; for proteomics MS, 2-DE, protein chips and yeast two hybrids and different structural proteomics tools such as X-ray and NMR.

X-ray and NMR are mainly employed for metabolic analysis, the metabolomics.

So as you can see here to generate the systems level information, the systems study requires different technologies which could be employed in the biological systems. At genome level by studying different type of technologies using HT-sequencing, HT-arrays, transcriptomics different type of transcriptome analysis using RNA sequencing and microarrays, proteome we discussed many methodologies, metabolome could be using either NMR or MS and in phenome which is studying about the images by using NMR method. So each level of these omic technologies can be useful for studying systems biology.

Let's now talk about how to model biological networks. To build a model in systems biology, first of all parts list can be generated by data sets derived from the systems biology approaches. The systems or subsystems model can be generated which can be used for systems model analysis. Now this could be applied for the real systems and by applying knowledge, using bioinformatical tools it can again applied back to original component which could be used to derive some hypothesis and validation of these data sets. It will work like a close loop.

To build the models in systems biology, information is generated at different levels. Level 1 such as DNA and gene expression, level 2 the intracellular networks, level 3 cell-cell and transmembrane signals, and level 4 integrated organ level information.

What are the frameworks required for modeling schemes? Different types of deterministic or stochastic models have been proposed. The compartmental variables or individual or functional variables have been studied. The especially homogenous or specially explicit models are generated which could be applied in the uniform time scale or separated time scale. This framework would involve single scale entities or cross scale entities.

As you can see here, this framework requires different level of information in very complex manner whether it is curation of the databases, how to align these information using bioinformatic tools to generate the predictive models which could be also developed by using the literature curated data sets or experimental data sets and finally it could be used to study the systems level properties.

Let's discuss the workflow of mathematical modeling. A paradigm can be proposed based on modify, model- measure-mine. Systematic experiments, different type of molecular genetics, chemical genetics and cell engineering approaches can be used for modifying and different level of measurements by applying microarrays, spectroscopy imaging and microfluidics based approaches from proteomics and genomics can be

used further for mining which involves bioinformatics databases and data semantics. Now these datasets could be used to derive the model, which could be reaction mechanistic, statistical or stochastic models. Starting from systematic experiments to reaching and deriving the quantitative models this workflow can be applied.

The modeling of probabilistic processes involves, let's say you want to study a biological system so some experiment has to be performed, the experimental datasets will be generated. From which some statistics can be applied this can be used for the comparison. Now different types of models can be generated using simulations and simulations datasets, which can be used for intermediate statistics and by comparing these two types of information and adjusting the parameter one can study the systems and derive probabilistic processes.

What are ordinary differential equations and stochiometric models? The quantitative analysis measures and names to make models for precise kinetic parameters of a systems network component. It also uses properties of network connectivity. The ODE is a mathematical relation that can be used for modeling biological systems. The quantitative models mostly use ODE to link reactants and products concentrations through the reaction rate constant. To develop the computationally efficient and reliable models of the underlying gene regulatory networks, these ODE models can be used. The stochiometric model, it is modeling a biological network based on its stochiometric coefficients, reaction rates and metabolite concentrations.

We will now discuss "Systems approaches for studying biological networks – from post-transcriptional control to drug discovery" with Dr. Sarath Chandra Janga, an Assistant Professor Indiana University & Purdue University Indianapolis (IUPUI), USA.

Let's try to integrate omics approaches with systems biology. Genome sequencing projects in genomics era from 1990s to 2000 accelerated the pace of omics research. Then from 2000 onwards proteomics field also got accelerated and new methodologies, new tools came in to study proteome. And the data derived from genomics, transcriptomics, proteomics, metabolomics and other omics approaches have now brought the integration of these data sets in the systems biology field.

The systems study requires obtaining dataset from different approaches and analyzing them. For example as shown in the slide the genome wide data sets can be derived at the genome level and looking at the expression of different transcripts or at the proteome level looking at different type of protein interactions. These datasets can be stored in the clinical databases and also it can be mined from the literature, literature manual curation then integration of orthogonal datasets further can be used for

validating the networks and deriving identifying therapeutic targets. Further, it can be used for experimental validation.

Studying systems cannot be done in isolation individual labs it requires different expertise and collaborations from scientists from different disciplines of biology, physics, engineering, chemistry, computer science, mathematics, medicine, statistic and many more.

So eventual aim of systems biology field is to employ the omics level information obtained from different levels from genome, transcriptome and proteome derive that information at the systems level, integrate and quantitate some models and then propose and use it for the understanding the physiology and apply that in medicine. This omics to physiology, flow can be well maintained by employing systems biology tools.

What are the challenges of systems biology? Systems biology is extremely challenging. The emphasis is to understand a system. Understanding dynamics of even simplest biological networks not only requires only the understanding of biology but also its modeling and simulation. The disintegrating study can be used for studying from cells to proteins to gene or integrative study could be used for putting these pieces back together again and then understanding and doing the prediction and control of functional biological process. All of these are very challenging but currently being addressed by applying various systems levels tools.

How proteomics and systems biology are integrated? Proteomics as we have studied, it useful to understand complex signaling networks in biological systems. It is very indispensable tool for systems biology. The global analysis of proteome is important however there are many limitations in each experiment only thousands of proteins can be studied therefore new approaches and systems level investigation and predictions are required. The system investigation is required to study the complex dynamic structure interaction with the biological systems whether it is at cellular level or at the organism level and ultimately it is responsible for their function and behavior.

In summary, today we discussed that how omics era the technological advancement in genomics, proteomics and metabolomics have generated large-scale date sets in all the aspects of biology. These large data sets has motivated the computational biologist and systems approaches with objective of understanding the biological system as a whole. While proteomics continues to generate the quality data at the proteome level so systems biology approach characterizes and predicts these dynamic properties of biological networks. In the next module we will focus on different type of proteomic technologies. Thank you.