LECTURE-1

Genomics to Proteomics

HANDOUT

PREAMBLE

A gene is a stretch of nucleotides that ultimately codes for a polypeptide, which in turn are of structural or functional significance to the cell or the organism. The entire sequence of an organism's hereditary information, including coding and non-coding regions, encoded in DNA is known as "genome". Studying genome, including function and interaction of all the genes of an organism is known as genomics. Success of genome sequencing projects has been remarkable; however, in spite of availability of the entire genome being sequenced, the complex biological processes cannot be unraveled until the role of each of the gene expression products is understood. Proteomics is study of the entire protein complement of an organism, under a given set of conditions. Proteomic studies rely on tools that entail understanding the biochemistry of proteins and the pathways in which these proteins participate in order to bring about a well-orchestrated and harmonious functioning of a given cell or organism in consideration.

OUTLINE OF LECTURE

1. Central Dogma concept
2. Overview of genomic studies and sequencing
3. Post-transcriptional modifications
4. Post-translational modifications

PROF. SANJEEVA SRIVASTAVA
DEPARTMENT OF BIOSCIENCES & BIOENGINEERING
IIT BOMBAY, MUMBAI, INDIA

5. Need for proteomics

6. Genomics vs. Proteomics

BOX FOR TERMINOLOGIES

- **Genome**: The entire sequence of an organism's hereditary information, including both coding and non-coding regions, encoded in its DNA is known as the genome.
- **Contigs**: A set of overlapping DNA fragments that are obtained from a single genetic source. These contigs are used to deduce the original DNA sequence.
- **Sequencing:** DNA fragments that have been amplified using the BAC vectors are sequenced to obtain the nitrogenous bases of each fragment. These are then used to deduce the original sequence of the intact DNA by aligning the fragments having overlapping end sequences.
- **Transcriptome**: The set of all RNA molecules, including mRNA, rRNA and tRNA, present in an organism is referred to as the transcriptome.
- **Proteome**: The entire complement of proteins expressed by the genome of an organism under specific defined conditions is known as the proteome. Like the transcriptome, the proteome of an organism also varies with external factors and conditions.
- **Genomic DNA:** The deoxyribonucleic acid polymeric sequence that acts as the store for genetic information and is essential for the synthesis of RNA and protein molecules, which are necessary for cellular functioning in all organisms.
- **Transcription:** The process by which the genomic DNA is converted into a chemically related molecule, the messenger RNA or mRNA. Several enzymes and other factors are involved in this process. All regions of the DNA, coding and non-coding, get transcribed into the corresponding mRNA.

- **Pre-mRNA:** The mRNA transcript that is produced from DNA as soon as transcription ends is known as the pre-mRNA. This contains both coding and non-coding sequences, is short lived and is further processed before translation.
- **BLAST:** Basic Local Allignment Search Tool for Gene matching as provided online by NCBI.
- **HMM:** Hidden Markov Model, which is a modified Bioinformatic based algorithm for sequence alignment.

1. CENTRAL DOGMA CONCEPT

The central dogma of molecular biology is the concept which governs the flow of biological information from gene to protein. It includes the major steps of DNA Replication, Transcription and Translation. Replication of DNA allows DNA duplication and ensures that a copy of entire chromosome complement is delivered to each of the resulting daughter cells after cell division. This is followed by a complexly controlled, enzyme-dependent step wherein the DNA is transcribed into mRNA. The mRNA strand thus generated is complementary to the DNA strand. The major difference between the two nucleic acids is that the backbone of RNA contains a Ribose sugar instead of the Deoxyribose sugar present in DNA. Besides, the nitrogenous base Thymine is replaced by Uracil in case of RNA. The RNA thus formed acts as a messenger of genetic information and is translocated from the nucleus (the site of transcription) to the cytoplasm. It is the cytoplasm, where the actual process of Translation or converting the genetic information into polypeptide takes place. The processes of transcription and translation are highly complex involving multiple components and are highly regulated. A departure from the normal flow of these steps, consequently affects further downstream processes.

PROF. SANJEEVA SRIVASTAVA
DEPARTMENT OF BIOSCIENCES & BIOENGINEERING
IIT BOMBAY, MUMBAI, INDIA

2. CONVENTIONAL GENOME SEQUENCING METHODS and NEXT-GEN SEQUENCING METHODS

The availability of various tools has made the completion of sequencing of several genomes possible. There have been several genome projects that have aimed at sequencing the entire genomes of organisms, including humans. Several databases are now readily available which have facilitated the identification of gene sequence of a protein, which has been sequenced by mass spectrometry.

Illustration: DNA Sequencing – Sanger's dideoxy method

A simple method for DNA sequencing devised by Frederick Sanger, where a collection of DNA fragments are synthesized by means of controlled interruption of enzymatic replication. Four DNA synthesis reactions are carried out simultaneously with the strand whose sequence is to be determined being used as the template. The reaction mixture consists of regular deoxynucleotides and DNA Polymerase along with a small amount of one labelled dideoxy nucleotide analog being added to each of the four reaction mixtures. A primer is added to begin the DNA synthesis and strand elongation continues until a dideoxy analog gets added instead of the regular dNTP. Chain termination occurs at this stage due to the absence of a 3' OH group for the formation of next phosphodiester bond. The synthesized strands are separated from each other, after which the differentially labelled strands of various lengths are separated by electrophoresis. The smallest fragments move further in the gel while the larger fragments remain close to the point of application. The different fluorescent labels of each ddNTP can then be detected by scanning the gel with a beam of laser. The output sequence obtained is complementary to the template strand, which can be used to deduce the original desired template sequence.

Illustration: Shotgun sequencing

Genomic DNA is cleaved using a suitable restriction endonuclease and the fragments are inserted into bacterial artificial chromosome vectors. These vectors enable the DNA fragments to be amplified. The genomic DNA fragments of the library are then organized into a physical map, after which individual clones are selected for sequencing. The selected BAC is amplified and these clones are sequenced using the Sanger's chain termination method. The sequence of the clone is then deduced by aligning them based on their overlapping regions. The entire genomic sequence is then obtained once each BAC is sequenced in this manner.

Illustration: Next-generation sequencing techniques

In Pyrosequencing multiple round of nucleotide addition are carried out on the immobilized template DNA using DNA Polymerase in the presence of ATP sulfurylase, luciferase and the nucleotide degrading enzyme apyrase. The release of an equal amount of pyrophosphate is determined by its conversion to ATP, by ATP sulfurylase enzyme, which is determined by the release of light on reaction with luciferase. The amount of light produced is determined by means of a CCD camera, which is used to determine the addition of nucleotides & therefore the sequence of the template DNA.

The nanopore sequencing offers a label-free approach for DNA sequencing. An exonuclease cleaves the single stranded DNA, one base at a time to release the nucleoside monophosphates. These NMPs pass through the nanopore under an applied potential, which is covalently coupled to an adapter molecule. Continuous movement of NMPs through the nanopore results in characteristic fluctuation of electric current that enables detection of various nucleotide bases.

Illustration: Genome sequencing projects

Genome sequencing projects have aimed to elucidate the complete genome sequence of organisms. The DNA sequences are identified by the shotgun sequencing technique and then aligned using suitable software to provide the complete genome sequence. The genome sequence of a large number of prokaryotic and eukaryotic organisms has been successfully deduced. The immense amount of information provided by the human genome motivated researchers to understand the nature and content of genetic material in great detail. The shotgun approach was the fundamental technique used for large-scale sequencing of the human genome, which also makes use of Sanger's sequencing. Progress made in sequencing was very rapid and by 2001, a draft of the sequence was ready covering around 83% of the genome. The genome sequencing studies successfully provided many novel findings about the human genome.

BOX: SEQUENCING METHODS

- **Maxam-Gilbert Sequencing:** This method brings about chemical modification of the bases and also requires that the DNA sequence be labeled radioactively at the 5'end. Chemical modification is brought about in such a way that four different types of modified products are generated. The final sequence is deduced by running a SDS-Polyacrylamide gel with the four products in four parallel lanes, which is followed by exposing the gel to X-ray film for autoradiography. Technique is limiting with respect to the use of radioactive labels.

- **Sanger Sequencing:** This is also known as the chain termination method, which relies on the fact that; a modified base (for instance dideoxynucleotides instead of the deoxynucleotides) when incorporated during *in vitro* replication causes the process to cease at that point. Each of the four modified bases is labeled with four different fluorescent labels, which are further detected in automated sequencers.

- The advent of numerous advancements and improvements in the genomics field has led to the development of high-throughput sequencing methods, which have made it possible that thousands of genes be sequenced at a time. These techniques mainly include: Pyrosequencing, Shotgun Sequencing, Sequencing by synthesis etc.

- Currently used sequencing techniques also employ Bacterial Artificial Chromosomes (BACs), which are DNA constructs that are useful for cloning purposes. These cloning vectors can carry DNA inserts of around 150-350 kbp and have been extremely useful in various genome-sequencing projects carried out. The genomic DNA is cleaved using 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. The information obtained through these sequencing projects is documented in genome databases and are extremely useful in correlating gene and protein sequences.

- The use of Next-Gen sequencing platforms has facilitated genomic studies to be carried out on a large-scale and has further expanded the scope of generating information about the structural and functional aspects of various genes. The plethora of attributes associated with the different coding genes can be explored only when the genomic data is supplemented with data obtained from proteomic studies.

- Third Generation Sequencing Technology relies on Scanning Tunnel Electron Microscopy, Fluorescence Resonance Energy Transfer, Single Molecule Real Time Sequencing and Protein nanopores.

3. POST TRANSCRIPTIONAL MODIFICATIONS

Post-transcriptional modifications mainly include the processing of heterogeneous RNA, which is generated as the product of transcription and converted to its mature form. This includes the 5' capping, splicing and 3' polyadenylation. In the 5' processing step, the 5' end of the primary transcript is capped with 7-methylguanosine and requires the aid of enzymes such as phosphatase and guanosyl transferase. This capping protects the primary transcript from the attack of Ribonuclease enzymes, which exhibit specificity for the 3'-5' phosphodiester bonds. In splicing, the intervening non-coding sequences (introns) are spliced out and the exons (coding sequences) are joined together to provide a continuous stretch of coding nucleotide sequence. In many cases, this process occurs concomitantly with transcription itself and makes use of a complex Spliceosome Assembly and small nuclear RNA to bring about splicing. The process by which a given pre-mRNA transcript is spliced differentially and hence is responsible for giving rise to different protein products is known as 'Alternative Splicing'. The 3' processing involves the cleavage of the 3'-end and addition of approximately 250 adenine residues, rendering a poly A tail at the 3'-end of the pre-mRNA. Polyadenylation usually occurs at a site where a polyadenylation signal sequence (5'-AAUAAA-3') is recognized near the pre-mRNA. This tail addition is also responsible for protecting the RNA transcript from the attack of Ribonuclease enzyme. The RNA thus generated after these steps of processing is termed as the mature RNA.

Illustration: Genomic DNA contains large stretches of non-coding regions

Pre-mRNA is synthesized from genomic DNA by the process of transcription. The gene to be transcribed is bound by transcription initiation factors and then by RNA Polymerase, which transcribes the gene in the 5' to 3' direction. The DNA strand that

PROF. SANJEEVA SRIVASTAVA
DEPARTMENT OF BIOSCIENCES & BIOENGINEERING
IIT BOMBAY, MUMBAI, INDIA

gets transcribed is known as the template strand. Once the termination sequences are reached, the enzyme and the newly formed mRNA transcript are released. The genomic DNA that gets transcribed into mRNA contains exons, the coding sequences, as well as introns which are intervening, non-coding sequences. This pre-mRNA has certain recognition sites within its intron sequence that allows the spliceosome assembly to recognize and bind to it. There is a conformation change that takes place upon binding of the protein-RNA complex. The remaining snRNPs bind following the conformation change of the pre-mRNA and there is cleavage at the GU site on the 5' end of the intron. It attaches to the branch site adenine nucleotide near its 3' end to form the lariat structure. The assembly cleaves the 3' end of the intron sequence containing the AG recognition element. The 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.

Illustration: A single gene can give rise to multiple protein products

Pre-mRNA transcribed from genomic DNA is often made up of several coding exons interspersed by non-coding introns. Alternative splicing, a common phenomenon observed in eukaryotes, allows the exons to be reconnected in multiple ways. There are several mechanisms for alternative splicing, the most common being exon skipping, wherein a particular exon may be included in the mature mRNA under specific conditions or in certain tissues and omitted from others. The mature mRNA produced then undergoes translation where it is bound to the ribosome and read as three letter codons. The corresponding amino acids are incorporated with the help of tRNAs. The ribosome moves along the mRNA and continues to incorporate the amino acids to the growing polypeptide until the termination codon is reached. The diversity of proteins encoded by a genome is greatly increased due to alternative splicing. Each mature mRNA formed gives rise to different protein products upon translation. Complexity of the proteome can be understood from the fact that a single gene can code for multiple proteins.

4. POST TRANSLATIONAL MODIFICATIONS (PTMs)

The polypeptide chain made up of several amino acid residues gets released at the end of the translation process and undergoes appropriate folding to attain its secondary and tertiary structures. If the protein is made up of multiple subunits, these come together to form the native protein structure. Many proteins undergo chemical modifications at some of their amino acid residues after translation. These are carried out by enzyme-catalyzed reactions and are essential for normal functioning of the protein. The protein that has undergone the required PTMs and is ready to function is the modified protein in its native, stable state conformation.

Some of the most commonly observed PTMs include:

a) **Phosphorylation:** The addition of a phosphate group, usually to serine, threonine or tyrosine residues of the protein. Protein phosphorylation and dephosphorylation is one of the most important control mechanisms for the inter-conversion of proteins between their functional and non-functional states.

b) **Glycosylation:** The enzymatic addition of saccharides to specific amino acid residues resulting in the formation of glycoproteins. Sugars like glucose and mannose are commonly added to either nitrogen atoms of aspargine, arginine or to hydroxyl oxygen atoms of serine, threonine, tyrosine etc.

c) **Methylation**: Addition of a methyl group, usually at lysine or arginine residues.

d) **Hydroxylation**: Addition of a hydroxyl (-OH) group by the hydroxylase enzymes. Proline is usually the principal residue that is hydroxylated resulting in hydroxyproline, an essential and abundant component of connective tissues like collagen.

Illustration: Post-translational modification of proteins

The protein obtained by translation undergoes folding and various PTMs such as phosphorylation, alkylation, glycosylation, hydroxylation etc. to give the final functional protein. This adds to the complexity of each protein since the functional protein product does not directly correspond to its gene sequence.

PROF. SANJEEVA SRIVASTAVA
DEPARTMENT OF BIOSCIENCES & BIOENGINEERING
IIT BOMBAY, MUMBAI, INDIA

5. NEED FOR PROTEOMICS

The field of proteomics originated from the research and development of the Human Genome Project to understand the proteome for the composition, structure, specific activity patterns and unique properties of proteins essential to provide data that complements the genomic information. Genomics deals only with the studies of the entire gene compendium of a particular cell or organism at a given point of time under a defined set of conditions. It is rightly perceived that Genomics is just the starting page in the book of understanding biological functions and the mechanisms that are responsible for maintaining a myriad of biological processes to run smoothly. The delineated role and function of a particular component cannot be completely understood until the gene products are studied in details. Genomics does not take into consideration the fact that a single stretch of nucleotides could give rise to different protein products when processed in different ways, namely in the intermediate steps that occur during the post-transcription and post-translation. This makes it necessary that the proteome of the cell or organism be studied in order to gain deeper insights into the structural and functional significance of the gene products. It is important to note that the proteome when studied, would include the entire protein complement of the cell or organism, and would involve proteins that have already undergone all the modifications that occur intermittently. Besides that, proteomic studies not only provide information about individual proteins but also about the interactions that occur between different proteins, which are in turn responsible for essential biological processes.

6. GENOMICS vs. PROTEOMICS

The proteome of a cell or organism is highly dynamic, which undergoes many changes in ontogeny and also in different pathological and physiological states that it is exposed to. These variations are appropriately attributed to different post-transcriptional and post-translational modifications that the proteins undergo. The expression data revealed by proteomics is directly related to cellular activity and hence gives the true picture of the cell under consideration. Genomics on the other hand deals with data that is revealed by studying it's genetic information. The genome does not correspond to the final expression product that bears functional significance. A comparative analysis of genomics and proteomics is provided in table.

Genomics	
Advantages	Limitations
• The PCR technique allows gene amplification, which can be further used for various purposes. • *De novo* gene synthesis has enabled the generation of oligonucleotides devoid of non-coding regions. • Availability of various Bioinformatics (e.g. BLAST) based tools has made it possible to identify genes that bear resemblance to each other but belong to different origins.	• The data obtained from gene expression studies is not directly related to the actual cellular activity as it does not take into consideration the post-transcriptional and post-translational modifications. • The gene expression products alone cannot be used to elucidate the functional aspects of the gene and hence require that the corresponding proteins be studied. • The enormous amount of data that is generated through whole genome sequencing requires extensive data curation.
Proteomics	
Advantages	Limitations
• Gives a real picture about the cellular activity, as all the data that is procured is about proteins that have either structural or functional significance to the cell. • Availability of protein sequencing techniques like Edman Degradation or Mass Spectrometry has enabled the elucidation of deeper insights into proteomic studies.	• There is practically no method available for protein amplification as of now. • For synthesizing a protein right from the beginning, it is essential to have the DNA sequence that encodes for the protein in question. • Availability of different Bioinformatic-based algorithms like HMM, which also takes into consideration the different conformations that would be a result of protein folding. • Proteomics studies have limitation with respect to reproducibility.

REFERENCES

1. Pandey, A. & Mann, M. Proteomics to study genes and genomics. Nature 2000, 405, 837-846.

2. Wilkins, M. R., Williams, K. L., Apple, R. D. & Hochstrasser, D. F. Proteome Research: New Frontiers in Functional Genomics 1–243 (Springer, Berlin, 1997).

3. Phizicky, E., Bastiaens, P. I. H., Zhu, H., Snyder, M., Fields, S., Protein analysis on a proteomic scale. *Nature 2003, 422,*208–215.