PROTEOMICS AND GENOMICS

UNIT 1

Introduction to Proteomics

1. Introduction to Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. It encompasses the identification, quantification, and analysis of the proteome, which is the entire set of proteins produced or modified by an organism.

Importance of Proteomics:

- Understanding Cellular Functions: Proteins are crucial for virtually every process within a cell.
- Disease Diagnosis and Therapy: Identifying disease-related proteins helps in developing targeted treatments.
- Drug Development: Proteins are common drug targets.

2. Protein Structure, Function, and Expression

Protein Structure:

- Primary Structure: Linear sequence of amino acids.
- Secondary Structure: Localized conformations like α-helices and β-sheets.
- Tertiary Structure: Three-dimensional folding of a single polypeptide chain.
- Quaternary Structure: Assembly of multiple polypeptide units.

Protein Function:

- Enzymatic Catalysis: Accelerating chemical reactions.
- Transport and Storage: Hemoglobin transports oxygen.
- Structural Support: Collagen provides structural integrity.
- Signaling: Hormones and receptors transmit signals.
- Immune Response: Antibodies protect against pathogens.

Protein Expression:

- Gene to Protein: Transcription (DNA to RNA) and Translation (RNA to protein).
- Regulation: Controlled at transcriptional, translational, and post-translational levels.

3. Proteome Analysis

High-Throughput Proteome Analysis with 2D-IEF:

- Two-Dimensional Gel Electrophoresis (2D-GE):
    - First Dimension: Isoelectric focusing (IEF) separates proteins based on isoelectric point (pI).
    - Second Dimension: SDS-PAGE separates proteins based on molecular weight.
- Applications: Comparing protein expression levels between different samples, identifying protein modifications.

4. Peptide Sequencing with MS-MS Methods

Mass Spectrometry (MS-MS):

- Principle: Measures mass-to-charge ratio of ionized peptides.
- Process:
    - Ionization: Peptides are ionized (e.g., ESI, MALDI).
    - Mass Analysis: Ions are separated based on mass-to-charge ratio.
    - Fragmentation: Peptides are fragmented, and fragment ions are analysed to deduce sequence.

MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization - Time of Flight):

- Principle: Ionizes peptides using a laser and measures their time of flight to a detector.
- Applications: Protein identification, post-translational modifications analysis.

5. Phage Display

Principle:

- Phage Display Technology: Uses bacteriophages to study protein interactions. Proteins or peptides are displayed on the surface of phages.
- Applications: Identifying protein-protein interactions, developing antibodies, and mapping epitopes.

6. Protein Chips

Principle:

- Protein Microarrays: Arrays of immobilized proteins or peptides on a solid surface.
- Applications: High-throughput analysis of protein interactions, enzyme-substrate relationships, and biomarker discovery.

7. Rational Drug Design

Principle:

- Target-Based Approach: Designing drugs based on the knowledge of the biological target's structure.
- Applications: Creating more effective and selective drugs with fewer side effects.

8. Lethal Mutants

Principle:

- Lethal Mutations: Genetic alterations that lead to non-viable organisms.
- Applications: Identifying essential genes, understanding gene function, and potential targets for antimicrobial drugs.

9. Significance and Applications of Proteomics in Biology

Applications:

- Disease Biomarkers: Identifying proteins associated with diseases for diagnostics.
- Functional Genomics: Understanding gene function through protein expression.
- Systems Biology: Integrating proteomics data with genomics and metabolomics for a comprehensive understanding of biological systems.
- Agriculture: Improving crop resistance and quality through proteomics research.

UNIT II

Protein Sequence Databases

Swiss-Prot

Description:

- A curated protein sequence database that provides a high level of annotation, including function descriptions, domain structures, post-translational modifications, and variants.

Key Features:

- High-quality annotations.
- Minimal redundancy.
- Links to other databases.

TrEMBL (Translated EMBL)

Description:

- A supplement to Swiss-Prot, containing protein sequences translated from the EMBL Nucleotide Sequence Database that are not yet integrated into Swiss-Prot.

Key Features:

- Automatically annotated.
- High coverage of newly sequenced organisms.

PIR (Protein Information Resource)

Description:

- Provides protein sequence data and functional information.

Key Features:

- Integration with genomic and proteomic data.
- Comprehensive protein annotation.

UniProt (Universal Protein Resource)

Description:

- Combines Swiss-Prot, TrEMBL, and PIR into a single comprehensive resource for protein sequence and functional information.

Key Features:

- Comprehensive resource.
- Extensive cross-references.
- High-quality annotations.

Structural Databases

CATH (Class, Architecture, Topology, Homologous superfamily)

Description:

- A hierarchical classification of protein domain structures.

Key Features:

- Classifies protein domains into evolutionary and structural relationships.
- Provides insights into protein function and evolution.

SCOP (Structural Classification of Proteins)

Description:

- A database that classifies protein structural domains based on similarities.

Key Features:

- Hierarchical classification.
- Provides a detailed and comprehensive description of the structural and evolutionary relationships.

Molecular Modeling

Homology Modeling

Principles:

- Predicts the 3D structure of a protein based on the known structure of a homologous protein.

Steps:

1. Identify homologous sequences.
2. Align target and template sequences.
3. Build a model based on the alignment.
4. Refine and validate the model.

Applications:

- Understanding protein function.
- Drug design and discovery.

Docking Studies

Principles:

- Predicts the preferred orientation of a small molecule when bound to a protein, which helps in understanding binding affinity and activity.

Using Molegro Virtual Docker:

- A tool for molecular docking that predicts the binding modes of ligands to their target proteins.

Steps:

1. Prepare the protein and ligand structures.
2. Define the binding site.
3. Run the docking simulation.
4. Analyze the docking results.

Key Features:

- High accuracy.
- Efficient search algorithms.

RASMOL

Description:

- A molecular visualization software used to view and analyze protein structures.

Key Features:

- Supports multiple formats (PDB, CIF).
- Allows interactive exploration of molecular structures.

Docking Analysis

Constraints

Description:

- Restrictions applied during docking to simulate real biological conditions.

Types:

- Distance constraints, angle constraints, and torsional constraints.

Data Analysis

Description:

- Evaluation of docking results to understand the binding interactions.

Tools:

- Visualization software, scoring functions, and interaction maps.

Sidechain Flexibility

Description:

- Considering the flexibility of side chains in the docking process to improve accuracy.

Template Docking

Description:

- Using a known protein-ligand complex as a template to guide the docking of similar ligands.

Drug Discovery

Target Identification

Principles:

- Identifying a molecule (usually a protein) that is involved in a disease process and can be targeted by a drug.

Methods:

- Genomic studies, proteomic studies, and bioinformatics analysis.

Target Validation

Principles:

- Confirming that modifying the target affects the disease.

Methods:

- Genetic studies, biochemical assays, and animal models.

Lead Identification

Principles:

- Finding compounds that have the desired biological activity against the target.

Methods:

- High-throughput screening, virtual screening, and fragment-based screening.

Lead Optimization

Principles:

- Modifying the lead compounds to improve their properties, such as potency, selectivity, and pharmacokinetics.

Methods:

- Medicinal chemistry, structure-activity relationship (SAR) studies, and computational modeling.

Clinical Trials

Phase I

Objectives:

- Assess safety, tolerability, pharmacokinetics, and pharmacodynamics.

Participants:

- Small number of healthy volunteers or patients.

Phase II

Objectives:

- Evaluate efficacy, side effects, and optimal dose.

Participants:

- Larger group of patients.

Phase III

Objectives:

- Confirm efficacy, monitor side effects, compare with standard treatments.

Participants:

- Large group of patients across multiple sites.

Pharmacodynamics

Principles:

- Study of the biochemical and physiological effects of drugs and their mechanisms of action.

Key Concepts:

- Dose-response relationships, drug-receptor interactions, and therapeutic index.

This e-content provides a detailed overview of protein sequence and structural databases, molecular modeling techniques, docking studies, and the drug discovery process, including clinical trials and pharmacodynamics. Each topic is covered with its principles, key features, and applications to give a comprehensive understanding suitable for study and research purposes.

UNIT-III: GENOMICS AND BIOINFORMATICS

The Human Genome Project (HGP)

Overview:

- A landmark international scientific research project aimed at mapping and understanding all the genes of the human genome.
- Initiated in 1990, completed in 2003.

Goals:

- Identify all the genes in human DNA (~20,000-25,000 genes).
- Determine the sequences of the 3 billion base pairs that make up human DNA.
- Store this information in databases.
- Improve tools for data analysis.
- Address ethical, legal, and social issues (ELSI) arising from the project.

Applications:

- Advancements in medicine through personalized medicine.
- Understanding genetic diseases and developing new treatments.
- Evolutionary biology and anthropology research.
- Enhancing agricultural practices.

Databases

International Nucleotide Sequence Database Collaboration (INSDC)

A collaboration among three major nucleotide sequence databases:

1. **GenBank (USA)**
2. **EMBL (Europe)**
3. **DDBJ (Japan)**

Purpose:

- To collect, annotate, and make available nucleotide sequence data to the scientific community.

GenBank

Overview:

- A comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation.
- Maintained by the National Center for Biotechnology Information (NCBI).

Features:

- Sequence submission and retrieval.
- Tools for sequence alignment and analysis.

European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database

Overview:

- A primary source for nucleotide sequences from Europe.
- Managed by the European Bioinformatics Institute (EBI).

Features:

- Integrates various types of biological data.
- Provides tools for sequence analysis and visualization.

DNA Data Bank of Japan (DDBJ)

Overview:

- A primary database for nucleotide sequences from Japan.
- Operated by the National Institute of Genetics (NIG).

Features:

- Collaborates with GenBank and EMBL to provide comprehensive data.

National Center for Biotechnology Information (NCBI)

Overview:

- A key resource for molecular biology information.
- Hosts GenBank and other databases like PubMed, Protein, and Genome.

Features:

- BLAST (Basic Local Alignment Search Tool) for sequence comparison.
- Entrez system for integrated access to literature, sequence, and structure data.

Sequence Comparison Techniques

BLAST (Basic Local Alignment Search Tool)

Principles:

- Compares nucleotide or protein sequences to sequence databases.
- Identifies regions of similarity to infer functional and evolutionary relationships.

Types of BLAST:

- **BLASTN:** Nucleotide vs. nucleotide.
- **BLASTP:** Protein vs. protein.
- **BLASTX:** Translated nucleotide vs. protein.
- **TBLASTN:** Protein vs. translated nucleotide.
- **TBLASTX:** Translated nucleotide vs. translated nucleotide.

Applications:

- Identifying homologous genes.
- Annotating genes and predicting functions.
- Understanding evolutionary relationships.

Comparative Genomics

Overview:

- The study of similarities and differences in the genomes of different species.
- Aims to understand the structure, function, and evolution of genomes.

Phylogeny

Principles:

- The study of evolutionary relationships among species.
- Uses genetic, morphological, and biochemical data to construct phylogenetic trees.

Applications:

- Tracing the evolution of genes and species.
- Understanding the genetic basis of traits and diseases.

Synteny

Principles:

- The conservation of blocks of order within two sets of chromosomes that are being compared from different species.
- Indicates evolutionary relationships and shared ancestry.

Applications:

- Comparing grass genomes to understand crop evolution.
- Identifying genetic elements crucial for certain traits.
- Enhancing crop breeding programs.

Databases and Tools in Comparative Genomics

INSDC Databases:

- Provide a vast repository of genetic sequences essential for comparative studies.

NCBI Tools:

- BLAST for sequence alignment and comparison.
- Genome Data Viewer for visualizing genomic data.
- HomoloGene for identifying homologs among different species.

Phylogenetic Analysis Tools:

- MEGA (Molecular Evolutionary Genetics Analysis)
- PhyML (Phylogenetic estimation using Maximum Likelihood)
- RAxML (Randomized Axelerated Maximum Likelihood)

Synteny Analysis Tools:

- SynMap from CoGe for synteny visualization and analysis.
- MCScanX for identifying syntenic blocks and duplications.

Functional Genomics

Expressed Sequence Tags (ESTs)

Definition:

- Short DNA sequences (~200-800 base pairs) generated by sequencing either one or both ends of an expressed gene.

Purpose:

- To identify gene transcripts.
- To provide a resource for gene discovery and annotation.

Applications:

- Gene mapping and genome annotation.
- Comparative genomics and evolutionary studies.
- Identifying gene expression patterns in different tissues or conditions.

Serial Analysis of Gene Expression (SAGE)

Principles:

- Captures short sequence tags from cDNA fragments.
- Tags are concatenated into long DNA molecules, cloned, and sequenced.
- Each tag corresponds to a specific transcript, allowing quantification of gene expression.

Steps:

1. mRNA isolation and cDNA synthesis.
2. Tagging cDNA with restriction enzymes.
3. Linking tags together (concatenation).
4. Cloning and sequencing concatenated tags.

Applications:

- Quantitative analysis of gene expression.
- Comparison of gene expression profiles between different samples.

Shotgun Libraries

Principles:

- Randomly breaking genomic DNA into small fragments.
- Cloning fragments into vectors for sequencing.

Process:

1. DNA fragmentation using mechanical or enzymatic methods.
2. Cloning fragments into vectors (e.g., plasmids).
3. Sequencing cloned fragments.
4. Assembling sequences using computational tools.

Applications:

- Whole genome sequencing.
- Metagenomics and environmental DNA studies.

Conventional Sequencing Methods

Sanger Sequencing

Principles:

- Chain termination method using dideoxynucleotides ($ddNTPs$).
- Incorporates $ddNTPs$ during DNA synthesis, terminating elongation.

Steps:

1. DNA denaturation and primer annealing.

2. Extension with a mixture of $dNTPs$ and $ddNTPs$.
3. Separation of fragments by gel electrophoresis.
4. Detection and sequence reading.

Applications:

- Sequencing single genes or small genomes.
- Validating sequences obtained by other methods.

Maxam & Gilbert Sequencing

Principles:

- Chemical cleavage method.
- DNA is chemically modified and cleaved at specific bases.

Steps:

1. End-labelling DNA with a radioactive tag.
2. Chemical treatment to modify specific bases.
3. Cleavage at modification sites.
4. Separation of fragments by gel electrophoresis.
5. Autoradiography to read the sequence.

Applications:

- Historical method with limited use today.
- Provided early sequencing data for small DNA fragments.

Automated Sequencing

Principles:

- Based on Sanger sequencing.
- Uses fluorescently labelled$ddNTPs$ for detection.

Process:

1. DNA denaturation and primer annealing.
2. Extension with fluorescently labeled $ddNTPs$.
3. Capillary electrophoresis to separate fragments.
4. Laser detection and computer analysis to read sequence.

Advantages:

- High throughput and accuracy.
- Suitable for large-scale sequencing projects, such as the Human Genome Project.

Analysis of Single Nucleotide Polymorphism (SNP) Using DNA Chips

Principles:

- DNA chips (microarrays) contain probes for specific SNPs.
- Hybridization of labeled DNA samples to the probes.
- Detection and analysis of hybridization patterns.

Steps:

1. DNA sample preparation and labeling.
2. Hybridization to the microarray.
3. Washing to remove non-specific binding.
4. Detection of bound probes (fluorescent or other labels).
5. Data analysis to identify SNPs.

Applications:

- Genotyping and identifying genetic variations.

- Studying genetic associations with diseases.
- Personalized medicine and pharmacogenomics.

Summary

Functional Genomics:

- **ESTs**: Identify and annotate gene transcripts, useful for gene mapping and evolutionary studies.
- **SAGE**: Quantitative gene expression analysis, comparing expression profiles.
- **Shotgun Libraries**: Random DNA fragmentation and sequencing for whole genomes and metagenomics.

Sequencing Methods:

- **Sanger Sequencing**: Chain termination method for single genes and small genomes.
- **Maxam & Gilbert Sequencing**: Chemical cleavage method, historically significant but now less common.

Automated Sequencing:

- High-throughput, accurate sequencing using fluorescent labels, essential for large-scale projects.

SNP Analysis Using DNA Chips:

- DNA microarrays detect and analyze SNPs, aiding in genotyping, disease association studies, and personalized medicine.

This content provides a comprehensive introduction to the key topics in functional genomics, including various sequencing techniques and methods for analyzing genetic variations. Each method's principles, processes, and applications are outlined to facilitate understanding and practical application in genomic research.