DEPTH OF BIOLOGY

Unit- 2

Lead Identification- combinatorial chemistry & high throughput screening, in silico lead discovery techniques, Assay development for hit identification.

Protein structure
Levels of protein structure, Domains, motifs, and folds in protein structure.

Computational prediction of protein structure: Threading and homology modeling methods. Application of NMR and X-ray crystallography in protein structure prediction

DEPTH OF BIOLOGY

High Throughput Screening (HTS)

High Throughput Screening is a scientific technique used in drug discovery that allows researchers to rapidly test thousands to millions of compounds for potential biological activity against a specific drug target.

Principle-

Uses automation, robotics, sensitive detection methods, and data analysis.
Relies on miniaturized assays (very small volumes in multi-well plates).
Measures biological responses such as enzyme inhibition, receptor binding, or cell survival.

DEPTH OF BIOLOGY

Process of HTS-

Preparation of Compound Library
Large sets of chemical compounds are collected (from natural products, combinatorial chemistry, or existing databases).

Assay Development
A biological test system (enzyme, receptor, protein, or cell line) is chosen.
Assays are designed to measure activity (fluorescence, luminescence, absorbance, etc.).

Screening
Compounds are tested in automated multi-well plates (96, 384, or 1536 wells).
Robotics and liquid-handling systems speed up the process.

DEPTH OF BIOLOGY

Data Collection and Analysis
Computer software records and analyzes results.
Active compounds are identified as “hits.”

Hit Validation
Hits are re-tested and confirmed for reproducibility.

Advantages-
Very fast (can screen thousands of compounds per day).
Cost-effective compared to manual testing.
Allows parallel screening of many targets.
Provides a large pool of lead molecules for further studies.
Limitations-
False positives and negatives are common.
Requires expensive equipment.
Hits still need further lead optimization.

2. In Silico Lead Discovery Techniques-

Definition

In silico lead discovery refers to the use of computer-based methods to identify, design, and optimize potential drug molecules before laboratory testing.

Major Techniques-

1. Virtual Screening (VS)

Screening large chemical databases computationally instead of experimentally.
Saves time and money compared to HTS.

2. Molecular Docking-

Predicts how a small molecule (drug) binds to the active site of a target protein.
Estimates the strength and stability of binding.

DEPTH OF BIOLOGY

3. Quantitative Structure–Activity Relationship (QSAR)-
Uses mathematical models to relate chemical structure to biological activity.
Helps predict the activity of new molecules.

4. Pharmacophore Modeling-
Identifies the essential chemical features (like hydrogen bond donors, hydrophobic groups, aromatic rings) required for activity.
Used to design new molecules with similar features.

5. Molecular Dynamics (MD) Simulations-
Simulates the behavior of drug–receptor complexes over time.
Provides insights into stability, flexibility, and conformational changes.

DEPTH OF BIOLOGY

Advantages

Reduces time and cost of drug discovery.

Allows virtual testing of millions of compounds.

Can identify potential toxicity or side effects early.

Supports rational drug design (designing drugs based on target structure).

Limitations

Accuracy depends on quality of databases and algorithms.

Needs experimental validation (cannot fully replace lab testing).

Comparison – HTS vs. In Silico Techniques

Aspect	High Throughput Screening	In Silico Techniques
Approach	Experimental (lab-based)	Computational (computer-based)
Speed	Very fast but still limited by lab work	Extremely fast (millions of compounds virtually screened)
Cost	Expensive (automation, robotics, reagents)	Cheaper once infrastructure is set
Accuracy	Can give false positives/negatives	Depends on software & models
Role	Identifies *hits* experimentally	Predicts *hits* and optimizes them before experiments

Assay Development for Hit Identification

An assay is an experimental setup used to measure the biological activity of compounds (e.g., enzyme inhibition, receptor binding, cell response).

In hit identification, assays are designed to test thousands of compounds and identify the first active molecules (hits) against a drug target.

Stages of Assay Development-

1. Target Selection and Validation

Choose a biological target (enzyme, receptor, protein, or pathway).

Confirm that the target is relevant to the disease.

DEPTH OF BIOLOGY

2. Assay Design-

Decide what kind of biological response will be measured.

Common types:

Biochemical assays → enzyme activity, binding affinity.

Cell-based assays → cell survival, signaling, gene expression.

Reporter gene assays → use fluorescent/luminescent markers to detect activity.

3. Optimization-

Adjust experimental conditions:

pH, temperature, incubation time.

Concentration of reagents and substrates.

Detection method (fluorescence, absorbance, radioactivity, luminescence).

Ensure the assay is sensitive, reproducible, and stable.

DEPTH OF BIOLOGY

4. Miniaturization & Automation-

Assays are adapted to multi-well plates (96, 384, 1536 wells).

Use of robotics and liquid-handling systems for large-scale screening.

5. Validation of Assay-

Test with positive controls (known active compound) and negative controls (inactive compound).

Evaluate assay quality using parameters like:

Z'-factor → statistical measure of assay robustness (values between 0.5–1.0 = excellent).

Signal-to-noise ratio.

Reproducibility.

DEPTH OF BIOLOGY

6. Hit Identification (Screening Stage)-

Large compound libraries are screened using the developed assay.

Compounds showing desired biological effect are recorded as hits.

Hits undergo secondary assays for confirmation and elimination of false positives.

DEPTH OF BIOLOGY

Protein structure Levels of protein structure, Domains, motifs, and folds in protein structure.

Proteins are linear polymers of amino acids that fold into specific three-dimensional shapes.

The structure of a protein at different hierarchical levels determines its stability and function (catalysis, binding, signalling, structure).

Understanding protein structure is central to biochemistry, molecular biology and drug design.

Levels of Protein Structure-

Primary Structure

The primary structure refers to the linear sequence of amino acids in a polypeptide chain. This sequence is determined by the genetic code in DNA. Amino acids are joined by peptide bonds, and even a small change in this sequence can alter the function of the protein. Example: sickle cell anemia occurs due to a single amino acid change in hemoglobin.

Secondary Structure

The secondary structure arises from local folding of the polypeptide chain due to hydrogen bonding between the backbone atoms. The two main secondary structures are the α-helix, a coiled spring-like structure, and the β-pleated sheet, a zig-zag sheet-like arrangement. These structures provide stability and the basic framework of proteins.

DEPTH OF BIOLOGY

Tertiary Structure-

The tertiary structure refers to the overall three-dimensional folding of a single polypeptide chain. It is stabilized by interactions between side chains such as hydrogen bonds, ionic bonds, disulfide bonds, and hydrophobic interactions. The tertiary structure is responsible for the protein's specific shape and biological activity. For example, the globular structure of myoglobin is its tertiary structure.

Quaternary Structure-

Some proteins are made up of more than one polypeptide chain. The arrangement and interaction of these chains form the quaternary structure. Each chain is called a subunit. Hemoglobin is a classic example, having four subunits (two α and two β chains).

DEPTH OF BIOLOGY

Domains in Protein Structure-

A domain is a part of a protein that can fold independently into a stable 3D structure.

It often has a specific function (like binding, catalysis, or regulation).

Features of Domains-

A single protein can have one or many domains.
Domains are usually made up of 100–250 amino acids.
Each domain works semi-independently but together they give the protein its full function.
Domains are often connected by short linker regions.

DEPTH OF BIOLOGY

Functions of Domains-

Binding domains → help attach proteins to DNA, RNA, or other proteins.

Catalytic domains → carry out enzymatic reactions.

Regulatory domains → control protein activity (turning it on/off).

Signal domains → allow proteins to sense and respond to signals.

Examples

Kinase proteins → have a catalytic domain (transfers phosphate groups) and a regulatory domain.

Antibodies → each has variable domains (bind antigens) and constant domains (for immune response).

Transcription factors → may have a DNA-binding domain (like zinc finger) and an activation domain.

DEPTH OF BIOLOGY

Importance-

Domains help proteins evolve new functions because new proteins can be formed by combining old domains.

They make proteins modular (like building blocks), so a single protein can perform multiple roles.

Motifs in Protein Structure

Definition

A motif is a short, recurring pattern in proteins, usually made of combinations of α-helices and β-sheets.

Motifs are also called supersecondary structures.

They are not stable on their own, but they appear in many proteins and often indicate a particular function.

Features

Small (10–40 amino acids).

Made by arranging secondary structures (like helix–turn–helix, β-hairpin, etc.).

Often involved in specific functions such as DNA binding or dimerization.

DEPTH OF BIOLOGY

Examples of Motifs-

Helix-turn-helix motif → found in DNA-binding proteins like transcription factors.

Zinc finger motif → binds DNA or RNA using a zinc ion for stability.

Leucine zipper motif → helps two proteins dimerize and bind DNA.

β-hairpin motif → connects two antiparallel β-strands.

Folds in Protein Structure-

Definition-

A fold is the overall 3D arrangement of secondary structures (α-helices and β-sheets) and motifs within a protein domain.

It gives the characteristic shape or framework of a protein.
Unlike motifs (small patterns), folds are larger, stable, and involve the whole domain.

DEPTH OF BIOLOGY

-Features of Folds

Formed by combining multiple motifs and secondary structures.

Provide the scaffold that determines protein stability and function.

Many proteins with different functions can share the same fold.

Folds are used to classify proteins into families and superfamilies.

-Types of Folds

Folds are often grouped into three main classes:

All-α folds → mostly α-helices. Example: globin fold (hemoglobin, myoglobin).

All-β folds → mostly β-sheets. Example: immunoglobulin fold in antibodies.

α/β folds → mixture of α-helices and β-sheets. Example: TIM barrel fold.

DEPTH OF BIOLOGY

Examples of Important Folds-

TIM barrel fold (α/β barrel) → common in metabolic enzymes like aldolase and isomerase.

Immunoglobulin fold → β-sheet sandwich structure found in antibodies.

Rossmann fold → binds nucleotides (NAD^+/FAD), common in dehydrogenases.

Globin fold → bundle of α-helices, found in hemoglobin and myoglobin.

β-propeller fold → arranged like blades of a propeller, found in enzymes and receptors.

DEPTH OF BIOLOGY

Importance of Folds-

Provide stability to protein structure.

Define specific functions (binding, catalysis, recognition).

Help in evolution → many proteins evolved by reusing common folds.

Useful in drug design and understanding diseases caused by misfolding.

Computational Prediction of Protein Structure

Proteins perform their function based on their 3D structure, but experimental methods like X-ray crystallography, NMR, and cryo-EM are costly and time consuming.

Therefore, computational methods are used to predict protein structures from their amino acid sequences.

DEPTH OF BIOLOGY

Types of Computational Prediction Methods-

1. Homology Modeling (Comparative Modeling)
Based on the idea that proteins with similar sequences have similar structures.

Steps:
Identify a known protein structure (template) with sequence similarity.
Align the unknown sequence with the template.
Build a 3D model using the template as a guide.

Advantages: Accurate if high similarity (>40%) exists.
Limitation: Poor accuracy if similarity is low.
Example: SWISS-MODEL server.

2. Threading (Fold Recognition)-

Used when no close homolog is available.
The unknown sequence is "threaded" through a library of known folds.
The best fit between sequence and fold is chosen.

Advantage: Works even with low sequence similarity.

Limitation: Accuracy is lower than homology modeling.

3. Ab initio (De novo) Modeling-

Predicts structure from scratch using only the amino acid sequence.

Based on physical and chemical principles like energy minimization, molecular dynamics, and statistical potentials.

Advantage: Does not need a template.

Limitation: Very computationally expensive and less accurate for large proteins.
Example: Rosetta software.

DEPTH OF BIOLOGY

4. AI-based Methods (Modern Approach)-

Use machine learning and deep learning to predict protein structure.
AlphaFold (DeepMind) and RoseTTAFold are highly successful.
These methods predict atomic-level accuracy directly from sequences.
Revolutionized biology by solving structures of thousands of proteins.

#Applications of Computational Prediction-

Understanding protein function.
Drug discovery → predicting binding sites for drug molecules.
Protein engineering → designing enzymes or therapeutic proteins.
Studying disease-causing mutations.

DEPTH OF BIOLOGY

Threading and homology modeling methods

Homology modeling is a computational method to predict the 3D structure of a protein using the known structure of a similar (homologous) protein as a template.

Principle

Proteins with similar sequences usually fold into similar structures.

Therefore, if a template protein with known structure exists, the unknown protein's structure can be modeled.

DEPTH OF BIOLOGY

Steps-

Template Identification → Search protein databases (like PDB) for a structure similar to the query sequence.

Sequence Alignment → Align the query sequence with the template sequence.

Model Building → Build the 3D structure of the query based on template geometry.

Model Refinement → Adjust side chains and loops to minimize errors.

Model Validation → Check accuracy using tools like Ramachandran plot.

DEPTH OF BIOLOGY

Advantages

Simple and fast.

Accurate if sequence identity is >40%.

Limitations

Low accuracy if sequence similarity is poor.

Cannot predict completely new folds.

Example: SWISS-MODEL server is commonly used.

Threading (Fold Recognition)-

Definition-

Threading (or fold recognition) is a method used to predict the 3D structure of a protein by "fitting" its sequence onto a library of known structural folds, even if no sequence similarity exists.

Principle -

The number of possible protein folds is limited.

Even if two proteins do not share sequence similarity, they might share the same fold.

The query sequence is "threaded" (placed) on different folds, and the best fit is chosen based on energy and compatibility.

DEPTH OF BIOLOGY

Steps-

Query sequence input.

Compare against a database of folds.

Thread sequence onto each fold.

Score each fit using energy functions and compatibility.

Select best fold as predicted structure.

Advantages-

Works even when no homologous template exists.

Useful for proteins with novel sequences.

Limitations-

Less accurate than homology modeling.

Computationally expensive.

Sometimes mispredicts folds.

Comparison (Homology vs Threading)

Feature	Homology Modeling	Threading (Fold Recognition)
Template need	Needs a homologous template with high sequence similarity	Works without homologous template
Accuracy	High (if similarity >40%)	Moderate
Best for	Proteins with known homolog structures	Proteins with no sequence similarity but possible common folds
Limitation	Fails for novel proteins	Lower accuracy and more complex
Example tool	SWISS-MODEL	Phyre2, I-TASSER

DEPTH OF BIOLOGY

1. X-ray Crystallography-

Principle-

Protein is first purified and crystallized.

When X-rays are passed through the crystal, they get diffracted by the electrons of atoms.

The diffraction pattern is collected on a detector.

Mathematical analysis (Fourier transform) converts it into an electron density map.

A 3D atomic model of the protein is built from this map.

DEPTH OF BIOLOGY

Applications

Provides high-resolution 3D structures of proteins (up to atomic detail).

Helps in identifying the geometry of active sites in enzymes.

Used for drug discovery and design → drug molecules can be modeled into binding pockets.

Determines structures of large proteins and protein complexes.

Helps to study the effect of mutations on protein folding and stability.

Allows classification of proteins into motifs, domains, and folds.

Helps in structural genomics – large-scale structure determination of unknown proteins.

DEPTH OF BIOLOGY

2. Nuclear Magnetic Resonance (NMR) Spectroscopy-

Principle

Proteins are dissolved in solution and placed in a strong magnetic field.

Nuclei like ^{1}H, ^{13}C, ^{15}N absorb and re-emit radiofrequency signals.

These signals provide information about distances and angles between atoms.

A 3D structure is built based on these distance constraints.

DEPTH OF BIOLOGY

Applications-

Determines 3D structure of proteins in solution (near physiological conditions).

Useful for small to medium proteins (<40 kDa).

Provides information on protein flexibility, folding, and conformational changes.

Can study dynamic interactions of proteins with ligands, nucleic acids, or other proteins.

Important in drug design → helps map binding sites of drugs on proteins.

Allows study of protein folding pathways and misfolding diseases.

Can provide data on intrinsically disordered proteins that cannot be crystallized.

Limitations -

Not suitable for very large proteins (spectra become too complex).

Requires a large amount of pure, isotopically labeled protein sample.

Resolution is lower compared to X-ray crystallography.

DEPTH OF BIOLOGY

Comparison (Detailed Points)

Feature	X-ray Crystallography	NMR Spectroscopy
Sample required	Crystalline protein	Protein in solution
Best for	Large proteins and complexes	Small to medium proteins (<40 kDa)
Resolution	Very high (atomic level)	Moderate
Structure type	Static (frozen in crystal)	Dynamic (movement, flexibility)
Applications	Active site study, drug design, mutation effects, large complexes	Dynamics, folding, ligand binding, protein interactions
Limitation	Crystallization is difficult; no dynamics	Size limitation; lower resolution