UCS2735 BIOINFORMATICS TECHNOLOGIES

Unit I

* **Bioinformatics** - a field of science in which biology, computer science, and information technology merge into a single discipline to analyze biological information using computers and statistical techniques.

Major Aspects : (i) well organized DBs

(ii) computationally - derived hypothesis

(iii) Web servers (online tools / applications)

(iv) Big data analysis

(v) Virtual screening of compounds for drug development

* **Central Paradigm of Bioinformatics**

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (symptoms)

* **Challenges understanding Genetic Information**

(i) Genetic info is redundant

(ii) Structural info is redundant

(iii) Genes and proteins are meta-stable

(iv) Genes and proteins are 1D but their function depends upon 3D structure

* Central Dogma of Molecular Biology

Genotype → Phenotype

(Aa) (pink)

replication →

Gene → ATCGATCGA

(DNA)

transcription ↓ reverse transcription

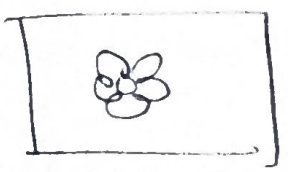
Messenger RNA → UAGUUCAGG...

↓ translation

Protein

↓

Trait



DNA (Carrier of Genetic Info) — Genomics — Structural, Function, Epigenomics, Metagenomics, Pangeomics

↓

mRNA - Genetic Information carrier — Transcriptomics

↓

Proteins - Genetic information expression — Proteomics

↓

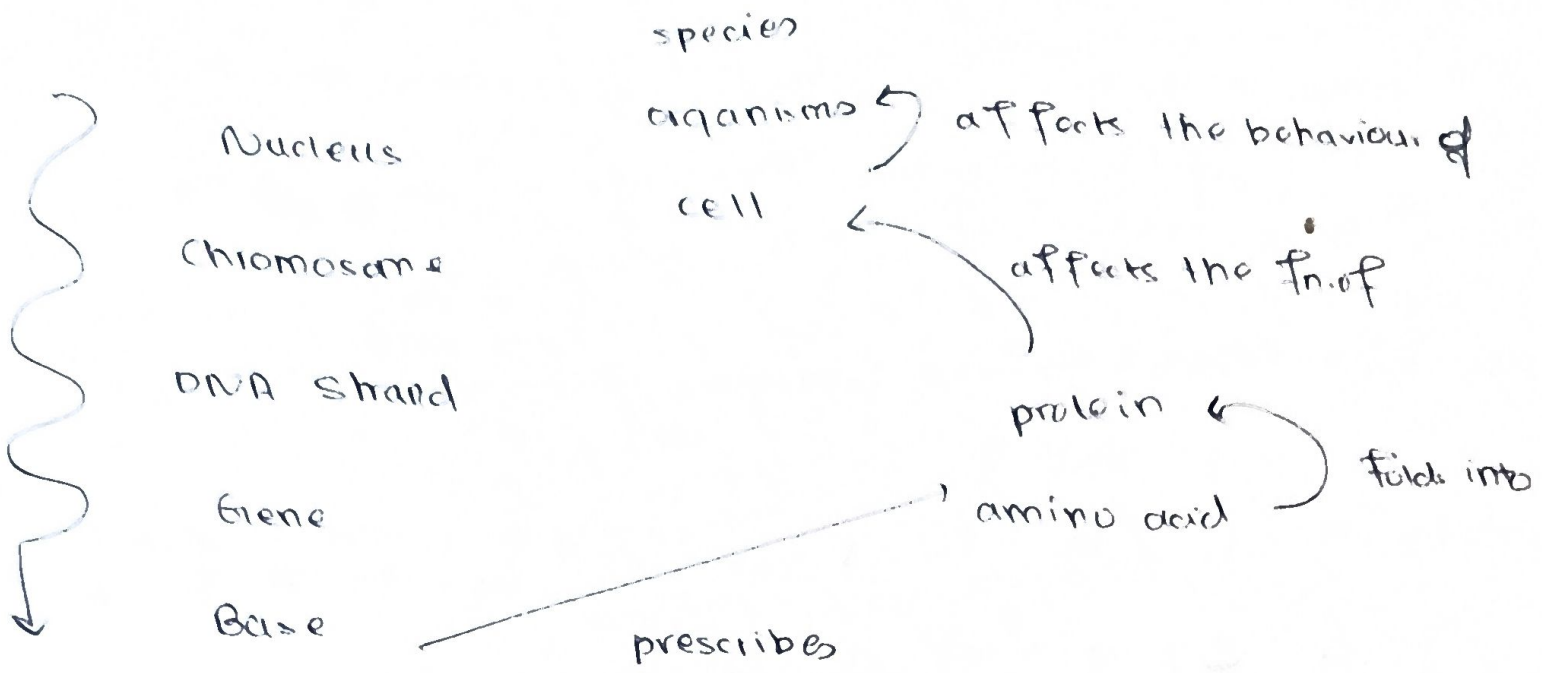
Metabolomics

Ionomics

↓

Traits — Phenomics

* Substructure and Effect



* Synthesis of Proteins

(i) DNA double helix is unzipped

(ii) One strand is transcribed to messager RNA

(iii) RNA acts as a template

(iv) Amino acid sequences fold into a 3d molecule

Transcription → write base counterparts

A ↔ T

G ↔ C

change T → U

Thymine to Uracil

For eg. G G A T G C C A A T G

C C T A C G G T T A G

C C U A C G G U U A G

Translation — Input triples of 3 base letters (codon)

— Output = amino acid

— eg. ACC becomes Threonine

A ↔ T
G ↔ C

eg Transcribe and Translate the following.

T C G G T G A A T C T G T T T G A T

A G C C A G T T A G A C A A A C T A

A G C — Ser → S

C A G — His → H

U U A — Leu → L

G A C — Asp → D

A A A — Lys → K

C U A — Leu → L

* Evolution of Genes — mutation
- radiation from toxicity
- Deletion, Insertion, Substitution

* Applications of Bioinformatics

(1) Pharmaceuticals

1. Biobank - biorepository that stores human biological samples
- used in genomics, personalized medicine
- helps accelerate the discovery and development of drugs

2. Veterinary Science / Animal Health - covers diagnostics, medicines and vaccines for farmed animals and pets

- due to a growing demand for animal proteins as well as a strong consumer need for companion animal health care

② Medical Implications

1. Pharmacogenics - not all drugs work on all patients, can cause death

- do gene analysis before treatment

- customized treatment

2. Gene Therapy - replace or supply the defective or missing gene

- eg. Insulin and Factor VIII or hemophilia

③ Diagnosis of Disease

- identification of genes which cause the disease will help detect disease at an early stage - eg. Huntington disease

↓

weird movements

personality changes & intellectual impairment

- excessively repeated sections of CAG

(4) Drug Design

→ expensive and time consuming

→ computational methods can improve testing methods

(5) Drug Discovery

→ Identify the molecule on which germs rely for its survival

→ develop a drug that will bind to the target

→ So the germ will not be able to interact with the target

→ Proteins are the most common targets

eg. HIV produces HIV protease which is a protein and which in turn eats other proteins

HIV protease has an active site where it binds to other molecules

(6) Phylogenetic Trees

→ understand our genes

→ genes are homologous

(7) Predicting Protein Structure

→ proteins fold to set up an active site

→ has a small but highly effective substructure

→ activity sites determine the activity of the protein

→ no one has found rules governing how proteins fold

* Need for Bioinformatics Technologies

interseaction of fields

computer and statistical fields

handle voluminous biological data.

identify interesting patterns

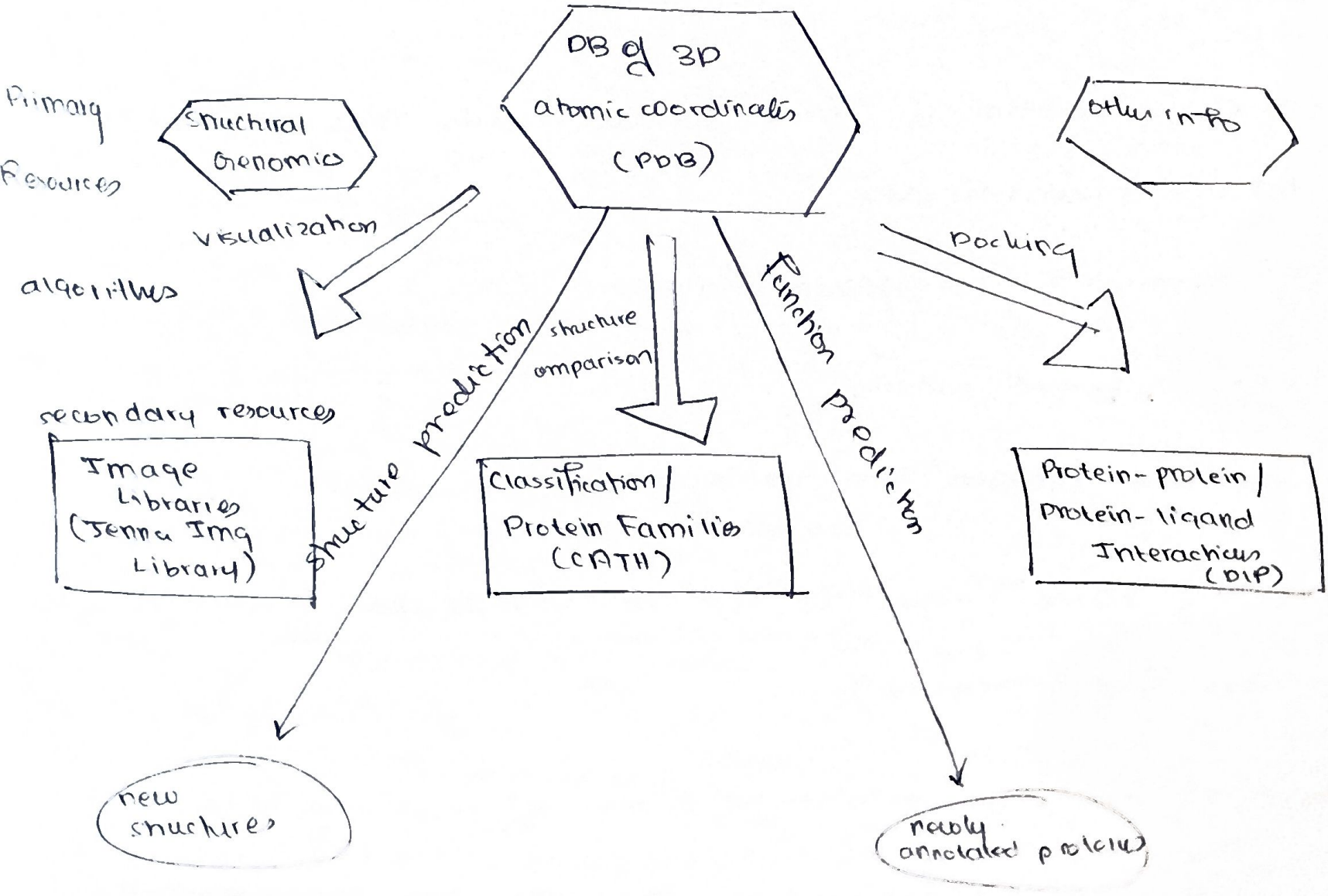
* Bioinformatics Technologies

KD
Sequence Analysis

Structure Analysis

Expression Analysis → grouping, profiling of genes

* Structural Bioinformatics



Structural bioinformatics represents the subset that deals directly or indirectly with the structure of macromolecules.

Primary Sources - contains atomic coordinates unimportant / uninformative to the majority of structural biologist

eg. PDB

- Protein Data Bank - stores 3D biological macromolecular structures
  - contains publicly available 3D structures of proteins, nucleic acids and a variety of other complex biomolecules determined by X-ray crystallography, NMR spectroscopy

Data Format

→ Consists of a collection of fixed format records that describe the
  (i) atomic coordinates
  (ii) chemical & biochemical features
  (iii) experimental details
  (iv) structural features such as H bonds

→ dictionary based format is the macromolecular crystallographic information file (mmCIF)

→ underlying data organization is a set of relational tables

→ mmCIF dictionary is an ontology that describes macromolecular structure & the various experiments used to derive it.

Data Processing and Quality Control

Data deposition

Data validation – access quality of deposited atomic models (structure validation) and assess how well they fit experimental results

Data annotation – process of adding info. to the entry

PDB uses accepted community standards to validate structures

Visualization - molecular graphics program

MICE

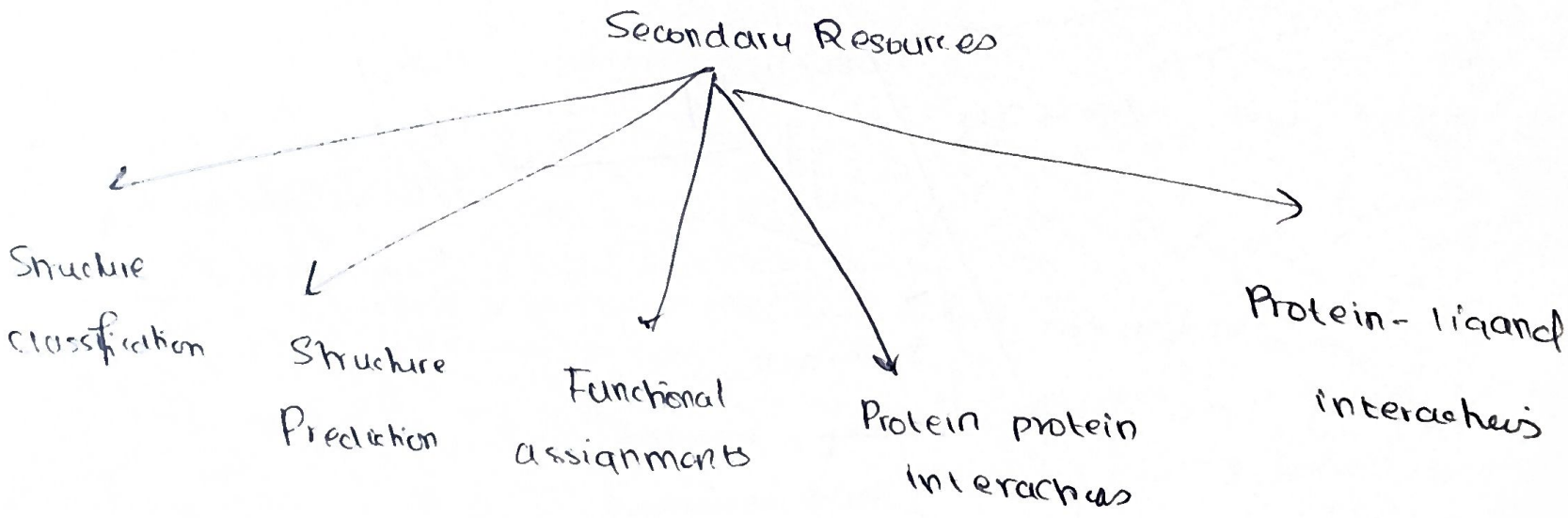
Mol Script

PyMOL

Secondary Structures

Resources

→ They are value-added structural databases - from data reducing using algorithms or human expertise



A. Structure Classification

→ Process of grouping proteins together by the level of 3D & sequence similarity

→ use structural comparisons

→ Divided into 3 major steps:

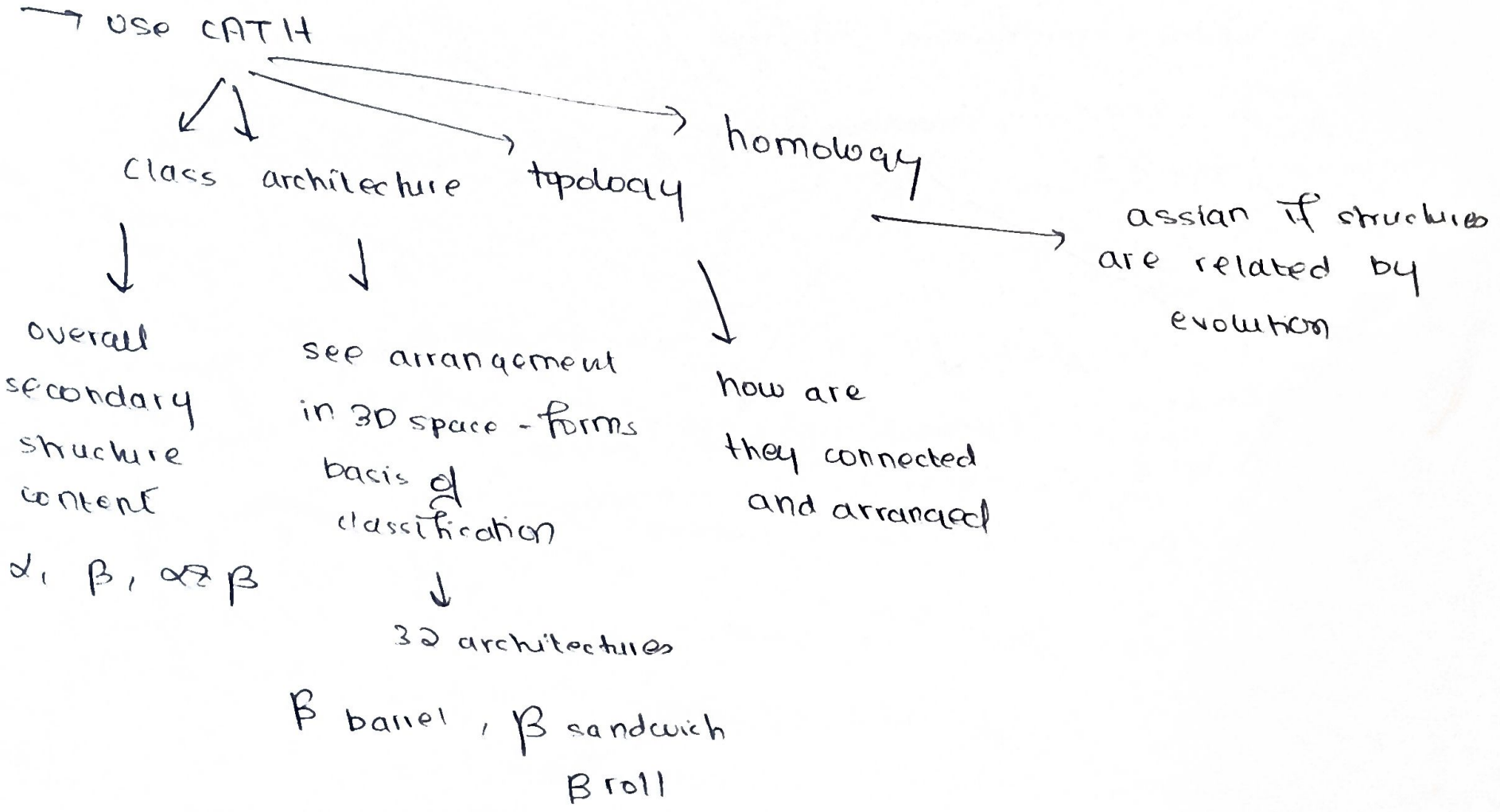
(i) representation of 3 structures in a coordinate-independent space

(ii) comparison & optimization

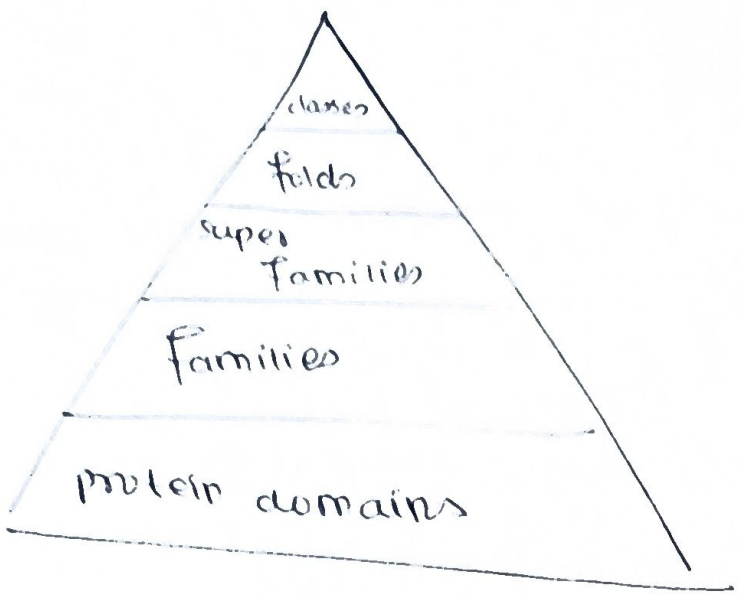
(iii) measure statistical significance of the alignment

→ Handle tradeoff between geometric alignment & its biological sequence.

→ use CATH



SCOP – Structural classification of proteins



B. Structural Predictions

Methods are

1. Homology modeling
2. Fold recognition
3. Ab initio

Experiments used are:

Critical Assessment of Structure Prediction (CASP)

"Fully Automated Structure Prediction (CAFASP)

C. Functional assignment

→ what exactly a protein does, when & where in the cell is it active, what are its interacting patterns.

→ assigns fn to the protein simply from its seq. or structure

→ proteins w/ similar sequences have similar functions

>40% → same structure, similar fn

<30 inconclusive 20-30 = twilight

<20 = midnight

D. Protein Protein Interactions

→ network of interacting components

→ infer using phylogenetic profiles gene fusion, and gene neighbourhood

→ if 2 proteins are present or both missing => likely to be involved in functional interactions

→ Interacting genes are placed closely to each other — correlated using STRING

E. Protein-Ligand Interactions

SMILES, MARVIN

→ use ligand-based (analog)
target-based (target) for drug discovery

→ use pharmacophore and quantitative structure activity relationships (QSAR)

2 methods docking - match from existing DB
building - generate new ligands by connecting atoms of molecular fragments

Drug Design

XRAY, NMR

↓

Structure of Target → Virtual Screening, Docking

↓

Structure of Lead with Target ← Lead Compound Generation

↕

Biochemical Assays to eval hits

↘

comp design of improved lead compound

Synthesis of New Compounds

Combinatorial libs

Computational de novo design of New Lead Compounds