# Bioinformatics

- **Course Details**
- 11 credit course
- 2 lectures and 1 practical

- **Introduction**
- Use data from biology and tools from informatics to get results
- Covid 19 - spike protein and ece2 interaction
- Sequence alignment, DNA and protein seq analysis, phylogeny, secondary and tertiary structure prediction, folding and stability, machine learning, structure based drug design, development of algorithms.
- Protein bioinformatics - Michael gromiha
- Practicals, Assignments(10-12, best of 8 or 9), 40% weightage to assignments. No quizzes. Endsem 60%

- The term “Bioinformatics” was coined by Paulien Hogeweg in 1979
- Life sciences provides data : DNA, protein, metabolic

- **Major Aspects**
- Databases
- Computational hypothesis
- Web servers, online tools and applications
- Virtual screening of  compounds for drug development
- Big data analysis

- For screening of compounds to make new molecules, we need data on target and ligands.

- **Complexity of biological systems**
- DNA strand
- Protein synthesis
- There are 6 reading frames, 3 in each direction
- GC and AT content
- Base stacking energy
- Flexibility energy
- Seq 2 Feature (sir’s website)

- **rases**
- They are organized digital collection of information
- Efficient storage and retrieval of data

- Data should be trustable

Characteristics:
- Decide the contents
- Ontology : list of definitions of terms used
- Schema - logical structures
- Format of data
- Routes for retrieval
- Tutorials

Nucleotide sequence databases
- DDBJ (DNA Data Bank of Japan)
- EMBL (European Molecular Biology Laboratory)
- Genbank (USA)

Literature databases
- PUBMED
- Scopus
- Google Scholar

H index:
H papers, cited h times

**Protein structure and function**
- Globular proteins
- Fibrous proteins
- Membrane proteins
- Membrane proteins can be in cytoplasm, inner membrane, periplasm and outer membrane and outer space
- Protein functions:
- Enzymes
- EC (enzyme commission) number:

**Top-level EC numbers**[4]

| Group | Reaction catalyzed | Typical reaction | Enzyme example(s) with trivial name |
|---|---|---|---|
| EC 1 *Oxidoreductases* | To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another | AH + B → A + BH (**reduced**)<br>A + O → AO (**oxidized**) | Dehydrogenase, oxidase |
| EC 2 *Transferases* | Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group | AB + C → A + BC | Transaminase, kinase |
| EC 3 *Hydrolases* | Formation of two products from a substrate by hydrolysis | AB + $H_2O$ → AOH + BH | Lipase, amylase, peptidase |
| EC 4 *Lyases* | Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved | RCOCOOH → RCOH + $CO_2$<br>or [X-A-B-Y] → [A=B + X-Y] | Decarboxylase |
| EC 5 *Isomerases* | Intramolecule rearrangement, i.e. isomerization changes within a single molecule | AB → BA | Isomerase, mutase |
| EC 6 *Ligases* | Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP | X + Y+ ATP → XY + ADP + Pi | Synthetase |

Brenda - Enzyme information system (activity, structure function)
Catalytic site atlas

Uniprot - more numerical values (not for enzymes)

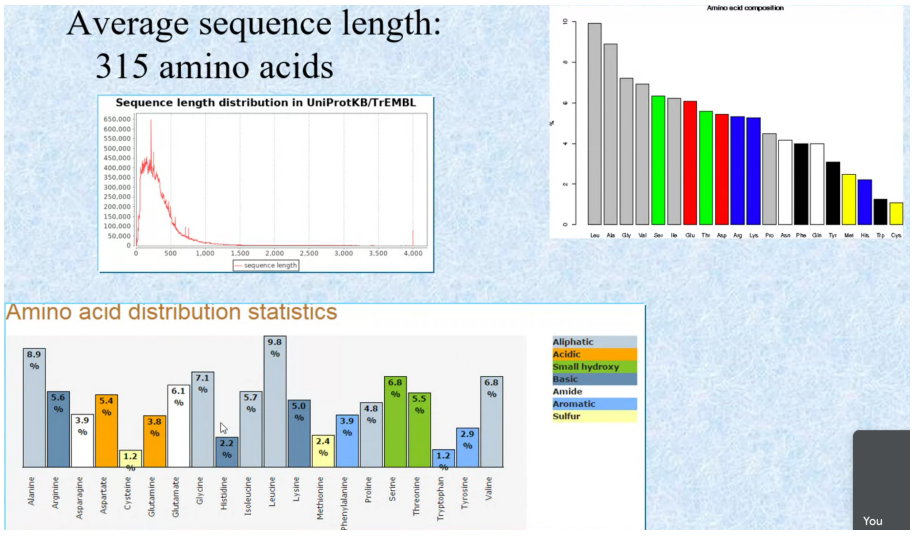
Transport proteins
Membrane proteins : alpha helical, beta barrel

Primary sequence :
Motifs
Insulin was first sequence

**Databases for protein sequences:**
UniProt
Swiss prot
TrEMBL
PIR - Protein information resource (Georgetown university)
iPro class - information reports on protein sequences

**3 Major aspects of uniport**
Minimal level of redundancy
High level of integration
High level of annotation

There are about 225.5 Million sequences in Uniport




**Contents of Uniprot**

- Name and origin
- Function

- Subcellular localization
- Pathology
- Post translational modifications
- Interactions
- Structure
- Primary sequence

**Pairwise alignment**

**Sequence comparison**

Dot plot : Generation exploration of your sequence
- Finding repeats
- Long insertions and deletions

Local Alignment : Comparing sequences with partial homology
- High quality alignments

Global alignment :
- Identify long insertions and deletions
- Identify mutations
- Check quality of your data

Dot plot:
- Plot dot if sequences are identical

**Simple Alignment**
- Mutations
- Insertions
- Deletion

- Gaps are added in cases of insertions and deletions
- Mutations are much more common than insertions or deletions

- Gap penalty is introduced as insertions and deletions are not common

- Continuous gaps are more common than scattered gaps.

1. Origination penalty : how many times gaps have to be made?
2. Length penalty : number of gaps

- Scores can be refined based on substitutions
- Transitions and transversions
- Scores can be assigned based on amino acids too.
- Amino acid size or chemical and physical properties
- Based on codons

- A common method for deriving scores is to observe the actual substitution rates occurring in nature
- One commonly scoring matrix on substitution rates is the point accepted mutation (PAM)
- PAM 1 is used to compare closely related sequences
- PAM 1000 can be used for distant relationships
- PAM 250 is the most commonly used.

**Blosum Matrix**

- Blocks Substitution Matrix, is obtained with statistical clustering techniques
- BLOSUM considers mainly conserved regions
- BLOSUM-X matrix is apt for comparing X% sequence identity
- Similar trends in both matrices

**BLAST**

- Basic local alignment search tool
- Divide the query into substrings of length k
- Use dynamic programming

- DP can only be used for local alignment.

- For optimal local alignment, add another condition (del, ins, subs and 0)
- This will make sure that the matrix has no negative values

Smith-Waterman is local
Needleman and Wunsch is global

**Database searches**

- Problems :
- Size of query sequence
- Number of sequence in database

**BLAST working**

Blast will find subsequences from a sequence data base for any query sequence
Blastp
Blastn

Tblastn (translated)
Tblastx
It may use PAM or Blossom matrices
It breaks down the query into subsequences
It used high scoring pairs, also uses gaps

# FASTA

**FASTA is another program for sequence similarity search and sequence alignment.**

**FASTA breaks the words into 4-6 nucleotides or 1-2 amino acids**

**Eg. Query sequence: FAMLGFIKYLPGCM**

| Word | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | 2 | 13 | | | 1 | 5 | | 7 | 8 | 4 | 3 | | 11 | | | | | | | 9 |
| | | | | | 6 | 12 | | | | 10 | 14 | | | | | | | | | |

**Target sequence: TGFIKYLPGACT**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T | G | F | I | K | Y | L | P | G | A | C | T |
| | -3 | -2 | 3 | 3 | 3 | -3 | 3 | -4 | -8 | 2 | |
| | 10 | 3 | | | | 3 | | 3 | | | |

**Large number of sequences have the number 3;**

**Offsetting with 3 gives a reasonable alignment**

```
FAMLGFIKYLPGCM
|||||||
TGFIKYLPGACT
```

# Alignment score and statistical significance

**The primary indicator of how similar the search results are to a query sequence is the alignment score (S).**

**Score is given with P or E value.**

**E-value is the expected number of sequences of score ≥ S that would be found by random choice**

**P-value is the probability that one or more sequences of score ≥ S would have been found randomly.**

**Low values of E and P indicate that the search result was unlikely to have been obtained by random chance, and thus is likely to bear an evolutionary relationship to the query sequence.**

**E values of less than $10^{-3}$ are often considered indicative of statistically significant results and search algorithms produce matches with E values on the order of $10^{-50}$.**

**p-value gives the likelihood of getting equally high scores by chance**
**E-value gives the expected number of sequences with high scores to be found by chance**

**Features of BLAST**

- Identifying sequence similar to the query
- Finding members of a protein family
- Finding proteins similar to a pattern
- Conserved domains
- Search for peptide motifs

**Blast options**

- Gi
- Accession number
- File formats, FASTA, NBRF(specify if protein of nucleic acid), GDE (same as fasta, starts with %)

**Multiple sequence alignment**

CLUSTAL : uses the fact that similar sequences are evolutionary related
Similar sequences align first, followed by distant sequences

MAFFT
MUSCLE
PROMOLS

**Conservation score**

- Some sequences are evolutionary conserved
- Conserved regions and variable regions

Steps:

- Positions specific amino acid frequency (unweighted, weighted, independent count)
- Calculation of score (entropy based, variance based score, sum of pairs measure)

AL2CO server, for multiple sequence alignment. Uses clustal format
ConSurf Server - pdb id, chain

**Protein sequence analysis**

- Size
- Amino acid composition
- PIR tools.

**Amino acid properties**

- Average hydrophobicity

- Identify different groups of proteins using deviation or correlation

- Hydrophobicity profile (plot of hydrophobicities)
- Using the plot, secondary structures can be identified
- Nozaki tanford jones scale and ponnuswamy gromiha
- Amphipathicity - periodicities in polar and non polar character of amino acid sequence (IMP)
- Periodicity is 4 because 3.6 residues per turn in alpha helix and 2 in beta strand

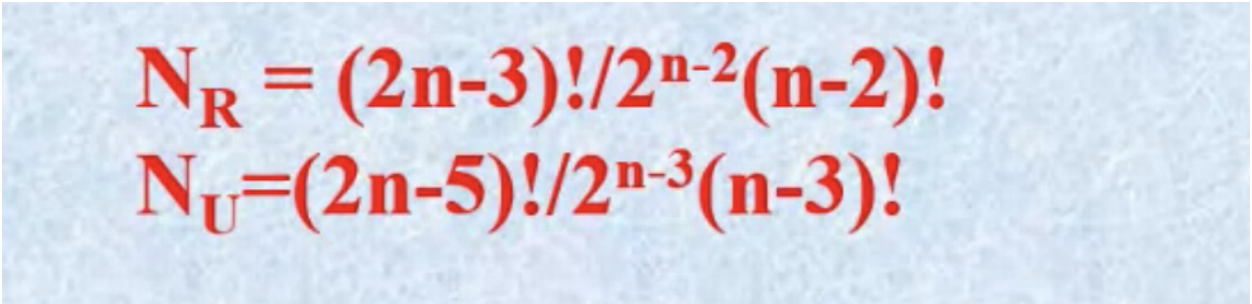
- Some patterns can be defined (like regex)
- PIR Pattern search or peptide search
- Patterns help recognize motifs
- Position specific scoring matrices (PSSM)
- Applications : secondary structure prediction, classification of proteins, identify binding sites

**Large scale analysis**

- Low identity sequences are redundant
- Programs to reduce redundancy : CD-HIT (cluster database at high identity with tolerance) Blastclust, PISCES
-
- Handle huge sequences, easy to download, quick results
- Greedy incremental algorithm
- Longest sequence first and check identity with shorter sequences
- Search for decapeptides, pentapeptides and so on..
- K means clustering (hamming distance -sum of differences in composition or euclidean distance)

**Phylogeny**

- Biological relationships expressed as a tree
- Assume homology
- Organisms with high degree of sequence similarity are closely related
- Newick format
- Rooted trees and unrooted trees

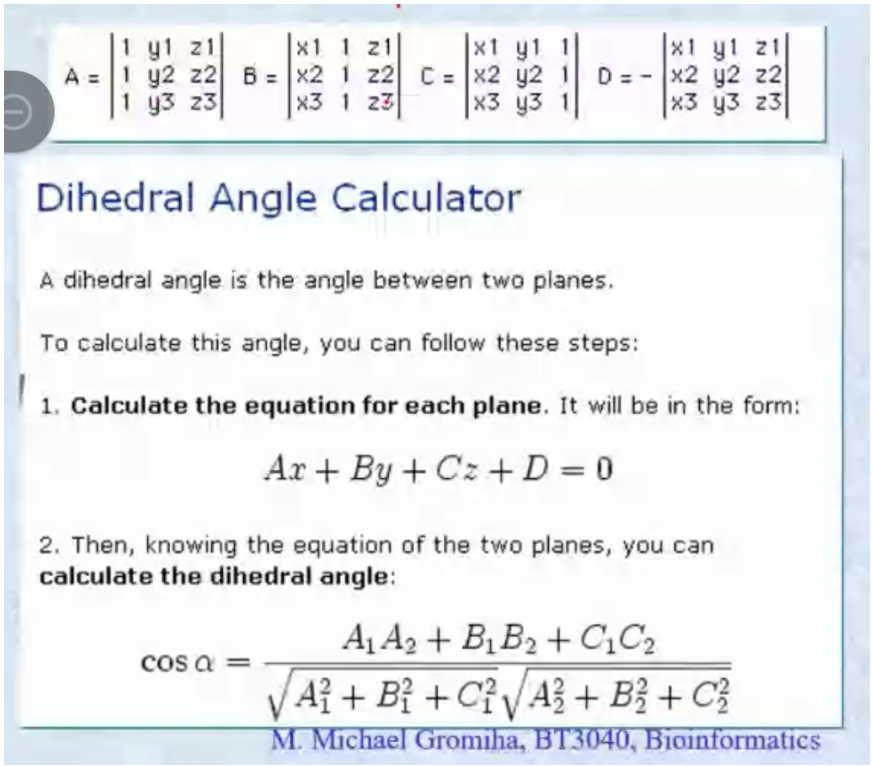



- Tree construction : UPGMA tree (unweighted pair group method with arithmetic mean)
- Construct a mismatch matrix

- Problems : distances are not additive
- Neighbor joining method
- Maximum likelihood method
- Phylip : Program to create trees, takes MSA as input
- Bootstrapping, accuracy measures

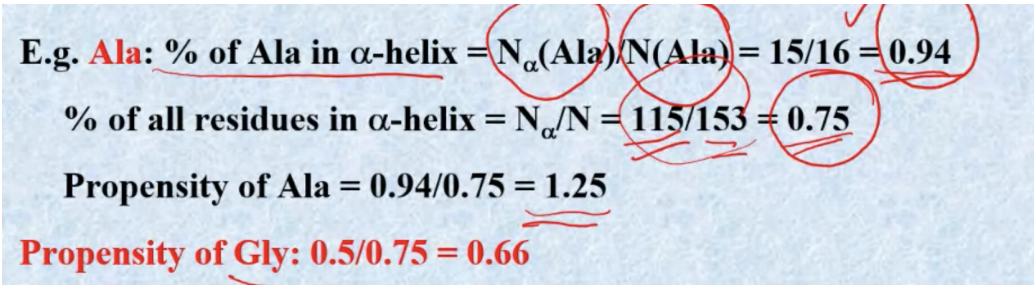
**Secondary structure**

- Alpha helices and beta strands
- Torsional angles calculator




DSSP : Dictionary of secondary structure of proteins (uses H-bonding pattern)
Secondary structure prediction : Hydrophobicity profile, propensity, information theory, MSA, Machine learning, consensus

- **Statistical analysis**
- Propensities




- Greater than 1 propensity will mean it is preferable else not preferable
- Alanine, glutamic acid has high propensities for helices.

**Secondary structure prediction methods**

- GOR : Based on information theory. Central residues, and then 8 residues on each side, window of 17, take average.

$$I(SS_i=X:\sim X;aa) = \ln( P(SS_i=X|aa) / P(SS_i=\sim X|aa) ) - \ln( P(S_i=X) / P(S_i=\sim X) ),$$

$SS_i \rightarrow$ secondary structure at position i in the sequence
$X \rightarrow$ any secondary state helices (H), sheets (E), turns (T) and coil (C)
$aa \rightarrow$ any amino acid residue

Use matrix and calculate j, j+1, j-1 etc.

**MSA for secondary structure prediction**

- Assign secondary structure by aligning it with known sequences.
- Confidence of prediction is related to conservation score

**Machine learning**

- Neural networks

**Consensus prediction**

- Take only consensus
- Meta prediction (ensemble + ML)

**Tertiary structure prediction**

- Find coordinates for each atom
- Experimental Methods : x ray crystallography, NMR, Cryo electron microscopy.
- PDB : First structure hemoglobin, followed by vitamin B12
- PDB Code : number, 3 letter, chain information
- Occupancy : conformations of side chains
- Temperature factor : How rigid the structure is due to vibrations (normalized as B factor)

**Secondary databases**

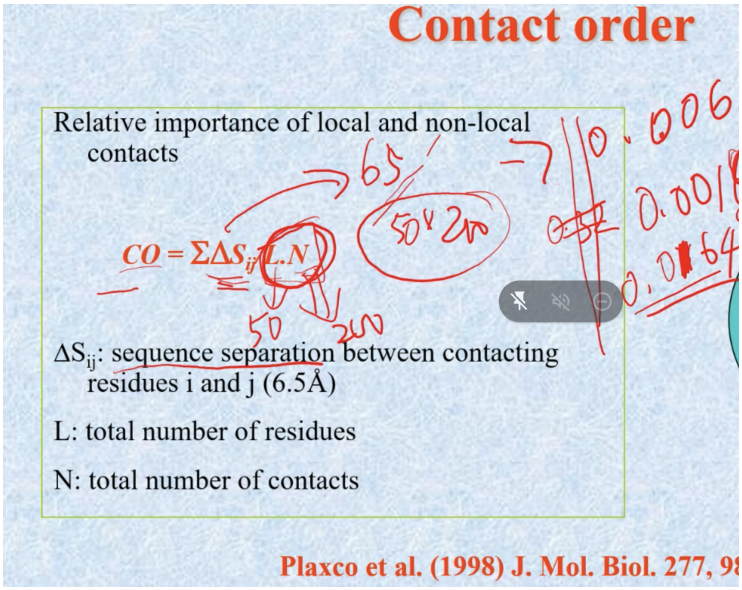
- Globular, fibrous and membrane proteins
- Membrane proteins are mostly helical
- Alpha helical membrane proteins : cytoplasm periplasm (hydrophilic residues)
- Beta barrel : periplasm and outside space (both hydrophobic and polar)
- PDBTM : protein database for transmembrane proteins
- SCOP structural classification of proteins : hierarchical levels, superfamily, fold, structural class
- CATH : Class, architecture, topology, homologous superfamily.
- Structure based parameters
    - Contact maps: distance between all possible pairs of residues, need distance threshold, which atoms?
    - 2d - representation of 3d

Types of contacts:

- Short range : < 3 residues away (diagonal)
- Medium range : 3 or 4 residues away (close to diagonal)
- Long range : > 4

**Solvent accessibility**

- Roll the water molecule on the surface
- Methods : ACCESS, NACCESS, ASC, DSSP
- ASA view : pictorial representation
- Percentage accessibility = asa folded / asa unfolded
- Ratio less than 5% then buried



## Long-range order

Obtained from the knowledge of long-range contacts (contacts between two residues that are close in space and far in sequence)

$LRO = \Sigma n_{ij}/N$; $n_{ij}=1$ *if* $|i-j| > 12$;

$= 0$ *otherwise.*

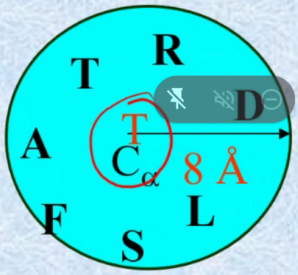
i and j: two residues in which $C_\alpha$ distance between them is ≤ 8Å

N:the total number of residues in a protein.


## Surrounding hydrophobicity

It characterizes the hydrophobic behavior of the 20 amino acid residues in protein environment.

$H_j = \Sigma\, n_{i,j}\, h_i$




A, C, G, M, Y: 1
F, I, L, V, W: 2
D, E, H, K, R: -2
N, P, Q, S, T: -1

$H_j$: Surrounding hydrophobicity of the central residue j
$n_{ij}$: Number of residues of type i around j
$h_i$: Experimental hydrophobicity of residue i

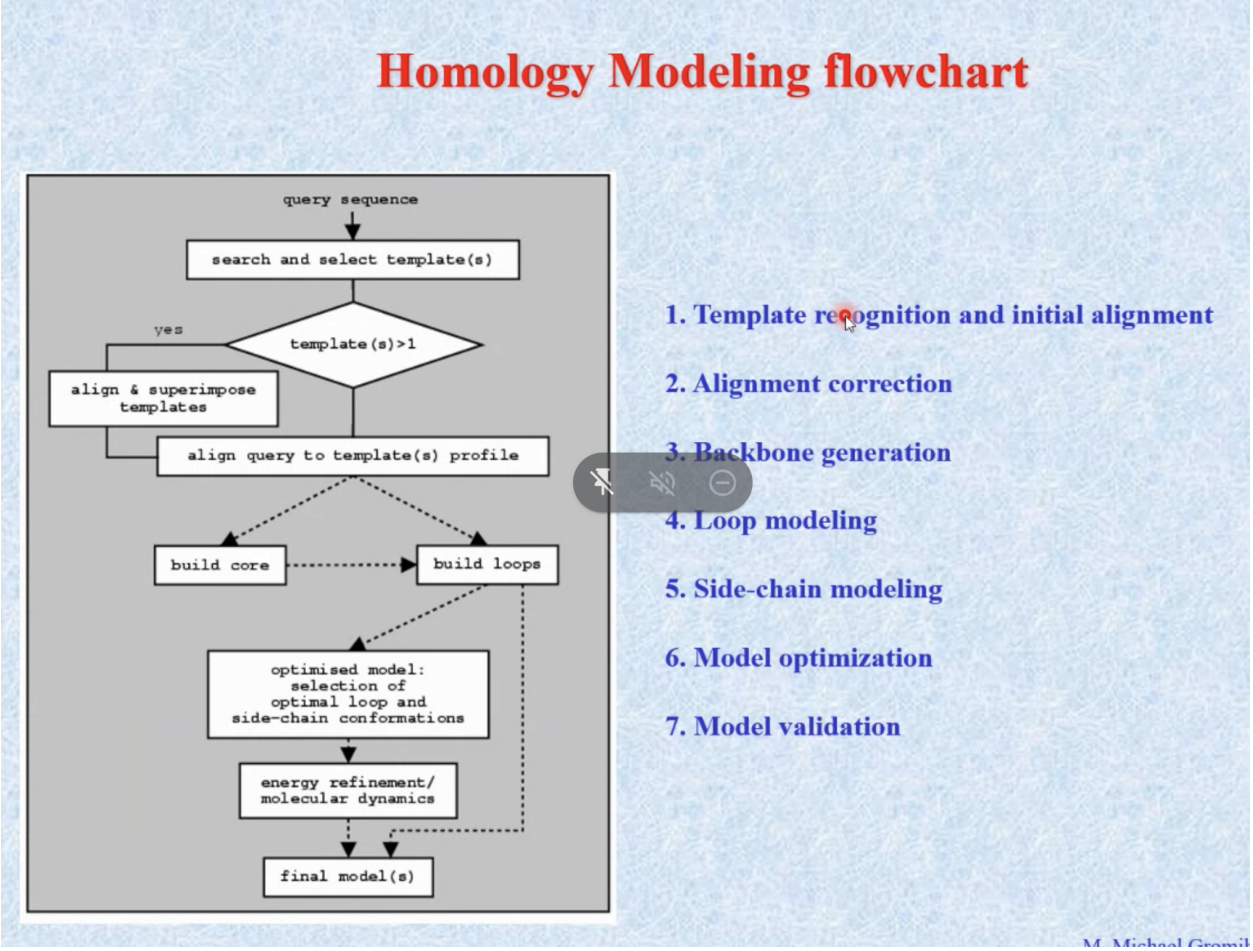
Surrounding hydrophobicity:

**Tertiary structure prediction:**

- Pdb structures : 190,000
- Homology modeling
- If sequence identity is not high:
- Use fold recognition
- Ab initio (energetic approach)

Homology modeling:

- Structure based drug design
- Interactions
- Antigenic behavior




**Aspects to consider while constructing and algorithm**

- Classification problems : secondary structures, stable or unstable, dna binding or not

Regression problems:

Sensitivity, specificity, accuracy, AUC