HI!

I’m Fathima Ridha
Research scholar at IITM
and “TA” for this course

Email: fathima.ridha.97@gmail.com
POLL

Have you attended Week 1 Session?
• Yes
• No, this is my first class
WEEK 2

Lec 1
• Protein structure and function

Lec 2
• Protein sequence databases

Lec 3
• Pairwise alignment
Lec 1 Recap

• What are the components of an amino acid?
  An “R” group, an amine, a carboxylic acid, and a hydrogen atom
• What type of bonds are used to link amino acids into a chain
  • Peptide bonds
• What are the 2 main classifications of amino acids?
The amino acids are classified into hydrophobic (hydrogen, aliphatic, aromatic, and sulfur containing) and hydrophilic (negatively charged, positively charged, and polar).
How do you memorize an amino acid one letter code?


Characteristics and Properties of Amino Acids

- **Polarity Ranking**: The ranking of polarity will depend on the relative ranking of polarity for various functional groups

  Amide > Acid > Alcohol > Amine > Ether > Alkane

- The number of alkyl groups also influences the polarity
Classification of proteins based on function

Match ??

1. Enzymes
   - a. Ovalbumin
2. Transport proteins
   - b. Thrombin
3. Storage proteins
   - c. Keratin
4. Defense protein
   - d. Sweetener
5. Contractile protein
   - e. Glucokinase
6. Structural protein
   - f. Myosin
7. Regulatory protein
   - g. Hemoglobin
8. Monellin
   - h. Insulin
Structural organization of proteins

Structural classification of proteins

- Fibrous proteins
- Globular proteins
- Membrane proteins
Lec 2 Recap

Databases for protein sequences:
• Protein Information Resource (PIR)
• SWISS-PROT and TrEMBL
• UniProt: The Universal Protein Resource
• Other protein sequence databases - EXProt, NCBI Entrez Protein database
Lec 3 Recap

- Protein Sequence Analysis - Dot Plot

Dot plots are useful as a first-level filter for determining an alignment between two sequences.
Alignments & Gaps

In the simplest case, in which gaps are not allowed, the alignment of two sequences is simply a matter of choosing the starting point for the shorter sequence.

Mismatch score = 0, match score = 1
The score for three alignments are 4, 1 and 3 respectively.

The two sequences that can be aligned without gaps in only three ways, can be aligned in 28 different ways when two gaps are allowed within the shorter sequence.

Mismatch score = 0, match score = 1, gap penalty = -1
The score for three alignments are 1, 3, 3 respectively.

Gap origination penalty = -2, length penalty = -1, Match score = 1, mismatch score = 0
The score for three alignments are -3, -1 and +1 respectively.
Scoring Matrices

• Several criteria can be considered in setting up a scoring matrix for amino acid sequence alignments
  • Similarity based matrices - physico-chemical (hydrophobicity, charge, electronegativity, and size), genetic code
  • Actual substitution rates - PAM matrix
PAM matrix

- The PAM matrices (also called Dayhoff PAM matrices) were first constructed by Margaret Dayhoff
- PAM stands for “point accepted mutation”
- A PAM unit is defined as 1% amino acid change or one mutation per 100 residues.

| TABLE 3.1. Correspondence of PAM Numbers with Observed Amino Acid Mutational Rates |
|---------------------------------|---------------------------------|---------------------------------|
| PAM Number | Observed Mutation Rate (%) | Sequence Identity (%) |
| 0         | 0                          | 100                          |
| 1         | 1                          | 99                           |
| 30        | 25                         | 75                           |
| 80        | 50                         | 50                           |
| 110       | 40                         | 40                           |
| 200       | 75                         | 25                           |
| 250       | 80                         | 20                           |

- To obtain matrices of different PAM numbers, for example, a PAM2 matrix, simply multiply a PAM1 by itself. Similarly, to obtain a PAM256 matrix, simply multiply PAM1 against itself 256 times.
Assignment Quiz
What is the one element that proteins have but carbohydrates and lipids do not have?

- a. Hydrogen
- b. Carbon
- c. Oxygen
- d. Nitrogen
What level of protein structure is influenced primarily by hydrogen bonding?

- a. Primary structure
- b. Tertiary structure
- c. Secondary structure
- d. Quaternary structure
What level of structure allows for the hemoglobin to have four identical subunits?

- [ ] a. Primary structure
- [ ] b. Tertiary structure
- [x] c. Secondary structure
- [ ] d. Quaternary structure
• Which type of hemoglobin is apparently more polar and soluble in water?

• Which type of hemoglobin is more non-polar and insoluble?

  a. Normal hemoglobin  
  b. Sickle cell anemia
Proteins, which are embedded in outer membrane are

- a. Globular proteins
- b. Fibrous proteins
- c. α-helical membrane proteins
- d. β-barrel membrane proteins
Main properties of channel proteins are

- (a) Gating
- (b) Selective ion conduction
- (c) Gating and selective ion conduction
- (d) None of the above
3. Olfactory receptors are related with

- (a) Visual perception
- (b) Immune response
- (c) Sense of taste
- (d) Sense of smell
Margaret Dayhoff developed the first protein sequence database called

- a. UniPROT
- b. Atlas of sequence and structure protein
- c. SWISS PROT
- d. PDB
__________ is a comprehensive database of enzymes

- (a) STRING
- (b) BREND
- (c) DDBJ
- (d) UniProt
Sickle cell anaemia is due to

- (a) Mutation of Glu 6 to Val in haemoglobin
- (b) Mutation of Val 6 to Glu in haemoglobin
- (c) Mutation of Glu 6 to Val in myoglobin
- (d) Mutation of Val 6 to Glu in myoglobin
One PAM means one accepted point mutation per

- a. $10^2$ residues
- b. 10 residues
- c. $10^3$ residues
- d. $10^4$ residues
PAM120, PAM80 and PAM60 scoring matrices are most suitable for aligning sequences

- a. 40%, 50% and 60% similarity respectively
- b. 60%, 50% and 40% similarity respectively
- c. 60%, 40% and 50% similarity respectively
- d. 100% each
Which of the following cases are commonly used?

- a. gap opening penalty = -2, gap extension penalty = -0.5
- b. gap opening penalty = -0.5, gap extension penalty = -2.0
- c. gap opening penalty = -100, gap extension penalty = 0
- d. gap opening penalty = -100, gap extension penalty = -100
Which one of the following best represents the central dogma of Bioinformatics?

- a. Sequence-Structure-Function
- b. DNA-RNA-Proteins
- c. Motifs-Domains-Superfamily
- d. Data-Databanks-Data mining tools
Rank the amino acids in increasing order of polarity: A,E,L,V,S

- a. L, V, A, S, E
- b. S, A, V, L, E
- c. E, V, L, A, S
- d. S, A, L, V, E
In the system of nomenclature which one of the following enzyme occupies 1st position

- a) Oxidoreductase
- b) Transferase
- c) Hydrolase
- d) Ligase
Most commonly used PAM matrix for alignment of protein sequences is:

- (a) PAM-1
- (b) PAM-250
- (c) PAM-500
- (d) PAM-1000
CSA stands for ___________

- a) Critical Sequence Assess
- b) Catalytic Site Atlas
- c) Catalyst Site Atlas
- d) None of these
The central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

- a. Genbank
- b. Uniprot
- c. NCBI
- d. Wikipedia
UniProtKB is made up of 2 large sequence collections: the _____ (manually annotated and reviewed sequences) and _____ (automatically annotated and not reviewed sequences)
Task

1. In UniProt, how many mouse \textit{(Mus musculus)} protein sequences are manually annotated?

2. How many of these manually annotated protein sequences are associated with PDB?

3. Map UniProt IDs of above manually curated mouse protein sequences with 3D structures to STRING database. How many STRING IDs are mapped?

4. Construct dot plot for the alignment of human and chicken hemoglobin β chain. Identify the segments, which are same in both sequences.

https://www.ebi.ac.uk/Tools/seqstats/emboss_dotpath/
THANKS!

Do you have any questions?

Next session: 12 FEB 2022, 6-7 PM