Protein Sequence Analysis using Transformer-based Large Language Model

Presenters : Bishnu Sarker, Sayane Shome

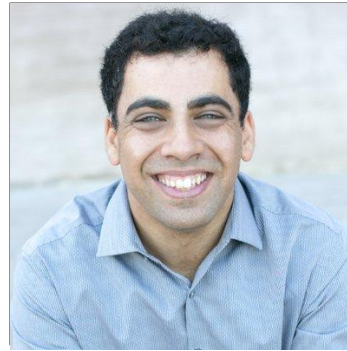**Date: 17-18 July, 2023**

Organizing team

Bishnu Sarker, Ph.D.
Assistant Professor
Computer Science and Data Science,
Meharry School of Applied Computational
Sciences,
Tennessee, USA

Sayane Shome, Ph.D.
Postdoctoral Fellow
Anesthesia and Pediatrics,
Stanford University School of Medicine,
California, USA

Farzana Rahman, Ph.D.
Assistant Professor
Applied Computer Science
(Data Science)
Faculty of Engineering, Computing
And Mathematics
Kingston University London, UK

Nima Aghaeepour, Ph.D.
Associate Professor
Anesthesia and Pediatrics,
Stanford University School of Medicine,
California, USA

Speakers

Bishnu Sarker, Ph.D.
Assistant Professor
Computer Science and Data Science,
Meharry School of Applied Computational Sciences,
Tennessee, USA
bsarker@mmc.edu

Sayane Shome, Ph.D.
Postdoctoral Fellow
Anesthesia and Pediatrics,
Stanford University School of Medicine,
California, USA
sshome@stanford.edu

Learning Objectives

- How to build basic machine learning models for sequence analysis.
- How to implement deep learning models such as Long-Short Term Memory and Recurrent Neural Networks (LSTM and RNN) in the context of biological sequence modelling.
- Fundamentals of transformer-based large language models.
- How to apply a pre-trained transformer language model for biological sequence analysis
- How to formulate and address biomedical problems using transformer-based large language models.
- What tools, frameworks, datasets, and programming libraries are available to work with transformer-based large language models for sequence analysis.

Link with all presentation materials/tutorials : https://sites.google.com/view/bishnusarker/ismb-eccb-2023-vt2

Tutorial Agenda: Part 1- Monday, July 17 (14:00 – 18:00 CEST)

Schedule (CEST time zone)	Topics covered in Part 1
14:00-14:30	**Instructor : Sayane Shome,PhD** ● **Introduction to the tutorial session** ● **Fundamental concepts about proteins from a biological perspective**
14:30-14:45	*Short Break and Q/As*
14:45-15:45	**Instructor : Sayane Shome, PhD** ● **Python Programming Refresher**
15:45-16:00	*Short Break and Q/As*
16:00-17:45	**Instructor : Bishnu Sarker,PhD** ● **Introduction to biological sequence analysis using Deep Learning in Python** ● **Building deep learning models (RNN, LSTM) for sequence analysis**
17:45 - 18:00	*Q/As*

Tutorial Agenda: Part 2- Tuesday, July 18 (14:00 – 18:00 CEST)

| Schedule (CEST time zone)	Topics covered in Part 2
14:00-15:00	**Instructor : Bishnu Sarker,PhD** ● **Introduction to Transformer-based Language models** ● **Transformers for biological sequence analysis**
15:00-15:15	*Short Break and Q/As*
15:15- 16:30	**Instructor : Sayane Shome,PhD** ● **Case study -1 : Protein Function Annotation**
16:30-16:45	*Short Break and Q/As*
16:45-17:45	**Instructor : Bishnu Sarker,PhD** ● **Case study -2 : Protein Metal-Binding Site Prediction**
17:45-18:00	**Q/As and Closing remarks**
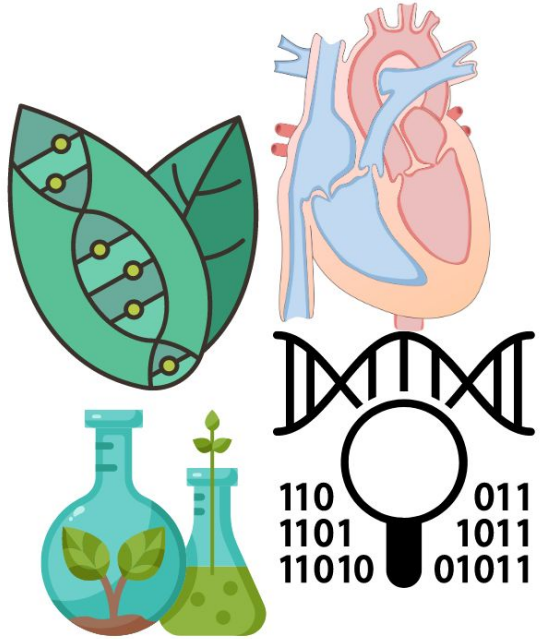
Fundamentals of Protein Biology

Biology Refresher

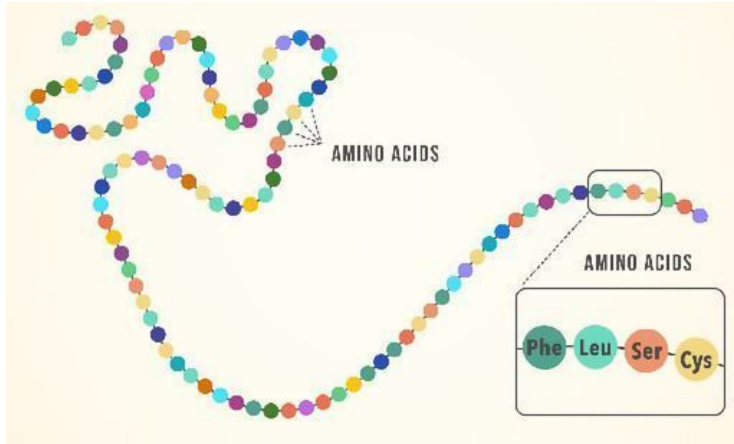# Learning Objectives of the session

To obtain a higher-level understanding of proteins and their role in living organisms from a biological perspective.
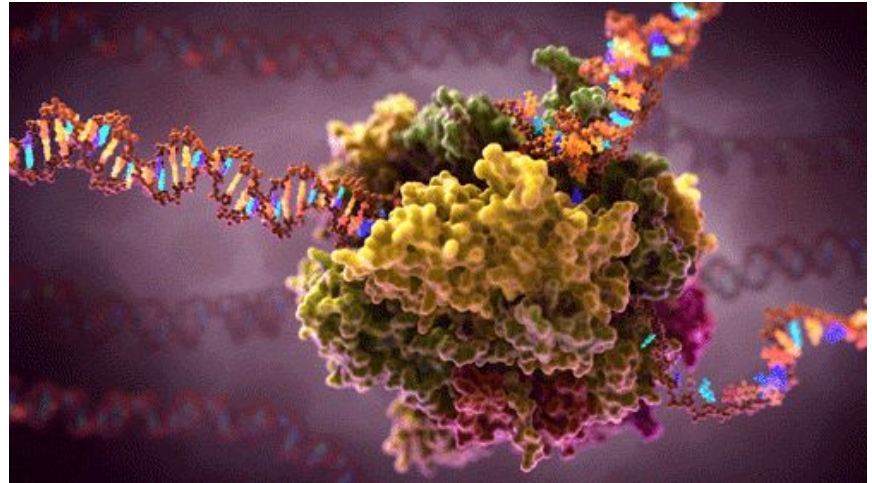
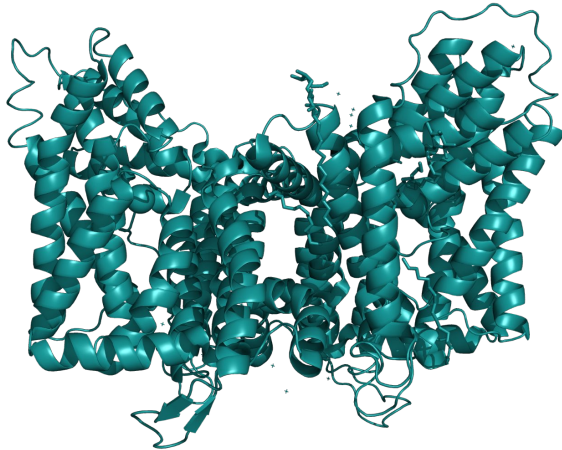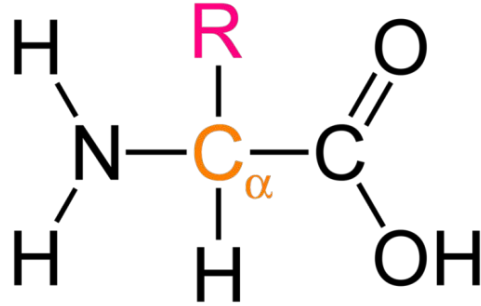# Proteins are building blocks of life
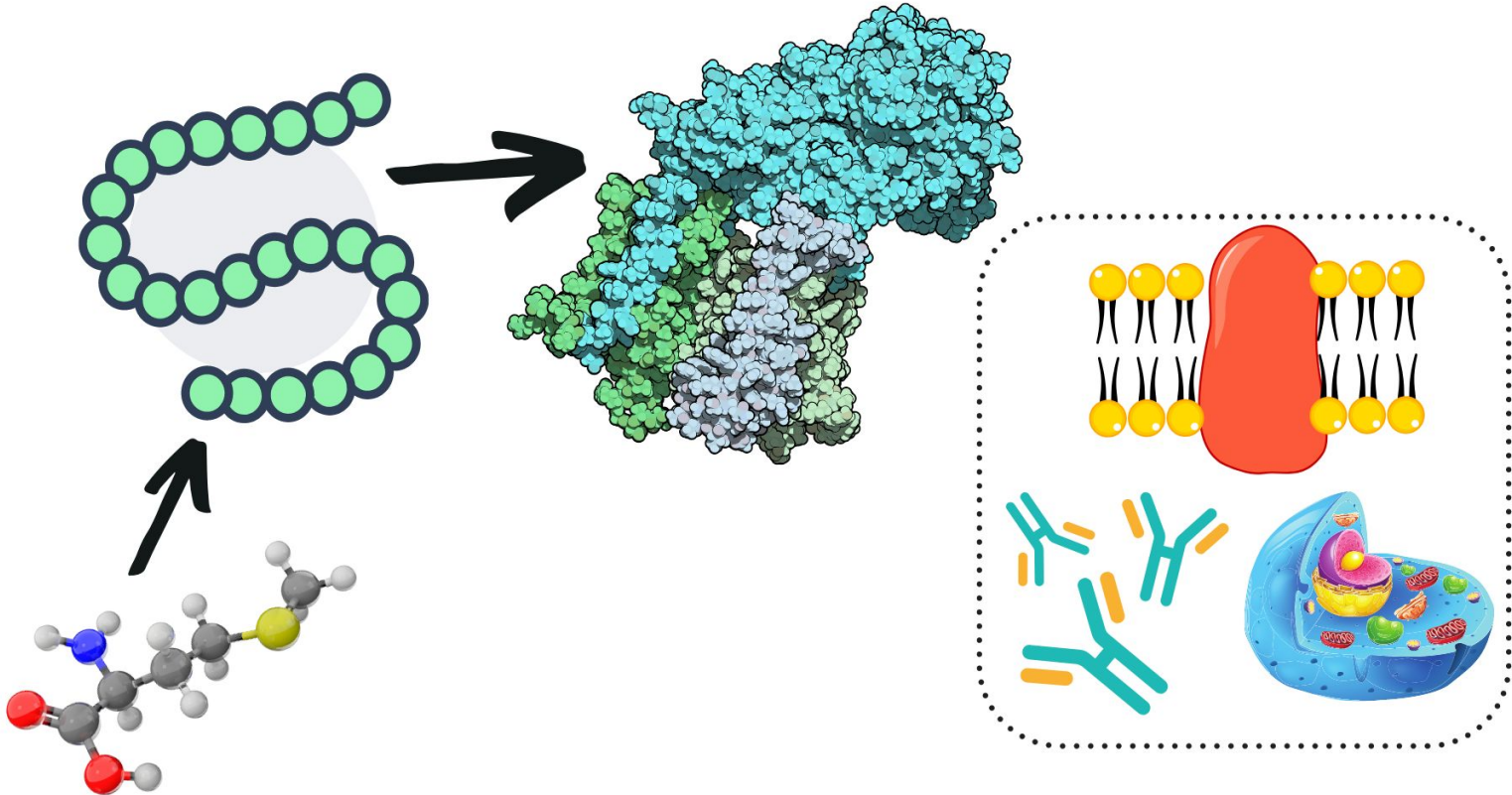
Proteins
Basic definition

AMINO ACIDS

AMINO ACIDS

Phe – Leu – Ser – Cys

- Proteins are macromolecules composed of amino acids.
- Proteins play a vital role in various biological processes.
- They are essential for the structure, function, and regulation of cells.

Amino acids and Protein structures

- Depending on the constituent of alkyl group in side-chain (R), there are 20 different amino acids.
- Amino acids are linked with peptide bonds leading to long polypeptide chains which fold in different manner to form tertiary structure of proteins.
- Different combinations and arrangements of amino acids result in a vast array of proteins.

Protein Domains and Protein Function

Domain 1
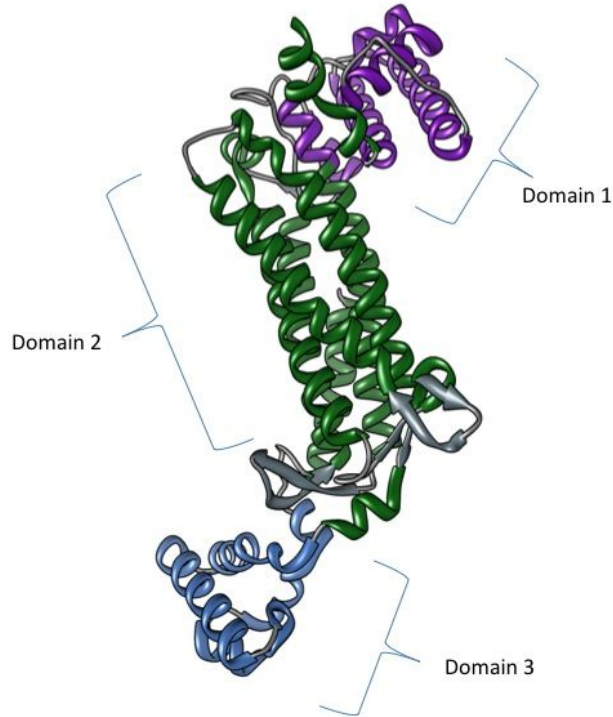
Domain 2

Domain 3

- Protein domains are distinct units within a protein that possess specific structures and functions.
- They provide modularity, flexibility, and evolutionary versatility to proteins, allowing for the development of complex functions and adaptation to diverse biological environments.
- They can be pose as sites for protein-protein interactions,metal binding sites and others.

Break !

We will reconvene in 15 mins. Meanwhile, we are available for Q/As

Next in line : **Python Programming Refresher**