ISMB ECCB 2023
LYON FRANCE
July 23-27
iSCB

Case Study 1 : Protein Function Annotation

Presenters : Bishnu Sarker, Sayane Shome

Date: 17-18 July, 2023

MEHARRY
SCHOOL OF APPLIED COMPUTATIONAL SCIENCES

Stanford
MEDICINE

Learning Objectives of the next two sessions

To expand the concepts we learnt in previous sessions into practical applications such as protein function prediction and metal binding site prediction in proteins.

Problem Definition

Given a protein sequence of length L,the objective is to assign functional terms such as Gene Ontologies or Enzyme commission number.

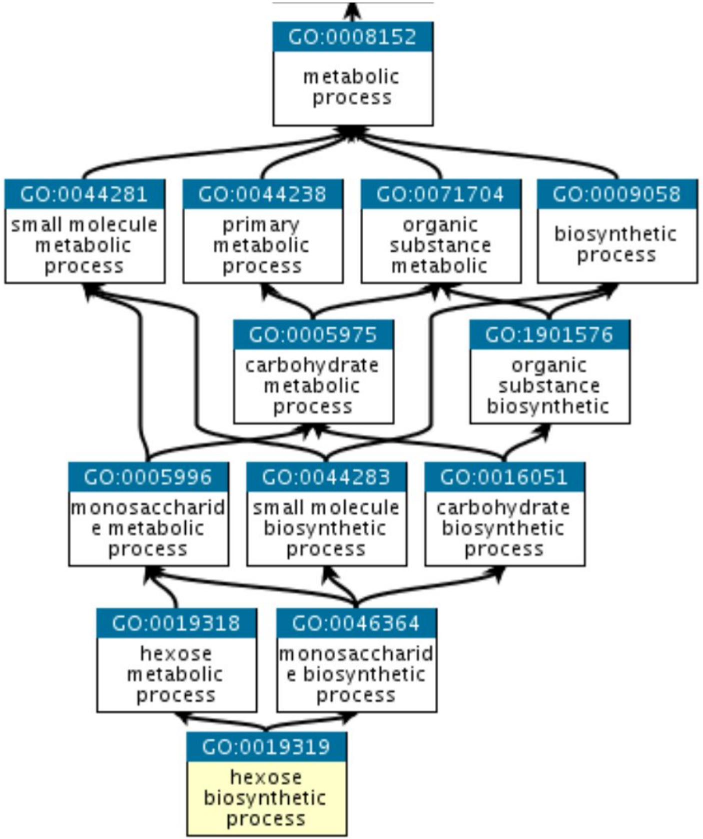
- Gene Ontologies(GO) is a standardized system that assigns functional terms to genes and gene products based on their known or predicted molecular functions, biological processes, and cellular components.
- Enzyme Commission (EC) numbers are a classification system used to categorize enzymes based on the reactions they catalyze. The EC number provides a unique identifier for each enzyme and is widely used in biochemistry and molecular biology.

Gene Ontologies

01 Biological Process

02 Molecular Function

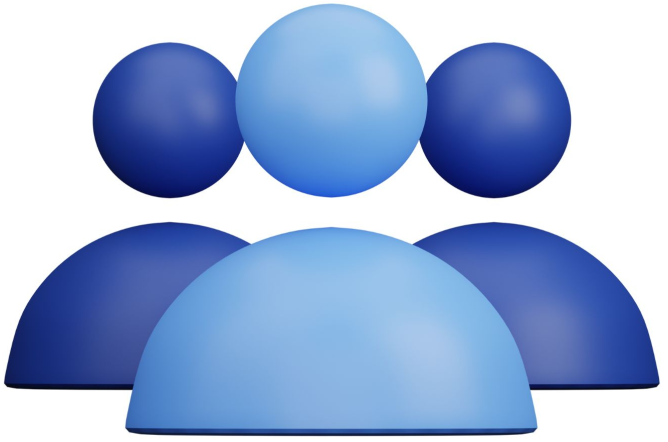
03 Cellular Component

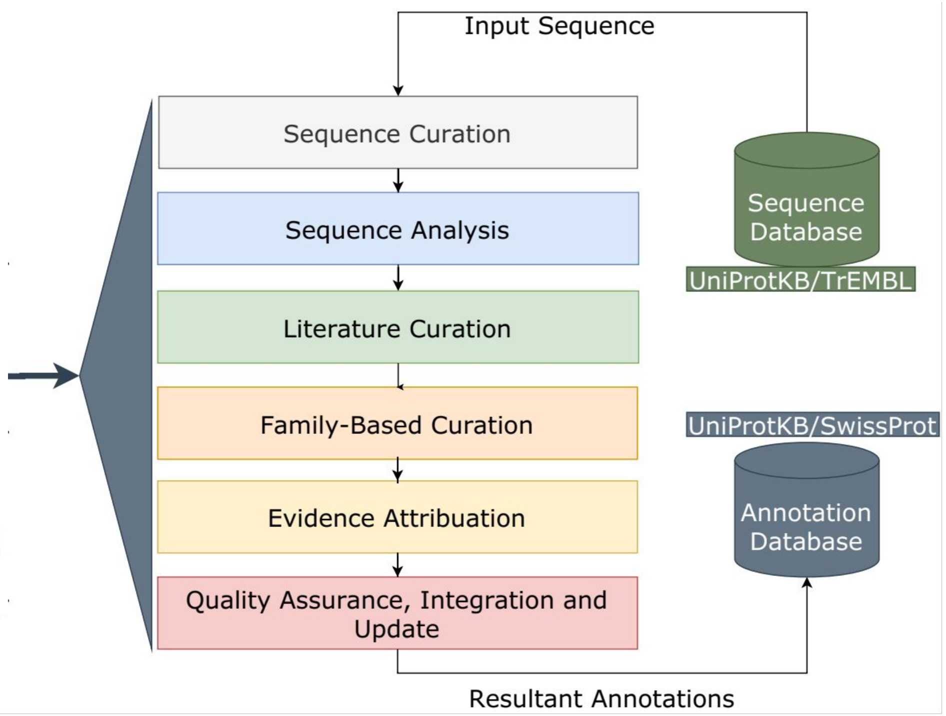

GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Background

Manual Annotation

Curators





Input Sequence

Sequence Curation

Sequence Analysis

Literature Curation

Family-Based Curation

Evidence Attribuation

Quality Assurance, Integration and Update

Resultant Annotations

Sequence Database

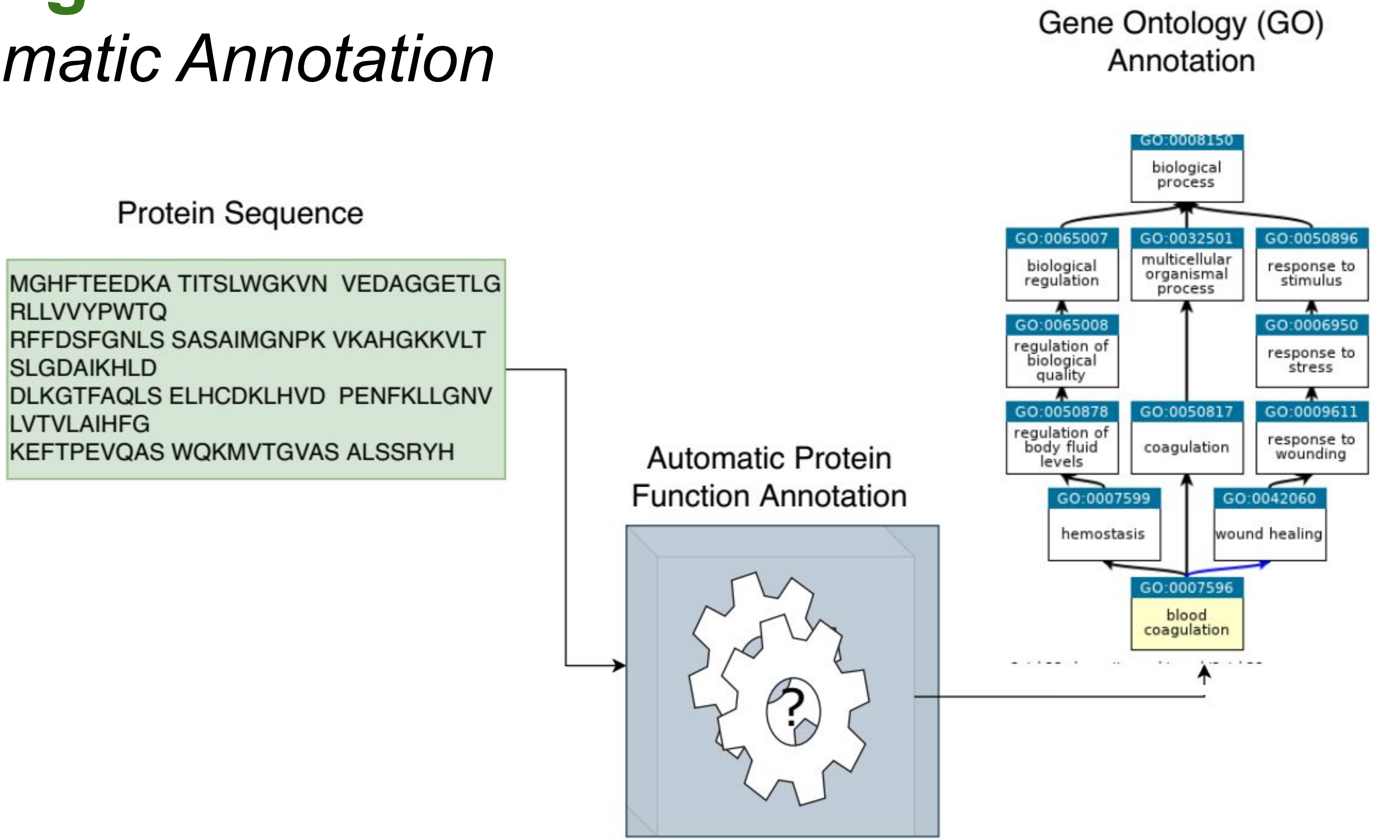
UniProtKB/TrEMBL

UniProtKB/SwissProt

Annotation Database

Background

Automatic Annotation



Protein Sequence

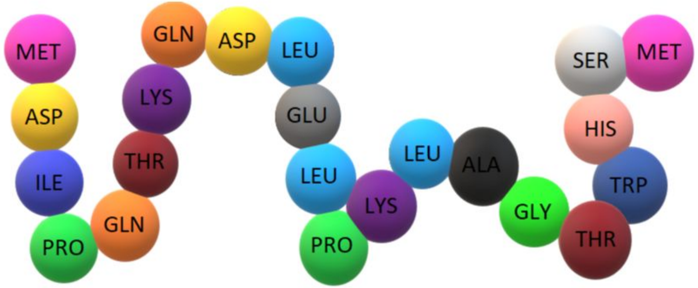
MGHFTEEDKA TITSLWGKVN VEDAGGETLG
RLLVVYPWTQ
RFFDSFGNLS SASAIMGNPK VKAHGKKVLT
SLGDAIKHLD
DLKGTFAQLS ELHCDKLHVD PENFKLLGNV
LVTVLAIHFG
KEFTPEVQAS WQKMVTGVAS ALSSRYH

Automatic Protein Function Annotation

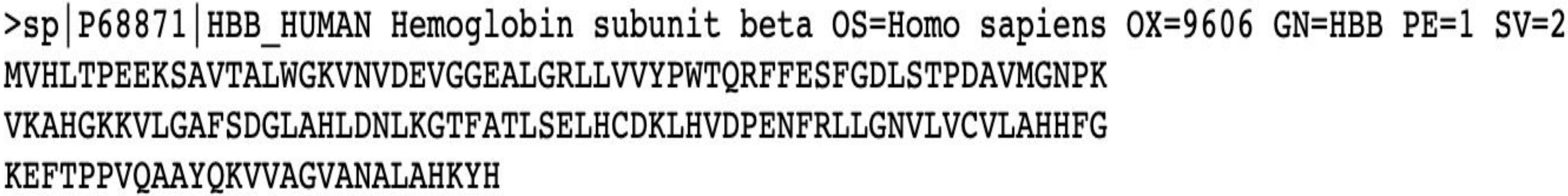
Gene Ontology (GO) Annotation

Protein Function Annotation

Input Data and Data Sources



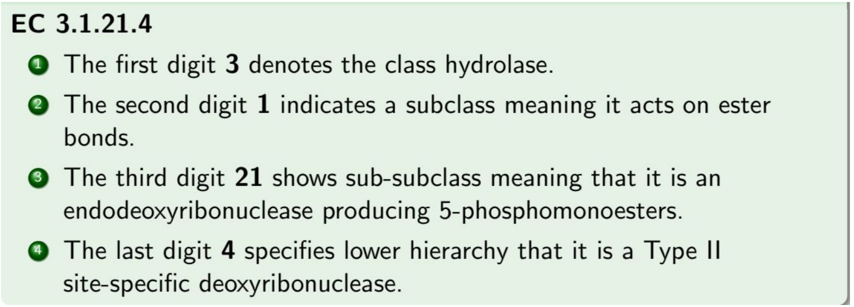
UniProt



```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

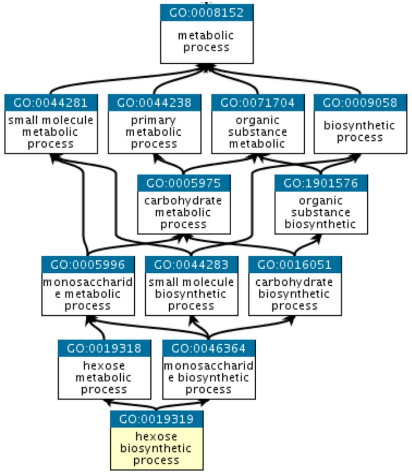
Protein Function Annotation

Output Data and Data Sources



EC 3.1.21.4

1. The first digit **3** denotes the class hydrolase.
2. The second digit **1** indicates a subclass meaning it acts on ester bonds.
3. The third digit **21** shows sub-subclass meaning that it is an endodeoxyribonuclease producing 5-phosphomonoesters.
4. The last digit **4** specifies lower hierarchy that it is a Type II site-specific deoxyribonuclease.


GO:0008152 metabolic process
GO:0044281 small molecule metabolic process
GO:0044238 primary metabolic process
GO:0071704 organic substance metabolic
GO:0009058 biosynthetic process
GO:0005975 carbohydrate metabolic process
GO:1901576 organic substance biosynthetic
GO:0005996 monosaccharide metabolic process
GO:0044283 small molecule biosynthetic process
GO:0016051 carbohydrate biosynthetic process
GO:0019318 hexose metabolic process
GO:0046364 monosaccharide biosynthetic process
GO:0019319 hexose biosynthetic process

BRENDA

GENEONTOLOGY
Unifying Biology

Protein Function Annotation

Approach

Obtaining protein sequence dataset from Uniprot and associated GO IDs/EC IDs

Obtaining pretrained embeddings for the protein sequence dataset from Uniprot

Using ML models for classifying the sequences with the GO IDs/EC IDs

Evaluating ML model performance using metrics

Protein Function Annotation

Future Challenges

01 Explainability

02 Multi-omics Integration

03 Computational Cost

Hands on Tutorial

Google colab notebook

Link : Colab-Notebook-Case-Study-1-Protein-Function-Prediction

Break !

We will reconvene in 15 mins.

Next in line : **Hands-on tutorial on Metal-binding site prediction**

ISMB
ECCB
2023
LYON
FRANCE
July 23-27
iSCB

ISMB ECCB 2023
LYON FRANCE July 23-27
iSCB

Case Study 2 : Metal Binding Site Prediction

Presenters : Bishnu Sarker, Sayane Shome
Date: 17-18 July, 2023

MEHARRY
SCHOOL OF APPLIED COMPUTATIONAL SCIENCES

Stanford
MEDICINE

Problem Definition

Given a protein sequence of length L and residue positions of the metal-binding sites in the protein,the objective is to find which metal ions will most likely bind to the sites.

We formulate this as a machine learning problem to be the focus of this hands-on tutorial.

Metal-Binding Site Prediction

Input/Output Data and Data Sources

UniProt

Input Data

1. Protein Sequences data
2. Protein residue positions at the binding sites

Output Data

1. Names of binding metal ions and ChEMBL ID

Metal-Binding Site Prediction

Approach

Obtaining protein sequence dataset from Uniprot and associated pretrained embeddings

Obtaining positional encodings for the residue positions encompassing the binding sites

Using ML models for predicting the metal ions binding at the sites

Evaluating ML model performance using metrics

Metal-Binding Site Prediction

Approach

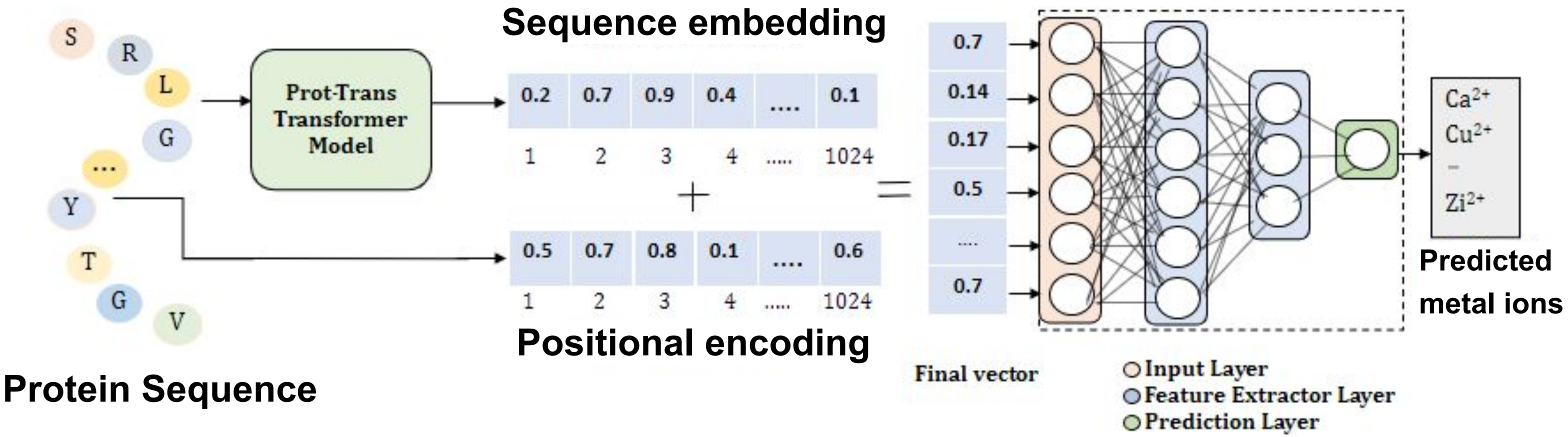
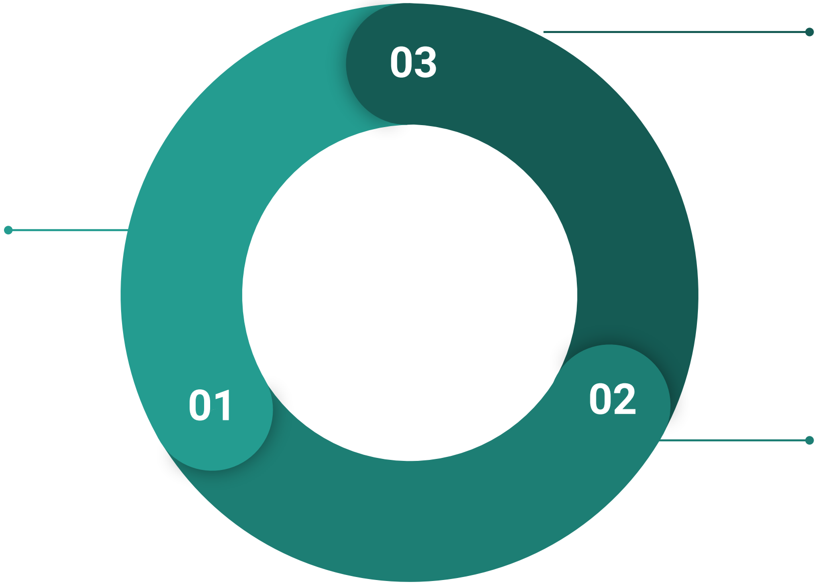


Figure from : *https://www.biorxiv.org/content/10.1101/2023.03.20.533488v1.full.pdf*

Metal-Binding Site Prediction

Current and Future Challenges



Explainability

Computational Cost

Metal binding site integration

Hands-on Tutorial

Google colab notebook

Link : Colab-Notebook-Case-Study-2-Metal-Binding-Site-Prediction

Acknowledgements

- ISMB/ECCB 2023 Tutorial committee chairs and reviewers
- Meharry Medical College,Tennessee,USA
- Stanford University,California,USA
- Kingston University,London,UK
- Participants !

ISMB
ECCB
2023
LYON
FRANCE
July 23-27
iSCB

MEHARRY
SCHOOL OF APPLIED
COMPUTATIONAL SCIENCES

Stanford
MEDICINE

Kingston
University
London

Thank you for joining us !

For any correspondence regarding questions about the materials and related topics :

- Bishnu Sarker (bsarker@mmc.edu)
- Sayane Shome (sshome@stanford.edu)

ISMB/ECCB tutorial Feedback Link!

Please provide your valuable feedback and suggestions!

SCAN ME