Hands-on case study on applications of LLMs in protein sequence analysis

Presenters : Cameron Pykiet, Jaylin Dyson, Bishnu Sarker

Date: 6th February, 2025

Learning Objectives

To expand the concepts we learnt in previous sessions into practical applications such as protein function prediction.

Problem Definition

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

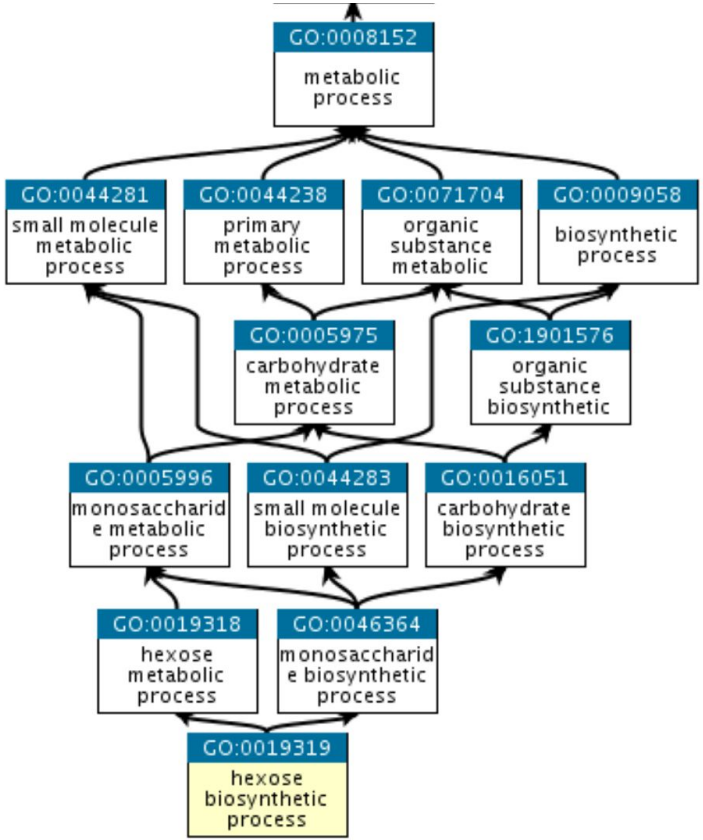
- Gene Ontologies(GO) is a standardized system that assigns functional terms to genes and gene products based on their known or predicted molecular functions, biological processes, and cellular components.
- Enzyme Commission (EC) numbers are a classification system used to categorize enzymes based on 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

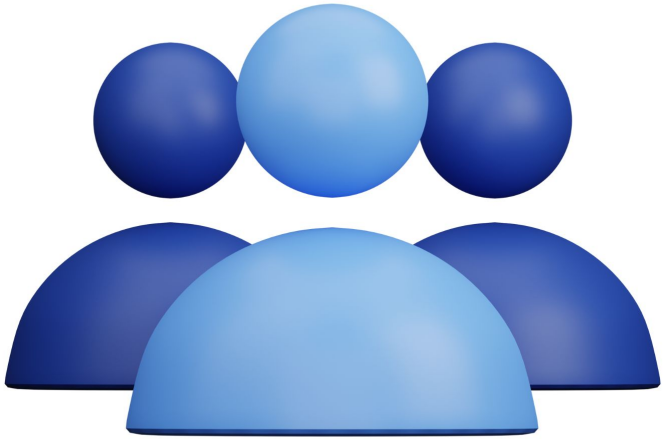
02 Molecular Function

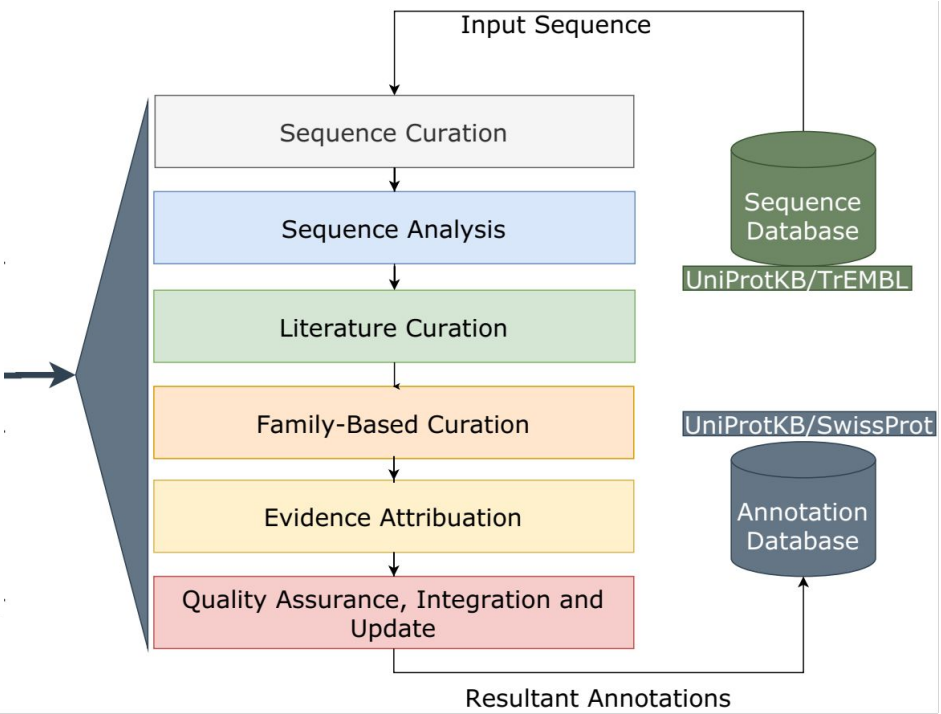
03 Cellular Component



Background

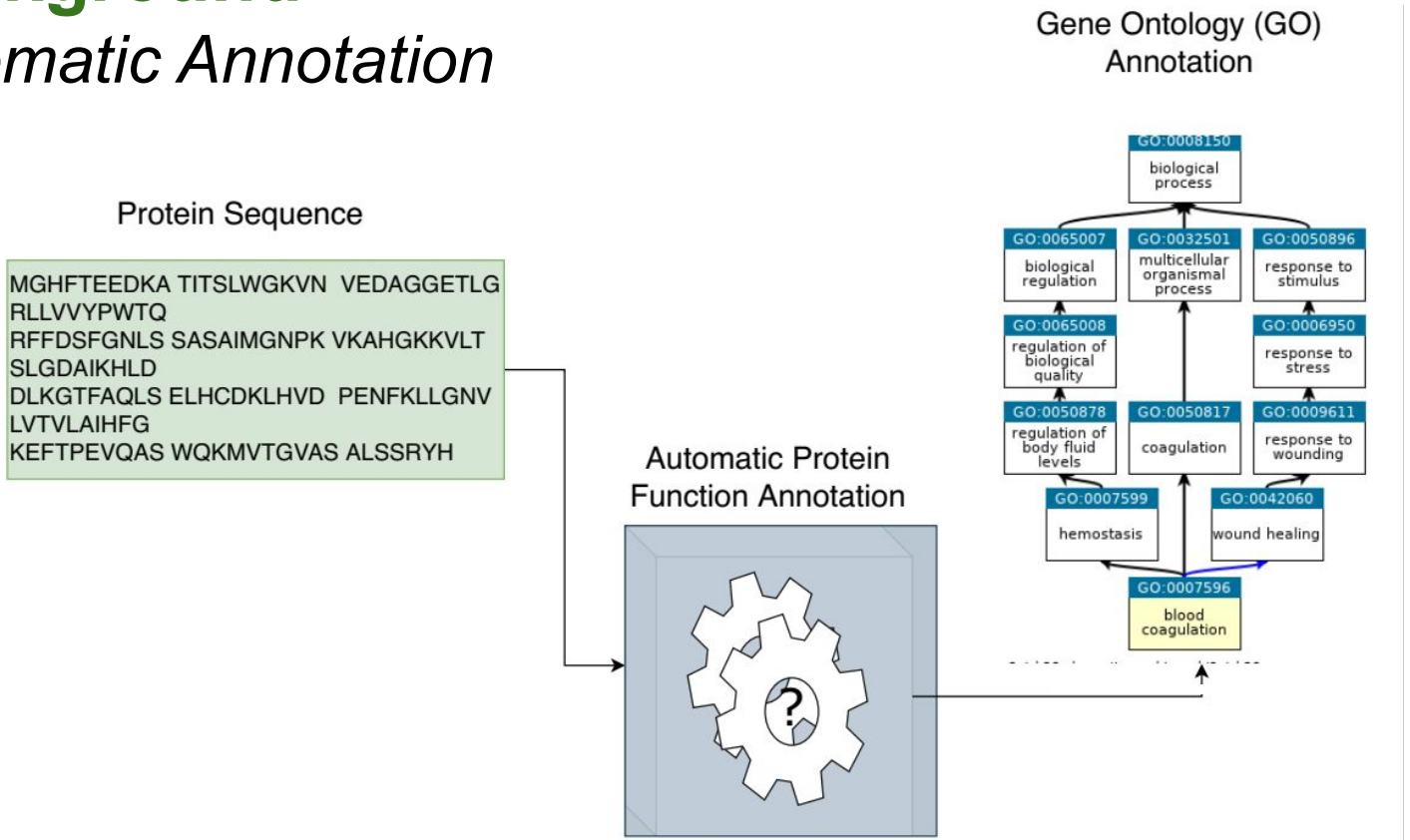
Manual Annotation





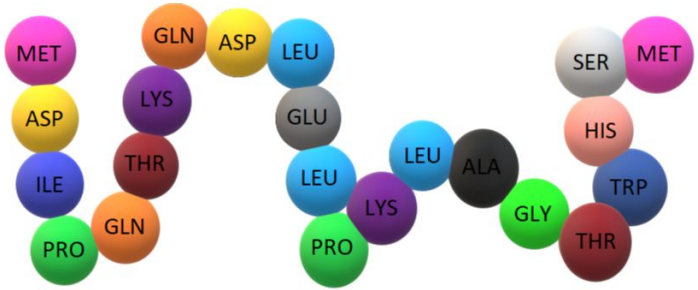
Background

Automatic Annotation



Protein Function Annotation

Input Data and Data Sources



UniProt

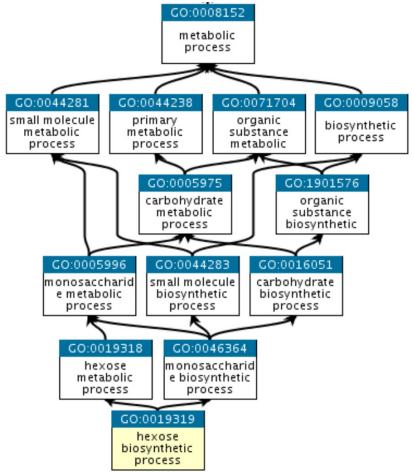
```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

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.



BRENDA

GENEONTOLOGY
Unifying Biology

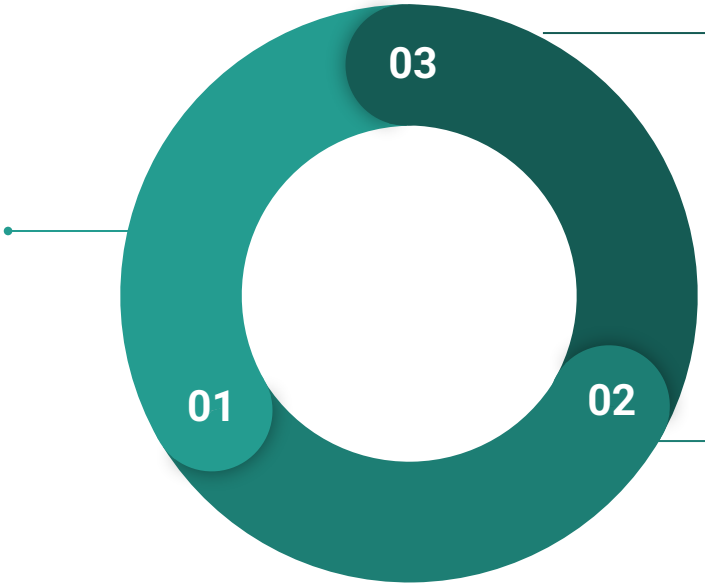
Protein Function Annotation

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



Computational Cost

Explainability

Multi-omics Integration

Q&A and Final Remarks!