Phase 1: UniProt open source Database

UniProtdao team

24/01/2025

Acknowledgment

The project was carried out by the *UniProtdao* team in partnership with Harvard University in Boston, USA.

Contents

1 Project description

Thanks to recent developments in genomic sequencing technologies, the number of protein sequences in public databases is growing enormously. In order to fully exploit this huge amount of data, protein sequences need to be associated with functional properties such as Enzyme Commission (EC) numbers and Gene Ontology terms. The UniProt Knowledge Base (UniProtKB) is currently the largest and most comprehensive resource for protein sequence and annotation data. According to the March 2018 release of UniProtKB, some 556,000 sequences are manually curated but over 111 million sequences lack functional annotations. The ability to automatically annotate protein sequences in UniProtKB/TrEMBL, the non-reviewed UniProt sequence repository, would represent a major step towards bridging the gap between annotated and unannotated protein sequences. The aim of this project is to first construct a graph representation of the UniProt protein database (or a part of it!) in which each node of the graph represents maintains some protein-related attributes. These attributes include, for example, protein taxonomy, function, and sequence. An edge between two nodes means that the linked proteins are similar (e.g., they belong to the same InterPro group) or they share the same protein signatures (e.g., domains and functional sites). Based on one or more attributes of interest, a protein graph may be constructed. The protein graph assigns to the reviewed proteins its attribute value. The unreviewed proteins are unlabeled nodes. The graph representation of the protein data may then be inputted to a label propagation algorithm that aims to infer attribute values of the unlabeled nodes.

2 Graph Construction

A common way of building the protein graph is to use the domain composition of protein data. This is a novel way of connecting the proteins using their constituent protein domains. Domains may be considered as natural building blocks of proteins. During evolution, protein domains have been duplicated, fused, and recombined in different ways to produce proteins with distinct structures and functions. Here, each node of the network represents a protein, while a link between two nodes means that the proteins exhibit a given minimum level of domain similarity. Thus, each node u is identified by a set of labels $L(u)$ (one or more annotations to propagate), has a set of neighbours $N(u)$, and for every neighbour v, it has an associated weight $W_{u,v}$.

3 Domain Based Protein-Protein Network (PPN)

To illustrate the construction of the PPN, let us consider five proteins with symbolic names P_1, P_2, P_3, P_4, and P_5. Let us assume that these proteins are composed of domains

$$D_1 = \{d_1, d_2, d_3, d_4\}, \quad D_2 = \{d_1, d_3, d_5\}, \quad D_3 = \{d_1, d_2, d_{10}\}, \quad D_4 = \{d_5, d_6, d_1\}, \quad D_5 = \{d_4, d_1, d_{10}, d$$

It is then evident that proteins P_1 and P_2 contain two domains in common, namely d_1 and d_3. Therefore, proteins P_1 and P_2 may be linked, and the number of shared domains may serve as the link weight such as

$$W_{P_1,P_2} = |D_1 \cap D_2| = |\{d_1, d_3\}| = 2.1$$

In a similar way, proteins P_1 and P_5 may be linked with a link weight of

$$W_{P_1,P_5} = |D_1 \cap D_5| = |\{d_1, d_4\}| = 2.2$$

In both cases, the link weight is two. However, the link weight computed in this way does not reflect the true strength of the relationship among the proteins. This is because, in the first case, there are a total of

$$|D_1 \cup D_2| = |\{d_1, d_2, d_3, d_4, d_5\}| = 5$$

different domains among the two proteins, and two are shared. In the second case, there are

$$|D_1 \cup D_5| = |\{d_1, d_2, d_3, d_4, d_{10}, d_{40}, d_7, d_9, d_{12}, d_{52}, d_{100}\}| = 11$$

different domains, of which two are again shared. Although two domains are shared in each case, P_1 is intuitively more aligned with P_2 than P_5.

Therefore, instead of using the above raw similarity score, we use the Jaccard index, or Jaccard similarity coefficient, to better reflect the similarity in composition. This is calculated as

$$\frac{|A \cap B|}{|A \cup B|},$$

where A and B are the two sets of constituent domains. Using the Jaccard coefficient, the link weights for P_1 and P_2 are calculated as:

$$W_{P_1,P_2} = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} = \frac{|\{d_1, d_3\}|}{|\{d_1, d_2, d_3, d_4, d_5\}|} = \frac{2}{5} = 0.4.3$$

In other words, according to the Jaccard measure, protein P_1 and P_2 are 40% similar in their domain composition.

Similarly, for P_1 and P_5, the Jaccard link weight is calculated as:

$$W_{P_1,P_5} = \frac{|D_1 \cap D_5|}{|D_1 \cup D_5|} = \frac{|\{d_1, d_4\}|}{|\{d_1, d_2, d_3, d_4, d_{10}, d_{40}, d_7, d_9, d_{12}, d_{52}, d_{100}\}|} = \frac{2}{11} \approx 0.18.4$$

In this case, P_1 and P_5 are roughly 18% similar in their domain composition.

The graph is a weighted undirected graph. It contains nodes that are labelled and nodes that are unlabelled.

4 Querying the Protein Database

The goal of this task is to propose to the user the possibility to query the protein database:

1. Search a protein/node by its identifier and/or name and/or description. As a result, the user can view the protein, its neighbors, and the neighbors of neighbors.

2. Compute some statistics such as:

   - The number of labelled and unlabelled proteins.
   - Isolated proteins (proteins with no neighbors).

3. Visualize a specific protein (and its neighborhood).

5 Protein Function Annotation Task

The goal of this task is to perform a classification/annotation task. The protein network contains a few nodes that are labelled with respective functions, whereas a large number of nodes do not have any function annotations. Annotated nodes can be labelled with terms from Gene Ontology (e.g., `GO:10004`) or Enzyme Commission numbers (e.g., `EC 1.3.5.25`).

This is a multi-label classification problem, as each node can have more than one label. The problem can be viewed as:

- A link completion or recommendation system task, where we need to recommend appropriate links to connect unannotated proteins with their respective functions, or
- A general multi-label machine learning problem, where annotated nodes are the training examples.

6 Project Data

6.1 Small Dataset

A pre-processed dataset that contains domain and EC information for Viruses is available in *Arche*.

6.2 Big Datasets

Raw data are accessible from the following sources:

1. `ftp://ftp.ebi.ac.uk/pub/databases/interpro/protein2ipr.dat.gz`: Here, each protein from UniProt is associated with their corresponding domain information in detail.
2. `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz`: This data contains GO annotations of each UniProt protein.
3. `ftp://ftp.expasy.org/databases/enzyme/enzyme.dat`: Contains EC information of UniProt proteins.
4. `http://purl.obolibrary.org/obo/go/go-basic.obo`: This file contains the GO Ontology. To realize the global picture of the annotations and inference, it will be very helpful to have a graph with all the information from these three files.

7 Project Evaluation

7.1 The Defense

Duration: 20 minutes
Structure: 5 minutes for technological choices and 15 minutes for system demonstration.

7.2 Mark

1. **Basic functionalities:** The creation of a NoSQL database for storing and querying protein graphs ensures a mark of 10/20. This mark takes into account the quality of the presentation, the motivations of technological choices, and the performance of the implemented system. Without basic functionalities, the attributed mark will be 0/20.

2. **Additional functionalities:**

   - Graphical interface: 3 points. The system proposes an intuitive GUI.
   - Good visualization of protein neighborhood: 2 points.
   - Personalized label propagation on protein graphs: 3 points. This task consists of performing the protein function annotation using several attributes (GO terms, EC numbers, etc.).
   - Statistics: 2 points.