Annotation of Proteins from LOD for a Viewer of Multiple Protein Sequence Alignment

Atsuko Yamaguchi[1] and Hiroyuki Toh[2]

[1] Database Center for Life Science (DBCLS),
Research Organization of Information and Systems (ROIS).
178-4-4 Wakashiba, Kashiwa, Chiba, 277-0871 Japan
atsuko@dbcls.rois.ac.jp
[2] Department of Biomedical Chemistry, School of Science and Technology,
Kwansei Gakuin University.
2-1 Gakuen, Sanda, Hyogo, 669-1337 Japan

Abstract. Recent progress in sequencing technology has caused so-called data flood of nucleotide and amino acid sequences. Such enormous amount of data can increase the reliability in prediction of structures and functions of proteins based on a multiple alignment. At the same time, however, it has become difficult to analyze a multiple alignment with the user-interactive manner due to the volume of sequences constituting the alignment. Therefore, we have worked on developing an alignment viewer that is referred to as ASHViewer. In this study, we focused on the annotation for a large number of proteins which is essential for the comparative study with alignment. To obtain various types of annotation for a large number of proteins all at once, we employ a method that uses federated query search along a path of class-class relationships. To clarify the effectiveness and problems using the method for alignment viewer, we implemented a prototype system of federated query search called PSurfer.

Keywords: Linked Open Data, database integration, federated query search, multiple protein sequence alignment

1 ASHViewer

In ASHViewer, a molecular phylogenyetic tree for a given alignment is constructed at first. Then, aligned sequences are divided into several groups, each of which corresponds to a subtree of the constructed tree. A coarse-grained tree with terminal nodes, each of which represents a collapsed subtree, is shown in a left panel of the window of ASHViewer. The corresponding consensus sequences of the alignment is shown in the right panel. The coarse-grained tree can be collapsed or deployed by clicking a terminal or internal node in the left panel. In accordance with the operation, the alignment of consensus sequences with different resolution is generated in the right panel. This method can improve the readability of large number of aligned sequenes. To obtain a knowledge from the alignment, Various information is necessary for annotation associated with the proteins, such as gene product properties like Gene Ontology (GO) terms, active sites, and chemical compounds that interact with the proteins. The required information is expected to be different among users because of the difference in their objectives for the comparative analyses. To obtain the annotation related to the aligned protein from multiple datasets, we adopted a method to extract the required information from LOD datasets as described below.

2 PSurfer

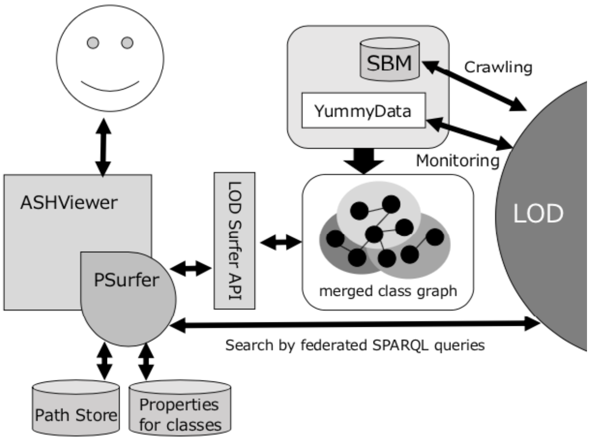

Fig. 1. Overview of PSurfer for ASHViewer

In our previous study, we developed the LOD Surfer API designed to easily develop a search system based on class-class relationships[1]. Using this API, a list of classes, a list of classes reachable using links from a given class, and paths between two given classes can be obtained. However, it is not easy for ordinary biologists to select an input class including IDs of proteins in MPSA, an output class with the required information related to the proteins in MPSA, and a path from an input class to an output class so that a SPARQL query automatically generated from the path can retrieve the required data. Therefore, we developed a system that connects ASHViewer with the LOD Surfer API called PSurfer. PSurfer uses a storage system of selected paths and a manually curated set of properties for some classes to show important human-readable information for the instances besides rdfs:label. Figure1 depicts an overview of the relations between ASHViewer, LOD Surfer API and PSurfer. Using PSurfer, a user of ASHViewer can obtain various type of annotations from LOD for the large number of aligned proteins at sequence or residue level, to extract novel knowledge from the alignment.

Acknowledgments This work was supported by JSPS KAKENHI grant numbers 17K00434 and by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

1. Yamaguchi, A., Kozaki, K., Yamamoto, Y., Masuya, H., Kobayashi, N.: LOD Surfer API: Web API for LOD Surfing Using Class-Class Relationships in Life Sciences. 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences CEUR Workshop Proceedings 2042, 2017