Semantic integration of the genome annotations

Toshiaki Katayama[1], Shinobu Okamoto[1], Shuichi Kawashima[1], Hiroshi Mori[2], and Takatomo Fujisawa[3]

[1] Database Center for Life Science, Research Organization of Information and Systems , Tokyo, Japan
ktym@dbcls.jp, {so,kwsm}@dbcls.rois.ac.jp
[2] Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Tokyo, Japan
hmori@bio.titech.ac.jp
[3] National Institute of Genetics, Research Organization of Information and Systems, Mishima, Japan
tf@nig.ac.jp

Abstract. Integration of the genome annotations is gaining more importance to interpret the biological meanings of large scale sequence data produced by the new sequencing technologies. Because existing annotations from public databases and literatures varies in terms of both categories and species, the Semantic Web technology has a great advantage for accumulating those wide-ranging information without having difficulty in the integration process. During the BioHackathon 2012, a new ontology to define the location of the annotations was defined as the Feature Annotation Location Description Ontology (FALDO). This ontology will be used to integrate annotations from INSDC (DDBJ/EMBL/GenBank) and UniProt databases, GFF3 formatted files, and many other bioinformatics resources. In parallel, Database Center for Life Science (DBCLS) and DNA Data Bank of Japan (DDBJ) has been jointly developing a RDF-based genome database which consists of the five layers. 1) RDF-based annotation data store, 2) SPARQL-based query engine, 3) RESTful API to retrieve genomic information, 4) HTML/CSS-based reusable web components and 5) integrated Web user interface. Our proposal is to standardize those layers so that every researcher can jointly use and/or update distributed genome annotations.

Keywords: genome annotation, database integration, semantic web, ontology

Introduction

Reliable genome annotations are essential for understanding the existing and newly sequenced organisms. A number of model organism databases have been already developed for from human to pathogens including prokaryotes. However, thanks to the rapid evolution of the next generation sequencers, the number and volume of the

sequenced organisms are growing exponentially. To understand the meaning of this vast amount of data, the importance of the reference annotation database is increasing. Even with the existence of many public databases, accumulation of those annotations is a difficult task because 1) annotations are still hidden in the literatures for most organisms and annotations of genes often require experimental verifications, 2) data formats used for the annotations were not always standardized, 3) there are no standard system to annotate any regions on the genomic sequence with a flexible yet controlled manner. The first problem requires text mining technologies and collaboration with biologists, the latter two issues can be resolved if our community agreed on a common standard. One of those efforts is the Generic Feature Format (GFF) and the Distributed Annotation System (BioDAS) which were introduced by the Generic Model Organism Database (GMOD) projects initially developed for the WormBase, FlyBase and some other model organism databases. Although the GFF format and the BioDAS protocol have been considered as standards for sharing genomic annotations, there still are some limitations. One is the lack of semantics in the GFF format which brought local variations such as the Gene Transfer Format (GTF). Also, especially for curators, it is very difficult to use the GFF format for describing sequence features not directly linked with the Sequence Ontology (SO) terms, or the semantic relations among sequence features such as interactions or regulations.

Results

To describe the missing semantics in the existing genome annotations or to add new heterogeneous annotations, we introduced the Semantic Web technology. It utilizes ontologies consisting of controlled vocabularies and their semantic relations, for describing objects to be annotated. Initially, we gathered reference knowledge from existing public data sources such as RefSeq, GFF3, and UniProt. We also merged annotations for prokaryote genomes from GTPS and MBGD databases and annotations of animal genomes from the H-inv database. All of those information are converted into RDF format and stored in a dedicated triple store. Along with this development, we are also developing a common genome annotation ontology.

During the NBDC/DBCLS BioHackathon 2012 held in Japan, developers of life science databases and applications gathered and agreed on to develop new ontology for describing locations of the objects. As a result, the Annotation Location Description Ontology (FALDO) was proposed and we converted the locations and positions of all annotations stored in our triple store to comply with this new standard.

Our proposed new RDF-based genome database is consisted of the five layers. 1) RDF-based annotation data store, 2) SPARQL-based query engine, 3) RESTful API to retrieve genomic information, 4) HTML/CSS-based reusable web components and 5) integrated Web user interface (Figure 1). To make the database generic as much as possible for any organisms, we introduced the following design principles.

First, to identify any object in the database, data provider must assign a unique URI for any object in the dataset. However, designing well-formatted URIs for every heterogeneous object without any confusion is a difficult task. Therefore, for this

purpose, we recommend to use unique URIs based on the Universally Unique Identifiers (UUIDs). Because a UUID can be independently and locally generated which will not collapse with the other ID in theory. One might think blank nodes in the RDF can be also applicable for this purpose, however, the use of UUID-based URIs can assure that the annotation for any object can be globally identifiable.

Second, we recommend to use existing ontologies such as Sequence Ontology (SO) and FALDO to identify the type and location of the annotated object on the sequence.

Third, it is recommended to develop a new ontology for each dataset which describes data source specific annotations such as the Feature/Qualifier used in the INSDC databases, specific tags in the H-inv database or the GFF3O ontology for the GFF3 data format. Then, each ontology should be linked with the new common genome annotation ontology so that users can use existing reasoning tools to generate interoperable triples which are subjected to be consumed by the genome database interface.

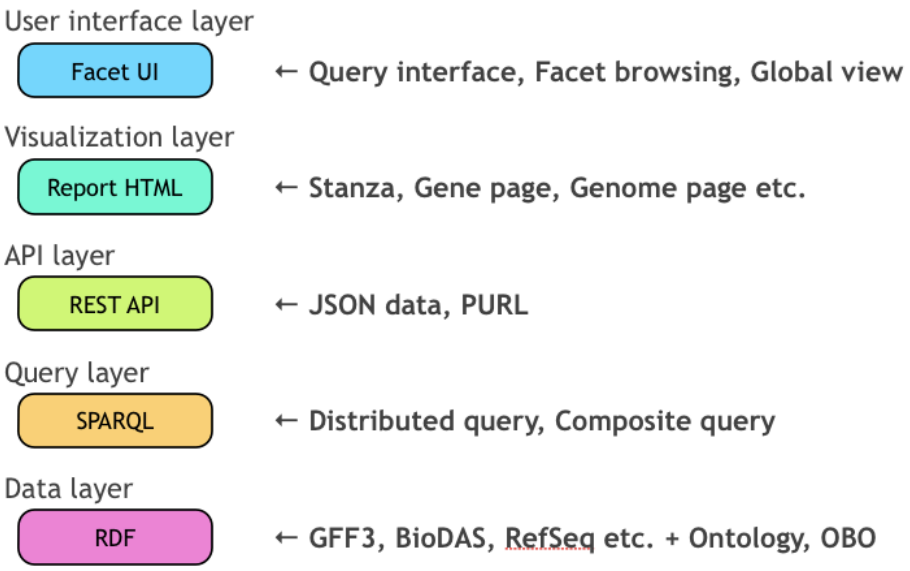

Figure 1. Layers of the RDF-based semantic genome database.

Discussions

We are still in a very early stage of the development. Therefore, we are inviting collaborators for standardizing each of the five layers within the life science community so that every researcher can jointly use and/or update distributed genome annotation databases.