Doctoral Dissertation

# Database Integration and Software Development for Bioinformatics using Computational Biosemantics

（計算バイオセマンティクスを用いた バイオインフォマティクスのための データベース統合とソフトウェア開発）

Yusuke KOMIYAMA

*Bioinformation Engineering Laboratory,*

*Department of Biotechnology,*

*Graduate School of Agricultural and Life Sciences,*

*The University of Tokyo*

Supervisor

Professor Kentaro SHIMIZU

25 December, 2013

博士学位論文

込山 悠介

*東京大学*

*大学院農学生命科学研究科*

*応用生命工学専攻*

*生物情報工学研究室*

指導教員

清水謙多郎 教授

2013 年 12 月 25 日

# Table of Contents

# Chapter 1 BACKGROUND

Chapter 1 describes the background and objective of this study. We define computational biosemantics as “a data-mining method from the Web for finding of biological novel discovery. It uses Semantic Web technologies.” “Biosemantics” is commonly utilized in the field of philosophy [1], which discusses the sense and value of a vital by symbolization of a vital phenomenon. Gene Ontology (GO) is an application in the life sciences [2,3]. Ontology is also a philosophy-related term, which defines principles that define the existence of things.

Bioinformatician is collecting data now cruising many databases on the scene that conducts data analysis. We have to shorten the period of drug design and medical studies by simplifying data set preparation. The active upgrade of a database with a relational database involving the redesign of the schema of a datum is not always ideal, and can have negative implications. Different standards and formats are used for different databases. Before extending a database record, a database engineer needs to consider and clarify possible database licensing issues. The license of dumping datum utilization was ambiguous in many cases.

On the other hand, Semantic Web can be used to express a database and convert resource description framework (RDF) using a graph structure [4]. We can add a novel record to the database being used by extending a network. It is collateralized by uniform resource identifier (URI) where an original datum belongs and is computer friendly. A licensable problem can solve Linked Open Data (LOD) using Web extensions [5]. We propose the use of Semantic Web by exploiting a life sciences database to overcome the disadvantages of the methods available till date. In addition, LOD should be continuously created using a project with different scales. For instance, BioHackathon, which is an international developer workshop [6–9], and LOD Challenge Japan [10,11] are members of very active communities. RDF is expected to serve as a standard format in the bioinformatics of the future, which is realized by changes in biosemantics. Furthermore, secure life-science data, in which personal information is enciphered, can predict the production of various applications. It is important for LOD other than life sciences to further develop for the benefit of society. This study aims to realize the database integration of proteomes and interactomes, which use biosemantics and tool development.

## 1.1 Introduction

### 1.1.1 Current Life Science Database

First, we describe the situation regarding existing life science databases. In Japan, the government organization National Bioscience Database Center (NBDC) [12] and the research institute Database Center for Life Science (DBCLS) are responsible for unifying life science data [13]. The protein structure database Protein Data Bank Japan (PDBj) [14,15], the genome database DNA Data Bank of Japan (DDBJ) [16], and the metabolism pathway database Kyoto Encyclopedia of Genes and Genomes (KEGG) [17,18] are well-known international biological databases. From 2008 to 2013, NBDC/DBCLS has hosted an annual international workshop called BioHackathon, which has seen participation of life science database researchers from other countries in addition to Japanese researchers. Different groups of researchers, programmers, and bio-curators Universal Protein Resource (UniProt) [19], ChEMBL [20], among others, collaborated under one platform [21]. They integrated the bioscience datum and are continually available to the public. The international bioscience consortium furthers the research and development of component engineering required for such databases. We believe that a global databank should be based on the historical circumstances of the past genome project. Therefore, in the database, computational biology is more advanced than other sciences and technological fields. However, realizing the globalization of Japan's experimental data is challenging because of the existing limitations to information disclosure required to acquire patents or publish papers. Moreover, biologists do not have resource such as capital, labor, skill, or time essential for converting experimental results to databases. In addition, situations exist where it is not possible to write a treatise even if the researcher updates the database. As a result, it is not easy to train personnel as bioscientific database researchers. In the field of bioscience database, the International Society for Biocuration organizes the international conference Biocuration.

### 1.1.2 Conventional Method using Relational Database (RDB)

In 1970, Edgar Frank Codd (1923–2003) proposed the concept of relational database (RDB) [22]. Nowadays, the RDB approach is the most widespread approach globally. The designer of the database determines its purpose, arranges the information, and designs the entity's relationship model. With respect to RDB, calculations that use relational algebra and related logic by defining the set of data according to their relations and attributes are possible. In the data modeling of RDB, it is necessary to regularize it to exclude data redundancy. The

entity (substance) included in the row (attribute) of the database should have a unique key. It does not permit an empty instance; repetition and inapt expressions can be prevented. Enterprises, governments, municipal offices, and educational institutions have been using RDB as a basic technology because it is very stable and reliable. However, with RDB, it is difficult to edit the table after databases are run; therefore, it limits database integration. SQL is one of the most famous query languages used for relational databases.

### 1.1.3 Proposed Idea of Biosemantics for Bioinformatics

Semantic Web is a new technique in the database field and is used to share data resources. Semantic Web is a standing rule for data mutual reference in the World Wide Web (WWW) and was proposed by Tim Berners Lee [23]. However, till date, this technique has not been achieved due to computational limitations. Moreover, it did not fit the current trends, and was not fully understood by many. However, with the rapid improvements in computer hardware technologies over the last few years, W3C has recommended RDF and web ontology language (OWL) based on Semantic Web. In the future a Web user will come to acquire information from the web of data like the present Internet (web of a document). Currently, the public has prompt access to data from many knowledge sources, and this information can be used by the application. In general, the packaging method for Semantic Web uses the RDF format. Triples are modeling of the subject, predicate, and object by using the directed graph. RDF can be defined such that it can refer to all resources as URI. This format is machine readable and shares the same data between different databases and Web applications. It is called Linked Data and combines the data generated by RDF as metadata. Moreover, the knowledge of semantics using RDF and OWL has been used to integrate the Linked Data to the WWW as the Semantic Web.

The following examples have been previously published in the field of computational biology. Belleau et al. made conversions from a major bioscience database to the integrated database Bio2RDF [24]. Chen et al. integrated information for drug development into the Linked Data of Chem2Bio2RDF [25] using public pharmaceutical databases. The developers of DBCLS created many Web applications that used Semantic Web technology, such as TogoWS [26], Semantic TogoDB [27], TogoGenome/TogoStanza [28], TogoAnnotation[29], and Allie [30].

The main theme of the international developer conference BioHackathon 2010 was Semantic Web [8]. In that workshop, international researchers discussed, developed, and consolidated their mutual opinions. At that time, they identified areas commonly accepted

and developed settings for the integration of bioscientific information. These results were called the “Odaiba Manifest” and were aimed at the international community of biocuration. The theme of BioHackathon 2011 in Kyoto was “Generation and utilization of Linked Data.” The theme of BioHackathon 2012 in Toyama was “Biomedical applications based on the Semantic Web technologies,” while the theme of BioHackathon 2013 in Tokyo was “Semantic interoperability and standardization of bioinformatics data and Web services.” The results of these workshops have been reported as a prepublication paper.

The formation of a research community in an international consortium and each region represented by advanced medical science research into the genome project is important. A researcher registers the research results into databanks, such as information regarding the arrangement of a genome, a library of protein conformation, and a chemical compound. Most of the information in these databases can be freely used as open data. Moreover, NCBI PubMed or PMC contains information that can be freely accessed [31]. At present, there is an increasing number of open-access journals to which contributions can be made. The evaluation index of related print magazines has also improved. In the informatics field, the rapid development of source code is encouraged due to public presentations and is inherited in the history of software development, i.e., it is open source. We believe that we should treat open data as well as open-source code. In Japan, the LOD Challenge began in 2011 for the technical promotion and education regarding LOD [10,11]. In recent years, there has been increased activity regarding open data activity, which has been encouraged by the Japanese government [32].

## 1.2 Purpose

Nowadays, utilizing biosemantics, e.g., Semantic Web, ontology, and LOD, even small-scale laboratories can quickly compile data. We constituted a graphical database of intermolecular interactions, focusing on the ligands of an enzyme (tyrosine kinase), carbohydrates (sugar), and lipids, which were subdivided using biosemantics. Moreover, we calculated the distances between amino acid residues and ligands in the Protein Data Bank (PDB) at the atomic level using octree [33], and modeled the result in RDF. We called this dataset Protein–Ligand Binding Site Pair (PLBSP). We later loaded PLBSP LOD during a triple store. Ontology-based machine-learning approaches have become effective for bioinformatics in the era of data-driven science. Machine learning is widely used for various prediction systems in bioinformatics. Among these, protein–protein and protein–ligand

interaction predictions (prediction of interactome) are widely used in drug design and green-biology research. However, a minute dataset is needed to train the predictor for machine-learning systems. Our system is aimed at the rapid development of prediction systems using the machine-learning approach. In addition, we have provided images that are useful for application development, and which are open to the public with an open license. We named this service UTProt-Image. We unified LOD, the predictor's pipeline, and image library for interactomics, and stacked them into the portal site **UTProt** (**U**niversity of **T**okyo **Prot**eins, http://utprot.net). Although UTProt currently has many developer-oriented contents, its main objective is the development of end user-oriented applications.

# Chapter 2 MATERIALS AND METHODS

Chapter 2 describes the materials and methods. This study introduces a method of utilizing RDF [34] and OWL [35] as datasets for the development of machine-learning predictors of interactomics. Moreover, using SPARQL (SPARQL Protocol and RDF Query Language) [36,37] we explain the process of implementing interactomics LOD in a graph database. RDF, OWL, SPARQL, and LOD are core techniques of the Semantic Web. In this research, we use RDF libraries such as Ruby [38], Python [39], and Perl [40] for RDF serialization of large-sized datasets. For medium-sized datasets, database schemes and data serialization are required from non-RDF data to RDF by Open Refine (former Google Refine) [41–43] and RDF Refine (a Google Refine extension for exporting RDF) [44–46]. For conversion to the Redland Raptor RDF Syntax Library [47] format, rdf2rdf [48] and ConvRDF [49] are useful.

These include proteins that interact with tyrosine kinase, part of the proteins that interact with sugar (carbohydrate) molecules, and identify a protein chain mapped by major protein databases such as UniProt [19,50] and PDB [51–55]. We use the RDF data of PDB, PDB-Ligand [56], and UniProt for the creation of interactome LOD. We add the RDFized dataset of European Bioinformatics Institute (EBI) Structure Integration with Function, Taxonomy and Sequence (SIFTS) [57]. Furthermore, we RDFize the interatomic distance between protein and ligands in PDB calculated using octree in Saad's research and other previous studies [33]. The total number of triples is 30 billion. Finally, we design three RDF schema models and enable access using AllegroGraph 4.11 [58] and Virtuoso 7 [59]. To operate RDF of the Giga triples class in these databases, we require a PC with at least 500 GB of main memory. We made the abovementioned LOD the backend database. We collect SPARQL endpoints on a single website on the WWW and wrote a document for the SPARQL query and provide services like those offered by the EBI RDF Platform [21]. It uses these endpoints and develops some interaction predictive tools between proteins and ligands in the short term.

This is the research position and development assistance of machine learning using interactome LOD. Users can use the developed predictive tool as a module of the workflow of a Web application. We use Galaxy for the Web framework of this workflow [60]. Moreover, we need to prepare image content beforehand for application development, and do not require an open-license image library. We propose a management method such that experimental data or report documents produced daily at each laboratory are not lost. This idea is also useful when the person is in change of personnel. We tackle the creation of the

appropriate structure for data accumulation. The data provider licenses the data and needs to enable it to refer to a recyclable form. However, it is necessary to control the timing of its release with regards to patents or papers made available to the scientific community. Therefore, there is a need for a data portal in a closed environment. The CKAN framework enables smooth system design since it is useful for closed data and open data [61]. This system has an option that enables the direct upload from the internal database of a laboratory to an international repository. We believe that the data portal is effective for use when there is a requirement for the data to be presented to the public. We report the development of the service for circulating these data and progress of employment.

## 2.1 How to Make a Domain Specific LOD?

### 2.1.1 Resource Description Framework (RDF)

RDF is a format used to describe the metadata on the Web and was recommended by W3C in 2004 [4]. The relation between the subject and object is connected by the predicate using the resources of the metadata with URI. RDF applies the subject and the object to the node, and the predicate is applied to the edge. Therefore, it can be expressed using the graph. The class of two nodes and one edge is called a triple as one unit. Moreover, a graph database is constructed by loading an RDF graph into the RDF store. The spreadsheets can convert the graphs. The rows of the table correspond to the subjects, the columns of the table correspond to the predicates, and the cells of the table correspond to the objects. **Figure 1** is a concept that represents the transformation to RDF data.

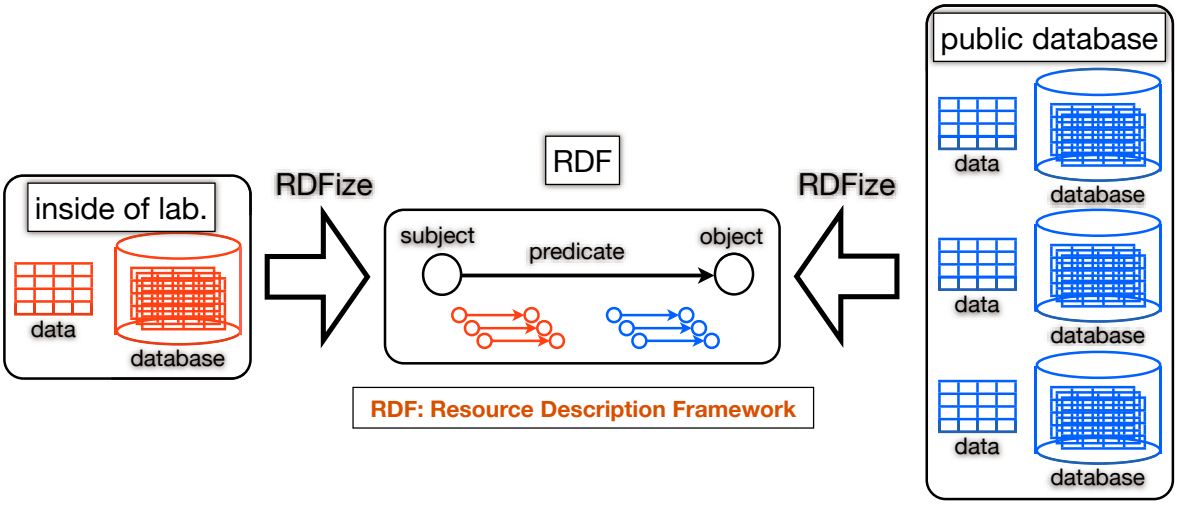


**Figure 1** RDF represents that which sets a subject and an object to nodes, and sets a predicate to an edge. It has the feature whereby a resource can refer to it using URI. By RDFizing data and the database, we can merge and perform repurposing of data between different domains. The information inside an organization is in orange, the external public database is in blue.

In addition, typical name spaces such as RDF Schema (RDFS) by W3C [34], Friend Of A Friend (FOAF) by FOAF Project [62], and Dublin Core (DC) have been recommended by The Dublin Core Metadata Initiative [63]. Permanent sharing of metadata on the Web is enabled by incorporating these into the data modeling like a standard library of the program. The OWL is one of the enhancing vocabularies of RDF for the ontology recommended by W3C in 2004 [35]. Modeling can be done on the relation of the class as well as the definition of the homonym and synonym, among others.

In this study, we made RDFs for the life science domain using UniProt [19], PDB [14], PDB-Ligand [56], and Bio2RDF. We converted the data of EBI SIFTS into RDF using the original RDF schema as non-RDF data. We used convertible tools such as Link Data (at light size) [64,65], Open Refine (at middle size) [41,46,66], and the RDF library of a script language such as Ruby/Python/Perl (at heavy size) to make RDF [38–40]. RDF has several formats such as N-Triples [67], RDF/XML [68], and Turtle [69]. It may also be useful to use the Redland Raptor RDF Syntax Library developed by Dave Backett to change these serializations [47].

### 2.1.2 Web Ontology Language (OWL)

Ontology is the dictionary of common vocabularies in a domain. Its role is to identify the same terms that exist in different domains. Various concepts utilize OWL for machine-readable purposes in Semantic Web technology. A namespace describes the data relation using RDF and OWL. For complex networks, it is difficult for an engineer to edit a script while imaging a model. In this case, the ontology editor has to attempt to visualize the conceptual modeling. The ontology editor often used in the life science field is Protégé [70], which was also used in our research. **Figure 2** shows the tree structure of a class in UniProt Core Ontology, which was obtained using Protégé. Each class is connected by property edges.

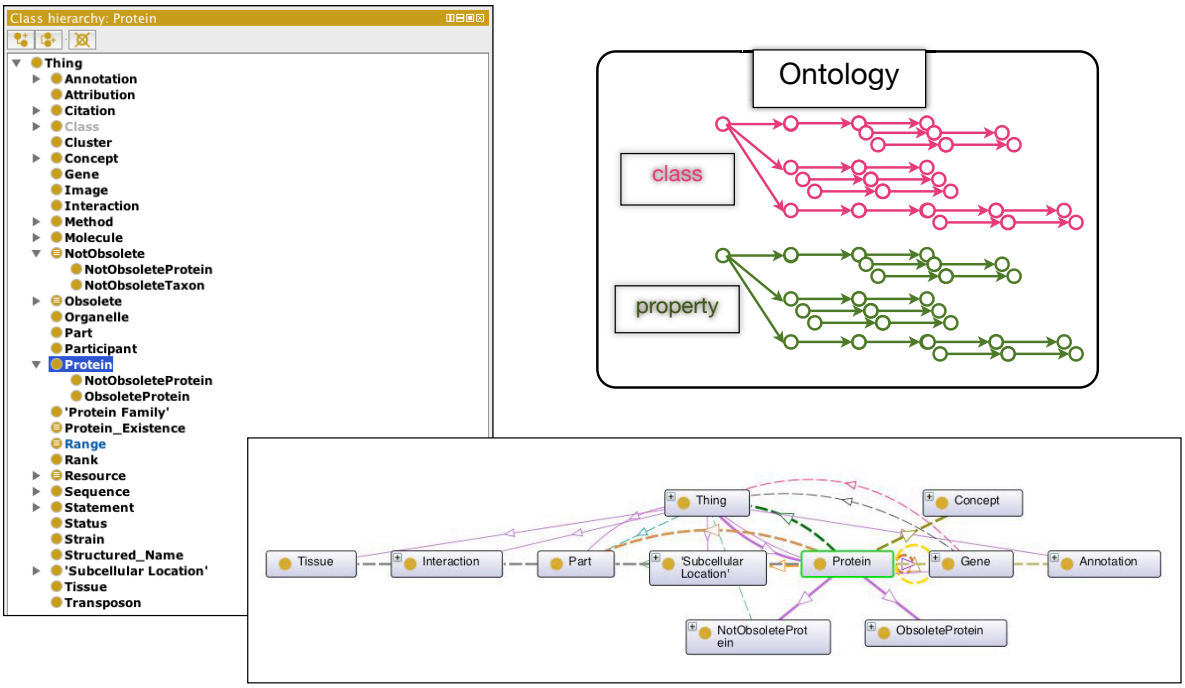


**Figure 2** Combination of a class and property expressed by a tree structure, which constitutes a network topology as ontology using OWL. This figure illustrates the owl:Thing's subclass of the UniProt core ontology using Protégé.

Here we introduce the ontology used in this research. We will utilize UniProt (Uniprot Core Ontology) [71], PDBo (PDB Ontology) [14,72], EDAM (EDAM Ontology of Bioinformatics and Data Formats) [73], FALDO (Feature Annotation Location Description Ontology) [74], DDI (Ontology for Drug Discovery Investigations) [75], and SIO (The Semanticscience Integrated Ontology) [76], among others, from the Website, NCBO BioPortal [77] or literature; we will model the novel RDF schema for the interactome LOD.

### 2.1.3 SPARQL Protocol and RDF Query Language (SPARQL)

SPARQL Protocol and RDF Query Language (SPARQL) was recommended by W3C as a query language of the RDF graph database in 2008 [36]. It specifies the name space in the RDF graph and operates it by giving a query corresponding to the subject, predicate, and object. This operation language can execute the edit, retrieval, merging, and inference by a unit of the graph or triples. SPARQL 1.1 was recommended by W3C in 2013 [37]. It executes a federated query between different endpoints using SERVICE functions. **Figure 3** is a SPARQL endpoint of AllegroGraph 4.11 [58].

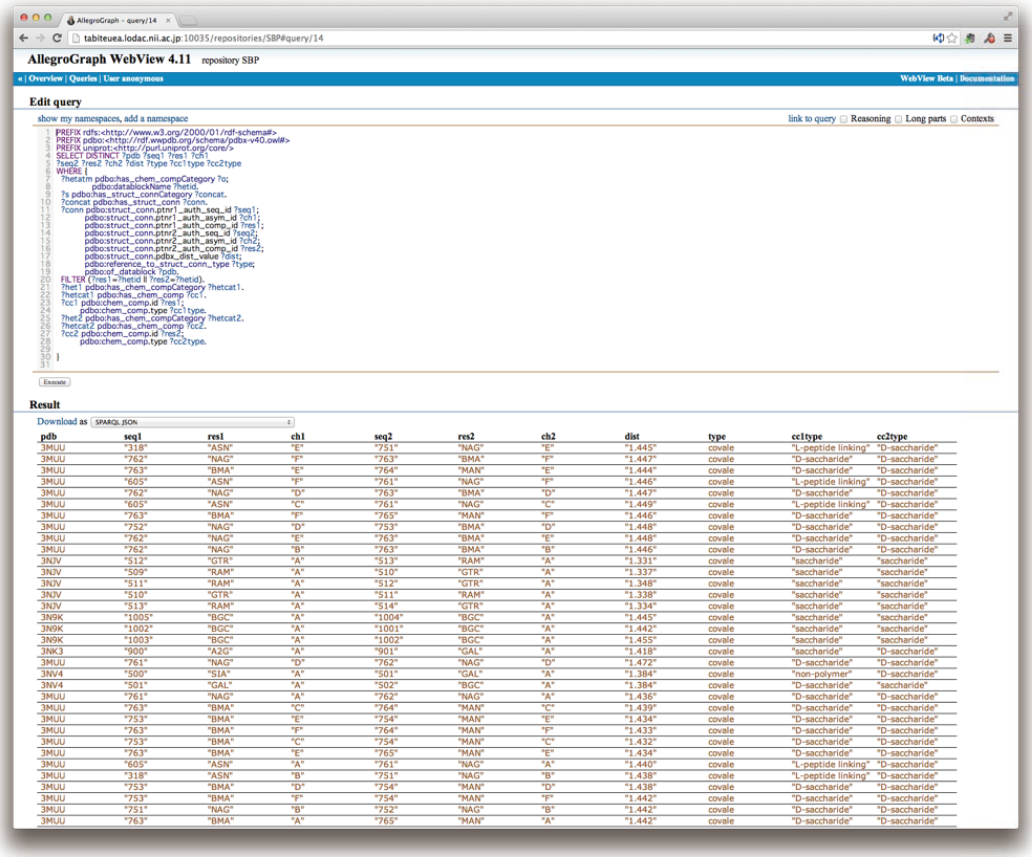


**Figure 3** This picture is an example of the use of the endpoint of AllegroGraph 4.11, which is used on the UTProt RDF platform. It is helpful for carefully considering the concept of programming to return a simple result in HTML.

### 2.1.4 Linked Open Data (LOD)

Linked Data is a database project connecting each RDF graph using hyperlinks. In 2007, Auer et al. developed the partial structured data (metadata) of Wikipedia to using Linked Data and named it DBpedia [78]. Information science, life sciences, social network multimedia, and government statistics, among others, were all merged with a possible cross-reference by making it a starting point and using RDF. **Figure 4** shows the latest Linked Open Data Cloud as of November 2011 [79]. The links became large, and the complete figure cannot be drawn. The pink domain represents the life science LOD. In science and technology fields other than informatics, the use of Linked Data is very advanced. It is also called the Web of data and was developed by Tim Berners Lee.

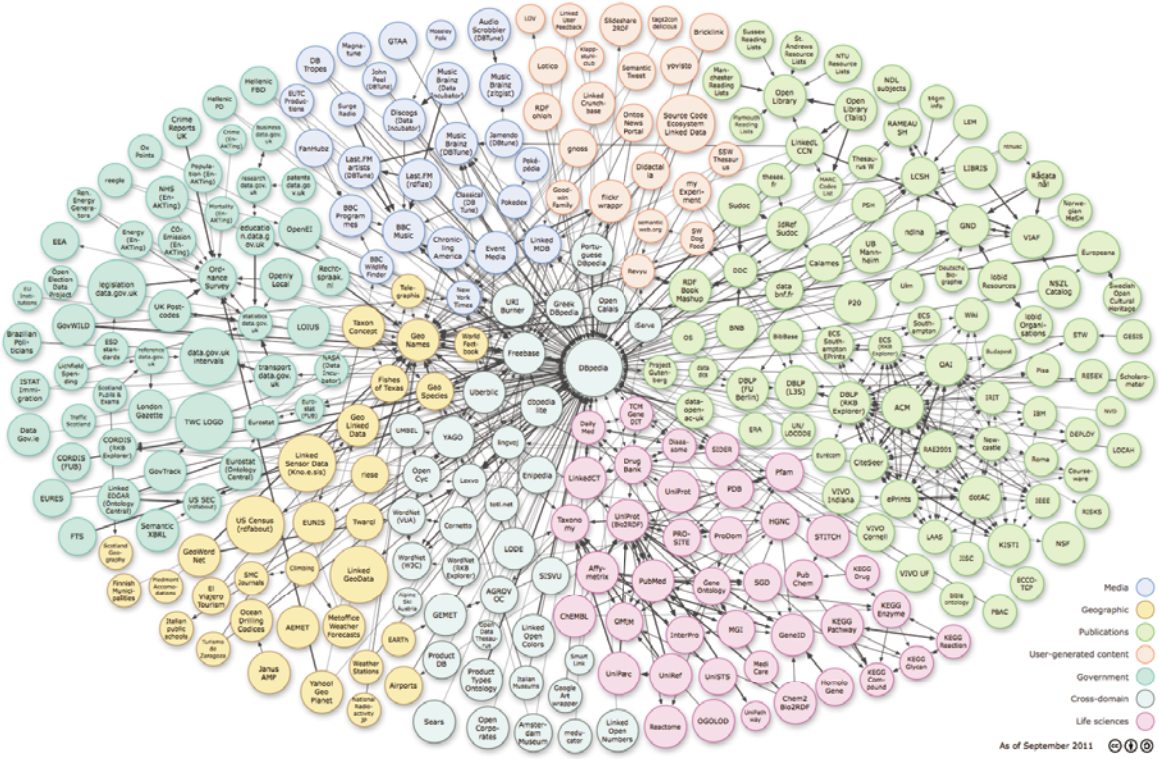


**Figure 4** Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/. In this figure, the pink zone is the LOD of the life science domain. This diagram's license is Creative Commons Attribution-ShareAlike.

Semantic Web acquires metadata on WWW and aims at data sharing between different applications. The abovementioned RDF and OWL are the base technology, and Linked Data aims to realize further achievements. The goal of the developer is that "All data should be able to be referred to open on the Web." Governments have been positively promoting the data archives in Europe and North America, and as a first step, government statistics, library-based book collections, scientific research data, and social nets have been put together for open access to the public. However, there is a problem regarding data that cannot be made available globally, such as papers and publications, patent filings, copyright, and privacy protection. We should also note the case where the producer of data does not desire to make the data public; the provider should be able to manually control the extent of the data's use, and the user can bypass the application for permission to use the data.

**Figure 5** is a flowchart showing database integration using Semantic Web and the LOD technology flowchart. This method combines internal information, public databases, application development, and data analysis. It should be expected that the application will propose various areas.

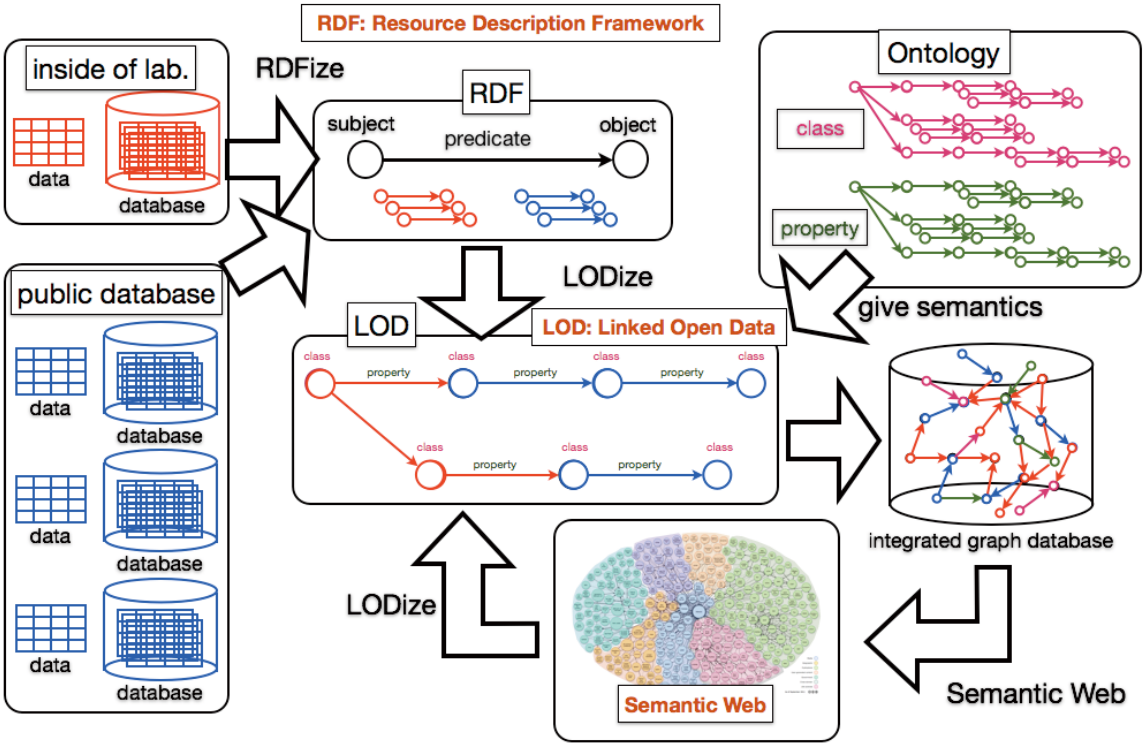


**Figure 5** Flowchart showing the practical use of Semantic Web in Biosemantics. We modified the internal data experimental results and the public databases for analysis into RDF and merged them. We employed a common vocabulary in ontology by OWL. We load serialized LOD into RDF store and open to the exterior. In WWW, an endpoint and a resource serve as Semantic Web.

## 2.2 Method for Controlling Large-scale LOD in RDF Store

The operation of over 10 Gtriples RDF such as PDB and UniProt requires a high-performance computing. For example, the main memory of the back-end server of the UTProt RDF Platform is 379 GB. A developer can convert data from comma spread value (CSV) and tab spread value (TSV), among others, to the RDF format. He can also convert it from one RDF format to another RDF format, such as N-Triples to RDF/XML and N-Triples to Turtle. Because one graph is located in a line with one line, it is easy to execute the script processing for N-Triples. However, since N-Triples are redundant, its file size is large. RDF/XML has many applications corresponding to the data structure of XML. Also, the file size of RDF/XML is less than that of N-Triples. Turtle is expressed as a triple with a brief sketch; it has a small file size and display letter counts. Turtle is also fit for printing in space. Large storage and high-speed CPUs are also necessary for editing and processing large amounts of data records that were loaded into the database. If huge master data is divided for load sharing the amount of storage consumption will double. Virtuoso 6 is an RDF database

(triple store) that was developed around 2011. Even with a PC that satisfied the requirements, in 2011, the time taken to load Virtuoso 6 from UniProt took at least one month. In 2011, Virtuoso 7 or AllegroGraph 4.11 was improved to allow UniProt to be loaded within several hours. However, the UniProt RDF platform grew from approximately 3 Gtriples to 10 Gtriples between 2011 and 2013. For RDF format conversion that exceeds 1 Gtriples, it may lead to a syntax error caused by a memory overflow with rapper (Redland Raptor RDF Syntax Library). convRDF is a transformation tool for large RDF such as UniProt at 2013. Moreover, there is a conversion Java tool called rdf2rdf, which also includes file compression. The compressed RDF file can also be loaded in Virtuoso 7 or AllegroGraph 4.11. A developer can select the RDF format and file compression condition from a self data repository. In the very efficient database mentioned above, a Web application for administrators is included. However, loading large data to a database via the Web is accompanied by the risk of server failure. It may be better to communicate with a server that had previously compressed RDF data. The data has to be loaded from a local server to a database. The methods used to monitor whether the loading is proceeding normally differs for different database systems. The loading can be monitored by an SQL function by typing isql at the command line in Virtuoso 7, and using the client of a command line in AllegroGraph 4.11 outputs a loaded log.

## 2.3 Database Integration and Tool Development for Interactome with Biosemantics

In this section, we explain the applied LOD for biosemantics. First, we develop the integrated database of the interactome LOD. Second, we develop Web applications with the interactome LOD. Third, we develop the image library with an open license as the database materials for human readability. Then, we develop a working support tool for researchers. Finally, we integrate these products into the portal site.

### 2.3.1 Interactome LOD

UTProt LOD was provided mainly for data-set creation, for machine learning of interactomics, and application development. In addition to major public databases (UniProt and PDB), this LOD comprises four graph databases: PLBSP, Tyrosine Kinase Interaction Pair (TKIP), Sugar-Binding Protein (SBP), and Structure Integration with Function, Taxonomy and Sequence, original data source by EBI (RDF-SIFTS). In this LOD, existing RDF resources were transferred to the new RDF schema model to guarantee the product

performance. Prior to RDF, we performed additional data compilation. We loaded approximately 30 billion triples into the graph database. A developer can operate RDF by loading graph databases from SPARQL endpoints. We also introduced detailed sample queries on the website. The graph databases currently used are AllegroGraph 4.11 and Virtuoso 7. These days, the target users of this database are bioinformaticians and database developers.

### 2.3.2 Predictor Tools and Pipeline for Interactome

Ontology-based machine-learning approaches can expedite predictor development for interactomics. We developed an instrument for binding prediction and the prediction of binding sites (protein residues that interact with ligands). UTProt LOD is used to collect the dataset for predictors with high precision. These predictors are exhibited on UTProt as stand-alone programs. Furthermore, we implemented parts of the workflow on the UTProt Galaxy pipeline for end users, and provided an operations manual. As an example of eScience (data-driven science), which uses Semantic Web technologies, we proposed a machine-learning tool for intermolecular interaction.

### 2.3.3 Open-License Image Library

A data bank of 2D/3D graphics would be useful for the development of an application that uses LOD. It is desirable to have mages that are continually provided by a particular URI. Furthermore, permission should be given for widespread public distribution, enabling it to be freely used for scientific graphics. In this manner, the databank data is recyclable by Creative Commons License Attribute (CC BY) [80]. The UTProt-Image is an image data bank for which an open license for interactomics was applied. We are currently developing this system based on Coppermine Photo Gallery 1.5.2.4 [81].

### 2.3.4 Data Portal for Closed Data with CKAN

Researchers are required to provide some documentation, such as a research document, program, or data, which can be used by other users. However, it is necessary to properly plan the timing with which a patent or paper is published. We carefully consider the placement of a data archive in a closed environment using the open-source data portal wear CKAN 1.8 [61].

## 2.4 List of Computer Specifications Used in this Study

This section describes the specification of each computer used in this experiment. The database server of the UTProt RDF Platform is an actual machine at the National Institute of Informatics (NII). The Web application server of UTProt (UTProt Galaxy, UTProt Image, BILAB Data Portal) was in the Cloud. The Cloud computing environment uses Amazon Web Service (AWS) Elastic Compute Cloud (EC2) [82].

Product Name: UTProt RDF Platform

CPU: Intel Xeon E5-2609 2.40GHz (6.40GT L2 10M/80W/TB SandyBridge-EP) 8Core (4Core x 2CPU)
Main Memory: ECC Registered DDR3-1333 Quad-Channel 384GB (16 GBx24)
Storage: HDD S-ATAII 2TBx3 (6TB)
Network Card: Dual GbE(intel i350) & Dual 10GBase-T(intel X540)

OS: Debian squeeze 6.0.7 (64bit)
Database: AllegroGraph 4.11 (Developer Version) and Virtuoso 7 (Open Source Version)

Product Name: UTProt & UTProt Galaxy

Cloud Environment: Amazon Web Service EC2
CPU (Instance Type): 8 ECU (Intel(R) Xeon(R) CPU E5645 @ 2.40GHz), 4vCPU, (m1.xlarge)
Main Memory: 15GB
Storage: HDD 8GB x 1, HDD 50GB x1
Network Performance: Middle (Dependent on an instance type)

OS: Ubuntu 12.04.3 LTS (GNU/Linux 3.2.0-40-virtual x86_64)
Framework: Concrete 5.6.1.2 and Galaxy (release_2013.02.08)

Product Name: UTProt Image

Cloud Environment: Amazon Web Service EC2
CPU (Instance Type): 6.5 ECU (Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz), 2vCPU, (m2.xlarge)
Main Memory: 17GB
Storage: HDD 32GB x 1, HDD 404GB x1, HDD 1TB x 1
Network Performance: Middle (Dependent on an instance type)

OS: Ubuntu 12.04.2 LTS (GNU/Linux 3.2.0-40-virtual x86_64)
Framework: Coppermine Photo Gallery 1.5.2.4

Product Name: BILAB Data Portal

Cloud Environment: Amazon Web Service EC2
CPU (Instance Type): Variable ECU (Intel(R) Xeon(R) CPU E5430 @ 2.66GHz), 1vCPU, (t1.micro)
Main Memory: 635MB
Storage: HDD 8GB x 1
Network Performance: Very Low (Dependent on an instance type)

OS: Ubuntu 12.04.3 LTS (Ubuntu 10.04.4 LTS (GNU/Linux 2.6.32-350-ec2)
Framework: CKAN 1.8

# Chapter 3 RESULTS AND DISCUSSIONS

We developed an integrated graph database for intermolecular interaction, the UTProt (University of Tokyo Protein) RDF Platform. It is a core database of PLBSP, which was created from PDB, PDB-Ligand, RDF-SIFTS from UniProt, and EBI SIFTS. In particular, RDF-SIFTS is an aggregate of seven small LODs that were constructed within a short time during BioHackathon 2013. It was developed based on the purpose-oriented subset LOD from the core databases such as SBP and PLBSP residue, which were developed by a bioinformatician. As of December 2013, the total number of UTProt triples was 30 billion. The global image of the UTProt database is shown in **Figure 6.** In the UTProt Project, the interactome LOD includes the public databases (blue) and original databases (green). Each sphere represents a database. The arrows show the links that exist between each database. To search LOD, the exclusive query language SPARQL is required. Next, we give search examples to the integrated database using SPARQL. We combined the SPARQL endpoint of the public database of PLBSP, RDF-SIFTS, and the exterior, and devised approximately 80 combined queries.

These databases supported the creation of the part predictive tool of the sugar interacting protein. The developed program is modularized in the pipeline. The pipeline was exhibited as the workflow system UTProt Galaxy, which can be used on the Web. Next, in the UTProt image, we provide an open license to the graphics, which are needed when developing applications. The image of the protein are registered for approximately 480,000 affairs and ligands, of which 400,000 are the 2D graphics of the protein–ligand binding site that execute parallel computing powered by Ligplot [83] using high-performance computing (HPC). We examined the requirement to maintain the above knowledge circulation. We propose the novel strategy that structures the data stored by a researcher daily, after which it is entered into the database. Using CKAN, we first store closed data; thus, we can specify the timing at our own discretion and can shift to open data. After applying this in a bioinformatics laboratory to the Bioinformation Engineering Laboratory (BILAB) Data Portal for around two years, a dataset comprising 33 affairs was accumulated. Because metadata and an open license are added to these datasets, we can open them to the public in future. Moreover, the search for a new researcher was required to continue the work done by the staff and student who had relocated. The portal site that merged the tools and databases for the interactome of this series is UTProt (http://utprot.net).

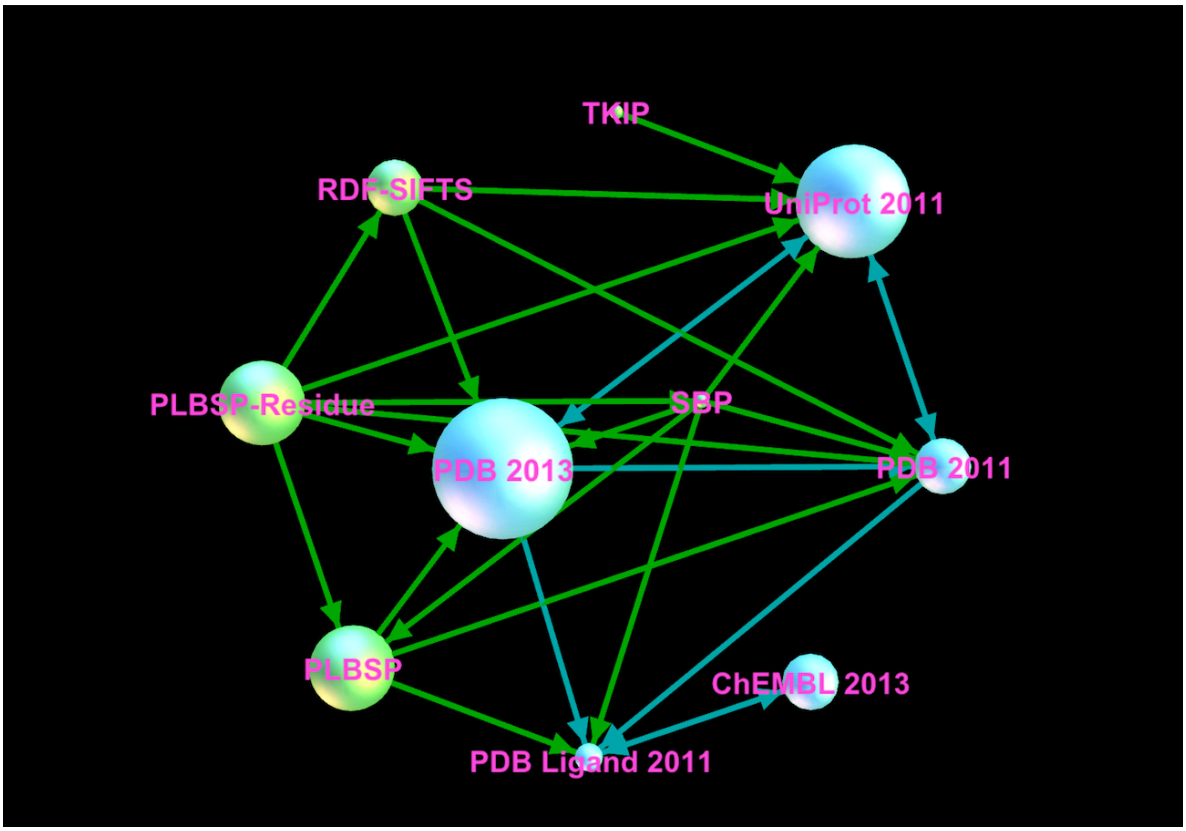


**Figure 6** In the UTProt Project, the interactome LOD includes the public databases (blue) and the original databases (green). Each sphere represents a different database. The arrows show the links between the different databases.

## 3.1 Summary of Integrated Graph Database for Interactomes

### 3.1.1 UTProt RDF Platform

The UTProt RDF Platform is an aggregate of the SPARQL endpoint for intermolecular interaction research. **Figure 7** describes the accessible SPARQL endpoints from the UTProt RDF platform. This product resembles the EBI RDF Platform. PLBSP has the interatomic distance of the protein–ligand in PDB calculated by previous research. According to Masaki Banno, we packed a sugar-interacting protein into the SBP database with respect to both sides of sequence and structure. According to Masayuki Yarimizu, TKIP is a protein-pair database comprising the receptor tyrosine kinase and a substrate ligand for protein–protein interaction. **Table 1** shows statistics of these RDF databases and the public databases used in this study. The number of entries (triples) exceeded 30 billion. This section explains two subprojects, PLBSP and RDF-SIFTS, within the UTProt project. Then, the PLBSP-Residue, which is a lightweight version of PLBSP, was developed in December 2013 for the UTProt Galaxy pipeline.

wwPDB/RDF does not have atomic coordinates distributed by PDB. When the RDF of PDB in which readers include atomic coordinates is required, it is necessary to acquire and change XMLST for a PDBML file and wwPDB/RDF. We created house data of PDB/RDF containing all atoms from 94,335 PDBML files on November 2013. We succeeded in loading it to Virtuoso 7 in spite of the fact that these files exceeded 25 Gtriples.

In 2011, the UTProt RDF Platform used the data of the integrated UniProt database of amino acid sequences. In 2013, we built a mirror of the ChEMBL data, which is the biological activity of a low molecular compound. These RDFs were loaded to AllegroGraph 4.11.

**Table 1** RDF statistics of the secondary database that was created by UTProt, and the public database used by the backend.

| UTProt database name | triples [#] |
|---|---|
| PLBSP: Protein–Ligand Bindig Site Pair (with atom coordinate) | 1,337,648,853 |
| PLBSP2PDB | 52,582,582 |
| TKIP: Tyrosine Kinase Interaction Pair | 440,806 |
| SBP: Sugar Binding Pair | 20,202,647 |
| RDF-SIFTS version 2.0 | 10,527,560 |
| PLBSP Residue | 96,532,835 |
| **name of inclusive public database** | **triples [#]** |
| UniProt 2011 | 3,104,363,044 |
| PDB 2013 (with atom coordinate) | 25,259,162,986 |
| PDB 2011 (no atom coordinate) | 612,736,982 |
| PDB Ligand 2011 | 28,040,694 |
| ChEMBL 2013 | 137,497,642 |
| Total | 30,659,736,631 |

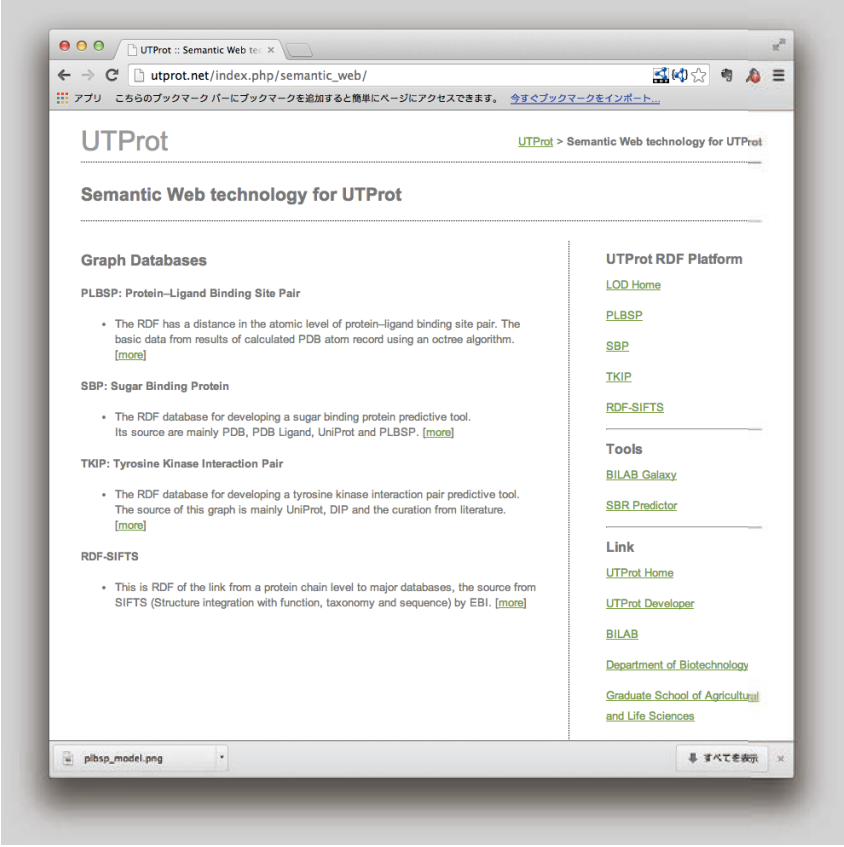


**Figure 7** LOD of PLBSP or RDF-SIFTS was loaded to the RDF store as a graph. A user can access those endpoints from the UTProt RDF platform.

### 3.1.2 Protein–Ligand Binding Site Pair (PLBSP) Database

PLBSP added and reconstructed the interatomic distances among ligand-binding residues using the 2011 edition of PDB and the data of the PDB-Ligand and PDB Ontology. The calculation of the interatomic interaction was as a result of inquiries made by Gul Saad in 2011. The serialization program of PLBSP was a Perl script that was implemented by Toyofumi Fujiwara in DBCLS together with the authors at the BioHackathon 2012 workshop held in Toyama, Japan. **Figure 8** shows the RDF schema of PLBSP. In this schema, we mainly use the PDB ontology. The URI of the name space is http://rdf.wwpdb.org/schema/pdbx-v40.owl, and the prefix is pdbo. We inputted the square of the interatomic distance, which has the property ddi:distance_to_atom of DDI. The URI of the name space of DDI is http://purl.bioontology.org/ontology/DDI, and the prefix is ddi. BILAB ontology was uniquely designed using the vocabulary of the existing ontology when the range and domain of the property were not in agreement. We defined the URI of the name space as http://www.bi.a.u-tokyo.ac.jp/ontology/owl, and defined the prefix is bilab. Others used the vocabulary of RDF http://www.w3.org/1999/02/22-rdf-syntax-ns#, RDFS http://www.w3.org/2000/01/rdf-schema#, and OWL http://www.w3.org/2002/07/owl#. We devised the design such that both the state of the existing interaction and its pair could be described between two atoms or two amino acid residues.

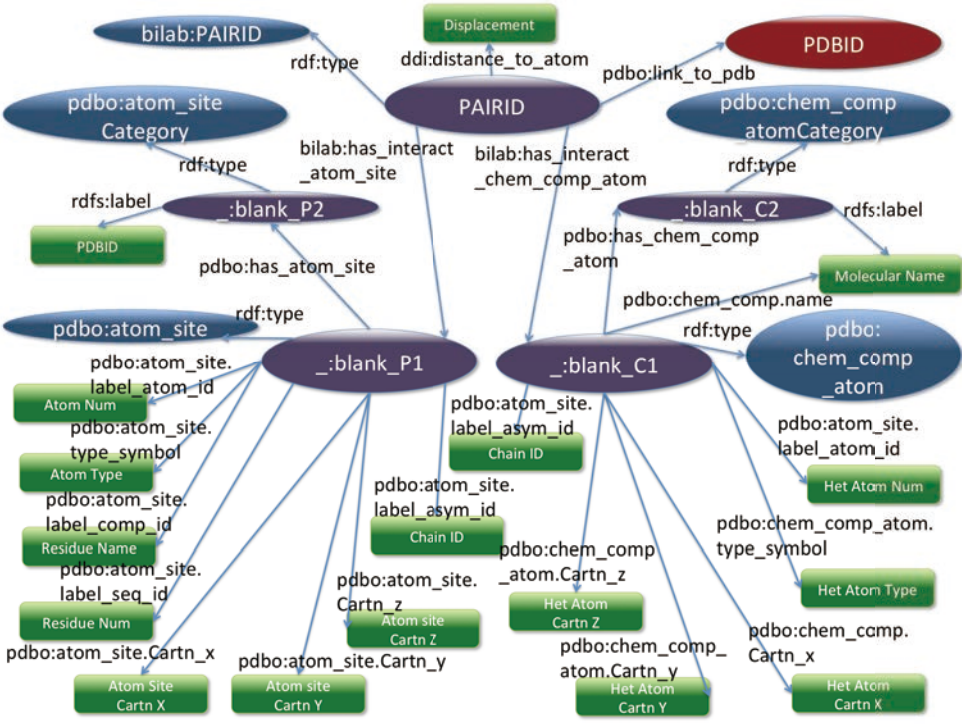


**Figure 8** RDF schema of PLBSP. It mainly uses PDB ontology. The blue node indicates the type class, the blank purple node indicates the resource ID, the literals are indicated by the green node, and the red node indicates the exterior LOD. The arrow shows the property.

### 3.1.3 RDF-SIFTS

Detailed cross-references such as chain-level information between UniProt and PDB require EBI SIFTS to present the dataset. We converted from SIFTS into the RDF format, and established a link between the smooth applications. SIFTS is a project that was used to map the concrete numerical value from the confirmed protein information on the chain level by EBI to the database of a function taxonomy and sequence. In SIFTS, they perform manual curation. We applied the precedence research to LOD by RDF-SIFTS.

**Figure 9** shows the schema of RDF-SIFTS. PDB, UniProt, EDAM, and FALDO, which are ontologies utilized for this schema design. The URI of the namespace of UniProt Core is http://www.uniprot.org/core/, and the prefix is uniprot. EDAM is an ontology of bioinformatics techniques proposed by Ison et al. The URI of the namespace is http://purl.bioontology.org/ontology/EDAM/, and the prefix is edam. The URI of ontology was used by FALDO to describe the detailed information regarding the nucleic acid and amino acid sequence discovered by Bolleman et al., and its namespace is http://biohackathon.org/resource/faldo, and the prefix is faldo. RDF-SIFTS was developed by the BILAB UTProt group during BioHackathon 2013, which was held in Tokyo.

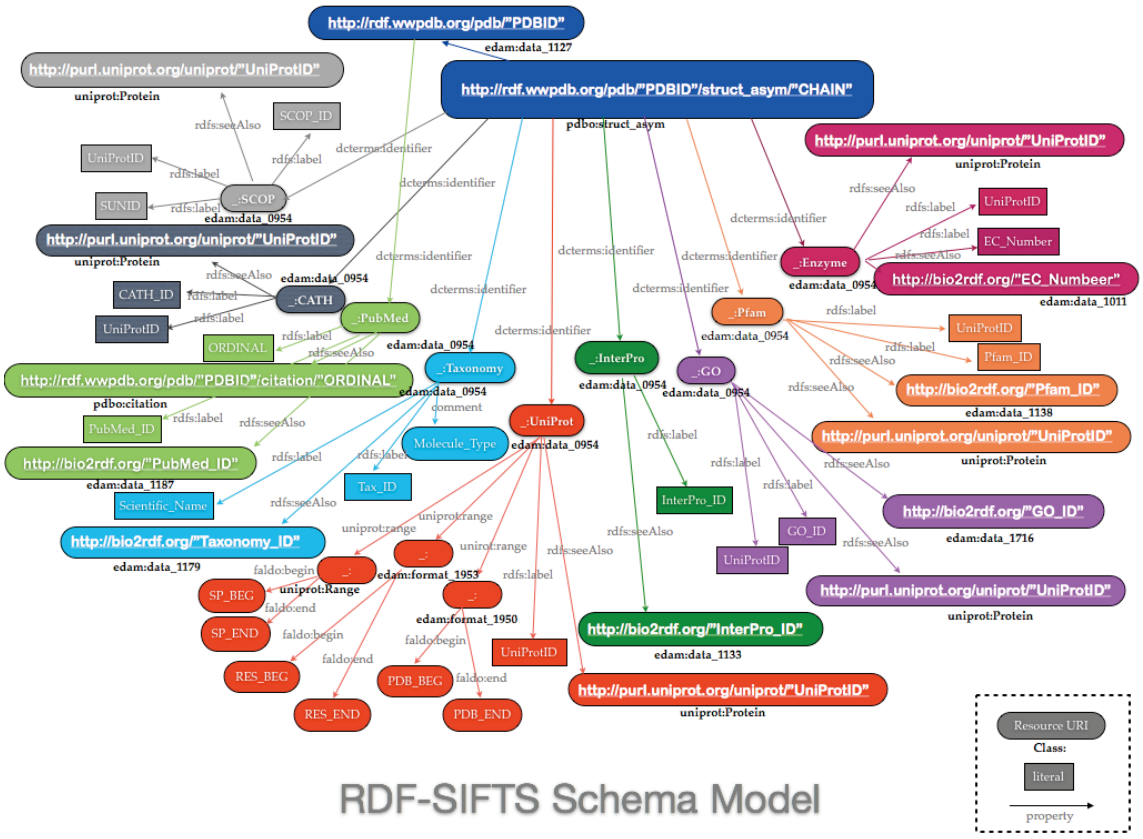


**Figure 9** RDF schema of RDF SIFTS. The root of the PDB chain URI links to nine subset graphs (UniProt, Taxonomy, Enzyme, PubMed, GO, SCOP, CATH, Pfam, and InterPro). The source data is EBI SIFTS.

### 3.2 Knowledge Discovery Brought about by Database Integration

In this section, we describe a method of knowledge discovery using the UTProt RDF platform. We used the query language SPARQL1.1 as the inquiry language of the database.

#### 3.2.1 Use Case of SPARQL Query for Proteome and Interactome

In this section, we discuss Code 1 to Code 3 based on the SPARQL endpoint of PLBSP_residue, which is the lightweight version of PLBSP. The URI of the PLBSP_residue repository of AllegroGraph 4.11 has the port number 10035 on the tabiteuea server in NII.

```
http://tabiteuea.lodac.nii.ac.jp:10035/repositories/PLBSP_residue#query/
```

**Code 1** is a query for web application UTProt Galaxy made using the UTProt RDF platform. The key sentence is “Select UniProt accession number of mannose-binding protein and the list of binding residue.” The result of this query enumerates UniProt AC (accession number) and the amino acid sequence numbers in PLBSP-Residue. **Data 1** is a header of the data that expressed the inquiry result of Code 1 using the Turtle format (.ttl). In this paper, although omitted, Data 1 returns the result of 8,616 affairs.

**Code 1** Select UniProt accession number of mannose-binding protein and the list of binding residue.

```
PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>
PREFIX dcterms:<http://purl.org/dc/terms/>
PREFIX edam:<http://edamontology.org/>
PREFIX sio:<http://semanticscience.org/resource/>

SELECT DISTINCT ?uniprot ?sqno
WHERE {
   ?het_res pdbo:atom_site.chem_comp.id "MAN".
   ?het_asym dcterms:isPartOf ?het_res;
              rdf:type sio:SIO_010432;
              rdfs:seeAlso ?pdb_res.
   ?pdb_res rdf:type edam:data_1756;
            rdfs:seeAlso ?unpres.
   ?unpres rdf:type edam:data_1756.
   ?unpres dcterms:isPartOf ?uniprot.
   ?unpres rdfs:label ?sqno.

} ORDER BY ?uniprot ?sqno
```

**Data 1** This is a header of the data indicating the inquiry result of Code 1 with Turtle (.ttl). In this paper, although omitted, Data 1 returns the result of 8,616 affairs.

```
@prefix dc:         <http://purl.org/dc/elements/1.1/> .
@prefix rs:         <http://www.w3.org/2001/sw/DataAccess/tests/result-set#> .
@prefix x:          <http://example.org/ns#> .
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

[]      rdf:type        rs:ResultSet;
    rs:resultVariable "uniprot";
    rs:resultVariable "sqno";
   rs:solution [
     rs:index 1 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/O61363> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "2551"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];
   rs:solution [
     rs:index 2 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/O61363> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "2553"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];
   rs:solution [
     rs:index 3 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/O61363> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "2555"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];

                                          :
                                          :
```

The feature of **Code 2** is to execute a database search of the external SPARQL endpoint from the PLBSP endpoint using the SERVICE function. In this case, the key sentence is "Select UniProt accession number and the amino acid sequence of mannose-bound protein." When using a SERVICE function, the user needs to understand the load of the server. In this example, the query has restricted the inquiry to the endpoint of UniProt LIMIT 100,000. It is common to acquire search results combining the OFFSET function and the LIMIT function in a database and shifting a start line. SPARQL also is not an exception. The results are shown in **Data 2**. We acquired an amino acid sequence of the mannose-bond protein with 24 affairs.

**Code 2** Select UniProt accession number and the amino acid sequence of mannose-bound protein. The query was for both PLBSP_residue and the original UniProt using the SERVICE function. Here we set to LIMIT 100,000 based on restrictions of the SPARQL endpoint of the original UniProt.

```
PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>
PREFIX dcterms:<http://purl.org/dc/terms/>
PREFIX edam:<http://edamontology.org/>
PREFIX sio:<http://semanticscience.org/resource/>
PREFIX up:<http://purl.uniprot.org/core/>

SELECT ?uniprot ?uniparc ?seq
WHERE {
  SERVICE <http://beta.sparql.uniprot.org/> {
    SELECT ?uniprot ?uniparc ?seq
      WHERE {
        ?uniprot a up:Protein.
        ?uniparc up:sequenceFor ?uniprot;
                 rdf:value ?seq.
      } LIMIT 100000
  }
  { SELECT DISTINCT ?uniprot {
      ?het_res pdbo:atom_site.chem_comp.id "MAN".
      ?het_asym dcterms:isPartOf ?het_res;
                rdf:type sio:SIO_010432;
                rdfs:seeAlso ?pdb_res.

      ?pdb_res rdf:type edam:data_1756;
               rdfs:seeAlso ?unpres.

      ?unpres rdf:type edam:data_1756;
              dcterms:isPartOf ?uniprot.
    }
  }
}
```

**Data 2** This is a header of the data indicating the inquiry result of Code 2 with Turtle (.ttl). In this paper, although omitted, Data 2 returns the result of 24 affairs.

```
@prefix dc:         <http://purl.org/dc/elements/1.1/> .
@prefix rs:         <http://www.w3.org/2001/sw/DataAccess/tests/result-set#> .
@prefix x:          <http://example.org/ns#> .
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

[]      rdf:type       rs:ResultSet;
    rs:resultVariable "uniprot";
    rs:resultVariable "uniparc";
    rs:resultVariable "seq";
  rs:solution [
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/P13689> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniparc/UPI000012514B> ;
                  rs:variable "uniparc" ] ;
     rs:binding [ rs:value
"MAPDLSELAAAAAARGAYLAGVGVAVLLAASFLPVAESSCVRDNSLVRDISQMPQSSYGIEGLSHITVAGALN
HGMKEVEVWLQTISPGQRTPIHRHSCEEVFTVLKGKGTLLMGSSSLKYPGQPQEIPFFQNTTFSIPVNDPHQVW
NSDEHEDLQVLVIISRPPAKIFLYDDWSMPHTAAVLKFPFVWDEDCFEAAKDEL" ;
                  rs:variable "seq" ] ];
  rs:solution [
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/Q9XFX3> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniparc/UPI00000AB8E8> ;
                  rs:variable "uniparc" ] ;
     rs:binding [ rs:value
"MGTSIKANVLALFLFYLLSPTVFSVSDDGLIRIGLKKRKVDRIDQLRGRRALMEGNARKDFGFRGTVRDSGSAV
VALTNDRDTSYFGEIGIGTPPQKFTVIFDTGSSVLWVPSSKCINSKACRAHSMYESSDSSTYKENGTFGAIIYGTG
SITGFFSQDSVTIGDLVVKEQDFIEATDEADNVFLHRLFDGILGLSFQTISVPVWYNMLNQGLVKERRFSFWLNRN
VDEEEGGELVFGGLDPNHFRGDHTYVPVTYQYYWQFGIGDVLIGDKSTGFCAPGCQAFADSGTSLLSGPTAIVT
QINHAIGANGVMNQQCKTVVSRYGRDIIEMLRSKIQPDKICSHMKLCTFDGARDVSSIIESVVDKNNDKSSGGIHD
EMCTFCEMAVVWMQNEIKQSETEDNIINYANELCEHLSTSSEELQVDCNTLSSMPNVSFTIGGKKFGLTPEQYIL
KVGKGEATQCISGFTAMDATLLGPLWILGDVFMRPYHTVFDYGNLLVGFAEAA" ;
                  rs:variable "seq" ] ];
  rs:solution [
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/P69327> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniparc/UPI0000036D3C> ;
                  rs:variable "uniparc" ] ;
     rs:binding [ rs:value
"MSFRSLLALSGLVCTGLANVISKRATLDSWLSNEATVARTAILNNIGADGAWVSGADSGIVVASPSTDNPDYFY
TWTRDSGLVLKTLVDLFRNGDTSLLSTIENYISAQAIVQGISNPSGDLSSGAGLGEPKFNVDETAYTGSWGRPQR
DGPALRATAMIGFGQWLLDNGYTSTATDIVWPLVRNDLSYVAQYWNQTGYDLWEEVNGSSFFTIAVQHRALVE
GSAFATAVGSSCSWCDSQAPEILCYLQSFWTGSFILANFDSSRSGKDANTLLGSIHTFDPEAACDDSTFQPCSP
RALANHKEVVDSFRSIYTLNDGLSDSEAVAVGRYPEDTYYNGNPWFLCTLAAAEQLYDALYQWDKQGSLEVTD
VSLDFFKALYSDAATGTYSSSSSTYSSIVDAVKTFADGFVSIVETHAASNGSMSEQYDKSDGEQLSARDLTWSY
AALLTANNRRNSVVPASWGETSASSVPGTCAATSAIGTYSSVTVTSWPSIVATGGTTTTATPTGSGSVTSTSKT
TATASKTSTSTSSTSCTTPTAVAVTFDLTATTTYGENIYLVGSISQLGDWETSDGIALSADKYTSSDPLWYVTVT
LPAGESFEYKFIRIESDDSVEWESDPNREYTVPQACGTSTATVTDTWR" ;
                  rs:variable "seq" ] ];

                                          :
                                          :
```

The following **Code 3** is a query that chooses the mannose-binding residue when it has protein sequence information. The result of **Data 3** returns 1,083 triples.

**Code 3** Search the mannose-bound protein from the target UniProt accession number and select the binding residue.

```
PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>
PREFIX dcterms:<http://purl.org/dc/terms/>
PREFIX edam:<http://edamontology.org/>
PREFIX sio:<http://semanticscience.org/resource/>

SELECT DISTINCT ?uniprot ?sqno
WHERE {
 ?het_res pdbo:atom_site.chem_comp.id "MAN".
 ?het_asym dcterms:isPartOf ?het_res;
             rdf:type sio:SIO_010432;
             rdfs:seeAlso ?pdb_res.
 ?pdb_res rdf:type edam:data_1756;
          rdfs:seeAlso ?unpres.
 ?unpres rdf:type edam:data_1756.
 ?unpres dcterms:isPartOf ?uniprot.
 ?unpres rdfs:label ?sqno.

VALUES ( ?uniprot ) {
    ( <http://purl.uniprot.org/uniprot/P13689> )
    ( <http://purl.uniprot.org/uniprot/P26213> )
    ( <http://purl.uniprot.org/uniprot/P12337> )
    ( <http://purl.uniprot.org/uniprot/P02867> )
    ( <http://purl.uniprot.org/uniprot/Q29451> )
    ( <http://purl.uniprot.org/uniprot/Q70KY3> )
  }
} ORDER BY ?uniprot ?sqno
```

**Data 3** This is a header of the data indicating the inquiry result of Code 3 with Turtle (.ttl). In this paper, although omitted, Data 3 returns the result of 1,083 affairs.

```
@prefix dc:         <http://purl.org/dc/elements/1.1/> .
@prefix rs:         <http://www.w3.org/2001/sw/DataAccess/tests/result-set#> .
@prefix x:          <http://example.org/ns#> .
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

[]      rdf:type       rs:ResultSet;
    rs:resultVariable "uniprot";
    rs:resultVariable "sqno";
  rs:solution [
     rs:index 1 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/P02867> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "69"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];
  rs:solution [
     rs:index 2 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/P02867> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "110"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];
  rs:solution [
     rs:index 3 ;
     rs:binding [ rs:value <http://purl.uniprot.org/uniprot/P02867> ;
                  rs:variable "uniprot" ] ;
     rs:binding [ rs:value "111"^^<http://www.w3.org/2001/XMLSchema#integer> ;
                  rs:variable "sqno" ] ];

                                          :
                                          :
```

## 3.3 Introduction of Applications using Biosemantics

In this study, we developed the graph database (UTProt LOD), the pipeline of predictor tools (UTProt Galaxy), and the image library (UTProt-Image) using an open license and based on interactome LOD. The portal site UTProt (http://utprot.net) was used to integrate these products. UTProt LOD has access to the graphics of approximately 30 billion triples. We are assembling data from UTProt LOD and have succeeded in designing predictors of carbohydrates and lipids using support vector machine within only a few weeks. The UTProt Galaxy pipeline was formed from four predictors and was combined with other bioinformatics processing tools so that it could execute a workflow.

### 3.3.2 UTProt Galaxy: Web Pipeline for Interactome Prediction and Analysis Tools

Galaxy is an open-source pipeline framework for the genome science and is used internationally. We built into it a predictive tool for interactome as a module. Although the main part of Galaxy is the program written by Python, we can use several other programming languages such as C/C++, Java, Perl, Python, Ruby, R, and web service API for the pipeline modules. We utilized the backend database that this Web service asks SPARQL queries.

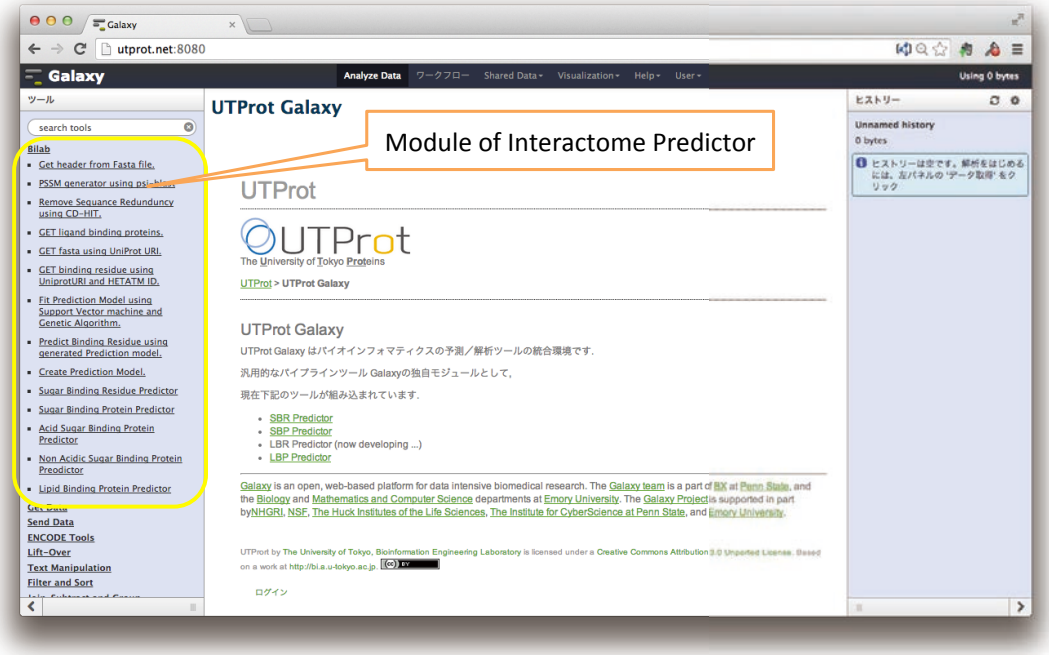


**Figure 10** Screenshot of UTProt Galaxy home. The menu on the left-hand sidebar has a machine-learning predictor module developed by BILAB in the UTProt project. The center indicates the user interface and results. The right-hand sidebar displays the status of the execution, the icon of inspection, and download of data.

The user can obtain access if http://utprot.net:8080 is inputted into UTProt Galaxy in the

address bar of a web browser. If we click the Bilab menu in the left-hand sidebar of the home screen, the user can access each module of the intermolecular interaction predictive tool. Now, the system has a predictor related to sugar-binding proteins and lipid-interacting proteins. In December 2013, the number of modules was 14. **Figure 10** is a screenshot of the UTProt Galaxy home screen. Next the list of those module functions is shown.

Bilab
- Get header from Fasta file.
- PSSM generator using psi-blast
- Remove Sequance Redunduncy using CD-HIT.
- GET ligand binding proteins.
- GET fasta using UniProt URI.
- GET binding residue using UniprotURI and HETATM ID.
- Fit Prediction Model using Support Vector machine and Genetic Algorithm.
- Predict Binding Residue using generated Prediction model.
- Create Prediction Model.
- Sugar Binding Residue Predictor
- Sugar Binding Protein Predictor
- Acid Sugar Binding Protein Predictor
- Non Acidic Sugar Binding Protein Preodictor
- Lipid Binding Protein Predictor

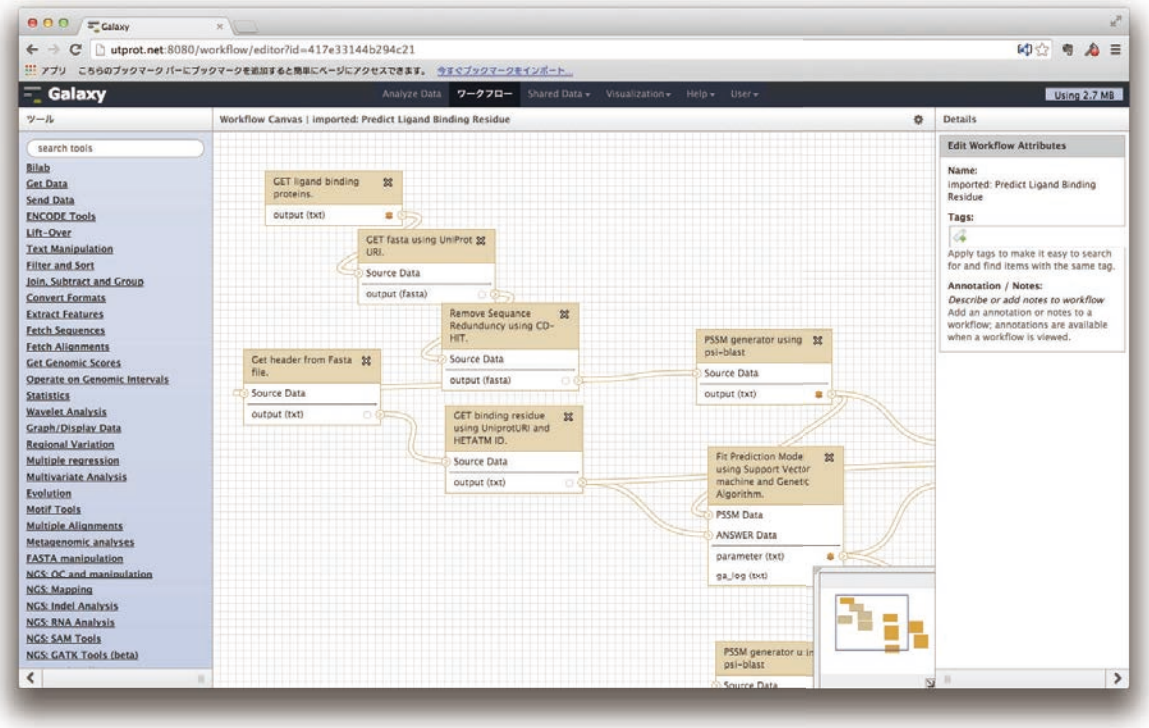


**Figure 11** In UTProt Galaxy, we can edit a module on a connection graphical user interface and can setup a workflow.

The execution of the module from the left-hand sidebar is performed using an interactive method. Next, we describe how these are pipelined. The user can perform a pipeline edit if “Workflow” is chosen from the menu in the upper part of a screen. The brown nodes can connect them freely by a pipe among each of the module functions in **Figure 11**. We believe that this edit function is user-friendly for wet biologists (non-bioinformaticians). The user can also use the default modules of Galaxy other than the collective predictive tools of UTProt.

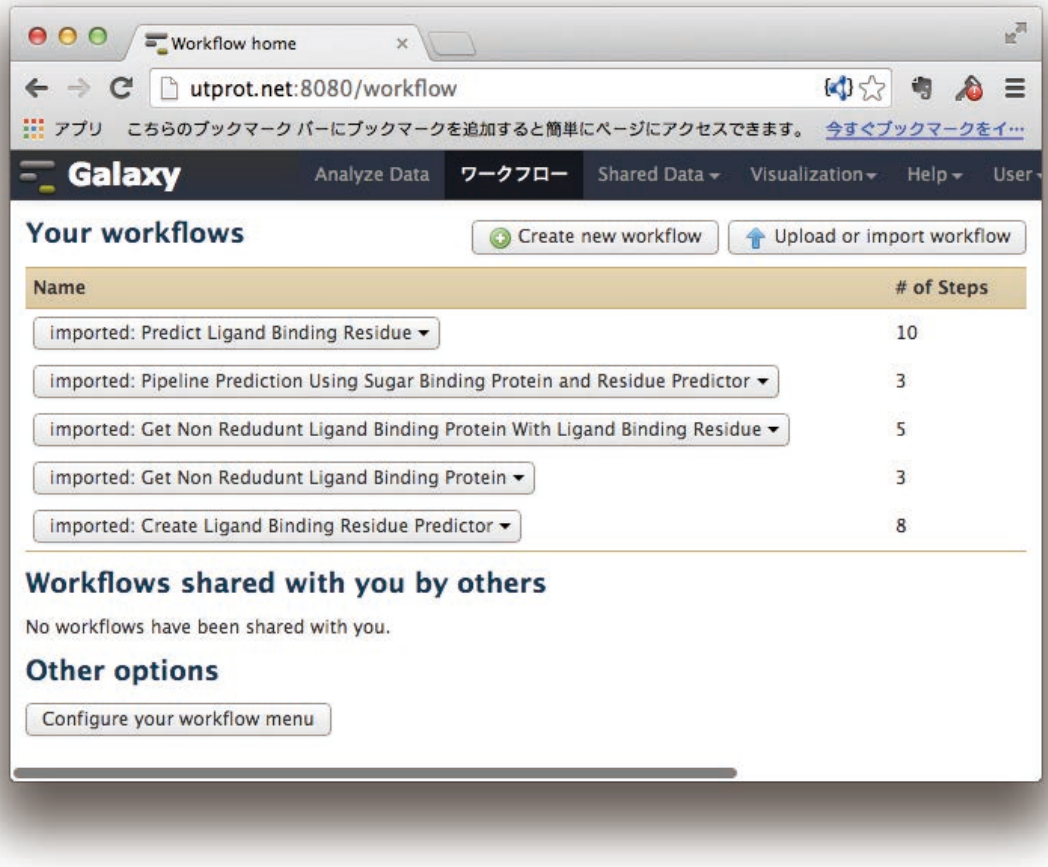


**Figure 12** Pipeline edited using the workflow of UTProt Galaxy can perform repeated operations using a single action. It can be shared with other users.

Moreover, the editor tool that was created reduces the number of workflows, and end users can perform it any number of times. Different users can also share the pipeline that was created. The workflow that was created by the development team of UTProt is prepared as the default. **Figure 12** is a screen showing the workflow execution.

### 3.3.3 UTProt Image: Open-license Image Library for Interactome for which CC BY was Applied to All Pictures

The development of the application for interactome requires the image data of the intermolecular interaction, which can be used immediately with an open license. In the UTProt-Image, a user can locate graphics from over 480,000 helpful images; these can be downloaded for interactomics using an open license. Raster graphics (in the PNG format) and vector graphics (in the SVG format) are provided as images in this service. **Figure 13** is the home screen of the UTProt Image service provided by us. We can peruse the Website by accessing http://www.utprot.org/resource/images/. There are three items (PDB thumbnail, Ligand thumbnail, and PLBSP Ligplot) at the top page in 2013. The PDB thumbnail is a two-dimensional figure of the solid structure, where the ligand drawn by the script of UCSF Chimera was emphasized. The ligand thumbnail shows the two-dimensional figures of the structural model drawn about the low molecular compound using the command line tools of ChemAxon [84]. PLBSP Ligplot is the result of 400,000 affairs obtained by Ligplot for the analysis of the ligand–residue pair of PLBSP.


**Figure 13** In 2013, the UTProt Image had three categories: the thumbnail of the protein structure of PDB, the chemical structure thumbnail of the ligand, and the two-dimensional picture of the interaction of a ligand and amino acid residue, which was calculated by Ligplot based on the result of PLBSP. The total exceeds 480,000 affairs.

**Figure 14** is a screenshot of an open page of the individual record. We display the protein using a wire model, which is indicated by a sphere as a low molecular compound. The protein color is distinguished by color and is based on the secondary structure. The helix is purple, the strand is green, and the coil is gray. The user first needs to do an ON collection of the PDB ID. The number of records is 84,638.

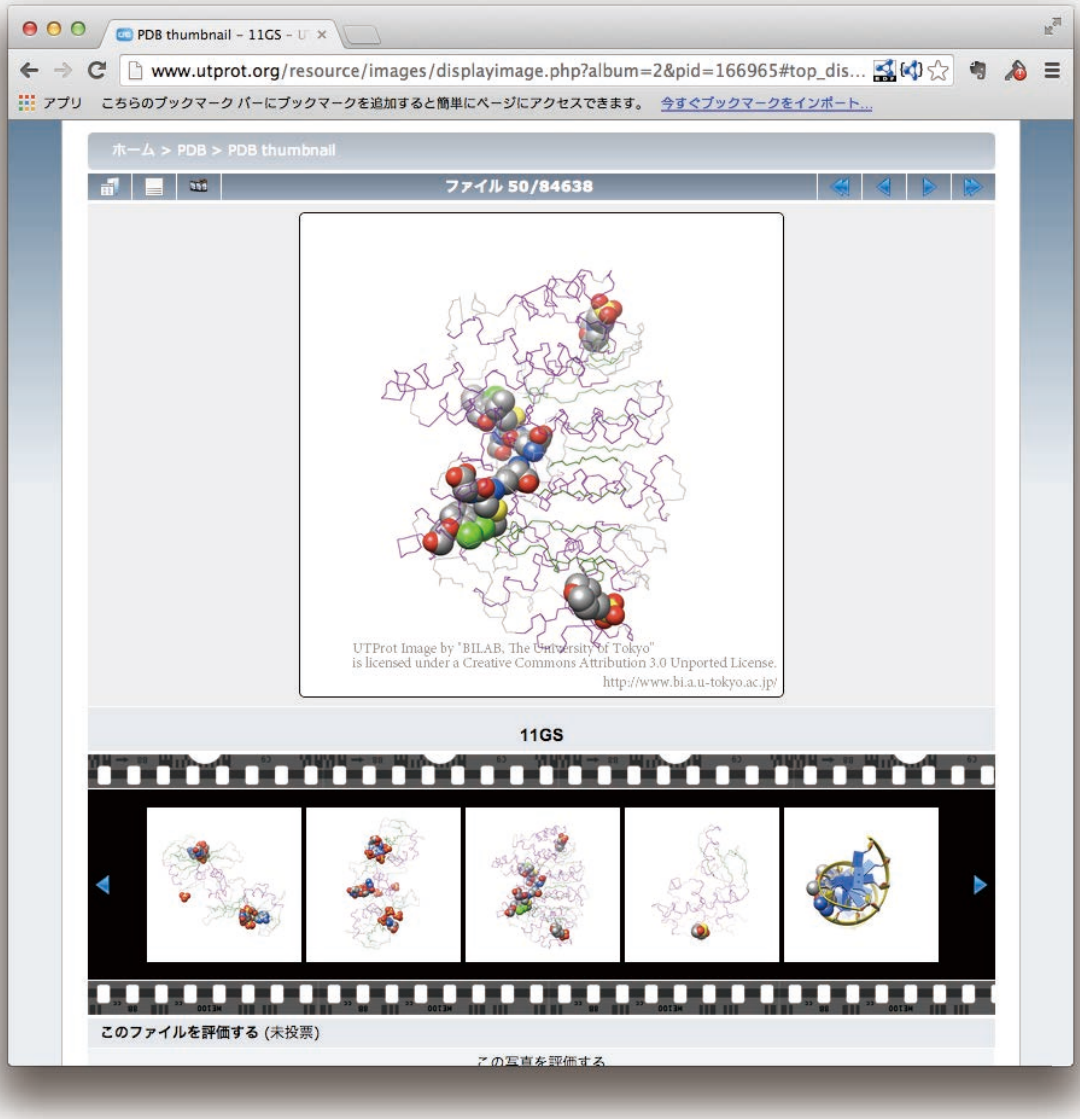


**Figure 14** This is a screenshot displaying 11GS using PDB ID in UTProt Galaxy. We can obtain two-dimensional images of a structure having 84,000 affairs by searching PDB ID at an input. A user can use all pictures under the condition of Creative Commons Attribution 3.0, which is an open-license tool.

**Figure 15** illustrates the chemical structure of sugar that was drawn using ChemAxon. The graphics of the ligand was tentatively created and has only 241 entries. They are registered by InChIKey, which is a hash key of International Chemical identifier (InChI) that identifies low molecular compounds [85]. The user first needs to do a search of the ON collection of the InChIKey.

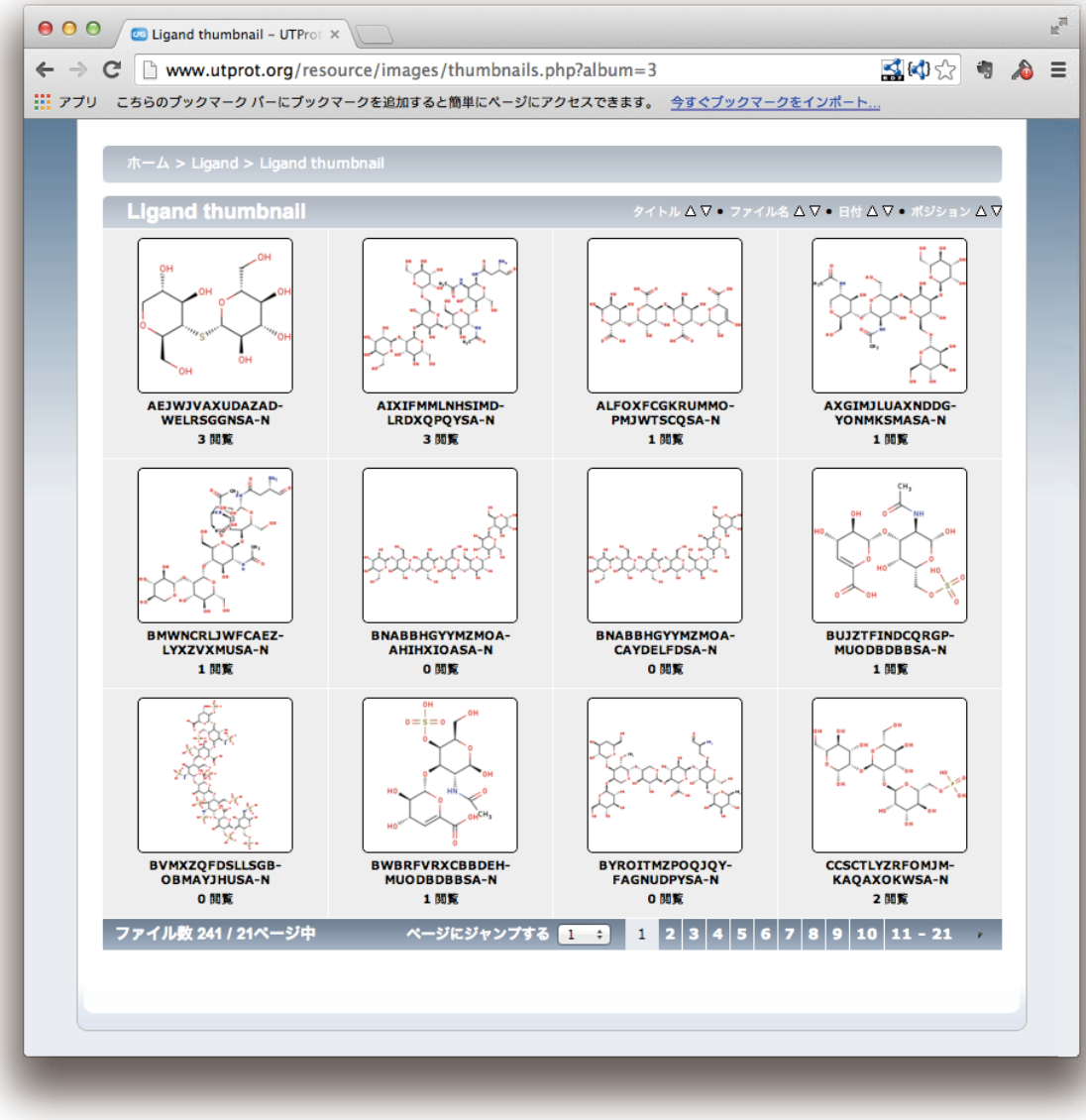


**Figure 15** This is a screenshot that searches the image for a ligand of sugar with a UTProt Image. A user can search an entry with InChIKey.

**Figure 16** is the result of a search for the mannose-bound amino acid residue with keyword “MAN” as the item of PLBSP-Ligplot in the UTProt Image. With respect to the result, there were 162 hits, and for 160 of them, files are included in PLBSP-Ligplot, and the remaining 2 files are included in the PDB thumbnail category.

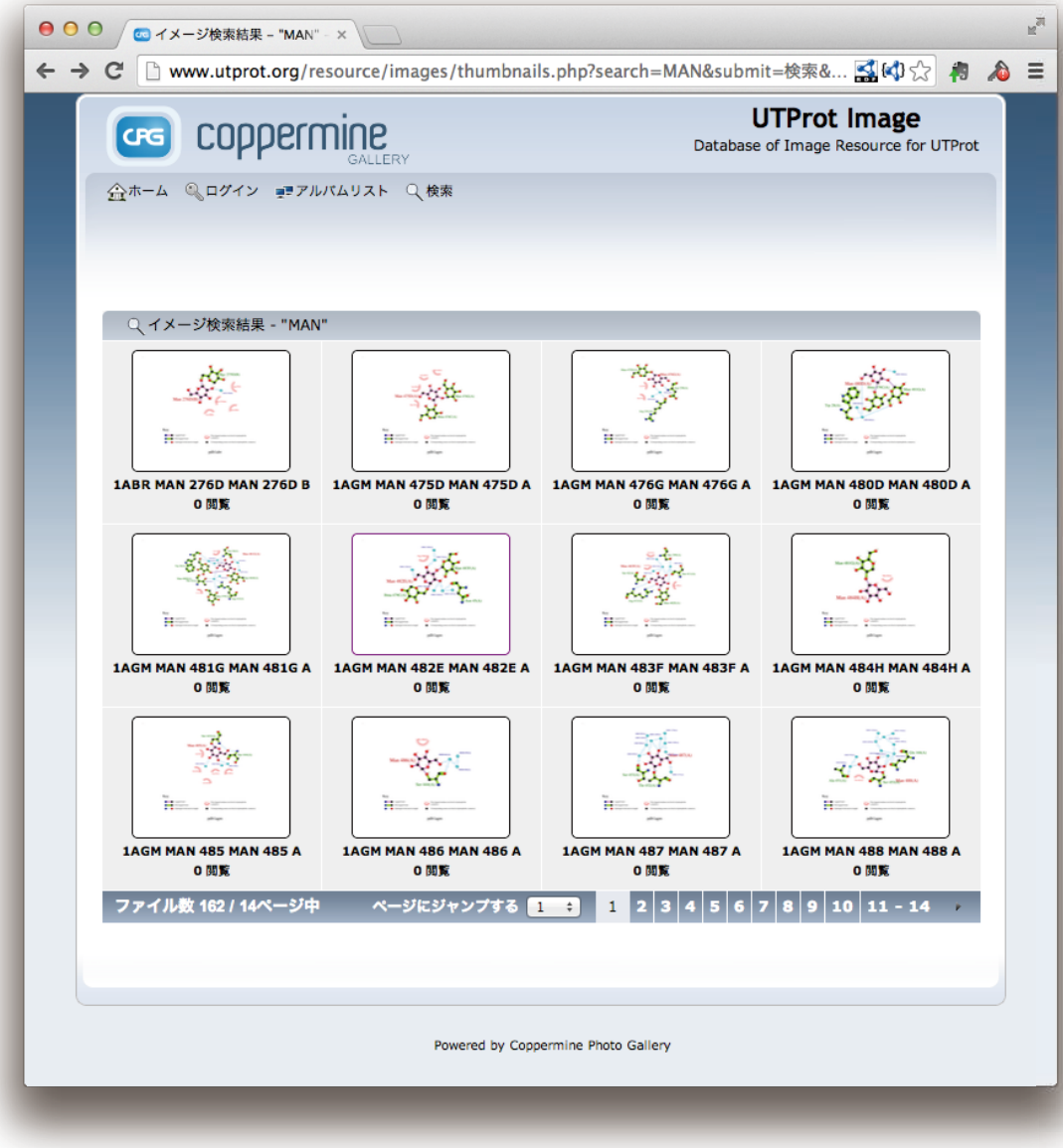


**Figure 16** This is a screenshot that searches with the item of PLBSP-Ligplot about mannose (MAN) in UTProt Image. There are 400,000 entries.

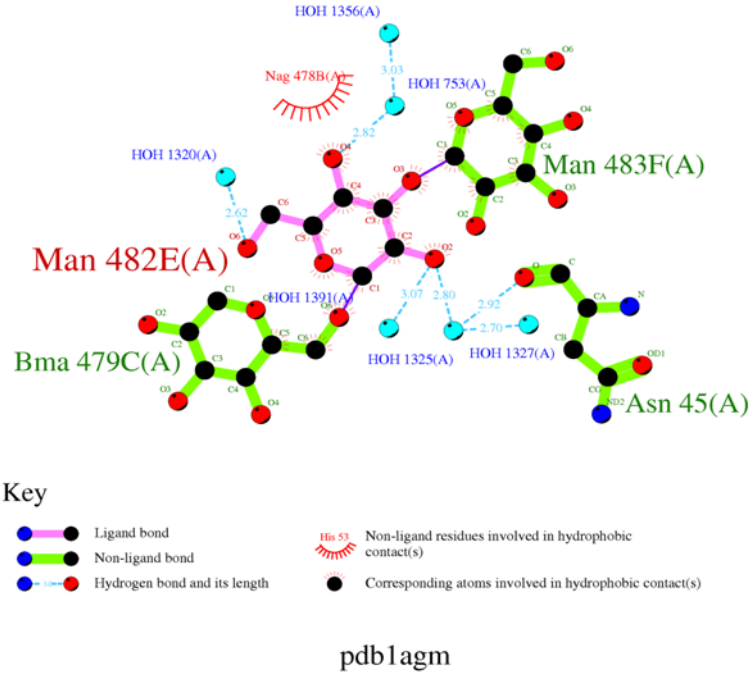


**Figure 17** As an example, a user acquires the two-dimensional picture of an interaction with the amino acid residue around the sequence number MAN 482 of A-chain of PDB ID 1AGM. The user can use these analysis results obtained by CC BY.

The interactome LOD quickly changed the machine-learning dataset concentrations. From this work, the developers can utilize UTProt, PDB, RDF-SIFTS, PDB Residue, and other databases. Furthermore, two or more predictors were extended to a user-friendly pipeline. The end users can use the open images in UTProt for research, education, and business under licensed conditions. One merit of data-driven science is the reformulation and reusability of data. We continue to enrich the contents of interactome in UTProt.

Here if an individual record is clicked, a detailed picture such as that in **Figure 17** will be displayed. This figure is drawn by the analysis of Ligplot in which red (pink) represents ligands and green represents the surrounding amino acid residues. Purple represents the covalent bonds, the light-blue ball represents $H_2O$, and the light-blue dashed line shows the hydrogen bond. The orange icon represents the hydrophobic contact. We calculated these using the unused processor on two-cluster computing systems such as tostogw2/tostogw4 series, which we own. **Table 2** shows the specifications of the cluster machine used for Ligplot's calculation. Moreover, we entered the configuration file of Ligplot for **Code 4**. In

this experiment, with the exception of the parameter file, we used the default value.

In addition, information such as the ID of PDB to input a ligand (or amino acid residue) name or a residue number are used in the PLBSP database. Here, this information is the most significant, which reuses LOD integrated by UTProt RDF Platform for HPC.

**Table 2** Specifications of a cluster machine that used the PLBSP origin data in UTProt Image for calculations by Ligplot.

| Clustar Number | Host Name | CPU | Memory [MB] | HDD [GB] | OS (kernel version) |
|---|---|---|---|---|---|
| Clustar 1 | tosto19-32 | Xeon 2.8 GHz | 4096 | 80 | SuSE 7.3 (2.4.20) |
| Clustar 2 | tosto55-58 | Xeon 3.47 GHz 6core*2 | 20540 | 240 | CentOS 5.5 (2.6.18-164.15.1.el5) |

**Code 4** This code has the parameter specification at the time of Ligplot execution. We omitted the portion of NOTE.

```
LIGPLOT v.4.0  -  Parameter file (ligplot.prm)
---------------------------------

PRINT OPTIONS
-------------
Y       <- Produce a colour PostScript file (Y/N)?
L       <- Orientation of plot: (P)ortrait or (L)andscape?
0.0         <- Rotation angle (clockwise) for final plot

PLOT PARAMETERS
---------------
Y       <- Include: Hydrophobic interactions - (Y/N)?
Y       <- Include: Water molecules - (Y/N)?
Y       <- Include: Non-ligand mainchain atoms - (Y/N)?
Y       <- Include: Linked residues listed below - (Y/N)?
Y       <- Plot: Hydrogen bonds - (Y/N)?
Y       <- Plot: Internal H-bonds in ligand - (Y/N)?
Y       <- Plot: External groups covalently bonded to ligand - (Y/N)?
N       <- Plot: Bonds showing hydrophobic interactions - (Y/N)?
N       <- Plot: Schematic ligand representation [see Note 1] - (Y/N)?
N       <- Plot: Schematic non-ligand residues [see Note 1] - (Y/N)?
N       <- Plot: Accessibility shading [see Note 2] - (Y/N)?
Y       <- Plot: Ligand atoms (as spheres) - (Y/N)? [see Note 4]
Y       <- Plot: Nonligand atoms (as spheres) - (Y/N)? [see Note 4]
Y       <- Plot: Double- and triple bonds - (Y/N)?
Y       <- Print: Key to symbols in PostScript output - (Y/N)?
Y       <- Print: Residue names/numbers - (Y/N)?
Y       <- Print: Atom names - (Y/N/C)? [see Note 4]
Y       <- Print: H-bond lengths on hydrogen bonds - (Y/N)?
Y       <- Print: Filename as title if title not explicitly defined - (Y/N)?
Y       <- Plot: Solid lines for covalent bonds to external groups - (Y/N)?
1       <- Non-bonded contacts option [see Note 3]
Y       <- Plot: Water atoms (as spheres) - (Y/N)? [see Note 4]
Y       <- Plot: Accessibility shading for the ligand only - (Y/N)?
N       <- Chemical notation   [see Note 4] - (Y/N/C)?

                                                    :
```

```
                                        (The rest is omitted.)
                                                :

LINKED RESIDUES (see Note below)
----------------
HIS-ASP   <- Residue-pair 1
HOH-***   <- Residue-pair 2
          <- Residue-pair 3
          <- Residue-pair 4
          <- Residue-pair 5
          <- Residue-pair 6
          <- Residue-pair 7
          <- Residue-pair 8
          <- Residue-pair 9
          <- Residue-pair 10

                                                :
                                        (The rest is omitted.)
                                                :

SIZES (All sizes are relative sizes given in Angstroms)
-----
0.33  <- Radius: Ligand atoms
0.33  <- Radius: Non-ligand atoms
0.33  <- Radius: Water molecules
1.15  <- Radius: Hydrophobic contact residues
0.80  <- Radius: Ligand residues in simple-residue representation
0.25  <- Line-thickness: Ligand bonds
0.35  <- Line-thickness: Non-ligand bonds
0.07  <- Line-thickness: Hydrogen bonds
0.07  <- Line-thickness: External covalent bonds


TEXT SIZES (Relative sizes in Angstroms)
-----------
1.20  <- Residue names: Ligand residues
1.00  <- Residue names: Non-ligand residues
0.50  <- Residue names: Water molecule IDs
0.50  <- Residue names: Hydrophobic-interaction residues
0.50  <- Residue names: in simple-residue representation
0.31  <- Atom names: Ligand atoms
0.31  <- Atom names: Non-ligand atoms
0.44  <- Hydrogen-bond lengths


COLOURS (Note: colour definitions are given at end of file)
-------
WHITE          <- Background colour of page
PINK           <- Ligand bonds [or ATOM - see Note]
LIME GREEN     <- Non-ligand bonds [or ATOM - see Note]
SKY BLUE       <- Hydrogen bonds
PURPLE         <- External covalent bonds
RED            <- Hydrophobic interactions
ORANGE         <- Accessibility shading: Buried atoms
YELLOW         <- Accessibility shading: Accessible atoms
BLUE           <- Nitrogen atoms
RED            <- Oxygen atoms
BLACK          <- Carbon atoms
YELLOW         <- Sulphur atoms
TURQUOISE      <- Water atoms
PURPLE         <- Phosphorus atoms
LILAC          <- Iron atoms
GREEN          <- All other atoms
BLACK          <- Atom edges
BLACK          <- Circles in simple-residue representation

                                                :
                                        (The rest is omitted.)
```

```
                                                                    :

TEXT COLOURS (Note: colour definitions are given at end of file)
------------
BLACK             <- Plot title
BLACK             <- Legends in key to symbols
BRICK RED         <- Residue names: Ligand residues
OLIVE GREEN       <- Residue names: Non-ligand residues
BLUE              <- Residue names: Water molecule IDs
RED               <- Residue names: Hydrophobic-interaction residues
BRICK RED         <- Atom names: Ligand atoms
OLIVE GREEN       <- Atom names: Non-ligand atoms
SKY BLUE          <- Hydrogen bond lengths


COLOUR DEFINITIONS
------------------
0.0000 0.0000 0.0000 'BLACK            '<- Colour 1
1.0000 1.0000 1.0000 'WHITE            '<- Colour 2
1.0000 0.0000 0.0000 'RED              '<- Colour 3
0.0000 1.0000 0.0000 'GREEN            '<- Colour 4
0.0000 0.0000 1.0000 'BLUE             '<- Colour 5
1.0000 1.0000 0.0000 'YELLOW           '<- Colour 6
0.8000 0.5000 0.0000 'ORANGE           '<- Colour 7
0.5000 1.0000 0.0000 'LIME GREEN       '<- Colour 8
0.5000 0.0000 1.0000 'PURPLE           '<- Colour 9
0.5000 1.0000 1.0000 'CYAN             '<- Colour 10
1.0000 0.5000 1.0000 'PINK             '<- Colour 11
0.3000 0.8000 1.0000 'SKY BLUE         '<- Colour 12
1.0000 1.0000 0.7000 'CREAM            '<- Colour 13
0.0000 1.0000 1.0000 'TURQUOISE        '<- Colour 14
1.0000 0.0000 1.0000 'LILAC            '<- Colour 15
0.8000 0.0000 0.0000 'BRICK RED        '<- Colour 16
0.5000 0.0000 0.0000 'BROWN            '<- Colour 17
0.9700 0.9700 0.9700 'LIGHT GREY       '<- Colour 18
0.1000 0.5000 0.0000 'OLIVE GREEN      '<- Colour 19
1.0000 1.0000 1.0000 'WHITE            '<- Colour 20


MINIMIZATION PARAMETERS
-----------------------
 10.00      <- Atom-atom clash parameter
  0.20      <- Bond-atom clash parameter
200.0       <- Bond-overlap score
  0.5       <- Weight for term giving H-bond deviation from ideal value
  0.01      <- Weight for term giving non-bond dist deviation from ideal value
 10.0       <- Weight for internal energies (atom clashes, etc)
 15.0       <- Furthest move-distance for H-bonded groups (in Angstroms)
  1.0       <- "Stretch factor" for H-bond lengths
1000        <- Number of loops for the minimization process
0.0         <- Terminate minimization if energy drops by less than this value
N           <- Random start for minimization routines - (Y/N)?
20.0        <- Weight for anchor-position energy term
5.0         <- Weight for interface-boundary energy term
2.0         <- Closest atom-distance to boundary representing interface
0.3         <- Weight for relative residue-positions energy term
```

### 3.4 Method of License Control for Circulatory Knowledge

In this section, we propose a method for collection and accumulation without losing important data results. From the first stage, it was necessary to deal with scientific data considering patent conditions and the privacy policy. However, after papers are published, if a patent is being sought, scientists and engineers are required to exhibit data related to the product. This is because the public presentation of data ensures the circulation of new knowledge. Then, we implemented a data repository in a closed environment that assumes open-source data from the beginning. We named as the BILAB Data Portal to it that is based on CKAN. CKAN is a data portal of global standard for open data as a framework of an open source made by the Open Knowledge Foundation (OKF). We develop this web service on AWS EC2. Accessible IP was restricted in a port level to access control and offered closed environment for the security. BILAB Data Portal makes the open data ones from closed data easy. BILAB Data Portal can register required memo, program, figure and report in the mush-up development during the project duration. Writing an experiment note of scientific research is important to continue. BILAB Data Portal supports succession of experimental data. A successor has a merit that can inherit a predecessor's work quickly by experimenting on data in structure per laboratory or project and accumulating them. Furthermore, an engineer may realize the idea which the researcher proposed. We think it to be the first step toward open data to share science data in an organization.

**Figure 18** shows the home screen of the BILAB Data Portal, on which we spent two years performing research and registered the project of 33 affairs as a dataset. If a keyword is inputted into a search form, we can perform the search for a dataset in retrieval by a natural sentence. For this purpose, it is necessary to provide the description of a dataset. Moreover, the user can refer to the tag given to the dataset and the file format of the data registered.


**Figure 18** This is a prototype of the BILAB data portal for accumulating and reusing the scientific data in a laboratory using CKAN. Thirty-three datasets were registered while using the closed environment for two years. The end users can use the search form: full-text search, facet search, or directory search.

In this system, the user can perform a directory search of the grouped dataset. In addition, the user can perform a facet search using the tag or file format. **Figure 19** shows a screenshot of the search screen.

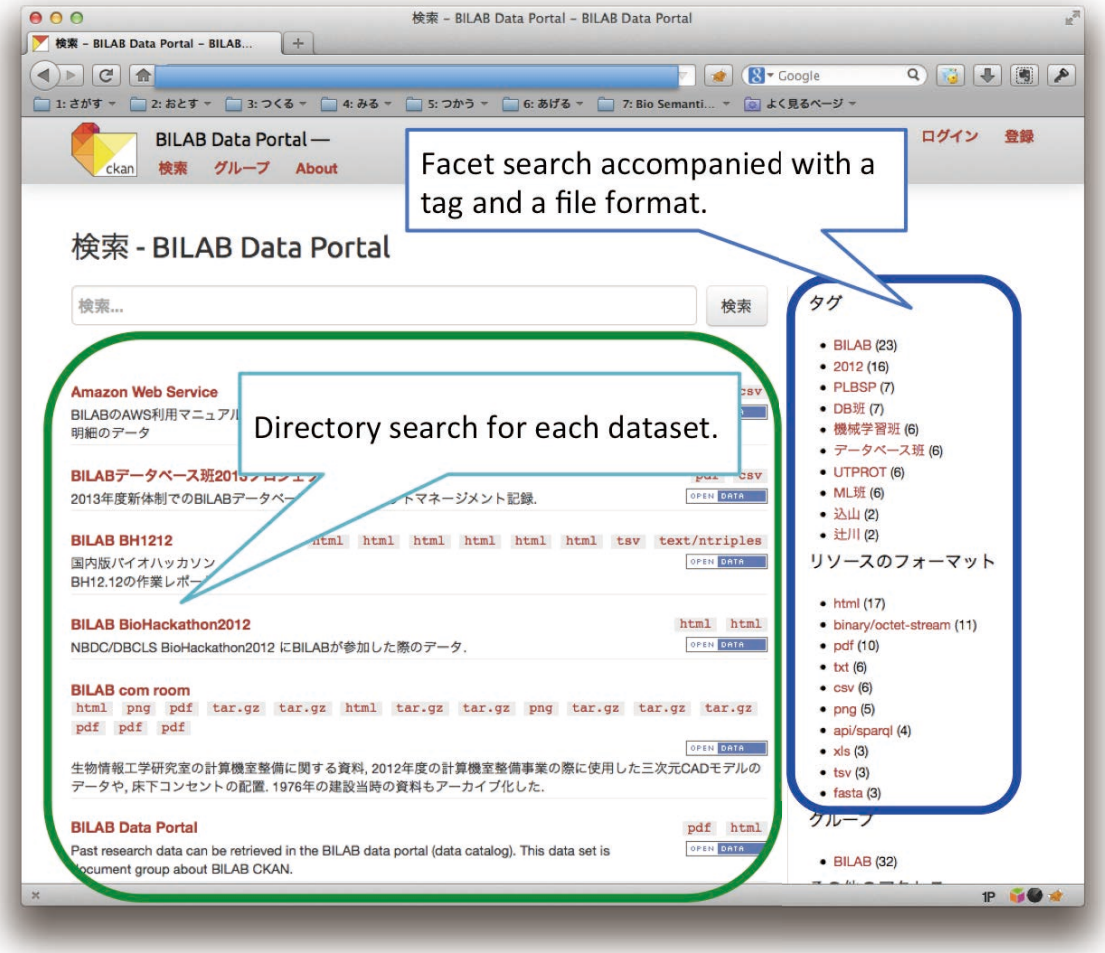


**Figure 19** We display the list of datasets in the BILAB Data Portal. The main area displays the title, description, license, and file format of the records. A user can browse the right-hand sidebar from the format group name of a tag resource.

A user registers data individually into a dataset. The user can link to data or a document, or perform file uploads or API registration. We registered the result of our research to the BILAB Data Portal. **Figure 20** is an entry of the project involving a ligand-binding site predictive tool.

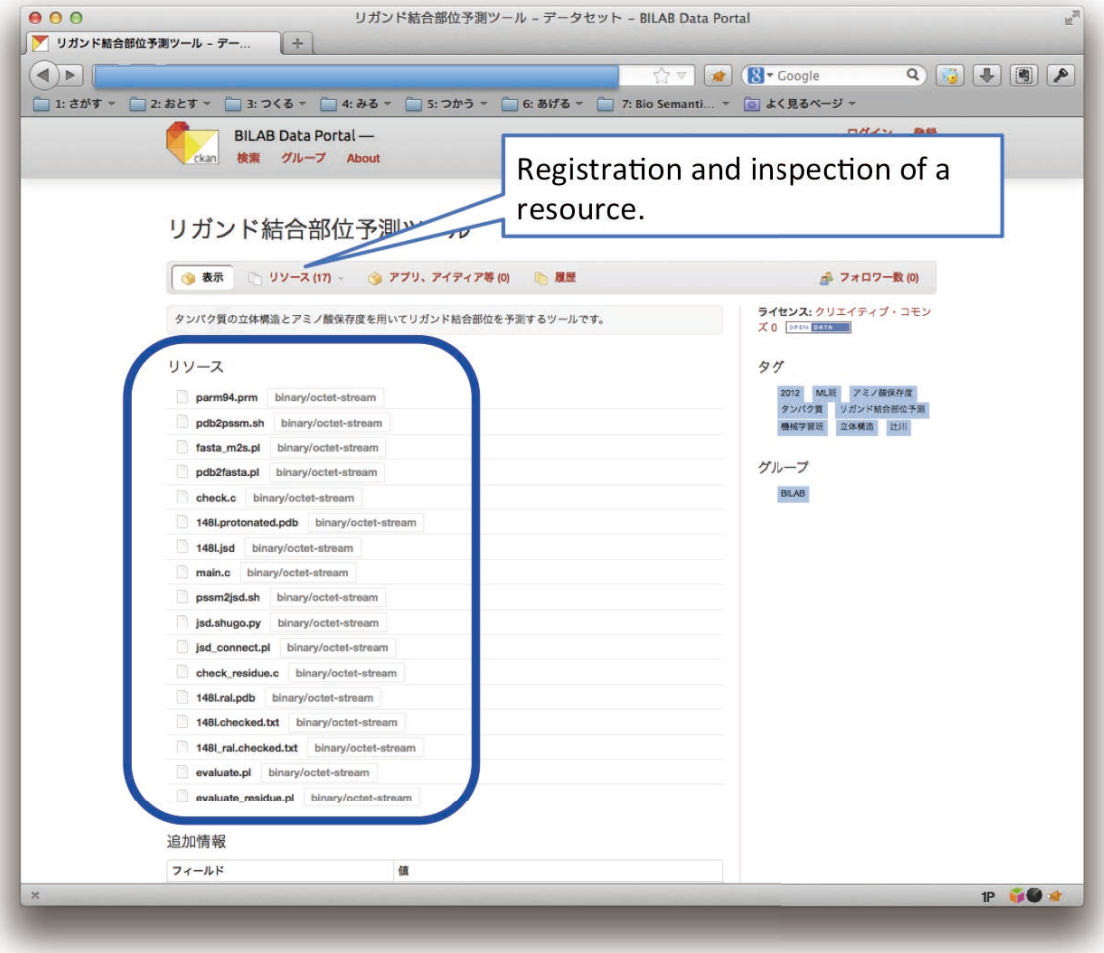


**Figure 20** When a researcher moves out of a laboratory, even if his successor does not ask permission, he/she can access previous information because of the open license assigned to the document and the program.

The file name, file format, license, preview, and metadata are displayed in the resource record. The user clicks the upper-right download button and acquires the relevant data. **Figure 21** is a screenshot displaying resources of the bottom of the heap.

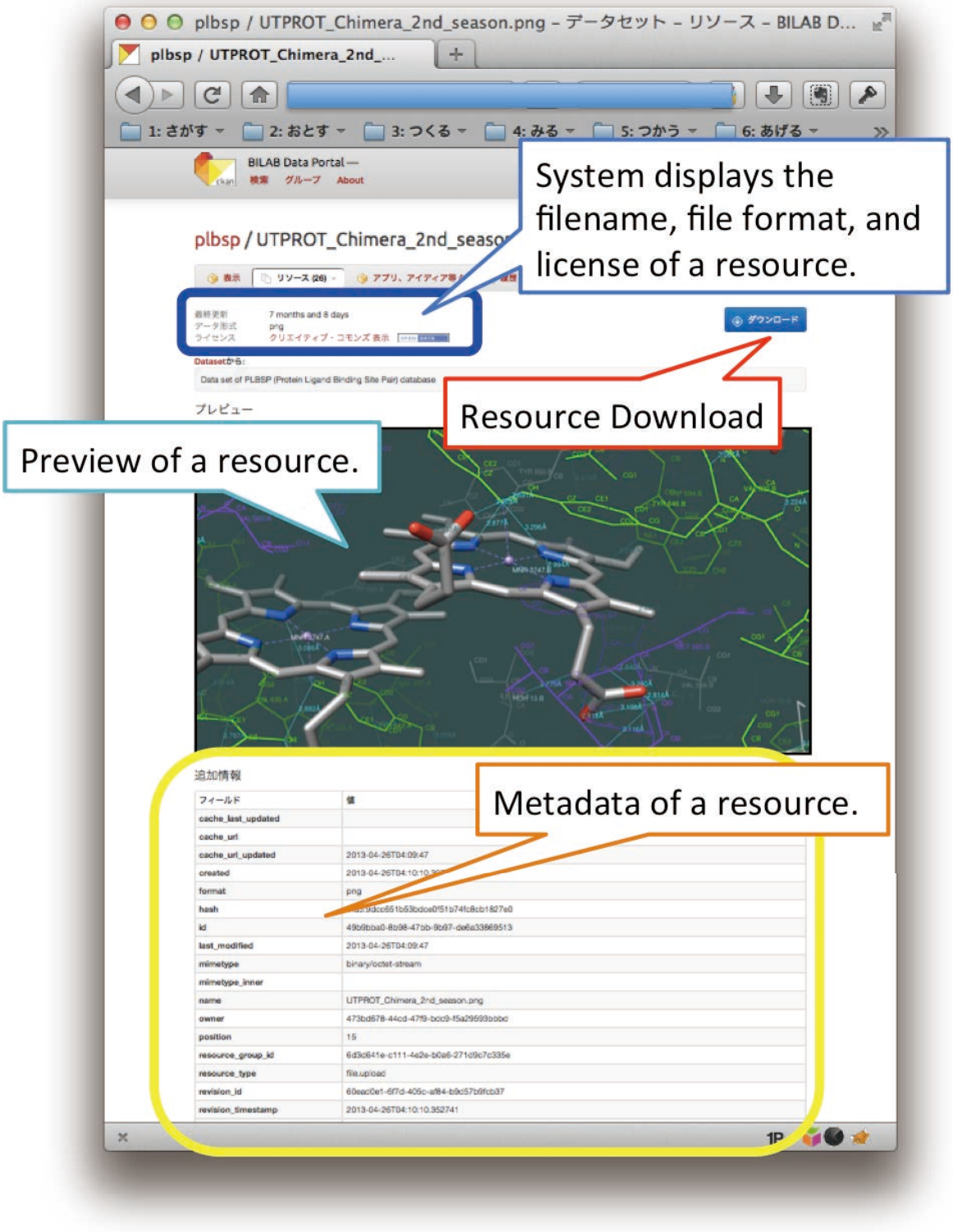


**Figure 21** In the BILAB Data Portal, we have assigned a license and metadata to the document program and the figure in the dataset (project). Such metadata are easily acquirable from REST API in the RDF format. The users can download data manually.

# Chapter 4 CONCLUSION

In conclusion, we developed a novel framework for reducing bioinformatics tasks using biosemantics that obtain interactome RDF data. We extended those base databases and supported the intermolecular interaction predictive tool development using machine learning. As a result, we could shorten the development period of the application. Interactome LOD can generate few predictive tools as modules of the workflow on the Web. We exhibited 480,000 images that provide common open licenses for database developers. It employs all licenses using the Creative Commons Attribution [80]. Furthermore, we proposed a method that stores in a laboratory life-science-related data, which is continually increasing and which can be used in future public presentations. We improved the ability to ensure the continuity of research by providing a structure and license using meta-information. We believe that biosemantics will be a basis of bioinformatics in the future.

## 4.1 UTProt: LOD-based Database Integration and Application Development for Interactome

We applied the concrete method of biosemantics to interactome. The result was an integrated UTProt database. The working efficiency of bioinformaticians can be increased by offering all with the RDF data. The short-term project of analytical (predictor) tools development of JST NBDC for four months is using the UTProt RDF Platform. In our project, we developed a module for three predictive tools and 14 pipelines. The UTProt RDF Platform can generate predictive tools as the modules of the workflow on the Web. We exhibited 480,000 images that have an open license for database developers. It employs all licenses by Creative Commons Attribution. Furthermore, we proposed a method of circulating knowledge to increase life science data in a closed organization. This idea will allow ease of registration to public databases in the future. We therefore improved the continuity of research by providing the structure and license using meta-information. We believe that biosemantics will be the basis of bioinformatics in the future.

## 4.2 Fusion of Data-driven Science and Bioinformatics in Big Data Era

Finally, we describe the short-term view about the public performance and society of science data. In the big data age, databases are indispensable to life science research. The design of drugs for medical treatment and new agricultural breakthroughs are expected due to advances in database integration. However, there remain some barriers. One is that biologists,

medical doctors, and chemists are not necessarily open to public presentation of their experimental data. They are usually interested only in registering to databases that are required to publish papers. Besides, in many cases, they may opt to hide the data when technology unique to the data is included. Using only an exclusive system for such a system, we consider that it will become impossible to clear future research tasks. Then, we proposed computational biosemantics, which combines open data, Semantic Web, bioinformatics, and a life science database. The most important aspect is that the open license is assigned to data document software. When a bioinformatician conducts a secondary analysis using the database, he/she can access and use it freely. We believe that open data and an open-access journal is required to reach other researchers. Other persons will be able to discover the value of data for which the data provider has not yet realized the value. Furthermore, an engineer may develop a tool that is useful for biologists and doctors, who are data providers. We believe that the integration and accumulation of open science data is helpful for purposes besides the promotion of scientific applications. We need to create a biosemantics foundation for molecular biology and structural biology to unify society. In December 2013, the Japanese government offered information related to machine-readable open data. Moreover, we have enabled the large-scale transfer of open data in various places, including private enterprises and civic activities other than universities and research institutions. All the results of the scientific results by a government subsidy should be indicated to people. Open data was useful for gathering information during the Tohoku earthquake and tsunami in Japan at 2011. Building and analyzing the basis for scientific information will contribute to the determination of national policy. Here, we specify that CC BY is given to the full text of this thesis. It is considered to be a step toward the generation of open science data.

# LIST OF FIGURES AND TABLES

## REFERENCES


1. Millikan RG (1989) Biosemantics. J Philos Vol. 86: pp. 281–297. Available:http://www.jstor.org/stable/info/2027123. Accessed 22 December 2013.

2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. Available:http://dx.doi.org/10.1038/75556. Accessed 11 December 2013.

3. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–61. Available:http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D258. Accessed 16 July 2012.

4. W3C (2004) RDF Primer. Available:http://www.w3.org/TR/rdf-primer/. Accessed 23 December 2013.

5. Tim Berners Lee (2006) Linked Data - Design Issues. Available:http://www.w3.org/DesignIssues/LinkedData.html. Accessed 23 December 2013.

6. Katayama T, Arakawa K, Nakao M, Ono K, Aoki-Kinoshita KF, et al. (2010) The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. J Biomed Semantics 1: 8. Available:http://www.jbiomedsem.com/content/1/1/8. Accessed 14 August 2012.

7. Katayama T, Wilkinson MD, Vos R, Kawashima T, Kawashima S, et al. (2011) The 2nd DBCLS BioHackathon: interoperable bioinformatics Web services for integrated applications. J Biomed Semantics 2: 4. Available:http://www.jbiomedsem.com/content/2/1/4. Accessed 14 August 2012.

8. Katayama T, Wilkinson MD, Micklem G, Kawashima S, Yamaguchi A, et al. (2013) The 3rd DBCLS BioHackathon: improving life science data integration with semantic Web technologies. J Biomed Semantics 4: 6. Available:http://www.jbiomedsem.com/content/4/1/6. Accessed 12 February 2013.

9. NBDC, DBCLS (2008) BioHackathon. Available:http://www.biohackathon.org/. Accessed 23 December 2013.

10. 乙守信行, 長野伸一, 佐藤宏之, 萩野達也 (2012) Linked Open Data チャレンジ Japan 2011 を振り返って. 人工知能学会誌 27: pp.518–526.

11. LOD チャレンジ実行委員会 (2013) Linked Open Data Challenge Japan 2013. Available:http://lod.sfc.keio.ac.jp/challenge2013/. Accessed 23 December 2013.

12. National Bioscience Database Center (2011). Available:http://biosciencedbc.jp/en/. Accessed 23 December 2013.

13. DBCLS (2007) Database Center for Life Science. Available:http://dbcls.rois.ac.jp/en/. Accessed 23 December 2013.

14. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, et al. (2011) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res: gkr811–. Available:http://nar.oxfordjournals.org/cgi/content/abstract/gkr811v1. Accessed 6 October 2011.

15. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2013) The future of the protein data bank. Biopolymers 99: 218–222. Available:http://www.ncbi.nlm.nih.gov/pubmed/23023942. Accessed 22 December 2013.

16. Kosuge T, Mashima J, Kodama Y, Fujisawa T, Kaminuma E, et al. (2013) DDBJ progress report: a new submission system for leading to a correct annotation. Nucleic Acids Res: gkt1066–. Available:http://nar.oxfordjournals.org/content/early/2013/11/04/nar.gkt1066.long. Accessed 22 December 2013.

17. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34. Available:http://nar.oxfordjournals.org/content/27/1/29.long. Accessed 22 December 2013.

18. Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27–30. Available:http://nar.oxfordjournals.org/content/28/1/27.long. Accessed 13 December 2013.

19. The UniProt Consortium (2012) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res: gks1068–. Available:http://nar.oxfordjournals.org/content/early/2012/11/16/nar.gks1068.long. Accessed 20 November 2012.

20. Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, et al. (2013) The ChEMBL database as linked open data. J Cheminform 5: 23. Available:http://www.ncbi.nlm.nih.gov/pubmed/23657106. Accessed 9 May 2013.

21. EMBL-EBI (2013) The EBI RDF Platform. Available:http://www.ebi.ac.uk/rdf/. Accessed 23 December 2013.

22. Cood EF (1969) A Relational Model of Data for Large Shared Data Banks. IBM Res Lab Calif. Available:http://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf.

23. Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner DJ (2006) Computer science. Creating a science of the Web. Science 313: 769–771. Available:http://www.sciencemag.org/content/313/5788/769.short. Accessed 14 December 2013.

24. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a

mashup to build bioinformatics knowledge systems. J Biomed Inform 41: 706–716. Available:http://www.ncbi.nlm.nih.gov/pubmed/18472304. Accessed 25 July 2010.

25. Chen B, Ding Y, Wang H, Wild DJ, Dong X, et al. (2010) Chem2Bio2RDF: A Linked Open Data Portal for Systems Chemical Biology. 2010 IEEE/WIC/ACM Int Conf Web Intell Intell Agent Technol: 232–239. Available:http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5616542. Accessed 23 July 2012.

26. Katayama T, Nakao M, Takagi T (2010) TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. Nucleic Acids Res 38: W706–11. Available:http://nar.oxfordjournals.org/cgi/content/abstract/38/suppl_2/W706. Accessed 12 August 2011.

27. 米澤明憲 (2011) 平成 23 年度 研究開発実施報告書 ライフサイエンスデータベース統合推進事業「基盤技術開発プログラム」. Available:http://biosciencedbc.jp/gadget/rdprog_over/K01_yonezawa_h.pdf. Accessed 23 December 2013.

28. 岡本忍, 藤澤貴智, 川島秀一, 片山俊明 (2013) ＴｏｇｏＧｅｎｏｍｅ／Ｔｏｇｏ Ｓｔａｎｚａ：ゲノム情報統合と再利用のためのプラットフォーム. 日本ゲノム微生物学会年会要旨集 7. Available:http://jglobal.jst.go.jp/public/20090422/201302273317628260#jgid201302273317628260. Accessed 23 December 2013.

29. Fujisawa T, Okamoto S, Katayama T, Nakao M, Yoshimura H, et al. (2013) CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes. Nucleic Acids Res: gkt1145–. Available:http://nar.oxfordjournals.org/content/early/2013/11/24/nar.gkt1145.long. Accessed 11 December 2013.

30. Yamamoto Y, Yamaguchi A, Bono H, Takagi T (2011) Allie: a database and a search service of abbreviations and long forms. Database (Oxford) 2011: bar013. Available:http://database.oxfordjournals.org/content/2011/bar013. Accessed 4 September 2011.

31. Sadeghi MR (2013) Pubmed/PMC as the First Line Resource in Biomedicine Field. J Reprod Infertil 14: 95. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3799274&tool=pmcentrez&rendertype=abstract. Accessed 22 December 2013.

32. 総務省 (2012) オープンデータの取組について | 電子政府の総合窓口 e-Gov［イーガブ］. Available:http://www.e-gov.go.jp/doc/opendata/. Accessed 23 December 2013.

33. Saad G (2011) Analysis and Classification of Protein-Ligand Binding Sites using Octree The University of Tokyo.

34. W3C (2004) RDF Vocabulary Description Language 1.0: RDF Schema. Available:http://www.w3.org/TR/rdf-schema/. Accessed 23 December 2013.

35. W3C (2004) OWL Web Ontology Language Overview. Available:http://www.w3.org/TR/owl-features/. Accessed 23 December 2013.

36. W3C (2007) SPARQL Query Language for RDF. Available:http://www.w3.org/TR/rdf-sparql-query/. Accessed 23 December 2013.

37. W3C (2013) SPARQL 1.1 Overview. Available:http://www.w3.org/TR/sparql11-overview/. Accessed 23 December 2013.

38. Bendiken A, Lavender B, Kellogg G (2010) RDF.rb: Linked Data for Ruby. Available:http://rdf.rubyforge.org/. Accessed 23 December 2013.

39. Python (2008) RdfLibraries - Python Wiki. Available:https://wiki.python.org/moin/RdfLibraries. Accessed 23 December 2013.

40. Backett D (2005) Redland librdf RDF API Library - Perl Interface. Available:http://librdf.org/docs/perl.html. Accessed 23 December 2013.

41. Huynh D (2010) google-refine - Google Refine, a power tool for working with messy data (formerly Freebase Gridworks) - Google Project Hosting. Available:http://code.google.com/p/google-refine/. Accessed 23 December 2013.

42. TheOpenRefineDevelopmentTeam (2012) OpenRefine. Available:https://github.com/OpenRefine. Accessed 23 December 2013.

43. Verborgh R, Max W (2013) Using OpenRefine | Packt Publishing. 1st New ed. Packt Publishing. p. Available:http://www.amazon.com/Using-OpenRefine-Ruben-Verborgh-ebook/dp/B00F3VNPN0/ref=sr_1_1?ie=UTF8&qid=1387794840&sr=8-1&keywords=using+openrefine.

44. Maali F (2011) Getting to the Five-Star: From Raw Data to Linked Government Data. National University of Ireland Galway. Available:https://docs.google.com/file/d/0B-jAmahMEbtJYjQ4OGIyM2EtMmZjOC00ODYwLWI2YTUtMDJkN2YwZTA0N2My/edit?hl=en.

45. Maali F, Cyganiak R, Peristeras V (2011) Re-using Cool URIs: Entity Reconciliation Against LOD Hubs. In Proceedings of the Linked Data on the Web Workshop 2011 (LDOW2011) Workshop at WWW2011. Available:http://events.linkeddata.org/ldow2011/papers/ldow2011-paper11-maali.pdf.

46. Maali F, Cyganiak R (2011) GRefine RDF Extension. Available:http://refine.deri.ie/.

Accessed 23 December 2013.

47. Backett D (2013) raptor rdf syntax library. Available:http://librdf.org/raptor/. Accessed 23 December 2013.

48. Minack E (2010) RDF2RDF - Converts RDF from any format to any. Available:http://www.l3s.de/~minack/rdf2rdf/.

49. Yamamoto Y (2013) ConvRDF. Available:https://github.com/dbcls/ConvRDF. Accessed 23 December 2013.

50. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011: bar009. Available:http://database.oxfordjournals.org/content/2011/bar009. Accessed 11 December 2013.

51. Berman HM (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242. Available:http://nar.oxfordjournals.org/content/28/1/235.full. Accessed 22 December 2013.

52. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10: 980. Available:http://dx.doi.org/10.1038/nsb1203-980. Accessed 22 December 2013.

53. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, et al. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr 54: 1078–1084. Available:http://www.ncbi.nlm.nih.gov/pubmed/10089483. Accessed 8 August 2013.

54. Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, et al. (2011) PDBe: Protein Data Bank in Europe. Nucleic Acids Res 39: D402–10. Available:http://nar.oxfordjournals.org/content/39/suppl_1/D402.long. Accessed 15 August 2013.

55. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, et al. (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res 40: D453–60. Available:http://nar.oxfordjournals.org/content/40/D1/D453.long. Accessed 7 December 2012.

56. Shin J-M, Cho D-H (2005) PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. Nucleic Acids Res 33: D238–41. Available:http://nar.oxfordjournals.org/cgi/content/abstract/33/suppl_1/D238. Accessed 23 August 2011.

57. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res 41: D483–9. Available:http://nar.oxfordjournals.org/content/41/D1/D483.long. Accessed 21 May 2013.

58. Franz (2013) Semantic Technologies AllegroGraph Triple Store RDF Web 3.0

Database,optimized SPARQL Query engine, Prolog and RDFS+ reasoner. Available:http://www.franz.com/agraph/allegrograph/. Accessed 23 December 2013.

59. Software O, OpenLink (2006) OpenLink Public Wiki. Available:http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/. Accessed 22 December 2013.

60. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19: Unit 19.10.1–21. Available:http://www.ncbi.nlm.nih.gov/pubmed/20069535. Accessed 21 May 2013.

61. ckan - The open source data portal software (2013). Available:http://ckan.org/.

62. Brickley D, Libby M (2010) FOAF Vocabulary Specification. Available:http://xmlns.com/foaf/spec/.

63. TheDublinCoreMetadataInitiative (2008) Expressing Dublin Core metadata using the Resource Description Framework (RDF). Available:http://dublincore.org/documents/dc-rdf/. Accessed 24 December 2013.

64. Cox RS, Nishikata K, Shimoyama S, Yoshida Y, Matsui M, et al. (2013) PromoterCAD: Data-driven design of plant regulatory DNA. Nucleic Acids Res 41: W569–74. Available:http://nar.oxfordjournals.org/content/41/W1/W569.long. Accessed 22 December 2013.

65. RIKEN (2011) Link and Publish your data｜Open data sharing & Download｜LinkData. Available:http://linkdata.org/. Accessed 24 December 2013.

66. OpenRefine (2012) OpenRefine. Available:http://openrefine.org/. Accessed 24 December 2013.

67. W3C (2013) RDF 1.1 N-Triples. Available:http://www.w3.org/TR/n-triples/. Accessed 23 December 2013.

68. W3C (2004) RDF/XML Syntax Specification (Revised). Available:http://www.w3.org/TR/rdf-syntax-grammar/. Accessed 23 December 2013.

69. W3C (2013) Turtle. Available:http://www.w3.org/TR/turtle/. Accessed 23 December 2013.

70. Rubin DL, Noy NF, Musen MA (2007) Protege: a tool for managing and using terminology in radiology applications. J Digit imaging Off J Soc Comput Appl Radiol 20 Suppl 1: 34–46. Available:http://www.springerlink.com/content/d066t61440n54g67/. Accessed 17 July 2012.

71. TheUniProtConsortium (2013) UniProt Core Ontology. Available:http://www.uniprot.org/core/.

72. PDBj (2011) PDBx ontology. Available:http://rdf.wwpdb.org/schema/pdbx-v40.owl. Accessed 23 December 2013.

73. Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, et al. (2013) EDAM: an ontology of

bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29: 1325–1332. Available:http://bioinformatics.oxfordjournals.org/content/29/10/1325. Accessed 31 May 2013.

74. Bolleman J (2012) FALDO. Available:https://github.com/JervenBolleman/FALDO. Accessed 23 December 2013.

75. Qi D, King R, Hopkins A, Bickerton R, Soldatova L (2010) An ontology for description of drug discovery investigations. J Integr Bioinform 7: 126. Available:http://journal.imbio.de/articles/pdf/jib-126.pdf.

76. Callahan A, Cruz-Toledo J, Dumontier M (2013) Ontology-Based Querying with Bio2RDF's Linked Open Data. J Biomed Semantics 4 Suppl 1: S1. Available:http://www.jbiomedsem.com/content/4/S1/S1. Accessed 23 December 2013.

77. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, et al. (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39: W541–5. Available:http://nar.oxfordjournals.org/cgi/content/abstract/39/suppl_2/W541. Accessed 21 March 2012.

78. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, et al. (2007) Dbpedia: A nucleus for a web of open data. Semant Web 4825: 722–735. Available:http://www.springerlink.com/index/rm32474088w54378.pdf. Accessed 4 November 2010.

79. Cyganiak R, Anja J (2011) The Linking Open Data cloud diagram. Available:http://lod-cloud.net. Accessed 23 December 2013.

80. Creative Commons — Attribution 3.0 Unported — CC BY 3.0 (2002). Available:http://creativecommons.org/licenses/by/3.0/.

81. Coppermine Photo Gallery (2003). Available:http://coppermine-gallery.net/. Accessed 24 December 2013.

82. Amazon (2002) Amazon Web Services, Cloud Computing: Compute, Storage, Database. Available:http://aws.amazon.com/. Accessed 23 December 2013.

83. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. "Protein Eng Des Sel 8: 127–134. Available:http://peds.oxfordjournals.org/content/8/2/127.short. Accessed 26 February 2013.

84. ChemAxon (1998) ChemAxon – cheminformatics platforms and desktop applications. Available:http://www.chemaxon.com/. Accessed 23 December 2024.

85. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I (2013) InChI - the worldwide chemical structure identifier standard. J Cheminform 5: 7. Available:http://www.jcheminf.com/content/5/1/7. Accessed 22 December 2013.

## ACKNOWLEDGMENTS

The research activities in my doctorate course were successful due to the excellent supervision by Professor Kentaro SHIMIZU. I also received help from Associate Professor Shugo NAKAMURA to overcome several difficulties. The “miracle wizard,” Assistant Professor Kazuya SUMIKOSHI, helped me overcome serious concerns related to the computer system used in this research. We would like to offer our sincere thanks for his coaching.

We are also grateful to Professor Hideaki TAKEDA, Associate Professor Itsuki OHMUKAI, Fumihiro KATO, and Seiji KOIDE, who is a research fellow at the National Institute of Informatics and who provided support and counsel regarding both the software and hardware perspectives for database construction and prepublication paper writing.

To all staff and colleagues who gave selfless support during my life as a researcher at the University of Tokyo, I want to say “Thank you very much.” I aim to repay your kindness by continuing to contribute to research in the future.

Yusuke KOMIYAMA dedicates his doctoral dissertation to his grandfather Hirosuke KOMIYAMA, who provided valuable support for Yusuke’s doctoral dissertation until the very end of his struggle with cancer. Mr. KOMIYAMA was an engineer who had superior skills and who was a manager with excellent judgment. Yusuke’s engineering mind was influenced by Hirosuke even after his death. In addition, Yusuke mentions his family “My job was successful because of you.”

This work was supported by JSPS KAKENHI Grant Number 12J07771 and 23300109.

This work was supported by Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

This research was supported by JST, NBDC.