Data Integration of Bioinformatics Database Based on Web Services

Yuelan Liu, Jian hua Wang
College of Computer, Harbin Normal University
Intelligent Education Information Technology Emphases Lab of Heilongjiang
Harbin, China
liuyuelan126@126.com

Yuefan Liu
Software Institute of Dalian Jiaotong University
Dalian, China
liuyuefan@163.com

Zhenwu Tan
The Fifth Survey and Design Institute (Group) CO., LTD. Of CRCC
Harbin, China
tzw_2603@sina.com

ABSTRACT: *With the development of human genome projects (HGP) in the world, a mass of genetic information is generated. Now there are hundreds of different kinds of important bioinformatics databases in the world. How to unify the bioinformatics database from different countries has become an important issue in Bioinformatics. In this paper, we present a data integration program based on Web Services of heterogeneous bioinformatics databases. The key technology of bioinformatics data integration has been researched and designed as well.*

Keywords: Web services, heterogeneous database, data integration, bioinformatics

Received: 2 April 2009, Revised 27 April 2009, Accepted 5 May 2009

1. Introduction

In recent years, with the development of human genome project(HGP) in the world, descendibity code that decodes the human beings and model biography has been an important subject in biographical field, while produced huge genome information. It is absolutely necessary content for the study of human genome to analyze these information and it also promotes the production and development of the bioinformatics.

Along with the large data accumulation of biological experiments, hundreds of databases of Bioinformatics are formed, such as the three major international databases of the nucleic acid: Genbank, the European Molecular Biology Laboratory (EMBL) database and the DNA Data Base of Japan (DDBJ) [1][2][3]. Bioinformatics on a wide range of databases, can be roughly divided into 4 major categories, namely: the genome database, the first-class structure of a protein and nucleic acid sequence database, three-dimensional structure database of the biological macromolecules (mainly proteins) and the secondary class database on the basis of the 3 above-mentioned category database for information and documentation. According to their respective targets, they collect and collate data of biological experiments and provide relevant data query, data-processing services.

At present, the bio-information database's owner only develops private system to provide users with data query and analysis services. Such as NCBI (National Center for Biotechnology Information. develops Entrez database query system which is used on Genbank. European Molecular Biology Laboratory develops SRS system .The key point is how to share those heterogeneous databases and make a common query platform for users.

In this paper, we propose a data integration program based on Web Services of heterogeneous bioinformatics databases.

2. Web Services Technology

Web services describes a standardized way of integrating Web-based applications using the XML(extensible markup language), SOAP(Simple Object Access Protocol), WSDL(Web Services Description Language) and UDDI (Universal Discover, Description and Integration) open standards over an Internet protocol backbone[4]. XML is used to tag the data, SOAP is used to transfer the data, WSDL is used for describing the services available and UDDI is used for listing what services are available. Used primarily as a means for businesses to communicate with each other and with clients, Web services allow organizations to communicate data without intimate knowledge of each other's IT systems behind the firewall.

Unlike traditional client/server models, such as a Web server/Web page system, Web services do not provide the user with a GUI. Web services instead share business logic, data and processes through a programmatic interface across a network. The applications interface, not the users. Developers can then add the Web service to a GUI (such as a Web page or an executable program) to offer specific functionality to users.

Web services allow different applications from different sources to communicate with each other without time-consuming custom coding, and because all communication is in XML, Web services are not tied to any one operating system or programming language. For example, Java can talk with Perl , Windows applications can talk with UNIX applications.

Web services are sometimes called application services.

2.1. Web Services Architecture

Web Services architecture is as followed in Figure 1.

Web Services consists of three components. Service providers: to provide services to register in order to make the service available; Service Agent: services exchange, the agent between service providers and the service requesters; service Requester: request services from service agent; use these services to create applications.

Web Services operations, including 3 operations. Publish / Unpublish: provider releases the registration of the registration service or releasing (removing) these services; Find: the requester asks service agents to find the implementation of the operation; the service requester describes the services; services agent distributes the matched result. Bind: Bind the service requester and service providers, so the requester can access and ask provider's services.

2.2. Web Services Standards and Related Technology

SOAP (Simple Object Access Protocol) is an XML-based communication protocol for information exchanges which is used in a decentralized and distributed network environment. Under this protocol, software components or applications can communicate through the standard HTTP protocol. The goal of the design is simplicity and scalability, which contributes to the process and a large number of heterogeneous interoperability among platforms, so that the existing applications can be used widely.

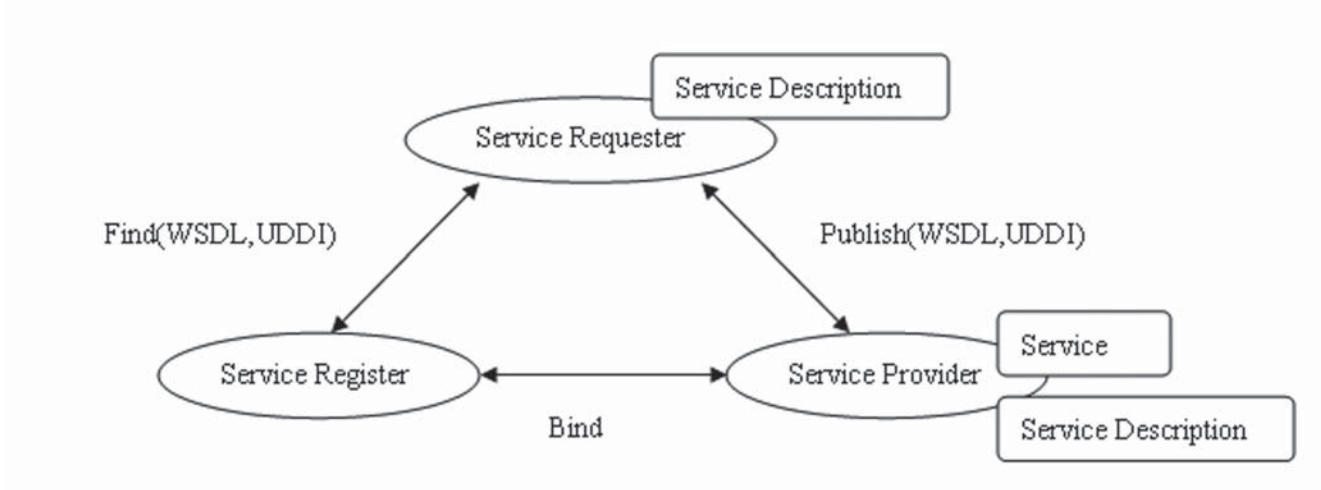

Figure 1. Web Services Architecture

Web services description language (WSDL) is an XML syntax, which provides a way of description for service providers, and it is a basic format of Web Services requests based on different protocols or encoding forms. WSDL services will be defined as a collection of network endpoints, or we can say a collection of ports.

UDDI(Universal Description, Discovery and Integration), A Web-based distributed directory that enables businesses to list themselves on the Internet and discover each other, similar to a traditional phone book's yellow and white pages.

UDDI is a directory service where companies can register and search for Web services.

- UDDI stands for Universal Description, Discovery and Integration
- UDDI is a directory for storing information about web services
- UDDI is a directory of web service interfaces described by WSDL
- UDDI communicates via SOAP
- UDDI is built into the Microsoft .NET platform

2.3. Application of Web Services

Web Service is a dynamic, integrated program. All services can be dynamically found, bound and used through the UDDI standard. It is easy to adapt to changes in the system and it has increased the flexibility and scalability of the system and overcome the shortage of using RPC (Remote Procedure Call) and API(Application Program Interface) integration technology, which meet the requirements of the loose coupling. Web Services has the following advantages in solving the traditional problems:

It is real cross-platform, solving the problem of the interoperability which the traditional components of middleware technology can not solve.

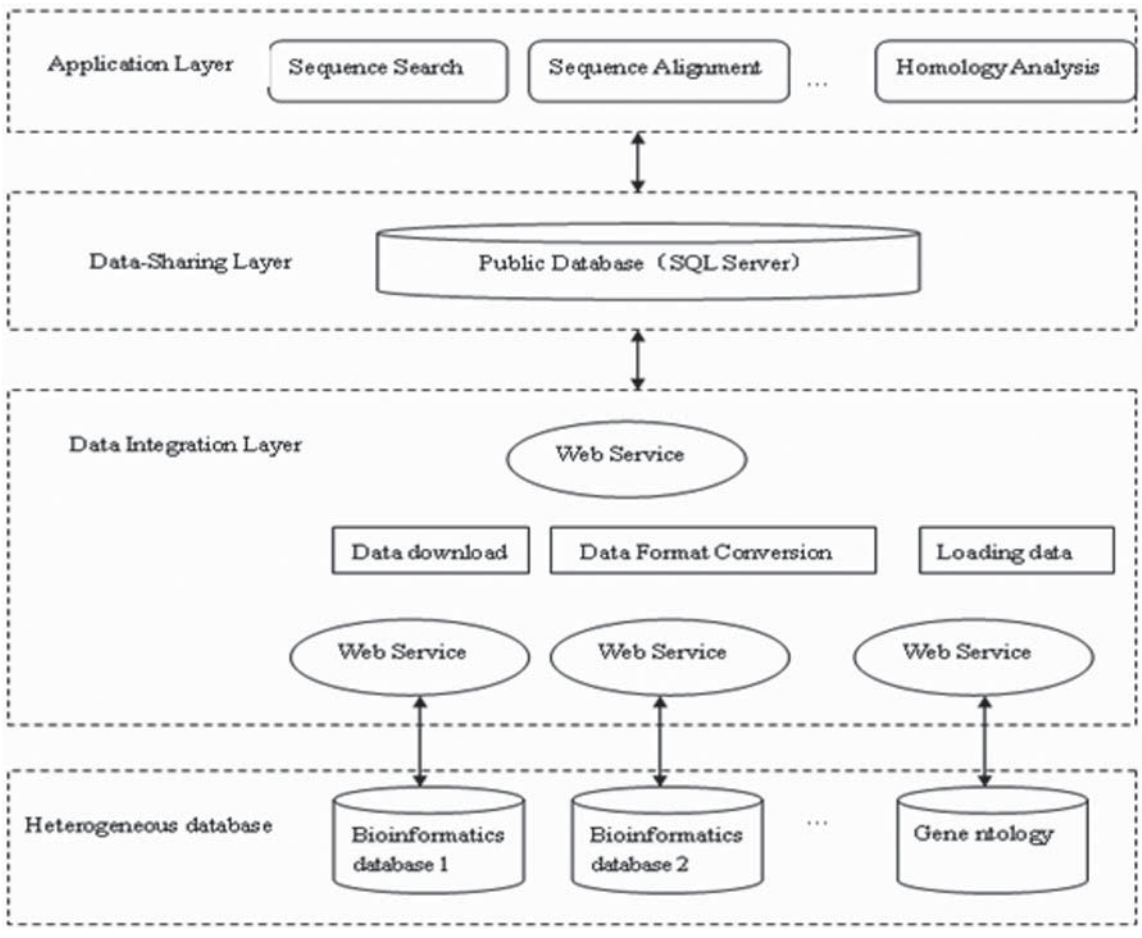

Figure 2. Data Integration Architecture of Heterogeneous Database

The data is loosely coupled. Data service providers will be able to choose data services provided by outside, and they can also provide only part of the data. It is flexible and easy to change. The data can be used at any time. Web Services can change the traditional point-to-point data integration thoroughly in the integrated approach.

By using Web Services we will develop data services even more quick, easy and low cost. In the long run, Web Services applications achieve the function or method integration among the application programs. At the same time, there are a lot of development tools to support Web Services, such as: NET, Delphi, C + + Builder, J2EE, and so on, and all of them have offered the conditions for the development of Web Services.

Web services are widely used on a lot of domains because of those advantages. Aiming to the characteristic of heterogeneous database of multiple application system in universities, the authors in reference [5] proposed the data integration architecture of heterogeneous database based on Web Services. But they didn't research the key technology of data integration.. It can only unify heterogeneous data format but not solve the data languages issue.

We have put forward a number of heterogeneous bioinformatics database integration program based on Web Services. We research the key technology of Bioinformatics data integration and solve the different language issue by Gene Ontology.

3. Heterogeneous databases of Bioinformatics integration program

Heterogeneous bioinformatics database integration is process to share or integrate the database from the Internet users and a variety of bio-information database to create a more useful application of bio-information database. Heterogeneous data integration is the key and the bases to Bioinformatics and databases. it has two implications: interactive, sharing. From the interaction point of view, it is to deal with the interactive data from one data source to another, thus

it is the basement of the process of the system. In the view of sharing the data, it is to collect the different sources, formats, characteristics of the heterogeneous data to focus on a series of subject-oriented, integrated, relatively stable, historical data for the system so as to provide a full range of data sharing. Bioinformatics database data integration architecture is in Figure2.

3.1. Heterogeneous database layer

It refers to a variety of bio-information database. These operation environment, as well as data formats, semantic may be very different.

3.2. Data integration layer

Heterogeneous database data integration layer, that is, the Web Services layer, is the key to this paper. It mainly downloads biological information coming from different databases in XML format.

XML stands for extensible markup language. It was designed to transport and store data. XML tags are not predefined. You must define your own tags. It is designed to be self-descriptive.

With the XML technology's development, XML is used by most of the main bioinformatics database such as Genbank, DDBJ, EMBL, Swiss-Prot, PDB, GO[3][6]. All of them provide data output and download in XML format. Then the XML format data is mapped into the public database. The system's public database uses relation-type database SQL Server. Data integration layer links the distribution of heterogeneous databases and shared databases. And it uses the heterogeneous nature, distribution, complexity and the data structure of the DBMS. Data synchronization achieves the synchronization between heterogeneous databases and public databases.Web Services make data communication, interaction and sharing between heterogeneous databases and public database becoming possible. It is based on middleware design ideas of XML and Web Services to ensure the general and high reusability of data integration layer.

Data integration steps:

(1) configure database sharing according to the distributing heterogeneous database structure information and the sharing requirement. grab sharing data from heterogeneous database automatically.
(2) transfer the data into XML document.
(3) send XML document to data integration server via Web Service. map the XML document into sharing database[7].

```
<?xml version=”1.0”?>
<!DOCTYPE INSDSet PUBLIC “-//NCBI//INSD INSDSeq /EN” http://www.ncbi.nlm.nih.gov/dtd/INSD _INSDSeq.
dtd >
<INSDSet>
<INSDSeq>
  <INSDSeq_locus>X87617</INSDSeq_locus>
  <INSDSeq_length>1497</INSDSeq_length>
  <INSDSeq_strandedness>double</INSDSeq_strandedness>
  <INSDSeq_moltype>DNA</INSDSeq_moltype>
  <INSDSeq_topology>linear</INSDSeq_topology>
  <INSDSeq_division>BCT</INSDSeq_division>
  <INSDSeq_update-date>21-JUL-1997</INSDSeq_ update-date>
   <INSDSeq_create-date>28-JUN-1996</INSDSeq_create-date>
  <INSDSeq_definition>Actinomycete (genus unknown) 16S ribosomal RNA</INSDSeq_definition>
  <INSDSeq_primary-accession>X87617</INSDSeq_primary-accession>
  <INSDSeq_accession-version>X87617.1</INSDSeq_accession-version>
  <INSDSeq_other-seqids>
    <INSDSeqid>emb|X87617.1|</INSDSeqid>
    <INSDSeqid>gi|1418292</INSDSeqid>
  </INSDSeq_other-seqids>
  <INSDSeq_keywords>
    <INSDKeyword>16S rRNA</INSDKeyword>
    <INSDKeyword>ribosomal RNA</INSDKeyword>
  </INSDSeq_keywords>
  <INSDSeq_source>Actinomycetaceae</INSDSeq_source>
  <INSDSeq_organism>Actinomycetaceae</INSDSeq_organism>
   <INSDSeq_taxonomy>Bacteria;Actinobacteria; Actinobacteridae;Actinomycetales; ctinomycineae</INSD-
Seq_taxonomy>
  <INSDSeq_references>
    <INSDReference>
      <INSDReference_reference>1</INSDReference_reference>
      <INSDReference_authors>
        <INSDAuthor>Radajewski,S.M.</INSDAuthor>
        <INSDAuthor>Blackall,L.L.</INSDAuthor>
        <INSDAuthor>Duxbury,T.</INSDAuthor>
      </INSDReference_authors>      <INSDReference_journal>Unpublished</INSDReference_journal>
    </INSDReference>
    <INSDReference>
      <INSDReference_reference>2</INSDReference_reference>
      <INSDReference_position>1..1497</INSDReference_position>
      <INSDReference_authors>
        <INSDAuthor>Radajewski,S.A.</INSDAuthor>
      </INSDReference_authors>
      <INSDReference_title>Direct Submission</INSDReference_title>
      <INSDReference_journal>Submitted (24-MAY-1995) S.A. Radajewski, The University of Sydney, Dept
of Micriobiology, N.S.W., 2006 Sydney, AUSTRALIA</INSDReference_journal>
    </INSDReference> . . . . . .
```

Figure 3. Example XML Document of a DNA sequence

For example, use one DNA sequence XML document, which belongs to GENBAK’s nucleic acid sequence database (see Figure 3)

We grab some major data and map into three tables of public database. See the results in table 1, table 2 and table 3.

GI	locus	length	moltype	topology	division	update-date	create-date	…	sequence
1418292	X87617	1497	DNA	linear	BCT	21-JUL-1997	28-JUN-1996	…	tgatcctggctcaggacgaacgctggcggt…

Table 1. ACCDB

GI	reference_id	author	journal	title
1418292	1	Radajewski,S.M. Blackall,L.L Duxbury,T.	Unpublished	
1418292	2(base 1 to 1497)	Radajewski,S.A.	Submitted (24-MAY-1995) S.A. Radajewski, The University of Sydney, Dept of Micriobiology, N.S.W., 2006 Sydney, AUSTRALIA	Direct Submission

Table 2. REFERENCE

GI	feature_key	location	qualname	qualvalue
1418292	rRNA	1..1497	product	16S ribosomal RNA
1418292	source	1..1497	organism	Actinomycetaceae

Table 3. FEATURE

3.3 Public database layer

We adopt large SQL Server database to create a Public database platform. Public database stores the data which need to be integrated and shared.

As there are different kinds of bio-information and the data of them are very confusing in biology meaning, one significant problem facing the integration the data is incompatible meaning. Biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research. This is hampered further by the wide variations in terminology that may be common usage at any given time, which inhibit effective searching by both computers and people.

This system uses Gene Ontology as the different definition's integration method[8]. GO(Gene Ontology) is established by Gene Ontology Consortium, aimed at creating a word and phrases standard, which can be widely used in viable species and can identify and describe the function of genome and protein.

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.

Figure 4. Complement Sequence tools

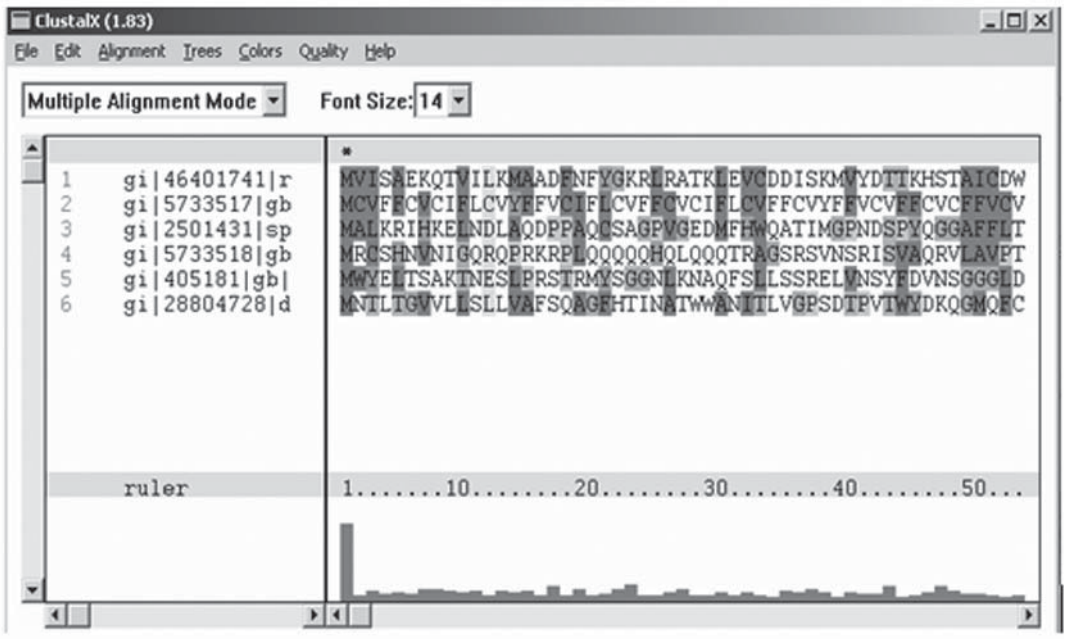

Figure 5. Multiple Sequence Alignment of Clustal

The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, development of tools that facilitate the creation, maintenance and use of ontologies.

Through the access and storage of the needed data from heterogeneous database, The Gene Ontology (GO) project can provide real-time data for the top floor - a unified application layer to ensure high-availability systems.

GO and NCBI GOA provide related database mapping table for GO.

The target to establish the relations between biology database and GO is to make Gene Ontology become the biology base in relation-typed database to establish the relationship between different data, so that we can finally integrate different database in the meaning of biology.

3.4 Uniform application layer

Based on bioinformatics database, we can provide users with a variety of applications, such as DNA and protein sequence search, the alignment right, homology analysis. Figure 4 shows the Complement Sequence tools , Figure 5 shows the Clustal Multiple Sequence Alignment .

4. Conclusion

We can use Web Services technology to realize the integration of the bioinformatics database and set up a unified platform for the users. It will effectively solve the problems in using a variety of heterogeneous bioinformatics database.

References

[1] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, David L. (2005,). Wheeler, “GenBank,” Nucleic Acids Res (Database issue), p. 34-38.

[2] Kulikova T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R. et al, (2004,). The EMBL Nucleotide Sequence Database, Nucleic Acids Res, 32, p.27–30.2

[3] Miyazaki S., Sugawara, H., Gojobori, T., Tateno, Y (2003). DNA Data Bank of Japan (DDBJ) in XML, Nucleic Acids Res, , 31, p.13–1656
[4] David B, Hugo H, Francis M, et al. Web Services Architecture: W3C Working Group Note 11 February 2004.[EB/OL]. http: //www.w3.org/TR/ws-arch/.
[5] NZhennao Cai, Yanchao Zhang, Li Wang (2008). Reserch on Data Integration of heterogeneous Database Based on Web Services, Wenzhou University Journal of Natural Science, Vol 29, No 4, p .25-29.
[6] Frederic Achard, Guy Vaysseix, Emmanuel Barillot. (2001). XML, bioinformatics and data integration, Bioinformatics Review, Vol. 17 no.2 p. 115-125.
[7] Yuelan Liu, Yanqing Zheng (2006). Bioinformatics data exchange on XML, Information Science, , p.52-54.
[8] The Gene Ontology Consortium (2000). Gene Ontology:tool for the unification of biology. Nature Genet.; 25: 25-29.