Genomics & Informatics Vol. 2(3) 142-146, September 2004

Web Services Based Biological Data Analysis Tool

Min Kyung Kim[1], Yo Hahn Choi[2], Seong Joon Yoo[2] and Hyun Seok Park[1,3]*

[1]Department of Computer Science and Engineering, Ewha University, Seoul 120-750, Korea, [2]School of Computer Science and Engineering, Sejong University, Seoul 143-747, Korea, [3]Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea

Abstract

Biological data and analysis tools are accumulated in distributed databases and web servers. For this reason, biologists who want to find information from the web should be aware of the various kinds of resources where it is located and how it is retrieved. Integrating the data from heterogeneous biological resources will enable biologists to discover new knowledge across the specific domain boundaries from sequences to expression, structure, and pathway. And inevitably biological databases contain noisy data. Therefore, consensus among databases will confirm the reliability of its contents. We have developed **WeSAT** that integrates distributed and heterogeneous biological databases and analysis tools, providing through Web Services protocols. In **WeSAT**, biologists are retrieved specific entries in SWISS-PROT/EMBL, PDB, and KEGG, which have annotated information about sequence, structure, and pathway. And further analysis is carried by integrated services for example homology search and multiple alignments. **WeSAT** makes it possible to retrieve real time updated data and analysis from the scattered databases in a single platform through Web Services.

Keywords: database integration, platform, web services

Introduction

Various databases contain a different subset of biological knowledge for example species such as Flybase and Wormbase (Fly consortium, 2003, Harris *et al.*, 2004) versus sequence and structure such as Genbank and PDB (Benson *et al.*, 2004; Westbrook *et al.*, 2004). Researchers who want to answer questions that span domain boundaries should be surfing the web. For example, structural information will be useful for the identification of interaction sites between interacting proteins, or orthologous protein data can be used in function prediction, and so on. Not only the quantity but also the quality of information will be increasing during the process of integration. Due to the characteristics of biological data, biological databases inevitably contain noisy data. Through the collected data from different sources, the reliability could be voted from the redundancy.

Thus, rapid growth of databases in the public biological domains makes it difficult to maintain and access updated data. On the other hand, data integration in the biological domain has been attempted through two different approaches, which are in striking contrast to another: federation and data warehousing (Stein, 2003).

Federation approach means one database would query the others in the terms understood by the others. It has advantages in aspect of up-to-date and maintenance. But its availability and reliability was not sufficient, when dealing with alteration of schema or address in data resources. The stability and persistence of URLs published in MEDLINE is not guaranteed and ~37% of these URLs and ~64% of ftp were not consistently available (Wren, 2004).

Data warehousing integrates diverse data to the one root with global schema. Still, data warehousing method provides an effective solution for extraction, usage and analysis of the biological information. Global schema design, which is based on current available information, is not feasible to agree without interruption. For example, interaction data not only increase the entity (*i.e.* protein interact with other partners) but also the content of information (i.e. which domain, condition, and so on).

ISYS is a CORBA based biological data analysis tool (Siepel *et al.*, 2001). ISYS provides an environment of integrated services such as sequence viewer and Z it does not support automated workflow method. Still, result of each service runs slowly and is not fed into next service module automatically.

Web Services is one of the most recent integration methods, which provide data through standard user

*Corresponding author: E-mail neo@ewha.ac.kr, Tel +82-2-3277-2831, Fax +82-2-3277-2306
Accepted 16 August 2004

interface, recording their up-to-date (Stein, 2002). It makes it possible to automatically up-to-date secondary databases which are distributed all over the world. In field of bioinformatics research, Web Services has been used on several web sites such as KEGG (Kanehisa *et al*, 2004), Ensemble (Cuff *et al*, 2004), Flybase, and etc. In aspect of secondary databases, the ability to access any data across multiple disparate databases and up-to-date data from different data sources at run-time is characteristic feature of Web Services approach.

Currently available integrated environments by Web Services protocols are myGRID (Stevens *et al.*, 2003) and BioMoby (Wilkinson *et al.*, 2002). myGRID is designed for data or service provider, which build applications for biologist. The scope of BioMoby is confined to service description, discovery, transaction, and simple input/output object type. Therefore, it is not a usable system for biologists.

As shown in Table 1, WeSAT is first trial to provide Web Services protocol integrated system to biologist, which has advantage (1) availability (2) real time updated data providing (3) analysis tool integrating. We have integrated several biological data servers including KEGG, DDBJ, EMBL, and PDB (Table. 2). Users can access sequence, structure, pathway and literature databases and several tools such as Blast, ClustalW, and TxSearch (Table 3). In addition, data accessed from one transaction can be automatically fed into next service modules.

Table 1. Integration systems that provide databases and/or tools in a single platform are listed.

Integration System	Method	Integrating Objects	User
Kleisli	Federated	Data	Biologist
ENSEMBL	Data warehousing	Data	Biologist
ISYS	CORBA	Data	Biologist
myGRID	Web Services	Data and tools	System developer
BioMOBY	Web Services	Data and tools	System developer
WeSAT	Web Services	Data and tools	Biologist

Table 2. List of WeSAT accessible databases.

External Module	Accessible Databases	Data format
KEGG	KEGG, EMBL	Text
DDBJ	DDBJ, SWISS-PROT, PDB, PIR	Text
EMBL	EMBL, PUBMED	XML
PDB	PDB	XML

Methods

System Architecture

Any program is callable by another program across the web in a way that is platform-independent, language-independent, and object model-independent. The Web Services architecture describes three roles: service provider, service requester, service registry. Three basic operations are: publishing, finding, and binding. These operations use next generation infrastructure standards: XML, SOAP, WSDL, and UDDI.

First, service provider publishes their services, which are described by WSDL (Web Services Description Language), at service registry. UDDI (Universal Description Discovery and Integration) enables enterprises to quickly and dynamically discover and invoke Web Services both internally and externally. Next, service requester could find their services in service registry. After services are found once in registry, this service could be found directly from service provider via SOAP (Simple Object Access Protocol) and these processes are called binding. SOAP is an XML notation for describing how messages are assembled and transmitted over HTTP between service requesters (clients) and service providers (servers). Web Services is methods of providing their data to the requester in unrestricted manner according to their schema or written language.

Figure 1 shows overall architecture of **Web** Service based **A**nalysis **T**ool (WeSAT). **WeSAT** has three components: External Application module, Local Web Services search module and Graphic Viewer module.

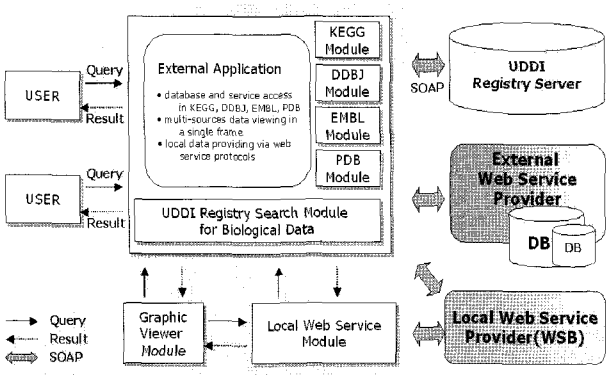

Fig. 1. System Architecture of **WeSAT.**

External Application and UDDI registry search module

External databases are linked through the external application module (KEGG, DDBJ, EMBL, and PDB) as

shown Table 2 (Sugawara *et al.*, 2003). Especially, DDBJ Web Services provides their own data and other services such as Blast, ClustalW, ExClustalW, Fasta, GetEntry, Gib, Gtop, PML, SRS, and TxSearch which are described by WSDL description. XEMBL contains annotated information of genes and BQS is literature retrieval system in EMBL. PDB module provides structural information of protein in the format of XML, PDBj-ML. Through the external application module in **WeSAT**, we can access KEGG, EMBL, DDBJ, SWISS-PROT, PDB, PIR, and PUBMED data. By using these soap servers, **WeSAT** is able to directly access biological objects from data sources, irrespective to the change of the data formats and addresses. UDDI registry search module will be added to the system and expected to perform the finding of updated Web Services in UDDI registry.

Local Web Services module

Besides the integration of provided Web Services, **WeSAT** also provides Web Services of specific methods. These methods make it possible to access categorized information in SWISS-PROT, EMBL, and Genbank data using Biojava. Through this module, users can pick the specific information, for example, FT or DR lines in SWISS-PROT. It can also be used as parsers, which could extract the desired fields in an entry for later analysis through a simple request.

Its implementation is based on SOAP protocols contained in JWSDP1.3 (Java Web Services Developer Pack 1.3) of Microsystems and web sever engine (TOMCAT4.1). By exchanging the WSDL document, we service this type of local Web Services to the service requester. Figure 2 shows a part of WSDL descriptions for this service.

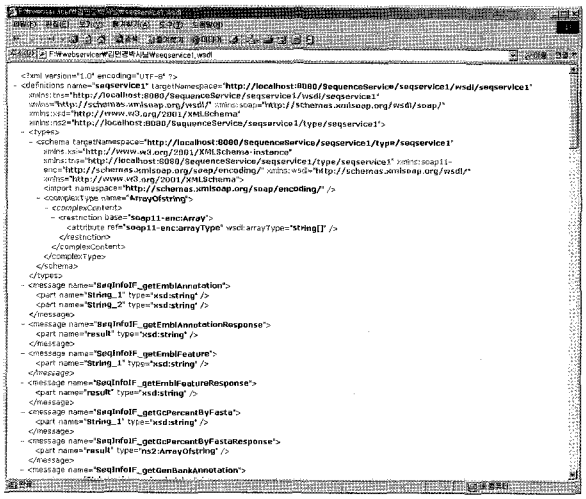

Fig. 2. WSDL description for local Web Services of **WeSAT.**

Graphic viewer module

Currently available SOAP data types are simple, complex, and binary, which could back up not only text but also graphic information. However, web serviced data types in the biological domain are limited in text types nowadays.

Usually biological data are understood more easily when graphical information such as pathway image in KEGG and genome browser for genome map data is provided together. For this reason, we developed graphic viewer module for pathway data. In **WeSAT**, KEGG pathway data are presented text and graphical mode that provided through graphic viewer mode or direct link via map URL as shown in figure 3. This module could be used in interaction network data, which will be added in the near future.

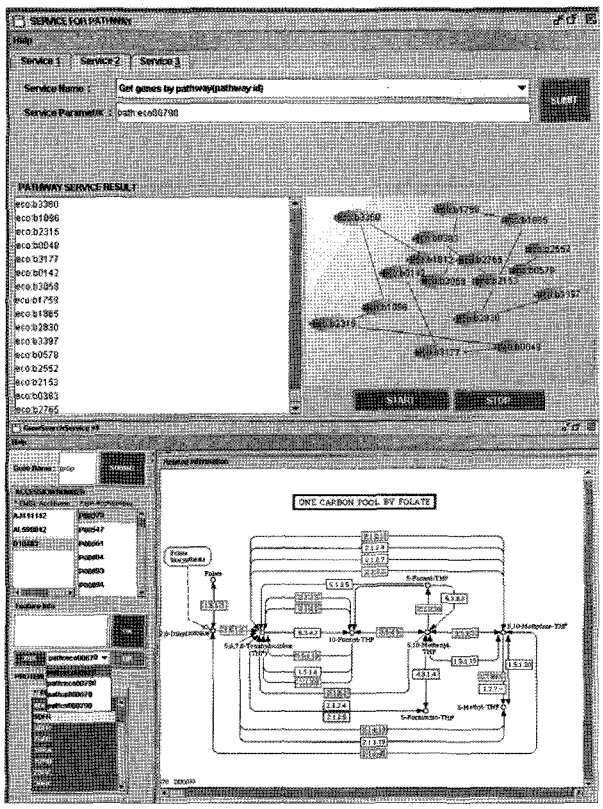

Fig. 3. Graphical view of pathway information in **WeSAT. WeSAT** provides pathway information, which is originated in KEGG, through two different modes. The upper is a snapshot of eco00300 path through the graphic viewer module in **WeSAT.** The lower is the other show that is represented as KEGG via URL.

User Interfaces

User Interface of **WeSAT** is shown in Figure 4. Top

menu consists of external/local/search service. External service menu are divided into KEGG, DDBJ EMBL, and PDB Web Services. Local services are category for our own developed web services. Search service menus are proposed to retrieve for UDDI registry.

Through the user interface, clients surfing the information across the subject for example sequence in EMBL, structure in PDB, pathway in KEGG and literature in Medline. Figure 4 illustrates the navigation of KEGG/DDBJ/SWISS-PROT in **WeSAT**.

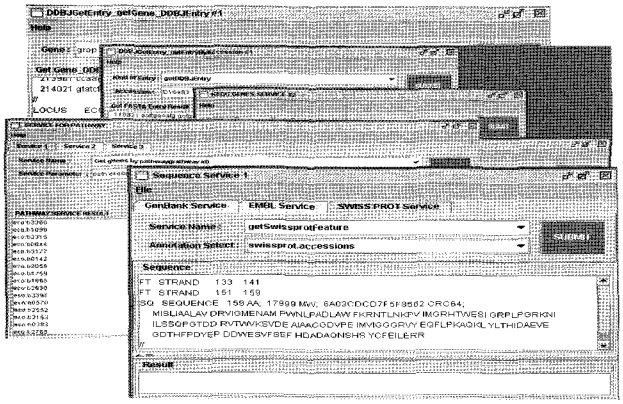

Fig. 4. User interface of **WeSAT** which makes it possible to access to heterogeneous subject databases.

WeSAT currently navigates databases via cross-references. Especially, we use SWISS-PROT that provides diverse cross-references except KEGG, which has cross- reference about SWISS-PROT id. Although only a small portion of proteins are not annotated in SWISS-PROT/ EMBL, they are not linked with across databases in **WeSAT**. For these proteins, local BLAST could be used, because sequences are a usable key for finding the same proteins.

Result and Discussion

WeSAT is a prototype version of web serviced data and service integration that could access multiple databases (KEGG, DDBJ, EMBL, SWISS-PROT, PDB, PIR, and PUBMED) and diverse services (Blast, ClustalW, FASTA, TxSearch, and etc). The client communicates with other servers and services through **WeSAT**. The client of **WeSAT** is biologist and it makes distinguishable from other web serviced systems.

WeSAT access scattered resources in remote condition, therefore maintenance is more convenient than centralized approaches. This integration approach is guaranteed to provide biologists' requirement for updated information and save bioinformaticians' time and effort for maintenance. Web Services approaches make it possible for bioinformatician to integrate biological sources, get over integration difficulties and focus on their effort in knowledge discovery/data mining from the integrated data.

Most distinguishable advantages of Web Services integrations are real time updated and reusability of data. The amount of accumulated biological data keeps growing at an exponential rate and therefore, the databases are continuously updated. Secondary databases are inevitably acted on concert with primary databases. Therefore, there is massive duplication of effort in multiple secondary databases originated from the same databases. Even more, the change of database schema is disastrous for bioinformatician who aggregate data from different sources. Bioinformatician and biologist could get more complete data set and it will be a basis for data mining and knowledge discovery. Although providing data via Web Services protocols is additional work for data provider, it will maximize reusability of data and minimize the obstacle of data assembly. Integration of data based on Web Services will get better as more databases and tools are web serviced. Because, the result of each service is automatically transferred to different analysis module unlike CORBA based system.

Protein or gene name could not be used as an identifier across databases because of synonyms or different types of representations. Identifying the same entity in different databases is not an easy task. For this reason, the function of allowing cross-database queries is not implemented at this moment, but will be implemented in the near future.

Acknowledgement

This work was supported by the Ministry of Information and Communication of Korea under grant number IMT2000-AB-05. **WeSAT** is available at http://www.skyhani.com.

References

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. *Nucleic Acids Res.* 32, D23-26.

Cuff, J.A., Coates, G.M., Cutts, T.J., and Rae, M. (2004). The Ensembl computing architecture. *Genome Res.* 14, 971-975.

FlyBase Consortium. (2003). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.* 31, 172-175.

Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D.,

Bradnam, K., Chan, J., Chen, C.K., Chen, W.J., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E.M., Van Auken, K., Wang, Q., Durbin, R., Spieth, J., Sternberg, P.W., and Stein, L.D. (2004). WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res.* 32, D411-417.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277-280.

Siepel, A., Farmer, A., Tolopko, A., Zhuang, M., Mendes, P., Beavis, W., and Sobral, B. (2001). ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics.* 17, 83-94.

Stein, L. (2002). Creating a bioinformatics nation. *Nature* 417. 119-120.

Stein, L.D. (2003). Integrating biological databases. *Nat. Rev. Genet.* 4, 337-345.

Stevens, R.D., Robinson, A.J., and Goble, C.A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19, i302-304.

Sugawara, H. and Miyazaki, S. (2003). Biological SOAP servers and web services provided by the public sequence data bank. *Nucleic Acids Res.* 31, 3836-3839.

Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H.M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489-491.

Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform.* 3, 331-341.

Wren, J.D. (2004). 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics* 20, 668-672.