INTEGRATION OF HETEROGENEOUS BIOINFORMATICS DATA THROUGH WEB SERVICES

Chaiwat Bootchai[1], Kanda Runapongsa Saikaew[1], Chumpol Ngamphiw[2],
Nuwee Wiwatwattana[3], Sissades Tongsima[2]
[1]*Computer Engineering, Khon Kaen University, Thailand, 40002*
[2]*National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand, 12120*
[3]*Pollution Control Department, Bangkok, Thailand, 10400*
Email: chaiwat.bootchai@gmail.com, krunapon@kku.ac.th, chumpol.nga@biotec.or.th, nuwee.w@pcd.go.th, sissades@biotec.or.th

Abstract

With the advent of comparative genomics, an increasing amount of genomic data from different organisms is to be compared. This paper presents an approach to integrate bioinformatics data from various source databases using web services and a private registry. The goal is to construct an efficient framework that correctly integrates genomic databases via web services and thus allows genomic browsers to connect and clearly presents data based on user criteria. The experimental result shows that for most queries our system with private registry can return correct answers with dramatically less response time than one without.

Keywords: Bioinformatics, Data Integration, Web Services, Private Registry

1. Introduction

After the human genome project was completed, several kinds of genomic data have continuously been deposited in publicly accessible databases. Even though most genomic databases similarly contain DNA sequences as the basic data unit, each database has its own web application interface to interact with users. Such applications are suitable for visualizing data from the sole corresponding database but are not ready for cross analysis, as in comparative genomics. In particular, differences among database schemas prevent web applications from accessing different databases and displaying data via a unified web interface. Thus, scientists have to manually compare data from different databases by repeatedly visiting different web pages several times. Moreover, scientists who wish to share their genomic data with a private group of researchers cannot do with ease. To address the problems, this paper presents an approach to integrate heterogeneous genomic data using web services and a private registry. The proposed protocol has been implemented and its effectiveness has been demonstrated in Section 4. After related work is presented in Section 2, we discuss about our system design in Section 3. Then, we explain the system implementation in Section 4 and experimental results in Section 5. Section 6 describes discussion and future work. Finally, we conclude in Section 6.

2. Related Work

Considerable research efforts have been conducted to introduce the web services and web services registry concepts to genomic databases. For example, DASRegistry [1] presents a public registry for web services that was implemented using DAS protocol [2]. BioMOBY [3] presents the bioinformatics tool services registry but does not support private registry. Unlike DASRegistry, our framework uses a private registry to offer a securely closed group participation. With a private registry, the system can dramatically reduce the response time since the system knows where the requested data is.

3. System Design

Our proposed system employs web services architecture which has the registry at the core to store and list all services for a given user, as shown in Figure 1.

In Figure 1, the provider hosts bioinformatics service and deploying web services. The provider also defines the DAS services and publishes them with the private registry. The private registry hosts the lookup information and descriptions of published services. The registry is primarily responsible for service registration and discovery of the web services. The private registry stores, lists, and manages privileges to help consumers

find and subscribe to the required services. The consumer locates bioinformatics services using the private registry and invokes the required services from the provider.

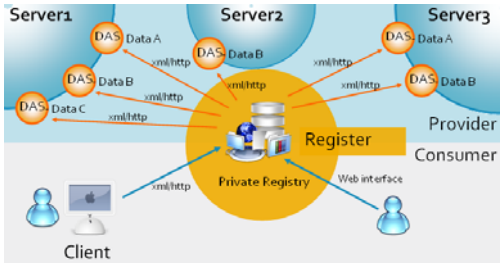

Figure 1: The overview of the system design

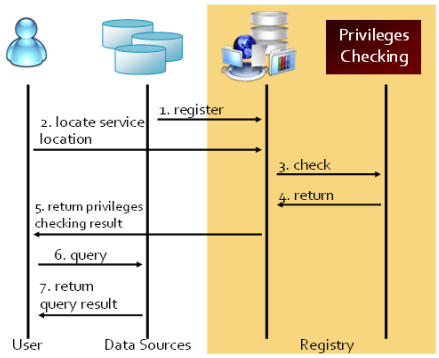

Figure 2: Private registry sequence diagram

4. System Implementation

The registry system was implemented using Java web services. Figure 2 shows the sequence diagram of the private registry system. (1) service providers register their services in the registry. (2) a user can query the registry looking for service locations. (3) the registry check user privileges (4) the registry returns the result of privilege checking to the main system (5) the main system sends user the privileges checking result. If the user has privileges, the user can follow (6) and (7). In step (6) the user requests the desired data from the locations that the system notifies. In step (7) the user then can acquire the output data.

Our private registry is designed to have these following functions:

1. A user can discover and register DAS sources via Web browser or Web services
2. A user can validates DAS sources to ensure that those sources are valid according to DAS schema
3. A user's authentication needs to be validated by using username and password to access this private registry via both Web browser or Web services

Then, we have implemented private registry to store and list the services of bioinformatics data. This service has these following available operations:

- GetAllDASSources
  - List all registered sources
- GetAllCategories
  - List all categories available in the registry
- GetAllOrganisms
  - List all organisms available in the registry
- GetAllCapabilities
  - List all capabilities that DAS supports
- GetAllLabels
  - List all labels of authority
- ValidateDAS
  - Validate the given DAS source
- GetDASSourcesByKeyword
  - Search DAS sources by keyword
- GetDASSourcesByCategory
  - Search DAS sources by keyword
- GetDASSourcesByOrganism
  - Search DAS sources by organism
- GetDASSourcesByCapability
  - Search DAS source sby capability that DAS supports
- GetDASSourcesByLabel
  - Search DAS sources by label of authority

5. Experimenal Results

In our experiment, we used existing bioinformatics open source tools. These tools include Apollo [4] and GBrowse [5] which are widely used as genomic features visualization tools. We configured GBrowse to query data and compare data from multiple sources of data A in server-1 and data B in server-3 as shown in Figure 1. We used Apollo to implement the private registry easily.

We compared the result of processing an original DAS source using GBrowse as the reference source (shown in Figure 3) with the result obtained from querying via private registry using Apollo (shown in Figure 4). For example, from Figure 3, the last line of the glyph has a reverse direction. The name of this feature is ORF and its type is ESTScan. The range position is 6 - 353 and the score is 267.0. In Figure 4, ORF has the reverse direction (the lower section). The type is orf:orf:WSTScan (the type field). The range position is 6 - 353 (the range field) and the score is 267.0 (the score field). Thus, we can see that querying via private registry (shown in Figure 4) returns the same result as querying from original DAS sources (shown in Figure 3).

To test the effectiveness of the private registry, we run 5 different queries over 5 data sources, with and without private registry. The sample queries and their responses which are return in XML format are

http://gi.biotec.or.th/cgi-bin/das/clone/features?segment=GL-N-STC02-0142-LF

and

http://coeservice.en.kku.ac.th/cgi-bin/das/clone/features?segment=HC-H-S01-1046-LF.

Comparison on real-world queries with and without the private registry is shown in Table1.

Table 1: Comparison on query response times

query	q1	q2	q3	q4	q5
With PR	4.885	4.183	55.84	8.79	61.338
Without PR	62.81	59.57	66.25	60.589	63.033

It turns out that with the private registry, the average response time is 27.001s while without private registry the running time is 62.450s. With private registry, query q5 takes a long time since its answer requires the returned results from four data sources. For most queries, our private registry can reduce response time because the system does not waste time to find the location of the requested data.

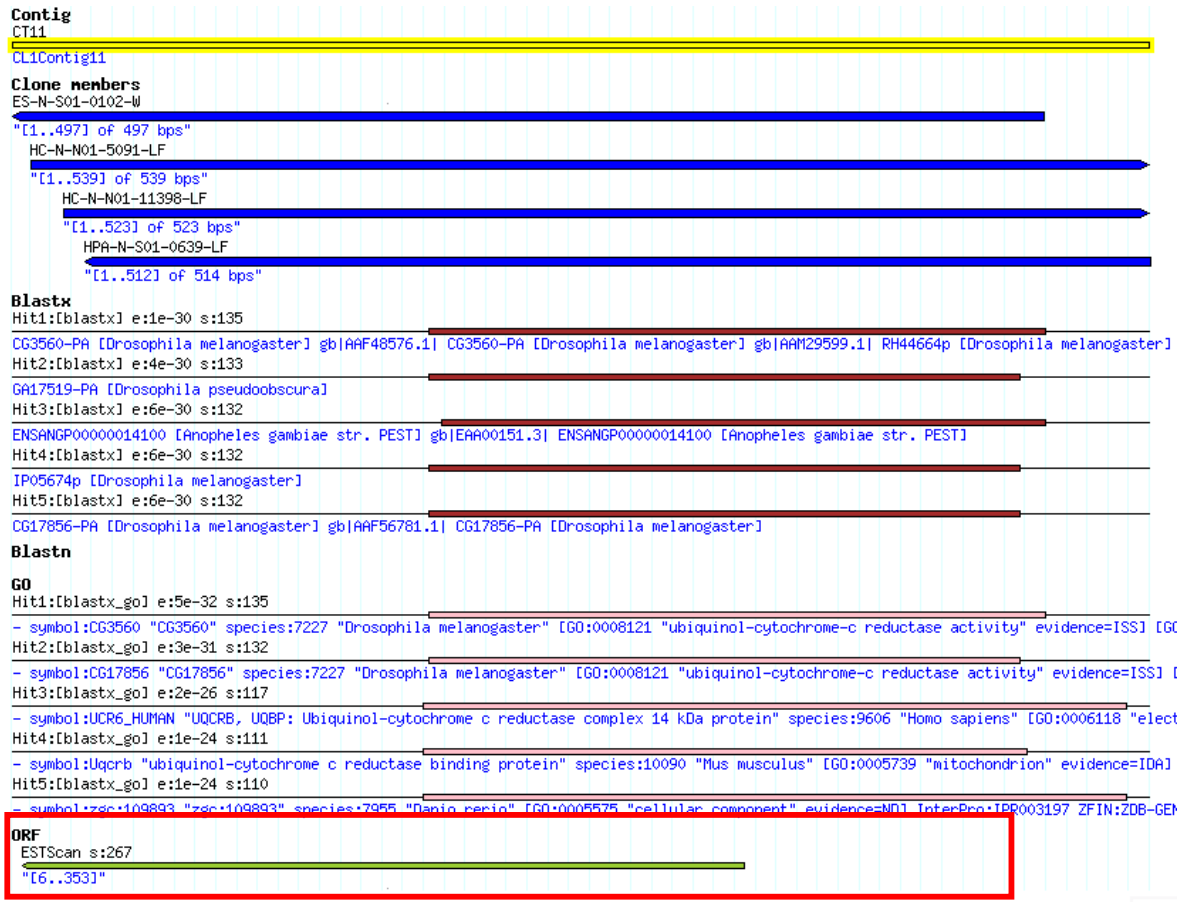

Figure 3: A result example from GBrowse which is queried from original DAS sources

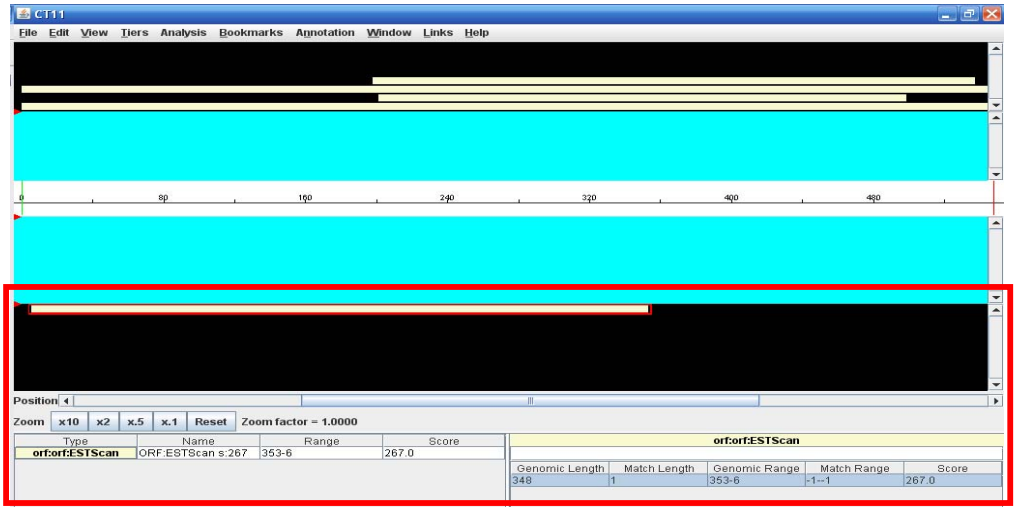

Figure 4: A result example from Apollo which is queried from our private registry

6. Discussion and Future Work

Web services can be exploited to integrate bioinformatics data via a standard protocol, such as DAS, and well-known tools, such as GBrowse and Apollo.

However, existing tools that do not have registry cannot help users to easily find the information. Some registry was implemented as a web application only. Thus, such registry cannot automatically be accessed by many applications.

It is not trivial in integrating heterogeneous data by using one application. One of typical approaches is to have the standard schema that various sources accept and use it. Such standard schema is DAS which becomes the protocol to communicate between client and server. DAS has been used in many applications. More than 250 projects that use DAS are available in a public registry for sharing public annotation to use in a distributed annotation system. Our system has developed and used a private registry which helps to increase security and protect data privacy.

Currently, the system is equipped with a simple authentication through using username and password. However, we plan to add more Web services mechanisms such as using WS-security to enhance security. Furthermore, we are interested to improve the network speed by using peer-to-peer between private nodes. For the private registry, we will add the feature to automatically choose and suggest the best time to access a given data source.

7. Conclusion

Our paper presents an approach to integrate bioinformatics data from various source databases using web services and a private registry. The goal is to construct an efficient framework that correctly integrates genomic databases via Web services and thus allows genomic browsers such as GBrowse and Apollo to connect and clearly presents data based on user criteria. The experimental result shows that for most queries our system with private registry can return correct answers with dramatically less response time than one without.

Acknowledgement

The authors would like to thank National Center for Genetic Engineering and Biotechnology (BIOTEC) for funding and supporting this research.

References

[1] Andreas Prlić, et. al, "Integrating sequence and structural biology with DAS", BMC Bioinformatics, Sep 12, 2007.

[2] Robin D Dowell, et. al, "The Distributed Annotation System", BMC Bioinformatics, Oct 10, 2001.

[3] Mark D. Wilkinson, et. al, "BioMOBY Successfully Integrates Distributed Heterogeneous Bioinformatics Web Services. (The PlaNet Examplar Case)", Plant Physiology vol 138, May 2005, pp. 5-7.

[4] SE Lewis, et. al, "Apollo: a sequence annotation editor", Genome Biology 2002, December 23, 2002.

[5] Lincoln D. Stein, et. al, "The Generic Genome Browser: A Building Block for a Model Organism System Database", Genome Res. (2002) 12: 1599 – 1610