

Bio-Sense: A System for Supporting Sharing and Exploration in Bioinformatics Using Semantic Web Services

Kai Xu¹, Ji Zhang¹, Mark Hepburn¹, Qing Liu¹ and Athman Bouguettaya²

¹ CSIRO ICT Centre, Hobart, TAS, Australia 7000

² CSIRO ICT Centre, Canberra, ACT, Australia 2601

{Kai.Xu, Ji.Zhang, Mark.Hepburn, Q.Liu, Athman.Bouguettaya}@csiro.au

Abstract

With a fast paced development of Bioinformatics in recent years, we have witnessed a rapid growth of the number of databases and tools available for aiding in scientific research and knowledge discovery for Bioinformaticians. Web service is an enabling technique to facilitate Bioinformaticians in this discovery process by integrating the databases and tools. In this paper, we will propose a novel system, called Bio-Sense, for supporting the sharing and exploration in Bioinformatics using semantic Web services. The system architecture and unique features of Bio-Sense will be discussed in this paper.

1 Introduction

Bioinformatics is an emerging research area that utilizes computational methods to address biological questions. It is the key factor of the recent breakthroughs in biological and medical science such as the sequencing of the human DNA and identify genetic causes of cancers. Bioinformaticians perform increasingly complex computation process to analyse the data, which involves the use of many different data sources and tools. Manual “assembling” of these resources is time consuming and error prone. There is a pressing need to address this problem systematically.

Service computing is commonly regarded as a potential solution to this problem. The integration is done through a loose coupling and does not require any change to the underlying data sources and tools. With the assistance of semantic information, it is also possible to automatically “compose” together the resources to perform the complex analysis required by the user. Web services—which employ Web as a medium for integration and communication—is particularly relevant to bioinformatics, as abundant data sources and tools are mostly available online.

*This research is supported by CSIRO Preventative Health Flagship program and Tasmanian ICT Centre. Tasmanian ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development and Tourism.

There is significant progress being made recently towards building Web services platforms to support Bioinformatics, with myGrid [3] and Bio-Moby [5] being the two prominent examples. They represent the efforts on integrating services from large research and computation centres and make them easily accessible for smaller labs and groups. However, some desired features are still missing or immature in their current implementations. First, myGrid, as well as Taverna, lacks a dedicated data manager, meaning the user has to have some external interaction with a data repository or locally stored data files [1]. Second, these systems use syntactics-based service description, discovery and composition, thus they are labor intensive and susceptible to human errors. Third, the prototype of a workflow construction environment, called Data Playground [1], has been recently developed as a plugin of myGrid. It allows for exploratory workflow construction where users, who do not have well-defined processes in mind, can interactively select the applicable Web services that the system discovers based on the current input and/or output of data. Nevertheless, it is not able to recommend possible construction schemes based on the workflows that have been constructed thus far. Finally, these systems are not equipped with the mechanism of provenance, which refers to the ability of tracing the evolution of information, e.g., from where and how the biological experimental results are obtained.

2 Bio-Sense

We are developing a novel system, called Bio-Sense, for supporting sharing and exploration of Bioinformatics in colorectal cancer research using semantic Web services. The architecture of Bio-Sense is shown in Figure 1. It is built in a multi-level structure that consists of the following five layers (starting from the lowest level to the highest one):

- **Databases.** This layer contains all the data sources, both internal (such as the Rat and Human CRC Genome Database) and external (such as the KEGG pathway and Literature). Also included are the Log Databases, which is important in supporting the provenance function.

- **Web Services.** In this layer, all the data sources (from the

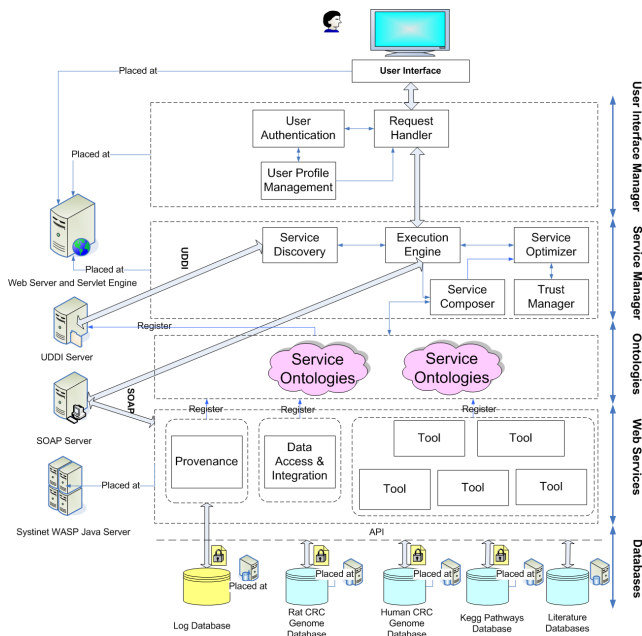


Figure 1. System architecture

Database layer) and tools (both internal and external) are wrapped up as Web services. This provides a unified way to integrate and access heterogeneous data sources and tools.

- **Ontologies.** In this layer, semantic description is added to the web services include in the previous layer. This information is essential for the system to “understand” the function of each services, and makes it possible to achieve automatic service composition based on user requirements.

- **Service Manager.** Service manager takes care of a number of Web services-related tasks in Bio-Sense. The “Service Discovery” component identifies the required services according to user inputs. After that, the “Service Composer” assembles them into a workflow, which is then executed by the “Execution Engine”. During the execution, the performance and security issues are looked after by the “Service Optimiser” and “Trust Manager” respectively.

- **User Interface.** This is the layer where users directly interact with Bio-Sense. User interface enables users to perform various activities such as data/service selection, service composition, service execution monitoring and result browsing, etc.

Bio-Sense has the following distinct characteristics and can cope with the limitations of the existing systems:

1. **Data as service.** Data as service is the reflection of the general principle adopted in Bio-Sense, that is *everything is service*. External or local databases are converted to services in order to facilitate data sharing. Data publication portal is being developed for streamlining this conversion. This contributes to a better management and utilization of data sources.

2. **Semantic service description and discovery.** Bio-Sense is built based on the WSMO/WSMX framework. Domain ontologies have been created in Bio-Sense using WSMO [4] to provide semantic description of the included services. WSMX [2] is used in Bio-Sense for composing and revoking services due to its potential to realize automatic service discovery, composition and execution.

3. **Interactive service composition.** Bio-Sense adopts a data-centric paradigm in composing services for supporting exploratory workflow construction. In each step of workflow construction, Bio-Sense provides a list of applicable services that can operate on the current input/output data and users can select one of them for extending the workflow until some tasks are fulfilled. Besides such service recommendation that has been proposed in Data Playground, Bio-Sense is able to recommend full-length workflows based on the workflows that have been constructed thus far. Workflow recommendation is more informative and useful from users’ perspective than service recommendation.

4. **Process provenance.** Bio-sense aims to provide process provenance support to *in silico* experiments. Biological process will be first captured in an automatic manner. These processes can be visually presented and efficiently retrieved. They can also be shared within domain researchers such as biologists and bioinformaticians. The scientific provenance management not only helps them to determine the data’s value, accuracy and authorship, but also facilitates them to interpret and understand results which can be more important than the actual results themselves.

3 Conclusions

In this paper, we presented Bio-Sense, a new semantic Web service system designed to support colorectal cancer research. It has an enriched multi-level architecture and a number of unique features such as data as service, semantic service description and composition using WSMO/WSMX framework, interactive workflow construction with recommendation and process provenance support. More advanced features such as gene linkage and knowledge discovery modules will be developed in the near future.

References

- [1] A. Gibson, M. Gamble, K. Wolstencroft, T. Oinn, and C. A. Goble. The data playground: An intuitive workflow specification environment. In *eScience*, pages 59–68, 2007.
- [2] A. Haller, E. Cimpian, A. Mocan, E. Oren, and C. Bussler. Wsmx - a semantic service-oriented architecture. In *ICWS*, pages 321–328, Washington, DC, USA, 2005.
- [3] L. Moreau, T. Payne, and M. Szomszor. Using semantic web technology to automate data integration in grid and web service architectures. In *CCGrid*, pages 189–195, 2005.

- [4] D. Roman, U. Keller, H. Lausen, J. de Bruijn, R. Lara, M. Stollberg, A. Polleres, C. Feier, C. Bussler, and D. Fensel. Web service modeling ontology. *Applied Ontology*, 1(1):77–106, 2005.
- [5] M. D. Wilkinson and M. Links. BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics*, 3(4):331–341, 2002.