TOWARDS THE INTEGRATION OF BIOINFORMATICS DATA AND SERVICES USING SOA AND MASHUPS

Elarbi Badidi[1] and Larbi Esmahi[2]
[1]*College of Information Technology, United Arab Emirates University, PO.Box. 17551, Al-Ain, United Arab Emirates*

[2]*School for Computing & Information Systems, Athabasca University*
1 University Drive, Athabasca, Alberta, T9S 3A3, Canada

Keywords: Data Integration, Application Integration, Mashups, Service Oriented Architecture, and Web Services.

Abstract: Worldwide biological research activities are generating publicly available biological data at a phenomenal pace. Data is usually stored in different formats (fasta, genbank, embl, xml, etc.). Therefore, retrieving, analyzing, parsing, and integrating these heterogeneous data require substantial programming expertise and effort that scientists do not have overall. Bioinformaticians have often considered several approaches to integrate heterogeneous data and software applications. Most of these integration approaches require significant computer skills. Recently, a new technology, called mashups, has emerged to simplify this integration. In this paper, we discuss widely used approaches for integrating data and applications in bioinformatics and our ongoing effort to use mashups in conjunction with Service Oriented Architecture (SOA) for integrating data and applications in Life Sciences.

1 INTRODUCTION

Data and application integration is one of the toughest problems facing bioinformatics today. Indeed, to enable the discovery of new biological insights, scientists have to interpret many types of information from a variety of sources including nucleotide and amino acid sequences, protein structures, and various other sources. Unfortunately, it is not easy for scientists to access all these information in an integrated way due to: a) the heterogeneity of data formats, b) the heterogeneity of interfaces, and c) the variety of bioinformatics software tools used by the underlying data sources.

The need for data and tools integration is widely acknowledged in the bioinformatics community. Successful data integration is one of the keys to enhanced productivity in biology and biopharmaceutical R&D. Worldwide biological research activities are increasingly generating data spread across public databases and throughout organizations in various formats (Fasta, Staden, Embl, Genbank, etc.). Furthermore, most of the software applications in bioinformatics are incompatible with one another as they use different input and output formats. Bioinformaticians have considered several approaches to address the challenges of data integration. As a result of that, numerous integration platforms and frameworks flourished in both academia and bio-industry.

In the last few years, SOA embodied by Web services has emerged as a key technology for providing services over the Web. Web services are interoperable across platforms and neutral to languages, which makes them suitable for access from heterogeneous environments. Web services technology has all the potential to be a major component in the integration endeavor because it provides a higher layer of abstraction that hides implementation details from applications. Using this technology, applications invoke other applications' functions through well-defined, easy-to-use interfaces. Each organization is free to concentrate on its own competence and still leverage the services that other research groups provide.

Recently, a new technology, called mashups, has emerged to simplify data integration. Mashups allow end-users to extract and integrate data from disparate sources using a visual approach. In this paper, we describe our work in progress that investigates how these emerging technologies may be applied to integrate biological data and services.

Badidi E. and Esmahi L. (2010).
TOWARDS THE INTEGRATION OF BIOINFORMATICS DATA AND SERVICES USING SOA AND MASHUPS.
In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 95-98
DOI: 10.5220/0003041200950098

2 BACKGROUND

2.1 Bioinformatics Data integration

Bioinformatics data sources (e.g. GenBank, Entrez, PDB, Swissprot, etc.) cover various domains, such as genes, proteins, or sequence annotations. To get the best out of such a massive amount of data, three main approaches for data integration have been commonly considered: database federation, adoption of XML as a common format for the exchange and storage of biological data, and adoption of ontologies.

Database federation approach relies on a middleware that acts a mediator amongst heterogeneous, disparate, and autonomous databases (Kemp et al., 2002). The middleware provides a uniform query interface, which allows end-users to store and retrieve data from multiple disparate databases using a single query. The system decomposes the query into sub-queries for submission to the appropriate constituent databases, after which the middleware integrates the result sets of the sub-queries. The advantage of this approach is that it does not require modification of the constituent databases of the federation.

XML, The Extensible Markup Language, is increasingly becoming a popular format for the exchange and storage of biological data (Achard et al., 2001). Most biological databases such as GenBank, PIR, and SWISS-PROT are XML-compliant. XML is used to describe biological data in a standardized format that can be shared and easily transported via standard Internet protocols.

To overcome the heterogeneous terminologies used in databases, ontologies are often used to standardize terms and concepts for which the meaning and relationships are explicitly defined by a standards organization. In bioinformatics, ontologies represent a new style of data integration as they set naming conventions for data elements stored in biological databases. The Gene Ontology (GO) from the Gene Ontology Consortium is an example of ontology in bioinformatics (Masseroli et al., 2006).

2.2 Bioinformatics Tools Integration

Bioinformaticians have considered several approaches and technologies to deal with the issue of bioinformatics tools integration. One of the approaches develops and employs locally interactive environments to facilitate bioinformatics analyses. Examples of such environments are Isys (Siepel et al., 2001) and Applab (Senger, 1999). The disadvantage of this approach is that the integration environment as well as the tools must be installed and configured locally, which requires substantial IT expertise. The second approach takes advantage of the growing use of the Web to make the tools available throughout Web interfaces using HTML forms and different scripting languages (Perl, CGI, etc.). The integration of tools is in most cases data-driven, as, for example, in Bionavigator (Littlejohn, 2001) and NCSA biology workbench (Unwin et al., 1998). The advantage of this approach is that the user does not need to install bioinformatics tools locally, but only requires a Web browser. A third approach is also Web-oriented, but it uses relatively new technologies at the server side to integrate various bioinformatics software tools. These technologies are Java RMI, EJB, and CORBA. Examples of integration frameworks using the above technologies are Anabench (Badidi et al., 2003), Applab (Senger, 1999).

The recent advances in distributed computing have lead to SOA and the Web and Grid services technologies, which promise to ease the integration and interoperability issues. In life sciences, the service orientation is seen as a promising approach for the integration of heterogeneous and incompatible bioinformatics applications available from different providers. It has quickly caught the interest of several organizations and research centers. Most of them have published their tools as Web services. The National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) has published its Entrez Utilities as Web Services. The European Bioinformatics Institute Web service (EBI; http://www.ebi.ac.uk/) provides programmatic access to data retrieval and analytical tools for several molecular databases. Besides, new environments and frameworks have emerged to enable the orchestration and the execution of biological Web services. The most notable systems include Taverna (Oinn et al., 2004) and BioMoby (Wilkinson et al., 2003).

2.3 Bioinformatics Mashups

Recently, the “Mashup” technology has emerged as a new approach for creating Web applications that fuse data from numerous sources to create a new service and to provide the user with an integrated experience. Data mashups typically provide visual representations of data available in public databases. Similarly, service mashups aim to support end-users in creating and building new and advanced Web

applications by means of easy-to-accomplish service compositions. Various mashup tools and frameworks have flourished in the last few years. They let end-users access and combine data from various sources (databases, Web services, feeds, etc.). Examples of well known mashup platforms are Yahoo Pipes (http://pipes.yahoo.com), and Google Mashup Editor (http://code.google.com/gme/).

Bioinformaticians have a wide range of tools and services at their disposal, each providing some value when used in isolation, but offering greater promise when integrated with others in the investigation of more complex scientific questions. Sophisticated scientific workflow management systems such as Taverna (Oinn et al., 2004) or Kepler (Ludäscher et al., 2006) can support almost any level of complexity. Mashups, however, may provide a middle ground between manual cut-and-paste interaction with the Web and these sophisticated workflow engines. They allow developing lightweight applications by combining data and services in accordance with the user requirements.

Recently, there were few attempts to develop some basic biomashups. Sumitomo et al. (Sumitomo et al., 2008) describe basic mashups, such as a mashup to translate Uniprot ID into Genbank Id, a Genbank to Pubmed mashup, and a biomashup to get the characteristics of a protein given its Uniprot ID. Kei-Hoi et al. (Kei-Hoi et al., 2008) discuss the potential of mashups and current Web 2.0 technologies to integrate data from disparate sources in the health care and life sciences (HCLS) domains.

We strongly believe that the proliferation of mashups platforms together with Web 2.0 techniques will contribute extensively to the wide adoption of mashups in Life Sciences. Life scientists empowered with Biomashups will be able to integrate local data with data from other sources (public biological databases, workbench portals, etc.). They will also be able to customize and exploit the results rather than consuming biological data in the way the publisher (organizations and research centers) presents that data.

3 BIOMASHUP ARCHITECTURE

3.1 Objectives

Given the limited number of biomashups that have been developed so far by the bioinformatics research community, and given the unprecedented advances made in the area of mashups technology and the Service Oriented Architecture, we set out to use an architecture that will allows us to:

1. Investigate the feasibility of applying "Mashup" technology in conjunction with SOA technologies in bioinformatics to cope with the issue of data and services integration.
2. Apply these technologies to some real-case scenarios (e.g. in proteomics and/or genomics) that typically require, from the scientist, to go through several steps by using various software tools or Web-based applications before getting an answer to his/her question.
3. Investigate how biological protocols can be implemented using mashups. Then, compare this mashups-based approach with established workflow systems and SOA service composition. Comparison will evaluate the following aspects: usability, composition effectiveness, support to the user, and extent of necessary technical knowledge.
4. Examine how mashup technology and SOA can contribute to data and services integration in the biomedical field in general. Healthcare and bioinformatics share many common aspects especially in terms of heterogeneity of data formats and data exchange protocols.

3.2 Architecture Overview

Figure 1 illustrates our proposed Mashups-based architecture for bioinformatics data and applications integration. The architecture relies on a central component, which is the *Mashup Server*. The Mashup Server may acquire data from various data sources including biological databases (Genbank, EMBL, Uniprot, Swissprot, …), biological Web services (BioMoby, MyGrid, Soaplab, NCBI,…), publication databases (Pubmed, …), and feeds. It then processes and combines acquired data to present the result as small components, called Mashlets, which can be integrated in portals, Web pages, wikis and blogs. Legacy applications may be integrated by being wrapped as Web services. In our previous work (Removed for blind review), we have implemented a system that allows wrapping command-line bioinformatics tools as Web services.

The mashup application will produce, then, a single interface replacing the manual process of accessing multiple data sources. Using mashup applications, scientists can collaborate more effectively, be more productive, and reduce the risk of error.

Enterprise Mashup servers have emerged recently following the success of mashup platforms in implementing interesting enterprise applications

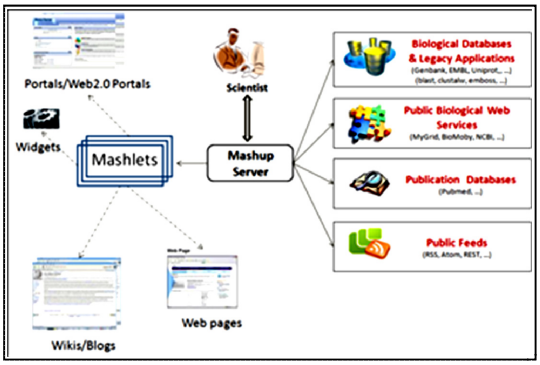

Figure 1: Integration of bioinformatics data and services using mashups technology.

We are considering to use open source mashup server such as WSO2 mashup server (http://wso2.org/projects/mashup), which support different data formats commonly used by biological databases. JackBe Presto Enterprise Mashup Server (http://www.jackbe.com/) is also a potential candidate as it provides many interesting features.

4 CONCLUSIONS

As a result of worldwide biological research activities, data is spread across disparate databases and organizations in various formats. Data Mining and analysis require comprehensive integration of these heterogeneous data. The need for data and services integration is widely recognized in the bioinformatics community. Successful data integration is one of the keys to enhanced productivity in biology and biopharmaceutical R&D. In this paper, we have provided an overview of existing approaches for data and applications integration in bioinformatics. We have also discussed the promise of emerging technologies, service oriented architecture and mashups. Furthermore, we have presented our work in progress in implementing an architecture that relies on a mashups server and Web services for integrating data and applications in Life Sciences.

REFERENCES

Achard, F., Vaysseix, G., and Barillot, E., 2001. XML, bioinformatics and data integration. *Bioinformatics*, 17(2):115-25.

Badidi, E., Salem, M. V., Bouktif, S., and Esmahi, L., 2009. A Web Services based Framework for Uniform Integration of Command-line Bioinformatics Software Tools. *International Journal on Web Services Practices (IJWSP)*, 4 (1) (2009). 36-43.

Badidi, E., De Sousa, C., Lang, B. F. and Burger, G., 2003. AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis. *BMC Bioinformatics*, 4: 63.

Kei-Hoi, C., Kevin Y. Y., Jeffrey, P. T., and Matthew, S., 2008. HCLS 2.0/3.0: Health care and life sciences data mashup using Web 2.0/3.0. *Journal of Biomedical Informatics*, 41(5): 694–705.

Kemp, G. J. L., Angelopoulos, N., and Gray, P.M.D., 2002. Architecture of a Mediator for a Bioinformatics Database Federation. *IEEE Transactions on Information Technology in Biomedicine*, 6(2): 116-122.

Littlejohn, T.G., 2001. Bioinformatics tools for genome projects. In *Molecular Breeding of Forage Crops, Spangenberg, G. (ed.)*, Kluwer Acad. Publ., The Netherlands, 83-99.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J., and Zhao, Y., 2006. Scientific Workflow Management and the Kepler System, *Concurrency and Computation: Practice & Experience*, 18(10): 1039-1065.

Masseroli, M., and Pinciroli, F., 2006. Using Gene Ontology and genomic controlled vocabularies to analyze high-throughput gene lists: Three tool comparison. *Computers in Biology and Medicine*, 36(7): 731-747.

Oinn, T., Addis, M. J., Ferris, J., Marvin, D. J., Greenwood, M., Carver, T., Wipat, A., and Li, P., 2004. Taverna, lessons in creating a workflow environment for the life sciences. *Paper presented at the GGF10*, Germany, 2004.

Senger, M., 1999. AppLab - A CORBA-Java based Application Wrapper. Retrieved from http://www.omg.org/docs/corbamed/98-03-08.pdf

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.

Sumitomo, J., Hogan, J. M., and Roe, P., 2008. BioMashups: The New World of Exploratory Bioinformatics? In *Proceedings of the Fourth IEEE International Conference on eScience*, 422-423.

Unwin, R., Fenton, J., Whitsitt, M., Jamison, C., Stupar, M., Jakobsson, E., and Subramaniam, S., 1998. Biology Workbench: A WWW-based Virtual Computing and Analysis Environment for the Biological Sciences. *Bioinformatics(Databases and Systems, S. Letovsky (Ed.))*, 233-244.

Wilkinson, M. D. and Links, M., 2003. BioMOBY: An open source biological web services proposal. *Briefings in bioinformatics*, 3(4): 331–341.