A FRAMEWORK BASED IN WEB SERVICES ORCHESTRATION FOR BIOINFORMATICS WORKFLOW MANAGEMENT

Luciano A. Digiampietri, Claudia B. Medeiros and Joao C. Setubal

Instituto de Computação, Universidade Estadual de Campinas
CP 6176, Campinas, SP 13084-971 BRAZIL
luciano@ic.unicamp.br

Abstract. Bioinformatics activities are growing all over the world, with proliferation of data and tools. This brings new challenges: how to understand and organize these resources and how to provide interoperability among tools to achieve a given goal. The purpose of this work is to define and implement a framework to help meet some of these challenges. There are four issues considered: the use of Web services as a basic unit; the notion of Semantic Web to improve interoperability at syntactic and semantic levels; the use of scientific workflows to coordinate services to be executed and their interdependencies and service orchestration.

This paper addresses the following topics
- Biological Databases, Data Management and Data Integration
- Algorithms and Software Tools for Computational Molecular Biology

1 Introduction

Bioinformatics activities are growing all over the world. Among the various problems, there is the question of providing a framework for inter-institutional cooperation. One of the directions considered is to use the new service technologies (Web services[1] and Grids[13]).

Web services are a good approach to solve heterogeneity problems. The use of XML[12] and standard Internet protocols contribute greatly to popularization and dissemination of Web services. Important issues are service and data discovery, service execution and coordination. Thus, there is a need for management mechanisms for data and services, and for supporting enhanced semantics.

The main goal of our work is to propose and develop a framework to solve some of these problems, for bioinformatics applications. In this specific application domain, there are already some incipient proposals that involve the coordination of distributed tasks by using workflows[6, 15, 18, 26] and Web services[3, 29]. These proposals suffer from problems described previously; moreover, there is a lack of standards for interfaces among tools used by end-users. Thus, besides contributing towards managing data and services, the framework will contribute to help tool interoperability.

The expected results are the specification and development of a framework for bioinformatics applications capable of: (i) specifying workflows, via composition of Web services, and storing these specifications; (ii) discovering services and workflows of interest, in a semantic way; (iii) managing workflow execution via service orchestration and (iv) auditing workflow execution.

The rest of this text is organized as follows. Section 2 shows related work. Section 3 describes the main aspects of the proposed framework. Section 4 contains conclusions and ongoing work.

2 Related work

2.1 Systems and frameworks

Many works consider the integration of bioinformatics data and tools. Some emphasize functionality for a specific research team whereas others concentrate on supporting cooperation on the Web, for teams within a given project. The main goals of these systems, summarized in Table 1 are: (i) to provide a set of bioinformatics tools, (ii) to allow data and tool integration and (iii) to build a framework for one specific bioinformatics project.

Systems that provide a set of bioinformatics tools make their tools available via Web sites and/or via a local program [4, 10, 15, 22]. Usually these tools are developed under specific standards, hampering the integration of tools built by different groups. The framework of Eckart and Sobral[10] employs Web services in the server side and an application on the client side. When the application is started, the list of available services is updated, allowing clients to invoke new services. Standard inputs and outputs allow the output of a service to be used as input of another. This framework does not yet allow automatic integration of tools.

Several systems provide some level of data integration. Some have the goal of integrating large volumes of available genomic data in one generic data model[19]. Other systems aim the modeling of any genomic project via a set of basic components[21]. Finally there are systems that link several kinds of services and data to facilitate the genomic annotation of one specific genome project[9]. Some systems handle the problem of tools integration. The specification of tasks interaction and interdependencies relations are typically designed using workflows[4, 15, 22]. The problem lies in workflow specification and execution.

There are two main kinds of frameworks for bioinformatics projects. In the first kind, all tools are developed for a specific project[17]. Whenever a new genome project is started, scientists need to adapt the entire framework. The second kind contains the frameworks formed by basic components[21]. The framework of each new genome project is constructed by combining components with low configuration costs. Both kinds of framework are especially good for genomic assembly and annotation of specific genomes, but their tools can not be accessed by other projects.

Characteristic	BioOpera[4]	Source[9]	Hall[15]	CMR[19]	GGB[21]	myGrid[22]
1- execution of a task in a distributed environment	x		x			x
2- maintenance of repository of bioinformatics tools		x			x	x
3- provide some level of tool integration	x	x	x	x	x	x
4- modeling workflows of a complex task	x		x			x
5- multi-institutional sharing of resources	x	x	x	x	x	x
6- multi-institutional development of tools						
7- coordination of workflow execution	x		x			x

Table 1. Some surveyed systems and their characteristics.

Our framework differs from the surveyed related wok in the following ways. First, it integrates all 7 characteristics of Table 1. Second, it allows user interaction while tools are executing. Third, it focuses multi-institutional development of tools using Internet standards. This means that these tools are available not only for one project but to any project that complies with these standards. Finally, tools are managed via service orchestration.

2.2 Related issues

Related work involves research on Web services and their orchestration, scientific workflows and bioinformatics tools and data.

Web services A Web service is "a software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML artifacts. A Web service supports direct interactions with other software agents using XML-based messages exchanged via Internet-based protocols"[30]. Some open topics are how to discover adequate services and service providers, how to automate Web services integration, how to minimize semantics ambiguity in services specifications and how to assign information about quality and reliability of the services offered by a provider[1]. Our work concentrates on the aspects of specification of interfaces for bioinformatics services and their orchestration.

Service orchestration is a centralized mechanism that describes how diverse services can interact. This interaction includes message exchange, business logics and order of execution. The most important works in the coordination of Web services involving BPEL4WS[5] and WSCI[28]. We will adopt BPEL4WS as a basis for specifying service orchestration.

Workflows A workflow denotes the controlled execution of multiple tasks in an environment of distributed processing elements. Workflows represent a set of activities to be executed, their interdependencies relations, inputs and outputs[20].

Bioinformatics workflows are scientific workflows, i.e., they differ from a usual workflow because they have some additional characteristics like high degree of flexibility, uncertainty and existence of exceptions[27]. Our work is concentrated in the execution of scientific workflows through the orchestration of Web services and user interaction. Open problems that will be attacked include communication protocols and interfaces among services to specify a workflow.

Bioinformatics tools and data There are many tools and databases for bioinformatics. Samples of tools include BLAST[2], Phred[11] and Consed[14]. These tools are geared towards sequence comparison, assembly and visualization. Other complex problems with dedicated tools involve fragment assembly of DNA (alignment and consensus), phylogenetic trees, database search, etc. Our research will concentrate on the applications for assembly, annotation and comparison of genomes. The choice of these applications was based in prior experience of the Laboratory for Bioinformatics (LBI)[17] at UNICAMP in assembly and genomic annotation. This choice is also common to several bioinformatics efforts using workflows[18], clusters[4] and grids[22].

3 The proposed framework

The framework will manage design and execution of scientific workflows that will support execution of distributed bioinformatics applications on the Web. One problem faced by scientists in such a context is integrating these procedures via adequate interfacing among tools. Our approach handles this problem by encapsulating data and tools by Web services. Figure 1 shows how the framework will support the main user activities in assembly of genomic data. There are three kinds of users that interact with our framework: software developer, user-developer and end user. Software developers design Web services and subscribe these services to the Service Catalog. Users-developers use our framework to design workflows that determine how complex tasks must be composed and executed. End users invoke a Web service or a Workflow designed by a user-developer. For instance, software developer 1 develops a phred Web service that is stored in the service catalog. User-developer 1 specifies an assembly workflow that invokes this phred service and is stored in a specific repository. When end user 1 requests execution of some assembly task, the service discovery module will inform the user there are 2 available workflows, and the user can then choose which workflow to execute. Workflow activities embed tools that are executed via services; data is also made available via services.

Figure 2 shows our system architecture that supports the integration of the tasks shown in Figure 1.

The **Service layer** manages the bioinformatics Web services that must provide basic operations such as assembly, matching and consensus, creation of descriptors and genomic annotation. They are amenable to composition to provide

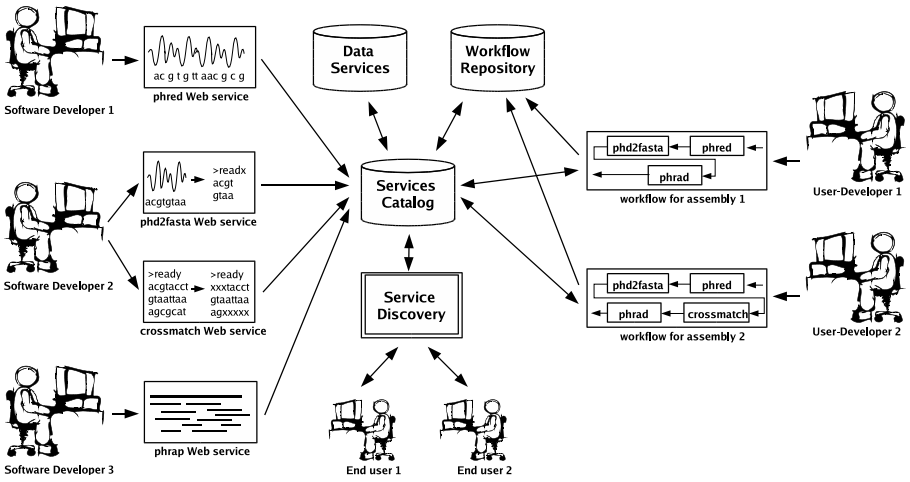

Fig. 1. User interaction overview

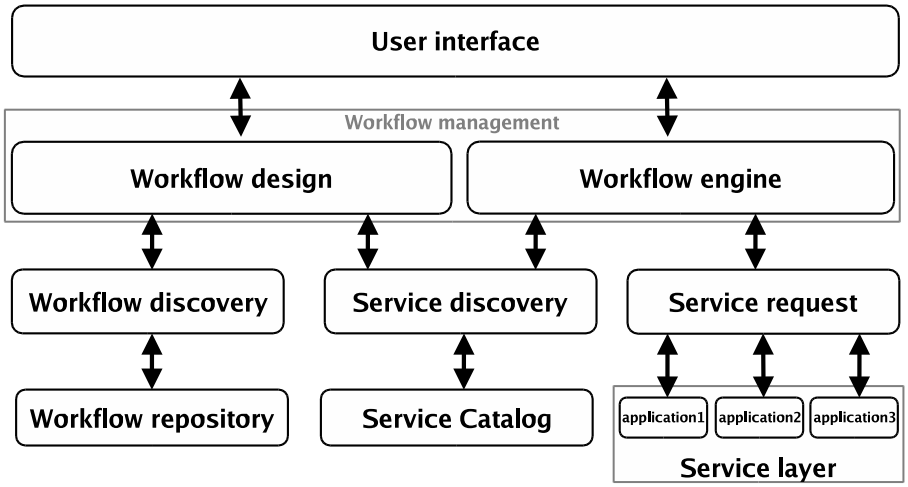

Fig. 2. Framework architecture.

more sophisticated functionalities - e.g. genomic comparison and gene family operations. Here, we started by transforming modules already available in LBI[17] into services.

Web services present some disadvantages for this work. They do not allow user interaction during their execution and they do not support auditing. Thus, when a workflow activity is performed by a Web service, it cannot be broken down (e.g., in sub workflows). Our framework will treat this problem by allowing workflows where activities can invoke external applications. Supporting human interaction is also a contribution of our proposal.

The **Service catalog layer** is responsible for storing Web services' syntactical and semantical descriptions, as well as the URI where each service can be found. This layer will utilize a schema of subscription / unsubscription to register the services. It must maintain a history of services availability and allow reuse of workflows.

Service discovery can be done in several ways. Our framework will allow search by functionality, context and syntax. Search by functionality and context will be done based in semantic data (metadata) assigned to the services. Search for compatible syntax will be based on the parameters of the service interface. These search methods will use techniques already discussed in the literature[7, 23] about syntactical, ontological and semantic matching.

The **Service request layer** will be responsible for the management of each Web service solicitation. This layer communicates with the Web services provider, sending input data and receiving results. It is responsible for detecting service failure such as unavailable service or time limit violation.

The **Workflow engine layer** is responsible for the controlled execution of all workflow tasks, via orchestration. The operation functions provided by the Workflow engine are interpretation of the process (or task) definition, creation and management of process instances, navigation between activities and supervisory and management functions[25].

The **Workflow design layer** must support workflow specification and edition. The facilities provided are: graphical interface for workflow edition, service list, interface description of selected services and syntactical check. It will use the scientific workflow editing tools developed at UNICAMP[20].

Our framework is being specified and developed using a bottom-up approach. We started with bioinformatics basic services specification and development encapsulating LBI tools into services[17] e.g. genomic annotation and comparison tools. This stage is also establishing the metadata types that must be associated with the services. The strategy for this stage requires initial definition of some bioinformatics basic services.

The second stage will be the study and development of techniques for service discovery and request using syntactic and semantics search mechanisms. The following step involves the specification and development of methods for workflow design and execution. Each workflow activity is a service or a bioinformatics application. This stage will make use of existing work on management of scientific and distributed workflows[16, 24] and tools developed at UNICAMP[20]. Here it will be necessary to specify and to implement an orchestration mechanism for these kinds of services (specific to workflows). Workflows data sources and providers will be encapsulated by services.

System tests will be based on large volumes of real data from LBI.

4 Conclusions and ongoing work

The main contribution of this work is the framework itself. It will allow multi-institutional cooperation via data, tools and workflow sharing. Various kinds of

users will be able to interact with our system and with each other to achieve a given goal. Other contributions lie in the solution of open problems in scientific workflow specification via composition of Web services and semantic specification of bioinformatics tasks. Another important contribution is the methodology for integration of these solutions for bioinformatics.

The work accomplished so far can be divided into research and practical work. The research was concentrated in the analysis of related work and tools utilized by bioinformatics research centers. The practical work was concentrated on development and utilization of LBI assembly and annotation genomic systems[17]. Furthermore, we modeled and implemented a comparative genomic system [8]. These activities allowed the understanding of bioinformatics applications in terms of types of data and applications involved. At this moment, we are specifying the Web service semantic description and encapsulating LBI tools.

Acknowledgements. This work was partially financed by grants from CAPES, CNPq and FAPESP, and projects MCT-PRONEX SAI and CNPq WebMaps and AgroFlow. The Laboratory for Bioinformatics, in which most of this work was done, is partly funded by FAPESP.

References

1. Alonso, G., Casati, F., Kuno, H., Machiaraju, V.: Web Services: Concepts, Architectures and Applications, *Springer* (2004)
2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, (1997) 3389–3402
3. Altunay M., Colonnese, D., Warade, C.: A framework for deploying bioinformatics applications as high-throughput Web services on the NC BioGrid. 2004, http://www-106.ibm.com/developerworks/webservices/library/ws-bioinfo.html
4. Bausch, W., Pautasso, C., Schaeppi, R., Alonso, G.: BioOpera: Cluster-aware Computing. *In Proc. of the 4th IEEE Int. Conf. on Cluster Computing (Cluster)*, (2002)
5. Business Process Execution Language for Web Services Version 1.1 (BPEL4WS). 2003, http://www-106.ibm.com/developerworks/library/ws-bpel/
6. Cannataro, M.; Comito, C.; Guzzo, A.; Veltri, P.: Integrating ontology and workflow in PROTEUS, a grid-based problem solving environment for bioinformatics *Proc. of the Int. Conf. on Coding and Computing*, **2**, (2004) 90–94
7. Cardoso, J., Sheth, A.: Semantic e-Workflow Composition. *Journal of Intelligent Information Systems*, **21**:3, (2003) 191–225
8. Digiampietri, L. A., Medeiros, C. M. B., Setubal, J. C.: A data model for comparative genomics. Proc. of the Second Brazilian Workshop on Bioinformatics (2003)
9. Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., Rees, C. A., Cherry, J.M., Botstein, D., Brown, P.O., Alizadeh, A. A.: SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **31** (2003) 219–223
10. Eckart, J. D., Sobral, B. W.: A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *Spring* **7**:1, (2003) 79–88
11. Ewing, B., Green, P.: Base-calling of automated sequencer traces using phred. II. Error probabilities. /em Genome Res. **8**:3, (1998) 186-94

12. Extensible Markup Language (XML) 1.0 (Third Edition). 2004, http://www.w3.org/TR/2004/REC-xml-20040204
13. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. *Morgan Kaufmann* (1999)
14. Gordon, D., Abajian, C., Green, P.: Consed: a graphical tool for sequence finishing. *Genome Res.* **8**:3, (1998) 195-202
15. Hall, D., Miller, J. A., Arnold, J., Kochut, K. J., Sheth, A. P., Weise, M.: Using Workflow to Build an Information Management System for a Geographically Distributed Genome Sequencing Initiative. *Bioinformatics Journal*, (1999)
16. Kim, K. H.: Workflow dependency analysis and its implications on distributed workflow systems *17th Int. Conf. on Advanced Information Networking and Applications*, (2003) 677–682
17. Laboratory for Bioinformatics, Institute of Computing, University of Campinas. http://www.lbi.ic.unicamp.br
18. Meidanis, J., Vossen, G., Weske, M.: Using workflow management in dna sequencing. *Proc. of the First Int. Conf. on Cooperative Information Systems*, (1996)
19. Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K, White, O.: The Comprehensive Microbial Resource. *Nucleic Acids Research* **29**:1 (2001) 123–125
20. Seffino, L.A., Medeiros, C.B., Rocha, J.V.R., Yi, B.: WOODS- a spatial decision support system based on workflows. *Decision Support Systems* **27** (1999) 105–123
21. Stein, L. D., Mungall, C., Shu S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J .E., Harris, T. W., Arva, A., Lewis, S.: The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**:10, (2003) 1599–610.
22. Stevens, R., Robinson, A., Goble, C. A.: myGrid: Personalised Bioinformatics on the Information Grid. *Bioinformatics*, **19**:1, (2003) 302–304
23. Sycara, K., Paolucci, M., Ankolekar, A., Srinivasan, N.: Automated discovery, interaction and composition of Semantic Web services *Web Semantics: Science, Services and Agents on the World Wide Web*, **1**:1, (2003) 27–46
24. The Workflow Management Coalition: The Workflow Reference Model. *Technical Report TC-1003*, (1995)
25. The Workflow Management Coalition: Workflow Management Coalition Terminology & Glossary (issue 3.0). 1999, http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf
26. Vouk, M.A." Integration of heterogeneous scientific data using workflows - a case study in bioinformatics Proc. of the 25th Int. Conf. on Interfaces, **16-19**, (2003) 25–28
27. Wainer, J., Weske, M., Vossen, G., Medeiros, C. B.: Scientific Workflow Systems, *Proc. of the NSF Workshop on Workflow and Process Automation Information Systems*, (1996)
28. Web Service Choreography Interface (WSCI) 1.0. 2002, http://www.w3.org/TR/wsci
29. Web Service for Bioinformatic Analysis Workflow. 2004, http://www.alphaworks.ibm.com/aw.nsf/reqs/wsbaw
30. Web Services Internationalization Requirements. 2003, http://www.w3.org/International/ws/ws-i18n-scenarios-edit/ws-i18n-requirements-edit.html

This article was processed using the LaTeX macro package with LLNCS style