Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

# A framework based on Web service orchestration for bioinformatics workflow management

**Luciano A. Digiampietri[1], Claudia B. Medeiros[2] and João C. Setubal[3]**
[1]Instituto de Computação, Universidade de Campinas, Campinas, SP, Brasil
[2]Departamento de Sistemas de Informação, Instituto de Computação, Universidade de Campinas, Campinas, SP, Brasil
[3]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA
Corresponding author: L.A. Digiampietri
E-mail: luciano@ic.unicamp.br

Genet. Mol. Res. 4 (3): 535-542 (2005)
Received May 20, 2005
Accepted July 8, 2005
Published September 30, 2005

**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. We defined and implemented a framework to help meet some of these challenges. Four issues were considered: the use of Web services as a basic unit, the notion of a Semantic Web to improve interoperability at the syntactic and semantic levels, and the use of scientific workflows to coordinate services to be executed, including their interdependencies and service orchestration.

**Key words:** Bioinformatics, Workflow, Web service, Semantic web

## INTRODUCTION

Bioinformatics activities are growing all over the world. Among the various problems that this tremendous growth has created, 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 (Alonso et al., 2004) and Grids (Foster and Kesselman, 1999).

Web services are a good approach to solve heterogeneity problems. The use of XML (W3C, 2004) and standard Internet protocols has contributed greatly to the popularization and dissemination of Web services. Important issues include service and data discovery, as well as service execution and coordination. Thus, there is a need for management mechanisms for data and services and for supporting enhanced semantics.

Our main goal was to propose and develop a framework to solve some of these problems for bioinformatics applications. There are already some incipient proposals that include the coordination of distributed tasks by using workflows (Meidanis et al., 1996; Hall et al., 2003; Vouk, 2003; Cannataro et al., 2004) and Web services (W3C, 2003; IBM, 2004) in this specific application domain. These proposals suffer from problems that have been previously described; moreover, there is a lack of standards for interfaces among the tools used by end-users. Thus, besides contributing towards managing data and services, the new framework will contribute to help tool interoperability.

The expected results are the specification and development of a framework for bioinformatics applications that is 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.

## RELATED WORK

### Systems and frameworks

Many projects consider the integration of bioinformatics data and tools. Some emphasize functionality for a specific research team, whereas others concentrate on supporting cooperation via the Web, for teams within a given project. The main goals of these systems (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.

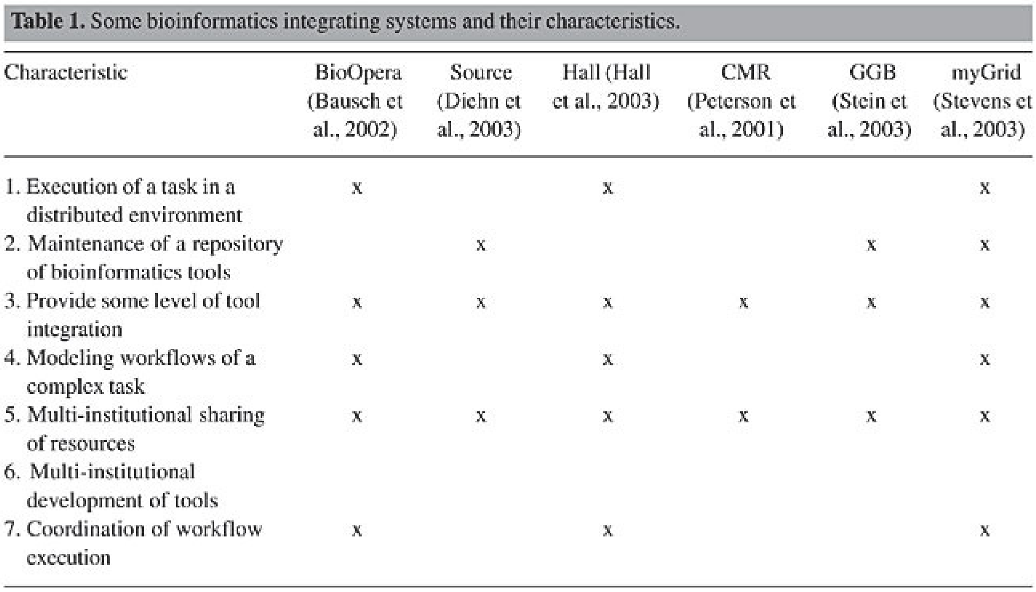

**Table 1.** Some bioinformatics integrating systems and their characteristics.

| Characteristic | BioOpera (Bausch et al., 2002) | Source (Diehn et al., 2003) | Hall (Hall et al., 2003) | CMR (Peterson et al., 2001) | GGB (Stein et al., 2003) | myGrid (Stevens et al., 2003) |
|---|---|---|---|---|---|---|
| 1. Execution of a task in a distributed environment | x |  | x |  |  | x |
| 2. Maintenance of a 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 |

Systems that provide a set of bioinformatics tools make their tools available via Web sites and/or via a local program (Bausch et al., 2002; Eckart and Sobral, 2003; Hall et al., 2003; Stevens et al., 2003). Usually these tools are developed according to specific standards, hampering the integration of tools built by different groups. The framework of Eckart and Sobral (2003) employs Web services in the server side and an application at the client side. When the application is initiated, 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 the 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 into one generic data model (Peterson et al., 2001). Other systems aim at the modeling of any genomic project via a set of basic components (Stein et al., 2003). Finally, there are systems that link several kinds of services and data to facilitate the genomic annotation of a specific genome project (Diehn et al., 2003). Some systems handle the problem of tools integration. The specification of task interactions and interdependency relations is typically designed using workflows (Bausch et al., 2002; Hall et al., 2003; Stevens et al., 2003). 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 (LBI, 2004). Whenever a new genome project is started, the scientists need to adapt the entire framework. The second kind contains frameworks formed by basic components (Stein et al., 2003). 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 cannot be accessed by other projects.

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

**Related issues**

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

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” (W3C, 2004). Some open topics are how to discover adequate services and service providers, how to automate Web service integration, how to minimize semantic ambiguity in service specifications and how to assign information about quality and reliability of the services offered by a provider (Alonso et al., 2004). We have concentrated on the 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 involve BPEL4WS (IBM, 2003), OWL-S (The OWL Service Coalition, 2003) and WSCI (W3C, 2002). We have adopted BPEL4WS as a basis for specifying service orchestration.

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 interdependency relations, inputs and outputs (Seffino et al., 1999).

Bioinformatics workflows are scientific workflows, i.e., they differ from a usual workflow because they have some additional characteristics, such as a high degree of flexibility, uncertainty and existence of exceptions (Wainer et al., 1996).

Our work is concentrated on the execution of scientific workflows through the orchestration of Web services and user interaction. Open problems that have been attacked include communication protocols and interfaces among services to specify a workflow.

There are many tools and databases for bioinformatics. Samples of tools include BLAST (Altschul et al., 1997), Phred (Ewing and Green, 1998) and Consed (Gordon et al., 1998). These tools are geared towards sequence comparison, analysis and visualization. Other complex problems with dedicated tools involve fragment assembly of DNA (alignment and consensus), phylogenetic trees, database search, etc. Our research concentrates on the applications for assembly, annotation and comparison of genomes. The choice of these applications was based on prior experience of the Laboratory for Bioinformatics (LBI) at UNICAMP in assembly and genomic annotation. This choice is also common to several bioinformatics efforts using workflows (Meidanis et al., 1996), clusters (Bausch et al., 2002) and grids (Stevens et al., 2003).

## THE PROPOSED FRAMEWORK

The framework manages the design and execution of scientific workflows that will support the 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 the various 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 the 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 that there are two available workflows, and the user can then choose which workflow to execute. Workflow activities embed tools that are executed via services; data are also made available via services.

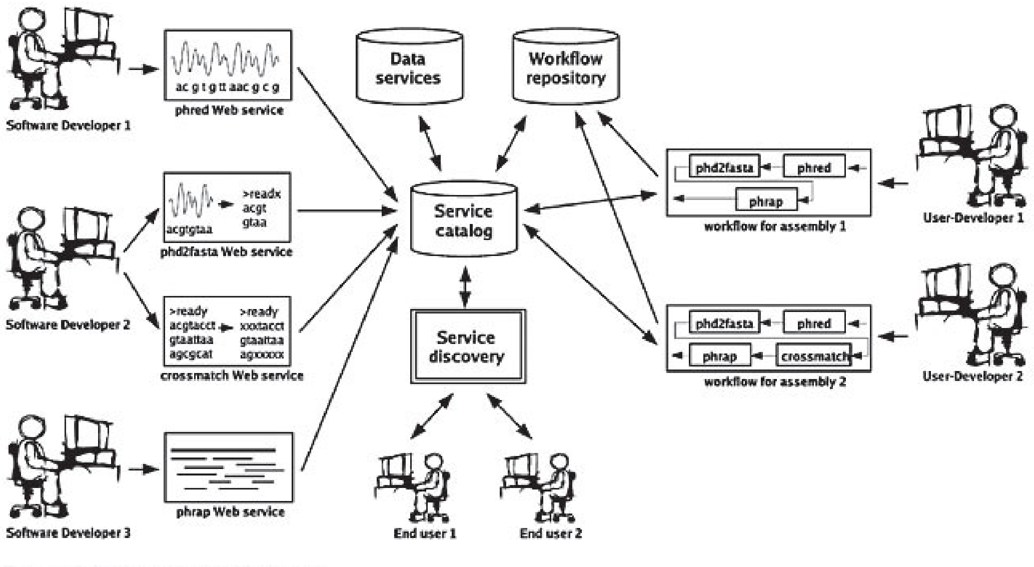

Figure 1. User interaction overview.

Figure 2 shows our system architecture, which supports the integration of the tasks shown in Figure

1.

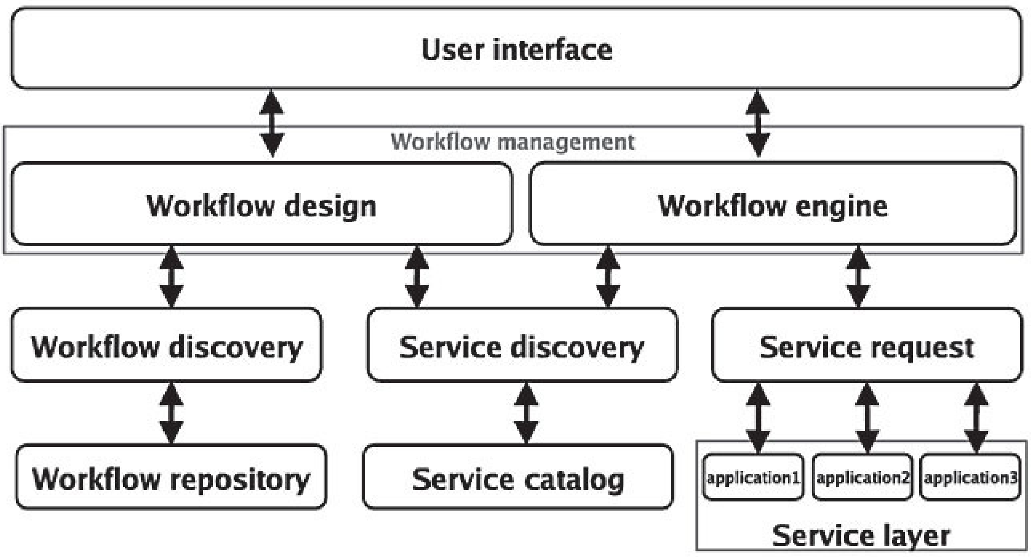

Figure 2. Framework architecture.

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 compositions to provide more sophisticated functionalities - e.g., genomic comparison and gene family operations. We started by transforming modules already available in LBI into services.

The Service catalog layer is responsible for storing Web services' syntactic and semantic 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 service availability and allow the reuse of workflows.

Service discovery can be accomplished in several ways. Our framework will allow search by functionality, context and syntax. Search by functionality and context will be done based on semantic data (metadata) assigned to the services. Search for a compatible syntax will be based on the parameters of the service interface. These search methods will use techniques already discussed in the literature (Cardoso and Sheth, 2003; Sycara et al., 2003) 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 service 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 (WFMC, 1999).

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 (Seffino et al., 1999).

Our framework is being specified and developed using a bottom-up approach. We started with the specification and development of bioinformatics basic services, encapsulating LBI tools into services, 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 basic bioinformatics services.

The second stage will be the study and development of techniques for service discovery and request, using syntactic and semantic 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 (WFMC, 1995; Kim, 2003) and tools developed at UNICAMP (Seffino et al., 1999). Here it will be necessary to specify and to implement an orchestration mechanism for these kinds of services (specific to workflows).

Workflow data sources and providers will be encapsulated by services.

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

## 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 on 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. Furthermore, we modeled and implemented a comparative genomic system (Digiampietri et al., 2003). 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.

## ACKNOWLEDGMENTS

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

## REFERENCES

**Alonso, G., Casati, F., Kuno, H.** and **Machiaraju, V.** (2004). *Web Services: Concepts, Architectures and Applications*. Springer, Heidelberg, Germany.

**Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W.** and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*: 3389-3402.

**Bausch, W., Pautasso, C., Schaeppi, R.** and **Alonso, G.** (2002). BioOpera: Cluster-aware Computing. *Proceedings of the 4th IEEE International Conference on Cluster Computing* (*Cluster'02*).

**Cannataro, M., Comito, C., Guzzo, A.** and **Veltri, P.** (2004). Integrating ontology and workflow in PROTEUS, a grid-based problem solving environment for bioinformatics. *Proceedings of the International Conference on Coding and Computing 2*: 90-94.

**Cardoso, J.** and **Sheth, A.** (2003). Semantic e-Workflow Composition. *J. Intelligent Information Systems 21*: 191-225.

**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.** and **Alizadeh, A.A.** (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res. 31*: 219-223.

**Digiampietri, L.A., Medeiros, C.M.B.** and **Setubal, J.C.** (2003). A data model for comparative genomics. *Proceedings of the 2nd Brazilian Workshop on Bioinformatics*, 38-46.

**Eckart, J.D.** and **Sobral, B.W.** (2003). A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *Spring 7*: 79-88.

**Ewing, B.** and **Green, P.** (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res. 8*: 186-194.

**Foster, I.** and **Kesselman, C.** (1999). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco, CA, USA.

**Gordon, D., Abajian, C.** and **Green, P.** (1998). Consed: a graphical tool for sequence finishing. *Genome Res. 8*: 195-202.

**Hall, D., Miller, J.A., Arnold, J., Kochut, K.J., Sheth, A.P.** and **Weise, M.** (2003). Using workflow to build an information management system for a geographically distributed genome sequencing initiative. In: *Genomics of Plants and Fungi* (Prade, R.A. and Bohnert, H.J., eds.). Marcel Dekker, New York, NY, USA, pp. 359-371.

**IBM** (2003). Business Process Execution Language for Web Services Version 1.1 (BPEL4WS). [http://www-106.ibm.com/developerworks/library/ws-bpel/]. Accessed July 12, 2004.

**IBM** (2004). Web Service for Bioinformatic Analysis Workflow. [http://www.alphaworks.ibm.com/aw.nsf/reqs/wsbaw]. Accessed July 12, 2004.

**Kim, K.H.** (2003). Workflow dependency analysis and its implications on distributed workflow systems. *17th International Conference on Advanced Information Networking and Applications*, 677-682.

**LBI** (2004). Laboratory for Bioinformatics, Institute of Computing, University of Campinas. [http://www.lbi.ic.unicamp.br]. Accessed July 15, 2004.

**Meidanis, J., Vossen, G.** and **Weske, M.** (1996). Using workflow management in DNA sequencing. *Proceedings of the 1st International Conference on Cooperative Information Systems*, 114-123.

**Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K.** and **White, O.** (2001). The comprehensive microbial resource. *Nucleic Acids Res. 29*: 123-125.

**Seffino, L.A., Medeiros, C.B., Rocha, J.V.R.** and **Yi, B.** (1999). WOODS - a spatial decision support system based on workflows. *Decision Support Systems 27*: 105-123.

**Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A.** and **Lewis, S.** (2003). The generic genome browser: a building block for a model organism system database. *Genome Res. 12*: 1599-1610.

**Stevens, R., Robinson, A.** and **Goble, C.A.** (2003). myGrid: Personalised bioinformatics on the information grid. *Bioinformatics 19*: 302-304.

**Sycara, K., Paolucci, M., Ankolekar, A.** and **Srinivasan, N.** (2003). Automated discovery, interaction and composition of semantic Web services. *Web Semantics: Science, Services and Agents on the World Wide Web 1*: 27-46.

**The OWL Services Coalition** (2003). OWL-S: Semantic Markup for Web Services. [http://www.daml.org/services/owl-s/1.0/owl-s.html]. Accessed July 13, 2004.

**Vouk, M.A**. (2003). Integration of heterogeneous scientific data using workflows - a case study in bioinformatics. *Proceedings of the 25th Internatinal Conference Interfaces 16-19*: 25-28.

**W3C** (2002). Web Service Choreography Interface (WSCI) 1.0. [http://www.w3.org/TR/wsci]. Accessed July 12, 2004.

**W3C** (2003). Web Services Internationalization Requirements. [http://www.w3.org/International/ws/ws-i18n-scenarios-edit/ws-i18n-requirements-edit.html]. Accessed July 12, 2004.

**W3C** (2004). Extensible Markup Language (XML) 1.0 (3rd edn.). [http://www.w3.org/TR/2004/REC-xml-20040204]. Accessed July 12, 2004.

**Wainer, J., Weske, M., Vossen, G.** and **Medeiros, C.B.** (1996). Scientific Workflow Systems. *Proceedings of the NSF Workshop on Workflow and Process Automation Information Systems.*

**WFMC** (1995). The Workflow Reference Model. *Technical Report TC-1003*.

**WFMC** (1999). Workflow Management Coalition Terminology Glossary (Issue 3.0). [http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf]. Accessed July 14, 2004.