Distributed Execution of Workflows in the INB

Ismael Navas-Delgado[1], Antonio J. Pérez[2],
Jose F. Aldana-Montes[1], and Oswaldo Trelles[2]

[1] University of Malaga, Computing Languages and Computer Science Department,
29071, Malaga, Spain
{ismael, jfam}@lcc.uma.es
http://khaos.uma.es
[2] University of Malaga, Computer Architectures Department,
29071, Malaga, Spain
ots@ac.uma.es, aperezp@uma.es
http://www.ac.uma.es

Abstract. Our workflow platform offers a view of the different tools available as a single and uniform pool of services readily available for enhancing query processing. This proposal is based on an architecture for publishing biological data and services, and is designed to be a flexible client for making use of Bio-MOBY servers, extending them with persistency of the information retrieved for each user. We also present in this paper some biological results, which have been obtained by taking advantage of the proposed workflow execution system. This work has been developed and implemented in the National Institute for Bioinformatics (INB) in Spain (available at http://www.inab.org/MOWServ).

1 Introduction

A Web-based service facilitates access to remote resources promoting the development and availability of highly diverse and specific tools. These new resource capabilities are of special interest in the bioinformatics domain where a variety of databases and services are required in order to produce a more complete view of biological problems. Unfortunately, the common bioinformatics research field becomes hard to operate since it can involve finding appropriate web services by collecting URLs of the useful ones, selecting the most popular or suitable ones, getting familiar with their specific interfaces (e.g. see the very popular sites like NCBI, EBI, ExPASy, etc.), copying and pasting data, manually selecting and combining partial results, and scheduling and pipelining tasks by-hand. Thus, the main component of a bioinformatician's daily work is to carry out a set of simple activities which they usually perform using a diverse set of tools for solving a problem. This implies interacting with different interfaces and storing partial results for use in another tool. This manual interaction with services is costly and error-prone.

To take full advantage of the amount of information available, researchers need to be able to access, link, combine, and query these biological data sets easily and efficiently, and then to integrate the significant number of tools which use these data sources. To address this problem, a growing effort is being made to develop common

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 224 – 231, 2006.

data-interchange methods, common reference ontologies and automated query engines. Data and service integration has become of particular interest in bioinformatics due to the potential payoff in terms of improved efficiency. Several groups have addressed general solutions for such integration infrastructures: TAMBIS [1], Model-Based Mediation [2], BioDataServer [3], PAT [4], BioBroker [5] and BioMoby [6]. However, these infrastructures have to be inter-related and this process requires human interaction, thereby unnecessarily increasing the time required to obtain a solution to the proposed problem, when the process could in fact be automated.

Thus, a workflow-based system is very useful when the tasks that the user needs to solve are (as is usually the case) predefined and relationships between the tasks are well known. Our proposal provides an execution environment for related tasks, so that users can develop a workflow defining a set of related tasks in order to solve a specific problem, and subsequently execute the workflow with specific inputs. This process can be repeated several times with different inputs in order to derive biological conclusions.

In this paper we focus on the description of a workflow management environment that provides a set of applications for storing, executing and monitoring workflows defined by means of XML files in the Scufl [7] representation language. Our proposal extends this approach with the capability of using authentication-based systems, in which data confidentiality is ensured. Besides the use of a scheduler, which has statistics about the services published in the system, our proposal offers an optimized execution process, optimized error handling and standardized view of data.

Our proposal is presented in Section 2, each element being described in detail. Section 3 illustrates a workflow example performing a homology search and a phylogenetic study. It will also be shown how by using this workflow we can obtain biological knowledge about a family of proteins. Finally we round off with some conclusions and future work.

2 System and Methods

The National Institute for Bioinformatics (INB) in Spain has addressed the integration problem in bioinformatics through the design of a simple, dynamic and extensible platform in order to represent, recover, process, integrate and discover knowledge. The description of biological input/output objects is coordinated and standardized by means of an object-ontology in such a way that services can communicate with each other, wiring natural bioinformatics workflows. Automatic interfaces and help system builders have been incorporated into the architecture to make it more cohesive and to facilitate user communication. Beyond traditional bioinformatics platforms, data persistence system, user management and scheduling abilities have created a new generation of bioinformatics platforms.

The INB system architecture is organized at three main levels: (a) a web-interface at the top of the architecture facilitates communication between the user and the platform (b) the architecture core including the services' interface though bioMOBY API; and (c) at the bottom of the scheme the services' providers.

A web interface manages user sessions with an authentication mechanism. An automatic web interface builder is able to dynamically build on interfaces for