Bioinformatics Web Portals

Mario Cannataro
Università "Magna Græcia" di Catanzaro, Italy

Pierangelo Veltri
Università "Magna Græcia" di Catanzaro, Italy

INTRODUCTION

Bioinformatics involves the design and development of advanced algorithms and computational platforms to solve problems in biomedicine (Jones & Pevzner, 2004). It also deals with methods for acquiring, storing, retrieving and analysing biological data obtained by querying biological databases or provided by experiments. Bioinformatics applications involve different datasets as well as different software tools and algorithms. Such applications need semantic models for basic software components and need advanced scientific portal services able to aggregate such different components and to hide their details and complexity from the final user. For instance, proteomics applications involve datasets, either produced by experiments or available as public databases, as well as a huge number of different software tools and algorithms. To use such applications it is required to know both biological issues related to data generation and results interpretation and informatics requirements related to data analysis.

Bioinformatics applications require platforms that are computationally out of standard. Applications are indeed (1) naturally distributed, due to the high number of involved datasets; (2) require high computing power, due to the large size of datasets and the complexity of basic computations; (3) access heterogeneous data both in format and structure; and finally (5) require reliability and security. For instance, applications such as identification of proteins from spectra data (de Hoffmann & Stroobant, 2002), querying of protein databases (Swiss-Prot), predictions of proteins structures (Guerra & Istrail, 2003), and string-based pattern extraction from large biological sequences, are some examples of computationally expensive applications. Moreover, expertise is required in choosing the most appropriate tools. For instance, protein structure prediction depends on proteins family, so choosing the right tool may strongly influence the experimental results.

Recently, there has been much interest from database community and computer science community for bioinformatics. Nevertheless, what is still missing is a high-level environment able to classify tools and provide Web-based easy to use application programming interfaces. In such a way, users can concentrate on the logic of application (i.e., biological aspects) leaving to such platform the work to compose applications, format input data, provide options and parameters, and collect results.

Another important requirement is the accessibility of such platform through a Web portal, that is, by using the user interfaces and protocols of the World Wide Web. A bioinformatics Web portal is thus a Web portal that allows access to bioinformatics tools and databases through a Web browser. Moreover, due to the complexity, diversity and a huge number of bioinformatics tools and databases, a bioinformatics Web portal should also support problem formulation, application composition and execution, results visualisation and annotation. A possible approach to solve these issues —high-level modeling and Web-based user interfaces—can be obtained by adding semantics links between biological problems and bioinformatics resources through ontologies (Baker, 1998), and by decoupling Web-based user interfaces from high-performance back-end platforms.

In this article we review main requirements of distributed bioinformatics applications and related bioinformatics Web portals, and report the proposal of a grid-based bioinformatics portal allowing choosing and composing of bioinformatics tools with the help of a domain ontology describing data and software resources.

BACKGROUND

Bioinformatics researchers, among the other directions, are investigating through: (1) data modeling to manage heterogeneous datasets (e.g., see HUPO, n.d., the HUPO, Human Proteome Organization—Proteomics Standard Initiative); (2) specialised services for protein sequences searching, and data mining techniques to extract meaningful information from datasets; (3) ontologies and metadata for a high-level description of the goals and requirements of applications; and (5) high performance computational platforms to execute distributed bioinformatics applications.

Many applications have been defined to support biological researchers for solving problems on different topics where large computing power is required. Grid community (Foster & Kesselman, 2003) has recognised that bioinformatics and postgenomic applications are both a challenge but especially

an opportunity for distributed high performance computing and collaboration. The Life Science Grid Research Group of the Global Grid Forum (see LSG, n.d.) aims to investigate how bioinformatics requirements can be fitted and satisfied by grid services and standards, and vice versa, what new services should grids provide to bioinformatics applications. Some bioinformatics grids projects are also appearing, for example, the EuroGrid project (EuroGrid, n.d.), the Bio-GRID work package (Bio-Grid, n.d.) used to access portal for biomolecular modeling resources, the *myGrid* (Stevens, Robinson, & Goble, 2003) system, and the Asia Pacific Grid (AsiaGrid, n.d.).

In recent years many platforms for developing bioinformatics applications, some of which dealing with ontologies and workflows, have been developed. Systems as *SpecAlign* (Wong, Cagney, & Cartwright, 2005), *MSAnalyzer* (Sashimi, n.d.), and those developed in Jeffries (2005), are all specialised in preprocessing, visualisation, and analysis of specialised datasets, that is, mass spectrometry data, but they do not support analysis of data and workflows composition, nor include domain ontologies. *LabBase* (Goodman, 1998) and similar laboratory information management systems are useful to manage experiments conducted in laboratory and related data, but are inadequate to support sophisticated analysis. More sophisticated bioinformatics platforms, like the *genomics research network architecture* (gRNA) (Laud, Bhowmick, Cruz, Singh, & Rajesh, 2002) and the *Pegasys* (Shah et al., 2004) bioinformatics system, offer some sort of configurable engine to pipeline a set of tasks and data. A special attention merits *myGrid* (Stevens et al., 2003), a powerful toolkit to build workflows of Web services that offers a large set of bioinformatics tools wrapped as Web services, leverages ontologies, and uses the powerful *Taverna* workflow editor (Oinn et al., 2004). General purpose workflow editors (see Yu & Buyya, 2005, for a survey), such as *Kepler*, *Pegasus*, and *Triana*, are all suitable to support the composition of bioinformatics workflows, but few of them use ontologies.

Finally, some bioinformatics Web portals are also appearing. Such systems, some of which are described in the following, offer a collection of bioinformatics tools and provide access to local and remote biological databases through a Web-based interface, but a few of them offer a machine-understandable semantic classification of the tools nor gives support for the design of complex workflows of such tools. The ExPASy (n.d.; Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics is dedicated to the analysis of protein sequences and structures (ExPASy). The grid protein sequence analysis (GPSA) is an integrated grid portal devoted to molecular bioinformatics and offers a user-friendly interface for the grid genomic resources on the EGEE grid. The Helmholtz Network for Bioinformatics (HNB, n.d.) offers access to numerous bioinformatics resources provided by many German bioinformatics research groups through a single Web portal. Mobyle (Neron, Tuffery, & Letondal, 2005), is an environment for running and defining bioinformatics analyses whose main objective is to enable biologists to access advanced features, such as pipelines or remote services discovery, without having to learn complex concepts nor installing sophisticated software.

In summary, an important trend in bioinformatics environments regards the increasing use of ontologies to model basic building blocks and the use of workflow systems to ease the application development and execution process in a distributed setting such as the grid. The decoupling between user interface and execution back-end is another important trend to move such environments toward bioinformatics Web portals.

REQUIREMENTS OF BIOINFORMATICS WEB PORTALS

From a computational point of view, bioinformatics applications present the following requirements:

1. They are often distributed, due to the high number of involved datasets.
2. They require high computing power, due to the large size of datasets and the complexity of basic computations.
3. They access heterogeneous data, where heterogeneity is in data format, access policy, distribution, and so forth.
4. They could access private data, thus should be based on a secure software infrastructure.

Current biological and biomedical research, for example, genomics and proteomics, makes full use of a plethora of tools and databases that address specific problems such as nucleotide/protein sequence alignment (e.g., see BLAST, n.d.), protein structure prediction, protein docking, mass spectrometry-based protein identification (e.g., see MASCOT, n.d.), molecule visualisation (e.g., see RasMol, n.d.), and so forth. Although many of those tools and databases are made available on the Internet, often researchers use them in a stand-alone way and if an experiment needs a composition of such tools, users need to manually insert input data and collect output results that in turn are used to feed another tool.

Current bioinformatics Web portals are just a collection of those tools and the more sophisticated provide also access to remote databases, but they do not offer support for the design and execution of complex "in silico" experiments.[1] Thus, next-generation bioinformatics Web portals need to support the entire lifecycle of in silico experiments, that is: