

Chapter 4.3

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

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/bioinformatics-web-portals/7970

Related Content

RDF(S) Store in Object-Relational Databases

Zongmin Ma, Daiyi Li, Jiawen Lu, Ruizhe Ma and Li Yan (2024). *Journal of Database Management* (pp. 1-32).

www.irma-international.org/article/rdfs-store-in-object-relational-databases/334710

Integrity Constraints in an Active Database Environment

Juan M. Aleand Mauricio Minuto Espil (2002). *Database Integrity: Challenges and Solutions* (pp. 113-143).

www.irma-international.org/chapter/integrity-constraints-active-database-environment/7880

Scalable Data Warehouse Architecture: A Higher Education Case Study

Dennis C. Guster, Christopher G. Brown and Erich P. Rice (2018). *Handbook of Research on Big Data Storage and Visualization Techniques* (pp. 340-381).

www.irma-international.org/chapter/scalable-data-warehouse-architecture/198770

Conceptual Data Modeling Patterns: Representation and Validation

Dinesh Batra (2005). *Journal of Database Management* (pp. 84-106).

www.irma-international.org/article/conceptual-data-modeling-patterns/3333

A Concept-Based Query Language Not Using Proper Association Names

Vladimir Ovchinnikov (2006). *Advanced Topics in Database Research, Volume 5* (pp. 374-400).

www.irma-international.org/chapter/concept-based-query-language-not/4401