Mining and Visualization of Logs of Bioinformatics Web Services *in silico* Experiments

Sérgio Manuel Serra da Cruz
Núcleo de Computação Eletrônica - Universidade Federal do Rio de Janeiro
serra@nce.ufrj.br

Abstract

The visualization of the bioinformatics datasets allow e-scientists to gain insight into the data and come up with new scientific hypotheses. So, in order to explore this opportunity we present VM-BioWSLogA (Visual-Mining Bioinformatics Web Services Log Architecture), a framework which relies on mining and on further visualization of Web Services log data captured on bioinformatics in silico experiments.

1. Introduction

Genome data and genome analysis initiatives are growing fast over the last years, giving rise to huge amounts of data available over the Internet [1]. Such amount of data and the complexity involved to analyze them have originated the area of Bioinformatics, where *in silico* scientific experiments are applied to solve Biological problems, encompassing multiple combinations of computational resources and mathematical algorithms. In bioinformatics environments, program composition is a frequent operation, requiring complex management [2]. A scientist faces many challenges when building an experiment: finding the right program to use, the adequate parameters to tune, managing input/output data, building and reusing workflows, and last but not least the visualization of the results in order to enhance its cognition about the problem. The emergence of Web services technology represents a significant contribution to the reuse of scientific applications, since it provides unprecedented infrastructure for connecting otherwise isolated computing resources. Recently, they have been pointed out in the bioinformatics area as a potential technology to allow heterogeneous distributed biological data to be fully exploited [2,3]. Another feature is related to the data's visualization, where huge amounts and a wide variety of formats are used in different scientific experiments; it often requires transformations and mining before it can be visualized on suitable tools.

2. Visualization of Web Services Utilization Logs

Recording the execution context of Bioinformatics Web services experiments into proper logs is an important phase of a scientific workflow execution. A concrete usage of a Web services log is to help e-scientists to avoid redundant efforts when repeating experiments. Usually, Bioinformatics tools have several input parameters, which can modify the behavior of their algorithm, and consequently modify the service results [5]. Another issue regarding service monitoring in bioinformatics environments is that e-scientists need efficiency. Some queries consist of hundred of sequences at a time and can take several days to run. Besides, due to the confidential nature of such experiments, it may be useful to keep of track of securities issues, such as recording the scientists that are submitting queries and when. Finally, e-scientists should also keep track program outputs they have used if they are willing to obtain faster results and able to reproduce the same experiments in another occasion. In order to solve those issues, we previously presented an architecture named BioWSLogA which supports a flexible log generation without changing the code of existing services, generating XML based logs repositories about services execution [5]. It can store complex and heterogeneous data structures and, at the same time, describe them through encapsulated metadata. In spite of this facility, e-scientists still need to understand the results of Bioinformatics experiments and workflows, so we propose the use of visualization techniques to help them to elucidate structures in complex datasets. This approach can play an important role in exploratory data analysis, where visual representations can help them to build up an understanding of the content of their *in silico* datasets and experiments.

The goal of this paper is to present a Visual Web Services Log Mining tool, named VM-BioWSLogA (Visual Mining-Bioinformatics Web Services Log Architecture), which relies on mining and on visualization of Web Services log data captured in a Bioinformatics environment. The contribution of VM-BioWSLogA is two-folded: first, it provides a human visual perception of biological experiments towards e-scientist personalization and auditing, supporting a new range of experimental strategies; and second, it also supports refined investigations about services quality monitoring, addressing administrative issues like performance, security and availability. VM-BioWSLogA uses both Web server logs and intercepted SOAP messages recorded by BioWSLogA [4, 5].

3. A Visual Web Services Log Mining Architecture

VM-BioWSLogA proposal outlines e-scientists needs for gathering knowledge about their *in silico* experiments. The architecture's prototype can be used to amplify the perception of rules, patterns, regularities and behaviors, It aids e-scientists to visualize four different aspects of Bioinformatics experiment data sets, such as: (i) the usage of suitable experiment parameters; (ii) A simple view of Bioinformatics Web services composition; (iii) A easy way to audit and track the Web services utilization; (iv) A feasible way to keep track of data provenance.

VM-BioWSLogA, a multi-layered architecture capable to deal with both Web services XML based logs and traditional Web server logs as input data. *The integration layer* is set of programs used to prepare data for further processing. For instance: extraction, cleaning, transformation and loading. This layer uses XQuery, XSLT and XML Schemas to feed the data repository, i.e., relational or XML native database. The Web server log parser component is used to parse and transform plain ASCII files produced by a Web server to a standard XML format. This component is important to make the architecture independent from the Web server supplier.

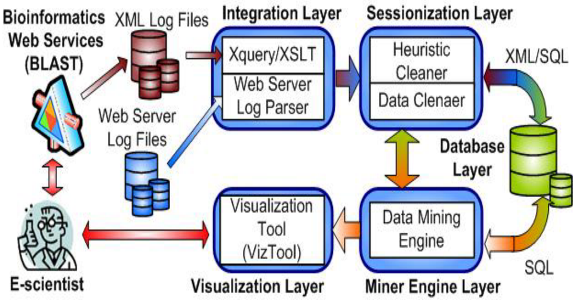

Figure 1 – An overview of VM-BioWSLogA

The *sessionization layer* is used to tie the instances of Web services and Web pages (through database foreign keys) to sessions and to user. This layer is important to investigate the usage of the Web services composition used through users sessions. The *Database layer* is a repository of input/output Bioinformatics experiments data. It also stores pre-processed logs, e-scientist sessions, and informations about the Web services execution.

The *Miner Engine Layer* is a data mining engine and is in charge of bulk loading XML data from database, executing SQL commands against it and execute the mining algoritms. *Visualization Layer* should be used to present implicit and useful knowledge from *in silico* experiments and Web services usage and composition. Data can be viewed at different levels of granularity and abstractions as paralled coordinates graphs [6, 7]. This visual model easily shows the interelationship and dependencies between different *n*-dimensions like users, experiments, services, parameters and results. Interactively, the model can be used to discover sensitivities and to do approximate optimization, provinding a simple decision support environment. The *Viztool* was implemented with Tcl/Tk scripts, we also extended some classes of the VTK package in order to present smarter 3D visualizations, such as the ones used to compare the results of Bioinformatics parameters, collate actual services composition patterns to expected patterns, or more generally, to compare and track services utilization.

4. Conclusion and Future Work

As far as we are concerned, there are no other initiatives of visualization of Bioinformatics Web Services logs or Services composition. VM-BioWSlogA is being tested with data originated by a collection of real world Bioinformatics Web services; we are involved in refining the architecture, which was implemented as java prototype using Tomcat/Axis as SOAP engine and VTK with its C++ classes as the visualization layer which can run on a variety of platforms. Future work will involve the tight integration of a number of more refined visualization techniques with traditional techniques from such disciplines as statistics, Data Warehousing and Web Mining [7]. Our ultimate goal lies on the enhancement of mining algorithms in order to bring the power of visualization technology to e-scientists desktops and allow a better, faster and more intuitive and cognitive exploration of experimental biological dataset resources.

10. References

[1] Hernandez, T., Kambhampati, S. "Integration of Biological Sources: Current Systems and Challenges Ahead", SIGMOD Record, 2004,Vol.33, No 3, 10 pp.

[2] Cavalcanti, M. C. ; Targino, R. ; Baião, F. Rössle S. C., Bisch, P. M., Pires P. F, Campos M. L .M, Mattoso M. "Managing Structural Genomic Workflows using Web Services". In Data & knowledge engineering, Elsevier, 2005, v. 53, n. 1, pages. 45-74.

[3] Clark, T. (2003), "Identity and interoperability in Bioinformatics", Briefings in Informatics 4(1), 2005, pages. 4-6.

[4] Cruz, S. M. S., Campos, M. L. M. Pires, P. F. Campos, L. M. "Monitoring e-business Web services usage through a log based architecture" In 2nd IEEE International Conference on Web services. San Diego, 2004, pages 61-67.

[5] Cruz, S. M. S., Campos, M. L. M. Pires, P. F. Dávila, A, Campos, L. M. "Monitoring Bioinformatics Web services Requests and Responses through a Log Based Architecture",2005 In XXXII SEMISH-SBC, São Leopoldo, 2005 pages 1787-1801.

[6] Inselberg, A. "Multidimensional detective", In IEEE Symposium on Information Visualization, 1997, vol. 00, p. 100-110 .

to visualize a comparison of a predicted gene structure

[7] Ware, C. 'Information Visualization: Perception for Design'. Morgan Kaufmann, New York, 2000.