BioWBI: an Integrated Tool for building and executing Bioinformatic Analysis Workflows

P. Leo[(1)], C. Marinelli[(1)], G. Pappadà[(2)], G. Scioscia[(1)], L. Zanchetta[(2)]

[(1)] Java Technology Center, IBM SEMEA Sud, via Tridente 42/14, 70125 Bari - Italy

[(2)] IBM External Contractor. The work of G. P. and L. Z.. has been supported by IBM through a stage fellowship

Keywords. Bioinformatic platform, Analysis workflow, Data integration, Web services

Introduction

Building integrated bioinformatic platforms is one of the most challenging tasks which Bioinformatics community is dealing with in recent years [1-2]. Facing this task, a number of specific problems arises connected to data integration, integration of specialized tools and algorithms. The solution described in this paper goes in the direction to solve this challenge. It is characterized by two original assumptions: 1) a quite sharp division between the data realm of a bioinformatics analysis and its components in terms of algorithms and processes, 2) the conception of a rigorous algebra that allows researchers to formalize their analyses in terms of *atomic process workflows*. As a result of this approach two bioinformatics web tools, *BioWBI* and *WEE*, have been designed and prototyped by our group to provide researchers with a virtual collaborative workspace in which defining their data-sources, drawing graphically as well as executing analysis workflows. These tools constitute the basic components of a much more general bioinformatic e-workplace.

Methods

1. Datasources

We name *datasources* all the sets of data used for bioinformatic analyses (biosequences or any other structured biological information). A simple categorization allows us to catalogue three different main datasource-types:
1) *Flat files*, i.e. files containing biosequences and information associated to them, stored in specific formats (e.g. FASTA, EMBL, GCG, etc.), or any other kind of structured pieces of information.
2) *Static datasources* consisting of the results saved by a biologist as a consequence of a query performed against public or private specialized databases.
3) *Dynamic datasources*, so called because each of them is not the result of a query, but the query itself. In this way, every time the biologist wants to use a dynamic datasource, the database is queried using the saved query.
Obviously different types of datasources can also be combined, so in any case the researcher can construct, in a suitable way, appropriate datasources for his analysis processes.

2. Algebra

Typical common exigencies of a biologist are: the possibility of executing the same algorithm on *different input files* through a single submission; executing the same algorithm on the same input letting the value of a parameter *change at each run*; executing in an appropriate order *analysis processes* consisting of more than one algorithm; *choosing the algorithm* to execute on the basis of the input data type; *suspending* the execution of an analysis process in order to verify the effectiveness of the partial results.
An appropriate algebra has been conceived to cope all the issues emerged above. It consists of a certain number of operators and rules that rigorously describe and constrain the analysis processes (i.e. the *workflows*) a biologist should formulate starting from a set of elementary algorithms and programs available for bioinformatics analyses. The fundamental operators required to describe the needs previously discussed are, in

the same order as the issues expressed above, the *iteration operator* $\mathcal{I}$, the *recursion operator* $\mathcal{R}$, the *pipe operator* $\mathcal{P}$, the *conditional operator* $\mathcal{C}$ and finally the *break operator* $\mathcal{B}$. By using an appropriate combination of these operators and their related rules, the biologist has the possibility to design process workflows as complex as he wants.

3. Algorithms Knowledge Base

In the developed platform, we integrated algorithms belonging to some of the largely diffused bioinformatic packages.For integration purposes, a XML file descriptor encapsulating the knowledge about each algorithm properties and features (name of its parameters, allowed values for them, etc.) has been setup. Such an approach, already used in previous solutions [3-4] although in a more simplified manner, makes simple the task of integrating new algorithms or updating existing ones. The set of these descriptors constitutes what we call the Algorithms Knowledge Base.

Tools

All three elements described before are the backbone of a bioinformatics tool able to assist researchers to build and execute their analysis process workflows. This tool consists mainly of two web-applications: the first, **Bio**informatic **W**orkflow **B**uilder **I**nterface (BioWBI), is a web application that supplies a web interface to create datasources and design workflows; the second one, **W**orkflow **E**xecution **E**ngine (WEE), is a back-end application able to elaborate and execute the submitted analysis requests from the first tool and give back results. These two applications communicate each other through Web Services technology. A schematic picture of such an architecture is shown in Fig. 1, where the synergic activity of both applications emerges. Both web-tools have been implemented by using Java Technology, adopting open standards running in a pure open source environment.

From a functional point of view, by using a web-browser, end-user can access to BioWBI[1], entering a working area consisting of three main sections where he can, respectively, manage the input datasources; design/re-use and execute workflows, and view results of previously submitted analyses.

Being BioWBI a collaborative environment for bioinformatic people, users may also share tools and workflows of common interest with all the platform's users (workflow library).

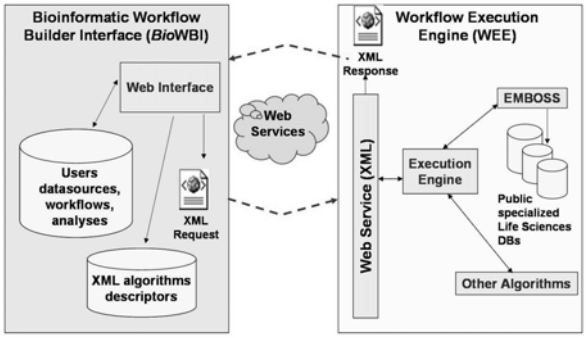

Fig. 1: Architecture for the BioWBI and WEE web-tools

Conclusions

The two basic assumptions, i.e. the decupling of data and processes in bioinformatic analyses and the introduction of the process workflow concept, have conferred flexibility and powerful analysis capabilities to the bioinformatic platform implemented by the web-applications discussed above.

References

[1] Stein, L.D., in *Nature Reviews/Genetics*, Vol. 4, 337-345, 2003.
[2] Limsoon Wong , in *Briefings in Bioinformatics*, Vol. 3, 389-404, 2002.
[3] Letondal, C., in *Bioinformatics*, 17(1), 73-82, 2001.
[4] D'Elia, D. et al., in ECCB 2003 Proceedings (Paris, France, September 27-30, 2003).

[1] A working prototype of BioWBI will be shortly available online from the *IBM alphaWorks* web site: http://www.alphaworks.ibm.com/