An Agent Approach to High Level Workflow Management in Functional Genomics

Giuliano Armano[1], Luciano Milanesi[2], Alessandro Orro[1,2], Eloisa Vargiu[1]
[1] DIEE, University of Cagliari, Piazza d'Armi, I-09123 Cagliari, Italy
[2] ITB-CNR, Via Fratelli Cervi 93, I-20090 Segrate (Milano), Italy

Most tasks releted to the analysis of genomics data can be barely carried out with a single standalone application. Most likely, a combination of several computational tools and data source is required to solve a particular task. Due to the diversity in formats and interfaces and the low diffusion of standard methodologies for data exchange, integration of heterogeneous computational and informative resource is difficult to obtain. The problem of resource integration has been tackled by different points of view: Design and implementation of workflows [2] [3], distribution of bioinformatics web services [5] and grid-oriented integration of computational services [4]. In this work we briefly describe an agent approach to support the composition, execution and management of bioinformatics workflows.

Materials and Methods

Current workflow strategies deal with a network of nodes in which each node represent a particular application. This is an application-oriented view, in which the user is required to know details of each application in order to choose the best parameters and to configure the pipeline. These problems become relevant especially in workflows with a large number of applications.

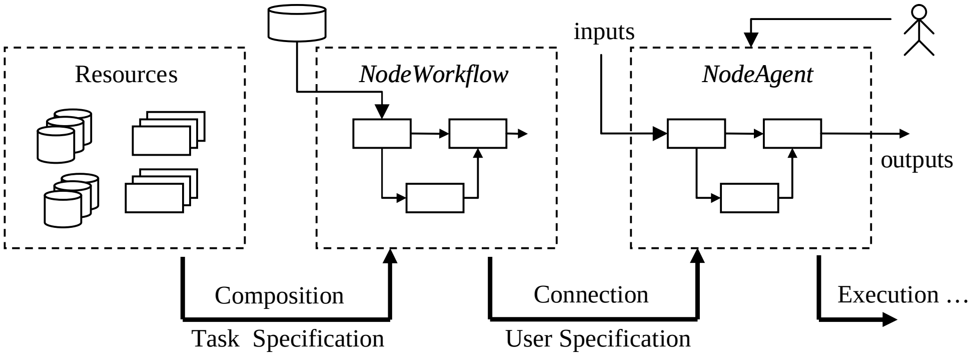

Figure 1: Behavior of the *TaskAgent*

We propose a task-oriented view of a bioinformatics process in which nodes are associated with a particular task, each task being assigned to an agent. In this context an agent (*TaskAgent*) is a software entity that plays the role of a node in a workflow and exports the behavior of the underlying application. Nevertheless, the agent has some additional features: (1) it makes use of knowledge domain information to select a suitable connection of resources for the related task, (2) it exposes only a set of high-level parameters that are more intuitive for the user, (3) it is able to interact with the user at execution time to permit the monitoring of the overall process. Figure 1 shows the behavior of *TaskAgent*s. It composes the underlying workflow with available resources and then creates the user interface in order to satisfy the specification of the task and the user respectively. Then the agent is inserted into the main workflow of the application where it behaves as a normal node in a workflow.

Case Study

We have experimented the above architecture in a bioinformatics workflow used to infer the molecular function of a unknown protein. The complete workflow is composed by four *TaskAgent*s devised to accomplish four tasks (see figure 2): (1) homology search for gathering sequences related to the target, (2) computing multiple alignment, (3) inferring the phylogenetic tree in which the target protein is involved, (4) predicting the protein function. In particular, let us consider with more details the agent associated with the multiple alignment task. It encapsulates several multiple alignment tools and composes a workflow based on the compatibility of the programs with the particular sequences to be aligned. For example in the region of low similarity it can be useful a program for secondary-structure induced multiple alignment [1]. In this case user interaction occurs in the visualization/editing operations. A possible workflow composed by the Multiple Alignment Agent is show in figure 2.

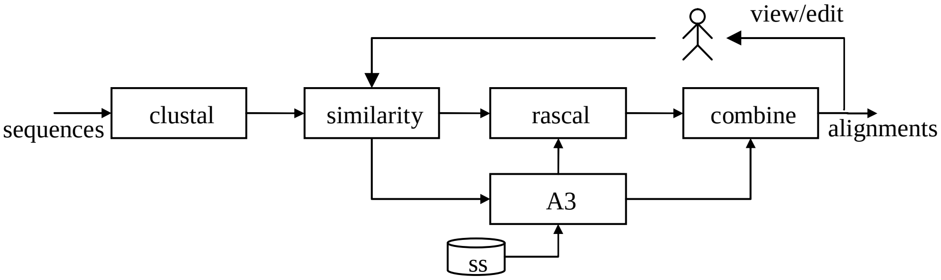

Figure 2: Multiple Alignment Agent composition

Acknowledgment

This work was supported by the MIUR-PNR "Bioinformatics for Genome and Proteome" and LitBio "Laboratory of Interdisciplinary Technologies in Bioinformatics" projects.

References

[1] Armano G, Milanesi L, Orro A. Using Secondary Structure Information to Perform Multiple Alignment. NETTAB 2004. Camerino, Italy.
[2] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004, 22;20(17):3045-54.
[3] Shah SP, He DY, Sawkins JN, Druce JC, Quon G, Lett D, Zheng GX, Xu T, Ouellette BF. Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics 2004, 5:40.
[4] Stevens R, Robinson A, Goble C: myGrid: personalized bioinformatics on the information grid. Bioinformatics 2003, 19:I302-I304.
[5] Wilkinson MD, Links M, BioMOBY: An open source biological web services proposal. Briefings in Bioinformatics. 2002, 3(4):331-41.