Oncology over Internet: integrating data and analysis of oncology interest on the net by means of workflows

P. Romano[1], G. Bertolini[2], F. De Paoli[2], M. Fattore[3], D. Marra[1], G. Mauri[2], E. Merelli[4], I. Porro[5], S. Scaglione[5], L. Milanesi[6,7]

[1]National Cancer Research Institute, Genoa, Italy, [2]University of Milan Bicocca, Italy,
[3]National Research Council, Genoa, Italy, [4]University of Camerino, Italy,
[5]University of Genoa, Italy, [6]National Research Council, Milan, Italy,[7]CILEA, Segrate, Italy

Introduction

The Human Genome Project has transformed biology by providing a list of all genes and proteins, but the field has since then expanded to the management, processing, analysis and visualization of large quantities of data from genomics, proteomics, medicinal chemistry and drug screening. This huge amount of data and the heterogeneity of software tools that are used for its distribution make the tasks of searching, retrieving and integrating the information very difficult. Data is retrieved and analysed by hand by accessing several bioinformatics servers and transferring the data by FTP clients or web browsers by the “cut and paste” technique.

The need is felt for a system that is able to improve the information accessibility. Such a system should be able to automate the accesses to the remote sites, in order to retrieve the information from the specific database or for using the appropriate software tools to achieve the desired analysis. At the same time, it should be able also to "understand" the information that it is managing. Among current ICT technologies, workflow management systems in connection with Web Services seem to be the most promising ones. Workflows are defined as “computerized facilitations or automations of a business process, in whole or part" (Workflow Management Coalition). Their goal is the implementation of data analysis processes in standardized environments and their main advantages relate to effectiveness, reproducibility, reusability of intermediate results and traceability.

Some workflow management systems have already been proposed and are being increasingly applied in the biomedical domain. Some of them are add-ons to other tools, like biopipe [1], a perl module designed to be used with bioperl, and GPipe, an extension of the Pise interface [2]. Other systems are autonomous applications that are being developed either by industries, like the Bioinformatic Workflow Builder Interface – BioWBI from IBM [3], and Pipeline Pilot from SciTegic, or by academic and research institutes, like Wildfire from the Singapore Bioinformatics Institute, and Taverna Workbench [4] from the European Bioinformatics Institute (EBI).

Web Services (WS) are software oriented network services which communicate usually by using SOAP (Simple Object Architecture Protocol, a framework for the distribution of XML structured information) over HTTP. They offer a good, standard solution for automated retrieval of information. Standards are available or have been proposed for their retrieval and identification, description and composition [5]. They allow software applications to access data in a semantic- aware way since being in the form of XML documents their contents can be made visible and when metadata is given, interpretation of semantic information becomes possible..

Many WS and WS deployment tools have recently been set up in the biomedical domain and perspectives for their widespread use has been proposed [6,7]. Among the providers of biology oriented WS are some of the most important network service providers such as the USA National Center for Biotechnology Information (NCBI), which has implemented API interfaces for accessing so-called e-utilities [8], the Bioinformatics Center of the National Cancer Institute [9], the Bioinformatics Center of Kyoto University, which has implemented the KEGG API [10], and the European Bioinformatics Institute (EBI).

We present here a prototype system which can support researchers in the execution of predefined and tested workflows of oncology interest by transparently accessing Web Services. A user-friendly web interface has been developed thus simplifying access to public in-silico analysis.

Methods

We have designed a general architecture of a system (see figure 1) for the remote execution of workflows of biomedical interest that are intended to access to and to retrieve data from various Web Services.

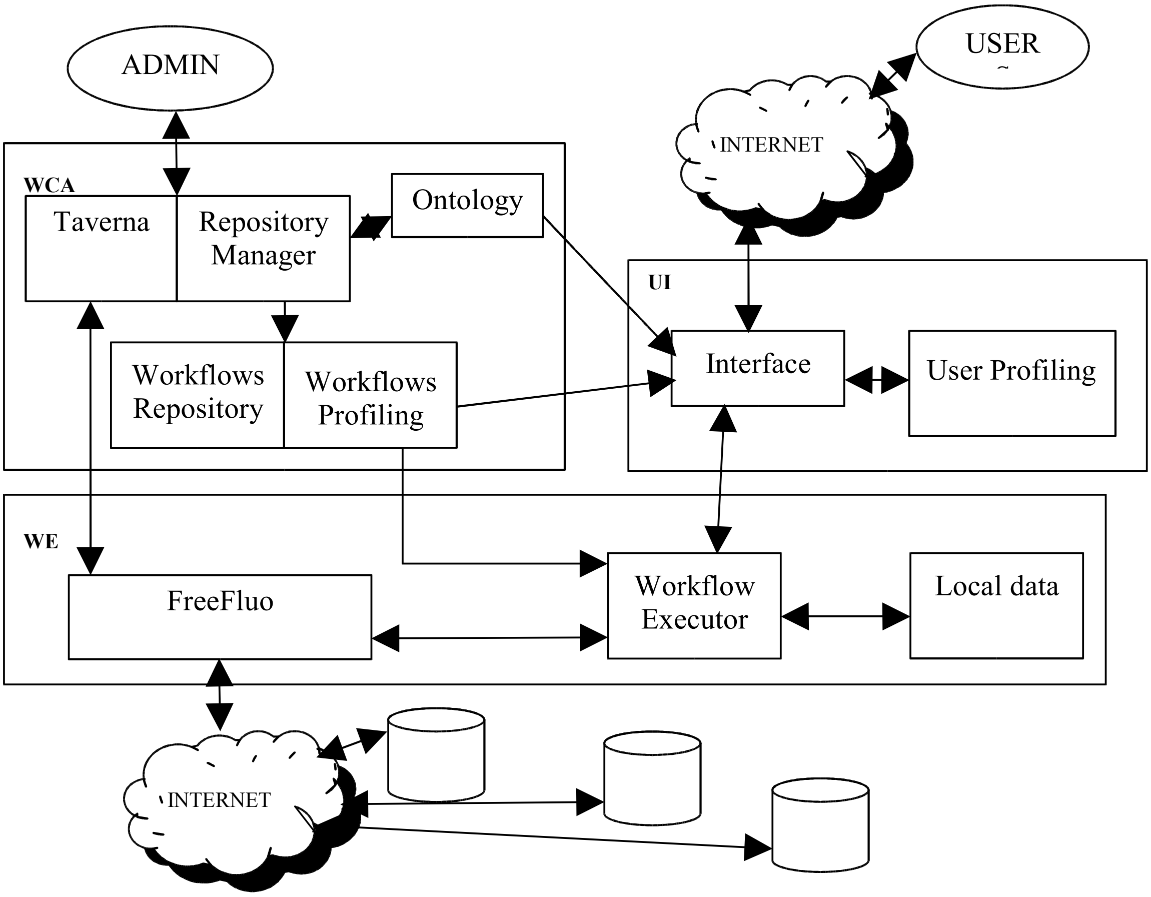


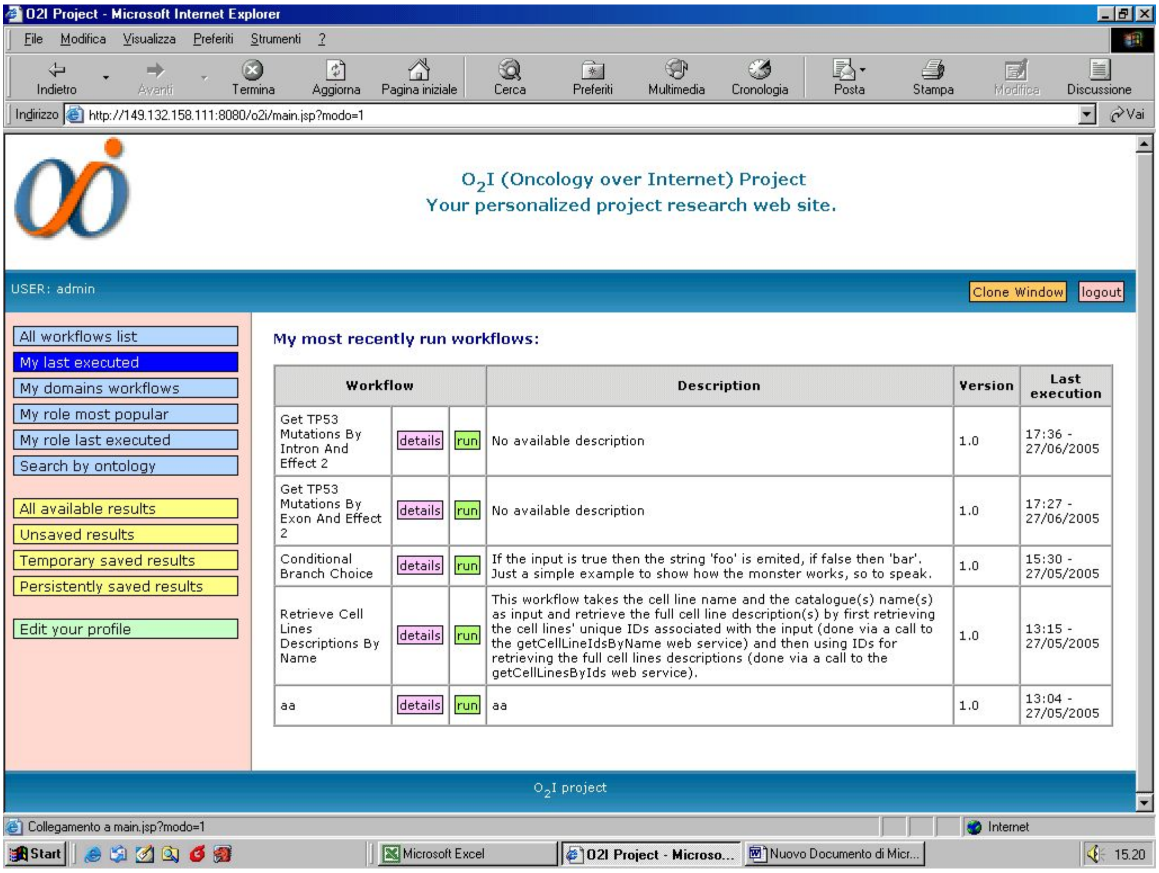
Figure 1: General architecture of the system.

The system is partially based on open source software tools. The Taverna Workbench is a workflow management system developed at the European Bioinformatics Institute (EBI) as part of the ^{my}Grid project [11]. It allows skilled end users to create complex analysis workflows, to access both remote and local processors of various kinds, to run workflows and to display results in different formats. In Taverna workflow execution is carried out by an associated tool, FreeFluo. Taverna also includes an ontology for bioinformatics data. Its only requirement is the availability of Java Run-time Environment (JRE) on a Windows XP or Linux box. A MySQL data management system [12] can optionally be used for local data storage.

The system includes three main blocks: the workflow manager, the user interface and the workflow executor. Workflows are created and tested by an administrator using the Taverna Workbench. They are then stored in a repository by adopting the Taverna Simple conceptual unified flow language (Scufl) format. At the same time, their main processing steps are annotated in workflows profiles by using a specially designed ontology. This ontology describes bioinformatics tasks on the basis of their input and output data, processing type and application domain. These main processing steps may actually represent more real tasks grouped to achieve a functionally significant processing step. The user interface supports end users authentication and profiling and allows for the selection and launch of workflows. Workflow selection can be assisted by the user profile and by searching through significant processors of annotated workflows. Workflows are executed by the third block that is based on FreeFluo and it is also able to store input and output data of actual workflow executions, so that they can later be analysed and possibly reused.

Preliminary results

At present, the general architecture, a preliminary version of the user interface and a set of new Web Services are available. Web Services implement access to IARC TP53 Mutation Database [12,13] and to CABRI catalogues of biological resources [14]. Workflows are being created and tested in various application domains. The ontology is being created starting from the Taverna bioinformatics data ontology.


O2I Project - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Avanti Termina Aggiorna Pagina iniziale Cerca Preferiti Multimedia Cronologia Posta Stampa Modifica Discussione
Indirizzo http://149.132.158.111:8080/o2i/main.jsp?modo=1
Vai
Clone Window
logout
All workflows list
My last executed
My domains workflows
My role most popular
My role last executed
Search by ontology
All available results
Unsaved results
Temporary saved results
Persistently saved results
Edit your profile
Collegamento a main.jsp?modo=1
Internet
Start
Microsoft Excel
O2I Project - Microso...
Nuovo Documento di Micr...
15.20

O_2I (Oncology over Internet) Project
Your personalized project research web site.

USER: admin

My most recently run workflows:

Workflow			Description	Version	Last execution
Get TP53 Mutations By Intron And Effect 2	details	run	No available description	1.0	17:36 - 27/06/2005
Get TP53 Mutations By Exon And Effect 2	details	run	No available description	1.0	17:27 - 27/06/2005
Conditional Branch Choice	details	run	If the input is true then the string 'foo' is emited, if false then 'bar'. Just a simple example to show how the monster works, so to speak.	1.0	15:30 - 27/05/2005
Retrieve Cell Lines Descriptions By Name	details	run	This workflow takes the cell line name and the catalogue(s) name(s) as input and retrieve the full cell line description(s) by first retrieving the cell lines' unique IDs associated with the input (done via a call to the getCellLineIdsByName web service) and then using IDs for retrieving the full cell lines descriptions (done via a call to the getCellLinesByIds web service).	1.0	13:15 - 27/05/2005
aa	details	run	aa	1.0	13:04 - 27/05/2005

O_2I project

Figure 2: list of most recently executed workflows

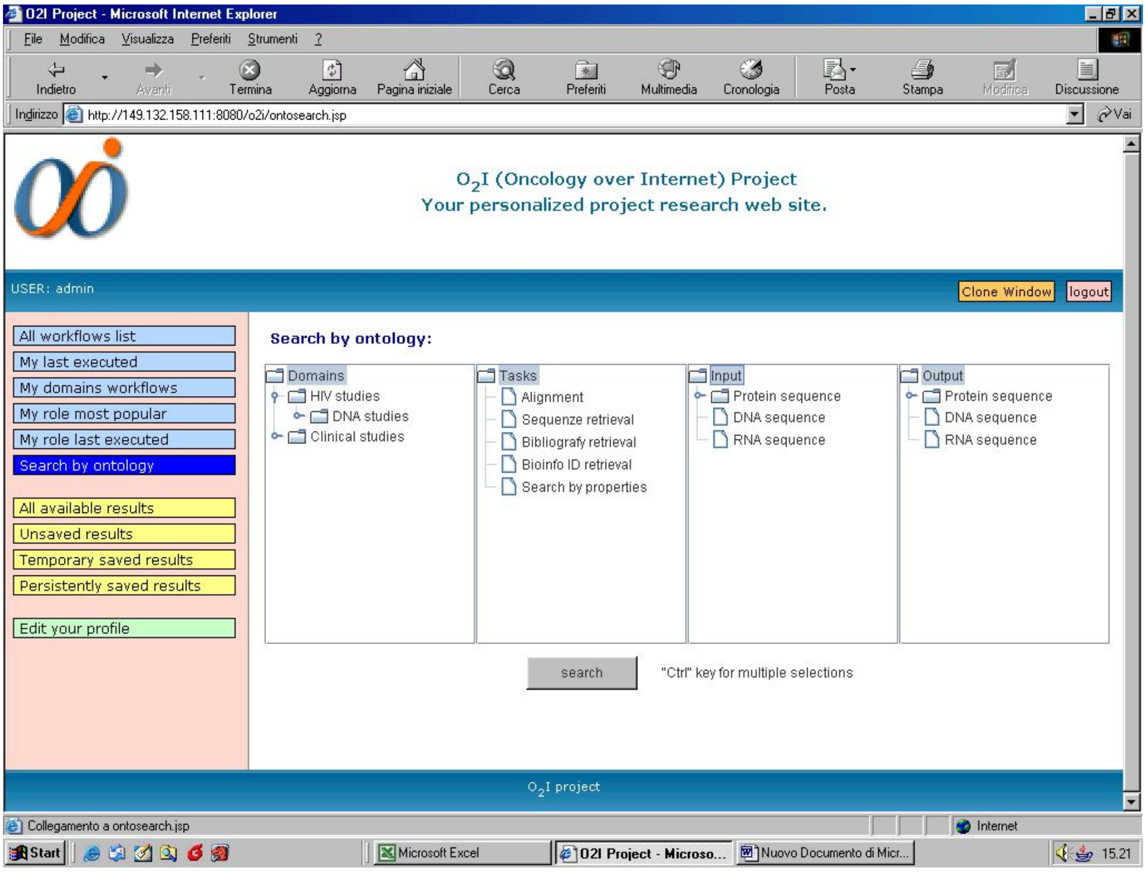
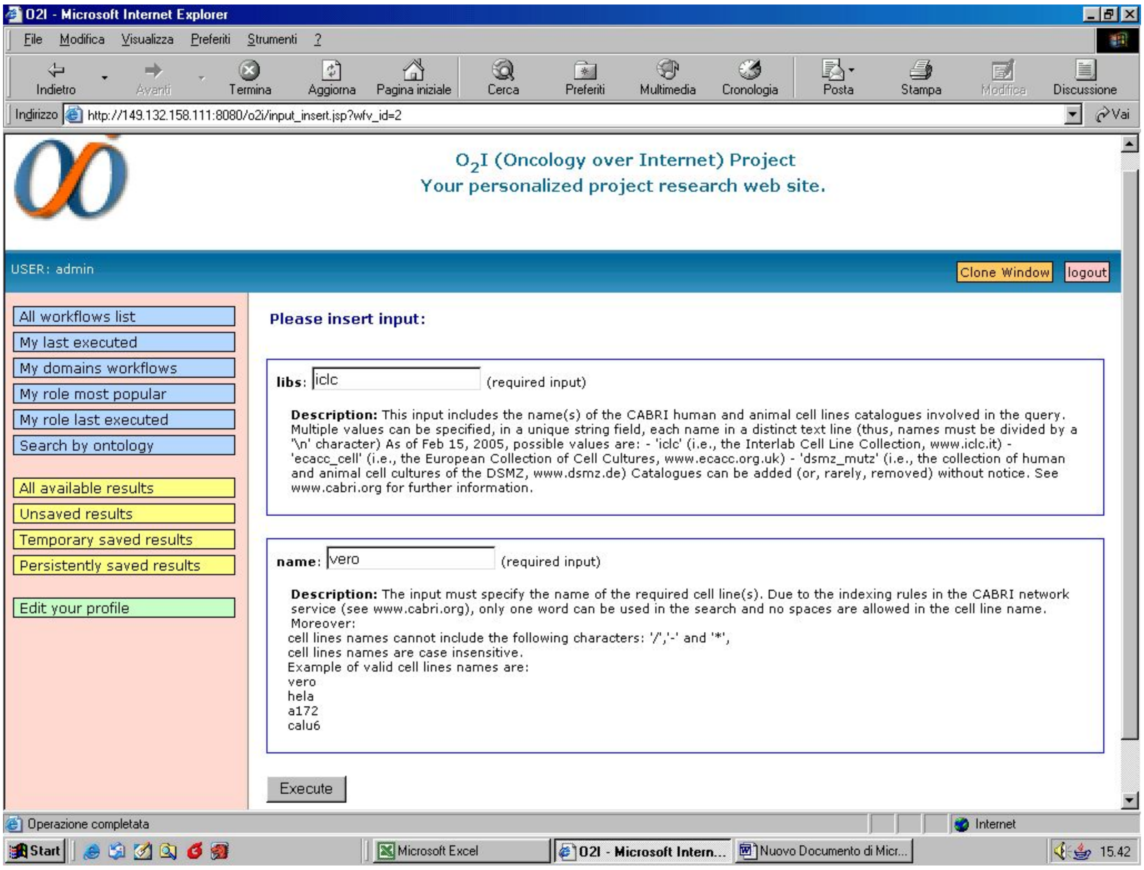


Figure 3: web page allowing search by the annotation of workflows.


O2I - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Avanti Termina Aggiorna Pagina iniziale Cerca Preferiti Multimedia Cronologia Posta Stampa Modifica Discussione
Indirizzo http://149.132.158.111:8080/o2i/input_insert.jsp?wfv_id=2
Vai
O_2I (Oncology over Internet) Project
Your personalized project research web site.
USER: admin
Clone Window
logout
All workflows list
My last executed
My domains workflows
My role most popular
My role last executed
Search by ontology
All available results
Unsaved results
Temporary saved results
Persistently saved results
Edit your profile

Please insert input:

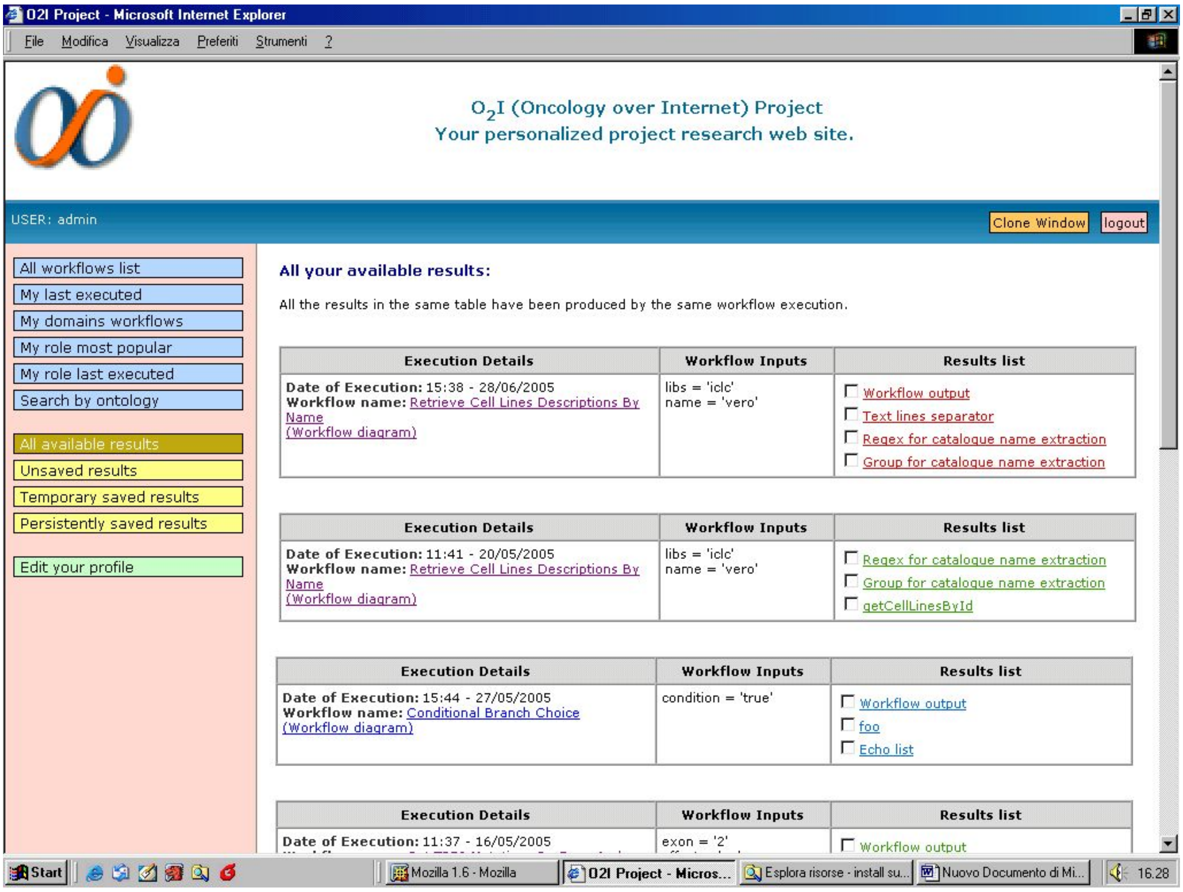
libs: iclc (required input)

Description: This input includes the name(s) of the CABRI human and animal cell lines catalogues involved in the query. Multiple values can be specified, in a unique string field, each name in a distinct text line (thus, names must be divided by a '\n' character) As of Feb 15, 2005, possible values are: - 'iclc' (i.e., the Interlab Cell Line Collection, www.iclc.it) - 'ecacc_cell' (i.e., the European Collection of Cell Cultures, www.ecacc.org.uk) - 'dsmz_mutz' (i.e., the collection of human and animal cell cultures of the DSMZ, www.dsmz.de) Catalogues can be added (or, rarely, removed) without notice. See www.cabri.org for further information.

name: vero (required input)

Description: The input must specify the name of the required cell line(s). Due to the indexing rules in the CABRI network service (see www.cabri.org), only one word can be used in the search and no spaces are allowed in the cell line name.
Moreover:
cell lines names cannot include the following characters: '/','-' and '*',
cell lines names are case insensitive.
Example of valid cell lines names are:
vero
hela
a172
calu6

Figure 4: Input form for the execution of a workflow



All your available results:

All the results in the same table have been produced by the same workflow execution.

Execution Details	Workflow Inputs	Results list

Date of Execution: 15:38 - 28/06/2005
Workflow name: Retrieve Cell Lines Descriptions By Name
(Workflow diagram)
	libs = 'iclc'
	name = 'vero'
		□ Workflow output
		□ Text lines separator
		□ Regex for catalogue name extraction
		□ Group for catalogue name extraction

Execution Details	Workflow Inputs	Results list

Date of Execution: 11:41 - 20/05/2005
Workflow name: Retrieve Cell Lines Descriptions By Name
(Workflow diagram)
	libs = 'iclc'
	name = 'vero'
		□ Regex for catalogue name extraction
		□ Group for catalogue name extraction
		□ getCellLinesById

Execution Details	Workflow Inputs	Results list
Date of Execution: 15:44 - 27/05/2005 **Workflow name:** Conditional Branch Choice (Workflow diagram)	condition = 'true'	□ Workflow output □ foo □ Echo list

Execution Details	Workflow Inputs	Results list
Date of Execution: 11:37 - 16/05/2005	exon = '2'	□ Workflow output

Figure 5: web page allowing for the examination of saved results

In figure 2, the web page listing workflows last executed by the user is shown. From this page, the user can launch workflows (button 'run') or retrieve related details (button 'details').

In figure 3, the web page allowing a search of workflows on the basis of their annotation is shown (in this figure, a fictitious ontology is used). Conditions can be set for each column and are linked by a logical AND. Within columns multiple selections are allowed and conditions are linked by a Boolean OR.

In figure 4, the input form for the execution of a workflow is shown. Input fields are described in details and suggestions for possible input values are reported. Required and optional fields are pointed out.

Results are displayed by using Taverna Workbench java applets and can be saved, either temporarily or definitely, and later reanalysed. In figure 5, the web page listing all saved executions' results and allowing for their further visualization is shown.

Conclusions

We have presented in this paper a general architecture for the implementation of a system that is able to execute workflows of biomedical interest remotely. We have presented as well the preliminary user interface.

The further development and implementation of Web Services allowing the access to and retrieval from an exhaustive set of molecular biology and biomedical databases being carried out by many research centres and network service providers in the biological and medical domains and the creation of effective and useful workflows by interested scientists through widely distributed workflows management systems such as those presented in this paper will significantly improve automation of *in-silico* analysis.

Acknowledgements

This work was partially supported by the Italian Ministry of Education, University and Research (MIUR), projects "Oncology over Internet (O_2I)" and "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)".

References

[1] S. Hoon, K. Kumar Ratnapu, J. Chia, B. Kumarasamy, X. Juguang, M. Clamp, A. Stabenau, S. Potter, L. Clarke, and E. Stupka, Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis, Genome Research, 13:1904-1915, 2003, doi:10.1101/gr.1363103

[2] C. Letondal, A Web interface generator for molecular biology programs in Unix. Bioinformatics, 17(1):73-82, 2001.

[3] Life Sciences Practice Team, BioWBI and WEE: Tools for Bioinformatics Analysis Workflows, IBM Business Consulting Services –AIS, 2004

[4] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat and P. Li, Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17):3045-3054, 2004

[5] Web Services activities at the World-Wide Web Consortium (W3C) – [http://w3.org/ws/]

[6] L. Stein, Creating a bioinformatics nation. Nature, 417:119-120, 2002

[7] D.C. Jamison, Open Bioinformatics (editorial). Bioinformatics, 19(6):679-680, 2003

[8] Entrez Utilities Web Service at NCBI – [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html]

[9] Web Services at NCICB – [http://ncicb.nci.nih.gov/]

[10]KEGG Web Services – [http://www.genome.jp/kegg/soap/]

[11]R. Stevens, A. Robinson and C. Goble, myGrid: personalised bioinformatics on the information grid, Bioinformatics, 19(1):i302-i304, 2003, doi:10.1093/bioinformatics/btg1041

[12]Olivier, M. et al. The IARC TP53 Database: new online mutation analysis and recommendations to users. Hum Mutat, 19(6):607-14, 2002.

[13]TP53 Mutation Database at IARC – [http://www.iarc.fr/p53/]

[14]CABRI catalogues – [http://www.cabri.org/]