Network integration of data and analysis of oncology interest

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

Summary

The Human Genome Project has deeply transformed biology and 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 implies the adoption on a very large scale of new, flexible tools that can enable researchers to integrate data and analysis on the network. ICT technology standards and tools, like Web Services and related languages, and workflow management systems, can support the creation and deployment of such systems. While a number of Web Services are appearing and personal workflow management systems are also being more and more offered to researchers, a reference portal enabling the vast majority of unskilled researchers to take profit from these new technologies is still lacking. In this paper, we introduce the rationale for the creation of such a portal and present the architecture and some preliminary results for the development of a portal for the enactment of workflows of interest in oncology.

1 Introduction

The Human Genome Project has transformed biology and 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. The research in domains like genomics and proteomics strictly depends on the creation, maintenance and use of huge databases. As an example, the size of the EMBL Data Library, the European primary databank of sequences of nucleotides, has reached 58,758,902 sequence entries, comprising 107,562,580,723 nucleotides, in its release 84 (issued in September 2005). This means that, with reference to the previous release, it grew of the 7,8% and of the 13,23%, respectively in entries and nucleotides. The increase in entries and in nucleotides were, respectively, of the 38,87% and of the 53,17% in the last year. An up-to-date statistics of contents of the EMBL Data Library can be retrieved from the web (see http://www3.ebi.ac.uk/Services/DBStats/).

Emerging domains, like analysis of mutations and variations, polymorphisms and metabolism, and high-throughput technologies, e.g., microarrays, will contribute with even huger amounts of data. ArrayExpress [1], the microarray experiments database maintained at the European Bioinformatics Institute (EBI), includes 1,187 experiments for a total size of about 800 Gb, as of December 21th, 2005, and it more than doubled its size from October 2004 to October 2005 (see http://www.ebi.ac.uk/arrayexpress/Help/stats/index.html).

A few databases are managed in a homogenous way under a coordination effort and they represent more an exception than the rule. E.g., databanks of nucleotidic sequences available at the EBI, the US National Center for Biotechnology Information (NCBI) and the Japanese

National Institute of Genetics (NIG), while using different data structures and database management systems, exchange their data on a peer to peer bases, so that the contents of their databases are always almost aligned [2]. This is carried out by applying a common policy in the framework of the International Nucleotide Sequence Database Collaboration (INSDC, see http://www.insdc.org/).

Information in secondary databases, whose data are partially retrieved from primary databases and undergo a careful process of analysis, removal of errors and duplications and a good and extended annotation and quality control, is of the highest quality and therefore they represent an essential resource for researchers. Also, many databases are highly specialized, e.g. by gene, organism, disease, mutation. Finally, it must also be taken into account that many databanks are created by small groups or even by single researchers. The supplemental issue of the Nucleic Acids Research journal [3] that is devoted to molecular biology databases gives a precise idea of this situation. In 2005, it listed 724 databases (see http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D5/DC1). Also, the list of public SRS sites (see http://downloads.lionbio.co.uk/publisrs.html) includes 1,300 different libraries (i.e., databases).

As a consequence of this sources heterogeneity, this huge amount of data is spread over hundreds of Internet sites where they are accessible by using different query methods. Data are also stored using different database management systems (DBMS) and data structures. This does not only imply that access to these data must be performed through many different user interfaces, all of which must be learned, but also, and especially, that there are no common information sets and that the semantics of data, i.e. the actual meaning associated to each piece of data, can be different, even when using the same or similar names, thus leading to potentially erroneous analyses.

The use of heterogeneous Information and Communication Technologies (ICT) tools for data distribution makes then the tasks of searching, retrieving and integrating information very difficult. As a consequence, data are often retrieved and analysed by researchers that make access to several bioinformatics servers through their web browsers and that then transfer the data by either using FTP clients or web browsers themselves. The “cut and paste” technique is widely used to transfer output from one web resource to another site where it is used as an input.

This heterogeneity is even more notable when considering specialist software programs that are essential for almost all analysis in molecular biology, such as sequence analysis, secondary and tertiary protein structure prediction, gene prediction and molecular evolution analysis. This software must of course interoperate with databases: records can be used as input and results of analyses can be seen as new data to be stored and further analysed. Although some integration efforts have already been carried out, like the creation of software suites (e.g., EMBOSS, see http://www.emboss.org/), and the creation of interfaces allowing data interchange between software tools and databases (e.g., Pise [4] and SRS [5,6,7]), the situation is far from being satisfactory.

Integration of heterogeneous data is anyway needed to achieve a better and wider view of all available information, but also in order to automatically carry out analysis and/or searches involving more databases and software and to perform analysis involving large data sets. Finally, only a tight integration of data and analysis tools can lead to a real data mining.

In such a context, 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 databases of interest or to use the appropriate software to achieve the desired analysis. At the same time, it should also be able to cope with many different systems and to "understand" the information that it is managing.

Integration of data and processes needs stability of the domain. This implies a deep knowledge of the domain and well defined information and data, both leading to a standardization of information schemas and formats. Also, essential is a clear definition of the goals. On the contrary, integration fears heterogeneous data and systems, uncertain domain knowledge, highly specialized and quickly evolving information, lacking of predefined, clear goals and originality of procedures and processes.

In biology, a pre-analysis and reorganization of the data is very difficult, because data and related knowledge change very quickly. Moreover, complexity of information makes it difficult to design data models which can be valid for different domains and over time. Finally, goals and needs of researchers evolve very quickly according to new theories and discoveries, this leading to frequent new procedures and processes. So, current integration methods, that are based on syntactical tools like explicit cross-references, implicit links (e.g., through names of biological entities) and common contents (achieved by using common vocabularies, reference lists and lexicons) are inadequate. Instead, new methods based on semantic links, such as those that can be derived by using metadata descriptions and reference ontologies, seem more adequate. Flexibility of systems, including the ability to support frequent changes of data, software and analysis, is mandatory.

Among current ICT technologies, workflow management systems, in connection with Web Services, seem to be the most promising ones. Web Services (WS) are network services that are based on the eXtensible Markup Language (XML, http://www.w3.org/XML/). As it is well known, XML allows for a machine readable description of the data that are described by using well defined document types. Many XML based markup languages for bioinformatics have already been defined, and some authors have already listed some of them and discussed pros and cons of their adoption [8,9,10,11].

WS are software oriented network services usually communicating by using the Simple Object Architecture Protocol (SOAP, see http://www.w3.org/2002/ws/), a framework for the distribution of XML structured information, over HTTP. They offer a standardized programming interface so that software tools can effectively make access to the information and services they are delivering. Standard protocols are available for their description (Web Services Description Language – WSDL, http://www.w3.org/2002/ws/desc/), retrieval and identification (Universal Description, Discovery, Identification – UDDI, http://www.uddi.org/), and composition (Web Services Flow Language - WSFL), just to mention a few. Hence, Web Services allow software applications to access data in a more "intelligent" way, since applications can identify and interpret the information and, possibly, when ontological metadata is added, the associated semantics.

Reasons for the setting up of Web Services in bioinformatics have recently been presented [12,13]. These include the need for overcoming the scaling problem arising from the use of high-throughput experimental protocols that provide such huge results that their analysis needs a "high-throughput" sequence analysis process in order to be studied in an adequate time scale. This could not be achieved through the traditional approach implying manual access to web sites, while software driven access to Web Services implementations of the required sequence analysis software could achieve it. Also, WS would offer bioinformatics the possibility of implementing a real distributed analysis environment, while protecting intellectual property rights for data, algorithms and source code, that would not be copied and would remain on the owners' information system.

WS have already been implemented by many Institutes and service centers in the biomedical field. Examples of WS available at some bioinformatics network service centers are Entrez Utilities at the National Center for Biotechnology Information (NCBI), (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html), Web Services at the

National Cancer Institute Center for Bioinformatics (NCICB, http://ncicb.nci.nih.gov/, KEGG Web Services (http://www.genome.jp/kegg/soap/) and the SoapLab implementation at the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/soaplab/), through which researchers can execute all tools included in the EMBOSS software suite. Lists of Web Services that are available for bioinformatics are available at the myGrid Wiki site (http://twiki.mygrid.org.uk/twiki/bin/view/Bioinformatics/BioinformaticsWebServices) and in the Taverna web site (http://taverna.sourceforge.net/index.php?doc=services.html)

BioMOBY is an open source software that implements an architecture for the discovery and distribution of biological data through web services; data and services are decentralised, but the availability of these resources, and the instructions for interacting with them, are registered in a central location called MOBY Central [14].

The notion of workflow is a central one in Web Services. 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. Effectiveness is achieved through automation of repetitive procedures: being an automatic procedure, a workflow can free bio-scientists from repetitive interactions with the web, at the same time supporting good practice. Reproducibility is also granted by the implementation of repetitive procedures, although it is limited by the frequent update of information sources; anyway, analyses can be replicated over time. Reusability is implemented by storing intermediate results and by allowing their use in subsequent workflows executions. Finally, traceability is achieved by storing intermediate results and allowing their analysis: the workflow is then carried out in a transparent analysis environment where data provenance can be checked and/or controlled. This is especially important when unexpected data are obtained.

Workflow management systems should not be compared to other integration systems, such as the Sequence Retrieval System (SRS, [5,6,7]) since they carry out tasks that are quite different. While SRS is able to perform limited, predefined operations (i.e., boolean and linking operations) on a local set of databases, a workflow management system is able to carry out any kind of elaborations and analysis on remote databases. Instead, an SRS site could be remotely queried through a properly programmed Web Service and its abilities, such as querying more databases at the same time, could therefore be added to a workflow. With workflow management systems, query processing on multiple sources can be achieved by carrying out parallel searches and later merging results. Alternate processing is also available with workflow management systems. This can be achieved by assigning the same task in a workflow to more services, by also providing them priority levels, and by invoking the services having the highest priority level first. Services with lower priority levels can then be invoked, when needed, if the previously called ones should fail.

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 [15], a perl module designed to be used with bioperl, and GPipe, an extension of the Pise interface [16]. Other systems are autonomous applications that are being developed either by industries, like the Bioinformatic Workflow Builder Interface – BioWBI from IBM [17], and Pipeline Pilot from SciTegic, or by academic and research institutes, like Wildfire from the Singapore Bioinformatics Institute, and Taverna Workbench [18] from the European Bioinformatics Institute (EBI).

These workflow management systems assume that end users know all bioinformatics resources they need, especially those resources that can be reached through a programatic interface, and are proficient, if not skilled, in programming computers and in the composition

of their own workflows. They are therefore not viable to the vast majority of biologists and researchers that are normally only skilled in the use of web interfaces.

We present here a prototypal system which can manage, organize and execute a set of predefined and tested workflows, and show its application in an oncology setting. The prototype presents a user-friendly web interface that is able to simplify access to such workflows and it therefore is viable to all end users. At the same time, it allows to profit from all advantages of the workflow management systems.

2 Methods

Our system is partially based on open source software, namely Taverna Workbench (http://taverna.sourceforge.net/), FreeFluo enactor engine (http://freefluo.sourceforge.net/) and mySQL database management system (http://www.mysql.com/). Workflows are created by using the Taverna Workbench and then stored in the Simple conceptual unified flow language (Scufl) format. The user interface has been created by writing some java servlets and it is delivered through a servlet engine, Apache Tomcat (http://tomcat.apache.org/). mySQL-connector (http://www.mysql.com/products/connector/j/) is used to get access to the mySQL database.

Taverna Workbench and FreeFluo have been selected for their various useful features. Taverna is a workflow manager developed at the European Bioinformatics Institute (EBI) in the frame of the myGrid project [19]. It is able to build complex analysis workflows, to access both remote and local processors of various kinds, to launch execution of workflows and to display different types of results, including text, web pages and various kinds of images. Workflows execution is carried out by an associated tool, the FreeFluo enactor engine.

Taverna Workbench includes an ontology for bioinformatics data that is used to describe the input and output data of each processor (i.e., a single elaboration step in the workflow). In our system, this was used as a starting point for the creation of an ontology that supports the carrying out of searches in the workflow repository by the selection of the features of the main elaboration steps in the workflows.

Taverna only requirements are availability of the Java Run-time Environment (JRE, http://java.sun.com/) on either a Windows XP or Linux operating system, and, in the latter case, of the graph visualization tool Graphviz (http://www.graphviz.org/).

Processors that can be used through the Taverna Workbench include Web Services, either described through their WSDL definition or accessed through a bioMOBY registry, and retrieval of information from BioMart databases [20,21] (http://www.biomart.org/). The workbench can of course execute any workflow that is defined by the XScufl language. Finally, local processors are included with Taverna for basic elaborations like simple list or string processing, definition of constant values, local input/output management. New local elaborations can be further defined and specialised by the user that is allowed to create and add scripts by using BeanShell (Lightweight Scripting for Java, http://www.beanshell.org/).

Our system (see fig. 1) includes three main blocks: the workflow manager, the user interface and the workflow executor.

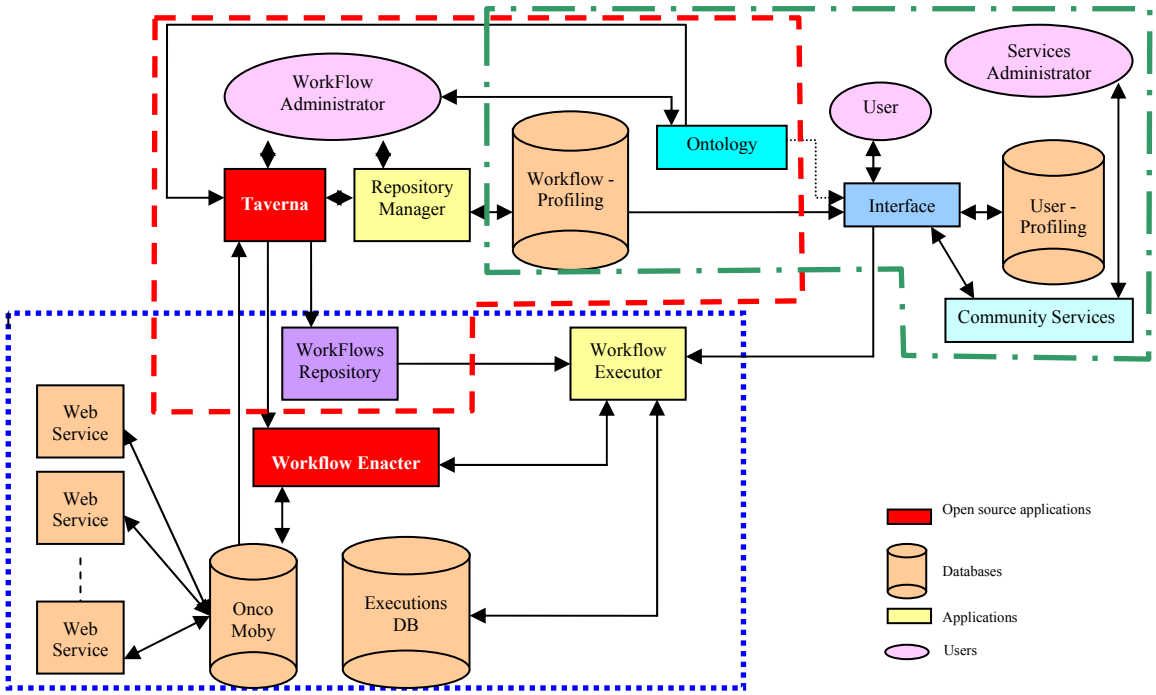

Figure 1: General schema of the Oncology over Internet (O_2I) system. It includes three main parts: a) workflows creation and annotation, that are performed by a special user, the workflow administrator (red dashed line), b) user interface (green dot-dashed line), and c) workflows execution (blue dotted line).

An administrator edits workflows off-line by using Taverna and then he stores them in a repository. The main processing steps of each workflow are also annotated on the basis of the input and output data, elaboration type and application domain. Annotations are defined by using an ontology for bioinformatics tasks and then they are stored in the workflow profiling database tables.

The user interface supports end users authentication and profiling. This information is stored in the User Profiling tables of the database. It includes some basic data such as the name and email address of the user and his classification on the basis of his role in his organization, such as “computer scientist”, “oncologist” or “molecular biologist”, and his domains of interest, like, e.g., “mutation analysis”.

The user interface also allows for the selection and enactment of workflows. Workflows selection can be assisted by users’ profiles and by searching workflows annotations. Users can request a list of all workflows in the system that have been annotated with reference to their role and/or with reference to their domains of interest. Workflows are executed by the third block that requests FreeFluo to enact the workflow and it is also able to store input and output data of actual workflows’ executions, so that they can late be analysed and possibly reused.

3 Results

We designed a web system that allows for the definition and the execution of a set of workflows of oncology interest. These workflows are designed to access to and to retrieve data from various Web Services. In our system, users’ registration supports retrieval of workflows on the basis of their role in the organization (e.g., researcher, clinician, computer scientist) and their domains of interest (e.g., mutation analysis, gene prediction). Search and identification of workflows of interest can also be achieved by means of the annotation of the workflows. This annotation is based on an ontology of processors that describe them on the

basis of their application domain, overall task and input / output data. Workflows can also be retrieved by date (last executed first). Finally, in our system workflow executions and related results can be stored and can be later retrieved.

The general architecture of the system has been defined and a prototype system has been developed and is currently under test at http://www.o2i.it:8080/o2i/.

In figure 2, the web page listing all workflows recently executed by the user is shown. The list includes the name and a short description of the workflows, together with their current version numbers and the last execution date. From this page, the user can enact workflows (button 'run') or retrieve related details (button 'details'). Similar web pages exist for all workflows available in the system and for workflows selected on the basis of the user's domain and role.

In figure 3, the web page allowing a search of workflows on the basis of their annotation is shown. Conditions can be defined on the application domain of the workflow, as well as on its type (the kind of elaboration or analysis that it performs) and the type of its input and output fields. Conditions can be set on each column and they are then combined by using a logical AND. When multiple conditions are put on the same column, these are combined by using a logical OR. An example query could be: find all workflows in the molecular biology domain (application domain) including elaboration steps that retrieve (retrieval task) DNA sequences (output) on the basis of a Genbank accession number (input). Of course, end users are not obliged to put conditions on every field: these can be left undetermined.

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

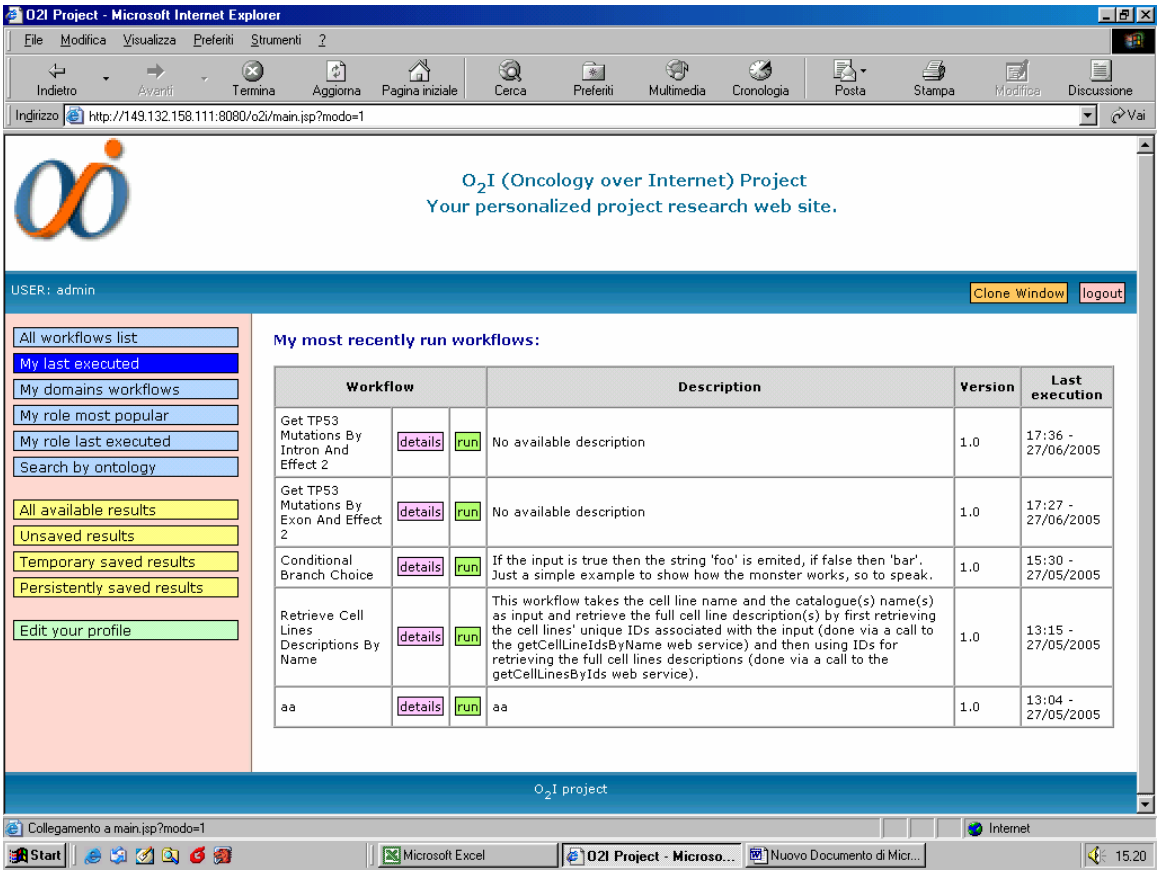

Figure 2: list of most recently executed workflows as they appear in the user interface

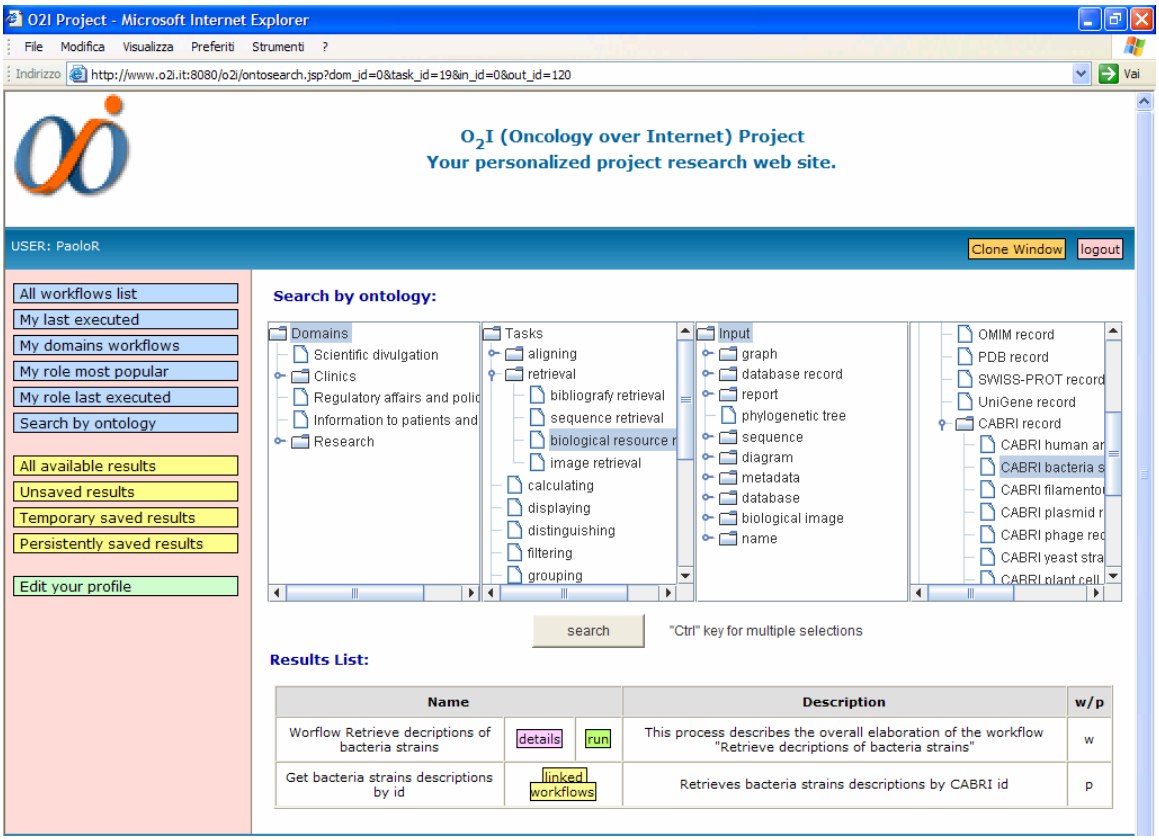

Figure 3: web page allowing search of workflows through their annotation and a list of results. Conditions were put on task (biological resources retrieval) and output (CABRI bacteria record)

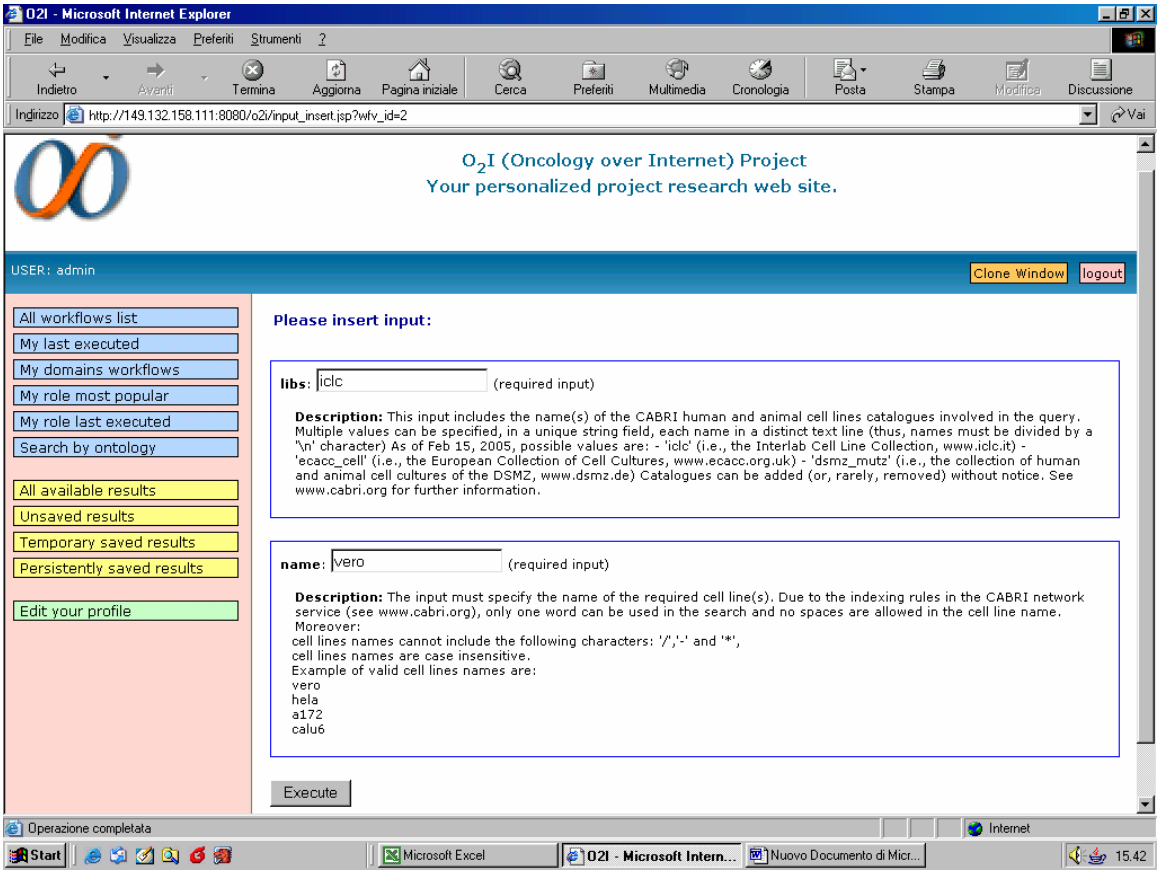

Figure 4: Input form for the execution of a workflow. The description of the input includes some examples. Inputs that are mandatory for the execution of the workflow are pointed out.

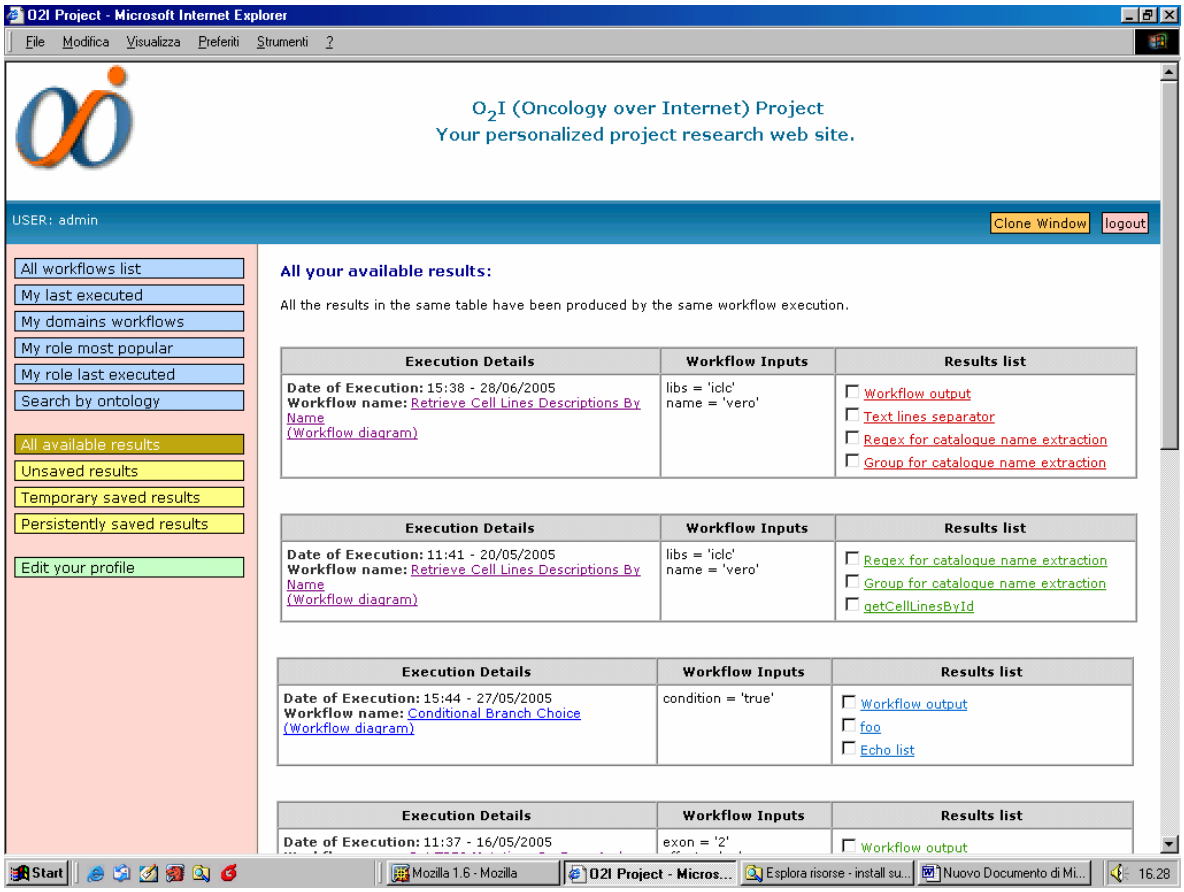

Figure 5: web page allowing for the examination of saved results

Results of the executions can be saved, either temporarily or definitively, and later reanalysed. In figure 5, the web page listing all saved results and allowing for their further visualization is shown. Results can currently be locally displayed by using a java library that must be downloaded and installed on the local computer where a version of java virtual machine must also be available and running. The visualization library is derived from Taverna Workbench.

A set of new Web Services has been developed and is available on-line at http://www.o2i.it:8080/axis/services/. They implement access to IARC TP53 Mutation Database (http://www.iarc.fr/p53/) [22] and to CABRI catalogues of biological resources (http://www.cabri.org/) [23], by using SoapLab (http://www.ebi.ac.uk/soaplab/) [24]. Workflows are being created and tested in various application domains (http://www.o2i.it/workflows/). The ontology is being developed starting from the Taverna bioinformatics data ontology. During next months, a first user interface will be made available on-line.

4 Conclusions

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

Such a system can implement predefined data analysis processes by remotely accessing bioinformatics Web Services. With reference to other integration systems, such as SRS, our system is able to offer a wider set of possible analysis and a more effective interface since it assumes no prior knowledge of available services and related data structures by end users.

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.

5 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)".

6 References

[1] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA, ArrayExpress - a public repository for microarray gene expression data at the EBI, Nucleic Acids Res. 2003 Jan 1;31(1):68-71.

[2] S. Brunak, A. Danchin, M. Hattori, H. Nakamura, K. Shinozaki, T. Matise, D. Preuss (2002), Nucleotide Sequence Database Policies, Science 298 (5597): 1333 15 Nov 2002

[3] Galperin, M.Y., The Molecular Biology Database Collection: 2005 update, Nucl. Acids Res. 2005 33: D5-D24 (doi:10.1093/nar/gki139)

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

[5] Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. Meth. Enzymol., 266, 114-128

[6] Zdobnov, E., Lopez, R., Apweiler, R. and Etzold, T. (2002) The EBI SRS server – new features. Bioinformatics 18(8), 1149-1150

[7] Zdobnov, E., Lopez, R., Apweiler, R. and Etzold, T. (2002) The EBI SRS server – recent developments. Bioinformatics 18(2), 368-373

[8] Guerrini, V.H. and Jackson, D. (2000) Bioinformatics and extended markup language (XML), Online Journal of Bioinformatics, 1:1-13

[9] Achard, F., Vaysseix, G., and Barillot, E. (2001) XML, bioinformatics and data integration. Bioinformatics 17(2):115-125

[10] Romano, P. (editor) (2001) Proceedings of NETTAB 2001 Workshop on "XML and CORBA: Towards a Bioinformatics Integrated Network Environment, Genova, May 17-18, 2001 (available from the editor)

[11] Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M. et al (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biology, 3(9):1-9

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

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

[14] Wilkinson, M.D., Links, M. (2002) BioMOBY: an open-source biological web services proposal. Briefings in Bioinformatics 3:4. 331-341.

[15] 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

[16] Garcia Castro A, Thoraval S, Garcia LJ, Ragan MA., Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator, BMC Bioinformatics, 6(1):87 (2005).

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

[18] 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

[19] 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)

[20] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439-40.

[21] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox and E. Birney. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. Genome Res. 2004 Jan; 14(1):160-169.

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

[23] Romano P., Kracht M., Manniello M.A., Stegehuis G., Fritze D., The role of informatics in the coordinated management of biological resources collections. Applied Bioinformatics 2005;4(3):175-86

[24] M. Senger, P. Rice, T. Oinn, Soaplab - a unified Sesame door to analysis tools, Proceedings, UK e-Science, All Hands Meeting 2003, Editors - Simon J Cox, p.509-513, ISBN -1-904425-11-9, September 2003