Annotation of cell lines by using workflows: a first experience

Domenico Marra, Barbara Parodi, M. Assunta Manniello and Paolo Romano
National Cancer Research Institute IST, Largo R. Benzi 10, I-16132, Genova, Italy

Introduction

To achieve the full potential of biotechnology innovation in Europe, there's a need to strengthen the underpinning scientific and technological infrastructure. The key element to this is the management of biological resources, meant not only as biological raw material, but also as related information. The Organization for Economic Co-operation and Development (OECD)[1] defines the Biological Resource Centres (BRCs) as follows: "*They consist of service providers and repositories of the living cells, genomes of organisms, and information relating to heredity and the functions of biological systems. BRCs contain collections of culturable organisms (…), replicable parts of these (…), viable but not yet culturable organisms cells and tissues, as well as data bases containing molecular, physiological and structural information relevant to these collections and related bioinformatics*". BRCs must meet the high standards of quality and expertise demanded by the international community of scientists and industry for the delivery of biological information and materials. They must provide access to those biological resources on which depends research and development in the life sciences and the advancement of biotechnology.

In the biomedical field, emerging domains (mutation and variation analysis, polymorphisms, metabolism) and technologies (microarrays, proteomics) are contributing with high amounts of data, which could be usefully linked to the information on biological material provided by BRCs (e.g., cell lines and tissue samples).

In this frame, the possibility of extending annotation on biological resources as a result of the carrying out of integrated queries and, possibly, analyses on more databases would provide a better and wider view of all available information, and, finally, a better quality service to the researchers.

The Interlab Cell Lince Collection (ICLC, http://www.iclc.it/), core facility of the National Institute for Cancer Research (IST) of Genoa, Italy, produces, collects and distributes human and animal cell lines, mainly derived from tumors, to the research community worldwide. In the last years, researchers have become always more demanding as to information on specific biological features of the cell lines, e.g., the p53 status of the cells. Moreover, original papers describing the cell lines available in the bank and further literature describing more biological functions and properties are essential for a deeper understanding of cell lines behaviour in experiments.

In order to satisfy this need of information, we designed and implemented an automated procedure that is able to improve cell lines annotation by carrying out a predefined set of queries on the Internet.

Methods

The procedure has been created by using the Taverna Workbench[2] (http://taverna.sourceforge.net), a very good tool able to build and enact workflows accessing different information sources and analyses on the net. It is available on-line in the Taverna Scufl format (http://www.o2i.it/workflows/) and can easily be executed by either using a local Taverna implementation or the FreeFluo workflow enactor.

Two main issues have been taken into account during the development of this workflow:

1) the need for an easy retrieval and integration of information from different sources. Without the use of automated procedures, the researcher can only query one source at a time. The integration of more queries often is the result of a manual operation of the researcher himself by means of time consuming and error prone "cut and paste" steps. In case of two or more initial results, this procedure must of course be iterated for every single data leading to a significant loss of time.
2) the simplification of the query by means of a reduction (in number and complexity) of query parameters. In order to carry out a fruitful search, the end user must accurately select those query

fields that are needed for his goals and, for each of them, the right data input format. The reduction of the complexity of the input simplifies the job of the researcher since he/she is no more compelled to acquire a deep knowledge of the data structure of the involved databases and query forms.

Starting from these considerations, we designed and implemented a workflow that is able to support annotation of cell lines available in CABRI catalogues (http://www.cabri.org/) by querying and retrieving information from three different sources: the cell line database (chosen from the CABRI catalogues), a database of mutations of the TP53 human gene (our SRS[3] implementation, http://srs.o2i.it/srs71/, of the IARC TP53 Mutation Database[4], http://www-p53.iarc.fr/index.html) and a reference literature database (Medline at NCBI, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed).

The input of the workflow is the name of one of the CABRI cell lines catalogue. Its output is a collection of records from the three source obtained on the basis of various interlinked queries.

The workflow is composed by the following steps:

1) Retrieval of the cell lines names described in the specified CABRI database
2) Retrieval of the complete list of sample names of the IARC TP53 Mutation Database
3) Comparison of the two above lists and identification of the common subset
4) Retrieval of complete CABRI records for cell lines present in the common subset
5) Retrieval of complete TP53 records for samples present in the common subset
6) Retrieval of PUBMED Unique Identifiers from TP53 records
7) Retrieval of the abstracts from PUBMED network service
8) Merging of results obtained from steps 4, 5 and 7

Figure 1 shows the diagram of the workflow.

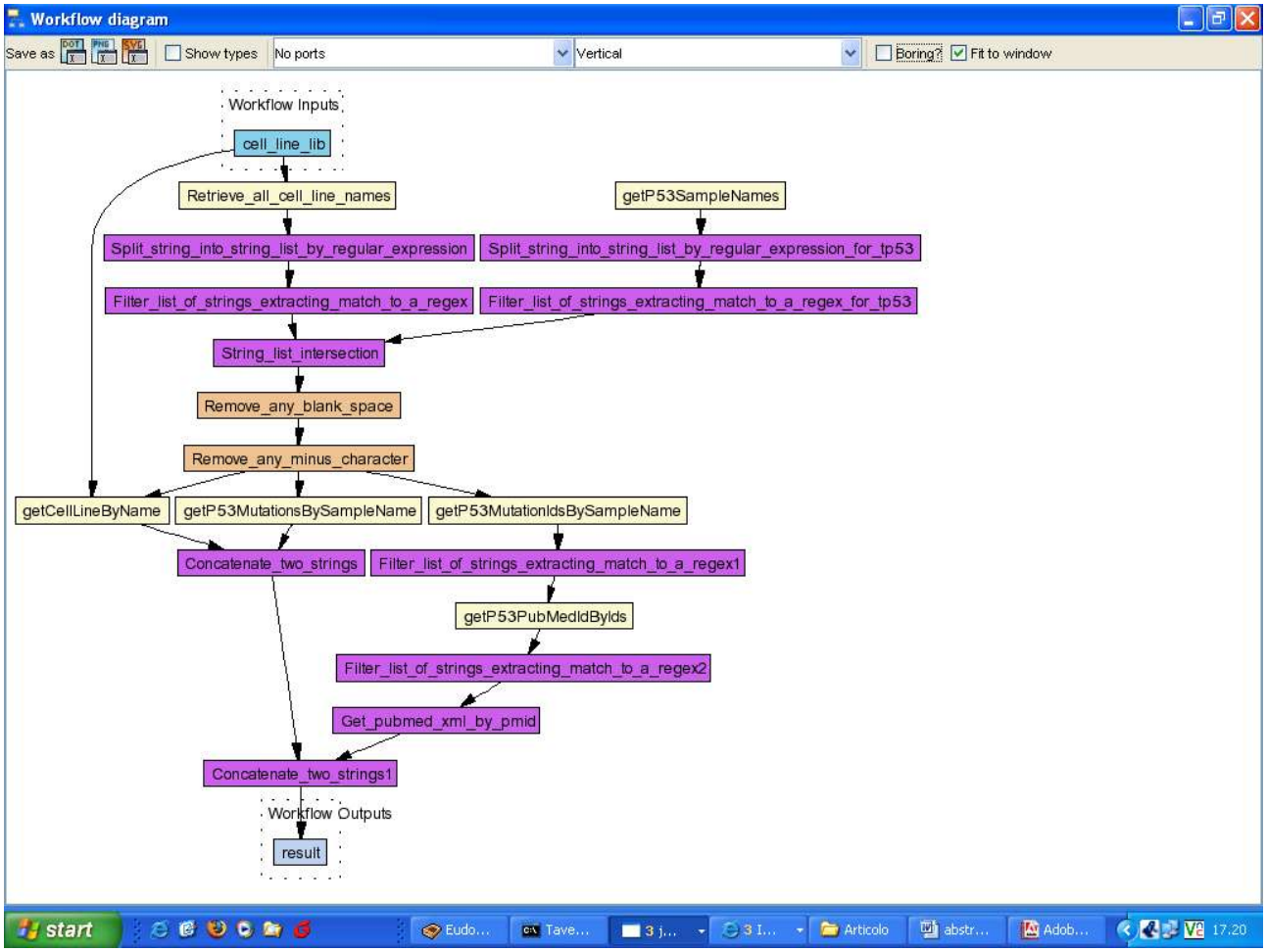

Fig.1 A graphical visualization of our workflow. For a better readability, all constant parameters have been omitted

During the enactment of the workflow, some other actions occur that are not included in this list; they carry out some string elaborations which are needed for the correct passage of data between the queries.
The final result is made by a list of cell lines' records each of which is complemented by all records of the mutation database that relates to that cell line. These are, in turn, expanded by including the abstracts of their literature references.
This result can easily be achieved without any specific database or query system knowledge, the only information that must be in the hands of the researcher is the acronym of the CABRI catalogue of their interest, all the other information comes automatically.

Conclusions

From the point of view of a BRC, applications based on workflows management are very promising since they allow integrated access to specialized databases and provide added value to the information that is available at the collection. The workflow implemented that is here described will be used in the routine activity of the collection for the annotation of cell line information with reference to information that is available as to their p53 status.
Results must carefully be analysed before incorporating them in the catalogues. The reason for this being that the association between cell lines in different databases in only based on their names. Instead, cell lines are living parts of organisms and as such they are subject to changes, both spontaneous and induced. It is normally considered that two cell lines identified by the same name in different collections should not be considered the same cell line, unless differently proved. So, a deeper identification of cell lines described in various databases, including, e.g., their origin, would simplify the work of the curators of the collections and would help in a proper annotation of information available in the catalogues.
A coordination between the repositories leaded by one authoritative institution might be the right answer to solve this problem since this could lead to a consensus on equivalence between cell lines available in different collections and it could also lead to a standard definition on all information needed to uniquely identify a cell line.

References

[1] OECD (2001) Biological Resource Centres Underpinning the future of Life Sciences and Biotechnology. OECD Science & Information Technology, May 2001, vol. 2001, no.7, pp.1-68
[2] Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20 (17):3045-3054, 2004
[3] SRS: information retrieval system for molecular biology data banks. Meth. Enzymol. 266:114-128, 1996
[4] The IARC TP53 Database: new online mutation analysis and recommendations to users. Hum Mutat, 19(6):607-14, 2002