

Integration of data from different sources: a prototype devoted to p53 mutations - (session: Database: Ontology and Integration)

A. Mucci[°], A. Cusmano[°], M. De Francisci[^], M. A. Manniello[^], D. Marra[^], P. Romano[^], G. Mauri[°]

[°]Università di Milano Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione - DISCO, Milano
[^]Istituto Nazionale per la Ricerca sul Cancro - IST, Genova

Oncology Over Internet is a project devoted to integration of data from different sources of oncology interest. The main focus of the project is on the software architecture, but important improvement for in silico biology research and for some clinical investigations are foreseen.

A prototype has been developed in order to test different technical solutions to access and integration issues and verify the overall feasibility of the system. The prototype is focussed on the database of mutations of the TP53 gene that is maintained by the International Agency for Research on Cancer (IARC). The TP53 gene expresses the p53 protein which has some important and well known influences in the control of cancer at the very preliminar steps and in the elimination of mutated cells. The database of the TP53 mutations has some implicit and explicit links with some molecular and cellular biology databases, such as sequence databases, literature databanks and human cell lines catalogues.

The prototype includes the user interface, the Java based search engine, that is in charge of carrying out of the queries and of gathering the information, a knowledge base and a database where the data retrieved from the various information sources is stored.

The user can ask for the execution of a query by submitting the proper parameters through an online form to the main server. The search engine will query every single database involved in the query, after checking the contents of the knowledge base; the results will finally be returned to the user.

The main application is structured in several blocks, each of which have a specific function:

- Query the knowledge base
- Preparation of the results' table
- Selection of the involved databases
- Preparation of the query
- Analysis of the sites hosting involved databases
- Query of the involved databases
- Gathering and restructuring of the results
- Displaying of results

The prototype is based on the knowledge base, whose goals are: 1) to select the information sources, 2) to carry out queries, 3) to extract the required information and 4) to integrate data.

The prototype has been developed by using open source softwares and products: MySQL database management system, Apache Tomcat web server and Java programming language. At present, the prototype is under beta test.