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

Paolo Romano (paolo.romano@istge.it)
National Cancer Research Institute,
Genoa, Italy
• Motivations for the system
• Goals of the system
• The system
  • architecture
  • ERA schema
  • users profiling
  • workflow annotation
Information in biology: well known facts

- Biomedical research produces an increasing quantity of new data and new data types
  - EMBL size: ver 82 7.3% more vs ver 81 (3 months), 112.9% vs ver 74 (24 months)
- Emerging domains, like mutation and variation analysis, polymorphisms, metabolism, as well as new technologies, e.g., microarrays, will contribute with even huger amounts of data
- Analysis softwares must interoperate with databases
  - Databases as input for softwares
  - Results as new data to record and analyze
Heterogeneity of databanks

- A few dbs are managed in an almost homogenous way (e.g., sequences at EBI, NCBI, DDBJ)
- Secondary databases are of high quality (good and extended annotation, quality control)
- Many databases are highly specialized, e.g. by gene, organism, disease, mutation, etc…
- Many databanks are created by small groups or by single researchers

- Databanks are distributed:
  - Different DBMS, data structures, query methods
  - Different information, semantics
Goals of the integration

In this context, data integration and work automation are needed to:

- Carry out analysis and/or searches involving more databases and softwares automatically
- Perform analysis involving large data sets effectively
- Achieve a better and wider view of all available information
- Carry out a real data mining
Data integration longevity

- **Integration needs stability**
  - Standardization……
  - Good domain knowledge
  - Well defined data
  - Well defined goals

- **Integration fears:**
  - Heterogeneity of data and systems
  - Uncertain domain knowledge
  - Fast evolution of data
  - Highly specialized data
  - Lacking of predefined, clear goals
  - Originality, experimentalism (“let me see if this works”)

P. Romano, Naples, October 6, 2005
Integration of biological information

In biology:
- A pre-analysis and reorganization of the data is very difficult, because data and related knowledge change very quickly.
- Complexity of information makes it difficult to design data models which can be valid for different domains and over time.
- Goals and needs of researchers evolve very quickly according to new theories and discoveries.

Integration must therefore be carried out by using flexible systems that are easy to adapt and to extend.
Workflows management

“A computerized facilitation or automation of a business process, in whole or part”. (Workflow Management Coalition)

Main goal is:
- the implementation of data analysis processes in standardized environments

Main advantages relate to:
- **effectiveness**: being an automatic procedure, it frees bio-scientists from repetitive interactions with the web and it supports good practice,
- **reproducibility**: analysis can be replicated over time,
- **reusability**: intermediate results can be reused,
- **traceability**: the workflow is carried out in a transparent analysis environment where data provenance can be checked and/or controlled.
Workflow management software

Workflow management softwares for bioinformatics applications:

- Biopipe, an add-on to bioperl
- GPipe, an extension of the Pise interface
- Taverna (EBI), a component of the myGrid platform
- Wildfire (Bioinformatics Institute, Singapore)
- Pipeline Pilot (SciTegic)
- BioWBI, Bioinformatic Workflow Builder Interface, from IBM

They all require knowledge of the systems and skills and time for development of the workflows.
We designed a web system that:

- allows for the carrying out of a set of predefined workflows (of oncology interest)
- supports workflows annotation by using a simple ontology for bioinformatics processors (domain, task, i/o)
- implements search of workflows on the basis of their annotation
- supports retrieval of workflows based on users’ registration and profiling
- allows storing and retrieval of workflows’ executions and related results
We designed a web system that:

- makes access to and retrieves data from Web Services and registries of Web Services
- stores workflows using the Simple conceptual unified flow language (Scufl) format
- is partially based on open source tools (Taverna WB, FreeFluo and mySQL)

Prototype available on-line by end of 2005: http://www.o2i.it:8080/portal/
Predefined workflows

Workflows are:
- created by internal staff using Taverna
- stored in Scufl format
- maintained (workflow vs version)
- submitted by:
  - users
  - service providers
Annotation of workflows

Workflows are annotated on the basis of:

- a simple ontology for bioinformatics processors:
  - application domains
  - task
  - inputs/outputs
- ontology derived from Taverna:
  - new structure
  - some additions (biological resources, images, …)
  - under further development
O₂I workflows annotation
Users’ registration and profiling

Users are profiled on the basis of:

- role in their organization
  - computer scientist / physician / researcher / patient / journalist / …
- domains of interest
- past workflows’ executions
O₂I (Oncology over Internet) Project
Your personalized project research web site.

Please login

username: 
password: 
login

New user? Please register!

Please install the new library. (Instructions)
Applets digital certificate.

P. Romano, Naples, October 6, 2005
### All workflows:

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Description</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional Branch Choice</td>
<td>This is a demo workflow distributed with Taverna Workbench (see taverna site). If the input is true then the string 'foo' is emitted, if false then 'bar'. Just a simple example to show how the conditional branch processor works.</td>
<td>1.0</td>
</tr>
</tbody>
</table>
| Retrieve Cell Lines Descriptions By Name      | 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). Both these web services are available at the soaplab system at http://www.o2i.it:8080/axis/services
A number of string or string list local elaborations are required: - returned IDs are in a string and this must be transformed in a list (done by the 'Separate_cell_line_ids' processor, that is implemented by using a Split_string_into_string_list_by_regular_expression local processor) - returned IDs include catalogue's names and this must be removed before their utilization for further processing (done by the 'Extract_ids_by_removing_catalogues_names' processor, that is implemented by using a Filter_list_of_strings_extracting_match_to_a_regex local processor) - returned IDs include a blank character and this must be substituted by a '_SP_' characters string before submitting the data to the 'getCellLinesByIds' web service (done by a trivial bash shell script).
Special requirements on input data are: - one or more of the following catalogue names can be specified: 'lclc', 'ecaco_cell', 'dszm_mut'. Other names may lead to errors: - when specifying more than one catalogue names, they must be in a unique input string but on distinct text lines, - cell lines names can only be made by a single word, excluding special characters as '/' and '*' - cell lines names are case insensitive.
Example of valid cell lines names are: - vero - hela - e172 - colu6                                                 | 1.0     |
O₂I (Oncology over Internet) Project
Your personalized project research web site.
**O₂I (Oncology over Internet) Project**

*Your personalized project research web site.*

---

**My application domains workflows:**

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Description</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve Cell Lines Descriptions By Name</td>
<td>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 getCellLinesBy1ds web service). Both these web services are available at the soalab system at <a href="http://www.o2i.it:8080/axis/services">http://www.o2i.it:8080/axis/services</a>. A number of string or string list local elaborations are required: - returned IDs are in a string and this must be transformed in a list (done by the 'Separate_cell_line_ids' processor, that is implemented by using a Split_string_into_string_list_by_regular_expression local processor) - returned IDs include catalogues' names and this must be removed before their utilization for further processing (done by the 'Extract_ids_by_removing_catalogues_names' processor, that is implemented by using a Filter_list_of_strings_extracting_match_to_a_regex local processor) - returned IDs include a blank character and this must be substitude by a '_SP_' characters string before submitting the data to the 'getCellLinesBy1ds' web service (done by a trivial bash shell script). Special requirements on input data are: - one or more of the following catalogues names can be specified: 'tcl', 'ebarr_cell', 'dsam_miz'. Other names may lead to errors, - when specifying more than one catalogue names, they must be in a unique input string but on distinct text lines, - cell lines names can only be made by a single word, excluding special characters as '/', ',' and ' ', - cell lines names are case insensitive. Example of valid cell lines names are: - vero - hela - a172 - calu6</td>
<td>1.0</td>
</tr>
<tr>
<td>Retrieve descriptions of bacteria strains</td>
<td>Retrieve full descriptions of bacteria strains from CABRI catalogues (see <a href="http://www.o2i.it:8080/axis/services">CABRI site</a>). Inputs of the workflow area the name of the involved CABRI catalogues (text/plain string with one catalogue name per line) and the scientific name of the desired bacteria strain (a text/plain string including genus and species separated by a blank space). Data is retrieved from CABRI Web Services in two steps. First, all bacteria strains IDs are retrieved by using the getBacterialIdsByName method, and after descriptions are retrieved by using the getBacterialBy1d method. Some list/text elaboration is required to remove catalogue names from returned IDs</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Simple demo workflow
O₂I (Oncology over Internet) Project
Your personalized project research web site.

**Workflow details**

**Name:** Retrieve Cell Lines Descriptions By Name

**Description:** 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 getCellLineIdsByNames web service) and then using IDs for retrieving the cell lines descriptions (done via a call to the getCellLinesByIds web service). Both these web services are available at the soaplab system at [http://www.o2i.it:8080/axis/services](http://www.o2i.it:8080/axis/services)

A number of string or string list local elaborations are required: returned IDs are in a string and this must be transformed in a list (done by the 'Separate_cell_line_ids' processor, that is implemented by using a `Split_string_into_string_list_by_regular_expression` local processor) - returned IDs include catalogues' names and this must be removed before their utilization for further processing (done by the 'Extract_ids_by_removing_catalogues_names' processor, that is implemented by using a `Filter_list_of_strings_expecting_match_to_a_regex` local processor) - returned IDs include a blank character and this must be substitutoe by a '_SP_' characters string before submitting the data to the 'getCellLinesByIds' web service (done by a trivial bash script).

Special requirements on input data are: one or more of the following catalogues names can be specified: 'iclc', 'ecacc_cell', 'dsmz_mutz'. Other names may lead to errors, - when specifying more than one catalogue names, they must be in a unique input string but on distinct text lines, - cell lines names can only be made by a single word, excluding special characters as '/', '-' and '*', - cell lines names are case insensitive.

Example of valid cell lines names are: - vero - hela - a172 - calu6

**Author:** Paolo Romano, IST, Genova, Italy (paolo.romano@istc.c.i)

**Roles list:**
- researcher - molecular biologist
- researcher - cellular biologist
- researcher - structural biologist
- researcher - microbiologist
- researcher - immunologist
- researcher - virologist

**Version details**

**Version:** 1.0

**Date:** 18:05 - 27/09/2005

P. Romano, Naples, October 6, 2005
### Version details

**Version:** 1.0  
**Date:** 18:05 - 27/09/2005  
**Author:** Paolo Romano, IST, Genova, Italy (paolo.romano@istge.it)  
**Filename:** RetrieveCellLinesDescriptionsByName.xml  
**Changes:** Original version  
**Image:** show in a new window

**Input list:**
- biological resource database (required)  
- cell line name (required)

**Output list:**
- CABRI human and animal cell lines record

**Domains list:**
- Microbiology  
- Cellular Biology

**Processors:**
Please be advised that is only a list of main components of the workflow. They are not ordered.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Task</th>
<th>Domains</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get cell lines id by name</td>
<td>Retrieve CABRI cell lines' IDs after a search in CABRI Web Services by cell lines' name</td>
<td>biological resource retrieval</td>
<td>Microbiology, Cellular Biology</td>
<td>biological resource database, cell line name</td>
<td>biological resource identifier</td>
</tr>
<tr>
<td>Get cell lines descriptions by id</td>
<td>Retrieves cell lines' descriptions by CABRI id</td>
<td>biological resource retrieval</td>
<td>Microbiology, Cellular Biology</td>
<td>biological resource database, biological resource identifier</td>
<td>CABRI human and animal cell lines record</td>
</tr>
</tbody>
</table>
**O₂I (Oncology over Internet) Project**
Your personalized project research web site.

**Please insert input:**

**CABRI Cell lines catalogues:** (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 Sep 15, 2005, possible values are:
- 'icl' (i.e., the Interlab Cell Line Collection, http://www.icl.it/)
- 'ecacc_cell' (i.e., the European Collection of Cell Cultures, http://www.ecacc.org.uk/)
- 'dsMZ_mutz' (i.e., the collection of human and animal cell cultures of the DSMZ, http://www.dsmz.de).

Catalogues can be added (or, rarely, removed) without notice. See the [CABRI site](http://www.cabri.org/) for further information.

**Cell line name:** (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 the [http://www.cabri.org/](http://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
O₂I (Oncology over Internet) Project
Your personalized research web site.
Your persistently saved results:

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

<table>
<thead>
<tr>
<th>Execution Details</th>
<th>Workflow Inputs</th>
<th>Results list</th>
</tr>
</thead>
</table>
| Date of Execution: 13:08 - 04/10/2005  
Workflow name: Show Gene Ontology Term Context  
(Workflow diagram) |

<table>
<thead>
<tr>
<th>Execution Details</th>
<th>Workflow Inputs</th>
<th>Results list</th>
</tr>
</thead>
</table>
| Date of Execution: 14:44 - 23/05/2005  
Workflow name: Retrieve Cell Lines Descriptions By Name  
(Workflow diagram) |

P. Romano, Naples, October 6, 2005
Some acknowledgements

IST, Genoa
Paolo Romano,
Ulrich Pfeffer,
Domenico Marra,
Valentina Mirisola,
M. Assunta Manniello

ISMAC, CNR, Genoa
Patrizio Arrigo,
Matteo Fattore

ITB, CNR, Milan
Luciano Milanesi

DISCo, University of Milan Bicocca
Guglielmo Bertolini,
Flavio De Paoli,
Giancarlo Mauri

DIST, University of Genoa
Ivan Porro,
Silvia Scaglione

DMI, University of Camerino (MC)
Emanuela Merelli

This work has partially been supported by the Italian Ministry for Education, University and Research (MIUR), project “Oncology over Internet” (2002 – 2005).