Intelligent access to genomic sources on the web

MD Devignes, Nizar Messai, Amedeo Napoli, Shazia Osman, Malika Smail
Orpailleur, LORIA (INRIA-CNRS-Université Nancy 1),
BP239, F-54506 Vandoeuvre-les-Nancy

Introduction

One of the objectives of the Orpailleur group (http://www.loria.fr/equipes/orpailleur/ ) is to
apply knowledge extraction processes and associated knowledge bases to different tasks
within the framework of the semantic web, e.g. web mining, intelligent information retrieval,
content-based document mining (Brachais et al., 2003).

Web sources are widely used in Life Sciences. Scientists are getting more and more
concerned with the problem of exploiting at best all the mass of biological information stored
in the numerous and heterogenous data sources that are available on the web. Many
functionalities are offered by the existing servers : access to the data, execution of programs
such as sequence alignment and analysis tools. Integrated systems offer unified access to
heterogeneous sources and resources (Entrez, SRS, PBIL…). Mediation architectures allow in
certain case-studies the automatic processing of complex queries (TAMBIS). However
solutions provided to the biologists faced with a given biological question remain limited. Our
approach is based on the distinction between two types of problems generated by a complex
query : first the localization and characterization of relevant sources in terms of availability
and query capabilities, second, the collection and integration of data from the selected
sources.

Previous work

Answering a complex biological question can be considered as the execution of a succession
of steps aimed at querying given sources or resources. Such steps have been described in our
previous work (Devignes et al., 2002) in a model that involves the following functions : (1)
selecting relevant sources (for a given step), (2) ranking sources (according to desired
criteria), (3) query construction and submission, (4) extraction of useful data from returned
documents, (5) iteration of (3) and (4) over the sources. Output data should finally be
integrated to build a global answer to the initial question. Chaining of the steps allows when
necessary output data from one step to become input data for the next step.

In practice, functions (1) and (2), as well as the description of the chaining of the steps,
constitute the definition of a scenario or workflow. This process clearly involves the user’s
cooperation and the optimization of the process relies on the user’s particular knowledge.
Various standards are emerging for the description of workflows such as Wf-XML and XPDL
by the Workflow Management Coalition (WFMC,
http://www.wfmc.org/standards/standards.htm), or WSFL used by MyGrid for workflows
involving web services (Wroe et al., 2003).

By contrast, functions (3) to (5) as well as integration of the data into a structured document,
represent the execution of the scenario, i.e. a data retrieval process that is easier to model and
can lead to the development of an application. Automation of this group of functions presents
several advantages : time-saving when answers are required for multiple entries, easy update
of the answers, facilitated exploitation of the answers because of the structured storing format.
The design of a generic data retrieval process (Xcollect project) has been based on two models (Devignes et al., 2004). The generic scenario model appears as a succession of steps. For each step, following information is specified: source name, type and location; input name, type and value (including parameters for appropriate query construction); output name and type; parameters (such as regular expressions) required to extract the useful data from the returned document. An XML DTD has been used to represent this model. A generic session-data model has also been written on the basis of the scenario DTD. It describes in a simple DTD the steps of the scenario with their respective input and output data. According to the desired usage of the data, appropriate XSL transformations allow easy conversion of this generic representation of the retrieved data into various biological meaningful structures.

The BioRegistry project

Current work deals with the formalisation and exploitation of knowledge about sources and their contents towards more efficiency and accuracy in the discovery of relevant sources. The goal is to build a biological registry (BioRegistry) gathering appropriate metadata about web data sources (Boudjida et al., 2000). A survey of various existing standards and resources (Dublin Core Standard, FGDC, DBCat...) led us to distinguish four metadata categories: identification, content, quality and access and availability. Moreover, the BioRegistry should also include metadata tracking and the description of relationships between sources. Well known biological ontologies can be associated to content metadata. At this stage of the work, an XML schema has been designed to implement the BioRegistry model that allows using any available ontology. The first instance documents involve MeSH and NCBI taxonomy. In the future, feeding the BioRegistry with information about novel sources could benefit from web mining procedures.

First exploitation of the BioRegistry is form-based querying, triggering structured information retrieval of the metadata. This should allow the biologist to formulate a multi-criteria query combining various metadata categories and to recover a sorted list of data sources. However, this method does not offer an overall view of the BioRegistry that would allow browsing during the process of source discovery.

An approach for this problem is being tested using hierarchical concept lattices. Actually, Galois lattice is constructed on the basis of boolean table of the form sources\times properties derived from BioRegistry resulting in hierarchical source organisation reflecting property sharing between data sources (Carpinetto et al., 2000; Messai, 2004). The querying of sources then take advantage of the lattice structure. The user query is first inserted into the hierarchy, thus allowing subsequent navigation in the neighbourhood of the best fitting sources.

Perspectives

Ultimately the various functionalities developed around BioRegistry can be implemented as web services enabling interactions with other web applications such as BioMOBY (Wilkinson et al., 2002) and myGRID projects (Wroe et al., 2003). In addition, BioRegistry could become the core component of a mediation architecture enabling invocation of selected sources and aiming at real semantic integration of collected data. This will obviously require a formalisation step of the useful domain knowledge and the definition of relevant inference mechanisms, for instance in OWL.
References


