A Generic Solution for Automated Collecting and Integration of Biological Data from Web Sources

Marie-Dominique Devignes, Yvan Norsa, Malika Smaïl-Tabbone, Philippe Collet, Lionel Domenjoud, Michel Dauça

A Generic Solution for Automated Collecting and Integration of Biological Data from Web Sources

Marie-Dominique Devignes[1], Yvan Norsa[1] and Malika Smaïl[1]
Philippe Collet[2], Lionel Domenjoud[2] and Michel Dauça[2]

[1]LORIA, UMR CNRS 7503, BP239, 54506 Vandœuvre-les-Nancy cedex – France
devignes@loria.fr

[2]University Henri Poincaré, EA 3446, BP239, 54506 Vandœuvre-les-Nancy cedex – France
Lionel.Domenjoud@scbiol.uhp-nancy.fr

Keywords. Integration of heterogeneous databases, source multiplicity, web genomic sources, scenario, XML structured document

Introduction

Web sources are widely used in bioinformatics. 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 public data sources. Many functionalities are proposed by the servers : access to the data, execution of programs such as sequence comparison and analysis tools. Integrated systems exist that offer unified access to heterogeneous sources and resources [1]. Mediation architectures allow in certain case-studies automatic processing of complex queries [2]. However general solution to answer any biological question via web sources do not yet exist. Our approach is based on the distinction between the two types of problems generated by a complex question : first the identification and characterization of relevant sources in terms of availability and query capabilities, second, the collecting and integration of data from the selected sources.
We present here a work dealing with automated collecting and integration of data along a user-defined scenario. Aspects such as query construction, query submission, parsing of returned document, filtering of desired data and storing them in a structured document have been considered as well as the chaining between the various steps of the scenario. Automation of the process allows to refresh the data in a time-saving manner in order to take into account the frequent changes in source contents. A configuration module distinct from the execution module allows to modify the scenario steps according to user preferences and/or source changes.

A Generic "User-Oriented" Approach

1. Two Distinct Problems

Answering a complex biological question can be considered as executing a succession of steps aimed at querying given sources or resources. Such steps have been described in our previous work [3] 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 constitute 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. This process clearly involves user's cooperation and its optimization relies on user's particular knowledge. On the contrary, 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.

We present here the modeling of a generic data retrieval process that has been implemented as the Xcollect application.

2. Modeling Data Retrieval along a Scenario

The first model is a generic scenario model. It 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) necessary to extract the useful data from the returned document. An XML DTD has been used to represent this model.
The second model is a generic session-data model. It describes the steps of the scenario with their respective input and output data. Rapid exploration of available XML DTD or schemas for biological data has revealed the absence of any satisfying standard solution for storing the retrieved data. Therefore a simple generic DTD has been written on the basis of the scenario DTD. Depending on the desired usage of the data, appropriate XSL transformations should allow easy conversion of this generic representation of the retrieved data into various biological meaningful structures.

3. Implementation

The Xcollect application is a java application composed of two modules : the configuration module and the execution module. In the configuration module an interface allows the user to enter manually all the information specifying his scenario. Entered data are stored into an XML document according to the generic scenario DTD. The execution module takes as input the XML scenario document, implements each step of the scenario and returns an XML session-data document containg the retrieved data structured according to the generic session-data DTD. A predefined style sheet converts the XML session-data document into an HTML document to allow visualisation of the session results.

Example : Retrieving the Genomic Context of Short Sequences of Interest

A specific scenario has been designed to answer the following question : « What are the genes localized on human genome in the vicinity of a given short experimentally determined sequence ? ». Automation of the data retrieval process was necessary because many sequences had been produced in the laboratory and their analysis was time-consuming, and also because human genome sequence and annotations undergo frequent updating. Obtained results have been exploited by the users in order to select interesting targets for promoter analyses.

Conclusion and Perspectives

The Xcollect generic application is a platform-independent flexible web application. Experimentation in progress reveals that it can be adapted to various types of scenario and to the frequent changes in source set and/or interface. Further work will deal with the problem of managing multiple answers to individual steps of a scenario and with scenario design. User's knowledge and formalized knowledge about sources should be combined in a complementary manner to address these important issues.

References

[1] S.B. Davidson, J. Crabtree, B.P. Brunk, J. Schug, V. Tannen, G.C. Overton and C.J. Stoeckert Jr, K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM systems journal*, 40:512-531, 2001.

[2] B.A. Eckman, Z. Lacroix and L. Raschid, Optimized seamless integration of biomolecular data. IEEE symposium on Bio-Informatics and Biomedical Engineering (BIBE'2001), Washington DC, pp 23-32, 2001.

[3] M. –D Devignes, A. Schaaf and M. Smaïl, Collecte et intégration de données biologiques hétérogènes sur le web – Xmap : application dans le domaine de la cartographie du génome humain. *Revue des sciences et technologies de l'information (RSTI) – Série Ingéniérie des systèmes d'information (ISI)*, 7:45-61, 2002.