INTEGRATION OF DATA SOURCES FOR PLANT GENOMICS

P. Larmande, C. Tranchant-Dubreuil, L. Regnier, I. Mougenot, T. Libourel
CIRAD, IRD, CINES, LIRMM, CNRS
UMR PIA, TA 40/03, av Agropolis Montpellier cedex 5, 34389, France

Keywords: Plant genomics, interoperability, integration, mediator.

Abstract: The study of the function of genes, or functional genomics, is today one of the most active disciplines in the life sciences and requires effective integration and processing of related information. Today’s biologist has access to bioinformatics resources to help him in his experimental research. In genomics, several tens of public data sources can be of interest to him, each source contributing a part of the useful information. The difficulty lies in the integration of this information, often semantically inconsistent or expressing differing viewpoints, and, very often, only available in heterogeneous formats. In this context, informatics has a role to play in the design of systems that are flexible and adaptable to significant changes in biological data and formats. It is within this framework that this paper presents the design and implementation of an integrated environment strongly supported by knowledge-representation and problem-solving tools.

1 INTRODUCTION

The study of the function of genes, or functional genomics, is today one of the most active disciplines in the life sciences. Researchers in the domain have to manage a constant flow of voluminous information originating from:
- experimental data,
- data that is structured or semi-structured according to pre-existing adapted schemas (databases),
- analysis results,
- annotations,
- output from models, etc.

For example, research into mutations in an organism is one of the study techniques of functional genomics. To identify a mutation in a gene and the consequences it can lead to, the biologist researcher has to consult and compare several information sources such as biological sequencing data obtained from public or private resources. He has also to compare experimental data and data that has undergone diverse procedures to be able to refine and perfect his analyses. Functional genomics is a research subject that therefore relies on several tools, expertise and resources to process and produce information on the function of genes. To this end, access to different necessary resources and services is imperative. However, there exist few integrating solutions for plant genomics and this lack is as much felt here as it is within the biomedical community, whose situation is very similar (Davidson et al., 1995).

One has to admit that the difficulties are many. To identify, sort and annotate sources is a task that is becoming more and more difficult. From a biological point of view, there is need (i) to identify the relevant sources for an analysis from a catalogue that does not stop growing in size (Galperin, 2004), (ii) to confer a degree of confidence to a source and to possibly annotate it, (iii) to keep track of regular updates and to possibly rerun one’s own analyses. From an informatics point of view, one is faced with a diversity of formats because, depending on the need and the concerned experiments, the biological sources can have very heterogeneous structures. For example, data is often in a flat file (e.g., Format GenBank, EMBL, ASN1) but could also be in an XML file or in the form of tabulated data. The structures of the sources also undergo changes and can present variations.

Any attempt to integrate diverse data, procedures, expertise and resources leads to problems of syntactic and semantic interoperabilities in the informatics domain. To these problems must be added the factor that this information is often sensitive (even confidential) and highly variable and subject to change. The goal of the present work is to present an integrated environment for knowledge representation that permits interoperability between different data sources by overcoming problems of heterogeneity. The in-
integration of data sources can be approached from different angles, each with its own advantages and drawbacks (Karp, 2003); we present a brief state of the art in the following section 2. Section 3 will be dedicated to an operational mediation system: Le Select. In the rest of the article, we will show how our exploratory approach is based on the contextualization of this system for use in the plant genomics domain.

2 STATE OF THE ART

The integration of heterogeneous data is a question that has occupied the information-systems scientific community for many years now. And the community has responded in many ways with various solutions. Within the confines of the database domain, the approach has gone from the distributed perspective to the federated databases perspective to the mediation perspective. The arrival of Web technologies has also strongly influenced this concept of integration by offering 'light' integration solutions as well as more complex solutions based on integrators of Web services. Within the genomics context, the solution that is the most handy is the light integration of sources in their original formats. References cross-coded using hypertext links are now found in a large number of biological sources (e.g., GenBank, SWISS-PROT) and searching them by browsing allows fast access to information.

Other, more complex, solutions integrate the sources by offering a common interface and query language to users. Several such systems have been implemented, for example, SRS1, the Entrez2 system developed at NCBI, DBGet3 in Japan or even the French ACNUC4 system (Gouy et al., 1985). In all these examples, the systems are analogous to huge access catalogues: the user selects the sources he want to query from amongst those already indexed.

Recently, proposals for stronger integration have been put forward. Architectures that allow transparent access to distributed sources by support- ing several data types: structured and semi-structured data, flat files, images, etc. It also handles programs executing several data types: structured and semi-structured data, flat files, images, etc. It also handles programs that use this data in the same manner. Le Select is an integration system of type mediator (Manolescu et al., 2002; Cavalcanti et al., 2002) developed at INRIA at Rocquencourt within the framework of the Caravel project5. This type of solution permits cleanup and annotations of imported data (Susan B. Davidson et al, 2001). Thanks to the rapid evolution of Web technologies, applications can now develop Web-based services (responding, for example, to a specific query). These Web services have the advantage of homogenizing the data exchange format, thus facilitating interaction between applications. Moreover, they can be included within a repository. This is the case with the BioMoby repository project which implements a service for locating resources based on a service ontology (P. Lord et al, 2004). Another project, myGRID, helps the user locate a sequence of services appropriate for any one analysis.

In the following sections, we will describe the mediation system that we have used and the stages involved in integrating the sources.

3 LE SELECT

Le Select is an integration system of type mediator (Manolescu et al., 2002; Cavalcanti et al., 2002) developed at INRIA at Rocquencourt within the framework of the Caravel project6. This system provides uniform access for integrating, publishing and interrogating distributed heterogeneous sources by supporting several data types: structured and semi-structured data, flat files, images, etc. It also handles programs that use this data in the same manner. Le Select is a mediator for access to distributed and heterogeneous resources. A Le Select server authorizes the publica-

---

1Sequence Retrieval System, www.infobiogen.fr/srs/
3www.genome.ad.jp/dbget/
4pbil.univ-lyon1.fr/search/query.html
5genoplante-info.infobiogen.fr
6http://www-caravel.inria.fr/LeSelect/
tion of a resource (data or program), while preserving the sources’ independence and format. It allows a certain amount of flexibility of use since the resources can be published incrementally in the system. Before a resource can be published, an adaptor (wrapper) specific to the resource has to be designed (Fig. 1.). The creation of wrappers is the responsibility of the administrator of the concerned resource but a wrapper library dedicated to standard resource types is slowly growing. Wrappers play multiple roles: as translators between the mediator and the resource by offering a homogenous representation based on the relational model and as exporters of statistical metadata on the resources they publish, such as the resource’s availability, the query execution time, the ability to execute a query (ability to join two resources or to test them for equality).

As for the mediator, it offers a pivot language for querying (close to the SQL standard). It includes a query management and optimization program which uses meta-information supplied by the wrappers to establish the execution plan of every query. At a general level, it breaks down the global query into elementary queries, assigning each elementary query to a source that can respond to it, submits the queries, then reconstructs the complete response from the elementary responses.

Finally, since the Le Select server is installed on a Web server, it possesses an interface which can be accessed by a Web browser. Published resources are thus accessible by simple browsing (access by links) or by the intermediation of SQL queries. The fact that Le Select communicates using SQL is of great interest because existing applications can be reused to connect to sources. For example, an application can be connected to the server network and communicate with a mediator via a JDBC bridge (Fig. 1). At the same time, several servers can be installed within the network (specially P2P peer-to-peer networks) and can co-operate to provide access to data and services.

### 3.1 Contextualization of Le Select

In our context, the exchange and distribution of information for the purposes of sharing it is of the utmost importance. Not only does data exchange allow the validation of data by other scientists running identical analyses, but data sharing also leads to generation of new data. Within the framework of a functional genomics project whose goal is to generate a collection of rice insertion lines (*Oryza sativa*) (Sallaud et al., 2004), a database has been created to store experimental data: Oryza Tag Line (OTL)\(^7\). In addition, another database, Rice BRC-db (BRC-db), has been developed for consolidating information on genetic and genomic rice resources.

The short term objective is to bring closer together the correlated information in these two databases, indeed to migrate part of the information from OTL to BRC-db, within the strict condition that the two databases should remain independent (DBMS and interface). In such a scenario, using a mediator approach seemed of interest to us. Before publishing the two information systems, we designed a virtual global conceptual model to solve the problems of semantic heterogeneity and to define correspondences between the entities. We encountered, for example, a case of homonymy (Karp, 1995): the class ‘plant’. In the OTL model, this referred to ‘mutant’ plants whereas in BRC-db they were ‘wild’. In this case, a new ‘plant’ class was created in the global schema to correspond to the ‘mutant’ individuals. To help limit problems of this nature, we referred to shared definitions within the domain: the ontologies (Gruber, 1993). The ones that are most commonly used in plant genomics are Gene Ontology (M.A. Harris et al., 2004) and Plant Ontology (The Plant OntologyTM Consortium, 2002).

Figure 2 shows the publication of OTL via Le Select’s interface. As can be seen, the page is divided into four parts. On the left are displayed the wrappers corresponding to the published sources. At the top, one can select the types of wrappers (wrappers & tables, views and programs). In the central part of the screen is displayed the data corresponding to a query posed in the Query area at the bottom. In the example shown, data is displayed from the TRAIT_VIEW table corresponding to the phenotypical characters observed in the collection of mutants. Data is directly extracted from the database, the table structure is not modified. The wrapper created for publishing this resource uses Java drivers (JDBC).

As we wanted to bring together data from the two databases in conformance with the studied virtual

\(^7\)http://genoplante-info.infobiogen.fr/OryzaTagLine
global conceptual model, we would have had to transform the schema of the sources before publication. But Le Select also offers a mechanism for viewing published sources. For the mediator, the views are also wrappers that execute a query on an already-published source. It is this feature that we have used by creating views of tables that have to be transformed. This establishes, in a simple manner, the correspondences between the OTL and BRC-db database schemas.

4 CONCLUSION

The information systems that researchers in functional genomics have to put together to fulfill their requirements need to preserve the resources’ independence and, very often, the confidentiality of at least a part of their information. The mediation solution is thus of relevance; it conserves the resources’ independence while allowing their distribution and it provides uniform access to information. In fact, even though the materialized approach is also a robust one, it does not handle well the changing character of genomic sources. Unavoidable changes in both systems, OTL and BRC-db, would entail numerous changes to the schema of the data warehouse and, subsequently, to the procedures for loading the underlying data.

The solution implemented using Le Select takes changes in the Oryza Tag Line schema in stride; they are incorporated directly by the mediator. And, by using the intermediary of views, the establishment of new correspondences with BRC-db is also relatively easy. We can thus think that the approach we propose is transferrable to other functional genomics applications. On the longer term, aside from incorporating the integration of programs, the systems should allow researchers to conduct online analyses by authorizing procedures on the data (access to both types of resources having been made transparent).

REFERENCES


