MBLabDB: a social database for molecular biodiversity data

Flavio Licciulli[1]✉, Domenico Catalano[2], Domenica D'Elia[1], Giorgio De Caro[1], Giorgio Grillo[1], Pietro Leo[3], Giuseppina Mulè[4], Paolo Pannarale[3], Graziano Pappadà[3], Francesco Rubino[3], Antonella Susca[4], Saverio Vicario[1], Gaetano Scioscia[3]

[1]Institute for Biomedical Technologies (ITB), National Research Council (CNR), Bari, Italy
[2]Institute of Plant Genetics (IGV), National Research Council (CNR), Bari, Italy
[3]IBM GBS BAO Advanced Analytics Services, Bari, Italy
[4]Institute of Sciences of Food Production (ISPA), National Research Council (CNR), Bari, Italy

Motivation and Objectives

The biodiversity is nowadays one of the main scientific area of interest because of its importance for a sustainable development in many technological domains such as biotechnologies as well as for agriculture and human health. For instance, plant genetic resources are the basis of food security and consist of diversity of seeds and planting material of traditional varieties or modern cultivars and crop wild relatives. These resources are used as food, feed for domesticated animals and in recent years for the identification of new chemical compounds to be used in clinical therapeutic protocols.

Biodiversity research communities have to deal with data coming from many different domains (e.g., biology, geography, evolutionary studies, genomics, taxonomy, environmental sciences, etc.). Collecting and integrating data from so many disparate resources is not a trivial task, data are extremely scattered, heterogeneous in format and purpose, often protected in repositories of diverse research institutes.

With the advent of next generation technologies, molecular biodiversity research is producing large amounts of data that researchers use for complex comparative analyses exploiting information present both in public databases (like GenBank) and in their personal repositories. Improving the management of molecular data and their integration with related information present in the genetic resources databases such as morphologic, geographic and ecologic data will lead to new valuable biodiversity knowledge.

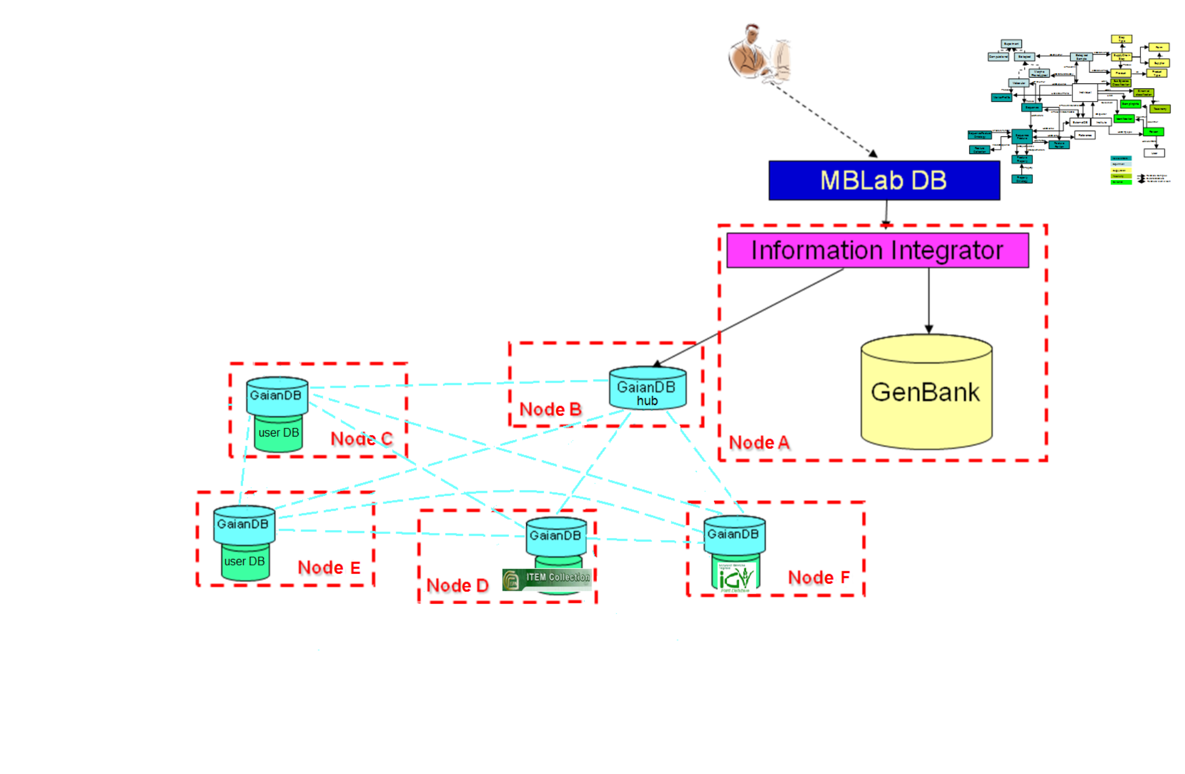
Driven by the widely diffused trend of the web of sharing information through aggregation of people with the same interests (social networks), and by the new type of database architecture defined as dynamic distributed federated database, here we present MBlabDB, a tool representing a new paradigm of data integration in the biodiversity domain.

Methods

MBLabDB uses a hybrid approach of data federation and data warehousing. The system architecture (Figure 1) is based on the integrated cooperation of several components: a robust Database Management System, managing the large volume of molecular data and information available in public resources such as GenBank; a set of federated databases implemented with GaianDB (Bent G. et al., 2008) tool, managing remote specialized biodiversity databases; the IBM Information Integrator, implementing the database conceptual schema and integrating all federated databases with public molecular data using a data warehouse approach.

The conceptual schema of MBLabDB named MolecularBiodiversity Database Schema (Pannarale et al.,2012), is tailored to biodiversity data collection, integrationand analysis. It is modeled on six main sections: Individual, MolecularData, Experiment, Collection, Supply chain and Taxonomy. The MolecularData section is structured following a Chado-like model (Mungall CJ et al., 2007), using Sequence Ontology (Eilbeck K et al., 2005) entities and relations. Similarly the Taxonomy section has been designed in order to incorporate and integrate more than one taxonomy, because of different reference taxonomies that could be related to a taxonomic kingdom.

The federated databases have been implemented by GaianDB (Bent G. et al., 2008), a Dynamic Distributed Federated Database of sources whose growth is regulated by biologically inspired principles and graph theoretic methods. The idea is to create a network of database nodes, each containing specialised collections of biodiversity data, and to expose their content



by means of a GaianDB data server. Information coming from the network nodes are collected by a GaianDB hub and are integrated with public data by means of the Information Integrator server. Two steps are needed to add a new GaianDB node: the installation of a GaianDB server instance and the writing of a wrapper for the mapping of the local schema with the general MBLabDB schema.

An efficient and reliable ETL (Extraction, Transformation and Load) module, implemented with CLIPS Rule Based Programming Language (Pannarale et al., 2012), has been used to integrate GenBank data in MBLabDB. The ETL procedure extracts information from the GenBank entries and fits them into the MBLabDB schema.

The MBLabDB graphical user interface (GUI) has been developed as a Java platform web application. In the GUI the public-private data integration is highlighted through the implementation of taxonomic and ontology based queries.

Results and Discussion

Currently, MBLabDB integrates 4,360,218 entries from the GenBank database and two biodiversity data collections: the ITEM Collection (http://www.ispa.cnr.it/Collection), located at the ISPA-CNR server (containing 9,181 specimen and 3,584 sequences), and the IGV Germoplasm Database (http://www.igv.cnr.it), located at the IGV-CNR server (containing 11,113 accessions). Furthermore the NCBI Taxonomy (www.ncbi.nlm.nih.gov/Taxonomy) and the Catalogue of Life (http://www.catalogueoflife.org/) taxonomic classifications have been included in the Taxonomy section.

Two search and retrieval modalities are available in MBLabDB, an advanced query mode, where search criteria and results can be combined using an incremental composition of "querying & filtering", and an ontology based retrieval that queries data using the biological concepts expressed by the Sequence Ontology.

Therefore, MBLabDB combines public molecular data with biodiversity data contained in genetic resource collections, that are typical of the biodiversity domain. By way of example, using MBLabDB a researcher can extract datasets of sequences related to specimen of his own interest using biodiversity criteria such as species/varieties, geolocation, morphology and passport data.

Using the MBLabDB paradigm of data integration, database hosting, management and information sharing strategy of specialised resources are left to the research group owner of the data collection. So the biodiversity research groups can contribute to the information network by sharing their data sources with a reasonable effort.

In this network, named Social Database for Molecular Biodiversity Data, information remains scattered, but knowledge are shared.

Acknowledgements

This work was supported by DM19410 - Bioinformatics Molecular Biodiversity LABoratory - MBLab (www.mblabproject.it).

References

1. Bent G. et al. (2008) A dynamic distributed federated database. Second Annual Conference of ITA, Imperial College, London
2. Eilbeck K et al. (2005) The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology 6:R44
3. Mungall CJ et al. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 23: i337-i346
4. Pannarale P et al. (2012) GIDL: a rule based expert system for GenBank Intelligent Data Loading into the Molecular Biodiversity database. BMC Bioinformatics 13 Suppl 4:S4