Linking Biological Data Across Organisms in Graph Databases

Luana Loubet Borges[1] and André Santanchè[1]

Institute of Computing – University of Campinas, Campinas, Brasil
luana@lis.ic.unicamp.br,santanche@ic.unicamp.br

Abstract. Representing data as networks have been shown to be a powerful approach for data analysis in biodiversity, e.g., interactions among organisms; relations among genes and phenotypes, etc. In this context, databases and repositories following a graph model (e.g., RDF) have been increasingly used to interconnect information and to support network-driven analysis. Usually, this kind of analysis requires gathering together and linking data from several distinct and heterogeneous sources. In this work, we investigate this challenge in the context of biological bases focusing on the characterization of living organisms, especially their phenotypes and diseases. It includes the rich diversity of Model Organism Databases (MODs) – repositories specialized in a particular taxon – widely used in the biological and medical studies. We exploit a lightweight integration approach, inspired in the Linked Open Data initiative, mapping several biological bases in a unified graph database – our BioGraph – and linking key elements to offer an interconnected view over the data.

The development of computational methods to collect, analyze and store biological data brought unprecedented opportunities to cross data from different organisms. However, there are two main challenges for this kind of analysis. First, data are stored in several distinct datasets, where each repository has its own representation, and they are not interconnected by themselves. Second, it is not trivial to analyze this high amount of data.

This research addresses the problem of crossing data from different organisms, resorting to several databases. It involves creating a – our BioGraph – database to support the search and analysis of the phenotypic data. Its main goal is to develop techniques to transform the phenotypic data from heterogeneous and distinct data sources into a homogeneous format, linking them and crossing phenotype information of different organisms.

The construction of BioGraph can be summarized as follows:

1. We have imported and linked several data sources: intermine [1], MODs – Model Organism Databases –, Uberon [2], Uberpheno, Human Disease Ontology (DO), the Symptom Ontology (SYMP) and other ontologies. There are several formats of data sources and their heterogeneity was a challenge. Each MOD and dataset used to build BioGraph have its specific format.

2. We have created a unified database containing data from all these data sources. To solve the heterogeneity problem, we developed a unified model to support different approaches to describe phenotypes.
3. We have interlinked data from several sources combining two strategies: (i) exploiting existing cross references among sources; (ii) importing bridges between ontologies: Uberon and Uberpheno.
4. With the interlinked graph, we have inferred new edges and nodes, generating knowledge.

Figure 1 shows an overview of BioGraph and how it is organized; it contains: descriptions of phenotypes; Uberon entities; terms of gene ontology; diseases; and symptoms. Edges with labels "link:uberon" and "link:uberpheno" indicate that these edges are derived from Uberon and Uberpheno respectively. In red, we highlight nodes and edges which we created by our inference process.

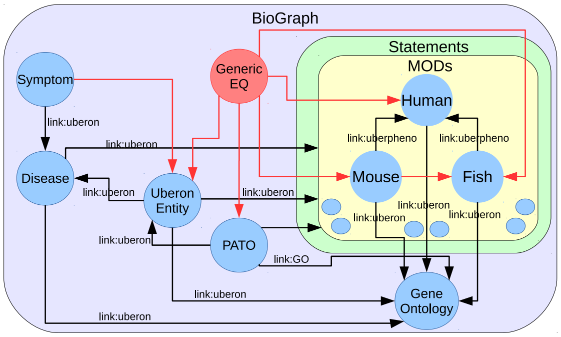

Fig. 1. Domain Model.

The main contributions of this work are: the unified model to support several descriptive approaches for phenotypes and the unified graph database, containing descriptions of phenotypes from 63 distinct data sources. Future work includes: to import genes, linking them with their phenotypes and diseases and to implement an interface for our system.

References

1. Smith, R.N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., Lyne, M., Lyne, R., Kalderimis, A., Rutherford, K., et al.: Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics **28**(23) (2012) 3163–3165
2. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A., et al.: Uberon, an integrative multi-species anatomy ontology. Genome Biol **13**(1) (2012) R5