Mapping-based Data Integration in Bioinformatics: The BioFuice System

Toralf Kirsten[†], Erhard Rahm[†‡]

[†]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
[‡]Dept. of Computer Science, University of Leipzig, Germany

Many bioinformatics applications require data from different sources to answer complex research questions. Integrating such highly diverse data is a major challenge in bioinformatics and often much too laborious and error-prone for scientists. Traditional integration approaches like data warehousing and mediators are often applicable but also time consuming to develop, to deploy and hard to maintain when source schemas change.

We introduce the BioFuice system [1] for interconnecting and integrating data from different autonomous sources. It is based on a decentralized peer-to-peer-like infrastructure. We utilize instance (object)-level correspondences between different sources which are often already available in the sources in the form of web links, e.g. based on accession ids. Moreover, such correspondences can be generated by applying tools, e.g. BLAST to associate similar objects based on its DNA/protein sequence similarity. Sets of such correspondences represent mappings between sources which describe objects of different types, such as genes, proteins, and their function. The object types and their corresponding mappings form the so called source mapping model. Mappings are also assigned a semantic mapping type. Together with object types they reflect the semantics of the domain within a so called domain model. The domain model can be used to categorize sources and mappings so that they can be selected and accessed according to application requirements.

To process queries and mappings we have devised a set of high-level operators. They can be used within script programs (workflows) to combine and analyze data from different sources. Furthermore, BioFuice provides a graphically user interface for explorative analysis and keyword search which automatically generates script programs from interactively specified queries.

Currently, BioFuice integrates data from more than 20 public molecular-biological sources and ontologies, such as Ensembl, GeneOntology, and HomoloGene, but also private sources as result of previous analyses or preferences. The integration approach is applied in various collaborative research projects ranging from analysis of microarray data (IZBI), the analysis of protein interaction networks (MPI MIS) to the detection non-coding RNAs and gene homologues (BioInf).

[1] Toralf Kirsten, Erhard Rahm: BioFuice: Mapping-based data integration in bioinformatics. Proc. 3rd Int. Workshop on Data Integration in the Life Sciences, Hinxton, 2006.