Poster B-3

A Graph-relational Integrated Database of Human and Mouse Omics data

Authors:
Junaid Gamieldien (*NBN South Africa*)
Darren Otgaar (*NBN South Africa*)
Nadeen Southgate (*NBN South Africa*)
Daniel A. Jacobson (*NBN South Africa*)

Short Abstract: We describe a graph-relational integrated data model of human and mouse data, which includes orthology, expression, trait/phenotype information, and several ontologies. Although a prototype, the existing model allows users to perform infinitely complex queries across the integrated resources in a way that approaches 'in-silico hypothesis generation'.

Long Abstract:
We have developed a graph-relational database, Digrabase (DGB) that simplifies the integration of data from semantically rich disciplines, of which biology is probably the foremost example. DGB stores data as a directed graph, where individual universal resources or atoms are stored as nodes/vertices and the associations between them stored are as directed edges. Of the many well-documented graph algorithms, transitive closure, which is probably the most useful to biological research, is currently supported. Instead of storing the actual data in the DGB, which is possible, the system has been designed to store pointers to data in other databases, relational or otherwise, which obviates the need for frustrating data migration efforts. Furthermore, the system has a custom graph query language (GQL) that enables infinitely complex queries to be performed on the data in the graph through the use of Boolean and unary operators. We are also in the process of distributing DGB queries and the graph itself to ensure that the system is scalable to integration projects of any size and complexity.

Our long-term objective is to employ the system to construct a massively integrated graph database of pointers to biological information, built around a large number of formal ontologies. The envisaged resource would be enable powerful queries across data types, knowledge domains, species and even kingdoms; and will result in an information-rich ontology for identifying novel relationships between biological entities and scenarios.

In order to get obtain insights into the longer-term 'real-world' functional requirements for our integration system, we have constructed a mammalian data model based on around the human and mouse data. We have included gene orthology, gene expression, trait/phenotype, gene ontology information and have used the following ontologies, all from OBO, as data integration 'scaffolds' and query startpoints: the GO, human and mouse development, mammalian phenotype, disease, cell type. Most of the human and mouse data was obtained from annotations and ancilliary information from NCBI's Entrez Gene, and additional mouse information (e.g. gene to phenotype associations) were obtained from the Jackson Labs' Mouse Genome Informatics resource (MGI). Human and mouse genes and

their gene orthology relationships were treated as the ‘central’ data objects in this prototype and were also mapped to various ontologies. In addition, approximately 200 Affymetrix microarray experiments from the NCBI’s Gene Expression Omnibus (GEO) database were manually curated to ‘biological sample description’ ontologies. Genes were transitively mapped to the samples and thus ontologies by using the representative probe's 'presence call' as evidence of its expression in its sample. The formal data model will be semantically richer and will incorporate rat and other mammalian information, other high-throughput datatypes, and other ontologies, e.g pathways. Online Mendelian Inheritance in Man (OMIM) and Online Mendelian Inheritance in Animals (OMIA) information will also be mapped to the Mammalian Phenotype ontology.

Although still a prototype, the existing model allows users to query these integrated resources in a way that approaches ‘in-silico hypothesis generation’, e.g.:

"Show all genes on human chromosomes 11 AND 17 whose products are located in the cytosol AND are involved in either cellular proliferation OR the cell cycle AND is expressed in both the mouse AND human brain AND are involved in a nervous system phenotype."

Similarly, users are able to provide a set of genes, to extract ontology terms they were directly or transitively associated based on biological experiments, in a single query. The latter has many applications, from further annotating BLAST results to providing insight to the biological context of a set of genes that result from a microarray experiment.

The preliminary prototype shows that a graph-relational datastructure is an effective solution to the difficulties that the ever-increasing biological data complexity presents RDBMS based bioinformatics data integration systems, particularly querying. Furthermore, once mature, our graph-integrated mammalian data model will have significant utility to research areas such as human and animal health, animal improvement, and particularly studies that employ high-throughput technologies that produce results that often require biological contextualization.