Poster B-45

Digr: A web interface for mining Digrabase graph databases

Authors:
Kieran O'Neill (*National Bioinformatics Network*)
Dan Jacobson (*National Bioinformatics Network*)
Hugh Murrell (*University of KwaZulu-Natal*)

Short Abstract: Digr is an extensible, platform independent, free open source interface to the graph mining functionality of DigraBase. Digr facilitates the finding of co-annotations of objects of interest to ontology terms. It has been applied to a molecular mammalian health model, and an endangered species tissue bank sample tracking system.

Long Abstract:

Digr is a free, open source web interface to the powerful graph mining functionality of the DigraBase graph database system, developed by the National Bioinformatics Network (NBN) of South Africa. Digr provides researchers with the ability to find potential associations between objects of interest, by finding objects co- annotated to combinations of ontology terms.

Digrabase (DGB) stores data in a digraph, with individual entries stored as vertices, and associations between them stored as directed edges. This has the advantage over a traditional relational database that well described graph mining algorithms can easily be applied to the data. At present, transitive closure is supported, although plans exist to implement others. Transitive closure takes a target object type, and a boolean combination of starting points, and finds all objects of the given type, reachable by directed edges from the combination of starting points. An example, from molecular biology, would be to search for all mouse genes expressed in the mouse brain but not in the mouse cerebellum, found on mouse chromosome 12, with analogues on human chromosome 14, involved in fatty acid metabolism. Such queries can be arbitrarily complex. To carry out similar searches, complex scripts to integrate data, or specially designed relational databases would be needed. These would still only serve for a particular query, or subset of queries. DGB permits such queries to be made up and executed "on the fly". To enable searches to be executed rapidly, DGB is clusterable. Furthermore, DGB is generic, and can be applied to any relational data that contains extensive categorical annotation.

This functionality has several advantages for research: When objects, such as genes, are co-annotated to multiple ontology terms, this co-annotation suggests an association. Thus, hypotheses for further research can be generated automatically. Furthermore, a researcher with a hypothesis, looking to find a set of target objects for investigation, can search for objects annotated to the terms of their hypothesis. Finally, a set of objects obtained from a high-throughput experiment, such as overexpressed genes from a microarray experiment, can be put into context by reversing the process, and finding the ontology terms they are annotated to.

Digr aims to provide a usable interface to this functionality, to allow researchers to use the system quickly and efficiently without needing to spend time learning it. To achieve this goal, the querying of DGB, which requires knowledge of a custom query language similar to SQL, is hidden from the user. Instead, the user is presented with ontology terms to query, and, as the terms build, only the ontology terms that have objects of interest co-annotated to the query, are displayed. At present, only intersection of ontology terms ("and") is supported, so as to keep complexity down. However, a more complex interface to allow more advanced users to construct more sophisticated queries will be implemented in future. The interface has been designed according to usability engineering heuristics, and will be refined based on the results of usability testing with researchers. In order to support expected web browser functionality, such as bookmarking and the use of the back button, the system has been built according to Representational State Transfer (REST)/W3C recommended architectural principles.

Digr also aims to be extensible and generic, as DigraBase has the capacity to be applied to an extremely wide range of problems. To achieve this, modern software engineering principles, such as model-view separation, have been employed in its design. Furthermore, Java Enterprise Edition technologies have been used as a framework for the system: Enterprise Java Beans have been used as controllers to a generic DGB client library; Java Server Faces has been used (in part) for HTML output, and Hibernate has been used for persistence. The system has built and tested on JBoss, a free open source J2EE server, but should be readily portable to commercial J2EE servers. As the system is implemented in Java, it can be deployed on most commonly used server platforms. The system also produces standards-compliant HTML, and has been tested on major web browsers. Finally, the incorporation of richer, web-based interfaces, such as Java Applets, Shockwave Flash, or AJAX, can be achieved for data that cannot be represented easily in HTML.

Several data models for mining are under development at the NBN. A mammalian genetic health model, incorporating mouse and human gene annotation data, has been produced as a prototype, and used for the initial Digr design. This model incorporates several public ontologies, including GO and some of the OBO ontologies, the annotated results of microarray experiments, and the direct annotations of genes to ontologies. These have all been parsed into a single DGB graph. The NBN is also developing a sample tracking system for Biobank, a South African endangered species tissue sample repository. This system will use DGB for data storage, thereby making data minable immediately following capture. The ontologies for this system are at a higher level than the first, including such concepts as species, population, and geographical location. However, once molecular biological analysis of the Biobank samples begins, the two models will be fitted together, enabling the discovery of associations between gene functions and other traits at the level of ecosystems. Additionally, a plant genetic model, and a malaria model are under development. This diversity of applications under development has motivated the emphasis on extensibility in Digr.

Digr will be released under a free, open source license, in accordance with the NBN's policy. It will most likely be available as a supplementary application to DGB, which will also be released as free OSS.