7th International Protégé Conference 2004
Presentation abstract

Topics:
Ontology development, medical applications, project management

Title:
"A Protégé-2000 Knowledgebase for domain dependent DNA-Microarray Expression Analysis"

Abstract:
We present the construction and use of a Protégé-2000 Ontology as knowledge-scheme for a molecular biology knowledgebase and show how different plug-ins are used to structure, query and visualize Affymetrix® Microarray expression data in the context of functional genomics.
Our understanding of the term "ontology" is focused on the representation of semantics rather than on standardisation.

Problem:
High-throughput technologies like DNA-Microarray Expression analysis produce massive amounts of experimental data which have to be *mined* for inherent information in a fast, accurate and intelligent way. Information reduction from ten-thousands of probe IDs representing gene-transcripts to a reasonable set with interesting features concerning the scientific context of the experimenter is necessary. This subset of data can then be analyzed and modelled in greater detail.
The three problems we feel are not sufficiently addressed in modern *Laboratory Information Management Systems* (LIMS) are the need for workgroup-dependent data annotation, data representation and querying to help information reduction.

Functional gene-annotations that come with the probes spotted on microarrays are often not suited to serve as search attributes. Because they come from public databases that are set up in a decentralized way, gene annotations are often dependent on the perspective and context of the workgroup that provided the annotation. Fundamental terms are often ambiguous in meaning and the same functions can be expressed in different orthographic variants. Querying an experiment for all "G-protein coupled receptors" for example will not find all genes which

code for these, because they are often annotated with other expressions like abbreviations ("GPCR") or synonyms ("seven transmembrane domain protein"). One even gets false positives, for example when there are annotations like "G-protein coupled receptor binding protein" where the search attribute is part of a description which is semantically different from what one is searching for.

Due to their traditional relational data model only weak data-semantics can be represented in most LIMS. This causes a complete separation of two processes that we think don't have to be separated necessarily - that is storing data on one hand and modelling knowledge on the other.
Usually experimenters store their data in one form, usually a relational database and then build a non-formal textual model of the discovered relationships of the data. The disadvantage is that although the domain dependent knowledge often exists it can't be incorporated into the data-query process because it is not formal enough.
Another drawback of table-oriented data representation commonly used in LIMS is the non-intuitive way of visualization. Context is hard to see in tables because relations aren't visualized.

Solution:
We think it's time to switch from simple data-stores to model-stores representing knowledge rather than data. This means the focus lies more on data semantics and the data-model.
We use the Protégé-2000 Knowledgebase Editor to assist ontology based knowledge management of Microarray expression data in addition to the Affymetrix® LIMS.
We went through a nearly complete knowledgebase development and ontology engineering process which included the following:
Definition of purpose and scope
Definition of application scenarios
Construction of a molecular biology centred domain ontology:

- Knowledge assessment: *mining* domain-specific text for potential concepts and slots
- Integration of existing ontologies and database-schemes
- Ontology creation: structuring concepts to an initial is-a-taxonomy using a text concordance, adding slots.

Knowledgebase creation: selection and import of instance-data.

We pre-structure the experimental data and probeset annotations in a relational Access® Database and then import it as Instance-data into the Protégé CLIPS ontology through the Datagenie plug-in which keeps the relations. For semantic enrichment the gene-instances are then "drag and dropped" manually under the ontological concepts describing their function. Here we use a domain dependent controlled vocabulary which is provided for each gene by domain specialists.

The resulting knowledgebase contains about 800 concepts, 90 slots and represents the 12 626 human genes on the Affymetrix® HU95A microarray with their functional annotation and other context data. This allows experimenters to query their expression data including own domain specific ontological concepts as abstract search attributes. Using the Protégé QueryTab users can not only query for probe IDs and context data, but also for relations between Instances and for implicit knowledge, say "show me all probe IDs under the concept "Enzyme" will also get probe IDs which fall under subconcepts of "Enzyme" like "Kinase". Such inference of implicit knowledge is the big advantage of ontology-based querying compared to simple text-based querying. Combined queries for different search criteria (connected through logical AND/OR statements) can easily be stated, as well as nested queries where an existing query is used for the formulation of a new one. All this through an intuitive query-interface which shields the user from underlying database complexity. The query interface is nearly self-explanatory, so the laboratory user doesn't have to learn a special query-language like MySQL. The ontology serves to constrain possible query-semantics. Once stated, queries can be stored in a query-library for later re-use.
To integrate query-results back into table oriented tools and the Affymetrix®-LIMS an Export-Button was added to the QueryTab, which allows tabular export of selectable slotvalues of the result-Instances.
Besides allowing semantically complex queries, storing data in a formal structured way has further advantages. The object-oriented and frame-based visualization of Affymetrix® probe IDs and context data is more intuitive than tabular visualizations and through the use of the URL-Slot-widget for each probe ID external websites can be shown within a frame representing a probe ID.
Laboratory experimenters can modify and extend the ontology according to their own needs for domain-dependent and intuitive concept-names. The frame-based visualisation can be configured and knowledgebase can be successively expanded according to the domain-specialists needs and current knowledge.

Further potential future uses are currently evaluated as there are visualization of signal-transduction paths and the use of the ontology for annotating free text.

Author:
Daniel Schober

Contact:
Max Delbrück Center for Molecular Medicine, Dept. for Bioinformatics
Robert-Rössele-Straße 10, 13092 Berlin-Buch, Germany
Tel.: +493094062833
Fax: +493094062834
E-mail: schober@mdc-berlin.de
Homepage: http://www.bioinf.mdc-berlin.de/~schober/