Ontology-based retrieval of bio-medical information based on microarray text corpora

Kim A. Hansen, Christian Theil Have and Sine Zambach
Research group PLIS: Programming, Logic and Intelligent Systems
Department of Communication, Business and Information Technologies
Roskilde University, P.O.Box 260, DK-4000 Roskilde, Denmark
{kallanh, cth, sz}@ruc.dk

Abstract. Microarray technology is often used in gene expression experiments. Information retrieval in the context of microarrays has mainly been concerned with the analysis of the numeric data produced; however, the experiments are often annotated with textual metadata. Although biomedical resources are exponentially growing, the text corpora are sparse and inconsistent in spite of attempts to standardize the format. Ordinary keyword search may in some cases be insufficient to find relevant information and the potential benefit of using a semantic approach in this context has only been investigated to a limited degree. We explore the possibilities of retrieving biomedical information from microarrays in Gene Expression Omnibus (GEO), of which we have indexed a sample semantically, as a first step towards ontology based searches. Through an example we argue that it is possible to improve the retrieval of biological information.

Key words: semantic annotation, ontologies, domain analysis

1 Introduction

Ontology-based information retrieval and extraction of semantic information in text corpora are techniques used in connection with text searches. These techniques can help to utilize biological domain knowledge in connection with the retrieval of gene expression experiments, which is an area where the need for shared information is increasingly important. Gene expression profiles are kept on so called microarrays, and are used for data analysis [1]. Though information exchange can be difficult due to the lack of standardization there does exist a metadata guideline that outlines the Minimum Information About a Microarray Experiment (MIAME) [2].

The National Center for Biotechnology Information database, Gene Expression Omnibus (GEO), is a public functional genomics data repository supporting MIAME-compliant data submissions[3], where it is possible to retrieve and download microarray data. Retrieval of data is mainly based on a keyword based text search, which is not always capable of finding all data of interest. Instead, the use of a bio-ontological oriented approach could be beneficial in such cases. However,
an ontological based text search in connection with microarray experiments has only been investigated to a limited degree.

This paper will focus on an experimental approach to a bio-ontological oriented search. By demonstrating this technique through an example, we want to show that it is possible to retrieve relevant information that will not be found using ordinary keyword search. Also, this might lead to a more intuitive approach to search for information in the microarrays. The biomedical data sources is based on metadata from the GEO database, which is imported into our local database. The imported metadata is indexed semantically in order to enable ontological inference. The method used to index the text corpus is similar to that in [4], which is closely connected to the SIABO project [5] and addresses problems about accessing the conceptual content of biomedical texts. Similarly, the microarray-oriented MGED Ontology [3] uses the same approach to fetch data from GEO as we do. They provide a framework which can be used by developers whose environments facilitates the use of ontologies in microarray metadata. It has not been convenient to use in our case, since it is difficult to extend with our method of semantic language processing [5].

2 Methods and Results

Gene expression experiments published in the GEO database are to be described in accordance with the MIAME standard [2]. Despite standardization efforts the amount and quality of the entered information varies a lot and searches in databases can therefore be quite a challenging task. In order to demonstrate the possibility to improve the retrieval of microarray information we analyze a sample GEO experiment using the logic programming language Prolog.

Fig. 1. Illustration of the data flow from fetching the data from GEO to generating Prolog code and returning results.

A GEO experiment is registered with a unique experiment id in the database and has links to microarrays, each registered with a unique microarray id. An experiment is conducted on a specific platform and even reuse of platforms can occur. These platforms are also described and each has a unique platform id and can be cross-referenced. To support our ontology-based search methodology, a process fetches the data from the GEO database. The data is converted into a Prolog format and saved in a local database. In this process entities and
Ontology-based retrieval of bio-medical information

Properties of those entities needs to be extracted from the textual descriptions associated with the experiments.

To illustrate how this method works, we make use of a human stem cells experiment (GSE6015) which has been fetched from GEO. The experiment is recorded according to the MIAME standard and is relatively well-described in all aspects. It would be expected to find such an experiment through ordinary keyword search. Ordinary keyword search would be expected to find such an experiment, but unfortunately this is not the case if we do a keyword search for “adult stem cell”, in which case GSE6015 will not be among the results. Embryonic fibroblast cells, which are mentioned in the experiment text, are in fact adult stem cells. Thus, from a user perspective, this GSE might be a relevant result when searching for adult stem cells. Consider the fragment of an ontology below, which models the adult stem cell relationship:

\[
\text{isa(stem\_cell,cell)}.
\]

\[
\text{isa(adult(stem\_cell),stem\_cell)}.
\]

\[
\text{isa(embryonic(fibroblast\_cell),adult(stem\_cell))}.
\]

Entities and properties are identified in the textual descriptions of the experiment and extracted as Prolog facts. For instance, in the sentence “Embryonic fibroblast cell” from the GSE6015 experiment, the compound noun results in an entity fact and the adjective in a property fact:

\[
\text{entity('GSE6015',1,fibroblast\_cell). property('GSE6015',1,embryonic).}
\]

Using a simple Prolog rule it is possible to compose more complex entities from the entities and properties of the textual descriptions:

\[
\text{composed\_entity(G,Id,Combined) :-}
\]

\[
\text{entity(G,Id,Name),}
\]

\[
\text{property(G,Id,Property),}
\]

\[
\text{Combined =.. [ Property, Name ].}
\]

This rule finds an entity (e.g. fibroblast\_cell) and associated property (e.g. embryonic) and combines them using the Prolog =.. operator to form a composed entity term, embryonic(fibroblast\_cell), suitable for matching the ontology. Consider for instance a sample query for “adult stem cells”:

\[
\text{query(G):- isa(X,adult(stem\_cell)),(entity(G,_,X);composed\_entity(G,_,X)).}
\]

Running the query(G) goal will then produce the result, G=’GSE6015’. The ontology is used to specialize the more general query “adult stem cell” to the more specific text “embryonic fibroblast cell” actually occurring in the textual description of the experiment. This example illustrates just one form of query expansion. For a thorough treatment of ontological query expansion see [6].

3 Discussion

We presented a sketch of methodology for performing ontology based search in the GEO database and demonstrated how such a search may be capable of retrieving information that eludes ordinary keyword based search. Our method
is in an early stage where the ideas are still being refined and there are a number of challenges to be solved.

We are investigating methods for the automation of extraction of entities and properties from the text, which present challenges such as ambiguity, inconsistent naming and anaphora. The matching of a query to a document relies on assumptions about the implied meaning of extracted relations and in the ontology. This approach is not semantically well-founded, but pragmatically useful since it enables users to retrieve the otherwise elucidating relevant experiments. However, traditional information retrieval evaluation measures are difficult to apply due to the lack of a golden standard. Instead, we plan to do a qualitative evaluation and continuous improvement in cooperation with biologists.

An important issue of the functionality of semantic search in microarray-corpora is that the textual, although usually in line with MIAME, is of differing qualities. Some, like the case of GSE6015, have even filled out most of the blanks with a lot of information, whereas others, as in GSE1310 (another embryonic stem cell study), is lacking a lot of information. It is written according to the MIAME standard, although very sparsely so. Sparse annotations is a complication, inherent to much of the microarray-corpora, where an ontology based approach may be advantageous since it can provide a means to bridge the query to the experiment through ontological inference.

The design of the ontology is a demanding task and even with a large effort it may not be elaborate enough to cover all relevant cases. As an alternative it may be interesting to explore a vector-space model [7].

Since it is naive to believe that the annotations will be complete in the future, we find that ontology-based search for microarray data can be useful and is an area that needs more investigating.

References