Using the NCBO Web Services for Concept Recognition and Ontology Annotation of Expression Datasets

Simon Twigger[1,2], Joey Geiger[2], Jennifer Smith[1]

[1] Human and Molecular Genetics Center and [2] Biotechnology and Bioengineering Center,
Medical College of Wisconsin, 8701 Watertown Plank Road,
Milwaukee, WI, 53226, USA
{simont, jfgeiger, jrsmith}@mcw.edu

Abstract. To provide enhanced access to expression datasets housed in the NCBI's Gene Expression Omnibus database and to enable new opportunities for data mining we are using the NCBO's Open Biomedical Annotator service to identify concepts and ontology terms in GEO records. Based on this first pass annotation we are curating these datasets using a variety of ontologies covering concepts of relevance to rat researchers, these include anatomy, rat strains, phenotypes and disease. We have built Gminer (http://gminer.mcw.edu) as a data exploration and curation tool for this work. Data from this project and the Rat Genome Database are available as RDF via the GMiner website for integration with other semantic web tools.

Keywords: Rat Genome Database, National Center for Biomedical Ontology, ontology, Gene Expression Omnibus, RDF

1 Introduction

"Are any of these genes associated with my disease or phenotype? Is this candidate gene expressed in my tissue of interest?" These are examples of common questions asked virtually every day by scientists attempting to identify genes contributing to human disease. Model Organism Databases such as the Rat Genome Database (RGD, http://rgd.mcw.edu) curate published data related to these questions but there is much more information available than can be manually curated. Much of this information is being deposited into large-scale data repositories but extracting useable information and knowledge from this stored data is a challenging problem. The goal of our project is two fold: 1 – to explore the use of ontologies and the National Center for Biomedical Ontology's Web service technologies to annotate large scale repositories such as NCBI's Gene Expression Omnibus (GEO). 2 – To build tools that enable researchers to use the resulting annotations to further their studies of the genetic causes of disease.

2 Results

Based on initial review of the GEO database, the text fields available and the desired queries and uses of rat researchers we decided to focus on a subset of ontologies for our initial work. These included the Mouse Gross Adult Anatomy ontology [1], the Rat Strain ontology [2], Medical Subject Headings (MeSH) [3] and the NCI Thesaurus [4]. GMiner [5], a custom Ruby on Rails web application was built to provide a data management and curation environment for the GEO annotation process. It uses a distributed queuing architecture based around RabbitMQ [6] to parallelize the annotation processing, providing the ability to scale up the annotation process to maximize use of the NCBO Open Biomedical Annotator (OBA) [7].

2.1 Results from initial annotation experiments

Text fields from a thirty GEO datasets and their associated Series and Sample records were annotated with these ontologies using OBA and the annotation results reviewed using GMiner. Tag cloud-based visualization and data exploration interfaces to the annotated data were developed along with a variety of curation interfaces to review and classify the OBA annotations to the GEO text. Annotations to the Rat Strain ontology had a high false positive rate due to the short length of many rat strain symbols (2-3 letters, eg. BN, ACI) and their similarity to a number of common words (AN, AT) and US state abbreviations (e.g. SD, Sprague Dawley rat strain and the state South Dakota). In contrast the Mouse Adult Gross Anatomy ontology provided a very low false positive rate, particularly when used on specific GEO text fields such as the 'Source name' field of the GEO Sample records.

Based on our success using the anatomy ontology to annotate the small-scale dataset we loaded all of the GEO Dataset, Series and Sample records for the various rat Affymetrix platforms housed in GEO and annotated specific fields using just the mouse anatomy ontology. This consisted of 192 datasets, 479 series and 12,012 sample records. In our initial pass we were able to get 9951 ontology associations to 7,609 records. A review of the anatomy annotation results was done for the GEO dataset records (Table 1) giving precision and recall values of 84.0% and 94.8%, respectively. This compares favorably to studies done by Shah et.al. [8] who obtained 83% precision and 86% recall when using the NCI Thesaurus to identify disease associations in GEO dataset records. These results are in contrast to those obtained for when using the rat strain ontology (Table 2, precision 32.3%, recall 55.4%). The poor performance of the rat strain annotation is in part due to the short strain names and clashes with common words and abbreviations, the variety of words used to describe rat strains many of which were not present as synonyms in the ontology (increasing the false negative rate) and problems with other, incorrect, synonyms in the ontology leading to a higher false positive rate. Improvements to the Rat Strain ontology and annotation workflow are in process to improve the accuracy of this process.

Table 1. Accuracy of identifying Mouse Anatomy annotations in the Title and Description fields of the Rat Affymetrix GEO dataset records loaded into GMiner.

	Correct	Incorrect	Total
Positive	384 (TP)	73 (FP)	397
Negative	13 (TN)	21 (FN)	94
Precision: 84.0%			Recall: 94.8

Table 2. Accuracy of identifying Rat Strain annotations in the Title and Description fields of the Rat Affymetrix GEO dataset records loaded into GMiner.

	Correct	Incorrect	Total
Positive	31 (TP)	65 (FP)	94
Negative	119 (TN)	25 (FN)	144
Precision: 32.3%			Recall: 55.4%

2.2 Linking anatomy annotations to gene-level expression data

Annotation of GEO sample records with anatomical terms provides an opportunity to explore extrapolating these results to the individual probeset or gene level in order to pull out additional data from the GEO record. Affymetrix platforms have a detection score, which indicates if a probeset was determined to have been present or absent in the given sample, i.e. if the transcript for that gene was present in the sample being studied. If the GEO sample record is tagged with an anatomy term it should then be possible to say that probeset X, corresponding to Gene A, was present in anatomical structure S, indicating that Gene A has been shown to be expressed in structure S. There are many caveats to this approach particularly as it glosses over the many varied experimental conditions being studied and the statistical significance of individual observations. That being said, even at a qualitative level this is useful information that scientists can now access as a result of this ontological indexing and evaluate for themselves before deciding how to act on this data. We are evaluating these results in comparison to known experimental tissue expression data such as that found in the Novartis BioGPS [8] to determine their utility.

2.3 Data availability and integration using RDF

Another goal is to make any data produced by the project available to the community as a whole. We will do this for our main audience on the web via the Rat Genome Database and the Ratmine data warehouse, however, we have also been evaluating RDF as a way to make these results available to the wider semantic web community. After a review of existing RDF resources and formats, we have created RDFizers for the rat gene data available at RGD, built around URI formats found at bio2rdf.org. We have added Affymetrix probeset to rat gene RGD ID mappings derived from

Ensembl's biomart and then our own preliminary probeset to mouse anatomy associations. This data was loaded into a local OpenRDF Sesame data store for evaluation. To allow us to explore the data more easily we have been investigating the available ruby RDF libraries to provide programmatic access to Sesame and other triple stores such as Virtuoso. We have had some success with ruby-sesame and ActiveRDF though neither have provided a comprehensive solution to date. RDF flat files are available at http://gminer.mcw.edu/rdf for RGD's rat gene dataset and for curated associations between GEO records and the Mouse Anatomy and the Rat Strain ontologies.

3 Discussion

The NCBO's OBA services have provided a very effective mechanism for us to rapidly annotate records from GEO, an example of a large data repository containing a lot of very useful information. As expected, subsequent manual review is necessary to validate and clean up the automated associations but this is made substantially less burdensome by the initial automated results. Our efforts to investigate integrating and publishing these results in RDF format have been both encouraging and frustrating. It has been encouraging to see many ongoing efforts to make biological data available as RDF, in particular the bio2rdf.org datasets have been very useful. However, it also illustrated the potential benefits of an 'approved' RDF model for standard objects such as genes, chromosomal locations, etc. with namespaces and predicates laid out for novice RDF producers to copy as a way to get their data linked in as painlessly as possible.

References

1. Mouse Adult Gross Anatomy ontology: Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. Genome Biol., 2005, **6**:R29
2. Rat Strain Ontology: http://bioportal.bioontology.org/ontologies/39234
3. Medical Subject Heading: http://www.nlm.nih.gov/mesh/
4. NCI Thesaurus: Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. J Biomed Inform. 2007 **40**:30-43.
5. GMiner: http://gminer.mcw.edu
6. RabbitMQ: http://rabbitmq.com/
7. Open Biomedical Annotator: http://bioportal.bioontology.org/annotator
8. Ontology-driven indexing of public datasets for translational bioinformatics: Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. BMC Bioinform. 2009 10:S2/S1
9. Novartis BioGPS: http://biogps.org