Introduction to Biomedical Ontologies 4:
Term Enrichment Analysis Using the RatMine Widgets

This video is the fourth in a series on biomedical ontologies. In the previous videos we introduced what an ontology is, talked about where ontology annotations come from and where they can be found, and looked in detail at the various components of ontology annotations and the valuable information those components contain. In this video, we will begin to look at how you can use ontologies to further your research objectives.

This video will introduce one of the common uses for ontology annotations, that is, analyzing the annotations assigned to a list of genes to determine whether or not there are ontology annotations that occur more frequently in that list compared to the annotations for a control set of genes. This is referred to as "Ontology Term Enrichment".

Term enrichment analyses are commonly used for results of experiments such as microarray analysis. For example, a list of genes which are found to be more highly expressed under the experimental conditions are examined to determine if they appear to have functions in common or to participate in any shared processes or pathways.

There are many tools available both on the web and as downloadable programs which do gene ontology term enrichment. The Gene Ontology Consortium website lists many such programs that are available for public use. We will only talk about a single tool, but the basic concepts can also be applied to other applications.

The tool that we will be examining is the InterMine ontology enrichment widgets. InterMine is an open access data warehousing, mining and analysis tool. The technology behind InterMine was originally developed to power the FlyMine application but is now being used by other groups including modMine which is in use for the modENCODE project and RatMine at the Rat Genome Database. For this video, we will look specifically at RatMine.

Ontology enrichment analyses in RatMine begin with a list of genes or proteins. These can come from queries done in the RatMine tool via the templates, or they can be uploaded to RatMine by the researcher. RatMine also has a number of useful lists already saved and available for use. For information about RatMine in general, check out the "Introduction to the RatMine Database" video on the RGD website at http://rgd.mcw.edu/.

To upload a list of genes, start at the RatMine home page and click on the "List" tab at the top of the page or on the link in the box below the "Templates" section. This will take you to the "Lists" page. If this is the first time you've been to the Lists page you will see the list upload tool. At the top of the page are tabs for both view and upload. Clicking the View tab will take you to a page which shows both RatMine's preloaded, publicly available lists and any lists that you have uploaded.

To load your own list, click the link for "Upload". On this page, select the type of objects in your list, in this case, gene. Type or paste a list of identifiers such as RGD IDs or gene symbols for genes or UniProtKB IDs for proteins into the text box, or use the "Browse" button to locate a file on your computer which contains a list of identifiers. Once you have your list entered, click "Create List". You will be prompted to give your list a name and then save it. If you are not logged into RatMine, any lists you save will be lost when your session ends. However, registering for a free account and logging in will allow you to save lists permanently to your account and re-use them at your convenience.

Once you have saved your list you will be taken to the List Analysis page for that list. You can see that at the top of the page, information is given about the genes in this list and about the list itself. You also have an option to convert the objects in the list from genes to their corresponding proteins.

Below this information are the ontology term enrichment widgets. Across the top of this section is the list of available widgets. By default they are all highlighted, meaning they are all being displayed. To turn one off, click the widget name in the list. To turn it back on, just click it again.

As you can see, unlike many online enrichment calculators which use only the Gene Ontology, RatMine includes data for all of the ontologies that are associated with the genes in RatMine. These include a Disease Ontology, RGD's Pathway Ontology, the Mammalian Phenotype Ontology and the KEGG pathways in addition to the Gene Ontology.

Ontology term enrichment analysis is a statistical calculation of how frequently a particular term appears as an annotation in an experimental group of genes or proteins compared to how often that term appears in the annotations for a control set of genes or proteins. RatMine currently uses the set of all rat genes as its control group. Although in some situations, such as when only a small group of genes was tested, this may not be the best control, even in those cases it can be used as a first approximation. And for large scale tests such as genome-wide association studies, the set of all rat genes is the appropriate control.

Although the RatMine widgets are not completely customizable, there are factors that you can change to increase or decrease the stringency of your analysis. These include the maximum p-value to display and which algorithm, if any, is used for multiple hypothesis correction. The cutoff for the p value defaults to 0.01, but can go as high as 1. As you can see, increasing the maximum p-value to be displayed increases the number of results returned.

RatMine widgets also include two algorithms for multiple testing: Benjamini and Hochberg, which is the default, and the Bonferroni correction, in addition to the option of not doing a multiple hypothesis test correction. Although a discussion of which algorithm might be better in a particular situation is beyond the scope of this video, we will say that often selecting "none" will increase the number of results returned whereas selecting "Bonferroni" which is generally considered to be a more conservative method, may reduce them.

The results of these widget analyses can be downloaded for use in other applications, or some or all of them can be viewed in a results table within RatMine. For instance, if we check the two terms related to "Bone Diseases" in this example, and choose "Display checked items in results table", we see the annotations for either those terms or more specific child terms underneath them in the ontology.

One of the great advantages of the RatMine tool, is the ability to then save this list and use it for further analysis—either using these same term enrichment widgets, or other RatMine list manipulation tools, such as looking for overlap between this list and another. Here we've saved the shorter list and found that the enrichment analysis for this subset of our original set of genes, not surprisingly, includes terms such as "ossification" and "skeletal system development" for the GO Biological Process Ontology and "hypothyroidism" and "Musculoskeletal Abnormalities" from the Disease Ontology.

Summary:

During this video, we've looked at how you can upload and save a list. We reviewed what term enrichment analyses RatMine can do on the genes or proteins in that list. We looked briefly at some of the factors that can be changed to refine your results, and we've seen how you can then use the results of one such analysis to further analyze a particular subset of your original list. For more information and to try out the RatMine ontology enrichment widgets for yourself, go to ratmine.mcw.edu.

For more information:

To try the RatMine ontology enrichment widgets for yourself, go to...
RatMine:
http://ratmine.mcw.edu

InterMine:

http://www.intermine.org/

FlyMine:
http://www.flymine.org/

modMine
http://intermine.modencode.org/

InterMine's documentation about Enrichment Widgets:
http://www.intermine.org/wiki/EnrichmentWidgets

The Gene Ontology Consortium

GOC Complete List of Tools
http://www.geneontology.org/GO.tools.shtml

Tools for Data Set Analysis
http://www.geneontology.org/GO.tools.microarray.shtm

GOC Searchable Publications list:
http://www.geneontology.org/cgi-bin/biblio.cgi

A few publications which may contain helpful information:

Biomedical ontologies: a functional perspective.
Rubin DL, Shah NH, Noy NF.
Brief Bioinform. 2008 Jan; 9(1): 75-90. Epub 2007 Dec 12. Review.
PMID: 18077472

Biomedical ontologies in action: role in knowledge management, data integration and decision support.
Bodenreider O.
Yearb Med Inform. 2008: 67-79.
PMID: 18660879

Ontologies for molecular biology and bioinformatics.
Schulze-Kremer S.
In Silico Biol. 2002;2(3):179-93.
PMID: 12542404

PCA2GO: a new multivariate statistics based method to identify highly expressed GO-Terms.
Bruckskotten M, Looso M, Cemiĉ F, Konzer A, Hemberger J, Krüger M, Braun T.
BMC Bioinformatics. 2010 Jun 21;11:336.
PMID: 20565932

Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology.
Xu T, Gu J, Zhou Y, Du L.
BMC Bioinformatics. 2009 Aug 5; 10: 240.
PMID: 19653916

GOstat: Find statistically overrepresented Gene Ontologies within a group of genes
Beissbarth T, Speed TP
Bioinformatics, 6.2004; 20(9): 1464-1465.
PMID: 14962934

GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly

enriched Gene Ontology terms associated with a list of genes.
Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G.
Bioinformatics. 2004 Dec 12;20(18):3710-5. Epub 2004 Aug 5.
PMID: 15297299

Controlling the false discovery rate: a practical and powerful approach to multiple testing
Benjamini, Yoav; Hochberg, Yosef
Journal of the Royal Statistical Society, 1995, Series B (Methodological) 57 (1): 289–300.

Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives
van der Laan, Mark J.; Dudoit, Sandrine; and Pollard, Katherine S.
Statistical Applications in Genetics and Molecular Biology: Vol. 3 : Iss. 1, Article 15, 2004.

What's wrong with Bonferroni adjustments.
Perneger TV.
BMJ. 1998 Apr 18;316(7139):1236-8.
PMID: 9553006

Funding for the Rat Genome Database is through grant #2R01HL064541, administered by the National Heart, Lung and Blood Institute.

This work was also supported through a Driving Biologial Project collaboration with the National Center for Biomedical Ontology (Grant U54 HG004028)