Text Mining with CoPub

This article was featured in *Library Notes* #49 (February 2009).

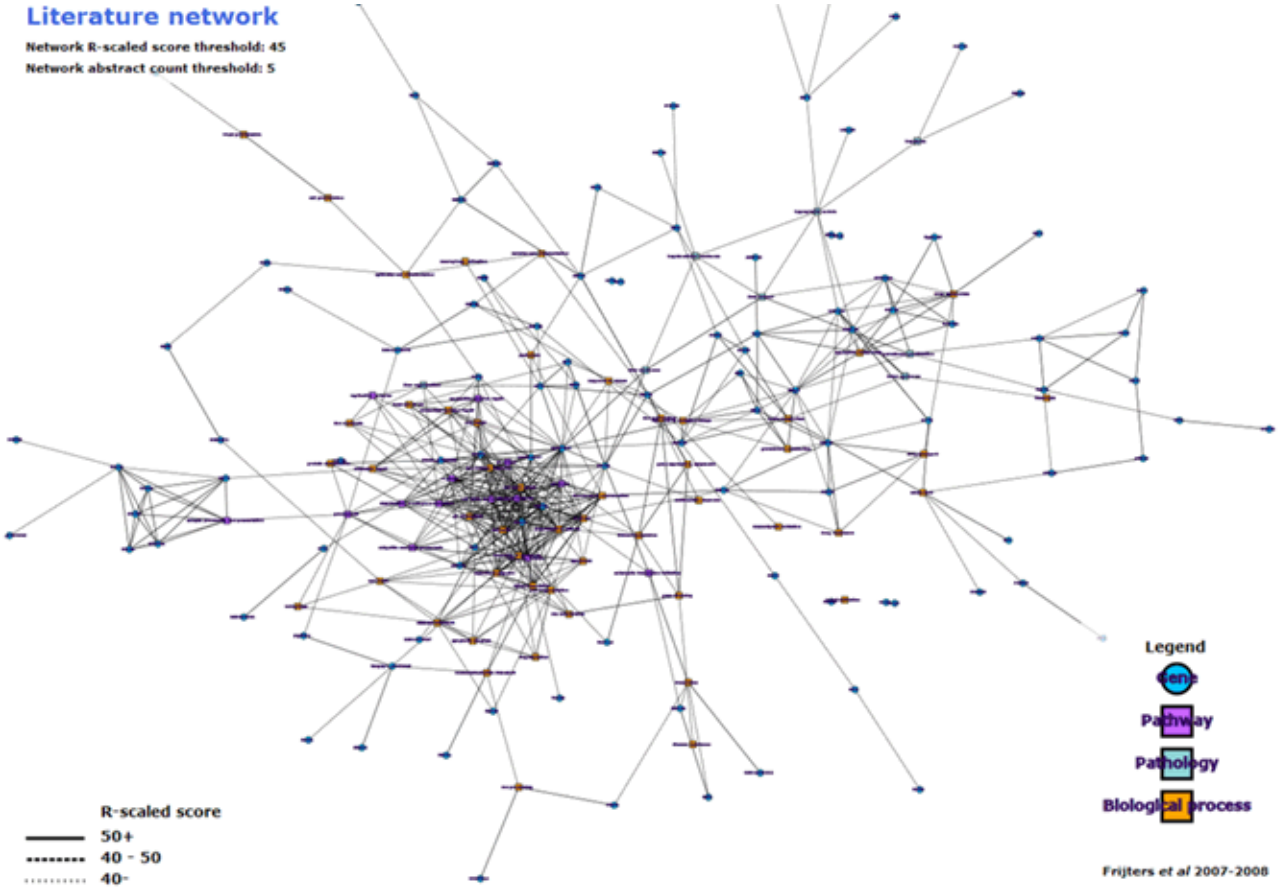
Free, online text mining tools are scarce and leave a lot to be desired in their functionality. Searching by gene identifiers can a frustrating and elusive task with many databases. You can use the Entrez Gene database or OMIM and link to PubMed results related to the gene you're investigating, but searching MEDLINE directly for gene identifiers returns results that lack both specificity and sensitivity.

CoPub is a search tool that was created to improve searches for genes and biological processes. It was developed by Raoul Frijters and Jan Polman at Organon (part of the Schering-Plough Corporation) and is now hosted and further developed by SARA with support of NBIC.

CoPub searches MEDLINE abstracts with sophisticated scripts that can detect gene symbol homonyms and exclude them from the search results, thus increasing specificity in gene name searching. The application uses Gene Ontology (GO) definitions for biological process, cellular components and molecular functions, and also returns results for pathways, drugs and diseases.

The user can search by gene, bioconcept or even by microarray sets. In the microarray search, CoPub will accept a set of Affymetrix gene identifiers and will search for the entire set of genes in MEDLINE. CoPub then returns a set of literature in which any subset of these genes are co-published (hence the name of the site: CoPub).

CoPub returns a table of results, based on the types of processes you selected. Each item in the table has a PubMed ID that is a link to the record in EMBL. This may seem unusual, but I suppose it is more efficient for the Netherland-hosted site to search the European-based EMBL instead of the U.S.-based PubMed. In the microarray search, CoPub will also return a network map of publications in which nodes are interactive links to the literature.



CoPub was featured in the *Nucleic Acids Research* 2008 Web server issue: Frijters R et al. CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res*. 2008 Jul 1;36(Web Server issue): W406-10. PMID: 18442992

While it is easy to search and returns useful results, CoPub has some flaws:

- It does not allow combined Boolean searching using and, or, not
- Its user manual, while helpful, is a pdf format, and thus not interactive
- The microarray search currently only allows Affymetrix identifiers as a search set
- It does not search up-to-the-minute MEDLINE contents: the current version (as of November 2008) searches the February 2008 Medline data set

What CoPub demonstrates, though, is that open-source text mining servers are getting more effective. This will streamline users' gene-based searches, which is always a good result.

Give CoPub a try.

Pamela Shaw
Biosciences & Bioinformatics Librarian
312-503-8689
e-mail Pam

The Biosciences Blog highlights new tools and news items of interest to the biosciences research community at Northwestern University.