GoPubMed:
Exploring PubMed with the GeneOntology

Andreas Doms and Michael Schroeder
Biotechnological Center and Dept. of Computer Science
TU Dresden, Germany
ms@biotec.tu-dresden.de
April 11, 2005

Abstract

The biomedical literature grows at a tremendous rate and PubMed comprises already over 15,000,000 abstracts. Finding relevant literature is an important and difficult problem. We introduce GoPubMed, a web server which allows users to explore PubMed search results with the GeneOntology, a hierarchically structured vocabulary for molecular biology.

GoPubMed provides the following benefits: First, it gives an overview over literature abstracts by categorizing abstracts according to the Gene Ontology and thus allowing users to quickly navigate through the abstracts by category. Second, it automatically shows general ontology terms related to the original query, which often do not even appear directly in the abstract. Third, it enables users to verify its classification because GeneOntology terms are highlighted in the abstracts and as each term is labelled with an accuracy percentage. Fourth, exploring PubMed abstracts with GoPubMed is useful as it shows definitions of GeneOntology terms without the need for further look up.

GoPubMed is online at www.gopubmed.org. Querying is currently limited to 100 papers per query.

1 Background

Limits of classical literature search. The biomedical literature grows at a tremendous pace. PubMed, the main biomedical literature database references over 15,000,000 abstracts. Due to this size, simple web-style text search of the literature is often not yielding the best results and a lot of important information remains buried in the masses of text.

Consider the following example: A researcher wants to know which enzymes are inhibited by levamisole. A keyword search for "levamisole inhibitor" produces well over 100 hits in PubMed. To find out about specific functions, the researcher has to go through all these papers. He/she is interested in the relevant enzymatic functions. From the first titles it immediately is evident that levamisole inhibits alkaline phosphatase. A less well-known fact is however still buried in the abstracts. The abstract The effect of levamisole on energy metabolism in Ehrlich ascites tumour cells in vitro with PMID 2947578 is ranked very low (position 89 on 7/2/2005) by PubMed. The abstract states that levamisole also inhibits phosphofructokinases. Most readers will miss this statement.

Even if the user would try to reduce the number of papers by filtering out the ones mentioning levamisole inhibitor (e.g. query PubMed for "levamisole inhibitor NOT phosphatase"), he or she would miss the less obvious hits like phosphofructokinase, if both terms occur in the same abstract. Thus, even advanced PubMed queries with boolean logic cannot always properly structure the search.

1Please note, that all examples in this paper depend on PubMed’s ranking of search results. Since the literature is growing, PubMed may return different articles for the same query at different time points. This means that GoPubMed may display different papers for the examples in this paper. All queries in this paper were checked on 8 Feb 2005.
Figure 1: User interface of GoPubMed. The screen-shot of GoPubMed displays the results for the query "levamisole inhibitor" limited to 100 papers. On the left, part of the GeneOntology relevant to the query is shown and on the right the abstracts for a selected GO term. The search terms are highlighted in orange and the GO terms in green.

Right of each abstract is a list with all the GO terms for that abstract ordered by an accuracy percentage. E.g. is the term P-glycoprotein, which is a synonym for the GO term xenobiotec transporting ATPase, found with 100% accuracy, while lung development matches only with 72%, as only the word “lung” occurs in the abstract. Synonyms, such as the term P-glycoprotein above, are displayed in dark grey and the synonymous term is given in a tool-tip (please note, that Mozilla based browsers do currently not break lines in tool-tips). Moving the mouse over the term displays the definition of the term in a tool-tip. The ontology on the left shows the paths from the root of the ontology - cellular component, biological process, and molecular function - to the currently selected GO term. The number in brackets behind each GO term in the ontology is the number of papers the GO term or any of its children occur in. In the figure, the path from molecular function to alkaline phosphatase is shown and the number 71 behind the term alkaline phosphatase indicates that there are 71 papers mentioning alkaline phosphatase. Clicking on the term displays the relevant abstracts, which confirm that levamisole inhibits alkaline phosphatase. Overall, the number of papers containing a term and its children is a very good indicator to let users select the most frequent terms and thus best representatives. Instead of using the ontology to browse through abstracts, users can also display all the abstracts in the same order as in PubMed with the additional benefit of displaying the GO terms and search keywords. Users can also search within the ontology using the input field at the bottom of the ontology. GoPubMed searches are currently limited to 100 papers per query. Answering a query takes around 20 seconds.
The GeneOntology. We propose to improve literature search by using ontologies, which are controlled, hierarchical vocabularies. The ontologies are used to categorize and explore abstracts. Currently, one of the most prominent ontology is the GeneOntology (GO) [2], which has been designed for the annotation of gene products. It comprises over 19,000 terms organized in three sub-ontologies for cellular location, molecular function and biological process.

GO was initially created to reflect Drosophila in the Flybase database, but has expanded to encompass many other genomes as well as sequence and structure databases. The hierarchical nature of GO allows one to quickly navigate from a rather abstract to very specific terms. As an example, there are maximally 16 terms from the root of the ontology to the deepest and most refined leaf concept in GO.

Extracting Terms from Abstracts. The main problem that needs to be solved before we can use ontologies for literature exploration is term extraction. Finding ontology terms exactly in the literature is rarely possible, as authors do not write their abstracts with the Gene Ontology in mind. Consider e.g. the excerpt ...

Example: Which enzymes are inhibited by levamisole? To illustrate the power of this approach let us consider the levamisole example again. Consider Fig. 1 and 2, which show screen-shots of the GoPubMed web server. The user wants to learn which enzymes are inhibited by levamisole. He/she submits "levamisole inhibitor" with GoPubMed. GoPubMed classifies the papers with GO and the user can explore the ontological classification of the papers:

- Out of the 100 papers some 50 papers mention terms, which are cellular components, some 90 papers mention biological processes, and some 90 molecular functions.
- Selecting molecular function and then catalytic activity, the user finds cyclases, transferases, isomerases, hydrolases, lases, small protein conjugating enzyme activity, and oxidoreductases.
- Consider Fig. 1. Hydrolases are mentioned in 81 papers. Refining this term, the user learns that there are 73 occurrences of phosphoric ester hydrolase activity, 72 occurrences of phosphoric mono-ester hydrolase activity, and finally 71 occurrences of alkaline phosphatase. The titles of these abstracts such as e.g. Effects of alkaline phosphatase and its inhibitor Levamisole... immediately sustain that levamisole inhibits alkaline phosphatase.
- Consider Fig. 2. Exploring the transferases, which occur in 14 papers, the user finds one article listed under phosphofructokinase activity. The abstract of this article states that levamisole directly inhibits tumor phosphofructokinase (PMID 2947578).
To summarize, GoPubMed allows users to quickly answer, which enzymes are inhibited by levamisole. The most obvious enzyme, alkaline phosphatase, is also the most frequently occurring in GoPubMed. The lesser known phosphofructokinases clearly show up in GoPubMed, while being deeply hidden in a classical PubMed search result list.

**Example: Author profiles.** GoPubMed is generally useful to gain an overview over a set of articles and to define a profile for these articles. This feature can be used to quickly get an insight into the topics a researcher is working on. Specifying e.g. the name and affiliation of a researcher as query to GoPubMed one will be able explore the researcher’s interest and focus of research. In particular, the induced GeneOntology can serve as a profile representing that researcher. As an example, consider Kai Simons in Dresden. The PubMed query "simons dresden" returns some 20 articles. The induced ontology for these papers indicates that he is working on cell organisation and biogenesis (within the process ontology) and in particular on lipid raft formation, a term that is found in 13 papers.

**Example: Actin.** Which term is most obviously related to actin? Many researchers will promptly reply myosin. In GoPubMed such obvious relationships can be identified by exploring the most frequently occurring GO terms. In the case of actin GoPubMed suggests that some 80 papers mention cellular components or any sub-terms, nearly 80 papers cell or sub-terms, some 70 intracellular, 67 cytoplasm, 57 cytoskeleton, 50 actin cytoskeleton and 9 myosin. Thus, in only 5 clicks the user can relate actin and myosin and even underpin this relationship through the statements of associated abstracts, such as PMID 15679101: *Syntrophin was also able to inhibit actin-activated myosin ATPase activity.*
Example: Rab5. After querying with "Rab5" the ontology shows among the biological processes the path physiological processes → cellular physiological processes → cell growth and/or maintenance → transport → vesicle-mediated transport → vesicle endocytosis. Selecting the papers mentioning vesicle endocytosis, there are statements such as

- PMID 15328530: The small GTPase Rab5 is a key regulator of clathrin-mediated endocytosis.
- PMID 15199393: Downregulation of several small GTPases, such as rab5, rac1, and rhoA, which regulate endocytosis, was found in CP-r cells.

Inspecting the ontology for cellular components there is a path cell → intracellular → cytoplasm → endosome → early endosome. Associated articles contain e.g. the statements:

- PMID 12876219: Rab5 small GTPase is a famous regulator of endocytic vesicular transport from plasma membrane to early endosomes
- PMID 14636058: Rabaptin-5 interacts with Rab5 and is an essential component of the fusion machinery for targeting endocytic vesicles to early endosomes.

Example: Molecular functions associated with osteoclast differentiation. Querying with "osteoclast differentiation bone resorption" the ontology shows a path molecular function → signal transducer activity → receptor activity → receptor binding → G-Protein coupled receptor binding → chemokine receptor binding with a descendent chemokine activity. The paper with PMID 15265944 supports this stating In this study, we examined the effect of MIP-1 gamma, a C-C chemokine family member, on receptor activator of NF-kappa B ligand (RANKL)-stimulated osteoclast differentiation, survival, and activation.

Example: MMP2 & VEGF. Which morphogenetic processes can be associated with the matrix metalloprotease MMP2 and the vascular endothelial growth factor VEGF? The query "MMP2 VEGF" results in an ontology with the path biological process → development → morphogenesis → organogenesis → blood vessel development → angiogenesis. For the latter the paper PMID 15389539 provides the following evidence which plays an important role in activation of MMP-2 and VEGF to induce angiogenic process and promotion of inflammation-associated adenoma formation in mice.

3 Comparison and Conclusion

GoPubMed is related to three other tools, namely Textpresso [3], XplorMed [4], and Vivisimo (vivisimo.com). Textpresso [3] is an information retrieval system based on a set of some 30 high-level categories, which can be seen as a very shallow ontology. Parts of the category members are based on the GeneOntology. Using these categories, Textpresso can answer queries like "Which abstracts mention an allele and a biological process in the title?" There are four main differences between Textpresso and GoPubMed: First, Textpresso uses only 30 categories for classification, while GoPubMed uses the fill GeneOntology, not limiting itself to the top concepts. Second, Textpresso returns a list of relevant abstract, while GoPubMed uses the deep ontology as vehicle to navigate through a large result set in a non-sequential order. Third, Textpresso is designed for full papers on C. elegans, while GoPubMed works on all the PubMed abstracts. Forth, Textpresso tries to find the category terms directly in the
text only allowing for some variations in lower/uppercase letters and plural forms. GoPubMed uses an algorithm, which allows for gaps within matches and considers the information content of words, which leads to a more refined term extraction. This is necessary, as most GO terms are not found directly in free text.

XplorMed [4] maps PubMed results to the eight main MeSH categories and then extracts topic keywords and their co-occurrences. For the query "levamisole inhibitor". XplorMed returns 22 relevant co-occurring words such as activity, protein, cell, which are, however, very general and do not shed any light on the enzymes inhibited, for example.

Vivissimo is closely related to GoPubMed as it also uses an ontology to explore search results. However, instead of the GeneOntology, Vivisimo automatically derives an ontology from the search results. While this ensures that the ontology closely matches the articles, the ontology itself cannot be as well structured as a hand-curated one like GeneOntology.

There are numerous other tools, which use the GeneOntology to explore data other than literature abstracts. Many of them cater for the annotation of gene expression data and are based on GOA, the Gene Ontology Annotation, which annotates sequences with GO terms. For a comprehensive list of these tools please refer to the GO web site www.geneontology.org.

Acknowledgment: We gratefully acknowledge support of the EU project REWERSE (IST-2004-506779). We kindly thank the reviewers for their comments, which helped to improve the paper.

References


