

GoPubMed: Exploring PubMed with the GeneOntology

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.

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.

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 results.

Extracting Terms from Abstracts. The main problem that needs to be solved before we can use ontologies for literature

The screenshot displays the GoPubMed web interface. At the top, there is a search bar with the query "levamisole inhibitor" and a "Go" button. Below the search bar, the interface is divided into two main sections. On the left, the "Induced Gene Ontology" is shown as a tree structure with terms like "levamisole inhibitor" (100/127) and "biological process" (100/127). On the right, the "Results for 'levamisole inhibitor' and GO term 'alkaline phosphatase activity'" are displayed. This section includes a list of abstracts with their titles, authors, and affiliations. The first abstract is titled "Increase of alkaline phosphatase in multidrug-resistant tumor cells and their cross-resistance to 6-thioguanine" and is by authors A. K. Ghosh, S. Ghosh, and S. Ghosh. The second abstract is "Ecto-alkaline phosphatase activity identified at physiological pH range on intact P19 and Hc-60 cells is induced by retinoic acid" by authors S. Ghosh, S. Ghosh, and S. Ghosh. The third abstract is "Possible involvement of a magnesium dependent mitochondrial alkaline phosphatase in the regulation of the 25-hydroxyvitamin D3-1 alpha- and 25-hydroxyvitamin D3-24R-hydroxylases in LLC-PK1 cells" by authors S. Ghosh, S. Ghosh, and S. Ghosh. The interface also shows a list of GO terms on the right side of the results, such as "alkaline phosphatase activity" (70/127) and "magnesium ion binding" (60/127).

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 *xenobiotic transporting ATPase*, is 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.

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.

For GoPubMed we have developed a term extraction algorithm. It uses local sequence alignment of words of the abstract and the words of GeneOntology terms.

Example: Which enzymes are inhibited by levamisole?

To illustrate the power of this approach let us consider the levamisole example again. Consider Fig. 1, which show screenshots 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*, *lyases*, *small protein conjugating enzyme activity*, and *oxidoreductases*.
- 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.
- 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 to 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*.

Successes

- GoPubMed to be published in Nucleic Acids Research (Impact Factor 6.575).
- GoPubMed presented at W3C Workshop on Semantic Web for the Life Sciences.
- GoPubMed attracted so far 1415 hits of 214 users worldwide although it is only propagated by word-of-mouth.
- GoPubMed and reasoning. The existing application already computes the induced ontology for a query. This will be refined in future releases.