GO2PUB

Use of Gene Ontology for Enriching Genetics-related Pubmed Queries

Abstract

Literature search on PubMed takes more and more time as PubMed growths. There is a need for automated search tools, which must have a better precision and recall than PubMed basic query system. We developped GO2PUB to answer this demand in the field of the genetics. Our purpose was to use the knowledge within the Gene Ontology (GO) to build PubMed queries and display results for a quick access to the information. GO annotations link genes products to the metabolic pathways in which these are involved. Genes annotated by a GO term concerning a pathway that interest us can be used as keywords for a PubMed query. Following the GO true path rule, it allows to query PubMed with many genes names from only a few numbers of GO terms. GO2PUB can build additional queries from users keywords to obtain more relevant results. For each result, it displays PMID, title, authors, date, abstract and journal. Name, symbol and synonyms of genes that have been used as keywords thanks to the GO term entered by the user are highlighted. GO2PUB is available at http://go2pub.genouest.org/

Contents

1 Introduction

1.1 Pubmed

PubMed is the most complete public database of biomedical literature. However, querying PubMed manually is a long process. It require first to build the best queries possible to obtain the more relevant literature. PubMed user has to read several dozens or hundreds of abstracts to select the relevant ones. Although beneficial for the scientist community, the growth of PubMed also means that these queries have to be repeated regularly in order to keep up with and an increase of the time and complexity of literature selection through PubMed querying system. Indeed, PubMed comprises near 20 million entries for biomedical literature from MEDLINE, life science journals, and online books.

1.2 The need

PubMed' size and growth call for automatic tools for helping the users build queries as precise as possible (hence complex) in order to minimize silence and noise. An approach to build such queries is based on query enrichment thanks to controlled vocabulary, such as ontologies. An important querying scenario concerns genes products involved in a specific metabolic pathway for different species. Gene Ontology annotations are a link between the genes products and the metabolic pathways in which these are involved. Over 30,000 GO terms describe biological processes and molecular functions of products of genes, as well as cellular components in which they are involved. These GO terms are different of Medical Subject Heading terms (MeSH) used by PubMed as biological keywords and are required to find genes products involved in a specific metabolic pathway. The names, symbols and synonyms of annotated genes are so many keywords for a PubMed search.

1.3 GO2PUB purpose

An ontology with its rules makes possible an automated approach to exploit the knowledge contened by GO for a literature search. GO terms are useful to collect genes products involved in a metabolic pathway, that we can use as keywords. Thanks to GO semantics, it is possible to query PubMed with many genes names from only a few numbers of GO terms, respecting the true path rule. The results obtained are more genetics oriented. GO2PUB purpose is to automatically generate all the queries and to compile the results.

2 Method

2.1 Use genes as keywords

GO2PUB creates an enriched PubMed query from the name, symbol and synonyms of genes annotated by one or several asked GO terms, for one or several species. The enrichment process follows the GO true path rule, adding to Pubmed query all genes annotated by all descendants of the asked GO terms following the ontology. For a better targeting, the user can precise the query with so much of wanted MeSH terms as keywords.

2.2 True path rule

A single term can annotate a lot of genes, directly or indirectly through the true path rule. So GO2PUB builds a query on the model "(n genes name, symbol or synonyms separated by OR) AND (m species) AND (p MeSH terms)". This big query is splitted into several smaller if it exceeds Pubmed server url length limitation. GO2PUB compiles results and displays all citations numbered and sorted by date.

2.3 GO2PUB output

GO2PUB yields literature about genetics in a field described by one or several GO terms, for one or several species. It run automatically the generated requests on PubMed, doing in a few minutes a work which would ask for hours manually. Results are formatted for a quick access to the information. Each citation obtained from PubMed is listed;

title, authors, date, abstract, journal, PMID and MeSH terms are displayed. The name, symbol and synonyms of gene annotated by the asked GO term(s) are highlighted in title and abstract.

3 How to fill the form

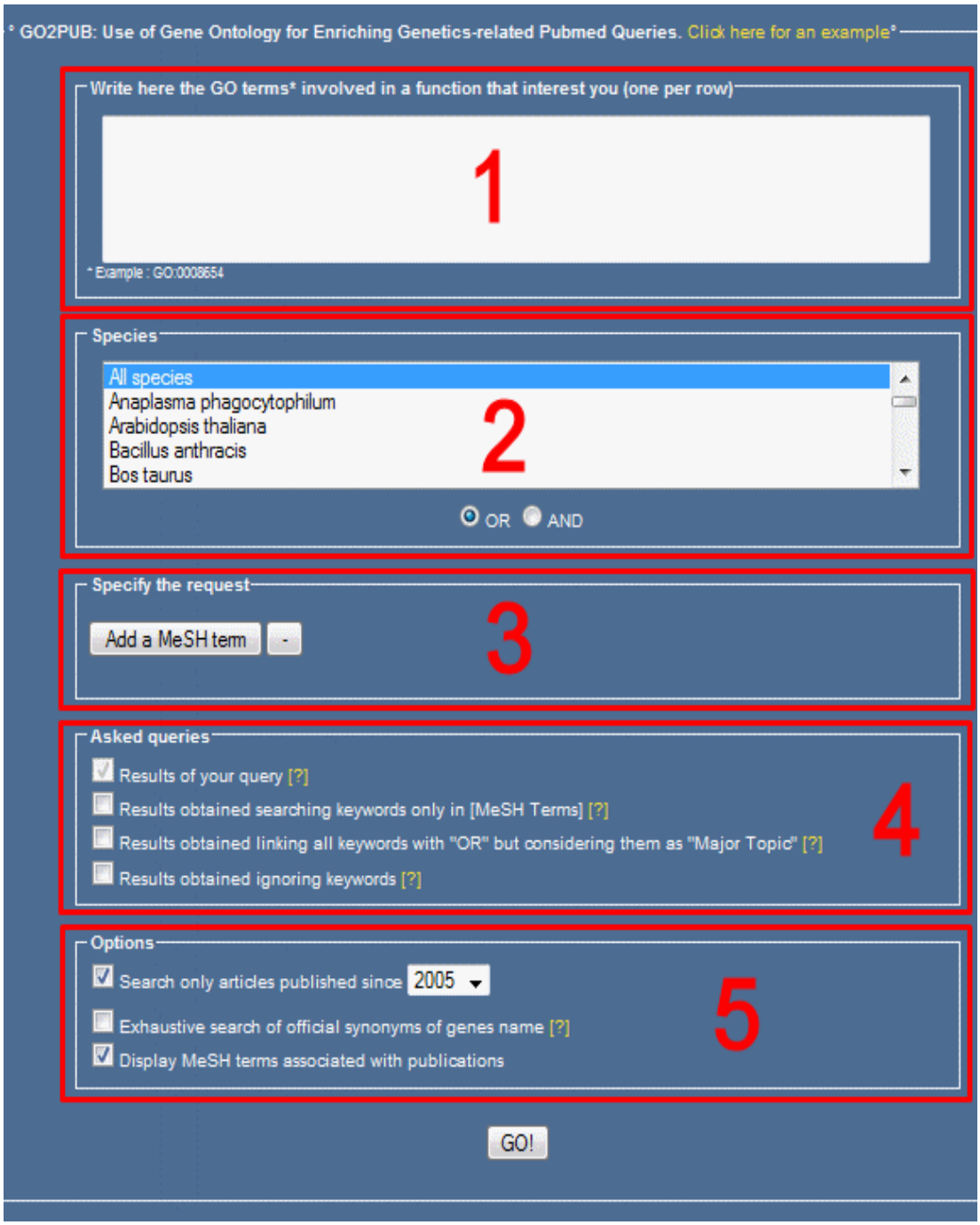

Figure 1: Screenshot of GO2PUB form. The 5 fields are described in next page.

3.1 Enter your GO term(s)

You can enter one or more GO term(s) here. A visit to Gene Ontology web browser (Amigo) is required to obtain exact GO terms and corresponding codes, which are needed by GO2PUB. Because of the true path rule of GO, when you enter a term, it will yield information from this term AND from all its descendents. So the more general is your term, the more results it will yield. We recommend to not use terms having a depth under 4 in GO hierarchy (too general term).

3.2 Select species

You can select a species here. You can select several species if you want by pressing Ctrl key while clicking on species. You can use "AND" or "OR" connectors between species. The more species you choose, the more the search spends time.

3.3 Precise the query with MeSH keywords

For a better targeting, you can enter some MeSH terms here. It decreases the number of false positives in results. A visit to MeSH page on NCBI web site is useful to find pertinent MeSH keywords in your research field. MeSH terms associated at the articles by PubMed are not all of same importance: some of them are "Major topic" (MAJR) classified. You can precise it for each keyword you enter. You can use "AND" or "OR" connectors between your keywords. At this point, you have builded your basic query.

3.4 Build additional queries

GO2PUB options proposes to build some derivated queries based on yours. Your main query will yield results the most close to those waited. However, it could be interesting to see if minor changes can bring you additionnal pertinent results. Indeed, 3 additionnal queries may be added. First, ignoring MAJR parameter and searching all keywords simply in PubMed [MeSH] tag. According to the number of MAJR tags initially presents, it brings more or less additionnal results. As MAJR terms are also MeSH terms, articles associated to them will still be found; so GO2PUB eliminates doubloons. The second derivated query available replaces "AND" connectors between keywords by "OR" connectors. However, as it can yield many more results, of which a large number of irrelevant, all keywords in this additionnal query will be tagged with MAJR. Species, that are normally searched in MeSH, will also be tagged with MAJR for this second additionnal query. The third and last derivated query available simply ignore keywords, and search species in MAJR. This option is to be carefully used, because it can return many hundred of results if the topic of the search is too general. It has of the interest only for the very precise subjects, for which the other requests return only few results.

3.5 Other options

GO2PUB proposes 3 other options. First one allows to set a publication year limit. The second option is about an exhaustive search of official synonyms of genes name. It consists in looking in Entrez Gene for all the synonyms recognized for a gene. As the authors sometimes use in their articles synonyms that are absent in the GOA database used to obtain genes information, it allows to build more complete PubMed queries to obtain more results. The last option concerns the display of the MeSH table associated with each article, which can be or not activated.

4 Credits

GO2PUB Development: Charles Bettembourg
Tests and evaluation: Olivier Dameron, Christian Diot, Anita Burgun
GO2PUB Hosting: GenOuest BioInformatics Platform
Banner realisation: Isabelle Stévant