W300–W304 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 22 May 2009*
doi:10.1093/nar/gkp429

GoGene: gene annotation in the fast lane

Conrad Plake[1,*], Loic Royer[1], Rainer Winnenburg[1], Jörg Hakenberg[1,2] and Michael Schroeder[1]

[1]Biotechnology Center, Technische Universität Dresden, 01307 Dresden, Germany and [2]Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

Received January 30, 2009; Revised April 21, 2009; Accepted May 11, 2009

ABSTRACT

High-throughput screens such as microarrays and RNAi screens produce huge amounts of data. They typically result in hundreds of genes, which are often further explored and clustered via enriched GeneOntology terms. The strength of such analyses is that they build on high-quality manual annotations provided with the GeneOntology. However, the weakness is that annotations are restricted to process, function and location and that they do not cover all known genes in model organisms. GoGene addresses this weakness by complementing high-quality manual annotation with high-throughput text mining extracting co-occurrences of genes and ontology terms from literature. GoGene contains over 4 000 000 associations between genes and gene-related terms for 10 model organisms extracted from more than 18 000 000 PubMed entries. It does not cover only process, function and location of genes, but also biomedical categories such as diseases, compounds, techniques and mutations. By bringing it all together, GoGene provides the most recent and most complete facts about genes and can rank them according to novelty and importance. GoGene accepts keywords, gene lists, gene sequences and protein sequences as input and supports search for genes in PubMed, EntrezGene and via BLAST. Since all associations of genes to terms are supported by evidence in the literature, the results are transparent and can be verified by the user. GoGene is available at http://gopubmed.org/gogene.

INTRODUCTION

High-throughput gene expression assays are nowadays common practice, for example, to compare expression levels of genes in a pathological state (diseased) to those in a control state (healthy) and then filter for genes that are significantly up or down-regulated between the two states. The outcome is usually a long list of genes that requires further investigation. Biologists analyze these genes, for instance, by exploring known GeneOntology (GO) annotations. However, manually curated GO annotations do not cover all genes and also do not cover relevant diseases, mutations, drugs, anatomical parts, etc. They also do not provide information to rank genes according to what is new or well-studied already and not all annotations have links to the literature, which is important for further study and finding related work.

GoGene associates all genes from different model organisms with concepts of GO and MeSH (Medical Subject Headings), two vocabularies that cover a variety of areas of biomedical research. The hierarchical structure of both vocabularies makes it possible to cluster and summarize long lists of genes. Because most knowledge is contained only in publications and not in databases, GoGene integrates manually curated gene annotations, literature references (GeneRIFs) and textual comments from UniProt and EntrezGene with text-mined annotations from all of PubMed. In doing so, more than 4 000 000 associations between genes and ontology concepts for the model organisms human, mouse, rat, worm, fruit fly, zebrafish, thale cress, baker's yeast, fission yeast and *Escherichia coli* are made available, thereby increasing the number of known GO annotations by one order of magnitude. Additionally, GoGene provides ~35 000 gene-mutation associations automatically extracted from PubMed abstracts, which are not contained in UniProt. All associations are linked to their origin (i.e. literature or database entries) for further investigation. By scanning the literature, GoGene also compiles publication histories for each gene that are used to rank genes according to what is new, what is widely studied, or what is of high impact (see Usage and examples section). All relevant concepts for a gene list are displayed as a tree that allows for quickly navigating through long lists of genes.

RELATED WORK

Many tools exist that facilitate the functional analysis of genes. Huang *et al.* (1) give a summary of 68 tools to study

*To whom correspondence should be addressed. Tel: +49 351 463 400 60; Fax: +49 351 463 400 61; Email: conrad.plake@biotec.tu-dresden.de

the enrichment of gene-related functions. However, those tools rely on annotation databases that are unlikely to be complete and not all of them are easily accessible via a web server. Babelomics integrates different approaches for the functional analyses of gene expression experiments including text-mined annotations but does not allow searching the literature directly (2). The Panther database contains a collection of protein families that have been further divided into functional subfamilies (3). It can be searched by keywords or gene lists to find related ontology terms, genes, or subfamilies together with family annotation information. However, as for Babelomics, no literature search is supported and neither Panther nor Babelomics provide disease or other medical annotations. On the other hand, text mining tools such as iHOP (4), GOAnnotator (5), EBIMed (6), or PolySearch (7) allow searching the literature for gene-related concepts, but they do not integrate known annotations from databases (except for GOAnnotator) nor do they provide means to filter and cluster genes based on their associated annotations. GoGene combines text-mined facts from all the literature in PubMed with annotations from the databases EntrezGene and UniProt. All annotations are displayed as a tree following the structure of GO and MeSH that can be seen as a table of contents to quickly find genes related to one or more categories.

USAGE AND EXAMPLES

GoGene accepts three kinds of queries: lists of EntrezGene identifiers or gene names, keyword queries that are sent directly to PubMed to find genes in the matching abstracts, or nucleotide or amino-acid sequences that are BLASTed against UniProt using its remote service interface (8). The type of a query can be specified by adding one of the following tags at the end: [gene], [pubmed], or [blast]. If the user does not specify the query type, GoGene tries to automatically find the best result. First it checks if the query resembles a nucleotide or amino-acid sequence. In this case, the BLASTX or BLAST service is invoked, respectively. Otherwise, the query is sent to both PubMed and EntrezGene and the larger gene list is returned. For EntrezGene results, genes are ranked as in the original result list. Results from a PubMed search are ranked by occurrence frequency in the matching abstracts in descending order. A BLAST search result is ranked by sequence similarity, with the most similar gene listed first. Each gene is associated with the bibliometric features: community, volume, and novelty. The community value represents the number of authors in PubMed who published about the gene. The volume is the sum of journal impact factor points from all publications mentioning the gene, and novelty is the sum of impact factor points, where later publications are weighted higher than older ones. We chose a yearly decrease of 50% towards the past. In GoGene, each gene list can be re-ranked according to these three bibliometric features by clicking the links/buttons in the result summary. A gene list (including all annotations) can also be exported to a file in different formats (SIF, GML, GraphML). The export links are located at the bottom of the result page.

Example 1: a search for rat genes related to osteoporosis and bone resorption

Each day PubMed grows by thousands of publications. Thus, keeping track of new developments in a research field can be very time-consuming. GoGene helps to search the literature for discussed genes and gene-related information in PubMed. Consider a biologist who is interested in rat genes related to osteoporosis and bone resorption. A keyword search for 'osteoporosis bone resorption' in the Rat Genome Database gives no results. The same query in EntrezGene results in only two rat genes (*Pth* and *Tnfsf11*). In PubMed, the query 'rat osteoporosis bone resorption' returns 857 publications. Reading 857 abstracts to identify genes is a cumbersome task. GoGene alleviates this task by automatically searching all relevant abstracts for genes and displaying the resulting gene list (here: 142 genes) to the user. In the tree, the biological process, bone resorption, the organism, rat, and the disease, osteoporosis, are listed as top categories. Selecting each as mandatory (green checkbox) augments the query and eventually leaves five rat genes (*Pth*, *Tnfsf11*, *Ctsk*, *Tnfrsf11b*, *Csf1*) for which literature references clearly state their role in osteoporosis and bone resorption. (All searches were performed on January 17, 2009.)

Example 2: finding genetic causes for the disease hepatoerythropoietic porphyria (HEP)

Hepatoerythropoietic Porphyria (HEP) is a metabolic disease caused by uroporphyrinogen decarboxylase deficiency in the liver and bone marrow. Searching GoGene for 'hepatoerythropoietic porphyria' shows the enzyme-encoding human genes, *FECH* and *UROD*, as the top hits. When looking at the detailed gene pages, one sees lists of mutations that where found in PubMed for those enzymes. For *FECH*, the most recent publication from 2007 reports on a novel missense mutation found in a patient with HEP, where an alanine was substituted by a threonine at position 185. For the gene *UROD*, one finds a publication from the same year where the authors report on a novel mutation causative for HEP (G168R). (All searches were performed on January 17, 2009.)

Example 3: exploring a pancreatic cancer microarray screen

When plotting the results of a microarray screen one often discovers many deregulated genes. The next step is to find interpretations for the deregulation, for example, by mapping genes to known functional annotations. Here, we consider a set of 400 genes found to be deregulated in pancreatic cancer taken from a genome-wide study by Grützmann *et al.* (9). After entering the EntrezGene identifiers into the query box the resulting genes are shown together with gene-related functions, processes, cellular components, diseases, etc. (Figure 1). Out of those 400 genes, 24 are linked to pancreatic cancer, and 248 genes are linked to cancer in general. Most of the genes encode membrane proteins and the most common process for

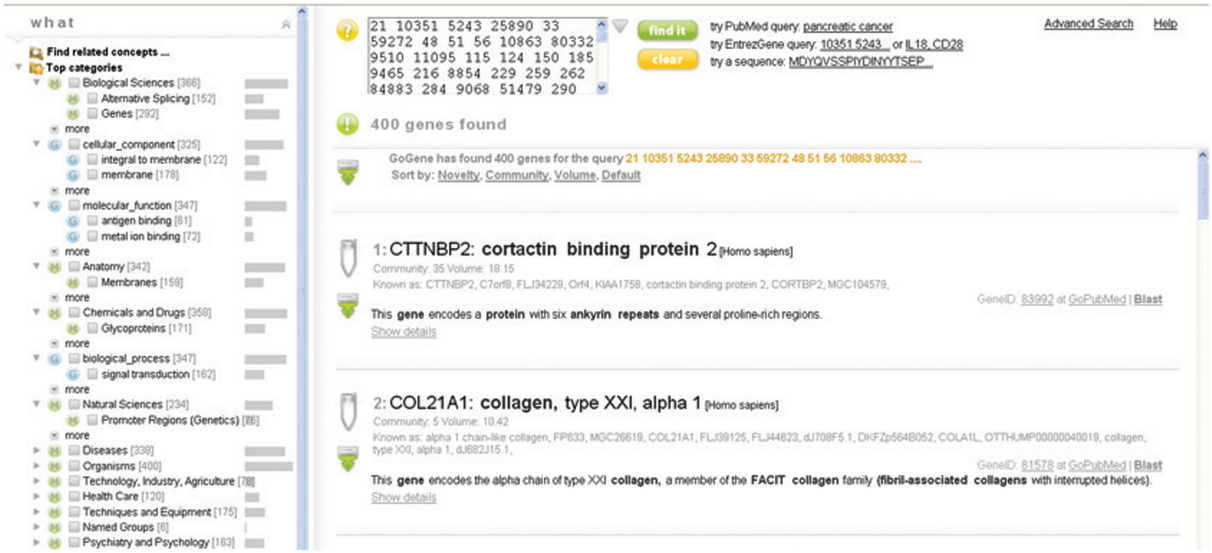

Figure 1. The GoGene web interface showing a result from a search for 400 gene IDs taken from an outcome of a microarray experiment on pancreatic cancer. On the left, all relevant concepts from GO and MeSH are shown. Clicking on a concept shows the related genes on the right. Each gene entry primarily consists of a title, a short summary and a link to a detailed gene page, where all information on a gene is summarized.

those genes is the transduction of signals. Clicking on one of the concepts shows only the related genes together with highlighted text snippets and hyper-links to relevant articles in PubMed. Thus, within seconds the user gets an overview what the genes in a gene set have in common and can quickly follow the links to the literature to find the relevant publications. (All searches were performed on January 17, 2009.)

Example 4: searching functional annotations for the uncharacterized human protein C15orf39

For the majority of proteins neither structural nor functional information is available. One approach to gain insights into the function of a protein is to compare its sequence to the sequence of other proteins where the function is already known. Proteins with similar sequences (e.g. homologs) are likely to share similar functions. For example, the uncharacterized human protein C15orf39 (UniProt-ID Q6ZRI6) is listed in UniProt without any GeneOntology annotation. After searching GoGene with its sequence, genes that code proteins with similar sequences are retrieved. When exploring the annotations of those genes by looking at the GO & MeSH tree, one finds almost 25% of genes are classified as phosphotransferases. The most frequent molecular functions are binding and catalytic activity, such as hydrolase activity. (All searches were performed on January 17, 2009.)

METHODS

Gene annotations derived automatically from literature are predictions that rely on the correct identification of genes and ontology concepts in text. For gene identification, we applied a context-sensitive algorithm that achieved an *F*-measure of 81% in the last BioCreative challenge for the task of human gene normalization and has been further improved to also normalize genes from other model organisms (10,11). Terms of GO and MeSH found in PubMed abstracts are provided by the GoPubMed webserver (12). For each term found, the corresponding text is implicitly mapped to all its ascendant terms according to the GO or MeSH hierarchy (transitive closure). For instance, if the abstract is annotated with pancreatic cancer, then it is also annotated with cancer. All possible gene-term pairs are assigned an association score based on the point-wise mutual information:

$$\mathrm{Score}_{g,t} = \mathrm{Log}_2 \frac{N \times n_{g,t}}{n_g \times n_t}$$

where N is the number of articles mentioning any gene and any term from an ontology branch (e.g. a disease), $n_{g,t}$ is the number of articles mentioning the gene and the term, n_g is the number of articles mentioning the gene and any term from the branch, and n_t is the number of articles mentioning the term and any gene. The higher the association score the more likely it is to observe the gene when the term is present (and vice versa). An association score of zero means that gene and term occur independently of one another. A negative score signals an under-representation of a pair in the literature.

To identify mutations in the literature, we apply our rule based tagger that finds textual descriptions of mutations as well as representations of the form *wNm*, where *w* represents a wild-type amino acid, *N* its position in the sequence of the protein, and *m* the mutant amino acid. If a mutation and a gene are found together in an abstract, the mutation is associated with the gene if any of its protein sequences contains the wild-type residue at the specified position. This approach has been shown

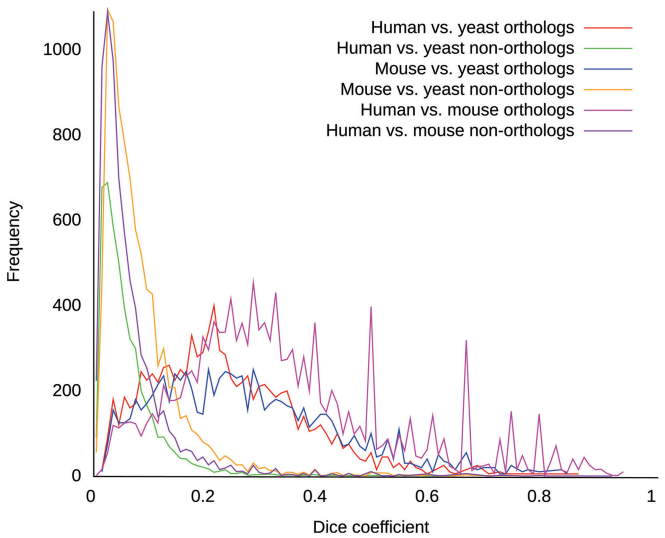

Figure 2. The distribution of Dice coefficients (see text for explanation) from pairwise species comparison. As expected, annotation sets of non-orthologous gene pairs show significantly lower Dice coefficients than from orthologous gene pairs.

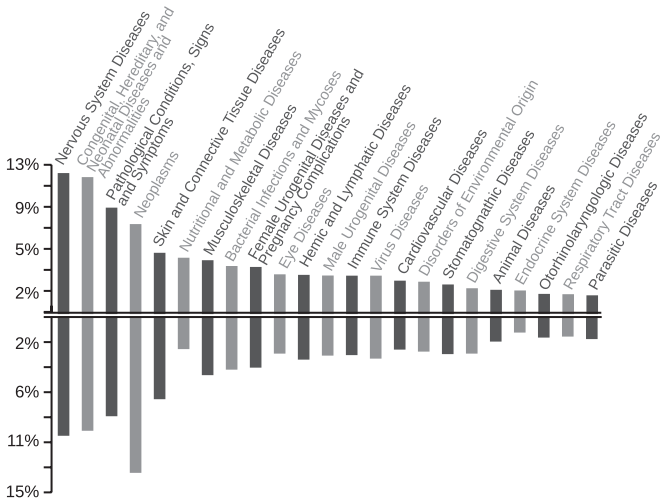

Figure 3. Disease associations mined from literature for all genes (top) compared to disease associations for known cell-cycle genes. Cell-cycle genes are enriched in neoplasms and depleted in nutritional and metabolic diseases.

to successfully predict mutations with implications in the stability of membrane proteins (13).

RESULTS

To measure completeness of text-mined gene annotations only, we compared them with gene-disease pairs contained in the OMIM database of genetic disorders in human and with annotations provided by the GeneOntology annotation project (GOA), and measured recall rates of 83% and 75%, respectively. Since GoGene contains diseases as MeSH concepts, we only considered those diseases in OMIM that are also contained in MeSH. We successfully mapped 358 MeSH concepts to their respective counterparts in OMIM.

Because OMIM and GOA are incomplete, precision cannot be measured reliably. Thus, we followed a different strategy by comparing text-mined annotations between orthologous and non-orthologous genes. We selected all orthologous pairs from the InParanoid database (14) and from NCBI Homologene. Non-orthologous pairs were selected randomly. To measure the similarity between two annotation sets A and B we compute the Dice coefficient as follows:

$$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

Figure 2 shows the distribution of Dice coefficients for gene-pairs from the species human, mouse, and yeast. The overlap between non-orthologous pairs is significantly smaller compared to orthologous pairs with Dice coefficients mostly between 0 and 0.2 (Mann–Whitney Test, $\alpha = 0.01$). As we expected, orthologous genes show a much higher overlap in annotations. In another approach to assess the biological consistency of results in GoGene, we looked at the enrichment in diseases. Figure 3 shows a ranking of text-mined disease categories from MeSH that are most often associated with genes. In comparison, the lower part in Figure 3 shows the same ranking for known cell-cycle genes only. Cell-cycle genes are involved in cell-cycle progression and defects in the cell-cycle are causative for cancers development. Our predictions show enrichment in neoplasms and depletion in nutritional and metabolic diseases.

CONCLUSION

GoGene helps to quickly find interpretations of results from high-throughput experiments together with relevant literature or to simply scan the literature for discussed genes. A sequence query lets users retrieve genes with similar protein sequences and explore their GO and MeSH annotations, for example, to find functional hints for yet uncharacterized genes. We presented four examples that describe typical scenarios for users of the GoGene web server, where the relevant information can be found only a few mouse clicks away.

GO and MeSH are evolving in terms of their vocabulary and structure, which makes it difficult to reflect new changes in a timely manner. We approach this problem by performing updates on a regular basis: once every year for MeSH after each new major release by the National Library of Medicine, and every six months for GO. Automated methods do not reach the high-quality of manual annotation. However, GoGene has a high recall of 75% and orthologous gene pairs can be distinguished from non-orthologous pairs purely based on text-mined annotations. GoGene complements manual annotations by providing more than 4 000 000 associations of genes with GO and MeSH terminology from PubMed, thus supporting interpretation of large gene lists resulting from high-throughput experiments, BLAST or literature searches.

FUNDING

We kindly acknowledge funding by the EU project SEALIFE and the BMBF project ForMaT. Funding to pay the Open Access publication charges for this article was provided by ForMaT.

Conflict of interest statement. None declared.

REFERENCES

1. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
2. Al-Shahrour,F., Minguez,P., Tárraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. and Dopazo,J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34(Web Server Issue)**, W472–W476.
3. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O., Campbell,M.J. *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **31(Database Issue)**, D284–D288.
4. Fernández,J.M., Hoffmann,R. and Valencia,A. (2007) iHOP web services. *Nucleic Acids Res.*, **35(Web Server Issue)**, W21–W26.
5. Couto,F.M., Silva,M.J., Lee,V., Dimmer,E., Camon,E., Apweiler,R., Kirsch,H. and Rebholz-Schumann,D. (2006) GOAnnotator: linking protein GO annotations to evidence text. *J. Biomed. Discov. Collab.*, **1**, 19.
6. Rebholz-Schuhmann,D., Kirsch,H., Arregui,M., Gaudan,S., Riethoven,M. and Stoehr,P. (2007) EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e44.
7. Cheng,D., Knox,C., Young,N., Stothard,P., Damaraju,S. and Wishart,D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36(Web Server Issue)**, W399–W405.
8. Patient,S., Wieser,D., Kleen,M., Kretschmann,E., Martin,M.J. and Apweiler,R. (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics*, **24**, 1321–1322.
9. Grützmann,R., Boriss,H., Ammerpohl,O., Lüttges,J., Kalthoff,H., Schackert,H.K., Klöppel,G., Saeger,H.D. and Pilarsky,C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, **24**, 5079–5088.
10. Hakenberg,J., Plake,C., Royer,L., Strobelt,H., Leser,U. and Schroeder,M. (2008) Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol.*, **S2**, S14.
11. Hakenberg,J., Plake,C., Leaman,R., Schroeder,M. and Gonzalez,G. (2008) Inter-species normalisation of gene mentions with GNAT. *Bioinformatics*, **24**, i126–i132.
12. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33(Web Server Issue)**, W783–W786.
13. Winnenburg,R., Plake,C. and Schroeder,M. (2008) Mutation tagging with gene identifiers applied to membrane protein stability prediction. In *Proceedings of the ECCB 2008 Workshop: Annotation, Interpretation and Management of Mutations (AIMM)*, Cagliari, Sardinia, Italy (September, June 22, 2008).
14. Berglund,A.-C., Sjölund,E., Ostlund,G. and Sonnhammer,E.L.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36(Database Issue)**, D263–D266.