BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 8 2011, pages 1187–1189
doi:10.1093/bioinformatics/btr073

Databases and ontologies

Advance Access publication February 23, 2011

Gene List significance at-a-glance with GeneValorization

Bryan Brancotte[1], Anne Biton[2,3,4,5], Isabelle Bernard-Pierrot[2,3], François Radvanyi[2,3], Fabien Reyal[2,3,6] and Sarah Cohen-Boulakia[1,*]

[1]Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud, F-91405 Orsay Cedex, [2]CNRS, UMR 144, Institut Curie, 26 rue d'Ulm, F-75248 Paris Cedex 05, [3]Institut Curie, Centre de Recherche, Paris, F-75248, [4]INSERM, U900, Paris, F-75248, [5]Mines ParisTech, Fontainebleau, F-77300 and [6]Institut Curie, Departement de Chirurgie, 6 rue d'Ulm, F-75005 Paris, France

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: High-throughput technologies provide fundamental informations concerning thousands of genes. Many of the current research laboratories daily use one or more of these technologies and end-up with lists of genes. Assessing the originality of the results obtained includes being aware of the number of publications available concerning individual or multiple genes and accessing information about these publications. Faced with the exponential growth of publications avaliable and number of genes involved in a study, this task is becoming particularly difficult to achieve.

Results: We introduce GeneValorization, a web-based tool that gives a clear and handful overview of the bibliography available corresponding to the user input formed by (i) a gene list (expressed by gene names or ids from EntrezGene) and (ii) a context of study (expressed by keywords). From this input, GeneValorization provides a matrix containing the number of publications with co-occurrences of gene names and keywords. Graphics are automatically generated to assess the relative importance of genes within various contexts. Links to publications and other databases offering information on genes and keywords are also available. To illustrate how helpful GeneValorization is, we will consider the gene list of the OncotypeDX prognostic marker test.

Availability: http://bioguide-project.net/gv

Contact: cohen@lri.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 15, 2010; revised on October 12, 2010; accepted on February 4, 2011

1 INTRODUCTION

High throughput technologies (e.g. comparative pan-genomic hybridization, gene expression, protein and methylation arrays, high-throughput DNA sequencing) are major, promising and very exciting tools to study biology. Each of them provide fundamental informations concerning thousands of genes such as their normal functions or specific alterations (e.g. DNA copy number alteration, loss of heterozygosity, change in expression, promoter methylation, mutation or post-translational modification).

Many of the current biological research laboratories daily use one or more of these technologies. Selecting genes of interest to design further experiments is of paramount importance. In this process, scientists need to access the latest publications concerning individual or multiple genes. Among the genes they may consider, researchers need to distinguish three categories of genes: (i) genes already clearly shown to be associated with the process they are studying; (ii) new and promising genes for a particular research field whose interest is well known in other research fields; (iii) genes that have not been studied. This task is becoming particularly difficult to achieve faced with the many gene lists which are retrieved and the exponential growth of publications available.

In this article, we introduce a web-based tool named GeneValorization. Given a list of gene names and a set of keywords describing the context of the study, GeneValorization provides a matrix with the number of publications cociting each gene name and keyword. GeneValorization thus gives very quickly a clear and handful overview of the bibliography corresponding to a gene list of interest within a given context of study. To illustrate how helpful GeneValorization is, we consider here the gene list of the OncotypeDX prognostic marker test (Paik *et al*., 2004), which is composed of 16 genes and used to determine the individual relapse risk of a breast cancer patient.

2 MAIN FUNCTIONALITIES

The main interface of GeneValorization is provided on Figure 1.

Basic queries: GeneValorization takes as input from the user a list of gene names and a list of keywords that we call filters. Filters are used to describe the context of the study: while the main filter represents the main context (e.g. Breast cancer), secondary filters are alternative ways of refining this context (e.g. Proliferation, Migration). Secondary filters can be as numerous as necessary. Given this input, GeneValorization provides a matrix of data where each line is dedicated to a gene and each column is dedicated to a filter. The first column considers the main filter only while the next columns consider both the main filter and secondary filters. More precisely, cells (x,y) of the first column of the matrix contain the numbers of publications cociting the main filter (on column y) and the gene name (on line x). Cells from the second column and the next ones contain the number of publications cociting the main filter and the secondary filter (on column y) and the gene name (on line x). Columns where a secondary filter is provided thus allow to subdivide the set of papers considered. In Figure 1, GeneValorization reported 5138 papers involving *PGR* in Breast Cancer (main filter). Among them, 658 papers are also related to Proliferation while 234 mention Apoptosis.

*To whom correspondence should be addressed.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

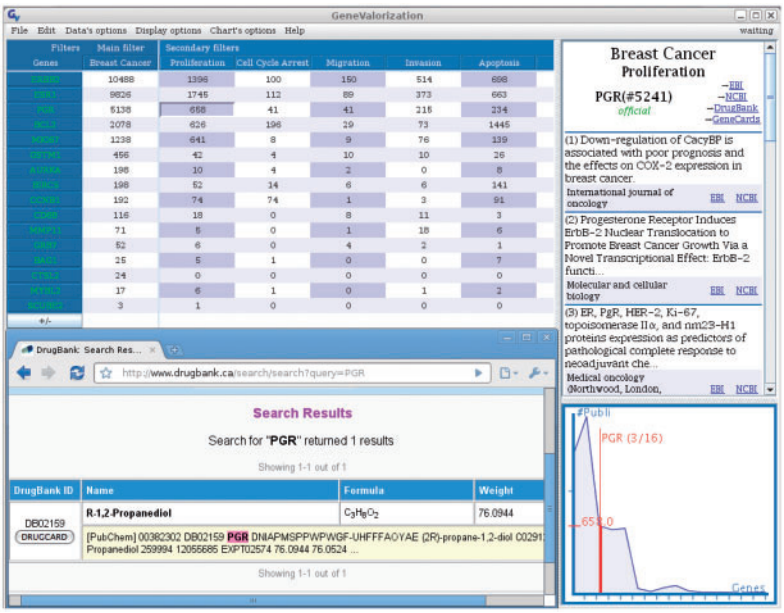

Fig. 1. Main interface of GeneValorization.

Specifying species and using aliases: GeneValorization is able to perform its search not only using the gene name *g* provided by the user but also using all aliases of *g* within a given species, by exploiting information from EntrezGene. The default species used is Human, which can be changed by the user. It is also possible not to consider aliases. All the menus available to make such choices are described in the manual.

Gene name disambiguation: GeneValorization provides assistance in the process of gene name disambiguation. Ids from EntrezGene can be directly entered (prefixed by #) instead of gene names. When users provide a gene name *g*, GeneValorization uses EntrezGene to check which of the three following cases occur: (i) *g* is an official gene name; (ii) *g* is not an official gene name but appears in the list of synonyms of one or several official gene names (within the same species); and (iii) none of the above. In case (i), GeneValorization runs normally, the gene name is green to indicate that it is official. In case (ii), the gene name appears in orange and by right clicking on the name, the user can access the list of official gene names having *g* in their aliases. Users are provided with links to EntrezGene web pages describing each alternative so that they can then choose one of them to remove any ambiguity (the orange gene name can thus be renamed using one of the alternative official names). In case (iii), the gene name appears in red to indicate that it is not an existing gene name.

By default, GeneValorization will still perform all searches using the value entered by the user. In case (ii), and if the user allows synonyms to be considered, GeneValorization will choose to consider the first official gene name provided by EntrezGene.

Visualizing results: clicking on a cell of the main grid allows users to measure the relative importance of genes and keywords. As an example, clicking on the cell corresponding to *PGR* and Proliferation generates the graph of Figure 1 (right-hand side) and states that *PGR* is the third most important gene (according to the number of publications) for the Breast Cancer and Proliferation topics, out of the 16 genes considered here. All the papers associated with the keywords are provided and can be accessed. Interestingly, it is also possible to compare the role of several secondary filters (several graphs can be displayed at the same time).

Accessing information from other sources: links to several databases are also provided. As for genes, information from PubMed, EntrezGene, GeneCards and DrugBank can be obtained (e.g. Fig. 1 shows the DrugBank web page for PGR). GeneValorization also makes calls to the NLM MeSH browser (http://www.nlm.nih.gov/mesh/meshhome.html) to match the keywords provided by the user with MeSH terms and places them within the MeSH ontology.

Data import/export: GeneValorization allows users to load and save gene lists and filters using various formats (e.g. text files, csv, xml). During the export, EntrezGene ids used to search for the aliases are added to the saved matrix.

Advanced queries: Users can express advanced queries involving wilcards '*' and 'AND' in gene names or filters. Considering gene names with or without aliases (from EntrezGene) is possible. Users may also indicate which part of the PubMed file should be queried (e.g. abstract, title).

3 TECHNICAL INFORMATION

GeneValorization is a Java webstart application. It is thus multiplatform and can be used without any specific installation. GeneValorization follows an on-the-fly querying process: queries are directly sent to portals (Entrez or SRS), and no warehousing is needed. Loading a cell information may take 1–4 seconds. It took 10 minutes (but each result is displayed as soon as it is available) to load the entire information associated to the gene list, considering all the aliases available in EntrezGene, of the OncotypeDX prognostic marker test, with the 19 secondary filters. To deal with long list of genes or filters, a caching system has been implemented to optimize the response time. It makes it possible to save previous loaded data

which can be updated on demand later on. Last, GeneValorization is currently able to run on the Entrez NCBI portal (http://www.ncbi.nlm.nih.gov/Entrez/) or the EBI SRS server (http://srs.ebi.ac.uk/); queries will be sent and interpreted by the respective portals. Among the differences, MeSH terms are automatically considered during this search when Entrez is used while it is a pure cooccurrence-based search in SRS (more information is available in the Supplementary Material).

4 PROOF OF CONCEPT

The OncotypeDX test quantifies the probability of distant recurrence in patients with node negative, estrogen receptor positive breast carcinoma. It is composed of 16 cancer-related genes, selected after a validation step on independent studies. This test has been included in the guidelines of the American Society of Clinical Oncology and the National Comprehensive Cancer Network.

We have used GeneValorization to analyze the literature corresponding to these 16 genes. Two queries have been uploaded with 'Cancer' and 'Breast Cancer' as main filter. The 19 secondary filters have been considered corresponding to examples of the most general items to depict the composition of a cancer gene list.

As a first result, GeneValorization allowed us to know that each of the genes has been refered in association with the term 'Cancer' and 'Breast Cancer' in 5 to 17 517 and 3 to 10 048 publications.

Second, GeneValorization made it possible to distinguish three sets of genes in the OncotypeDX test. Five genes (*ERBB2*, *ESR1*, *PGR*, *BCL2*, *MKI67*) were highly studied (>1000 publications associated) while four genes (*BAG1*, *CTSL2*, *MYBL2*, *SCUBE2*) were not very actively studied (less than 50 publications) and the seven remaining genes (*GSTM1*, *AURKA*, *BIRC5*, *CCNB1*, *CD68*, *MMP11*, *GRB7*) had an intermediate level. All the genes of the list are thus not equally known to be involved in breast cancer. Genes associated to only a few or no publications are then of particular interest since they have been selected to be in the signature while not being studied extensively. Our study suggests to conduct new experiments on some of these genes to better demonstrate how connected to breast cancer they may be.

Third, GeneValorization underlined that the secondary filters 'Proliferation', 'Apoptosis', 'Invasion', and 'Angiogenesis' were strongly associated with the gene list while 'Immunity', 'Cell-cycle arrest', 'Epigenetic' or 'microRNA' were much less often associated. This underlines the fact that the OncotypeDX signature is a publication-based molecular signature which contains genes involved in processes commonly known to be linked to breast cancer prognosis but does not contain genes known to be related to new trails in breast cancer studies (as the role of immune response in cancerogenesis).

5 DISCUSSION

Mining PubMed abstracts and ranking publications have been of particular interest in the last years. Approaches are mostly based on Text-Mining techniques [see for instance, Vellay *et al*. (2009) or Krallinger *et al*. (2008) and the references therein]. When available, softwares able to analyze genes such as PDQWizard (Grimes *et al*., 2006), GoGene (Plake *et al*., 2009) and CoPubMapper (Alako *et al*., 2005) differ from GeneValorization in several aspects. From a technical perspective, (i) they make use of data warehouses to store publications, which poses the major problem of updating the local databases and makes it impossible to benefit from all of new publications daily available and (ii) they may not consider simultaneaous requests which makes the response time too long. From a functionality perspective, (i) they may not be flexible, e.g. providing only predefined lists of keywords and (ii) they may not consider gene aliases. From a user perspective, they may not provide results in a concise and/or graphical manner and may not allow users to easily store and load their results at any time. A more complete related work (eight tools compared with GeneValorization) is available in the Supplementary Material.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their helpful comments to improve the manuscript.

Funding: The CNRS (Brasero project) (to B.B.) the CNRS, the Institut Curie, the Ligue Nationale Contre le Cancer (associated laboratory), and the Drop-Top FP6 European project (LSHB-CT-2006-037739) (to A.B., I.B.P., F.Ra., and F.Re.); the Institut National du Cancer (to A.B.).

Conflict of Interest: none declared.

REFERENCES

Alako,B.T. *et al*. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.

Grimes,G.R. *et al*. (2006) PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature. *Bioinformatics*, **22**, 2055–2057.

Krallinger,M. *et al*. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Paik,S. *et al*. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.

Plake,C. *et al*. (2009) GoGene: gene annotation in the fast lane. *Nucleic Acids Res.*, **37**, W300–W304.

Vellay,S.G. *et al*. (2009) Interactive text mining with Pipeline Pilot: a bibliographic web-based tool for PubMed. *Infect. Disorders Drug Targets*, **3**, 366–374.