GeneFuncster: A Web Tool for Gene Functional Enrichment Analysis and Visualisation

Asta Laiho[1,⋆,⋆⋆], András Király[2,⋆⋆], and Attila Gyenesei[1]

[1] Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Tykistökatu 6A, 20520 Turku, Finland
[2] University of Pannonia, Department of Process Engineering, P.O. Box 158 Veszprém H-8200, Hungary

Abstract. Many freely available tools exist for analysing functional enrichment among short filtered or long unfiltered gene lists. These analyses are typically performed against either Gene Ontologies (GO) or KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) database. The functionality to carry out these various analyses is currently scattered in different tools, many of which are also often very limited regarding result visualization. GeneFuncster is a tool that can analyse the functional enrichment in both the short filtered gene lists and full unfiltered gene lists towards both GO and KEGG and provide a comprehensive result visualisation for both databases. GeneFuncster is a simple to use publicly available web tool accessible at http://bioinfo.utu.fi/GeneFuncster.

Keywords: functional enrichment analysis, pathway analysis, gene expression analysis.

1 Introduction

The technological advance in the field of biotechnology during the last decade has led to the increasing generation of functional genomics data. Especially the well established DNA microarray technology and more recently developed high-throughput short read sequencing technology are producing large data sets that require automated means for analysing and visualising the results by taking the gene functions into account. As a result, functional enrichment analysis has become a standard part of the analysis of high-throughput genomics data sets and many freely available tools with different approaches have been developed during the recent years (see [4] for a good review). These tools vary for example based on the kind of input and organisms supported, the statistical tests used for carrying out the enrichment tests, the selection of databases available to conduct the enrichment analysis against and the way the results are reported and visualised. Many of the tools provide useful and unique features, and thus the researchers typically need to use several tools in order to gain a more complete view on the biological significance behind the list of genes under inspection. For

⋆ Corresponding author.
⋆⋆ Equal contributors.

D. Gilbert and M. Heiner (Eds.): CMSB 2012, LNCS 7605, pp. 382–385, 2012.

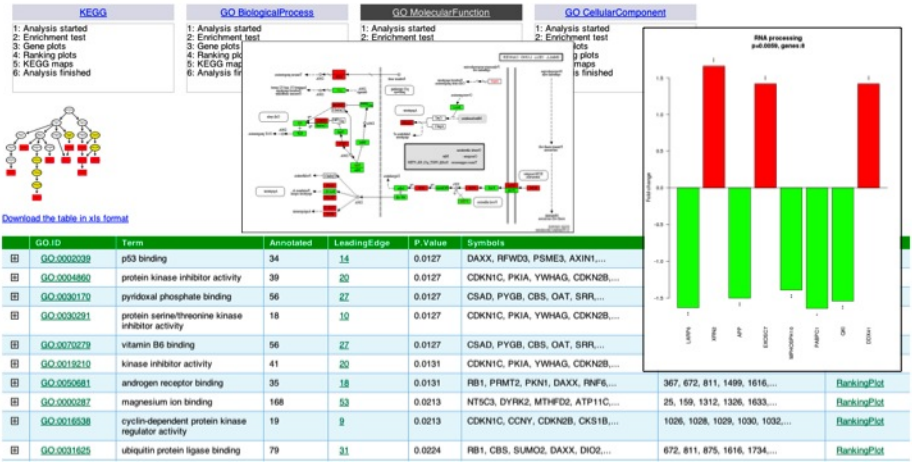
example, many of the available tools take a short filtered list of genes as input to be compared to a given background set of genes and then perform a statistical test (typically based on a hypergeometric or binomial model) to detect whether genes belonging to certain functional categorizations appear in the input gene set more often than would be expected by chance alone.

While analysing the functional enrichment among the filtered genes (e.g. most differentially expressed ones) is very useful, the choice of filtering thresholds can have a significant effect on the analysis outcome. When none or only a few of the many related influenced genes are regulated strongly enough to meet the filtering criterion, some important functional categorizations may be missed. As a solution, other tools take the full unfiltered gene list as input and employ threshold free ranking based approaches applying for example a non-parametric Kolmogorov-Smirnov test. These types of methods are efficient in detecting subtle but consistent changes among genes belonging to the same functional category. Thus the two approaches should be regarded as complementary and optimally applied in parallel to gain a complete view of all affected functional categorizations.

While majority of the available tools simply report the results as a table of category terms ranked according to the test p-value, some tools also provide ways to visualise the results in the context of the functional category information. An informative way to present the GO enrichment results is to provide a view on the directed acyclic GO term graph (DAG) in which each gene product may be annotated to one or more terms. Colouring the terms by the enrichment significance allows an easy detection of the clusters of affected closely related GO terms. Similarly, a good way to visualise the KEGG enrichment results is by presenting the pathway maps where the nodes representing genes or gene complexes are coloured. As many of the described useful features are scattered across various tools, there is a clear need for a combined method. In this article, we present GeneFuncster that is able to analyse functional enrichment in both short filtered gene lists and full unfiltered gene lists as well as represent the results by both GO hierarchical graphs and KEGG pathway maps. If fold-change and/or p-value data is provided by the user, gene level bar plots as well as colouring of the KEGG graph gene nodes according to the direction of the regulation becomes available. These kinds of more advanced and extremely useful visualisations are currently missing from most freely available tools.

2 Methods

GeneFuncster takes advantage of several R/Bioconductor packages including topGO, GOstats and gage [1,3]. GeneFuncster is able to run functional enrichment analyses on both short filtered and full unfiltered gene lists. A traditional over-representation analysis is employed to compare a short filtered gene list to a background provided by the user. Alternatively, the list of all known genes of a specific organism can be used as background. With the full unfiltered gene list enrichment analysis the question of how to rank the genes becomes important.



	GO.ID	Term	Annotated	LeadingEdge	P.Value	Symbols		
⊞	GO:0002039	p53 binding	34	14	0.0127	DAXX, RFWD3, PSME3, AXIN1,...		
⊞	GO:0004860	protein kinase inhibitor activity	39	20	0.0127	CDKN1C, PKIA, YWHAG, CDKN2B,...		
⊞	GO:0030170	pyridoxal phosphate binding	56	27	0.0127	CSAD, PYGB, CBS, OAT, SRR,...		
⊞	GO:0030291	protein serine/threonine kinase inhibitor activity	18	10	0.0127	CDKN1C, PKIA, YWHAG, CDKN2B,...		
⊞	GO:0070279	vitamin B6 binding	56	27	0.0127	CSAD, PYGB, CBS, OAT, SRR,...		
⊞	GO:0019210	kinase inhibitor activity	41	20	0.0131	CDKN1C, PKIA, YWHAG, CDKN2B,...		
⊞	GO:0050681	androgen receptor binding	35	18	0.0131	RB1, PRMT2, PKN1, DAXX, RNF6,...	367, 672, 811, 1499, 1616,...	RankingPlot
⊞	GO:0000287	magnesium ion binding	168	53	0.0213	NT5C3, DYRK2, MTHFD2, ATP11C,...	25, 159, 1312, 1326, 1633,...	RankingPlot
⊞	GO:0016538	cyclin-dependent protein kinase regulator activity	19	9	0.0213	CDKN1C, CCNY, CDKN2B, CKS1B,...	1026, 1028, 1029, 1030, 1032,...	RankingPlot
⊞	GO:0031625	ubiquitin protein ligase binding	79	31	0.0224	RB1, CBS, SUMO2, DAXX, DIO2,...	672, 811, 875, 1616, 1734,...	RankingPlot

Fig. 1. Overview of various result reports generated by GeneFunscter. Detailed descriptions can be found in the tool web site. The most enriched terms/pathways are listed for KEGG and each main GO category. Result visualisations include coloured KEGG pathway maps, GO term graphs and gene-level plots.

In the context of gene expression profiling data, the goal is to rank the genes according to the strength of evidence for differential gene expression between the sample condition groups. Some tools, like GOrilla, take a list of ranked gene symbols as input while others, like GSEA, start from the matrix of normalized gene expression values across all samples and then perform the statistical analysis between the sample condition groups and rank the genes according to the test statistics. In GeneFuncster, the user may provide a ranked list of gene identifiers, or attach fold-changes and/or p-values and then choose to have the gene list ranked according to either of these. As a unique feature of GeneFuncster, the user may also choose to use the so called average rank method for ranking the genes. This method first ranks the genes separately based on fold-changes and p-values, and then calculates the average ranks based on both of them.

Primary input for GeneFunscter consists of a list of Entrez gene identifiers or gene symbols. The list can be pasted directly to the input form or uploaded from a file with optionally included fold-changes and p-values to be used in visualisations and for allowing the ranking of genes. The user can optionally give a background gene list to be used in the filtered list analysis. There are many parameters available for fine tuning the analysis and result visualisation. Several organisms are currently supported and many others can easily be added when requested.

Results are reported on a summary html page in tables of terms sorted according to the term p-values, separately for all analysed main categories. These tables contain links to official term description pages, GO term graph and KEGG pathway maps where the associated genes are coloured and additional gene level

plots taking advantage of the potentially available fold-change and/or p-values for genes. Overview of the various result reports produced with GeneFuncster is shown in Fig. 1.

3 Conclusion

Functional analysis has become a standard tool in elucidating the underlying biology within short unsorted or long sorted lists of genes. Coupled with informative visualisation of the results in a biological context, it has a huge potential in serving the biological research community. GeneFuncster provides functional enrichment analysis with an emphasis especially on the result reporting and visualisation. An earlier version of the tool has been successfully used in several studies including [5].

References

1. Gentleman, R., et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 10(5), R80 (2004)
2. Subramanian, A., et al.: GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 23, 3251–3253 (2007)
3. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008)
4. Huang, D.W., et al.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 1(37), 1–13 (2009)
5. Koh, K.P., et al.: Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell 8(2), 200–213 (2011)