GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information

Ruben Nogales[1,*] (ruben.nogales@fdi.ucm,es), Pedro Carmona-Saez[1,*] (pcarmona@fis.ucm.es), Miguel Vazquez[2] (miguel.vazquez@fdi.ucm.es), Cesar Vicente[1] (cesar.vicente@fis.ucm.es), Xiaoyuan Yang (), Francisco Tirado[1] (ptirado@dacya.ucm.es), Jose M. Carazo[3] (carazo@cnb.csic.es) Alberto Pascual-Montano[1] (pascual@fis.ucm.es)

[1] Computer Architecture Department, [2] Software Engineering Department, Complutense University. Madrid. Spain. [3] Biocomputing Unit. National Center for Biotechnology. CNB-CSIC. Madrid. Spain.

* Both authors contributed equally to this work

GeneCodis: http://genecodis.dacya.ucm.es/

Keywords: functional analysis, enrichment analysis, gene ontology

ABSTRACT

GeneCodis is a web server application for functional analysis of gene lists able to integrate annotations from different sources and find modular patterns of interrelated annotations. This tool differs from other approaches in the way the heterogeneous information is integrated to search for annotations that frequently co-occur in a set of genes. This integrative approach has proved to be useful for the functional interpretation of high-throughput experiments and, therefore, a new version of the system has been developed to expand its functionality and scope. This new version GeneCodis not only provides functional information but also regulatory patterns with the potential to integrate both sources of information in the same analysis. In addition, user-defined annotations can also be included. To cover most of the necessities of the research activity, traditional singular enrichment has been implemented and more organisms and gene identifiers are now supported. The application has been also re-engineered to improve performance, accessibility and scalability. In addition to the web browser, GeneCodis now can be accessed through a public web services interface, enabling users to launch jobs from their own scripts and workflows. The application is freely available at http://genecodis.dacya.ucm.es

INTRODUCTION

High-throughput experiments such as DNA microarrays or proteomics techniques have been widely used during the last decade and currently they are standard technologies in many research centers. Although these methodologies generate huge amounts of data, the challenge does not lies in the analysis or data processing, in which significant advances have been done by the bioinformatics community, but rather in interpreting such datasets to get biological knowledge and meaningful information to formulate new hypothesis.

An essential task in this context is to translate gene signatures into information that can assist in the understanding of the biological mechanisms. In the last few years several methods and tools have been developed to interpret large lists of genes or proteins using information available in biological databases. The common idea in most of these methods is to find functional descriptors that are significantly enriched in the gene signature with respect to the entire genome or other reference list. Annotations from different resources such as Gene Ontology (GO) (Ashburner, Ball et al. 2000) or KEGG (Kanehisa, Araki et al. 2008) are commonly used in this context.

The first type of methods that emerged in this field were focused on evaluating the frequency of individual annotations and apply an statistical test to determine what annotations are significantly enriched in a input list with respect to a reference list, usually the whole genome or all genes in the

microarray. Several tools were developed following this idea, and although each application introduces variations such as different statistical tests, sources of annotations or supported organisms, they all performed the same type analysis and offer slightly variations in the results. Good reviews of such methods can be found in (Khatri and Draghici 2005; Dopazo 2006). A fresh line of research appeared with the observation that the use of thresholds to select the significant genes could lead to underestimate the effect of significant biological effects during the functional analysis. This derived in a new and different analytical concept in which the distribution of annotations is evaluated in the whole list of genes, sorted by their correlation with the phenotype. Different methods followed this approach after the pioneer work of Tamayo and Subramayian published in 2007 (Subramanian, Kuehn et al. 2007), the gene set enrichment analysis (GSEA). However, both standard enrichment tools and GSEA methods evaluate each annotation independently from the others without taking into account the potential relationships between them. Nevertheless, most of the annotations in biological databases are interconnected because they are associated to common genes. Patterns that contain these relationships among annotations can provide invaluable information and extend our understanding of biologic events associated to the experimental system. It is therefore highly desirable to incorporate these relationships among annotations in the functional analysis of gene lists and there are new tools that attempt to extract this type of information (see a review in (Huang da, Sherman et al. 2009))

In 2007 we introduced GeneCodis (Carmona-Saez, Chagoyen et al. 2007), a tool for modular enrichment analysis oriented to integrate information from different sources and find enriched combinations of annotations in large lists of genes or proteins. This approach represented a step forward in the functional enrichment analysis because of its capacity in integrating heterogeneous annotations and discovering significant combinations among them. Since its original publication, this tool has achieved more that 25,000 jobs submissions from all over the world and has been referenced in different works. Even a mirror has also been recently set up at the Center for Bioinformatics in Peking University to facilitate the access in the Asian region. In this work we present a new version of the software with an improved functionality, performance, accessibility and extended scope. Firstly, we have expanded the type of information that can be analyzed by the application incorporating new types of annotations such as microRNAs and transcription factors. In this way, the new version of GeneCodis offers the possibility to mine not only functional information but also regulatory patterns with the potential to integrate both sources of annotations in the same analysis. Moreover, the application now allows researchers to submit their own annotations and perform a joint analysis with the rest of annotations provided in the application. As in the first version, this release of the software finds concurrent annotations enriched in the input list, but we have also included the analysis of individual annotations (singular enrichment analysis). With this new feature GeneCodis covers more possible analytical scenarios than in the first version. We have also increased the type of gene identifiers that are supported including commercial platforms such as Affymetrix Probe Ids and new organisms.

From a technical point of view, GeneCodis has been completely reengineering making it faster, flexible and efficient. The algorithm to retrieve sets of concurrent annotations has been improved and implemented to run in a multi-grid environment that is more suitable to handle large number of jobs simultaneously, improving in this way the performance and the throughput of the system. The application now can be accessed in a programmatic way by Web Services. This allows researchers to include its functionalities in data analysis pipelines. Finally, the new friendly interface is designed to facilitate the use of the tool and now results include more useful information through different graphs and file formats.

FEATURES AND FUNCTIONALITY

GeneCodis has gone through a lot of changes since its first publication, but the tool works in a similar way. In this web-based application the users have to select the organism and the biological annotations that will be considered in the analysis and then load the gene list to be analyzed. The whole genome is used by default as a reference set but users can also load their own reference list. No more information or user's parameters are mandatory at this point. As advance options, it is possible to select the preferred statistical test among three possible alternatives: hypergeometric test (default), chi-square test or both. Annotations will be considered in the concurrent analysis only if they appear in at least a minimum number of genes. This is known as minimum support and it is set to three by default. And finally users can omit or select the multiple hypothesis correction method between two alternatives: false discovery rate method (default) or permutation-based correction. In addition, a new feature of this version allows users to submit a file with a list of user-defined annotations that can be considered in the analysis together with the rest of selected annotations.

Functional Analysis In GeneCodis

As we have commented previously, the enrichment analysis of individual annotations was the first method introduced for functional analysis of large lists of genes or proteins, being also the most popular one. There is a large collection of tools that implements this type of enrichment analysis, most of them focused on the analysis of Gene Ontology annotations. The arrival of GSEA methods turned the enrichment analysis of individual annotations from a gene-centered to a gene-set based analysis. These methods, although extremely useful for the interpretation of gene lists, do not exploit the inter-relationship that exists among gene annotations. In this context, tools such as GeneCodis provide a new way to analyze functional information by taking into account the relationships among annotations associated to common genes in the list. This analysis offers different advantages with respect to singular enrichment methods. Joint terms may contain unique biological meaning for a given study, not held by individual terms. For example the combination of two terms such as *Apoptosis* and *Mitochondria* may be enriched in a gene list while the individual annotations are not significant if evaluated independently in the same list of genes. This may indicates that mitochondria apoptotic pathways can be more significant in the system than other apoptotic processes. But probably more interesting is the possibility to integrate and jointly analyze information that covers different aspects of the biology of the genes.

GeneCodis is one of the few tools that offer these options for the functional analysis of gene lists by implementing an algorithm based on the extraction of frequent itemsets (see (Carmona-Saez, Chagoyen et al. 2006) for details). Nevertheless, one of the weak points of this type of analysis is that it can increase the complexity of the results because redundant patterns may be generated. Therefore, it would be beneficial for researchers to have a unified system that integrates both type of analysis, the evaluation of individual annotations and co-annotations for a better exploration of the high throughput experimental results.

In this new version of GeneCodis we have included the singular enrichment analysis precisely to take into account these two types of analysis. As results, besides the combinations of annotations enriched in the list of genes, in this new version GeneCodis also provides the analysis of individual annotations, showing those from the selected categories that are enriched in the input list. For example, if two categories of annotations are selected for the analysis, the results will include three different lists: one with the concurrent analysis results and two with singular enrichment information of each one of the types of annotations. Moreover, these results are provided in different formats: html tables, tabulated text files, xml files, pie charts and bar graphs.

New Sources of Annotations: Integrating Functional and Regulatory Information

Among the multiple resources of information, Gene Ontology is by far the most popular one for functional analysis of gene lists. This is reasonable due to the rich content of GO in terms that describe the functional role of genes at a molecular level and the initiative of different consortiums to annotate complete genomes with GO terms. Earlier enrichment tools were mainly based on the analysis of GO terms although annotations from other sources such as KEGG or Biocarta (www.biocarta.org) were being incorporated in more recent applications. Although GO covers three aspects of the biology of genes; Biological Process, Cellular Component and Molecular Function, the former is the information in which researchers have been mainly interested for the functional characterization of gene lists. This is in part because biological process annotations provide explicit information to interpret the biological mechanisms that may be associated to the experimental system. Indeed, pathway information contained in KEGG or Biocarta covers similar aspects of the biology of genes than GO Biological Processes annotations.

Nevertheless, beyond functional information there are other properties of genes and proteins that can be also very useful to interpret biological systems, such as information related to transcriptional mechanisms. There are different sources of regulatory information that have been incorporated to enrichment tools (Dennis, Sherman et al. 2003; Al-Shahrour, Minguez et al. 2006; Abascal, Carmona-Saez et al. 2008, Guruceaga, Segura et al. 2009)). In GeneCodis we have included annotations related to transcriptional information from different sources. miRBase (Griffiths-Jones, Saini et al. 2008) contains putative targets of microRNAs, molecules that in the last few years have been shown as key regulators in many biological systems. From this database we have extracted the microRNAs associated to genes in different organisms; *B. taurus, C. elegans, D. rerio, D. melanogaster, G. gallus, H. sapiens, M. musculus* and *R. norvegicus*.

Using information from TransFac we have annotated genes from *H. sapiens, M. musculus* and *R. norvegicus* with transcription factors that have biding motifs in their promoter sequences. In addition, we have used results from chip-on-chip experiments (Carmona-Saez, Chagoyen et al. 2006) to annotate genes in *S. cerevisiae* with transcription factors that bind to their promoter regions. This new annotations allow users to perform enrichment analysis with regulatory information in addition to functional information.

Nevertheless, beyond the independent analysis of different properties of genes the integration of heterogeneous sources of information can provide a more complete picture of the analyzed system. In this sense, the new sources of information offer the possibility to mine gene lists to discover associations among regulatory and functional information. This turns the potential of GeneCodis to integrate different kinds of annotations in one of its most useful features. This is briefly illustrated in Figure 1 with a very simple example. It shows a screenshot of the GeneCodis results from the analysis of the ribosomal gene set in KEGG. Using also the information of transcription factors in the analysis we can see how co-annotations of functional and regulatory information are evaluated. Indeed, enriched annotations contain transcription factors such as RAP1, FHL1 or SFP1 that are well known to play important roles in the regulation of ribosomal genes.

Organisms and Identifiers Supported

To extend the scope of GeneCodis we have also increased the number of supported organisms and gene identifiers. GeneCodis works with most of the model organisms in biological research. The whole list of organisms supported includes *Arabidopsis Thaliana, Bos Taurus, Caenorhabditis elegans, Candida albicans, Danio rerio, Drosophila melanogaster, Escherichia coli, Gallus gallus, Homo sapiens, Leishmania major, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe* and *Vibrio cholerae*. Alternatively, to facilitate the usability, we have extended the supported gene identifiers to cover the most common ones used in gene microarrays studies, including proprietary Ids of commercial platforms such as Affymetrix, Agilent, Codelink and Illumina. The backbone of the system that is now based on Ensembl gene ids has been extracted through biomart (http://www.biomart.org).

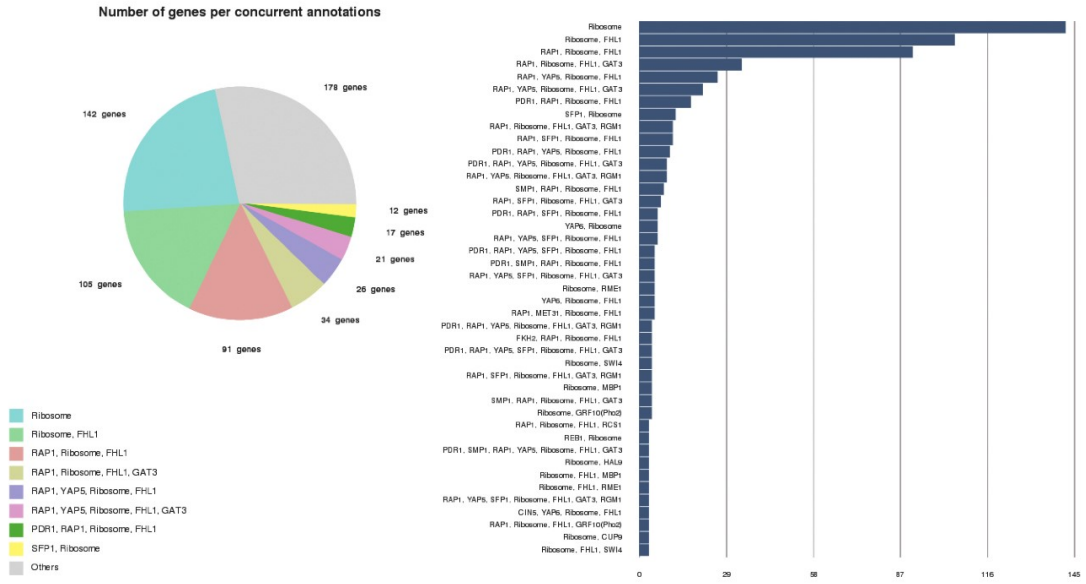

Figure 1. Example of the graphic generated in a typical GeneCodis analysis. Enriched combinations of significant annotations are represented in a pie and bar graphs, where the length of the bars and size of the slices are proportional to the number of genes supporting the significant combination of annotations.

Interface

The design of a web tool is an important part to take into account in its development. The new friendly interface has been designed to facilitate the usability of the tool. Now the application is more fresh and intuitive. Results can be explored more dynamically and are presented in different formats including pie charts and bar graphs images that allow users to explore and interpret the data in a quick look (see figure 1 for an example). XML files with structured results are also provided permitting its use in data process pipelines. Finally, tabulated text files and html tables are also included in the report. These different file formats would help researches to use GeneCodis results in other applications.

In another context, in this new release big efforts have been done to facilitate the access of GeneCodis functionality by different ways. In addition to the classical web-server access, the tool can now be used programmatically trough the Web Services technology. Using this technology, researchers can insert functional analysis in their data mining pipelines or in other bioinformatics systems in a very straightforward manner (see figure 2 for a very simple example). There are many advantages in the use of Web Services, it is a platform-independent and language-independent technology and it is very adequate for loosely coupled systems with different architectures working together. In this way it is guaranteed that there is no any prerequisite imposed for its use. A complete tutorial can be found in the site including example scripts.

```ruby
require 'soap/wsdlDriver'

#### Test arguments ###
#
org = "Sc" #Saccharomyces Cerevisiae
algorithm = 1 # Concurrent analysis
test = 0 # Hypergeometric p values
correction = 1 # FDR method
minsupport = 3 # Minimum support of 3 is considered in the analysis
annotations = ["GO_Biological_Process","GO_Molecular_Function"]
reference_list = [] # whole genome
input_list = ["S000004295", "S000005284", "S000000598", "S000006205",
              "S000006183", "S000004982", "S000005662", "S000001387",
              "S000002555", "S000005668", "S000002585", "S000003736"] #input list


#### Connecting to server ###
#
WSDL_URL = "http://genecodis.dacya.ucm.es/static/wsdl/genecodisWS.wsdl"
driver = SOAP::WSDLDriverFactory.new(WSDL_URL).create_rpc_driver

### Submit tha analysis ###
#
job_id = driver.analyze(org,algorithm,test,correction,minsupport,input_list,annotations,reference_list)

### Check job state ###
#
status = driver.status(job_id)
while status == 1
  puts "Waiting ..."
  sleep 2
  status = driver.status(job_id)
end

# If error
if status < 0
  error_message = driver.info(job_id)
  raise "Finished with error #{ error_message }"
end

# Get results
results = driver.results(job_id)
puts results
```

Figure 2: Example of a ruby client code to invoke the GeneCodis Web Service. The access only needs three main steps: submit the analysis, ask for the job status and get the results.

IMPLEMENTATION

The algorithmic core of GeneCodis has been reviewed and drastically improved. There is a new implementation of the algorithm to extract closed itemsets based on more efficient and faster methods (Zaki and Hsiao 2005). The performance of the system depends on the number of annotations related with the genes in the input list. The new algorithm can deal with large sets of annotations in a faster and more efficient way than the algorithm implemented in the first version of GeneCodis. This is especially evident if multiple annotations are included in the same analysis and when the permutation-based test is used for the multiple hypothesis correction, in which the computing time has decreased drastically with respect to the previous version.

The throughput and performance of the entire system has also been improved by implementing the new algorithm and the Web Service technology in the context of a multi-grid computational environment. The new system is able to handle all submitted jobs simultaneously without queuing any of them for long periods of time. The current implementation takes advantage of one cluster with two Quad-Core Intel Xeon processors of 64 bits and two independent grid infrastructures (CyTED, http://www.cyted.org/, and EELA2, http://www.eu-eela.org/) integrated by the GridWay metascheduler (Huedo, Montero et al. 2005)

When users submit a query through the web site, the system launch one job for the concurrent analysis and one job for each annotation's category selected for the singular enrichment analysis. All these jobs will be executed in parallel. Users can also submit a query directly from a script using the Web Services. In all cases the flow of the system is the same. A meta-scheduler determines, depending on the computational load of the cluster and the grids, in which computational environment the jobs will be executed. When the cluster is free, then the jobs are queued and executed in the cluster. Otherwise, the jobs are sent to the less workloaded grid resource. This approach represents a cost-effective alternative to improve the throughput of the application and to guarantee its real-time performance, which is a critical step in every popular web-server application.

CONCLUSIONS AND DISCUSSION

High-throughput experimental techniques have demonstrated to be very useful allowing the study of biological systems from a global perspective. In many cases, these techniques generate huge amounts of data in the form of large gene or protein. An essential task in this context is to translate these lists to functional information that aids researchers in the interpretation of the underlying biological processes. But this interpretation is not always a trivial step; there is a lot of biological information distributed in different databases that is necessary to extract the full meaning of the data. Methods based on the enrichment analysis have proved to be very useful tools for the analysis and interpretation of such lists of genes.

GeneCodis is a tool designed to expand the enrichment analysis of annotations by adding the possibility of extracting not only individual annotations, but also significant combinations of them. Since its creation, this tool has proved to be a useful resource for the research activity. This encourages us to improve it by adding more functionality that expands its scope, performance and accessibility.

In summary, the new version of GeneCodis finds combinations of annotations but also includes in the output the traditional singular enrichment analysis. New annotations have been included such as microRNAs or transcription factors to extend the functionality of GeneCodis towards the analysis of regulatory information. Moreover, GeneCodis allows researchers to jointly analyze regulatory and functional information and to extract association patterns between both data sources. In addition, users can submit their own annotations that can be considered in the analysis together with the rest of the annotations provided in the application. The algorithm to retrieve sets of concurrent annotations has been improved and implemented to run in a multi-grid environment that is able to handle all submitted jobs simultaneously, improving in this way the performance and the throughput of the system. Finally, GeneCodis now can be accessed in a programmatic way by a web services interface; which allows researchers to include its functionalities in data analysis pipelines.

One of the disadvantages of the modular enrichment analysis used in GeneCodis is the intrinsic redundancy of the combined functional annotations that are usually extracted. This is due to the unavoidable natural redundancy of the information that is already known about genes and proteins and the

nature of the frequent itemsets mining algorithm that extract all possible combinations. Several methods still need to be implemented to filter out intrinsic repetitions and constitute a necessary future functionality. Going in the direction of creating a more self-contained application, new future releases of GeneCodis will also include Gene Set Enrichment Analysis methods to allow users the selection of all possible flavours of functional analysis in the same environment.

The new version has been running since August 2008. Extensive tests have been carried out using synthetic and real datasets for which the outcome of the software is known. The diverse functionalities supported by this tool have been fully tested by real users who have provided feedback on issues that have helped in improving the application. We hope the renewed GeneCodis will be of interest to the scientific community.

GENECODIS AVAILABILITY

This application can be freely accessed through its main site at http://genecodis.dacya.ucm.es. A mirror in the Asian region has also been recently set up at the Center for Bioinformatics in Peking University. The mirror is available at http://genecodis.cbi.pku.edu.cn/.

ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, CYTED-505PI0058, TIN2005-5619 and PS-010000-2008-1. This work also makes use of results produced by the EELA-2 project (www.eu-eela.eu), co-funded by the European Commission within its Seventh Framework Programme". We would like to acknowledge Luis Canet for his technical help. Special thanks also to Prof. Jinchu Luo from the Center for Bioinformatics at Peking University for his help in setting the Asian GeneCodis mirror. A.P.M. acknowledges the support of the Spanish Ramón y Cajal program.

REFERENCES

Abascal, F., P. Carmona-Saez, et al. (2008). "ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data." Bioinformatics **24**(9): 1208-9.

Al-Shahrour, F., P. Minguez, et al. (2006). "BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments." Nucleic Acids Res **34**(Web Server issue): W472-6.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Carmona-Saez, P., M. Chagoyen, et al. (2006). "Integrated analysis of gene expression by Association Rules Discovery." BMC Bioinformatics **7**: 54.

Carmona-Saez, P., M. Chagoyen, et al. (2007). "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists." Genome Biol **8**(1): R3.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Dopazo, J. (2006). "Functional interpretation of microarray experiments." OMICS **10**(3): 398-410.

Griffiths-Jones, S., H. K. Saini, et al. (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-8.

Guruceaga, E., V. Segura, et al. (2009). "FactorY, a bioinformatic resource for genome-wide promoter analysis." Computers in Biology and Medicine (in press).

Huang da, W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res **37**(1): 1-13.

Huedo, E., R. S. Montero, et al. (2005). "The GridWay Framework for Adaptive Scheduling and Execution on Grids." Scalable Computing - Practice and Experience **6**(3): 8.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-4.

Khatri, P. and S. Draghici (2005). "Ontological analysis of gene expression data: current tools, limitations, and open problems." Bioinformatics **21**(18): 3587-95.

Subramanian, A., H. Kuehn, et al. (2007). "GSEA-P: a desktop application for Gene Set Enrichment Analysis." Bioinformatics **23**(23): 3251-3.

Zaki, M. J. and C.-J. Hsiao (2005). "Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure." IEEE Transaction on Knowledge and Data Engineering **17**(4): 12.