

# Gene Analogous Finder: a GRID solution to find functionally analogous gene products

Angelica Tulipano, Giacinto Donvito, Flavio Licciulli, Giulia De Sario, Giorgio Maggi, Andreas Gisel



[www.eu-egee.org](http://www.eu-egee.org)



<http://grid-it.cnaf.infn.it/>



# Functionally analogous gene products

We developed a project for finding, within the same or different species, functional analogous gene products, that is the gene products with similar functions but not necessarily similar sequences.

Usually researchers compare genes by sequence similarity, but similar function does not always mean similar sequence:

to find functional analogies between gene products it is necessary to compare them according to the information of their function within the gene description.

Gene Ontology (GO) offers a controlled vocabulary for the description of the gene products: the molecular functions they have, the biological processes they are involved in, the cellular components they are associated to.

i2

We developed a project to compare gene products within the same or different organisms on the level of their functions. Normally researchers use sequence similarity to find similar gene products, but similar function does not always mean similar functions. We suggested to use the description of the gene products and their functions to make this comparison. Gene Ontology offers us a controlled vocabulary of descriptive terms for the description of the gene products: the molecular function they have, the biological processes they are involved in, the cellular components they are associated to.

itb; 18/04/2007



# Gene Ontology

**GO** is an international standard to annotate genes:

- is structured as a directed acyclic graph with three independent branches with top-level terms 'molecular function', 'biological process' and 'cellular component'
- the descriptive terms (*GO* terms) are nodes in the graph.
- data are available in a public database ([www.godatabase.org/dev](http://www.godatabase.org/dev))
- more than 1.700.000 gene products are described by the *GO* terms associated
- more than 20000 *GO* terms ending up with >7.000.000 associations

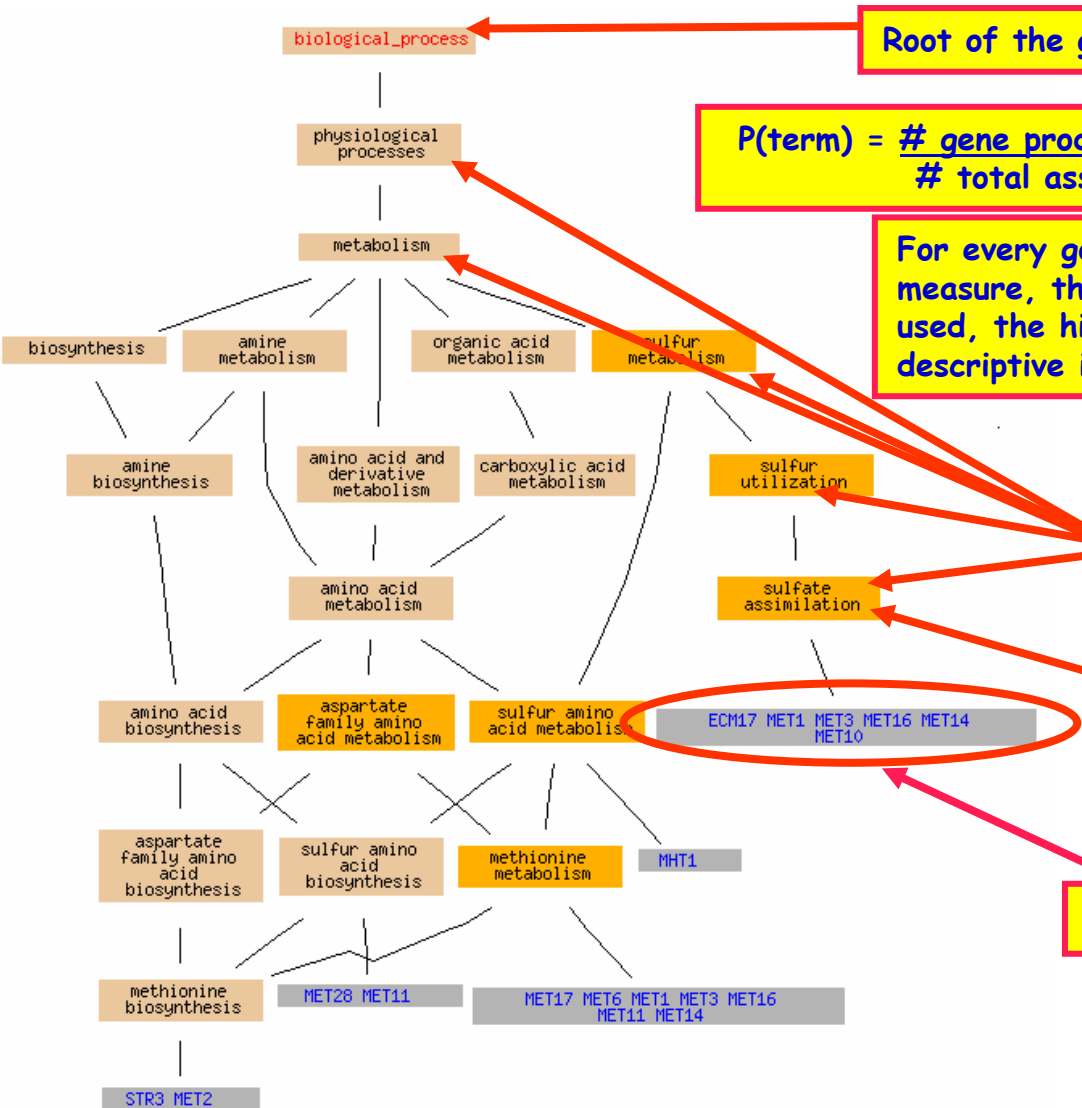
The consortium produces an ongoing effort to find new associations, improving the existing descriptions and creating new ones.





# Graph of Gene Ontology associations

BioinfoGRID



Root of the graph

$$P(\text{term}) = \frac{\# \text{ gene products associated to the term or any of its children}}{\# \text{ total associations between all GO terms and gene products}}$$

For every go term we calculated its semantic similarity measure, the p(term): the less a term of a vocabulary is used, the higher the information content and the more descriptive is.

GO terms of the path

GO term directly associated

Gene products annotated

i4

here it is showed the graph of some GO associations: this is one of the three branches of the GO, starting with the top level term 'biological process'. All the go paths of the descriptive go terms, which are the nodes of the graph, are linked to the top term. In the blue boxes there are the names of the gene products which are annotated with the directly associated go terms but also with all the related path of go terms indirectly associated up to the root.

Of course not all the terms are equally descriptive: moving along the path of a specific GO term away from the root, the detail of the information content of a given term increases from node to node. So we considered the 'rule' that the less a go term is used in the associations, the higher the information content and the more descriptive is. For this we weighted each term with a  $p(\text{term})$ , which indicates its semantic similarity measure: we calculated the  $p(\text{term})$  for every term by counting the number of the gene products associates to a term or any of its children, divided by the number of total associations between the GO terms and gene products.

itb; 18/04/2007



# Algorithm of the search

Through a  $\chi^2$  statistical test we compare gene product A and gene product B:

- we count the number of the GO terms directly or indirectly associated which are common and uncommon to two genes;
- we weight each term with  $1-p(\text{term})$ , giving more importance to specific terms.

	# go terms in A	# go terms not in A
# go terms in B	$O_{11}$	$O_{12}$
# go terms not in B	$O_{21}$	$O_{22}$

Table of the observed frequencies

The higher the  $\chi^2$  value is, the bigger the probability of functional dependence between the two gene products A and B is.

The algorithm of the statistical comparison was implemented in a perl script.

## Problem:

The comparison of all the gene products annotated is very data-intensive (>1.000.000 gene products) and time-consuming (a single comparison occupies one CPU for 30-45 min, the whole search ~55 CPU years!)

- i5      In the algorithm for the comparison we use a statistical chi-square test which tells us how probable is that two samples are independent or not. In this case we compare two gene products at a time, counting the go terms, directly or indirectly associated to each one which are in common, which are not in common being present in one or in the other, and all the go terms that are not present in both. To these frequencies we applied a chi square test: the higher the value of the chi-square, the higher the probability that the two gene products are not independent on the level of their functional description, that is to say they are probably functional analogues. The algorithm has been implemented in a perl script.
- Our aim was to find functional analogous gene products by comparing all the gene products annotated with the Gene Ontology. This is of course a very data-intensive process (there are more than 1 million gene products to compare each other) and time-consuming operation (a single comparison occupies one CPU for 30-45 min, and the whole search would take 55 years!)

itb; 18/04/2007



# Approach

Run the search on the INFN GRID, splitting it into several smaller independent jobs.

Each job works on a sub-list of the gene products of interest.

The jobs were submitted to the GRID by the User Interface and distributed to the available worker nodes assigned by the Resource Broker.

- Each worker node has its own local source of data
- An output text file with 100 best hits is compiled as an additional DB

i6

The approach we were following to solve these problems was to run the search on the INFN GRID, splitting it into several smaller jobs running in parallel. To perform the analogous gene search, the entire list of gene products to compare was splitted into sub-lists with each job working on its own sub-list. The total list of gene products to examine is stored into a central Mysql db, which keeps track of the completed comparisons of gene products, the failed and the running ones using a Job Application Monitoring Tool. Generally speaking, the speed of the search is limited by the access time to the data source, so what is very important is distribute the data source on each worker node running the job, so that each one has its own local source of data. The jobs were submitted to the GRID by the User Interface by means of a perl script running on it.

itb; 18/04/2007



# GRID (db distribution)

We selected a list of ~ 80000 gene products of 13 different organisms to compare with all the other 1.000.000 gene products

Each job recreated locally, on its own worker node, the entire GO MySQL database and installed the perl libraries: this operation took about 6% of the total execution time.

This search was completed using up to 950 WNs in ~3 days, instead of 5 years!

The set up of the running environment, data base and perl libraries installation, could be very useful for other data-intensive application in bionformatics.

i8

The first search considered a list of 80000 gene products to compare against the other 1 million gene products. Each job recreated on its worker node the entire GO MySQL database to retrieve the needed information for the search. This was an important development provided by this project: the possibility to distribute and temporally install a relational db such mysql on the same wn where the job is running. After having installed the db the perl libraries were installed and set-up on the wn, so that they can be used by the script of the algorithm. This operation of distribution was not very time consuming, it too only 6% of the total execution time. Since in bioinformatics the use of Perl and its modules is widely spread, such a distributed installation can be of high interest also for other application. This search was completed in 3 days using up to 950 worker nodes, instead of 5 years on a single cpu!

itb; 17/04/2007



i7

itb; 17/04/2007

i10

On the left we see a very simple scheme of the data flow in a job run: the list of gene products is splitted into sub-lists which were distributed to the worker nodes, also the mysql db is distributed locally on each worker node as data source. The results of each job are reported as output files. On the right, a scheme of the Grid process. The user interface distributes jobs on every worker node, the submission is performed via the resource broker finding the worker nodes available. The execution on the farms firstly initiates a download of the required data from a storage element (SE) followed by the effective computing. In parallel a job monitoring procedure (JAM) is started to test the success of the executed job. Then the results can be recovered from the farm via the user interface.

itb; 18/04/2007



# GRID distribution (text files distribution)

We compare all gene products (>1000000) against all, running the jobs on the EGEE infrastructure within the VO Biomed.

We downloaded all the needed information in text files and distribute them to the worker nodes:

- the GO terms associated to each described gene are extracted from the GO DB and stored in a text file
- the text file is transferred from one of the available SE's to the WN.

In this search we submitted more than 42000 jobs: the submission uses 3 RB's in a round robin algorithm in order to avoid the overload of a single RB and that the failure of a single RB can stop the submission of jobs

This search was completed in ~30 days, instead of 55 years!!

i9

In a second time we performed all the search comparing 1 million of gene products against all. This time we used as data source text files with all the needed information: the go terms associated to each described gene are extracted from the godb and stored in a text file. This file is transferred from one of the available storage elements distributing them to the worker nodes. The submission uses 3 resource brokers to avoid the overload of a single resource broker or that the failure of one of them can stop the submission of jobs. The whole search was completed in about 1 month, instead of 55 years on a single cpu!

itb; 18/04/2007



# Results

This method finds most of the orthologous gene products and members of the same gene family, but also finds functional analogous gene products not belonging to the same family with low level of sequence similarity but a high number of common GO terms and sharing therefore similar functions.

Example:

BCL2\_HUMAN, a well studied apoptosis gene.

In the list of its 30 best analogous gene products:

- 12 gene products belonging to its same family
- 4 gene products belonging to another apoptosis family with already a lower sequence similarity
- the other 14 hits (45%) are all gene products related to apoptosis which are not in a similar family and have low levels of sequence similarity with BCL2\_HUMAN, but were selected because of their similar description.

i11

We demonstrated that with this method we are able to find most of the gene products orthologous to the gene product of interest which are members of the same family (that is to say with a good level of sequence similarity), but also we find in the list of the best hits gene products not belonging to the same family, therefore with low level of sequence similarity but having a high number of GO terms in common with the gene product of interest and sharing for this similar functions with it. As an example of our search results, I have reported here the list of the 30 gene products which are the most probable functionally analogous to BCL2-hUMAN, a well studied gene involved in the death of a cell (apoptosis). 12 are gene products belonging to its same family, 4 belonging to another apoptosis family with already a lower level of sequence similarity, the other 14 hits are all gene products related to apoptosis which are not in a similar family and have low levels of sequence similarity with bcl2\_human, but were selected because of their similar description.

itb; 18/04/2007



# Benefit for the Biologist

This data set offers to the scientist:

- a list of functional similar gene products over a broad range of well- and non-well known organisms
- an help to understand the functionality and probable proprieties of his gene of interest
- a support for evolutionary studies to understand the strategies of development of the same function in different gene families



# Future plans

- Gene Ontology is continuously improving its associations, using new GO terms and describing new gene products;

- ~ every month an update:

gene products would be more and more accurately described and our method will be more precise.

Now we are working on:

- a new algorithm to create an efficient updating procedure to profit from the new monthly GODB versions and increasing knowledge;
- a MySQL dump for distribution of the analogous gene products with the monthly GODB release.

i12

Gene Ontology is continuously improving its associations and gene products descriptions, using new go terms and also annotating new gene products. Every month there is an update of the GOdb with all this new information. Thanks to these efforts to describe more accurately the gene products, our search will be in time more and more precise. Now we are working on a new algorithm to create an efficient procedure to take advantage from the monthly updated version of the godb. We are working also on the creation of a mysql dump for the distribution of the results of the search of the analogous gene products with the monthly GOdb release.

itb; 18/04/2007



# Acknowledgments

- **Giacinto Donvito<sup>1</sup>, Giorgio Maggi<sup>1</sup>**

**For technical aspects and grid distribution**

- **Andreas Gisel<sup>2</sup>, Giulia De Sario<sup>2</sup>, Flavio Licciulli<sup>2</sup>,  
Angelica Tulipano<sup>1,2</sup>**

**For bioinformatical aspects**

**<sup>1</sup> INFN, Bari**

**<sup>2</sup> CNR-ITB, Bari**