

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

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



www.ba.itb.cnr.it

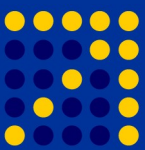


Enabling Grids for
E-science in Europe

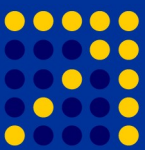
www.eu-egee.org



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



- The functional analogous gene products
- The approach for the Grid environment
- Results
- Future plans



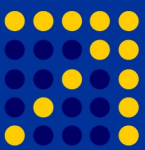
Functionally analogous gene products

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.

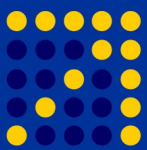


Gene Ontology

GO is an international standard to annotate genes:

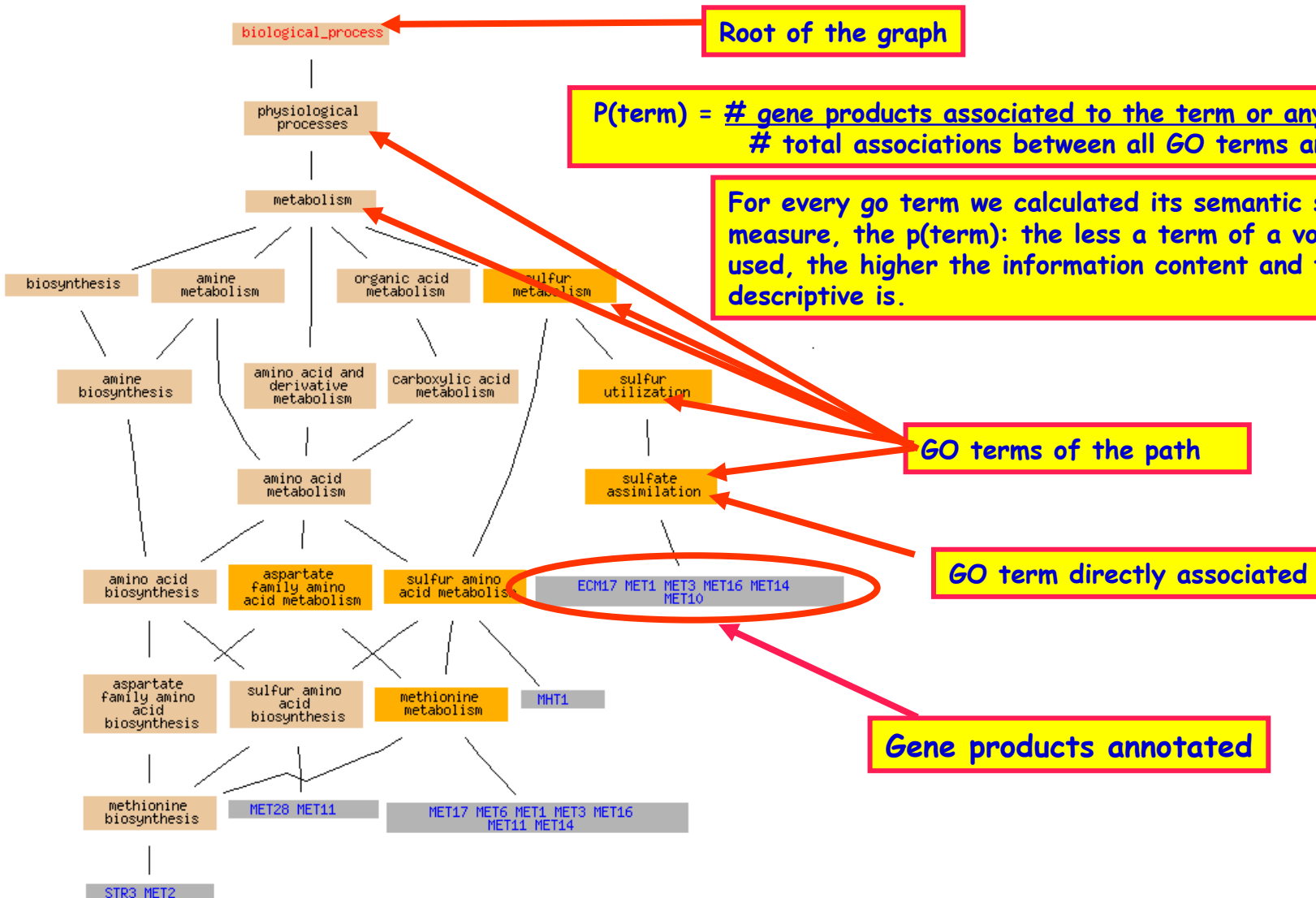
- is structured as a directed acyclic graph with three independent branches with top-level terms 'molecular function', 'biological process' and 'cellular component'
- the descriptive terms (*GO terms*) are nodes in the graph.
- data are available in a public database (www.godatabase.org/dev)
- more than 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

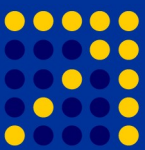
$P(\text{term}) = \frac{\# \text{ gene products associated to the term or any of its children}}{\# \text{ total associations between all GO terms and gene products}}$

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



Algorithm of the search

BioinfoGRID

- Through a χ^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}

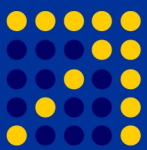
Table of the observed frequencies

The higher the χ^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!)

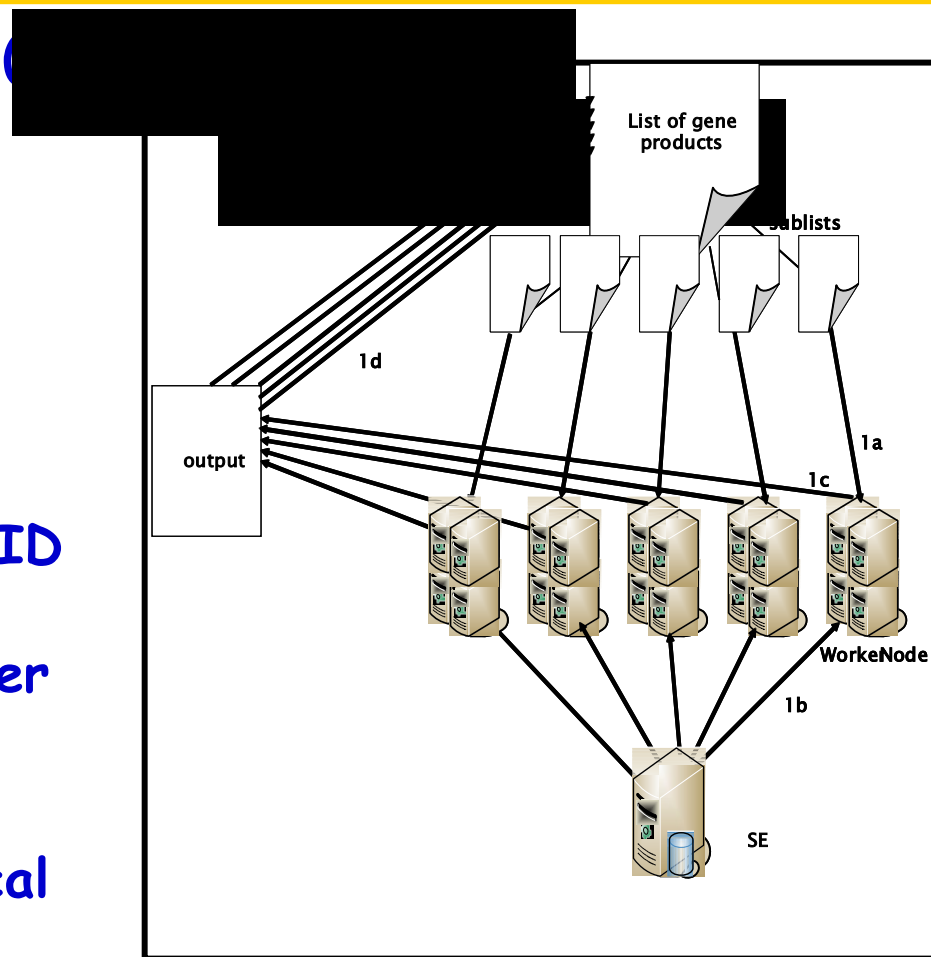


Run the search on the INFN GRID (bio) and EGEE infrastructure (biomed), splitting it into several smaller independent jobs.

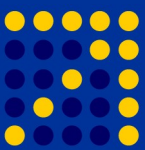
Each job works on a sub-list of the gene products of interest.

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



Scheme of the data flow in a job run



GRID (db distribution)

BioinfoGRID

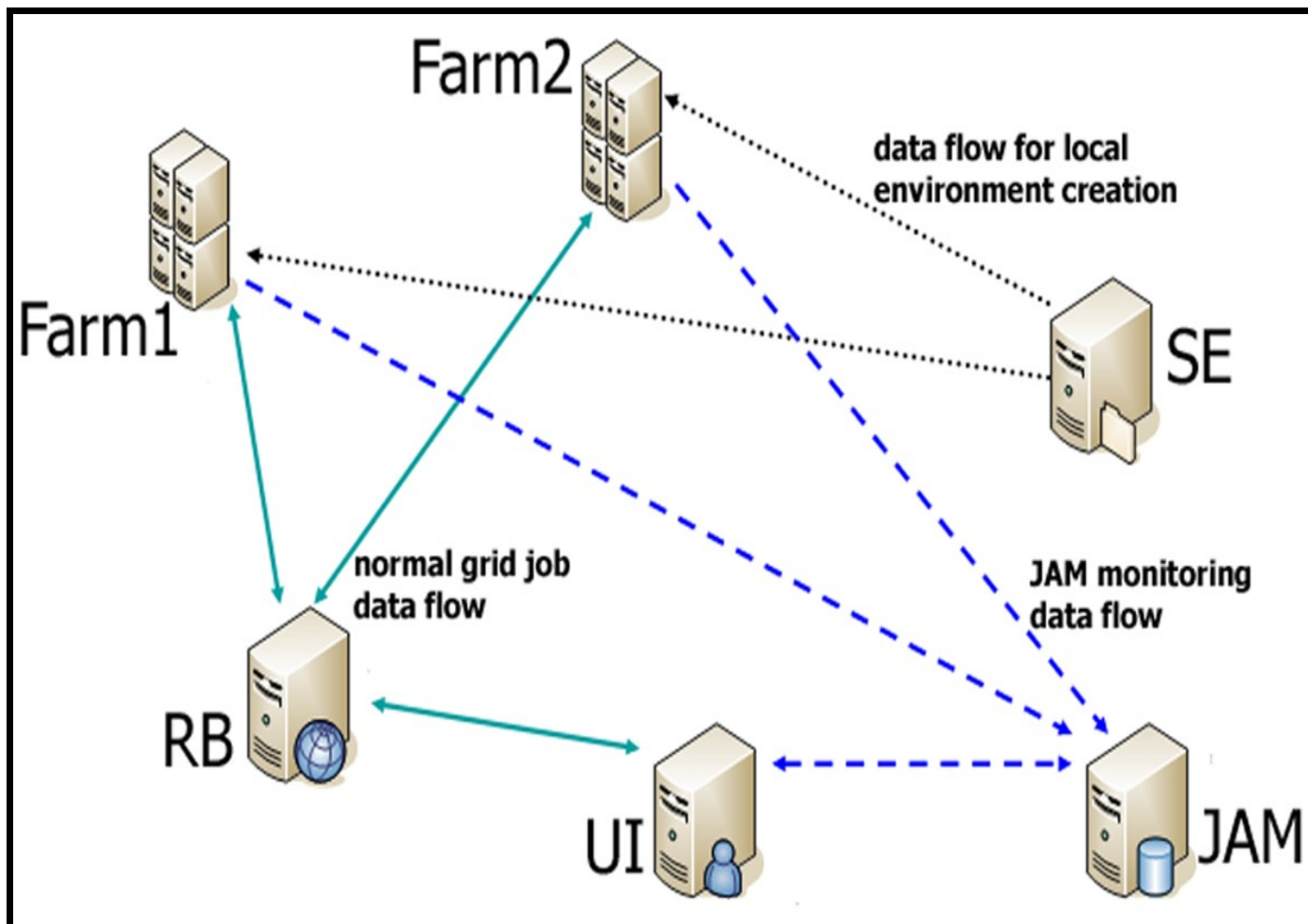
We selected a list of ~ 80000 gene products of 13 different organisms to compare with all the other 1.000.000 gene products

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.

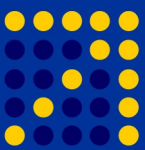
This search was completed using up to 950 WNs in ~3 days, instead of 5 years!

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

GRID distribution



Scheme of the Grid process



GRID distribution (text files distribution)

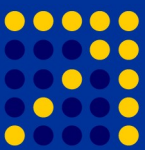
We compare all gene products (>1000000) against all.

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 ~67000 jobs and 42000 jobs were completed succesfully: 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!!



Results

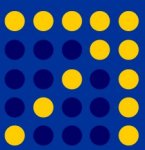
This method finds most of the orthologous gene products and members of the same gene family, but also finds functional analogous gene products not belonging to the same family with low level of sequence similarity but a high number of common GO terms and sharing therefore similar functions.

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.

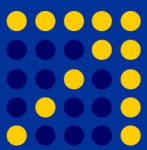


Benefit for the Biologist

BioinfoGRID

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

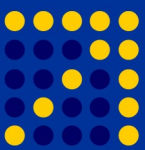
- Gene Ontology is continuously improving its associations, using new GO terms and describing new gene products;

- ~ 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.



Acknowledgments

BioinfoGRID

- **Giacinto Donvito¹, Giorgio Maggi¹**

For technical aspects and grid distribution

- **Andreas Gisel², Angelica Tulipano^{1,2}, Flavio Licciulli²,
Giulia De Sario²**

For bioinformatical aspects

¹ INFN, Bari

² CNR-ITB, Bari