BioinfoGRID

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

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

www.ba.itb.cnr.it

Enabling Grids for E-science in Europe

www.eu-egee.org

INFN GRID

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

Outlines

- 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, these are 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, and the cellular components they are associated to.

Gene Ontology

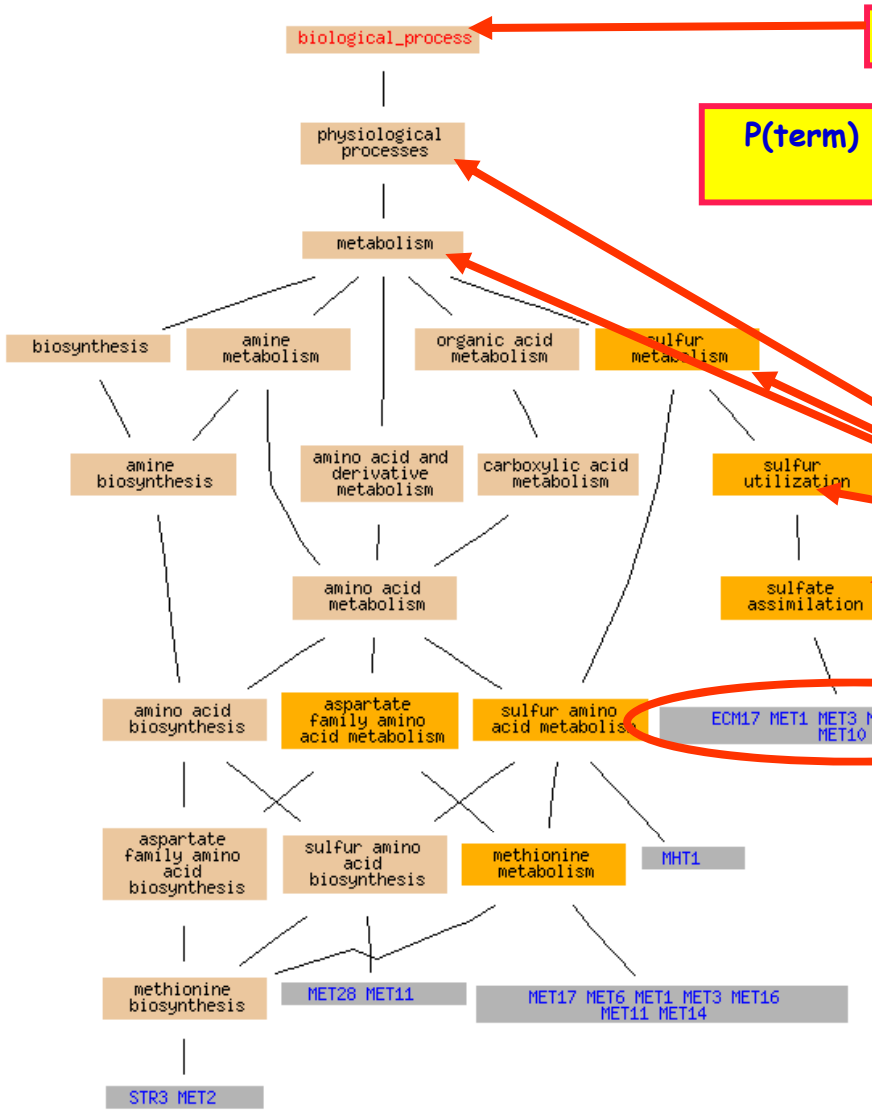
GO is an international standard to annotate gene products:

- 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 actually working with version go_02_08)
- By now more than 3.300.000 gene products are described by the GO terms associated covering more than 135000 species
- more than 24000 GO terms ending up with >14.500.000 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(term) = # gene products associated to the term or any of its children / # 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 is the information content and the more descriptive it is.

GO terms of the path

GO term directly associated

Gene products annotated

Algorithm of the search

Through a χ^2 statistical test we compare two gene product A and 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(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 is the probability of functional dependence between the two gene products A and B.

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

Problem of the search

Problem:

The comparison of all the gene products annotated is very data-intensive (>3.300.000 gene products) and time-consuming (a single comparison occupies one CPU for 30-45 min

the whole search ~180 CPU years !

Approach step I

BCL2_HUMAN

Ontologies	
GO	GO:0005741; Cellular component: mitochondrial outer membrane *(inferred from direct assay from UniProtKB).* GO:0031965; Cellular component: nuclear membrane *(inferred from direct assay from UniProtKB).* GO:0051434; Molecular function: BH3 domain binding *(inferred from physical interaction from UniProtKB).* GO:0002020; Molecular function: protease binding *(inferred from direct assay from UniProtKB).* GO:0046982; Molecular function: protein heterodimerization activity *(inferred from physical interaction from UniProtKB).* GO:0042803; Molecular function: protein homodimerization activity *(inferred from direct assay from UniProtKB).* GO:0006916; Biological process: anti-apoptosis *(inferred from direct assay from UniProtKB).* GO:0051607; Biological process: defense response to virus *(inferred from direct assay from UniProtKB).* GO:0007565; Biological process: female pregnancy *(non-traceable author statement from UniProtKB).* GO:0006959; Biological process: humoral immune response *(traceable author statement from UniProtKB).* GO:0032848; Biological process: negative regulation of cellular pH reduction *(inferred from direct assay from UniProtKB).* GO:0051902; Biological process: negative regulation of mitochondrial depolarization *(traceable author statement from UniProtKB).* GO:0051402; Biological process: neuron apoptosis *(traceable author statement from HGNC).* GO:0046902; Biological process: regulation of mitochondrial membrane permeability *(inferred from sequence or structural similarity from HGNC).* GO:0000074; Biological process: regulation of progression through cell cycle *(inferred from direct assay from UniProtKB).* GO:0043497; Biological process: regulation of protein heterodimerization activity *(inferred from direct assay from UniProtKB).* GO:0043496; Biological process: regulation of protein homodimerization activity *(inferred from direct assay from UniProtKB).* GO:0001836; Biological process: release of cytochrome c from mitochondria *(non-traceable author statement from UniProtKB).* GO:0010039; Biological process: response to iron ion *(inferred from direct assay from UniProtKB).* GO:0035094; Biological process: response to nicotine *(inferred from direct assay from UniProtKB).* GO:0009314; Biological process: response to radiation *(non-traceable author statement from UniProtKB).* GO:0009636; Biological process: response to toxin *(inferred from direct assay from HGNC).*

Non-redundant list of GO terms ➔ Description of gene product

Approach step I

3.3 million gene products (UniProt) are described by 87438 descriptions

Reduce a 3 million by 3 million matrix to a 87500 by 87500 matrix

Factor of ~ 35 on the level of data to process

Factor of ~ 30 on the level of process time per comparison

Total factor of ~ 1000

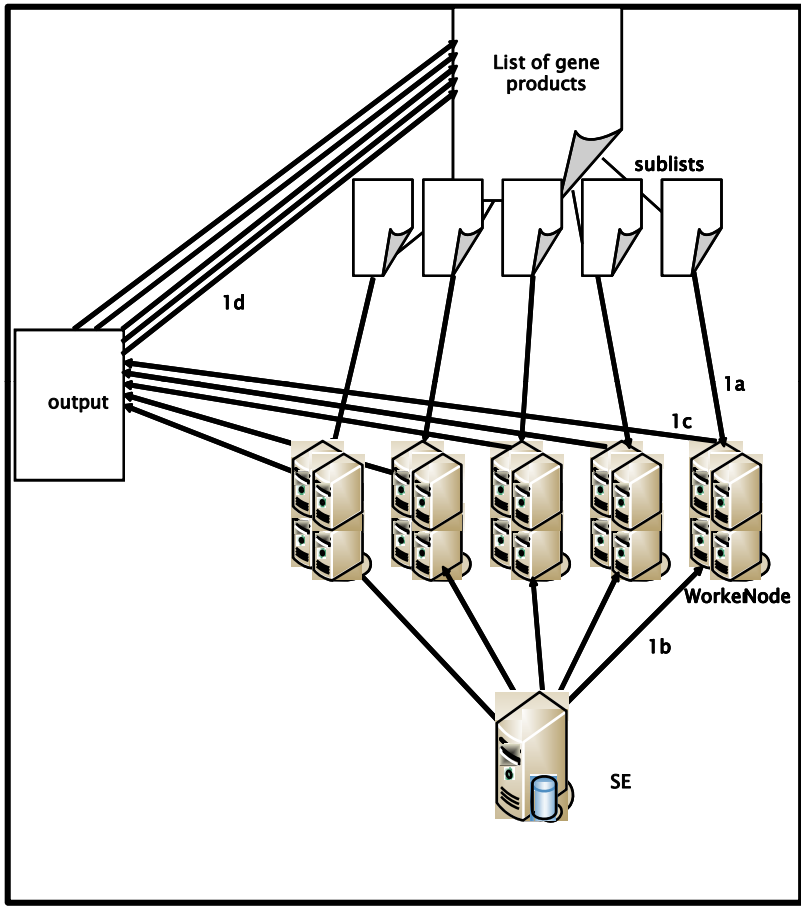
Reduction from 180 CPU years

to 2 CPU months

Approach step II

Run the search on the INFN GRID (VO 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.
- Each worker node has its own local source of data



Scheme of the data flow in a job run

GRID distribution

We compare all gene products (>3300000 gene products resp. 87500 descriptions) against all.

We downloaded all the needed information in text files stored on SE for further distribution to the worker nodes:

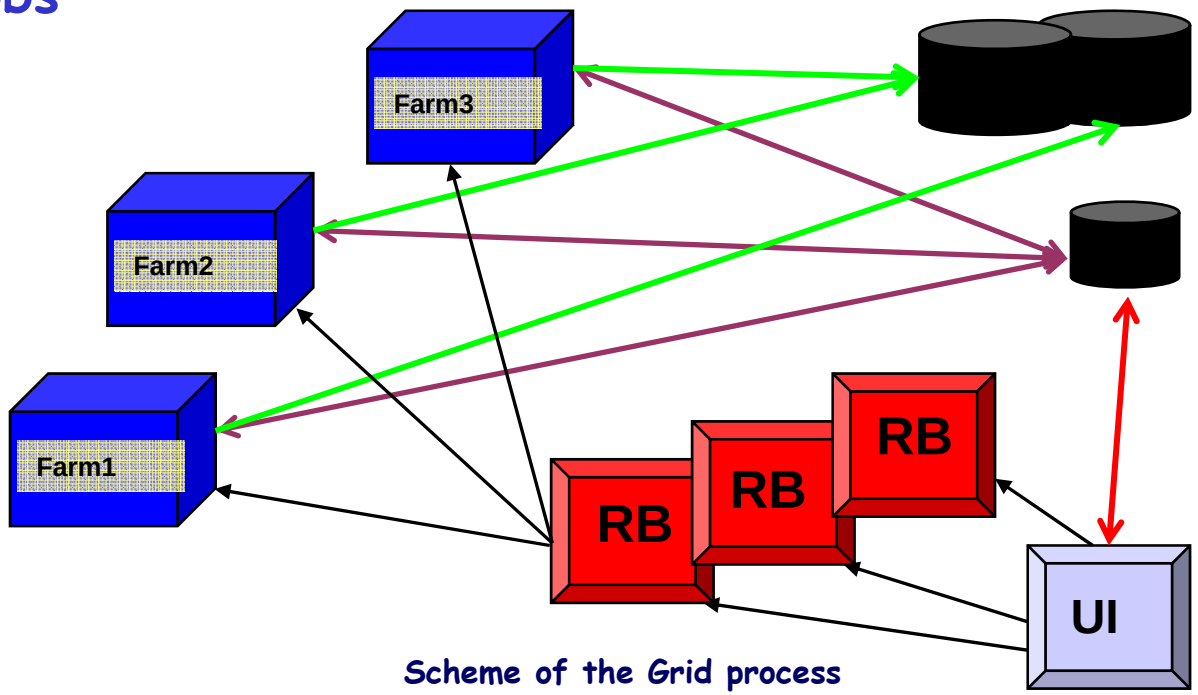
This search was split into ~ 4400 jobs terminated successfully having submitted ~6000 jobs : the submission uses 3 RB's in a round robin algorithm in order to avoid the overload of a single RB and that the failure of a single RB can stop the submission of jobs.

A job submission tool (JST) was used to submit and control the search so that failed jobs were resubmitted.

This search was completed in less than a day instead of 65 days using up to 200 WN.

GRID distribution

A job submission tool [*] was used for the submission of a large number of jobs



Scheme of the Grid process

[*] http://webcms.ba.infn.it/cms-software/index.html/index.php/Main/JobSubmissionTool

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.

Results

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

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 and species

Future plans

- Gene Ontology is continuosly improving its associations, using new GO terms and describing new gene products;
- ~ every month an update to provide:
  - a MySQL dump for distribution of the analogue gene products with the monthly GODB release.

- Graphical web interface to access to the gene analogue data

Acknowledgments

- Giacinto Donvito[1], Giorgio Maggi[1]

For technical aspects and grid distribution

- Angelica Tulipano[1,2], Giulia De Sario[2], Andreas Gisel[2],

For bioinformatical aspects

[1] INFN, Bari

[2] CNR-ITB, Bari

Tulipano A, Donvito G, Licciulli F, Maggi G, Gisel A. (2007) Gene analogue finder: a GRID solution for finding functionally analogous gene products. *BMC Bioinformatics*, 8(1):329

Acknowledgments

Thanks!!!!

… and questions??