RESEARCHER'S DIGEST: USING GENE ANNOTATIONS TO CLASSIFY GENE LISTS INTO FUNCTIONAL GROUPS

Juan Carlos Triviño, Juan Carlos Sánchez-Ferrero and Juan Carlos Oliveros

(jctrivino@cnb.csic.es, jcsanchez@cnb.csic.es, oliveros@cnb.csic.es)

Service of Bioinformatics for Genomics and Proteomics (BioinfoGP), Centro Nacional de Biotecnología (CNB-CSIC). Darwin 3, Campus de Cantoblanco, 28049, Madrid, Spain.

Introduction

DNA Microarrays are used in a routine way to measure the transcription level of thousands of genes in several experimental conditions. The main result of these experiments usually consists on one or several lists of genes whose transcription levels are induced (or repressed) compared against a reference, or share similar transcription patterns in different cellular states.

While there are many robust statistical methods to obtain these gene lists, today there is a growing interest on the development of new bioinformatic tools that would assist the researcher in the interpretation of their results going beyond the pure numerical classification.

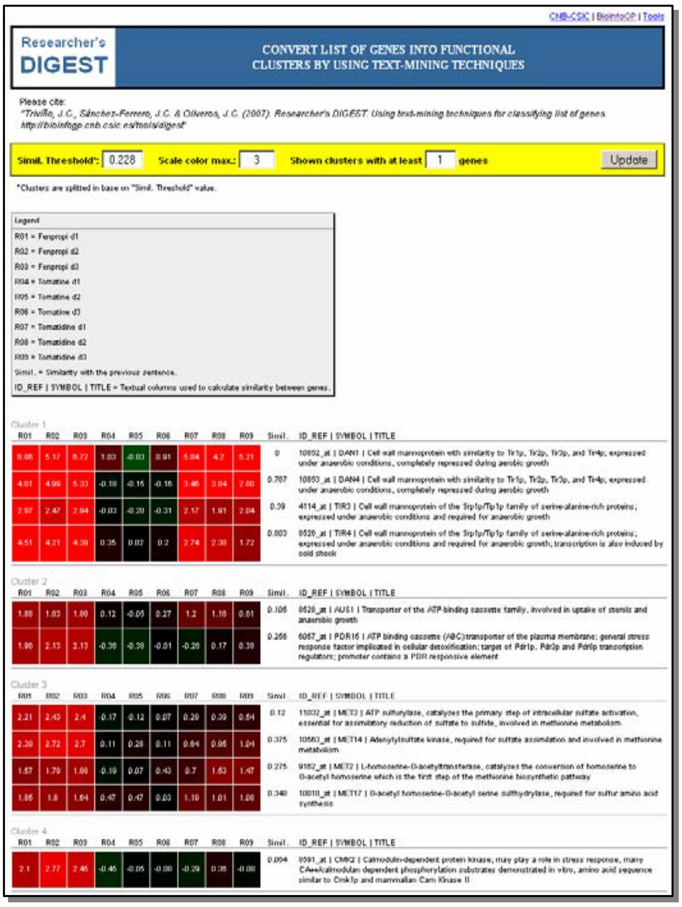
Researcher's Digest is an on-line tool to classify gene lists, obtained from microarray experiments, into functional groups. To do that, the system makes use of text-mining techniques applied to the gene annotations.

The program:

1. Classify the words in the gene annotations in biological categories (enzymes, accession codes, gene families, etc.).
2. Evaluate the degree of similarity between all pairs of sentences (the more relevant terms are shared between two annotations, the more similar are they).
3. Sort all genes starting from the first one in the list so the more similar gene is placed the next and so on.
4. Estimates a similarity threshold that best separates the gene groups.

The results…

…are presented in a user friendly web page where the user can modify the similarity threshold used to separate the groups and change some visualization options.



CNB-CSIC | BioinfoGP | Tools
Researcher's DIGEST
CONVERT LIST OF GENES INTO FUNCTIONAL CLUSTERS BY USING TEXT-MINING TECHNIQUES
Update

Please cite:
"Triviño, J.C., Sánchez-Ferrero, J.C. & Oliveros, J.C. (2007). Researcher's DIGEST. Using text-mining techniques for classifying list of genes. http://bioinfogp.cnb.csic.es/tools/digest"

Simil. Threshold*: 0.228 Scale color max.: 3 Shown clusters with at least 1 genes

*Clusters are splitted in base on "Simil. Threshold" value.

Legend
R01 = Fenpropi d1
R02 = Fenpropi d2
R03 = Fenpropi d3
R04 = Tomatine d1
R05 = Tomatine d2
R06 = Tomatine d3
R07 = Tomatidine d1
R08 = Tomatidine d2
R09 = Tomatidine d3
Simil. = Similarity with the previous sentence.
ID_REF | SYMBOL | TITLE = Textual columns used to calculate similarity between genes.

Cluster 1

R01	R02	R03	R04	R05	R06	R07	R08	R09	Simil.	ID_REF | SYMBOL | TITLE
6.05	5.17	6.72	1.03	-0.03	0.91	5.84	4.2	5.21	0	10852_at | DAN1 | Cell wall mannoprotein with similarity to Tir1p, Tir2p, Tir3p, and Tir4p; expressed under anaerobic conditions, completely repressed during aerobic growth
4.01	4.99	5.33	-0.18	-0.15	-0.16	3.46	3.04	2.00	0.707	10853_at | DAN4 | Cell wall mannoprotein with similarity to Tir1p, Tir2p, Tir3p, and Tir4p; expressed under anaerobic conditions, completely repressed during aerobic growth
2.97	2.47	2.94	-0.03	-0.20	-0.31	2.17	1.91	2.04	0.39	4114_at | TIR3 | Cell wall mannoprotein of the Srp1p/Tip1p family of serine-alanine-rich proteins; expressed under anaerobic conditions and required for anaerobic growth
4.51	4.21	4.39	0.35	0.02	0.2	2.74	2.38	1.72	0.803	8520_at | TIR4 | Cell wall mannoprotein of the Srp1p/Tip1p family of serine-alanine-rich proteins; expressed under anaerobic conditions and required for anaerobic growth; transcription is also induced by cold shock

Cluster 2

R01	R02	R03	R04	R05	R06	R07	R08	R09	Simil.	ID_REF | SYMBOL | TITLE
1.88	1.63	1.80	0.12	-0.05	0.27	1.2	1.10	0.61	0.106	8528_at | AUS1 | Transporter of the ATP-binding cassette family, involved in uptake of sterols and anaerobic growth
1.90	2.13	2.13	-0.38	-0.38	-0.01	-0.28	0.17	0.38	0.256	6067_at | PDR16 | ATP binding cassette (ABC)transporter of the plasma membrane; general stress response factor implicated in cellular detoxification; target of Pdr1p, Pdr3p and Pdr8p transcription regulators; promoter contains a PDR responsive element

Cluster 3

R01	R02	R03	R04	R05	R06	R07	R08	R09	Simil.	ID_REF | SYMBOL | TITLE
2.21	2.43	2.4	-0.17	-0.12	0.07	0.20	0.39	0.64	0.12	11002_at | MET3 | ATP sulfurylase, catalyzes the primary step of intracellular sulfate activation, essential for assimilatory reduction of sulfate to sulfide, involved in methionine metabolism
2.30	2.72	2.7	0.11	0.28	0.11	0.64	0.06	1.04	0.375	10560_at | MET14 | Adenylylsulfate kinase, required for sulfate assimilation and involved in methionine metabolism
1.67	1.79	1.98	-0.19	0.07	0.43	0.7	1.63	1.47	0.275	9182_at | MET2 | L-homoserine-O-acetyltransferase, catalyzes the conversion of homoserine to O-acetyl homoserine which is the first step of the methionine biosynthetic pathway
1.86	1.8	1.64	0.47	0.47	0.03	1.19	1.01	1.08	0.348	10010_at | MET17 | O-acetyl homoserine-O-acetyl serine sulfhydrylase, required for sulfur amino acid synthesis

Cluster 4

R01	R02	R03	R04	R05	R06	R07	R08	R09	Simil.	ID_REF | SYMBOL | TITLE
2.1	2.77	2.46	-0.46	-0.05	-0.08	-0.29	0.38	-0.08	0.004	8591_at | CMK2 | Calmodulin-dependent protein kinase; may play a role in stress response, many CA++/calmodulan dependent phosphorylation substrates demonstrated in vitro, amino acid sequence similar to Cmk1p and mammalian Cam Kinase II

Digest is one of the few bioinformatics tools that uses free-text as source of data for grouping genes and can compare up to ~1000 genes in few seconds. It is accessible at:

http://bioinfogp.cnb.csic.es/tools/digest