Co-expressed Gene Groups Analysis (CGGA): An Automatic Tool for the Interpretation of Microarray Experiments

Ricardo Martinez, Nicolas Pasquier, Claude Pasquier, Martine Collard, Lucero Lopez-Perez

Co-expressed gene groups analysis (CGGA): An automatic tool for the interpretation of microarray experiments

Ricardo Martinez[1], Nicolas Pasquier1[1], Martine Collard[1], Claude Pasquier[2] and Lucero Lopez-Perez[3]

[1]Laboratoire I3S; 2000, route des lucioles, 06903 Sophia-Antipolis cedex, France;
{rmartine,pasquier,mcollard}@i3s.unice.fr

[2]Laboratoire Biologie Virtuelle; Centre de Biochimie, Valrose; 06108 Nice cedex 2, France;
claude.pasquier@unice.fr

[3]INRIA Sophia Antipolis; 2004, route des Lucioles; 06903 Sophia-Antipolis cedex, France;
lucero.lopez@gmail.com

Summary

Microarray technology produces vast amounts of data by measuring simultaneously the expression levels of thousands of genes under hundreds of biological conditions. Nowadays, one of the principal challenges in bioinformatics is the interpretation of this large amount of data using different sources of information. We have developed a novel data analysis method named CGGA (Co-expressed Gene Groups Analysis) that automatically finds groups of genes that are functionally enriched, i.e. have the same functional annotations, and are co-expressed. CGGA automatically integrates the information of microarrays, i.e. gene expression profiles, with the functional annotations of the genes obtained by the genome-wide information sources such as Gene Ontology. By applying CGGA to well-known microarray experiments, we have identified the principal functionally enriched and co-expressed gene groups, and we have shown that this approach enhances and accelerates the interpretation of DNA microarray experiments.[1]

Keywords: Microarray, Ontology, Co-expression, Genes and Functional Annotations.

1 Introduction

One of the main challenges in microarray data analysis is to highlight the principal functional gene groups using different sources of genomic information. These sources of information, constantly growing by an ever-increasingly volume of genomic data, are:

- Taxonomies, thesaurus and ontologies providing the semantic information for the genes, for example: Gene Ontology (GO)[2] , Unified Medical Language System (UMLS), Medical Subject Headings (MESH), Universal Protein Ressource (Uniprot), etc.

[1]CGGA program is available at `http://www.i3s.unice.fr/~rmartine/CGGA`
[2]Gene Ontology project: `http://www.geneontology.org/`

- Literature and bibliographic databases (articles, on-line libraries, etc.) covering the results of previous analysis: Pubmed, Medline, etc.
- Experience databases: Arrayexpress, Gene Expression Omnibus (GEO), etc.
- Nomenclature databases: human (HUGO), fruit fly (flybase), yeast (SGD), etc.

A variety of statistical and data analysis approaches, identifying groups of co-expressed genes based only on the expression profiles, i.e. without taking into account prior knowledge, have been reported: [4], [6], [8], [22]. A common characteristic of purely numerical approaches is that they determine gene groups (called clusters) of potential interest; however, they leave to the expert the task of discovering and interpreting biological similarities hidden within these groups.

These methods are useful, because they guide the analysis of the co-expressed gene groups. Nevertheless, their results are often incomplete, because these approaches do not include biological considerations and also, they reject heterogeneous functional groups i.e. that belong to various functional groups [21]. Actually, one of the major goals in bioinformatics is the automatic integration of biological knowledge from different sources of information with gene expression data [2]. A first assessment of the methods developed to answer this challenge was proposed by Chuaqui [5].

Nowadays, one of the richest sources of biological annotations is contained on structured and controlled vocabulary such as ontologies. These annotations can be functional, relational and syntactic information on genes. We target here the enrichment of two recently developed research orientations, *sequential* and *a priori*, that exploit multiple sources of annotations such as Gene Ontology.

The sequential axis methods build co-expressed gene clusters (groups of genes with a similar expression profiles). Then they detect co-annotated gene subsets (sharing the same annotation). Afterwards, the statistical significance of these co-annotated gene subsets is tested. Among the methods in this axis let us quote *Onto Express* [7], *Quality Tool* [9], *EASE* [10], *THEA* [15] and *Graph Modeling* [21].

The a priori axis methods first finds functionally enriched groups (FEG), i.e. groups of co-annotated genes by function. Then they integrate the information contained in the profiles of expression. Later on, the statistical significance of the FEG is tested by an enriched score [14], a pc-value based on a hypergeometric distribution [3], or a z-score test [11].

Our approach, called CGGA (Co-expressed Gene Groups Analysis), is inspired by the a priori axis: the FEG are initially formed from the Gene Ontology, next a function, which synthesizes the information contained in the expression data, is applied in order to obtain an arranged gene list. In this list, the genes are sorted by decreasing expression variability. The statistical significance of the FEG obtained is then tested using a similar hypothesis proof as presented in *Onto Express*. Finally, we obtain co-expressed and statistically significant FEG.

The IGA algorithm [3] is a method from the a priori axis that allows to find the FEG of most expressed genes, leaving out all the FEG made up of less expressed genes that have however a similar level of expression and thus can be related later. Our CGGA method is an extension of the IGA algorithm that finds all subsets FEG of significant co-expressed genes with similar level of expression.

This article is organized in the following way: in section 2 we describe the validation data as well as the tools used: databases, ontologies, statistical packages; our algorithm CGGA is described in section 3; the results obtained are presented in section 4 and the last section presents our conclusions.

2 Data and Methods

2.1 Dataset and Statistical Pretreatment

In order to evaluate our approach, the CGGA algorithm was applied to the DeRisi dataset which is one of the most studied in this field [6]. This dataset measures the variations in gene expression profiles during the cellular process of diauxic shift for the yeast *Saccharomyces Cerevisiae*. When inoculated into a glucose-rich medium (anaerobic growth), the budding yeast can convert the glucose to ethanol (aerobic respiration), the shift from anaerobic fermentation of glucose to aerobic respiration of ethanol is the so-called *diauxic shift*.

The technique used is double channel microarray, obtained by two color fluorochromes with distinct emission spectra Cy3 and Cy5. The DeRisi dataset contains the expression levels of 6199 ORF's, opening reading frame, of the yeast (an entirely sequenced organism), for 7 temporal points that correspond to samples harvested at successive two-hour intervals after an initial nine hours of growth.

The dataset was pretreated by taking the log_2 ratios (to consider cellular inductions and repressions in a numerically equal way) and applying the imputation algorithm of k-nearest neighbors [12] in order to treat the missing values (1.9% of the total).

2.2 Ontology and Functionally Enriched Groups (FEG)

In order to fully exploit data, knowledge discovery systems rely on a formal representation of information based on a well-defined semantic [19]. These formal requirements have led to the utilisation of the well structured ontology Gene Ontology (GO) and the nomenclature database SGD[3]. Sructure of Gene Ontology (GO) and the annotations of Saccharomyces Cerevisiae Genome with GO terms were retrieved from the GO database web site[4] on may 2006. Automatic annotations not reviewed by curators (IEA evidence code) were discarded. For each gene product, we have stored all the functional annotations of the gene product and his parents preserving the hierarchical structure of GO.

Gene Ontology (GO)

GO is a controlled vocabulary developed by a consortium of scientists to address the need for consistent descriptions of gene products in different databases. It can be used to annotate a gene or gene product by a *GO-term*, with regard to its molecular functions (GO:MF), cellular localizations (GO:CL) and biological processes (GO:BP).

GO-terms are organized in structures called directed acyclic graphs (DAGs), which differ from hierarchies in that a child, or more specialized, term can have many parent, or less specialized,

[3] Saccharomyces Genome Database: `http://www.yeastgenome.org/`

[4] `http://www.godatabase.org/dev/database/`

terms. Annotators can assign properties of gene products at different levels, depending on how much is known about a gene [1].

Genome Data

In order to be congruent with GO annotations files and among the multiple yeast gene identifiers, we have used the yeast *Saccharomyces cerevisiae* database. SGD is a scientific database of the molecular biology and genetics of the yeast [24].

Functionally Enriched Groups (FEG)

Queries carried out on the GO database have built the whole set of the FEG: each FEG corresponds to a couple made up of a *GO-term* and of the list of genes annotated by this one.

2.3 Expression Profile Measure of the Genes

In order to incorporate the expression profile of the genes, we have used a measurement of their variability of expression, *f-score*, which is more robust than other measurements such as *anova, fold change* or *t-student* statistics [17].

This measurement enables us to build a list of genes, *g-rank*, ordered by decreasing expression variability. We have used the SAM program [23] to calculate the *f-score* associated with each gene.

3 Co-expressed Gene Groups Analysis (CGGA)

The CGGA is based on the idea that any resembling change (co-expression) of a gene subset belonging to an FEG is physiologically relevant. We say that two genes are co-expressed if they are close in the sense of the metric given by the expression variability (*f-score*). The CGGA algorithm computes a *pc-value* for each FEG that estimates its coherence (according to the *g-rank*) and thus allows to detect the statistically significant groups.

3.1 CGGA Algorithm

The CGGA algorithm first builds the *g-rank* list from the expression levels and the FEG from the GO database. For each FEG of n genes, the algorithm determines the $n(n+1)/2$ gene subsets that we want to test for co-expression. For each subset we compute the *pc-value* corresponding to the test described below in order decide whenever the genes of the subset are co-expressed.

Let H_0 be the hypothesis that x genes from one of these subsets were associated by chance, given their place on the *g-rank* list. If H_0 is rejected, there are good chances that the genes belonging to the subset are improbably close on the list because they have a very similar expression profile.

To compute the probabilty that H_0 is true for a fixed subset FEG or class, let us ask the question, how likely is to find x members from the class placed this way on the *g-rank* list? The answer to this question is given by the following hypergeometric distribution:

$$p(X = x|N, R_{g(x)}, n) = \frac{\begin{pmatrix} R_{g(x)} \\ x \end{pmatrix} \begin{pmatrix} N - R_{g(x)} \\ n - x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

where:

$$p(X = 0|N, R_{g(x)}, n) = 0$$

with:

- N : total number of genes in the dataset,
- n : number of genes in the FEG,
- x : position of the gene in the FEG (previously ordered by rank),
- $r_{g(x)}$: absolute rank of the gene of position x in the *g-rank* list,
- $R_{g(x)}$: number of ranks (in the *g-rank* list) between the gene of position x from its FEG predecessor. $R_{g(x)}$ is calculated from the absolute ranks $r_{g(x)}$ according to the formula:

$$R_{g(x)} = r_{g(x)} - r_{g(x-1)} + 1 \; where \; R_{g(0)} = r_{g(x)} = 1.$$

The *pc-value* corresponding to this hypothesis test is (refer to [7] for details):

$$pc - value(x) = 1 - \sum_{k=1}^{x} p(X = k|N, R_{g(k)}, n).$$

In order to accept or reject H_0 we will use the following significance threshold:

$$p - value = Min \left\{ \frac{1}{N}, \frac{1}{|\Omega|} \right\},$$

where $|\Omega|$ is the cardinality of the set of functional annotations. So, for each FEG, if $pc - value(x) < p - value$ then H_0 is rejected, i.e. the FEG is statistically significant.

Pseudo-code for CGGA algorithm is presented on Figure 1. The algorithm has been implemented in Perl (language). It takes as input the list of annotations for each gene (generated by a query on the database GO database containing all the GO annotations) and the ordered *g-rank* list of the N genes. It returns as output the list of the groups of significant co-expressed genes.

The algorithm begins by computing the *p-value* (stage 2) and generating the FEG from the GO annotations (stages 3 to 9). Then it considers successively each FEG (stages 10 to 18). For each FEG, it takes all non-empty subsets and computes the pc-value for each of them (stages 11 to 16). If the computed *pc-value* is less than *p-value*, the subset is added to the FEG results list (stages 13 to 15). The added subsets that are non-maximal according to the inclusion are deleted (stage 17).

Input: List of annotations for each gene G: annotations(G).
Ordered list of N genes: *g-rank*.
Output: Results set containing the FEG of co-expressed genes: $results(FEG_A)$.

```
1   Begin
2       compute p-value
3       for each annotation A of the GO do
4           for each gene G do
5               if A ∈ annotations(G) then
6                   FEG_A ← FEG_A U G
7               end if
8           end for
9       end for
10      for each FEG_A do
11          for each subset S of FEG_A do
12              compute pc − value(S)
13              if pc − value(S) < p − value then
14                results(FEG_A) ← results(FEG_A)US
15              end if
16          end for
17          delete from results(FEG_A) the non maximal S according to inclusion
18      end for
19      results ← U_{i=A} results(FEG_i)
20  End
```

Figure 1: CGGA Algorithm

For example, let the FEG_A annotated set,

$$FEG_A = \{g_1, g_2, g_3\},$$

thus we have:

$$results(FEG_A) = \{\{g_1\}, \{g_2\}, \{g_3\}, \{g_1, g_2\}, \{g_2, g_3\}, \{g_1, g_2, g_3\}\}.$$

Then, all the subsets of $\{g_1, g_2, g_3\}$ are deleted from results(FEG_A). Finally, the total result consists of all the groups of co-expressed and significant genes (stage 19).

3.2 Example

An example of the CGGA applied to a group of co-annotated genes is presented in Table 1. The data used in the example is from the experiment carried out by DeRisi (see section 2.1), where the diauxic shift process of the yeast, *Saccharomyces Cerevisiae*, was analyzed.

The ordered *g-rank* list was computed using the *f-score* obtained with the SAM program (see section 2.3). The data of the FEG, annotated "vacuolar protein catabolism", was obtained from the GO database (see section 2.2). This FEG contains 4 genes ($n = 4$) whose rows in the total *g-rank* list vary from 6 to 424.

In Table 1 we show the values of the parameters needed to determine the significant gene subsets within the FEG. We have highlighted the subset of genes: $\{1, 3\}$, from vacuolar protein catabolism FEG, found significantly co-expressed by CGGA.

List *g-rank*	x	Gene ID (SGD)	GO Annotation	$r_{g(x)}$	$R_{g(x)}$
1				1	
2				2	
6	1	S000000490	VACUOLAR PROTEIN CATABOLISM	6	1
7				7	
8	2	S000001586	VACUOLAR PROTEIN CATABOLISM	8	3
69	3	S000000786	vacuolar protein catabolism	69	62
424	4	S000006075	vacuolar protein catabolism	424	356
N				N	

Table 1: CGGA Analysis for the FEG of genes annotated "vacuolar protein catabolism"

CGGA tested for H_0 the (4*5)/2=10 FEG subsets computing their *pc-value* and comparing it to the *p-value*. For example, the *pc-value* corresponding to the subset {S000000490, S000001586} of rank 6 and 8 in *g-rank* is $2.63E^{-05}$ (cf. Table 2). This *pc-value* being lower than *p-value*, fixed at $6.88E^{-04}$ (cf. section 3.1), CGGA rejected H_0 and the group of genes {S000000490, S000001586} is then labelled statistically significant and co-expressed. We see that the subset with genes of rank 6 and 8 is very close and then co-expressed. On the other hand the genes of rank 69 and 424 are rather distant from their closer neighbours, i.e. the groups that contain them are not co-expressed significantly.

4 Results

In order to evaluate our method, we compared the results obtained by DeRisi [6], IGA [3] and CGGA. The results obtained using CGGA for the over-expressed and under-expressed genes are presented in Table 2 and Table 3 respectively. As expected, almost all groups identified as significantly co-expressed by the DeRisi method have also been identified by the CGGA. The groups identified by CGGA and DeRisi are in **bold**, the ones identified only by CGGA are in *italics*, and the only group identified also by IGA is in SMALL CAPS.

In the case of over-expressed genes (Table 2), CGGA found seven of the nine groups obtained manually by DeRisi [6]. The two annotated groups "glycogen metabolism" and "glycogen synthase" have not been identified by CGGA because they are expressed only at the initial phase of the process. However CGGA identified eight other statistically significant and coherent groups. Only one of these eight other groups has also been identified by IGA and none of them by DeRisi.

Functionally Enriched GO Group	n genes	x Over-expressed genes	$pc - value$
proton-transporting ATP synthase com-plex	2	2	$4.38E^{-06}$
invasive growth (sensu Saccharomyces)	5	3	$6.13E^{-06}$
signal transduction during filamentous growth	2	2	$8.77E^{-06}$
respiratory chain complex II	4	4	$3.75E^{-05}$
succinate dehydrogenase activity	4	4	$3.75E^{-05}$
mitochondrial electron transport	4	4	$3.75E^{-05}$
aerobic respiration	36	10	$3.30E^{-05}$
tricarboxylic acid cycle	14	5	$5.09E^{-05}$
tricarboxylic acid cycle	14	5	$6.54E^{-05}$
gluconeogenesis	12	2	$9.64E^{-05}$
response to oxidative stress	10	3	$1.55E^{-06}$
filamentous growth	8	4	$9.06E^{-05}$
VACUOLAR PROTEIN CATABOLISM	4	2	$2.63E^{-05}$
respiratory chain complex IV	8	2	$4.05E^{-04}$
cytochrome-c oxidase activity	8	2	$4.05E^{-04}$

Table 2: Over-Expressed FEGs obtained by CGGA with a $p - value = 6.88E^{-04}$

For the case of under-expressed genes (Table 3), CGGA has found seven of the eight gene groups selected manually by DeRisi. As for over-expressed genes, the group annotated "ribosome biogenesis" was not identified by CGGA, because it was only expressed during the final phase of the process. CGGA have also identified seven other statistically significant and coherent groups which were not identified on the DeRisi analysis nor by IGA.

The three groups identified by DeRisi that CGGA did not identify, namely the over-expressed groups "glycogen metabolism" and "glycogen synthase", and the under-expressed group "ribosome biogenesis" share two important properties. First, they contain genes belonging to a heterogeneous structure, i.e genes that appertain to several functional groups. Second, these FEG are not expressed throughout the entire process but only during a specific phase. Detect these groups will only be possible by integrating information on the metabolic pathways ontologies such as: KEGG, EMP, CFG, etc.

5 Conclusion

The CGGA algorithm presented in this article makes it possible to automatically identify groups of significantly co-expressed and functionally enriched genes without any prior knowledge of the expected outcome. CGGA can be used as a fast and efficient tool for exploiting every source of biological annotation and different measure of gene variability.

In contrast to sequential approaches such as [7], [9], [10], [15] and [21], CGGA analyze all the possible subsets of each FEG and does not depend on the availability of fixed lists of expressed genes. Thus, it can be used to increase the sensitivity of gene detection, especially when dealing with very noisy datasets. CGGA can even produce statistically significant results without any

Functionally Enriched GO Group	n genes	x Under-Expressed genes	$pc - value$
chromatin modification	6	5	$2.35E^{-06}$
mitochondrial inner memb. prot. inser. complex	3	2	$3.60E^{-06}$
regulation of nitrogen utilization	4	2	$7.20E^{-06}$
acid phosphatase activity	4	2	$7.20E^{-06}$
histone acetylation	4	4	$7.95E^{-06}$
nucleolus	52	10	$3.41E^{-05}$
rRNA modification	10	3	$2.75E^{-05}$
transcription initiation from RNA poly. II prom.	14	3	$1.00E^{-05}$
mitochondrial matrix	15	3	$1.25E^{-05}$
processing of 20S pre-rRNA	11	2	$1.97E^{-04}$
ribosomal large subunit biogenesis	9	4	$3.17E^{-04}$
small nucleolar ribonucleoprotein complex	20	3	$2.52E^{-04}$
cytosolic large ribosomal subunit	69	13	$2.87E^{-04}$
ribosomal large subunit assembly and maint.	21	2	$2.52E^{-04}$

Table 3: Under-Expressed FEGs obtained by CGGA with a $p - value = 6.88E^{-04}$

experimental replication. It does not need that all genes in a significant and co-expressed group change, so it is therefore robust against imperfect class assignments, which can be derived from public sources (wrong annotations in ontologies) or automated processes (naming errors, spelling mistakes, etc.).

The automated functional annotation provided by our algorithm reduces the complexity of microarray analysis results and enables the integration of different sources of genomic information such as ontologies.

CGGA can be used as a tool for platform-independent validation of a microarray experiment and its comparison with the huge number of existing experimental databases and the documentation databases. Experimental results show the interest of our approach and make it possible to identify relevant information on the analyzed biological processes. In order to identify heterogeneous groups of genes expressed only in certain phases of the process, we plan to integrate the information concerning the metabolic pathways ontologies for future work.

References

[1] Ashburner M., Ball C., Blake J., et al.: Gene Ontology: tool for the unification of biology. *Nature Genetics*, Vol. 25. (2001) 25-29.

[2] Attwood T., Miller C.J.: Which craft is best in bioinformatics? *Computer Chemistry*, Vol. 25. (2001) 329-339.

[3] Breitling R., Amtmann A., Herzyk P.: IGA: A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *Bioinformatics*, Vol. 5. (2004) 34.

[4] Cho R., Campbell M., Winzeler E. et al.: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, Vol. 2. (1998) 65-73.

[5] Chuaqui R.: Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*, Vol. 32. (2002) 509-514.

[6] DeRisi J., Iyer L. et Brown V.: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, Vol. 278. (1997) 680-686.

[7] Draghici S., Khatri P., et al. (2003). Global functional profiling of gene expression. *Genomics*, 81:1-7.

[8] Eisen M., Spellman P., Brown P., Botsein D., et al.: Cluster analysis and display of genome wide expression patterns. Proc. Nat. Acad. Sci., 95 Vol. 25. (1998) 14863-8.

[9] Gibbons D., Roth F., et al.: Judging the quality of gene expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, Vol. 12. (2002)1574-1581.

[10] Hosack D., Dennis G., et al.. Identifying biological themes within lists of genes with EASE. *Genome Biology*, Vol. 4. (2003) R70.

[11] Kim S., Volsky D. et al.: PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, Vol. 6. (2005) 144.

[12] Little R. and Rubin D.,*Statistical Analysis with Missing Data*. John Wiley & Sons, New York. 2002.

[13] Masys D., et al.: Use of keyword hierarchies to interpret gene expressions patterns. *BMC Bioinformatics*, Vol. 17. (2001) 319-326.

[14] Mootha V., Lindgren C., Eriksson K., Subramanian A. et al.: PGC-l'alpha-reponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, Vol. 34(3). (2003) 267-273.

[15] Pasquier C., Girardot F., Jevardat K., Christen R.: THEA : Ontology-driven analysis of microarray data. *Bioinformatics*, Vol. 20(16). (2004).

[16] Quackenbush J.: Microarray data normalization and transformation. *Nature Genetics*, Vol. 32 (suppl.). (2002) 496-501.

[17] Riva A., Carpentier A., Torresani B., Henaut A.: Comments on selected fundamental aspects of microarray analysis. *Computational Biology and Chemistry*, Vol. 29. (2005) 319-336.

[18] Robinson M., et al.: FunSpec : a Web based cluster interpreter for yeast. *BMC Bioinformatics*, Vol. 3. (2002) 35.

[19] Simoff S., Maher, M.: Ontology-based multimedia data mining for design information retrieval. Computing in Civil Engineering, Proceedings of the International Computing Congress. VA: ASCE. (1998), 212-223.

[20] Storey J., Tibshirani R.: Statistical significance for genomewide studies. Proc. Natl. Acad. Sci., Vol. 100 (16). (2003) 9440-5.

[21] Sung G., Jung U., Yang K.: A graph theoretic modeling on GO space for biological interpretation of gene clusters. *BMC Bioinformatics*, Vol. 3. (2004) 381-386.

[22] Tamayo P., Slonim D., et al.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci., Vol. 96. (1999) 2907-2912.

[23] Tusher V., Tibshirani R., Chu G., et al.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci. USA, Vol. 98 (9). (2001) 5116-21.

[24] Weng S., Dong Q., Balakrishnan R., Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res*, Vol. 31. (2003), 216-218.