Interpreting Microarray Experiments Via Co-expressed Gene Groups Analysis (CGGA)

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

[1] Laboratoire I3S, 2000, route des lucioles,
06903 Sophia-Antipolis cedex, France
{rmartine, pasquier}@i3s.unice.fr
[2] Laboratoire Biologie Virtuelle, Centre de Biochimie, Parc 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

Abstract. 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 huge data using different sources of information.

We propose 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 (GO)[1].

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.[2]

1 Introduction

One of the main challenges in microarray data analysis is to highlight the principal functional gene groups using distinct sources of genomic information. These sources of information, constantly growing by an ever-increasingly volume of genomic data, are: semantic (taxonomies, thesaurus and ontologies), literature and bibliographic databases (articles, on-line libraries, etc.), experience databases (ArrayExpress, GEO, etc.) and nomenclature databases (HUGO: human, Flybase: fruit fly, SGD: yeast…).

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

N. Lavrač, L. Todorovski, and K.P. Jantke (Eds.): DS 2006, LNAI 4265, pp. 316 – 320, 2006.

Actually, one of the major goals in bioinformatics is the automatic integration of biological knowledge from distinct sources of information with genomic data [1]. A first assessment of the methods developed to answer this challenge was proposed by Chuaqui [3]. We target here the enrichment of two recently developed research axes, *sequential* and *a priori*, that exploit multiple sources of annotations such as GO.

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* [5], *Quality Tool* [6], *EASE* [7], *THEA* [11] and *Graph Modeling* [15].

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* [10], a *pc-value* [2], or a *z-score* test [8].

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* [5]. Finally, we obtain co-expressed and statistically significant FEG.

This article is organized in the following way: in section 2 we describe the validation data as well as the tools used: databases, ontologies and 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

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 [4]. DeRisi experience measures the variations in gene expression profiles during the cellular process of diauxic shift for the yeast *Saccharomyces Cerevisiae*. This process corresponds to the transition from fermentation to respiration that occurs when fermenting yeast cells, inoculated into a glucose-rich medium, turn to the utilization of the ethanol (aerobic respiration).

2.2 Ontology and Functionally Enriched Groups (FEG)

In order to fully exploit Gene Ontology (GO) we have generated: SGOD database. Our database contains all GO annotations for every yeast gene using Saccharomyces Genome Database (SGD)[3] nomenclature. We have stored all the functional annotations of each gene and his parents preserving the hierarchical structure of GO. Queries carried out on the SGOD database have built the whole set of the FEGs.

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

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* [13], which is more robust than other measurements such as *anova*, *fold change* or *t-student* statistics [13].

This measurement enables us to build a list of genes, *g-rank*, ordered by decreasing expression variability. We have used the SAM program [16] 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 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 SGOD. 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.

Let H_0 be the hypothesis that x genes from one of the subsets were related by chance. The probability that H_0 is true follows from the hypergeometric distribution[4]:

$$p(X = x \mid N, \mathrm{R}_{g(x)}, n) = \frac{\binom{R_{g(x)}}{x}\binom{N - R_{g(x)}}{n - x}}{\binom{N}{n}} \text{ where } \quad p(X = 0 \mid N, \mathrm{R}_{g(x)}, n) = 0\,, \tag{1}$$

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), $\mathrm{r}_{g(x)}$: absolute rank of the gene of position x in the *g-rank* list and $\mathrm{R}_{g(x)}$: number of ranks between the gene of position x from its FEG predecessor. $\mathrm{R}_{g(x)}$ is calculated from the absolute ranks $\mathrm{r}_{g(x)}$ according to the formula: $R_{g(x)} = \mathrm{r}_{g(x)} - \mathrm{r}_{g(x-1)} + 1$ where $R_{g(0)} = \mathrm{r}_{g(0)} = 1$.

The *pc-value* corresponding to this hypothesis test is [5]:

$$pc - value(x) = 1 - \sum_{k=1}^{x} p(X = k \mid N, R_{\mathrm{g(k)}}, n)\,. \tag{2}$$

In order to accept or reject H_0 we will use the following significance threshold: $p\text{–}value = Min\ \{N^{-1}, |\Omega|^{-1}\}$, where $|\Omega|$ is the cardinality of the set of functional annotations. So, for each FEG, if $pc\text{–}value(x) < p\text{–}value$ then H_0 is rejected, i.e. the FEG is statistically significant.

[4] For more details on the computation of this probability, refer to [17].

4 Results

In order to evaluate our method, we compared the results obtained by DeRisi [4], IGA [2] and CGGA. The results obtained using CGGA for the over-expressed and under-expressed genes are presented in Table 1. As expected, 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 underlined.

Table 1. Over-expressed FEGs obtained by CGGA with a *p-value* = 6.88E-04

Functionally Enriched GO Group	*n* genes	*x* over-exp. genes	*pc-value*
proton-transporting ATP synthase complex	2	2	4.38E-06
invasive growth (sensu Saccharomyces)	5	3	6.13E-06
signal transduction 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

In the case of over-expressed genes (Table 1), CGGA found seven of the nine groups obtained manually by DeRisi [4]. 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. Similar results, available at CGGA web page, were obtained for the under-expressed FEGs.

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 [5]-[7], [11], and [15], 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 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 (spelling or naming errors).

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. 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 pathway ontologies for future work.

References

1. Attwood T. and Miller C.J.: Which craft is best in bioinformatics? Computer Chemistry, Vol. 25. (2001) 329-339.
2. Breitling R., Amtmann A., Herzyk P.: IGA: A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics, Vol. 5. (2004) 34.
3. Chuaqui R.: Post-analysis follow-up and validation of microarray experiments. Nature Genetics, Vol. 32. (2002) 509-514.
4. DeRisi J., Iyer L. and Brown V.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, Vol. 278. (1997) 680-686.
5. Draghici S. et al.: Global functional profiling of gene expression. Genomics. (2003). 81:1-7.
6. 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.
7. Hosack D., Dennis G., et al.: Identifying biological themes within lists of genes with EASE. Genome Biology, Vol. 4. (2003) R70.
8. Kim S., Volsky D. et al.: PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics, Vol. 6. (2005) 144.
9. Masys D., et al.: Use of keyword hierarchies to interpret gene expressions patterns. BMC Bioinformatics, Vol. 17. (2001) 319-326.
10. Mootha V., et al.: PGC-l α-reponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, Vol. 34(3). (2003) 267-273.
11. Pasquier C., Girardot F., Jevardat K., Christen R.: THEA : Ontology-driven analysis of microarray data. Bioinformatics, Vol. 20(16). (2004).
12. Quackenbush J.: Microarray data normalization and transformation. Nature Genetics, Vol. 32 (suppl.). (2002) 496-501.
13. Riva A., Carpentier A., Torresani B., Henaut A.: Comments on selected fundamental aspects of microarray analysis. Computational Bio. and Chem., Vol. 29. (2005) 319-336.
14. Robinson M., et al.: FunSpec: a web based cluster interpreter for yeast. BMC Bioinformatics, Vol. 3. (2002) 35.
15. 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.
16. 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.
17. Martinez R., et al.: CGGA: An automatic tool for the interpretation of gene expression experiments. Accepted (to appear) on the Journal of Integrative Bioinformatics. 2006.