Minguez et al. 2006 Genome Informatics Series 17 : 57-66

A function-centric approach to the biological interpretation of microarray time-series

Pablo Minguez [1] **Fátima Al-Shahrour** [1]
pminguez@cipf.es falshahrour@cipf.es

Joaquín Dopazo[1,2]
jdopazo@cipf.es

[1] Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Autopista del Saler 16, E46013, Valencia, Spain
[2] Functional Genomics Node, INB-CIPF, Autopista del Saler 16, E-46013, Valencia, Spain

Abstract

The interpretation of microarray experiments is commonly addressed by means a two-step approach in which the relevant genes are firstly selected uniquely on the basis of their experimental values (ignoring their coordinate behaviors) and in a second step their functional properties are studied to hypothesize about the biological roles they are fulfilling in the cell. Recently, different methods (e.g. GSEA or FatiScan) have been proposed to study the coordinate behavior of blocks of functionally-related genes. These methods study the distribution of functional information across lists of genes ranked according their different experimental values in a static situation, such as the comparison between two classes (e.g. healthy controls versus diseased cases). Nevertheless there is no an equivalent way of studying a dynamic situation from a functional point of view.

We present a method for the functional analysis of microarrays series in which the experiments display autocorrelation between successive points (e.g. time series, dose-response experiments, etc.) The method allows to recover the dynamics of the molecular roles fulfilled by the genes along the series which provides a novel approach to functional interpretation of such experiments. The method finds blocks of functionally-related genes which are significantly and coordinately over-expressed at different points of the series. This method draws inspiration from systems biology given that the analysis does not focus on individual properties of genes but on collective behaving blocks of functionally-related genes.

The FatiScan algorithm used in the method proposed is available at: http://fatiscan.bioinfo.cipf.es, or within the Babelomics suite: http://www.babelomics.org. Additional material is available at: http://bioinfo.cipf.es/data/plasmodium

Keywords: time series, functional interpretation, gene ontology, *Plasmodium falciparum*

1 Introduction

DNA microarray technology has been extensively used to obtain snapshots of the expression of genes in different samples, tissues, experimental conditions, etc. While typical microarray assays are designed to study static experimental conditions, there is a class of experiments, time series, in which a temporal process is measured. Time series offer the possibility of identifying the dynamics of gene activation, which allows to infer causal relationships. Such relation-ships can be used to infer models of regulatory networks [1] either directly [2] or through the identification of activators and repressors [3].

An important difference between these two types of experiments is that while static data sampled from a population (e.g. diseased cases, healthy controls, etc.) are assumed to be independent, time series data are characterized by displaying a strong autocorrelation between successive points [4]. Initially, time series were analyzed using methods originally developed for independent data points [5-7]. More recently, algorithms were developed to specifically address this type of data. Data analysis now address issues such as the

alignment of temporal data sets [8, 9] and the identification of differentially expressed genes [10]. Also different clustering methods specific for time series have been recently proposed. Among these, clustering based on the dynamics of the expression patterns [11], clustering using a hidden Markov model [12] or clustering specifically devised for short time series [13].

The final aim of a typical microarray experiment is to find a molecular explanation for a given macroscopic observation, which in the case of time series is the dynamic behavior of a system such as a cell cycle [6, 14], responses to temperature changes or stresses [15], immune response [16], etc. Commonly, a two-step approach is used for the analysis and further functional interpretation of such experiments. In it, genes showing an "interesting" behavior (differentially expressed, clustered, etc.) are firstly selected, usually ignoring the fact that these genes are acting cooperatively in the cell and consequently their behaviors must be coupled to some extent [17]. In this selection process, stringent thresholds to reduce the false positives ratio in the results are usually imposed. In a second step, the selected genes of interest are compared to the background in order to find enrichment in any functional term. Several popular programs, such as FatiGO [18], Ontoexpress [19], and others [20], use this schema. Under a systems biology perspective this procedure is far away from being optimal because much information is lost due to the fact that a large number of true positives is lost as false negatives in order to reduce in the possible the number of false positives. Methods inspired in systems biology can use lists of genes ranked by differential expression and directly search for the distribution of blocks of functionally related genes across the list without imposing any threshold that ignore the cooperative behavior of such blocks. Recently, a family of threshold-free methods devised to find groups of functionally related genes with a coordinate over- or under-expression across a list of genes ranked by differential expression have recently been proposed [21-24].

Here we present a conceptually different approach in which the aim is to detect the biological processes operating along a time series by using a threshold-free approach [21]. We applied the method to a time series obtained for the intraerythocytic developmental cycle of the parasite *Plasmodium falciparum* [14]. The results obtained are not simple lists of genes, but the pattern of temporal activations and deactivations of the different biological roles that shape the developmental cycle of the parasite.

2 Data

The data used is a microarray time series experiment of *Plasmodium falciparum* during the 48 hours of the intraerythrocytic developmental cycle (IDC) with samples taken every hour [14]. The dataset consists of 7462 probes (70mer oligonucleotides), representing 4488 of the 5409 ORFs of the *Plasmodium falciparum* strain 3D7. A total of 46 time points (covering 48 hours, sampling time points in intervals of one hour; time points 23 and 29 had no data in the original dataset) were used in the study.

The asexual blood stage of the parasite *Plasmodium falciparum* causes the pathogenesis of the parasite in human, therefore understanding its gene expression profile is crucial for drug discovery and vaccine design. In blood, the parasite undergoes a 48 hours cycle characterized by three developmental forms, Ring (1-17 time points, being 1 time point equal to 1 hour), Trophozite (18-29 time points) and Schizont (30-48 time points). The mature Schizont suffers an asexual division to form up to 32 merozoites that are released to blood to invade new erythrocytes, this produces a crisis known as malaria fever.

3 Methods

3.1 Microarray data preparation

Microarray data used were the log-ratios of the normalized values of expression (see [14] for details). The data were arranged into a matrix of gene expression values where columns represent time points and rows represent genes. Here, time point 1 (first column) was taken as reference for the analysis. Genes with no information in this column were removed from the analysis. Then, a transformed matrix containing as many columns as time points minus one (the initial time point) was obtained by subtracting each time point from the reference time point (both are log ratios). Each column so obtained accounts for the relative differences in expression of each gene with respect to its original expression value at time 1 (or in other words, the log ratios of the gene expression values with respect to their respective value in the initial time).

3.2 Functional analysis using the FatiScan, a threshold-free method

The aim of the analysis is to find biological roles (according to GO annotations) that are activated or deactivated across the time series. To this end, all the columns were analyzed to detect blocks of functionally-related genes constitutively over- or under-expressed with respect to the initial condition.
The FatiScan method [21, 25] is a segmentation, threshold-free test for detecting the biological roles fulfilled by the genes across an ordered list of gene expression values. The aim of the test is to detect, across an ordered list of genes, groups that are cumulated in one of the extremes of the list (see Figure 1C). The biological meaning is as follows: if the genes are ordered by differential expression between two conditions, a block of genes on the top would point to its collective over-expression (or under-expression if in the bottom of the list). The blocks can be formed using the GO annotations, so we would be able of finding the collective behavior of blocks of genes functionally related. In particular, the method consists on the sequential application of an Fisher's exact test to consecutive partitions of an ordered list of genes to detect asymmetrical distributions of GO annotations in the genes at both sides of the partition. As previously mentioned, finding such asymmetry would point to an over-representation (or under-representation) of a GO term in one of the sides of the list. If the list represents the differences in gene expression observed in time t with respect to the initial time, this particular GO term can be considered to be constitutively and significantly activated at time t.
The significance of this asymmetry is obtained through a Fisher's exact test (one-tailed in this case) applied over a contingency table. This will detect significantly over or under represented GO terms when comparing the upper side to the lower side of the list, as defined by any partition. The number of partitions used was of 50, which was previously shown that produces optimal results in terms of sensitivity and results recovered [21]. Multiple testing effect due to the massive testing and assignation of biological terms was corrected by the widely accepted FDR [26]. The FatiScan applied to a list of genes ranked by expression level will detect blocks of functionally related genes constitutively over- or under-expressed. In this study FatiScan was used to search for significant GO terms from to the three main categories (biological process, molecular function and cellular component) at different levels of the GO hierarchy (levels 3, 4, 5, 6 and 7).

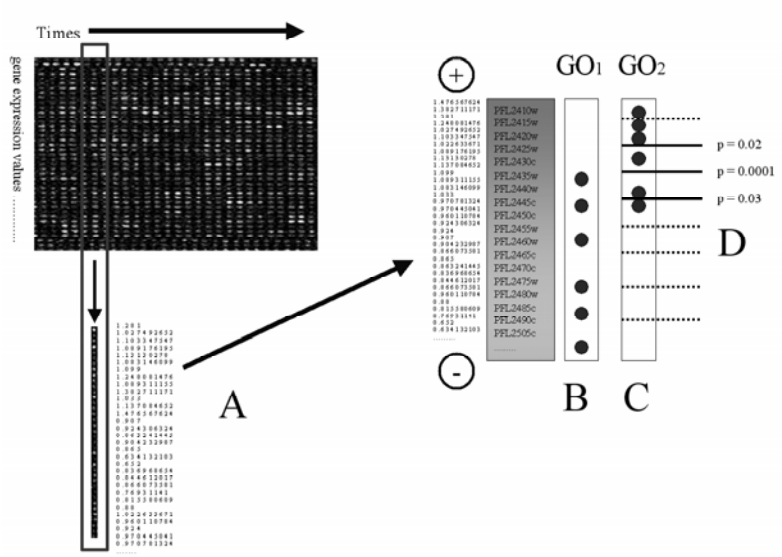

Figure 1: Procedure for the functional annotation of a time series. See text for an explanation.

3.3 Functional analysis of the time series.

Each column, corresponding to any time point beyond the initial time, was independently ordered from higher to lower values of relative expression (log ratios of the expression with respect to time 1). The positions occupied by the genes in the ranking previously obtained for each column are related to their relative contribution to the biological processes operating in this particular time point. A FatiScan analysis of

each time point will render the GO terms whose corresponding genes displayed a significant coordinated high expression. These GO terms provide a detailed view on the biological roles active at the time points at which they have been detected. The dynamics of these functional roles within the cell can be understood by plotting the GO terms significantly over-expressed along the time axis.
Figure 1 illustrates the procedure followed. Genes at each time point are ranked from highest (top) to lowest (bottom) relative expression with respect to time 1 (Figure 1A). Then, for each list of ranked genes generated in any time point, the significant over-represented GO terms in the tail corresponding to the highest expression values are recorded. Figure 1B shows a GO term not related to high expression at this time point. Conversely, the GO term in Figure 1 C is significantly over-represented in high expression values. The partitions used to decide that a given term is significantly over-represented in the upper tail of the list with respect to the lower part are used for the graphical representation. The proportion of genes annotated with a significant GO term in the most significant partition is finally plotted in the graphical representation of the GO dynamics. In the example in Figure 1 D, the most significant partition, with p=0.001, captures the maximum divergence according to the test (with 4 out of a total of 6 terms), which would correspond to a value of 66.67%.
The way in which the lists are ordered determines the hypothesis to be tested. In this case we are testing over-expressions with respect to a initial situation (point t=1), but other arrangements are possible (e.g. any point with respect to the previous one, etc.)

4 Results and discussion

Plasmodium falciparum life cycle is highly complex, involving two different hosts (mosquito, human), different tissues (liver, blood and mosquito), intra and extra cellular location and three developmental stages. Many attempts have been made to explain the gene expression profile of this the intraerythrocytic developmental cycle of the parasite due to its implication in human health [14, 27].

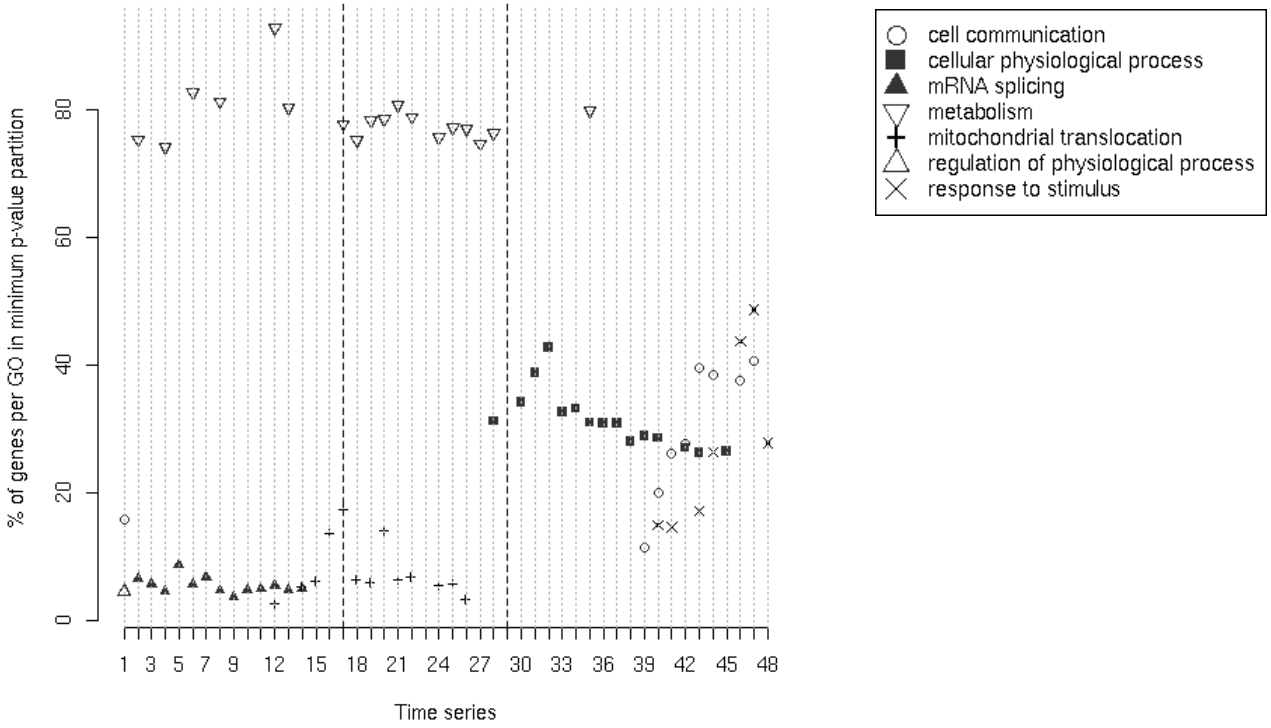

Figure 2. Dynamics of the biological process GO category at level 3 along the intraerythrocytic developmental cycle of the parasite. The blue vertical lines mark the transition between ring to thropozyte and from this stage to schizont, respectively.

Here we take a new approach based on the direct analysis of the dynamics of the biological roles fulfilled by the genes, as described by the GO categories.
More than 40 GO terms over-represented at high expression values were obtained for some GO categories and levels, which constitute an unprecedented wealth of information on the functional behavior at molecular level of *P. falciparum* blood cycle across time. Results for each ontology class may give different type of information. Thus, plots of GO cellular component terms show how the parasite activity moves from some cellular compartments to others (for example from nucleus in the initial stages of the cycle to membrane and host in the final stage) while molecular function and biological processes GO categories are related to biological roles of different nature coordinately played by the genes in the cell
Here we provided only a summarized discussion of several aspects of the dynamics of the biological roles and subcellular locations at which the genes are carrying out their activities. A more detailed discussion is beyond the scope of this paper.

4.1 Dynamics of the biological roles along the cell cycle

The plots of the different GO terms found as significantly over-represented at different times clearly illustrates the dynamics of the different roles and how the cell carries out a sequence of functional steps during its life cycle.
Figure 2 illustrates how different biological roles switch on and off along the time points. These biological processes account for the molecular events that govern the transitions between the developmental stages of the intraerythrocytic cycle of the parasite. Just to cite a few examples, GO terms related to metabolism are found in early staged of the cycle, whereas GO terms related to signaling occur preferentially at the end, in the schizont stage.

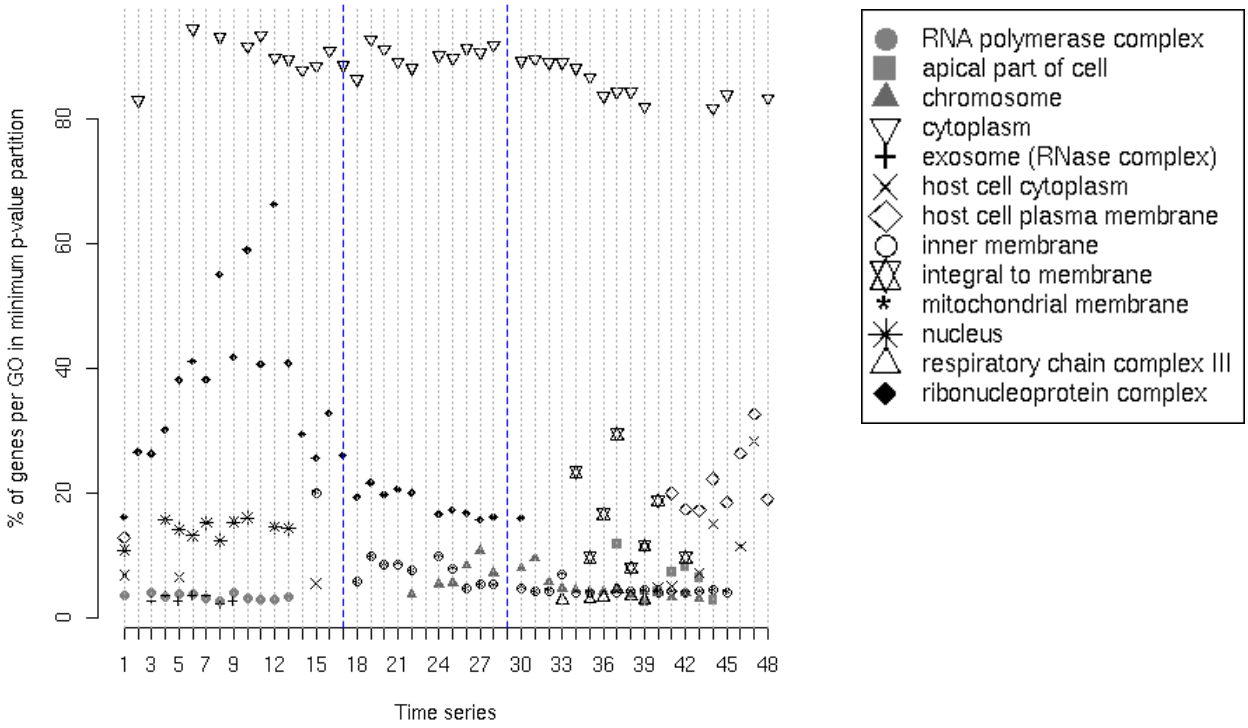

Figure 3. Dynamics of the subcellular location GO category at level 4 along the intraerythrocytic developmental cycle of the parasite. The blue vertical lines mark the transition between ring to thropozyte and from this stage to schizont, respectively

Figure 3 gives information on the location, at subcellular level, where the above mentioned biological roles are taking place. It is very illustrative the fact that during the initial stages much activity occurs around

locations related to replication (RNA polymerase complex, nucleus, ribonucleoprotein complex), while in the later stages terms related to invasion and with the interaction with the host cells are found (host cell cytoplasm, host cell plasma membrane).
In the next section a more detailed explanation of some terms in relation to the stage of the cycle in which they appear is provided. Also, in the additional information web page the three GO categories at different levels can be found. More precise terms, descendant in the GO hierarchy of the terms shown in the figures, account with more precision for the biology of the parasite and can be found in the web page of additional information in http://bioinfo.cipf.es/data/plasmodium. Additionally, this page contains links to the genes in the GO categories found as significantly over-expressed.

4.2 Biological roles in the different developmental stages

A summary of the GO terms (not all but the ones that are in consonance with previous findings as well as other relevant in this context) over-represented during the *Plasmodium* Intraerythrocytic Cycle follows. A complete list of the GO terms found is available in the additional material pages.

4.2.1 Ring and early-Trophozoite:

Basic metabolism is on the top of the cell activities during this part of the cycle in which the parasite is starting the maturation process. This is reflected by a high gene activity, firstly related to transcription and then to translation. Some GO terms found as over-represented at this stage that support these evidences are: *transcription, DNA-dependent, mRNA splicing, RNA metabolism, transcription, RNA modification, RNA metabolism, nucleotide metabolism, nucleus, translation, biosynthesis, protein biosynthesis, amino acid activation, amino acid metabolism, tRNA metabolism, regulation of protein biosynthesis, regulation of translation, regulation of metabolism, organic acid metabolism, RNA polymerase complex* and *tRNA and amino acid metabolism*.
There is also an over-representation of terms that express the increase of the ribonucleotide biosynthesis such as: *pyrimidine base metabolism* and *pyridine nucleotide metabolism*. The presence of a high metabolic activity can be deduced from the over-representation of *metabolism* and other related terms like *macromolecule biosynthesis* and *protein biosynthesis*, all with a high number of genes involved. The high representation of *protein biosynthesis* could explain the over-representation of the *proteasome complex* term at this stage (proteosome as a quality control system). Interestingly, we found in time point 15 over-represented the terms *host and host cell cytoplasm*. The major stage of interaction with host comes later on in the cycle.

4.2.2 Trophozoite and early-Schizont:

We have found terms associated to DNA metabolism significantly over-represented in this developmental stage: *DNA metabolism, DNA replication, DNA replication factor, replisome, replication fork* and *chromosome*. Although, the initial analysis of the microarray data suggested that DNA metabolism was active from the beginning of the cycle until the early schizont stage, our results seems to extend the importance of this process along almost all the schizont stage (evidences in term *DNA replication* -Biological process category- and *replisome* -Cellular component category-).
Metabolism seems to be also very important in this developmental stage as shown by terms such as: *carboxylic acid metabolism, protein biosynthesis* and *macromolecular metabolism*. Regarding subcellular location terms, cell activity has an important location in mitochondria, as shown by the terms: *mitochondrion, mitochondrial membrane, mitochondrial inner membrane, ion transport* and *respiratory chain complex III*. This, together with the over-representation of the terms *plastid* and *apicoplast*, supports the theory that in this period translation activity moves from nucleus to plastid and mitochondria.

4.2.3 Schizont:

GO terms associated to protein catabolism become significant: *protein catabolism, proteolysis and peptidolysis, proteasome complex* and *macromolecule catabolism*. Other terms as DNA replication and cell proliferation, indicate that cells are in a high division activity stage. At the end of this period the main

intracellular activities of the parasite are located close to membrane (as indicated by-terms such as *membrane, host cell membrane* and *infected host cell surface knob*) and much more plasmodium specific terms associated to invasion and interaction with host become important (*cell communication, cell-cell adhesion, response to stimulus, response to biotic stimulus, response to external stimulus, heterophilic cell adhesion, defense response, evasion of host defense response, host pathogen interaction* and *cytoadherence to microvasculature*)
More specific terms are in consonance to previous analysis saying that proteases, kinases and actin-myosin motors have an important role in invasion (*kinase activity, myosin, actin cytoskeleton* and *peptidase activity*) The activity also increases in the plastid genome.

4.2.4 Early Ring:

Invasion-specific biological activities still remain significant in the earliest hours of Ring stage as shown by GO terms such as: evasion of host cell response and host cell plasma membrane. In the original paper [14] the authors alert on the possibility of some contamination from the previous stages affecting to the ring stage, so the results obtained for this stage must be taken carefully. Terms such as *cell communication, cell invasion, response to stimulus, cell adhesion* and other similar become visible already in the very first time points of the analysis.

5 Conclusions

As we learn more on the functional basis of the cooperative behaviors of groups of genes, systems biology approaches gain more importance in our attempt to decipher cell biology relevant questions. Following this, the interpretation of genome-scale experiments is starting to focus more in groups of functionally-related genes than on the properties of individual genes. Recently, different procedures that make use of functional annotations for the direct selection of functionally-related groups of genes have recently been proposed in the context of microarray data [21-24, 28]. These procedures use a list of genes ranked by differential gene expression and, without imposing any threshold based on the experimental values, study global over- or under-expression of blocks of functionally related genes. Nevertheless, its application has been restricted to static experimental designs. Here we show how to expand this concept to the functional analysis of a time series. This analysis gives dynamic information on continuous behaviors occurring across the series of experiments analyzed. By means of the procedure presented in this paper it is easy to understand the sequence of functional events taking place in particular moments of the period studied. It is important to remark that, for the first time, we have directly addressed the temporal evolution of the biological roles fulfilled by the genes, and not the behavior of individual genes, which might (and actually do) contribute to more than one functional category.
As previously mentioned, the use of different GO categories (or other functional terms) allow explaining different aspects of the biology of the cell (e.g. the biological roles fulfilled by the genes by using the biological process GO category, or where these roles took place, by using the subcellular location GO category, etc.). The proposed methodology addresses systems biology-inspired questions on the behavior of groups of functionally-related genes. The method provides results with a statistical support. The so obtained significance p-values make reference to the functional blocks of genes, but not necessarily to individual genes. The method can be applied using any kind of gene annotation beyond the GO terms here used, such as KEGG pathways, Interpro domains, etc.
Summarizing, we have presented a method for the functional analysis of microarrays series with some dependence of autocorrelation (time series, dosage series, etc.) The method allows to recover the dynamics of the significantly over-represented functional terms along the series. A proper understanding of the biology of the cell from the perspective of systems biology need of approaches like the presented here, which tackle global functional properties cooperatively carried out by groups of genes.
The method can easily be applied with a web-based tool, the FatiScan (available at http://www.babelomics.org).

Acknowledgements

This work is supported by grants from Fundació La Caixa, MEC BIO2005-01078 and NRC Canada-SEPOCT Spain. The Functional Genomics node (INB) is supported by Fundación Genoma España.

References

1. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al*: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**(11):1337-1342.
2. Herrero J, Díaz-Uriarte R, Dopazo J: **An approach to inferring transcriptional regulation among genes from large-scale expression data**. *Comparative and Functional Genomics* 2003, **4**:148-154.
3. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**(7006):308-312.
4. Bar-Joseph Z: **Analyzing time series gene expression data**. *Bioinformatics* 2004, **20**(16):2493-2503.
5. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**(3-4):601-620.
6. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. *Mol Biol Cell* 1998, **9**(12):3273-3297.
7. Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, Futcher B: **Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth**. *Nature* 2000, **406**(6791):90-94.
8. Aach J, Church GM: **Aligning gene expression time series with time warping algorithms**. *Bioinformatics* 2001, **17**(6):495-508.
9. Liu X, Muller HG: **Modes and clustering for time-warped gene expression profile data**. *Bioinformatics* 2003, **19**(15):1937-1944.
10. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS: **Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes**. *Proc Natl Acad Sci U S A* 2003, **100**(18):10146-10151.
11. Ramoni MF, Sebastiani P, Kohane IS: **Cluster analysis of gene expression dynamics**. *Proc Natl Acad Sci U S A* 2002, **99**(14):9121-9126.
12. Schliep A, Schonhuth A, Steinhoff C: **Using hidden Markov models to analyze gene expression time course data**. *Bioinformatics* 2003, **19 Suppl 1**:i255-263.
13. Ernst J, Nau GJ, Bar-Joseph Z: **Clustering short time series gene expression data**. *Bioinformatics* 2005, **21 Suppl 1**:i159-i168.
14. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum**. *PLoS Biol* 2003, **1**(1):E5.
15. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes**. *Mol Biol Cell* 2000, **11**(12):4241-4257.

16. Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES, Young RA: **Human macrophage activation programs induced by bacterial pathogens**. *Proc Natl Acad Sci U S A* 2002, **99**(3):1503-1508.
17. Wolfe CJ, Kohane IS, Butte AJ: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks**. *BMC Bioinformatics* 2005, **6**:227.
18. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. *Bioinformatics* 2004, **20**(4):578-580.
19. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression**. *Genomics* 2003, **81**(2):98-104.
20. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* 2005, **21**(18):3587-3595.
21. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information**. *Bioinformatics* 2005, **21**(13):2988-2993.
22. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 2004, **20**(1):93-99.
23. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment**. *BMC Bioinformatics* 2005, **6**:144.
24. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.
25. Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J: **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W472-476.
26. Benjamini Y, Yekutieli D: **The control of false discovery rate in multiple testing under dependency**. *Annals of Statistics* 2001, **29**:1165-1188.
27. Ben Mamoun C, Gluzman IY, Hott C, MacMillan SK, Amarakone AS, Anderson DL, Carlton JM, Dame JB, Chakrabarti D, Martin RK *et al*: **Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite Plasmodium falciparum revealed by microarray analysis**. *Mol Microbiol* 2001, **39**(1):26-36.
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.