OMICS A Journal of Integrative Biology
Volume 10, Number 3, 2006

Functional Interpretation of Microarray Experiments

JOAQUÍN DOPAZO

ABSTRACT

Over the past few years, due to the popularisation of high-throughput methodologies such as DNA microarrays, the possibility of obtaining experimental data has increased significantly. Nevertheless, the interpretation of the results, which involves translating these data into useful biological knowledge, still remains a challenge. The methods and strategies used for this interpretation are in continuous evolution and new proposals are constantly arising. Initially, a two-step approach was used in which genes of interest were initially selected, based on thresholds that consider only experimental values, and then in a second, independent step the enrichment of these genes in biologically relevant terms, was analysed. For different reasons, these methods are relatively poor in terms of performance and a new generation of procedures, which draw inspiration from systems biology criteria, are currently under development. Such procedures, aim to directly test the behaviour of blocks of functionally related genes, instead of focusing on single genes.

INTRODUCTION

An increasing corpus of evidence reveals that genes do not operate alone within the cell, but in an intricate network of interactions that we only recently start to envisage (Hallikas et al., 2006; Rual et al., 2005; Stelzl et al., 2005). It is widely accepted that co-expressing genes tend to fulfil common roles in the cell (Lee et al., 2004; Stuart et al., 2003), and in fact, this causal relationship has been used to predict gene function from patterns of co-expression (Mateos et al., 2002; van Noort et al., 2003). In this scenario, there is a clear necessity for methods and tools to aid the interpretation of large-scale experiments such as microarrays, and to formulate genome-scale hypothesis from a systems biology perspective (Westerhoff and Palsson, 2004), in such a way that the collective properties of groups of genes are taken into account.

Among the genome-scale experimental methodologies DNA microarray technology can be considered the paradigm due to its popularity and characteristics. Although many different biological questions can be addressed though microarray experiments, there are usually three types of objective in this context: "class comparison," "class prediction," and "class discovery" (Allison et al., 2006; Golub et al., 1999). The first two objectives fall into the category of supervised methods and usually involve the application of tests to define differentially expressed genes, or the use of different procedures to predict class membership on the basis of the values observed for a number of "key" genes. Clustering methods belong to the last category, also known as unsupervised analysis, because no previous information about the class structure of the data set is used in the study.

Department of Bioinformatics, and Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.

Interestingly, if strategies for microarray data analysis are considered from a historical perspective, an initial period can be distinguished in which almost all publications were related to reproducibility and sensitivity issues. Thus, most early microarray papers dating from the late nineties were proof-of-principle experiments (Eisen et al., 1998; Perou et al., 1999), in which only cluster analysis was applied in order to check whether differences at gene expression level could reproduce macroscopic observations. Later, as the focus moved towards finding genes differentially expressed among experimental conditions specificity became a main concern. It soon became obvious that genome-scale experiments needed to be carefully analysed, as when large amounts of data were explored many apparent findings occurred by mere chance (Ge et al., 2003), especially if liberal procedures, such as the fold change were used for biomarker selection purposes. In this context, different statistical tests along with methods for the adjustment of *p*-values (which are considered standard today) started to be extensively used (Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2003). More recently, the use of microarrays for building predictive models of clinical outcomes (van 't Veer et al., 2002), although not free from criticism (Simon, 2005), has fuelled the use of this technology because of its practical clinical implications.

While intra-platform reproducibility is nowadays high (Moreau et al., 2003), there are still some concerns with the cross-platform coherence of results. Recent studies show that, despite the overlap between genes with significant differential expression found across platforms is poor, the biological roles these genes fulfill are highly consistent (Bammler et al., 2005). This fact clearly highlights the importance of the interpretation of experiments in terms of their biological implications rather than restricting them to a mere comparison of lists of gene identifiers (Al-Shahrour and Dopazo, 2005; Al-Shahrour et al., 2005b).

The extensive use of microarray technology has brought about the need to develop functional annotation tools that essentially look for functional terms over-represented in groups of genes defined by the experimental values. Examples of widely used terms with functional meaning are Gene Ontology (GO) (Ashburner et al., 2000), KEGG pathways (Kanehisa et al., 2004), CisRed motifs (Robertson et al., 2006), predictions of transcription factor binding sites (Wingender et al., 2000), Interpro motifs (Mulder et al., 2005), etc. Programmes such as ontoexpress (Draghici et al., 2003), FatiGO (Al-Shahrour et al., 2004), GOMiner (Zeeberg et al., 2003), etc., can be considered representatives of a family of methods that use these terms to find clues for the interpretation of the results of microarray experiments (Khatri and Draghici, 2005). These methods are used *a posteriori* on the genes of interest that were previously selected in the first step. Typical criteria for selection are differential expression (class comparison) and co-expression (class discovery), for example. By means of this simple two-step approach, a reasonable biological interpretation of a microarray experiment can be attained. Nevertheless, this approach has a weak point: the list of genes of interest, which is generally incomplete. This is due to the fact that the definition of this list is affected by many factors including, among others, the method of analysis and the threshold imposed. In the case of class discovery analysis, biological annotations have also been employed as a cluster validation criteria (Bolshakova et al., 2005).

The difficulties for defining repeatable lists of genes of interest across laboratories and platforms even using common experimental and statistical methods (Bammler et al., 2005; Qiu et al., 2006) has led several groups to propose different approaches which aim to select genes taking into account their functional properties. The Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005), although not criticism-free (Damian and Gorfine, 2004), has pioneered a family of methods devised not to find individual genes but to search for groups of functionally related genes with a coordinate (although not necessarily high) over- or under-expression across a list of genes ranked by differential expression between two classes, compared in microarray experiments. Different tests have recently been proposed for microarray data, with this aim in mind (Al-Shahrour et al., 2005a; Goeman et al., 2005, 2004; Kim and Volsky, 2005; Smid and Dorssers, 2004; Tian et al., 2005) and also for ESTs (Chen et al., 2006) and some of them are available on web servers (Al-Shahrour et al., 2005a,b). In particular, the FatiScan procedure (Al-Shahrour et al., 2006, 2005b), which implements a segmentation test (Al-Shahrour et al., 2005a), can deal with ordered lists of genes independently from the type of data that originated them. This interesting property allows for its application to other types of data apart from microarrays. Another recent development is the use of biological information (Huang and Pan, 2006; Pan, 2006) or phenotypic information (Jia and Xu, 2005) as a constitutive part of clustering algorithms in the case of class discovery (clustering) analysis.

METHODS

Threshold-based functional analysis

The final aim of a typical microarray experiment is to find a molecular explanation for a given macroscopic observation (e.g., which pathways are affected by the deprivation of glucose in a cell, what biological processes differentiate a healthy control from a diseased case). In the first generation of approaches proposed, the interpretation of microarray data is usually performed in two steps: in a first step genes of interest are selected (because they co-express in a cluster or they are significantly over- or under-expressed when two classes of experiments are compared) using different procedures, whose description is beyond the scope of this manuscript (Azuaje and Dopazo, 2005; Draghici and Kuklin, 2003). The selection process does not take into account the fact that these genes are acting cooperatively in the cell and consequently their behaviour must be coupled to some extent. In this selection process, under the unrealistic simplification of independence among gene behaviours, 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 with a background (typically the rest of the genes) in order to find enrichment in any functional term. This comparison with the background is required otherwise the significance of a proportion (even if high) cannot be determined. This comparison to the background is essential because sometimes apparent high enrichment in a given functional term is nothing but a reflect of a high proportion of this particular term in the whole genome and, consequently, has nothing to do with the set of genes of interest. The procedure for the interpretation of genes selected by significant differential expression among two pre-defined classes of experiments in illustrated in Figure 1A. Similarly, Figure 1B shows the two-step procedure applied to co-expressing genes found by clustering.

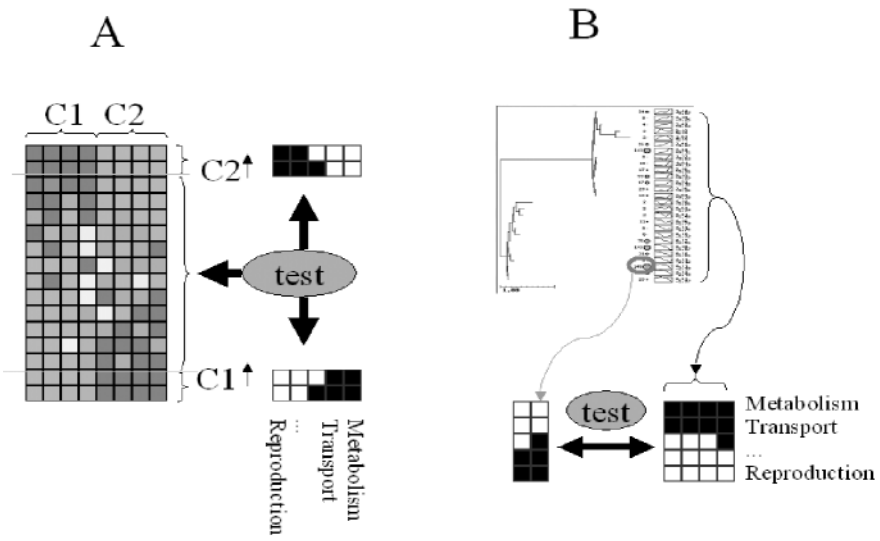

FIG. 1. Two-steps procedure for the functional interpretation of distinct microarray experiments. (**A**) The supervised approach: functional annotation of genes differentially expressed among two classes (C1 and C2) of experiments. The figure represents a list of genes (rows) ordered by differential expression when classes C1 and C2 are compared. Genes on the top are more expressed on class C1 (red color) than in C2. Conversely, genes on the bottom are more expressed on class C2. There is a gradient of differential expression between the two extreme situations. Typically genes are arranged by means of a test (e.g., a *t*-test) and those with a value of the statistic over a given threshold are declared as significant (right part). Then the distribution of functional terms (e.g., GO terms) among the differentially expressed genes and the rest is compared by means of another test (e.g., Fisher's exact test). (**B**) The unsupervised approach: functional interpretation of clusters of co-expressed genes. The functional terms found in the set of genes of interest (clusters of co-expressed genes) are compared to the background: the rest of genes.

This second step of comparison between the selected genes and the background can be carried out by means of the application of other, equivalent tests such as the hypergeometric, χ^2, binomial, Fisher's exact test, etc., implemented in different available tools, reviewed in (Khatri and Draghici, 2005). Among these tools, the most popular ones (most quoted in literature) are Onto-express (Draghici et al., 2003) (http://vortex.cs.wayne.edu/ontoexpress/) and FatiGO (Al-Shahrour et al., 2004) (www.fatigo.org). These tools use different biological terms with functional meaning such as GO (Ashburner et al., 2000), KEGG pathways (Kanehisa et al., 2004), and other terms of biological relevance. Table 1 shows a list of the most popular tools for analysing enrichment in biologically relevant terms. Surprisingly, there are still some tools that do not take into account widely established procedures such as multiple testing corrections (Table 1). As a general comment, FDR-based multiple testing adjustments are less conservative than Bonferroni or Sidak counterparts. Thus the package Babelomics (Al-Shahrour et al., 2006; Al-Shahrour et al., 2005b), which includes FatiGO (Al-Shahrour et al., 2004), and the Onto Tools (Draghici et al., 2003; Khatri et al., 2005) would be optimal in terms of biological information content and testing strategies. DAVID/Ease (Dennis et al., 2003; Hosack et al., 2003), FunSpec (Robinson et al., 2002), only for yeast, and GeneMerge (Castillo-Davis and Hartl, 2003) would be attractive from the point of view of the biological information although a bit conservative in terms of multiple testing correction. On the other hand, BayGO (Vencio et al., 2006), GOMiner (Zeeberg et al., 2003; Zeeberg et al., 2005), GOstat (Beissbarth and Speed, 2004), GOSurfer (Zhong et al., 2004) and Ontology Traverser (Young et al., 2005) use proper multiple-testing corrections although only provide GO terms for the annotation of the experiments. Other tools such as GO:TermFinder (Boyle et al., 2004) only provide GO and are conservative in the multiple testing adjustment or even fail to provide such adjustment, like FuncAssociate (Berriz et al., 2003), GOTM (Zhang et al., 2004) or CLENCH (Shah and Fedoroff, 2004).

This two-step approach is the natural choice for analysing clusters of genes, and is implemented in some packages, such as the GEPAS (Herrero et al., 2003, 2004; Vaquerizas et al., 2005), where clusters of genes obtained can be directly transferred to FatiGO (Al-Shahrour et al., 2004) for their functional analysis. Nevertheless, the application of a two-step strategy to the interpretation of differential gene expression in class comparison experiments causes an enormous loss of information as a large number of false negatives is accepted in order to preserve a low ratio of false positives (and the noisier the data the worse the effect).

Threshold-free functional analysis

The interpretation of a microarray experiment using the two-step approach is far from efficient when considered from a systems biology perspective. Methods directly inspired in systems biology focus on collective properties of the genes more than on individual gene expression values. Functionally related genes simultaneously fulfil their roles in the cell and, consequently, they are expected to display a coordinated expression. It is a long recognized fact that genes with similar overall expression often share similar functions (Eisen et al., 1998; Lee et al., 2004; Wolfe et al., 2005). This observation is consistent with the hypothesis of modularly-behaving gene programmes, where sets of genes are activated in a coordinated way to carry out functions. In this scenario, a different type of inference can be made based on testing hypothesis centred on blocks of functionally related genes, instead of testing one gene at a time.

Thus, genes can be ranked by using their differential expression values when comparing predefined classes (e.g., cases and healthy controls) by means of any appropriate statistical test (e.g., *t*-test). The order of the genes (that cooperatively act to define pathways, functional classes) in this ranked list must be related to its participation in the trait studied in the experiment. Consequently, each functional class "responsible" for the differences between the classes will be found in the extremes of the ranking with highest probability. Under this perspective the previous imposition of a threshold based on the rank values, which does not take into account the cooperative behaviour of the genes, is thus avoided. Figure 2 illustrates the threshold-free analysis strategy. Genes are arranged by differential expression between the classes N (normal) and T (test). On the right-hand side of the figure, there are labels for two different functional categories at the points in the list where genes fulfilling the corresponding roles are situated. Functional category A is completely unrelated with the experiment because different genes, belonging to this functional category, appear over-expressed in classes N and T and also in intermediate positions. Conversely, functional category B is predominantly fulfilled by genes with high expression in class N (red values corresponding to highest

Table 1. Compilation of Tools for Functional Annotation of Sets of Genes

Tool	*Statistical model*	*Correction for multiple experiments*	*Functional labels*	*Site (web-based applications)*	*Reference*
Babelomics	Fisher's exact test, *t*-test, Kolmogorov-Smimov	FDR, *q*-value	GO, KEGG, protein domains, swissprot keywords, Transfac motifs, CisRed motifs, chromosomal location, tissues, bioentities (text-mining)	http://www.babelomics.org	(Al-Shahrour et al., 2006; Al-Shahrour et al., 2005)
BayGO+	Hypergeometric	Bayesian	GO	—	(Vencio et al., 2006)
DAVID/EASEonline	Fisher's exact test	Bonferroni	GO, pathways, diseases protein domains, interactions	http://david.abcc.ncifcrf.gov/	(Dennis et al., 2003; Hosack et al., 2003)
FatiGO+	Fisher's exact test	Step-down minP, FDR	GO, KEGG, protein domains, swissprot keywords, Transfac motifs, CisRed motifs, chromosomal location, tissues	http://www.fatigo.org	(Al-Shahrour et al., 2004)
FuncSpec	Hypergeometric	Bonferroni	GO, phenotypes, protein interactions, etc. (only for yeast)	http://funspec.med.utoronto.ca/	(Robinson et al., 2002)
GeneMerge	Hypergeometric	Bonferroni	GO, KEGG, chromosomal location, other.	http://genemerge.bioteam.net/	(Castillo-Davis & Hartl, 2003)
GO:TempFinder	Hypergeometric	Bonferroni	GO		(Boyle et al., 2004)
GoMiner	Fisher's exact test	FDR	GO		(Zeeberg et al., 2003; Zeeberg et al., 2005)
Gostat	χ^2 Fisher's exact test	FDR, Holm	GO	http://gostat.wehi.edu.au/	(Beissbarth & Speed, 2004)
GoSurfer	χ^2	*q*-value	GO		(Zhong et al., 2004)
GO ToolBox	Hypergeometric, binomial, Fisher's exact test	Bonferroni	GO	http://crfb.univ-mrs.fr/GOToolBox/index.php	(Martin et al., 2004)
Ontology Traverser	Hypergeometric	FDR	GO	http://franklin.imgen.bcm.tmc.edu/rho-old/services/OntologyTraverser/	(Young et al., 2005)
Onto-Tools	χ^2 binomial, hypergeometric Fisher's exact test	Sidak, Holm, Bonferroni, FDR	GO, KEGG	http://vortex.cs.wayne.edu/projects.htm	(Draghici et al., 2003; Khatri et al., 2005)
FuncAssociate	Fisher's exact test	—	GO	http://llama.med.harvard.edu/cgi/func/funcassociate	(Berriz et al., 2003)
GOTM	Hypergeometric	—	GO	http://bioinfo.vanderbilt.edu/gotm/	(Zhang et al., 2004)
CLENCH	Hypergeometricx, χ^2 binomial	—	GO (only for *A. thaliana*)	— —	(Shah & Fedoroff, 2004)

For web-based applications, the URL is also provided. Although the most commonly tools have been included here, this list does not intend to be exhaustive.

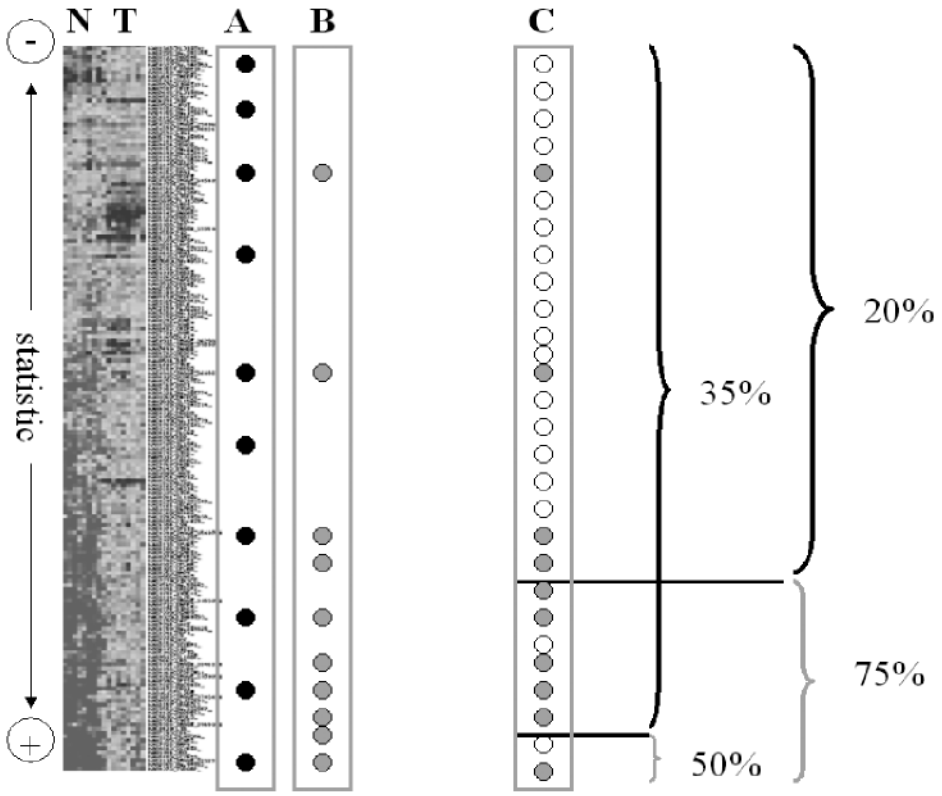

FIG. 2. Threshold-free procedure for the functional annotation of class comparison experiments. On the left: genes ordered by differential expression between classes N (normal) and T (test). (**A**) Functional term unrelated to the experiment from which the rank of genes was obtained. (**B**) Functional term related to the experiment. (**C**) Schematic representation of two partitions of the segmentation test.

expression), but scarcely appears among genes highly expressed in class T. This observation clearly points to functional category B as one of the molecular basis of the macroscopic observation made in the experiment. Instead of trying to select genes with extreme values of differential expression, systems biology-inspired methods will directly search for blocks of functionally related genes significantly cumulated in the extremes of a ranked list of genes.

There are different methods which have been proposed for this purpose such as the GSEA (Mootha et al., 2003; Subramanian et al., 2005) or the SAFE (Barry et al., 2005) method that use a non-parametrical version of a Kolmogorov-Smirnov test. Other strategies are also possible, such as the direct analysis of functional terms weighted with experimental data (Smid and Dorssers, 2004) or model-based methods (Goeman et al., 2004). With similar accuracy, conceptually simpler and quicker methods have also been proposed such as the parametrical counterpart of the GSEA, the PAGE (Kim and Volsky, 2005) or the segmentation test, Fatiscan (Al-Shahrour et al., 2005a).

FatiScan: a segmentation test

A simple way of studying the asymmetrical distribution of blocks of genes across a list of ranked genes is to check if, in consecutive partitions, one of the parts is significantly enriched in any biological term with respect to their complementary part. Figure 2C illustrates this concept in an ordered list of genes. In this list, gray circles represent genes annotated with a particular functional category and open circles represent genes with any different annotation. In the first partition, the differences (50% versus 35%), cannot be considered significant. Nevertheless, in the second partition, the differences in the proportions are high enough

to be declared significant (75% versus 20%): the vast majority of the genes annotated with the functional category are on the lower side of the partition.

The segmentation test used for threshold-free functional annotation consists on the sequential application of the FatiGO (Al-Shahrour et al., 2004) test to different partitions of an ordered list of genes. The FatiGO test uses a Fisher's exact test over a contingency table for finding significantly over or under represented biological terms when comparing the upper side to the lower side of the list, as defined by any partition. Previous results have shown that a number between 20 and 50 partitions often gives optimal results in terms of sensitivity and results recovered (Al-Shahrour et al., 2005a). Given that multiple terms (T) are tested in a predefined number of partitions (P), the unadjusted p-values for a total of $T \times P$ tests must be corrected. The widely accepted FDR (Benjamini and Yekutieli, 2001) can be used for this purpose. Nevertheless, carrying out a total of $T \times P$ tests would correspond to the most conservative scenario, in a situation in which no *a priori* functional knowledge of the system is available. Usually many terms can initially be discarded from the analysis due to prior information or just by common sense.

The FatiScan test presents two fundamental advantages when compared to other alternative methods based on Kolmogorov-Smirnov or related tests. On one hand, this method does not require an extreme non-uniform distribution of genes. FatiScan is able to find different types of asymmetries in the distribution of groups of genes across the list of data. On the other hand, and still more important, this method does not depend on the original data from which the ranking of the list was derived. The significance of the test depends only on the ranking of the genes in the list and the strategy used for performing the partitions. This means that, in addition to DNA microarray data, this method can be applied to any type of genome-scale data in which a value can be obtained for each gene. FatiScan is available within the Babelomics package (Al-Shahrour et al., 2006, 2005b) for functional annotation of genome-scale experiments (www.babelomics.org).

RESULTS AND DISCUSION

Functional analysis of gene expression in human diabetes

A case study of gene expression in human diabetes (Mootha et al., 2003) is used to illustrate the application of the threshold-free strategy. In the experiment a comparison between two classes (17 controls with normal tolerance to glucose versus 26 cases, composed of 8 subjects with impaired tolerance and 18 more subjects with type 2 diabetes mellitus, DM2) failed to detect even a single gene with significant differential expression.

Genes were ordered according to their differential expression between cases and controls. A t-test, as implemented in the T-Rex tool from the GEPAS package (Herrero et al., 2003, 2004; Vaquerizas et al., 2005) was used for this purpose. The value of the statistic was used as ranking criteria for ordering the list. As in the original analysis (Mootha et al., 2003) no genes with a significant differential expression (with a FDR-adjusted p-value lower than 0.05) were found. A total of 50 partitions of the ranked list were analysed with the FatiScan algorithm for over- or under-representation of KEGG pathways and GO terms. Table 2 shows the functional terms significantly over-expressed in the comparison; among them, the most relevant are *oxidative phosphorylation*, *ATP synthesis* and *Ribosome*, that were found to be significantly over-expressed in healthy controls versus diseased cases. Conversely, *Insulin signalling pathway* was over-expressed in diseased cases. GO terms and other biologically relevant terms with similar meaning were also found.

Comparison among different threshold-free approaches

Other alternative methods give similar results (Table 2). Oxidative phosphorylation and mitochondrion are found by GSEA (Mootha et al., 2003), PAGE (Kim and Volsky, 2005) and other statistics (Tian et al., 2005). Nucleotide biosynthesis can be assimilated to other terms found by these three methods (Kim and Volsky, 2005; Mootha et al., 2003; Tian et al., 2005) based on a set of functional categories developed by (Subramanian et al., 2005).

The example showed how different methods detected similar biological terms, even though some of the terms were defined in different repositories and different contexts. This agreement in the results is in bold

Table 2. Different Biological Terms Found as Significantly Up- and Down-Regulated When Healthy Controls versus Diabetic Cases by Distinct Threshold-Free Methods

Healthy vs. diabetic	*Pathway*	*FatiScan*	*GSEA*	*PAGE*	*Tian et al.*
Up-regulated	Oxidative phosphorylation[1,3,4]	Yes[1,4]	Yes[3]	Yes[3]	Yes[3]
	ATP synthesis[1]	Yes	NA	NA	NA
	Ribosome[1]	Yes	NA	NA	NA
	Ubiquinone[2]	Yes	NA	NA	NA
	Ribosomal protein[2]	Yes	NA	NA	NA
	Ribonucleoprotein[2]	Yes	NA	NA	NA
	Mitochondrion[2,4]	Yes[2,4]	Yes[3]	Yes[3]	Yes[3]
	Transit peptide[2]	Yes	NA	NA	NA
	Nucleotide biosynthesis[3,4]	Yes[4]	Yes[3]	Yes[3]	Yes[3]
	NADH dehidrogense (ubiquinone) activity[4]	Yes	NA	NA	NA
	Nuclease activity[4]	Yes	NA	NA	NA
Down-regulated	Insulin signalling pathway[1]	Yes	NA	NA	NA

Some of the terms refer to similar concepts and contain essentially the same genes, but are defined in different repositories (e.g., GO, KEGG).

Functional blocks constructed according to [1]KEGG pathways; [2]Swissprot keywords; [3]Classes defined in GSEA; and [4]GO terms.

contrast if compared to the poor overlap reported in different studies where the aim was the selection of genes (Bammler et al., 2005) instead of blocks of functionally related genes. Nevertheless a rigorous comparative study is still necessary to decide on the most efficient one. From a practical point of view, the preferred methods would be those which use more biological terms and software packages, such as FatiScan or GSEA.

There are, however, situations in which the threshold-based, two-step approach is the natural choice. For example, in clustering of genes, clusters are firstly defined and then, are analysed by functional annotation methods. Table 1 provides an overview for the relative merits of the available programmes. As previously mentioned, the package Babelomics (Al-Shahrour et al., 2006, 2005b), which includes FatiGO (Al-Shahrour et al., 2004), and the Onto Tools (Draghici et al., 2003; Khatri et al., 2005) would be optimal in terms of biological information content and testing strategies.

On genes and blocks of functionally related genes

Microarray technologies allow for data to be obtained on thousands of genes. Nevertheless, functional genomics hypotheses on class comparison are still tested by artificially decomposing the problem into thousands of independent problems, one for each gene, and by later exploring some functional links among them. The tests are repeatedly applied to one gene at a time and ignore all the available knowledge about their cooperative behaviour accumulated over the last years, stored in different repositories (e.g., GO, KEGG, protein interaction databases). Since this strategy involves carrying out a large number of tests, severe corrections must be imposed to reduce the number of false positives. Later, the genes selected by means of these procedures are functionally analysed in a second step. It is evident that with this strategy a significant amount of information is lost in both steps. Recently, however, different methods have been proposed that take into account the properties of the genes and address different biological questions not in a gene-

centric manner but in a function-centric manner for class comparison (Al-Shahrour et al., 2005a; Goeman et al., 2005, 2004; Kim and Volsky, 2005; Mootha et al., 2003; Subramanian et al., 2005; Tian et al., 2005), class assignation (Lottaz and Spang, 2005) and class discovery (Huang and Pan, 2006; Jia and Xu, 2005; Pan, 2006).

Last, but not least, algorithms are used if they are implemented in programmes. Given the number of steps necessary for the proper analysis of a microarray experiment (normalisation, the analysis itself, and the functional annotation), integrated packages are preferable in order to avoid problems derived from the change of formats. Among the most complete packages available on the web is the GEPAS (Herrero et al., 2003, 2004; Montaner et al., 2006; Vaquerizas et al., 2005) which offers different options for normalisation, supervised and unsupervised analysis (www.gepas.org) and is linked to the Babelomics suite (Al-Shahrour et al., 2006, 2005b) for the functional annotation of genome-scale experiments (www.babelomics.org), where different tests for two-steps of threshold-free functional annotation are implemented.

CONCLUSION

Ending up with a mere list of genes of interest is only half way to the result of a microarray experiment. Apart from the utility that functional annotation can have as an external criterion to check the reliability of the results of a microarray experiment -for example, in cluster quality assessment (Bolshakova et al., 2005)-, it constitutes itself an unavoidable final step of any microarray analysis.

The aim of the methods for the functional interpretation of microarray experiments is to find a functional explanation at molecular level that accounts for the macroscopic observation related to the hypothesis that originated the experiment (e.g., why a number of genes are responsible for the physiological differences between healthy and diseased people). This is achieved through the study of the over-representation of some type of functionally relevant labels in the genes detected as important in the experiment. Two main approaches are currently in use: threshold-based and threshold-free. In the first case, the conclusions are reached by means of a two-steps process where the important genes are firstly selected based upon their experimental values (e.g., using a test for differential expression between two classes or a clustering method for finding co-expressed genes, etc.) Then, this selection is analysed for the significant enrichment of biological terms with functional meaning (e.g. GO, KEGG) using different tests (Khatri and Draghici, 2005). Several authors have pointed out that the first step of such strategy, where genes are selected without taking into account their cooperative behaviour, would constitute its Achile's heel. If the genes are considered as independent and tested one at a time, then very stringent thresholds need to be used to reduce the rate of false positives (Al-Shahrour et al., 2005a; Mootha et al., 2003). The obvious consequence if this is the reduction of the sensitivity in the second step. Threshold-free methods avoid this first step of selection by not considering genes alone, but in functional blocks. These methods have proved to be much more sensitive that the threshold-based alternatives (Al-Shahrour et al., 2005a; Mootha et al., 2003). Moreover, from the point of view of the systems biology threshold-free methods are far more consistent because they directly test pre-defined functionally-related blocks of genes. These blocks are formed by genes that share functional labels, which are supposed to account for the cooperative roles fulfilled by these genes in the cell. Such functionally-related blocks of genes would provide a molecular-level explanation for the macroscopic traits studied in the microarray experiment.

The functional interpretation of microarray experiments can be considered an emerging conceptual area in which a number of issues need still to be addressed. Two main aspects are susceptible of improvement: the definition of blocks of functionally-related genes and the interpretation of data other than simple ranked lists of genes. Blocks of functionally-related genes refer to biological meaningful terms that have been defined by curators in different repositories (e.g., GO, KEGG) or can be defined by the users. These blocks can be considered categorical variables in the sense that a gene belongs (or not) to a given class. Partial or conditional membership are not considered. While this definition could be applicable to some functionally-related classes, such as the "ribosomal proteins" which show a tight coordinated expression, in other classes this level of coordination in the expression cannot be expected from all the members. Thus, the introduction of some weight that consider distinct degrees of membership or the use of different tests that account

for non categorical classes would improve the resolution of the methods for functional interpretation of microarray experiments.

Not all the experimental outcomes in microarray data analysis can be represented as a list of ordered genes. This representation is suitable for class comparison or for the study of a continuous parameter (e.g., the level of a metabolite) or survival studies, in which a threshold-free approach can be applied. Nevertheless there are situations in which this list does not have a so simple interpretation, as is the case of multiclass comparison. Another interesting situation is when multiple variables are simultaneously studied. In this case instead of a uni-dimensional list the resulting representation could be imagined as a multi-dimensional space in which accumulation of biologically relevant terms must be studied. Also a network of transcriptional interactions could be represented as a graph or as a matrix. In any case, the functional interpretation by threshold-free strategies of these different gene arrangements is something that must be addressed in the future. As previously mentioned, class discovery methods are also using systems biology inspired strategies. Recently, biological information (Huang and Pan, 2006; Pan, 2006) or phenotypic information (Jia and Xu, 2005) has been used as a constitutive part of clustering algorithms.

A systems biology approach to the analysis of microarray data in the future will tend to use more information on the cooperative properties of the genes beyond their simple functional definition. Thus it is expected that a deeper knowledge on both the interactome and the transcriptional network of the genomes will contribute to the formulation of more realistic biological questions using microarray experiments and to obtain more complete and accurate answers.

ACKNOWLEDGMENTS

This work is supported by grants from MEC BIO2005-01078, NRC Canada-SEPOCT Spain, and Fundación Genoma España.

REFERENCES

AL-SHAHROUR, F., DIAZ-URIARTE, R., and DOPAZO, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics **20,** 578–580.

AL-SHAHROUR, F., DIAZ-URIARTE, R., and DOPAZO, J. (2005a). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics **21,** 2988–2993.

AL-SHAHROUR, F., and DOPAZO, J. (2005). Ontologies and functional genomics. In *Data analysis and visualization in genomics and proteomics*. Azuaje, F. and Dopazo, J., eds. (Wiley, New York), pp. 99–112.

AL-SHAHROUR, F., MINGUEZ, P., VAQUERIZAS, J.M., et al. (2005b). BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res **33,** W460–464.

AL-SHAHROUR, F., MINGUEZ, P., TARRAGA, J., et al. (2006). BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res (in press).

ALLISON, D.B., CUI, X., PAGE, G.P., et al. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet **7,** 55–65.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25,** 25–29.

AZUAJE, F., and DOPAZO, J. (2005). *Data Analysis and Visualization in Genomics and Proteomics* (Wiley, Chichester, UK).

BAMMLER, T., BEYER, R.P., BHATTACHARYA, S., et al. (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods **2,** 351–356.

BARRY, W.T., NOBEL, A.B., and WRIGHT, F.A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics **21,** 1943–1949.

BEISSBARTH, T., and SPEED, T.P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics **20,** 1464–1465.

BENJAMINI, Y., and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics **29,** 1165–1188.

BERRIZ, G.F., KING, O.D., BRYANT, B., et al. (2003). Characterizing gene sets with FuncAssociate. Bioinformatics **19,** 2502–2504.

BOLSHAKOVA, N., AZUAJE, F. and CUNNINGHAM, P. (2005). A knowledge-driven approach to cluster validity assessment. Bioinformatics **21,** 2546–2547.

BOYLE, E.I., WENG, S., GOLLUB, J., et al. (2004). GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics **20,** 3710–3715.

CASTILLO-DAVIS, C.I., and HARTL, D.L. (2003). GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19,** 891–892.

CHEN, Z., WANG, W., LING, X.B., et al. (2006). GO-Diff: Mining functional differentiation between EST-based transcriptomes. BMC Bioinformatics **7,** 72.

DAMIAN, D., and GORFINE, M. (2004). Statistical concerns about the GSEA procedure. Nat Genet **36,** 663; author reply 663.

DENNIS, G., Jr., SHERMAN, B.T., HOSACK, D.A., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol **4,** P3.

DRAGHICI, S., KHATRI, P., MARTINS, R.P., et al. (2003). Global functional profiling of gene expression. Genomics **81,** 98–104.

DRAGHICI, S., and KUKLIN, A. (2003). Data analysis tools for DNA microarrays, CRC Press, Boca Raton).

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., et al. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95,** 14863–14868.

GE, H., WALHOUT, A.J., and VIDAL, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet **19,** 551–560.

GOEMAN, J.J., VAN DE GEER, S.A., DE KORT, F., et al. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics **20,** 93–99.

GOEMAN, J.J., OOSTING, J., CLETON-JANSEN, A.M., et al. (2005). Testing association of a pathway with survival using gene expression data. Bioinformatics **21,** 1950–1957.

GOLUB, T.R., SLONIM, D.K., TAMAYO, P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286,** 531–537.

HALLIKAS, O., PALIN, K., SINJUSHINA, N., et al. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell **124,** 47–59.

HERRERO, J., AL-SHAHROUR, F., DIAZ-URIARTE, R., et al. (2003). GEPAS: A web-based resource for microarray gene expression data analysis. Nucleic Acids Res **31,** 3461–3467.

HERRERO, J., VAQUERIZAS, J.M., AL-SHAHROUR, F., et al. (2004). New challenges in gene expression data analysis and the extended GEPAS. Nucleic Acids Res **32,** W485–491.

HOSACK, D.A., DENNIS, G., Jr., SHERMAN, B.T., et al. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol **4,** R70.

HUANG, D., and PAN, W. (2006). Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics **22,** 1259–1268.

JIA, Z., and XU, S. (2005). Clustering expressed genes on the basis of their association with a quantitative phenotype. Genet Res **86,** 193–207.

KANEHISA, M., GOTO, S., KAWASHIMA, S., et al. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res **32,** D277–D280.

KHATRI, P., and DRAGHICI, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics **21,** 3587–3595.

KHATRI, P., SELLAMUTHU, S., MALHOTRA, P., et al. (2005). Recent additions and improvements to the OntoTools. Nucleic Acids Res **33,** W762–765.

KIM, S.Y., and VOLSKY, D.J. (2005). PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics **6,** 144.

LEE, H.K., HSU, A.K., SAJDAK, J., et al. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res **14,** 1085–1094.

LOTTAZ, C., and SPANG, R. (2005). Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics **21,** 1971–1978.

MATEOS, A., DOPAZO, J., JANSEN, R., et al. (2002). Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res **12,** 1703–1715.

MONTANER, D., TARRAGA, J., HUERTA-CEPAS, J., et al. (2006). Next station in microarray data analysis: GEPAS. Nucleic Acids Res In press.

MOOTHA, V.K., LINDGREN, C.M., ERIKSSON, K.F., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet **34,** 267–273.

MOREAU, Y., AERTS, S., DE MOOR, B., et al. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet **19,** 570–577.

MULDER, N.J., APWEILER, R., ATTWOOD, T.K., et al. (2005). InterPro, progress and status in 2005. Nucleic Acids Res **33,** D201–205.

PAN, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics **22,** 795–801.

PEROU, C.M., JEFFREY, S.S., VAN DE RIJN, M., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A **96,** 9212–9217.

QIU, X., XIAO, Y., GORDON, A., et al. (2006). Assessing stability of gene selection in microarray data analysis. BMC Bioinformatics **7,** 50.

ROBERTSON, G., BILENKY, M., LIN, K., et al. (2006). cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res **34,** D68–D73.

ROBINSON, M.D., GRIGULL, J., MOHAMMAD, N., et al. (2002). FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics **3,** 35.

RUAL, J.F., VENKATESAN, K., HAO, T., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature **437,** 1173–1178.

SHAH, N.H., and FEDOROFF, N.V. (2004). CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. Bioinformatics **20,** 1196–1197.

SMID, M., and DORSSERS, L.C. (2004). GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. Bioinformatics **20,** 2618–2625.

STELZL, U., WORM, U., LALOWSKI, M., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell **122,** 957–968.

STOREY, J.D., and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100,** 9440–9445.

STUART, J.M., SEGAL, E., KOLLER, D., et al. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science **302,** 249–255.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **102,** 15545–15550.

TIAN, L., GREENBERG, S.A., KONG, S.W., et al. (2005). Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A **102,** 13544–13549.

VAN 'T VEER, L.J., DAI, H., VAN DE VIJVER, M.J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature **415,** 530–536.

VAN NOORT, V., SNEL, B., and HUYNEN, M.A. (2003). Predicting gene function by conserved co-expression. Trends Genet **19,** 238–242.

VAQUERIZAS, J.M., CONDE, L., YANKILEVICH, P., et al. (2005). GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res **33,** W616–620.

VENCIO, R., KOIDE, T., GOMES, S., et al. (2006). BayGO: Bayesian analysis of ontology term enrichment in microarray data. BMC Bioinformatics **7,** 86.

WESTERHOFF, H.V., and PALSSON, B.O. (2004). The evolution of molecular biology into systems biology. Nat Biotechnol **22,** 1249–1252.

WINGENDER, E., CHEN, X., HEHL, R., et al. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res **28,** 316–319.

WOLFE, C.J., KOHANE, I.S., and BUTTE, A.J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics **6,** 227.

YOUNG, A., WHITEHOUSE, N., CHO, J., et al. (2005). OntologyTraverser: an R package for GO analysis. Bioinformatics **21,** 275–276.

ZEEBERG, B.R., FENG, W., WANG, G., et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol **4,** R28.

ZEEBERG, B.R., QIN, H., NARASIMHAN, S., et al. (2005). High-Throughput GoMiner, an "industrial-strength" integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC Bioinformatics **6,** 168.

ZHANG, B., SCHMOYER, D., KIROV, S., et al. (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics **5,** 16.

ZHONG, S., STORCH, K.F., LIPAN, O., et al. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Appl Bioinformatics **3,** 261–264.

Address reprint requests to:
Dr. Joaquín Dopazo
Department of Bioinformatics
INB
CIPF
Avda. Autopista del Saler 16
Valencia E-46013, Spain

E-mail: jdopazo@cipf.es