ABSTRACT
Motivation: The analysis of genome-scale data from different high throughput techniques can be obtained to use lists of genes ordered according to their different behaviours under distinct experimental conditions corresponding to different phenotypes (e.g. differential gene expression between diseased samples and controls, different response to a drug, etc.). The order in which the genes appear in the list is a consequence of the biological roles that the genes play within the cell, which account, at molecular scale, for the macroscopic differences observed between the phenotypes studied. Typically, two steps are followed for understanding the biological processes that differentiate phenotypes at molecular level: first, genes with significant differential expression are selected on the basis of their experimental values and subsequently, the functional properties of these genes are analysed. Instead, we present a simple procedure which combines experimental measurements with available biological information in a way that genes are simultaneously tested in groups related by common functional properties. The method proposed constitutes a very sensitive tool for selecting genes with significant differential behaviour in the experimental conditions tested. Results: We propose the use of a method to scan ordered lists of genes. The method allows the understanding of the biological processes operating at molecular level behind the macroscopic expression from which the list was generated. This procedure can be useful in situations where it is not possible to obtain statistically significant differences based on the experimental measurements (e.g. low prevalence diseases, etc.). Two examples demonstrate its application in two microarray experiments and the type of information that can be extracted.


Contact: jdopazo@ochoa.fib.es

1 INTRODUCTION
High throughput methodologies, such as gene expression (Eisen et al., 1998), produce vast amounts of data that can be used to relate molecular level differences in gene behaviour in response to macroscopic phenotypes. Microarray gene expression data can be processed using different methods, based on distinct criteria, such that ordered lists of genes can be generated. The rationale behind this step is that the disposition found for the genes must be a direct consequence of the distinct functions carried out by them in the cell under the different experimental conditions or distinct phenotypes studied. In order to understand the molecular basis that differentiate the phenotypes, two steps are typically followed. First, gene expression values are examined by means of an appropriate statistical test. Genes with extreme differential expression values (over the threshold defined in the test) are believed to have a differential behaviour in the experiment (that can involve discrete classes—e.g. healthy versus diseased samples—or continuous traits—such as the concentration of a metabolite, drug dosages, time series or even survival data). Then, available biological information for these genes are used, a posteriori, in order to understand the biological processes operating at molecular level that account for the experiment. For example, van’t Veer et al. (2002) selected 231 prognostic reporter genes according to their correlation coefficients with an average good prognosis gene expression profile. The functional annotations of the genes were then studied to gain insight into the underlying biological mechanism leading to rapid metastases. Genes involved in cell cycle, invasion and metastasis, angiogenesis and signal transduction were found to be significantly upregulated in the poor prognosis signature (van’t Veer et al., 2002). Other authors have used DNA microarrays to identify clinically relevant subtypes of cancers (see, for example, Lapointe et al., 2004) and then found genes differentially expressed among the classes that are associated with distinct biological properties. In particular, Culilife et al. (2003) used gene ontology to highlight functionally distinct biological responses to different mitogens in a study of response of breast cancer to growth regulators.

Despite the prior availability of this biological information, thresholds for defining differential behaviour of genes are obtained relying exclusively on experimental measurements of gene expression, with functional study as the subsequent step. Only a
few attempts have been made to use the biological information as part of the gene selection process. Mootha et al. (2003) used Gene Set Enrichment Analysis (GSEA) to detect modest but coordinate changes in the expression of groups of functionally related genes. They applied GSEA to a dataset in which no genes were over conventional thresholds, based only on the comparison of gene expression values. They were able to find a group of genes involved in phosphorylation with decreased expression in human diabetic muscle.

Here we present a different procedure that uses biological information in combination with the experimental measurements to understand the biological processes operating at molecular level that explain macroscopic differences between phenotypes, or experiments using DNA microarray or other high throughput techniques, such as proteomics. The procedure proposed here uses Gene Ontology (GO) (Ashburner et al., 2000) terms as functional labels for genes. GO provides information on molecular function, biological processes and cellular components for a number of different organisms. In addition we use information on metabolic pathways taken from KEGG (Kanehisa et al., 2004).

The procedure consists of the application of a test for extracting significant over- or under-represented terms associated with groups of genes (see Al-Shahrour et al., 2004). The test is sequentially applied to different partitions of an ordered list of genes, previously obtained by studying their differential expression according to the phenotypes analysed. Then, labels accounting for biological information are used to test if groups of genes functionally related are simultaneously over- or under-expressed (or, in other words, if these genes tend to be concentrated in one of the ends of the list instead of being uniformly distributed across it.) Contrarily to GSEA, and other alternative methods based on comparison of distributions through Kolmogorov–Smirnov or related tests, this method does not require an extreme non-uniform distribution of genes. It is capable of finding different types of asymmetries in the distribution of groups of genes across the list of data. The strong multiple-testing nature of the statistical contrast has been taken into account. Two examples show the use of the procedure and demonstrate its sensitivity. The proposed procedure seems to have a better performance than GSEA.

2 METHODS

2.1 General description of the method

The proposed method has three steps (Fig. 1). First, a list of genes is ordered using experimental information on their differential expression according to the phenotype studied in the experiment (Fig. 1A). For example, genes can be ordered on the basis of their differential expression between two experimental conditions (pre- and post-drug administration, healthy versus diseased samples, etc.). The second step (Fig. 1B and C) involves the study of the distribution of functional terms in different partitions of this list using a test described below (Al-Shahrour et al., 2004). Then, a plot with the significant terms obtained upon the application of the test can be used to detect significant asymmetrical distributions of genes, responsible for different biological processes, across the list (Fig. 1D). Finding significant asymmetrical distributions of functional terms across the list strongly suggests that groups of genes having in common functional labels are significantly over- or under-expressed as a block.

The last significant partition defines the border for the number of genes belonging to a functional class that can be considered to display a significant non-uniform distribution across the list. For example, if we assume that in Figure 1 the second partition is the last significant one for genes in the right column (Fig. 1E), the conclusions would be: (1) this function is asymmetrically distributed across the rank and (2) six of these genes can be considered to have a coordinated, significant non-uniform distribution, while the other four do not show an asymmetrical behaviour with respect to the rank. It is not unusual to find such a tail of genes. The reason for this is that many genes participate in more than one biological function and they might be carrying out different biological roles unrelated to the experimental conditions that originated the rank of genes.

2.2 Extraction of significantly under- and over-represented functional terms in a set of genes: the FatiGO/FatiWise algorithm

The core of the method proposed is based on an algorithm to test whether a set of genes, labelled with terms (biological information), contain significant enrichments on one or several of these terms with respect to another set of genes of reference. We have used the FatiGO (Al-Shahrour et al., 2004) tool to perform such test. FatiGO works with tables of correspondence between genes and GO terms. Kegg pathways are implemented in the program FatiWise, included in the GEPAS (Herrero et al., 2004) and BABELOMICS (Al-Shahrour et al., 2005) packages. The application of FatiGO and FatiWise to consecutive partitions of the list of genes constitutes the step represented in Figure 1B. Both programs use a Fisher’s exact test for 2 × 2 contingency tables for comparing two groups of genes and extracting a list of GO terms whose distribution among the groups is significantly different. Given that many GO terms are simultaneously tested, the results of the test are corrected for multiple-testing to obtain an adjusted P-value. FatiGO and FatiWise return adjusted P-values based on three different ways of accounting for multiple testing: step-down minP method (Westfall and Young, 1993) and two False Discovery Rate (FDR) methods (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) (Fig. 1C).

2.3 Step 1: preparing the data: obtaining lists of genes ordered according to phenotypes

Typical DNA microarray data usually consist of a matrix of gene expression values where columns represent different experiments, corresponding to

---

Fig. 1. Summary of the steps of the proposed method: (A) list of genes ordered according their differential behaviour in the experimental conditions studied by means of a statistical test, (B) establishment of different partitions across the list, (C) extraction of the biological terms differentially expressed in each partition and (D) plot of the super-adjusted terms with significant differential expression. (E) Genes labelled with two different terms. The gene on the left is uniformly distributed across the list whereas the gene on the right is mainly present in one of the tails of the list and, consequently, this term is likely to be related to the experimental trait studied.
different phenotypes, and rows are genes. Phenotypes are understood here in a broad sense: columns can correspond to discrete classes (e.g. case control studies, distinct tissues, etc.), to continuous traits (e.g. the level of a metabolite, survival data, etc.), to the different responses of a cell culture to distinct drugs, etc.

The first step of the method is to obtain an ordered list of genes based on their differential expression with respect to the trait studied. This ordered list can be obtained by means of the application of tests for differential expression between two classes (t-test) or among multiple classes (ANOVA). Continuous traits can be studied by correlation, and survival data can be analysed using cox models. The POMELO tool, included in the GEPAS package (Herrero et al., 2003) implements all these tests. The value of the statistic can be used as ranking criteria (Figure 1A).

2.4 Step 2: scanning ordered lists of genes with FatiGO/FatiWise and final selection of significant terms

Once an ordered list of genes has been obtained, the method checks whether blocks of genes with common functional properties are uniformly distributed across it or, on the contrary, are either cumulated in one of the tails or present in any sort of asymmetrical distribution. A uniform distribution of genes with a particular functional property across the list implies that this property is most probably unrelated to the trait studied. Contrarily, an accumulation of genes with a common function in the upper or lower range of the list is a clear indication of a possible relationship between this function and the trait (Fig. 1E). And this is true at the level of blocks of functionally related genes even in the case of no one single gene satisfies an individual test based solely on expression values.

The second step involves the study of different partitions of this list using the FatiGO/FatiWise test described above (Fig. 1B). Here we have to perform one test per partition and, consequently, we confront a problem with a severe multiple-testing component. The procedure proposed here implies producing as many partitions of the list of genes as defined by the step selected. Since FatiGO adjusts P-values for only one partition at a time and we are conducting P partitions, we need a more rigorous adjustment. We then have to correct multiple terms (M) in multiple partitions (P). In other words, we have to correct unadjusted P-values for \( M \times P \) tests.

We have used the multtest package (Pollard et al., 2004) of bioconductor (Gentleman, 2004, http://www.bioconductor.org) to obtain adjusted P-values by FDR (Benjamini and Yekutieli, 2001) from the unadjusted P-values. Multtest inputs all the uncorrected P-values obtained, and outputs FDR-adjusted P-values.

Finally, a plot with the significant terms obtained for each partition upon the application of the FatiGO/Fatiwise test is produced. This plot can be used to detect significant asymmetrical distributions of blocks of functionally related genes across the list (Fig. 1D). Only terms significant after the super-adjustment process will be plotted. Since the test is based on the differences of percentages, the plot will represent these differences as a function of the value of the statistic. If some biological process is activated as a block, genes playing this role will tend to be cumulated in the most extreme values of the statistic (which measures differential expression). As a general trend, significant differences in percentages will tend to be higher for extreme values of the statistic and will tend to be lower as partitions tend to move to values of the statistic nearer to zero. This happens because we dilute the term by adding other genes less differentially expressed. Actually, what is important in the results is not the plot itself, but the term which is found as differentially distributed in the list of genes. The plot helps to establish a scale of trustworthiness for the terms. Thus, terms significant across a wide range of partitions are more trustworthy than terms that are significant only for a unique partition.

2.5 How many partitions? A heuristic approach

The decision on the number of partitions to be made is an heuristic one. More the partitions more number of tests (which will cause an artificial decrease in the power and the sensitivity of the method after the adjustment of P-values).

On the other hand, a very small number of partitions (say ten or less) can cause a loss of sensitivity because we could not detect regions of the list enriched in genes belonging to some functional category.

Partitions can be made in different ways (based on percentiles, the value of the statistic, etc.). In the case of studying differential gene expression the use of the value of the statistic is probably the best choice, because it reflects differences in expression as well as significance of these differences. Our experience shows that a number between 20 and 50 partitions often gives optimal results in terms of sensitivity and results recovered. The method aims to find blocks of genes (belonging to a functional category) asymmetrically distributed across an ordered list of genes. Consequently, the sensitivity of the method is related to the number of genes detected belonging to these categories. An example with a diabetes dataset (Mootha et al., 2003) was used to illustrate the influence of the number of partitions on the number of genes detected for the above mentioned functionally related blocks. In the example, different functional terms (pathways in this case) showed a significant asymmetrical distribution (i.e. they were related to the trait studied). The goal is to find the level of granularity at which the maximum number of genes belonging to a functional category can be detected by the FatiGO/Fatiwise algorithm. Figure 2 shows how, in some cases, even 20 partitions were sufficient to find all the genes belonging to a functional category asymmetrically distributed. In all cases, the selection of a window step that produces 50 partitions detected a number of genes for each category that is not further improved by using more partitions. Based on empirical results we suggest the use of a level of granularity between 20 and 50 partitions.

3 RESULTS

3.1 Differential gene expression in human diabetes samples

To illustrate the use of the method, we had chosen a study of gene expression in human diabetes (Mootha et al., 2003). The authors applied the GSEA (a version of the Kolmogorov–Smirnov test for comparing distributions) to the data composed of 43 samples (17 with normal tolerance to glucose, 8 with impaired tolerance and 18 with type 2 diabetes mellitus, DM2) and found that genes related to
We ordered the genes according to their differential expression among normal tolerance to glucose versus impaired tolerance together with DM2. A t-test, as implemented in the POMELO tool from the GEPAS package (Herrero et al., 2003) was used for this. The value of the statistic was used as the ranking criteria for ordering the list. As in the case of Mootha et al. (2003) we were unable to find individual genes with a significant differential expression (differentially expressed genes with an adjusted P-value lower than 0.05).

We have used the value of the t-statistic to carry out the partitions. Starting from the highest values, which correspond, in this case, to genes showing higher expression in normal tolerance to glucose samples (right-hand side of the plots in Figs 3 and 4), we produced ~50 partitions of the whole set of genes (by using steps of 0.2 U of the t-statistic). Then, for each partition we compared the genes with highest expression in normal tissues versus the remaining ones using the FatiWise algorithm from the GEPAS package (Herrero et al., 2004). Whenever a biological term corresponding to any KEGG pathway (Kanehisa et al., 2002) is significantly over- or under-represented in one of the compared partitions versus the other one, it was plotted in the figures.

When we studied the distribution of genes corresponding to different pathways (Fig. 3) we found oxidative phosphorylation over-represented in normal tissue (and consequently under-represented in diseased tissue) with a clear trend of reduction towards diseased tissues. Actually, 82 out of a total of 95 genes labelled as oxidative phosphorylation were found in the last significant partition which approximately correspond to half of those highly expressed in healthy samples. This was already reported by Mootha et al. (2003). But, in addition, our approach detected another term: ubiquinone biosynthesis. This term is also under-represented in diseased tissues. Ubiquinone is an endogenous antioxidant and is a part of the transport electron chain which is altered in diabetes processes (Wold et al., 2003). In fact, supplements of this coenzyme are used as therapy in DM2. In this case, we found all the genes (30 in this case) labelled with this term. The proposed approach seems to be more sensitive to the GSEA (Mootha et al., 2003), which was unable to identify ubiquinone biosynthesis probably because of the low number of genes with this annotation.

We have found molecular function terms at level 4, such as oxidoreductase activity acting on NADH or NADPH or NADH dehydrogenase activity, accounting, as described by means of GO categories, for the same biological roles as oxidative phosphorylation in pathways and displaying the same trend as in the plot of pathways in Figure 3. Terms, such as cation transporter activity, metal ion transporter activity and primary active transporter activity are also significantly lower in diseased patients. Alterations in transport activities might be a consequence of the oxidative stress caused by the diabetes.

This case is a representative example of a quite common situation in gene expression data analysis of clinical samples. In many cases, the biological variability among patients’ samples is so high that it is difficult to find genes showing significant differential expression attending only to their experimental values of gene expression. Using biological information in combination with experimental data allows focusing on sets of related genes, instead of on individual genes, and thereby find significant differential behaviours.

### 3.2 Roles carried out by genes related to survival in breast cancer

Additionally we have studied the results by van’t Veer et al. (2002) in which gene expression patterns of 98 primary breast cancers were analysed. Two patients groups could be distinguished: 34 of them...
developed distant metastases within 5 years (bad prognosis group) and the other 44 continued free of the disease after a period of at least 5 years (good prognosis group). The authors selected 231 prognostic genes according to their correlation coefficients with an average good prognosis profile.

As in the previous example, we ordered the genes according to their differential expression with respect to the prognosis group using a t-test implemented in the POMELO tool. The value of the statistic was used for ordering the list. We performed approximately 50 partitions and we analyzed GO terms corresponding to molecular function and biological process.

Figure 5 shows biological processes significantly over-represented. Here the interpretation of the plot is the opposite to one in the previous section. FatiGO performs a two-tailed test and in this case we found the significance on the part corresponding to the bad prognostic. Negative values of the statistic correspond to genes highly expressed in the bad prognosis group. Since differences in percentages are calculated as the percentage in the good prognosis part of the partition minus the percentage in the bad prognosis part, a negative difference implies that these genes are significantly over-represented in the bad prognosis part (or under-represented in the good prognosis part). Thus, genes labelled as biological processes, such as cell proliferation (GO level 4) and cell growth and/or maintenance (GO level 3), are over-represented among the genes over-expressed in the bad prognosis part of the plot. On the other hand, genes labelled as response to biotic stimulus (GO level 3), which mainly correspond to genes labelled as cell communication and response to external stimulus (GO level 4) are under-represented among the genes highly expressed in bad prognostic patients.

Anecdotally, this last term (response to external stimulus) illustrates a curious effect that sometimes occur when the source of biological information used to label genes has a hierarchical structure in which upper levels correspond to more general definitions and lower levels have a higher degree of precision (GO terms follow exactly this structure). The use of very general terms (high in the hierarchy) can produce bimodal distribution of genes because two children terms of this particular term can have opposite behaviours. A typical example would be the term apoptosis, which has among their descendants anti-apoptosis and induction of apoptosis, defined as ‘a process that directly inhibits any of the steps of the apoptosis’ and ‘a process that directly activates any of the steps of the apoptosis’, respectively. It is easy to imagine an experiment in which genes corresponding to both children terms will be activated, but in the opposite sides of the rank. In the genes that are studied with the general term apoptosis, they will clearly show a bimodal distribution. Nevertheless, due to the few partitions in which the terms response to biotic stimulus, cell communication and response to external stimulus showed a significant asymmetrical distribution it cannot be discarded that this finding is due to chance.

When functions are studied, it can be observed that terms accounting for the significant biological processes also appear as significant. Figure 6 shows how genes belonging to GO functional categories related to communication, such as receptor activity (GO level 3), transmembrane receptor activity (GO level 4) or transcription factor activity (GO level 3) are significantly under-represented among the genes highly expressed in the bad prognosis patients. On the other hand, functional terms related to cell proliferation, such as: nucleotide binding (GO level 3), more accurately defined as purine nucleotide binding (GO level 4), or transferase activity (GO level 3), in particular transferase activity transferring phosphorus-containing groups (GO level 4) or kinase activity (GO level 3), are over-represented in genes highly expressed in bad prognosis patients.

Both the observed trends, those corresponding to biological processes and to molecular functions, are consistent with the biology of the metastatic cells. On one hand, we observe a significant activation of proliferation processes (cell proliferation and cell growth and/or
maintenance) as well as the corresponding molecular functions. On the other hand, processes related to communication tend to be switched off. Both biological processes are the basis of the metastatic uncontrolled cellular growth. This example is a good biological proof of concept of the efficiency of the method proposed.

4 DISCUSSION

As genomic technologies are becoming popular, the possibility of studying problems at the system level instead of at the gene level is becoming a reality. A more interesting question than ‘what is this gene?’ within the context of system biology is ‘what do the genes in this group have in common?’ Unfortunately, most of the data analysis tools were designed under the paradigm of ‘one gene at a time’ and cannot properly address these questions. By means of two examples based on microarray data we have tried to show how a very simple method can be used to answer functional questions relevant for sets of genes. Many phenotypic traits can be studied at gene expression level by designing series of experiments. Genes can be ranked according to their differential expression with respect to the trait (differential expression in cases versus controls, correlation with respect to the level of a metabolite, etc.). These ordered lists of genes can be studied at the level of blocks of genes having in common some functional labels by the proposed method. This method can also be applied to other large-scale genomic techniques, such as proteomics, interactomics, etc.

Genome-scale data analysis must be carried out with caution due to the high likelihood of finding relationships just by chance (Ge et al., 2003). Multiple testing is an important issue that, unfortunately, is not very often addressed (Slonim, 2002). If the multiple-testing nature of the statistical contrast is not taken into account an increase in the rate of false positives (i.e. terms identified as over- or under-represented that, in reality, are not significant) occurs. Statistical tests for finding differentially expressed genes in multiple-testing problems require both high differences in the values and a large number of samples to provide reliable results. Unfortunately, there are cases in which the differences in experimental values are too low to be detected by conventional tests. There are also circumstances in which the availability of a sufficient (usually large) number of samples is a serious problem (e.g. in the case of low prevalence diseases) that impairs the application of conventional statistic approaches. The method presented here provides a statistical framework to study the biological processes that accounts, at molecular level, for the traits studied in gene expression experiments. The method combines experimental and biological information in a way that focuses on groups of genes, related by means of a functional label, instead of focusing on single genes. Contrary to other approaches based on direct comparison of distributions (e.g. using a Kolmogorov–Smirnov test, or the GSEA), our method is able to find asymmetrical distributions of genes with a common biological label across a ranking provided this asymmetry is not too extreme. The argument in favour of using Kolmogorov–Smirnov-based tests is that only one test per term is required. This is an advantage considering the step of multiple-testing adjustment. Nevertheless, finding significantly asymmetrical distributions by this method requires considerable differences in the distribution of terms with respect to the uniform distribution (which is the null hypothesis).

On the other hand, the advantage of using the proposed method is that, despite the apparent disadvantage due to the necessity of performing a strong adjustment in the P-values, the sensitivity in detecting modest asymmetries clearly counterbalances the global sensitivity. The final outcome is a procedure of improved sensitivity, able to detect asymmetrical distributions of groups of genes, even in the case of modest asymmetries.

In the example shown, the procedure proposed here has been more sensitive than the GSEA previously proposed by Mootha et al. (2003) when analysing the same dataset.

The procedure described here is of a straightforward application with publicly available software. FatigO (Al-Shahrour et al., 2005), FatiWise and Pomelo are part of the of GEPAS (Herrero et al., 2003, 2004), a suite of web tools for microarray gene expression data analysis; (http://www.gepas.org) and can be used in combination with the multitest package (Pollard et al., 2004, http://www.bepress.com/ucbbiostat/paper164) of Biocductor (Gentleman, 2004, http://www.biocductor.org) to apply the method described here to microarray data.

ACKNOWLEDGEMENTS

F.A. is supported by Grant BIO2001-0068 from MCYT. This work has been partly supported by grants from Fundación Ramón Areces, Fundación BBVA and Fundació La Caixa. The Functional Genomics node from the INB is supported by Fundación Genoma España.

REFERENCES