

THE USE OF GO TERMS TO UNDERSTAND THE BIOLOGICAL SIGNIFICANCE OF MICROARRAY DIFFERENTIAL GENE EXPRESSION DATA

Ramón Díaz-Uriarte, Fátima Al-Shahrour, and Joaquín Dopazo

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas, (CNIO), (Spanish National Cancer Centre), Melchor Fernández Almagro 3, 28029 Madrid, Spain; email: {rdiaz, falshahrour, jdopazo}@cnio.es

Abstract: We show one way of using Gene Ontology (GO) to understand the biological relevance of statistical differences in gene expression data from microarray experiments. To illustrate our methodology we use the data from Pritchard *et al.* [2001]. Our approach involves three sequential steps: 1) analyze the data to sort genes according to how much they differ between/among organs using a linear model; 2) divide the genes based on "how much or how strongly" they differ, separating those more expressed in one organ vs. those more expressed in the other organ; 3) examine the relative frequency of GO terms in the two groups, using Fisher's exact test, with correction for multiple testing, to assess which of the GO terms differ significantly between the groups of genes. We repeat steps 2) and 3) using a sliding window that covers all the sorted genes, so that we successively compare each group of genes against all others.

By using the GO terms, we obtain biological information about the predominant biological processes or molecular functions of the genes that are differentially expressed between organs, making it easier to evaluate the biological relevance of inter-organ differences in the expression of sets of genes. Moreover, when applied to novel situations (e.g., comparing different cancer conditions), this method can provide important hints about the biologically relevant aspects and characteristics of the differences between conditions. Finally, the proposed method is easily applied.

Key words: Gene Ontology, Unigene, DNA microarray, Fisher's exact test, multiple testing, ANOVA, linear models

1. INTRODUCTION

DNA array technology [e.g., Brown & Botstein, 1999] allows us to analyze the behaviour of several thousands of genes simultaneously in a unique experiment. This number of genes constitutes a significant proportion of all the genes being expressed in a tissue and, consequently, gives us the complete picture of the biological processes active in the sample studied. The molecular basis for the different phenotypes (tissues, diseases, organs, etc.), can be attributed to those genes showing a significant differential expression among the phenotypes. This significance is usually obtained by means of a test that provides evidence that the observed differences in behaviour are unlikely to be observed by chance (the p -value). Nevertheless some differences in the gene behaviour with a clear biological meaning could occur at a level indistinguishable from a random difference and vice versa. If the significance in the distribution of biological processes or functions is used to determine which differences in the expression level are to be considered significant, the test becomes a tool to set a threshold with biological meaning; likewise, if the differential distribution of terms that relate to the biological process or molecular function is used to label sets of genes based on how much they differ (for example, between organs), then we can use these terms to try to understand the biological relevance of differential gene expression.

Gene Ontology provides a powerful source of information to be used to infer differences in gene expression based on biological background. We have used Gene Ontology to investigate the biological relevance of differential gene expression among three mouse organs, and will specifically use the differences between kidney and testis as our main example.

2. GENERAL OUTLINE OF THE METHODOLOGY

Pritchard *et al.* [2001] present microarray gene expression data for three different organs (testis, kidney, liver) of six different mice, with four replicates per organ per mouse. We want to examine what characterizes the genes that differ between the three organs or between pairs of organs (e.g., kidney vs. testis).

2.1 First step: sorting differentially expressed genes

The first step of our method analyzes the microarray experiment data to sort the genes according to how much they differ between organs. In this

case, we could use a linear model where we model gene expression as a function of the relevant experimental design features and organ. For instance, we could model

$$y_{ijkl} = \mu + dye_i + mouse_j + organ_k + error_{ijkl} \quad (1)$$

where y_{ijkl} is the $\log_2(\text{Experimental}/\text{Control})$ ratio that is usual in microarray data analysis, i is the index for dye ($i = 1$ meaning Cy3 on the experimental channel data, and $i = 2$ meaning Cy5 on the experimental channel data), j is the index for mouse (i.e., $j = \{1, 2, 3, 4, 5, 6\}$), k is the index for organ (e.g., $k = 1$ for kidney and $k = 2$ for testis), l is the index for replicate (i.e., $l = \{1, 2, 3, 4\}$ for each tissue within mouse), μ is a common intercept term, and $error_{ijkl}$ is the random error term. (Some more comments about this model, and its relation to other models, will be discussed below).

The greater the differences between the two organs, the larger the coefficient for organ should be; however, instead of looking directly at the coefficient, we should use the common t statistic from a linear regression model which is the coefficient divided by its standard error. This coefficient allows us to compare different genes: genes that are much more expressed in the kidney will have a large positive t statistic, and those that are much more expressed in the testis will have a very small (very large in absolute value, but with negative sign) t statistic.

In summary, in this first step we sort genes by how much they differ between organs.

2.2 Second step: formation of two groups of genes based on the sorted differences

In the second step we use this sorting information to group genes. We could form one group of genes with all those with $t > 15$ and a second group of genes with all those with $t < -15$. Thus, we would form a first group of genes that are much more expressed in the kidney and a second group that are much more expressed in the testis. Instead of comparing extreme groups (i.e., $t > 15$ against $t < 15$), we might want to compare those genes with $t > 15$ against all other genes; we would therefore be searching for differences between genes that are much more expressed compared to all other genes. The latter is the approach we will use, given that it allows us to highlight what makes a group of genes different (compared to all other genes). In contrast, if we only compare extreme groups (e.g., $t < -15$ vs. $t > 15$), the interpretation of the differences becomes more and more difficult as the threshold (the value of the t statistic) becomes smaller.

2.3 Third step: analysis of differential frequency of GO terms

In the third step we examine the relative frequency of Gene Ontology terms between the two groups that we formed in the second step. **Gene Ontology** (GO), [Ashburner *et al.*, 2000] provides a structured vocabulary for the annotation of genes and proteins. GO terms are structured in a hierarchy (actually, a directed acyclic graph), ranging from more general to more specific. Therefore, in these ontologies a term at a lower level (e.g., level 4) has one (or more) parent term(s) at the preceding level (level 3). GO is structured in three ontologies, corresponding to biochemical function, cellular processes, and cellular components. In Gene Ontology, data can be annotated at varying levels, depending on the information available.

In our study we first selected the GO level for our query. We chose level 3 as it is the best compromise between quantity and quality of GO information (Conde *et al.*, 2002). We queried the Unigene Id (for mouse genes) for each GO term at that level of resolution and we obtained the GO terms associated with each of our two groups of genes. For each GO term, we tested whether the two groups differ in the frequency of that GO term using Fisher's exact test for 2x2 contingency tables. For each GO term we can represent the data as a 2x2 contingency table with rows being presence/absence of the GO term and each column representing each of the two groups of genes. In other words, the numbers in each cell of the 2x2 contingency table are the number of genes of the first group where the GO term is present, the number of genes of the first group where the GO term is absent, the number of genes of the second group where the GO term is present, and the number of genes of the second group where the GO term is absent.

Fisher's exact test returns the p -value for that contingency table. However, we cannot directly use the individual p -value of each GO term, because we are testing multiple hypotheses, one for each GO term. If we were to use the p -value directly and declare all GO terms with a p -value < 0.05 as significantly differentially represented, we would have a great number of false rejections (i.e., we would end up considering as differentially represented many more GO terms than we should). Thus, we need to account for multiple testing.

There are several methods available to account for multiple testing (reviews in [Dudoit *et al.*, 2002a and b; and Westfall & Young 1993]). We have chosen to control the **False Discovery Rate (FDR)** using the method of Benjamini & Hochberg [1995]. The FDR is the expected proportion of false rejections (i.e., true null hypothesis that are incorrectly rejected)

relative to the total number of rejections. The control of FDR is probably a more appropriate method than other alternatives that control the overall Family Wise Error Rate (the probability that there is one or more false rejections over all the tests conducted), because it is less conservative, and thus it is more appropriate in studies that have a large exploratory component (Benjamini *et al.* 1999). In addition, compared to the resampling-based step-down minP and maxT methods of Westfall & Young [1993]; see also Dudoit *et al.*, [2002a and b], control of FDR using Benjamini & Hochberg's method does not require random permutation of data and is, therefore, much faster computationally. This is an important advantage in a case such as ours, where we repeat the process several hundred times using a sliding window. Using the FDR method, we only considered a GO term as being differentially represented in the two groups of genes if its FDR was smaller than 0.1 (we used the 0.1 level, instead of 0.05 because of the low power of Fisher's exact test and due to the exploratory nature of our study).

In summary, from the third step, we obtain GO terms which have a statistically significantly different relative proportion in the groups of genes.

We can now repeat steps two and three using a different threshold. For instance, if our first division in two groups used thresholds $t > 15$ and $t \leq 15$, we could now form another set of two groups by using instead $t > 10$ and $t \leq 10$. Now we would compare the relative frequency of GO terms between these two new groups. We can repeat the process for a whole set of thresholds. However, since setting up fixed thresholds in the absence of other knowledge might be somewhat arbitrary, we have instead used a sliding window. In this way, by moving over the set of ordered genes, we successively form groups of genes that can be compared to others¹.

In the sections that follow we provide additional details regarding the procedure, using as an example the comparison between kidney and testis. We later comment on alternative approaches that use slightly different steps one and two.

3. DATA PREPROCESSING, STANDARDIZATION, AND REDUCTION OF MISSING VALUES

We obtained the corrected data from Pritchard *et al.* [2001] from the CAMDA'02 web site. In addition, from the authors web site

¹ Strictly, we are no longer maintaining control of the family wise error rate, because of the sequence of tests, but we use this as a heuristic device to identify which set of groups have certain terms (e.g., gametogenesis, DNA binding) which are over- or under-represented.

(<http://www.pedb.org>) we downloaded “Mouse_array_merge_full.txt”, containing the Unigene Ids for each of the clones; we used the Unigene Ids provided by the authors to obtain the GO terms corresponding to each clone.

3.1 Preprocessing and normalization

We first subtracted background from foreground intensity levels for each sample, for both the experimental and control data. Next, using the information about the dye of the experimental channel data, we took the log (base 2) of the ratio experimental/control. Finally, we set all samples with a flag of -50 as missing data. To allow for comparisons among arrays, we standardized the data using the global median for each array (i.e., each of the \log_2 ratios was divided by the median \log_2 ratio of its array).

3.2 Reduction of missing values

Some analyses of organ effects can be seriously affected by an imbalance in the data sets [Milliken & Johnson, 1992; Miller, 1997], questioning the use of standard F-tests. In addition, and of particular importance for the analyses reported here, the interpretation and use of the t-tests from the regression models is simplified if there is balance in the organ sample sizes; finally, balance in the data makes it easy to compare the differences between pairs of organs.

Our main concern here is the imbalance in the representation of each organ. Thus, we have filtered the data to only use genes where the organ with the smallest sample size has a sample size that is at least 87.5% of the sample size of the organ with the largest sample size (in the case of a gene where at least one of the organs has no missing data, the organ with the largest number of missing data would be allowed to have at most three missing values: 21/24). After applying this filter, we are left with a total of 4140 genes. (This simple criterion additionally ensures that, of these, 3998 genes have at most a total of six missing values.) More restrictive criteria could be used (e.g., eliminating all those genes with a total of more than six missing values or filtering also with respect to mouse), but we have tried to achieve a balance between reliability and interpretability of results, ease of implementation of the procedure, and keeping most of the genes in the data set. Finally, after excluding those genes without Unigene Id information, we were left with 3736 clones for further analyses.

4. FITTING THE LINEAR REGRESSION MODEL

The model we fitted to kidney and testis data was discussed in section 2.1:

$$y_{ijkl} = \mu + dye_i + mouse_j + organ_k + error_{ijkl}. \quad (2)$$

In contrast to Pritchard *et al.* [2001], the dye term was fitted independently to each gene. This dye term can be included in the model because dye swapping was used in the experiment. We choose not to include interactions in this model to facilitate the interpretation of the coefficients. We parameterized the above model so that a large positive coefficient (and thus t statistic) for 'organ' means that the gene is much more expressed in the kidney and a large negative coefficient for organ means that the gene is much more expressed in the testis.

Some of the terms in our model are equivalent to those in the model by Kerr & Churchill [2001 a, b], where they use ANOVAs for the analysis of microarray data: in our analyses, the 'dye' term is equivalent to their DG interaction term, the 'mouse' term is equivalent to the AG interaction term, and the 'organ' term is equivalent to the VG interaction term. As in the models of Kerr & Churchill, we use the 'dye' and 'mouse' term to control for systematic effects that would, otherwise, be assigned to the error term (thus decreasing statistical power), but the 'dye' and 'mouse' terms are of no intrinsic interest. Our interest lies in the organ term, which is the one that measures the strength and direction of testis vs. kidney effect. Instead of the model above, we could have applied the mixed-effects model of Wolfinger *et al.* [2001], and we would then have used their GT term (where their T, for treatment, would have corresponded to our tissues).

We applied the model above to each of the 3736 genes and we divided genes in groups by choosing a threshold. We can interpret the meaning of this threshold examining the magnitude and sign of the t value. Large and positive values correspond to genes that are much more expressed in the kidneys and large and negative values to genes that are much more expressed in the testis. We used a sliding window of 150 genes, which was moved in jumps of 10 genes over the list of all genes ordered from minimum (largest and negative) t value to maximum t value. Given that the windows are of 150 genes and we move in jumps of 10 genes, there is an overlap of 140 genes between successive windows. A window of 150 genes was chosen as a compromise between low power of the Fisher test (see below and discussion) and too large a group that would become too heterogeneous. The number of genes with GO terms in this data was around 20%, so

choosing a window of 150 results in about 30 genes with GO in the given window (and about 700 genes with GO in the reference group). A smaller window might result in too few genes with GO in the window group and thus very low power to detect differences in the frequency of terms; on the other hand, too large a group might result in groups of genes that could be too heterogeneous to show a common pattern with respect to biological meaning. Similar qualitative results were obtained with a sliding window of 300 genes.

We might have preferred to use (unadjusted) p -values to order genes instead of the t statistic, because there are minor differences in the sample sizes, and thus degrees of freedom. We used the t statistic directly, however, because these differences in sample size are minor (see above), and thus the relation t statistic-- p -value is essentially the same for most t statistics; in this case, using the t statistic results in a much simpler procedure with a more straightforward interpretation than using the p -values taking into account the sign of the t statistic. Moreover, we are using the threshold as a device to form groups, but the ultimate judgment concerning the relevance of the groups is derived from the results with the GO terms.

4.1 Difference in the frequency of GO terms in each group

For each sliding window we compared the frequency of GO terms of the genes included in the window vs. the frequency of GO terms in the genes outside the window. In other words, we examined if there were significant differences (adjusted for multiple testing) in the representation of GO terms in the two groups. We only considered those GO terms that had an FDR-adjusted p -value of less than 0.1² as differentially represented. We used the 0.1 level, instead of 0.05 because of the low power of Fisher's exact test and due to the exploratory nature of our study. In addition, we only show those terms that had an adjusted p -value of less than 0.1 in at least three windows, to try to minimize spurious significant results. The results are shown in Figure 1.

The results from molecular function at level 4 are straightforward: the terms "globin", "oxidoreductase, CH-OH group", and "oxygen transporter" are significantly more common, with a difference of about 10%, in those genes that are much more strongly expressed in the kidney. The results for molecular function at level 3 are much richer; genes that are much more expressed in the kidney show more abundance of the "oxygen binding" and "oxidoreductase" terms but show a smaller frequency of the "hydrolase"

² Similar qualitative results were obtained using 0.05 instead of 0.1 as the adjusted p -value cut-off.

and “nucleic acid binding” terms. Genes that are more strongly expressed in the kidney also have a somewhat higher frequency of the term “blood coagulation factor.” Genes more strongly expressed in the testis show higher prevalence of the terms “structural constituent of ribosome,” “structural constituent of cytoskeleton” and “ribonucleoprotein.” Finally, genes that are not expressed differentially between the kidney and testis show a higher frequency of the GO term “neurotransmitter binding,” which might be a case of a false positive. Most of these results are in agreement with our biological knowledge of the differences between these two organs.

Regarding the biological process, the results for level 3 indicate that the genes that are much more expressed in the kidney have a higher prevalence of the term “transport” but a smaller frequency of the term “metabolism,” whereas those much more expressed in the testis have a higher frequency of the term “reproduction.” It is somewhat surprising, however, that genes that are not differentially expressed show a lower frequency of the term “metabolism.” Level 4 unravels a more complex picture. Genes much more expressed in the testis show a higher frequency of the term “gametogenesis” (whose parent term at level 3 is “reproduction”), and genes more expressed in the testis, but not so extremely, show a higher frequency of the term “M phase” and “muscle contraction.” Genes much more expressed in the kidney show a higher frequency of “gas transport” (whose parent term at level 3 is “transport”) and a lower frequency of “nucleobase, nucleoside, ...” (whose parent term is “metabolism”). Genes moderately more expressed in the kidney are enriched in the term “phosphate metabolism,” and those genes only slightly more expressed in the kidney have a higher frequency of the term “pattern development.”

Again, these results agree with our biological knowledge. Notice how the genes that show the highest expression in the kidney compared to the testis are those which have an under-representation of GO terms related to the processes involved in the formation of spermatozooids, but show an overrepresentation of terms involved in transport. Conversely, genes that show the highest expression in the testis have an over-representation of terms related to reproduction and cell cycle (reproduction, gametogenesis, M phase.) Interestingly, other terms (pattern specification and phosphate metabolism) seem to be over-represented in genes that are more (but not extremely more) expressed in the kidney than in the testis. Metabolism being under-represented in genes that show almost no difference between the two organs might be either a false positive, or an artifact due to these sets of genes being enriched in genes with very low expression in most conditions, and thus unlikely to be involved in metabolic processes (which are common to all cells.)

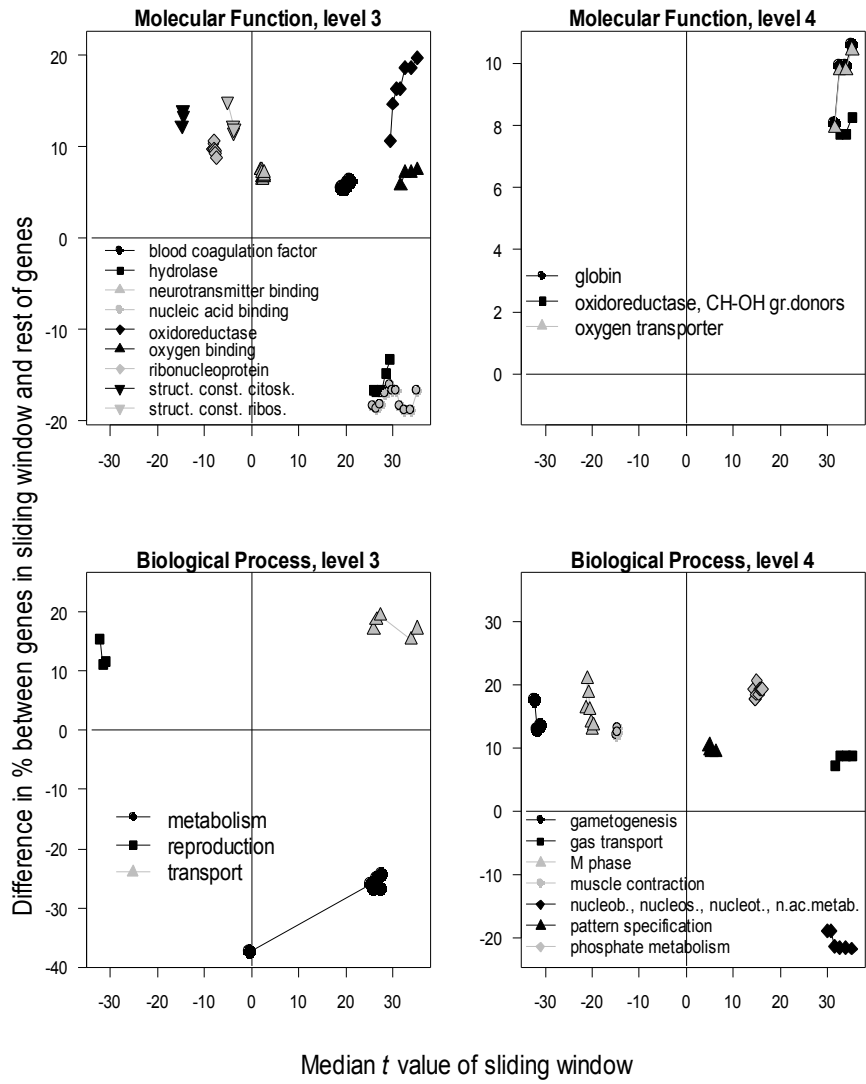


Figure 1. Results for all GO terms that had at least three adjusted p-values < 0.1. Large negative t values correspond to genes much more expressed in the testis, and large positive t values to genes much more expressed in the kidney.

5. ALTERNATIVES TO STEPS ONE AND TWO

5.1 An alternative way of comparing two organs

Instead of using the signed t statistic, we could form groups based upon the p -value. We would use a sliding window over the genes ordered by increasing p -value, so that as we move over the window we would progressively examine genes that are less differentially expressed between organs, regardless of whether they are more or less expressed in one organ or the other. Here we would be comparing differentially expressed genes against the reference group of all other genes, whereas with the linear regression approach we compared genes that are more expressed, with similar strength, in one organ vs. the rest of the genes.

Related to this approach, instead of using the t statistic from the model in equation (1) we could use the F-ratio for organ; genes with a large F-ratio for organ will be those where there are large differences in expression among organs. Of course, using the p -value from the linear regression model above would be equivalent to fitting an ANOVA model and using the p -value for organ.

The problem with these approaches is that, when comparing two groups, the F ratio and the p value do not provide information on directionality of the difference ---i.e., whether it is the testis or the kidney that show larger expression. However, most biologists will generally be interested in the directionality of the difference because it allows for much easier interpretation; thus the use of the t statistic will generally be more appropriate than the use of the F ratio or the p value.

Another alternative is not to use a sliding window, but to compare successive pairs of extreme groups of genes as defined by a threshold; for instance, we could compare genes with a t value < -20 vs. genes with a t value > 20 or, in other words, genes that are much more expressed in the kidney vs. genes that are much more expressed in the testis. However, we suspect that using the sliding window approach we have employed here might better highlight specific biological aspects of groups of genes that differ, in a similar way, between the two organs.

5.2 Comparing three organs

We could fit an ANOVA model for the three organs simultaneously, and using a sliding window over the ordered F values, order genes based on how

strongly they show differential expression among the three organs. This is similar to the use of p -values in the previous section, except interpretation might be more complicated because, in addition to losing information on directionality (in which organ is expression the greatest?), the comparison is among three organs.

5.3 Are there multiple testing issues when sorting genes?

Multiple testing issues are not very relevant in the second step, since the ultimate arbiter of the biological significance is the analysis of GO terms. Threshold values for the statistics could be chosen so that they are significant even when considering multiple testing, but genes that are statistically significantly different are not necessarily biologically significant or relevant.

6. DISCUSSION: MERITS, LIMITATIONS, CONCLUSION

The suggested procedure is easy to use. The first two steps can be easily implemented, for instance, using a statistical package with programming facilities; we have used R, but other statistical packages are probably also suited. The third step might seem the most complicated, but we have built a tool, FatiGO [Al-Shahrour *et al.*, in prep.], now available at <http://fatigo.bioinfo.cnio.es>, that carries this out (searching for GO terms from a list of gene id's ---Unigene Id or Gene Symbol---, and performing Fisher's exact test plus the permutation test for step-down minP multiple testing).

There are two main limitations of the method. The first one is related to the amount and quality of annotation. On the one hand, many genes have no GO annotation; for instance, at threshold 30, only 25 of the 88 genes from the kidney group have an annotation at level 4. In these examples, about 16% to 25% of the genes are annotated at level 4, and about 20% to 30% are annotated at level 3. Of course, the validity of the method with non-annotated genes relies on the non-annotation being independent of the GO term (i.e., that a gene not having a GO annotation be due to reasons unrelated to the value that the GO term would have had). One of the effects of this lack in data is that it makes it hard to detect GO terms that have different frequencies between two groups (an effect that is probably more serious when the original groups of genes are already small ---e.g., with a

threshold of 30 in our case). This can also explain some inconsistencies and the fact that some terms might be significant at a given threshold, not significant at the next, and significant again at the following threshold. Finally, this limitation, as it affects smaller groups more strongly, might have its most adverse effects when examining the most extreme differences in expression, which might be those that are of more interest to us. On the other hand, the quality of annotation varies: some genes are annotated manually and carefully, whereas other genes are annotated automatically and based on similarities with other genes, which can lead to dubious annotations (Gene Ontology includes information related to the type of annotation that allows ranking of quality of annotation). We expect some of these limitations to be of lesser importance with time as more genes are annotated, and the annotation of other genes is reviewed. In the present analyses, we have ignored the issues of lack of annotation and of dubious annotations. By using a Bayesian approach, however, it should be possible to explicitly deal with (incorporate) the lack of annotation and annotations of different quality. We are currently working on this.

The second limitation was previously mentioned: since we repeat the process of examining the significance of GO terms many times (one for each sliding window), we are no longer maintaining strict control of the false discovery rate. Providing strict control of the FDR over all the comparisons is probably unwarranted. We suggest that our methodology be used as a heuristic device to identify for what set of groups certain terms (e.g., gametogenesis, DNA binding) are over- or under-represented (providing control of the error rate at each of the threshold levels.) The patterns identified should probably show consistency across adjacent threshold levels and between gene ontology levels, whereas spurious results (due to chance from multiple testing) are likely to show up as occasional significant results that are not consistent across thresholds or gene ontology levels. Given the exploratory nature of the method, and the loss of power associated with the relative scarcity of GO annotations, we prefer to decrease Type II errors (missing a real biological difference) even if that means slightly increasing Type I error rates (declaring significant a term that is not really differentially represented); this way, we provide an increased ability to understand the biological meaning of the detected differences, while limiting (via control of FDR at each window) the number of false detections.

6.1 Similar approaches

Our approach is similar to many other recent proposals that are trying to incorporate annotation data to gene expression data [e.g., Zhou *et al.*, 2002;

Gibbons and Roth, 2002; etc]. The most similar was that suggested by Pavlidis *et al.* [2002]. These authors first use a statistical model (in their case an ANOVA) to obtain the p -values for each gene. While Pavlidis *et al.* [2002] compared tumor types and brain regions, we could compare pairs of organs, or the three organs together with an ANOVA. They compute the score for each “GO class,” where a GO class is the group of genes that share a GO term; the score for a GO class is the average of the minus \log_{10} p -value of the genes in the GO class, so that GO classes with a large score are composed of genes with lower p -values. Finally, Pavlidis *et al.* select classes using those with a relevant score based on a permutation test. Their approach differs from ours on several counts. First, since the p -value is based on an ANOVA, directionality information is lost, so a given GO class can be composed of some genes that are much more expressed in one organ and some other genes that are much more expressed in other organs. This, however, could be solved using the t statistic when comparing two classes, and requiring GO classes to have only t statistics with the same sign. Second, and more importantly, their approach is not designed to answer the general question “what GO terms differentiate this set of genes from that set of genes,” but the question “which GO terms are associated with genes that, on average, show strong differential expression.” Note that the approach of Pavlidis *et al.* might detect a GO class that includes some genes that show no differential expression between conditions (if the other genes in the GO class do show strong effects). Thus, their approach is a different, and complementary, way of incorporating GO information to gene expression data.

To conclude, we think that our proposed method is promising in identifying biological features of genes that are differentially expressed between/among organs or conditions, since it incorporates information on the known (annotated) biological processes and molecular functions associated with the genes that belong to different groups.

ACKNOWLEDGMENTS

A. Dopazo and L. Lombardía answered questions about GENEPIX. We also thank assistants at CAMDA '02 for their thought-provoking questions, and the editors and anonymous reviewers for comments that improved the manuscript. A. Wren revised the English. R.D-U is contracted by the Ramón y Cajal programme from the MCYT. F. A. is supported by contract BIO2001-0068 from the MCYT.

REFERENCES

- Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Traver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., and Rubin, G. S. G. 2000. Gene ontology: tool for the unification of biology gene ontology: tool for the unification of biology. *Nat. Genet*, 25:25--29.
- Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society, Series B*, 57: 289-300.
- Benjamini, Y., Drai, D., Kafkafi, N., Elmer, G., Golani, I. 1999. Controlling the false discovery rate in behavior genetics research. (available from <http://www.math.tau.ac.il/~ybenja>).
- Brown, P., O., and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Biotechnol*, 14:1675--1680.
- Conde, L., Mateos, Á., Herrero, J. & Dopazo, J. 2002
Unsupervised reduction of the dimensionality followed by supervised learning with a perceptron improves the classification of conditions in DNA microarray gene expression data. *Neural Networks for Signal Processing XII*. IEEE Press (New York). Eds. Bouldard, Adali, Bengio, Larsen, Douglas. pp. 77-86
- Dudoit, S., Shaffer, J. P., Boldrick, J. C. 2002 Multiple hypothesis testing in microarray experiments. Technical report # 110. Division of Biostatistics, UC Berkeley.
- Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. 2002 Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12: 111-139.
- Gibbons, F. D., Roth, F. P. 2002. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12: 1574-1581.
- Kerr, M. K., and Churchill, G. A. 2001 a. Experimental design for gene expression microarrays. *Biostatistics*, 2: 183-201.
- Kerr, M. K. and Churchill, G. A. 2001 b. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77: 123-128.
- Miller, R. G.. 1997. *Beyond Anova* Chapman & Hall.
- Milliken, G. A. and Johnson, D. E. 1992. *Analysis of Messy Data*. Chapman & Hall.
- Pavlidis, P., Lewis, D. P., Noble, W. S. 2002. Exploring gene expression data with class scores. *Proc. Pacific Symp. Biocomputing*, 2002: 474-485.
- Pritchard, C. C., Hsu, L., Nelson, P. S. 2001. Project normal: defining normal variance in mouse gene expression. *PNAS*, 98:13266—13271.
- Westfall, P. H. and Young, S. S. 1993. *Resampling-based multiple testing: examples and methods for p-value adjustment*. John Wiley & Sons.
- Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. S. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8: 625-637.
- Zhou, X., Kao, M.-C. J., Wong, W. H. 2002. Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, 99: 12783-12788.