

## Ontologies and functional genomics

*Fátima Al-Shahrour and Joaquín Dopazo*

*Bioinformatics Unit, Spanish National Cancer Centre (CNIO), Madrid, Spain*

### **Abstract**

High-throughput methodologies have increased by orders of magnitude the possibility of obtaining data in orders of magnitude. Nevertheless, translating data into useful biological knowledge is not an easy task. We review how bioontologies, and in particular, gene ontology, can be used to understand the biological roles played by genes that account for the phenotypes studied, which is the ultimate goal of functional genomics. Statistical issues related to high-throughput methodologies, such as the high occurrence of false or spurious associations, are also discussed.

### **Keywords**

Gene ontology, multiple testing, annotation, functional genomics

### **7.1. Information mining in genome-wide functional analysis**

Molecular biology has addressed functional questions by studying individual genes, either independently or a few at a time. Despite it constituted a reductionistic approach, it was extremely successful in assigning functional properties and biological roles to genes and gene products. The recent possibility of obtaining information on thousands of genes or proteins in one sole experiment, thanks to high throughput methodologies such as gene expression (Holloway et al., 2002) or proteomics (MacBeath, 2002), has opened up new possibilities in querying living systems at the genome level that are beyond the old paradigm “one-gene-one-postdoc”. Relevant biological questions regarding gene or gene products interactions or biological processes played by networks of components, etc., can now for the first time be addressed realistically. Nevertheless, genomic technologies are at the same time generating new challenges for data analysis and demand a drastic change

in the habits of data management. Dealing with this overabundance of data must be approached cautiously because of the high occurrence of spurious associations if the proper methodologies are not used (see Ge et al., 2003, and chapter 12 for discussions in some related aspects).

Traditional molecular biology approaches tended to mix up the concepts of data and information. This was partially due to the fact that researchers had a great deal of information previously available about the typical data units they used (genes, proteins, etc.). Over the last few years the increasing availability of high throughput methodologies has amplified in orders of magnitude the potential of data production. One direct consequence of this revolution in data production has been to clarify how fictitious the equivalence between data and information actually used to be. Systems biology approaches emerge then to convert the flood of data into information and knowledge (Bassett et al., 1999; Ge et al., 2003).

There are, however, several problems related to massive data management. One of them is the lack of accurate functional annotations for a considerable number of genes. Another non-negligible difficulty stems from the fact that, even in the instance of availability of proper functional annotations, processing all the information corresponding to thousands of genes involved in a high-throughput experiment is beyond the human capabilities. Automatic processing of the information therefore becomes indispensable to draw out the biological significance behind the results in this type of experiment. As previously mentioned, the occurrence of false or spurious associations is common when dealing with thousands of elements. Unfortunately, these spurious associations are often considered as evidence of actual functional links, leading to misinterpretation of results. All these features of genomic data must be taken into account for any procedure aiming to properly identify functional roles in groups of genes with a particular experimental behaviour.

## **7.2. Sources of information: free text versus curated repositories.**

Any approach using biological information for functional annotation purposes uses two main sources: free text or curated repositories.

The use of techniques of automatic management of biological information to study the coherence of gene groups obtained from different methodologies has been addressed in

recent years (Oliveros et al., 2000; Raychaudhuri et al., 2002b; Pavlidis et al., 2002). Considerable effort has been focused on developing automatic procedures for extracting information from biomedical literature. Information extraction and text mining techniques in particular have been applied to the analysis of gene expression data (Jenssen et al., 2000; Oliveros et al., 2000, Tanabe et al., 1999). It has been claimed that free text processing, essentially using PubMed abstracts as a source of information, has the advantage of providing numerous gene-to-abstract correspondences. Nevertheless, text mining methodologies still present many drawbacks (Blaschke et al., 2002) such as: problems of interpreting terms due to the context of the sentence in which the gene is cited; the lack of a standardised nomenclature of genes in literature that makes it difficult to find all the citations for all the synonyms used for them; there is a profuse use of acronyms in literature that makes it hard to find accurate citations of genes (for example, for the term STC, corresponding to gene *secretin*, 158 citations were found in the year 2001, 121 of them corresponded to Stem Cell Transplantation and the rest to other concepts such as Spiral Computerised Tomography, Solid Cystic Tumour, etc., and only one of them was a real reference to the gene *secretin*); there are problems surrounding orthography and boundaries for identifying gene names; and finally there are irrelevant terms related to non-functional features that appear, nonetheless to be associated to gene names.

On the other side of the spectra are the repositories with curated functional information, which contain less gene-to-term correspondences although these are reliable, consistent and standardised. There are diverse repositories such as pathway databases, among which KEGG database (Kanehisa et al., 2004) is the paradigm, protein interaction databases (see DIP, Xenarios et al., 2002), protein motifs databases (see, for example, InterPro, Mulder et al., 2003), etc.

The most valuable resource is most probably the Gene Ontology (GO) database of curated definitions (Ashburner et al., 2000) and the annotations based on GO.

Diverse genome initiatives and databases are annotating genes according to GO terms (Camon et al., 2003; Xie et al., 2002), constituting a priceless resource for information mining implementations.

Although some direct applications of free-text mining to data analysis have been proposed (Tanabe et al., 1999; Oliveros et al., 2000; Raychaudhuri et al., 2003), the future of the practical application of these technologies probably resides in its use by information repositories curators to help in the annotation process. Methods for predicting GO categories from the analysis of biomedical literature (Raychaudhuri et al., 2002a) have therefore been proposed and there are also similar methods based on the study of different biochemical and physical protein features (Jensen et al., 2003; Schug et al., 2002).

### **7.3. Bio-ontologies and the Gene Ontology in functional genomics**

Applied ontologies are centred around a specific domain of knowledge. These ontologies endeavour to represent a system of categories accounting for a particular vision of a given area, in order to establish rules that describe relationships between these categories, and to instantiate the objects in the categories. In practical terms, these ontologies provide an organizational framework of concepts about biological entities and processes in a hierarchical system in which, associative relations which provide reasoning behind biological knowledge, are included. One of the most powerful features of an ontology is the implementation of a controlled, unambiguous vocabulary. This is extremely useful in an inherently complex and heterogeneous discipline such as biology, where a great deal of sophisticated knowledge, in most cases of a hierarchical nature, needs to be integrated with molecular data (Bard and Rhee, 2004). The most important ontologies in the domain of biology are included under the umbrella of the Open Biological Ontologies (OBO) initiative, which constitutes a *de facto* standard for them (see <http://obo.sourceforge.net/>). Some known examples are the Unified Medical Language System (UMLS, <http://www.nlm.nih.gov/research/umls/>), which implements a hierarchy of medical terms used in the indexation of PubMed, or the Microarray Gene Expression Data Society (MGED) Ontology (<http://mged.sourceforge.net/>), recently popularised by the increasing production of gene expression data with microarrays, etc. Nevertheless, the most relevant ontology in the area of functional genomics is, undoubtedly, the Gene Ontology (GO, <http://www.geneontology.org/>), which provides a controlled vocabulary for the description of molecular function, biological process and cellular component of gene

products (Ashburner et al., 2000). GO terms are used as attributes of gene products by collaborating databases, facilitating uniform queries across them. Because of the existing homologies between proteins among different taxa, GO terms can be thoroughly used across species (Ashburner et al., 2000).

The controlled vocabularies of terms are structured in a hierarchical manner that allows for both attribution (assignment of gene products to particular terms) and querying at different levels of granularity. This hierarchical structure constitutes the representation of the ontology within which each term is a node of a Directed Acyclic Graph (DAG), which is very similar to a tree - the only difference being that in a DAG it is possible for a node to have more than one parent. The deeper a node is in the hierarchy, the more detailed the description of the term. In GO, child to parent relationships can be of two types: “is a”, meaning the child is an instance of the parent (e.g. *chloroplast envelope* GO:0009941 is a *membrane* GO:0016020) and “part of” when the child is a component of the parent (e.g. *inner membrane* GO:0019866 and *outer membrane* GO:0019867 are part of *membrane* GO:0016020).

The success of an ontology relies largely upon the approval received from the scientific community. The most important achievement of GO is perhaps that the GO consortium has been able to attract a large number of collaborating databases which are actively mapping gene products onto GO terms. These databases with controlled and curated annotations, which can easily be queried by computers, constitute an invaluable though not yet fully exploited resource, for the scientific community. Additional information on the quality of the annotation of gene products is provided by the collaborative databases through the evidence codes (<http://www.geneontology.org/GO.evidence.html>). The codes represent different types of evidence used in the annotation. Among them, the highest quality codes are for GO-gene correspondences supported by experimental functional assays (IDA, IMP codes) and the lowest quality corresponds to correspondences inferred from electronic annotations (IEA code).

As previously mentioned, the representation of GO resides in its hierarchy. There are different tools available that are useful to browse this hierarchy (see a comprehensive list in: <http://www.geneontology.org/GO.tools.html>). Such GO browsers allow the viewing of all gene products annotated with a given GO term, or searching for a gene product and

view all its associations. In addition, by browsing the ontologies it is also possible to view relationships between terms.

#### **7.4. Using GO to translate the results of functional genomic experiments into biological knowledge.**

Functional genomics experiments allow the scaling of the classical functional experiments to a genomic level. Comparison of phenotypes (e.g. patients versus controls, studies of different clinical outcomes, etc.) by means of techniques such as DNA microarrays or proteomics provides insight into their molecular basis. Nevertheless, the data obtained in these experiments are measurements of the gene or protein expression levels. To translate this data into information numerical analyses are firstly required to determine which genes (among the thousands analysed) can be considered as significantly related to the phenotypes (see chapter 12). The second step is to interpret roles played by the targeted genes. The availability of GO annotations for a considerable number of genes helps interpret these results from a biological point of view. The rationale commonly used is as follows: if some genes have been found to be differentially expressed when comparing two different phenotypes (or are correlated to a given continuous phenotypic trait, or to survival, etc.) it is because the roles they play at molecular level account (to some extent) for the phenotypes analysed. The GO annotations available for the genes that present the same asymmetrical distribution or correlation serve as a more or less detailed description of these biological roles. For example, if 50 genes from an array of 6,500 genes are differentially expressed and 40 of them (80% - a high proportion) are annotated as response to “external stimulus” (GO:0009605), it is intuitive to conclude that this process must be related to the phenotypes studied. In addition, if the background distribution of this type of gene in the genome is, lets say, of the 4%, one can conclude that most of the genes related to “external stimulus” have been altered in their expression levels in the experiment.

There are many tools listed on the GO consortium web page that extract lists of GO terms differentially represented when comparing two sets of genes (see <http://www.geneontology.org/GO.tools.html>) and, in some cases, provide scores or even individual tests for comparisons between two sets of genes. For example, GoMiner

(Zeeberg et al., 2003), MAPPFinder (Doniger et al., 2003), GFINDER (<http://www.medinfopoli.polimi.it/GFINDER/>) or eGON (<http://nova2.idi.ntnu.no/egon/>), just to cite a few, generate tables that correlate groups of genes to biochemical, molecular functions or GO terms. Some of them are specific for organisms, such as FunSpec (Robinson et al., 2002), which evaluates groups of yeast genes in terms of their annotations in diverse databases or CLENCH (Shah and Fedoroff, 2004) for *A. thaliana*. Nevertheless, differences in the distribution of GO terms between groups must, in addition to being spectacular (which is quite a subjective concept), be also significant (which is an objective statistical concept related to the probability of drawing your observations purely by chance).

### **7.5. Statistical approaches to test significant biological differences.**

As previously mentioned much caution should be adopted when dealing with a large set of data because of the high occurrence of spurious associations (Ge et al., 2003). Table 1 has been constructed using ten datasets obtained by the random sampling of 50 genes from the complete genome of *Saccharomyces cerevisiae*. For each random set, the proportions of all the GO terms (at GO level 4) have been compared between both partitions (50 genes with respect to the remaining ones), and the GO term showing the most extreme differential distribution was displayed in each case (rows of the table). The first column shows the percentage of genes annotated with the GO term in the random partition of 50 genes, the second column represents the corresponding percentage in the rest of the genome and the third column shows the p-value obtained upon the application of a Fisher's exact test for 2x2 contingency tables. For many people it still seems staggering that most of the random partitions present asymmetrical distributions of GO terms with significant individual p-values (column 3). This apparent paradox stems from the fact that we are not conducting a single test in each partition, but as many tests as GO terms are being checked (several hundreds). Nevertheless, in this situation the researcher tends to forget about the many hypothesis rejected and only focus on the term for which an apparent asymmetrical distribution was found. In some cases this situation is caused by the way in which some of the above mentioned programs work. To some extent the fact that many tests are really being conducted is hidden to the user and the result is presented as if it were the case of a unique test. If we conduct several hundreds of tests

simultaneously, the probability of finding an apparently asymmetrical distribution for a given GO term increases enormously. A very simple example can be used here to illustrate this concept: let's imagine you flip a coin 10 times and you get 10 heads. You would certainly suspect that something was wrong with the coin. If the same operation was repeated with 10.000 different coins one or even several occurrences of 10 heads would not be considered surprising. We intuitively accept this because of the probability of having an unexpected result just by chance is high. If we were interested in checking whether an observation is significantly different from what we could expect simply by chance in a multiple testing situation then the proper correction must be applied. The fourth column of Table 1 shows an adjusted p-value using one of the most popular multiple-testing corrections, the False discovery Rate (FDR; Benjamini, Yekutieli 2001), and it is obvious that none of the situations depicted in columns 1 and 2 can be attributed to anything other than random occurrence.

**Table 1.** GO terms found to be differentially distributed when comparing ten independent random partitions of 50 genes sampled from the complete genome of yeast. See text for an explanation.

% in random set	% in genome	p-value	adjusted p-value	GO term
8.33	1.86	0.0752	1	ion homeostasis (GO:0050801)
10.00	31.34	0.0096	0.6735	nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0006139)
3.33	0.24	0.075	1	One-carbon compound metabolism (GO:0006730)
4.04	8.00	0.0177	0.6599	energy pathways (GO:0006091)
3.45	0.22	0.0669	1	metabolic compound salvage (GO:0043094)
5.88	0.67	0.024	1	vesicle fusion (GO:0006906)
6.45	1.60	0.09	1	negative regulation of gene expression, epigenetic (GO:0045814)
13.79	3.97	0.028	1	response to external stimulus (GO:0009605)
16.13	4.23	0.0097	1	response to endogenous stimulus (GO:0009719)
2.70	0.13	0.054	1	host-pathogen interaction (GO:0030383)

Table 1 shows how random partitions, for which no functional enrichment should be expected, yield apparent enrichments in GO terms because the most asymmetrically

distributed GO term among several hundreds are chosen *a posteriori*. These values occur simply by chance and cannot be considered as either biologically authentic or statistically significant. This clearly shows beyond any doubt, that multiple testing adjustment must be used if several hypotheses are simultaneously tested.

Multiple testing has been addressed in different ways depending on particular cases and the number of simultaneous hypotheses tested. Thus, corrections such as Bonferroni or Sidak are of very simple application but are too conservative if the number of simultaneous tests is high (Westfall and Young, 1993). Another family of methods that allow less conservative adjustments are the family wise error rate (FWER), that controls the probability that one or more of the rejected hypotheses (GO terms whose differences cannot be attributed to chance) is true (that is, a false positive). The minP step-down method (Westfall and Young, 1993), a permutation-based algorithm, provides a strong control (i.e., under any mix of false and true null hypothesis) of the FWER. Approaches that control the FWER can be used in this context although they are dependent on the number of hypotheses tested and tend to be too conservative for a high number of simultaneous tests. Aside from a few cases in which FWER control could be necessary, the multiplicity problem in prospective functional assignment does not require protection against even a single false positive. In this case, the drastic loss of power involved in such protection is not justified. It would be more appropriate to control the proportion of errors among the identified GO terms whose differences among groups of genes cannot be attributed to chance instead. The expectation of this proportion is the False Discovery Rate (FDR). Different procedures offer strong control of the FDR under independence and some specific types of positive dependence of the tests statistics (Benjamini and Hochberg ,1995), or under arbitrary dependency of test statistics (Benjamini and Yekutieli 2001).

We have shown how important multiplicity issues are in finding functional associations to clusters of genes. Any procedure that does not take this into account is as consequence considering a high number of spurious relationships as reliable.

## **7.6. Using FatiGO to find significant functional associations in clusters of genes.**

The FatiGO (Fast Assignment and Transference of Information using GO, available at <http://fatigo.bioinfo.cnio.es>) tool was the first application for finding significant differences in the distribution of GO terms between groups of genes taking the multiple testing nature of the contrast into account (Al-Shahrour et al., 2004). FatiGO takes two lists of genes (ideally a group of interest and the rest of the genome, although any two groups, formed in any way, can be tested against each other) and convert them into two lists of GO terms using the corresponding gene-GO association table. Since distinct genes are annotated with more or less detail at the different levels of the hierarchy, it is meaningless to test for different terms that are really descriptions in different detail of the same functional property (e.g. why test apoptosis versus regulation of apoptosis?). To deal with this, FatiGO implements the “inclusive analysis”, in which a level in the DAG hierarchy is chosen for the analysis. Genes annotated with terms that are descendant of the parent term corresponding to the level chosen therefore take the annotation from the parent. Figure 1 illustrates this procedure. If the level corresponding to, for example, apoptosis was selected, any gene annotated as either apoptosis or as any children term was considered in the same category (apoptosis) for the test. This increases the power of the test. There are less terms, each with more genes, to be tested.

A Fisher's exact test for 2x2 contingency tables is used. For each GO term the data are represented as a 2x2 contingency table with rows being presence/absence of the GO term, and each column representing each of the two clusters (so that the numbers in each cell would be the number of genes of the first cluster where the GO term is present, the number of genes in the first cluster where the GO term is absent, and so on).

In addition to the unadjusted p-values (which are given just because they are obtained as part of the process, but should not be considered as evidence of significant differential distribution of GO terms between clusters), FatiGO returns adjusted p-values based on three different ways of accounting for multiple testing: FDR under independence (Benjamini and Hochberg, 1995), or under arbitrary dependency of test statistics (Benjamini and Yekutieli, 2001) as well as FWER control by the minP step-down method (Westfall and Young, 1993). Results are arranged by p-value to facilitate the

identification of GO terms with a significant asymmetrical distribution between the groups of genes studied.

### **7.7. Other tools**

Recently, other tools have included some multiple-testing possibilities. For example, the latest versions of Onto-Express (Khatri et al., 2002) include Bonferroni and Sidak corrections as well as a permutation test, or GeneMerge (Castillo-Davis and Hartl 2003), which implements Bonferroni correction. New tools such as FunAssociate (<http://llama.med.harvard.edu/cgi/func/funcassociate>) include unspecified permutation tests, although other ones include more established multiple testing controls such as FDR, which is the case of GoSurfer (<http://biosun1.harvard.edu/complab/gosurfer/>) or GOSTat (Beissbarth and Speed, 2004) that has exactly the same functionalities than FatiGO.

### **7.8. Examples of functional analysis of clusters of genes**

As previously mentioned a research scientist is continuously interested in understanding the molecular roles played by potentially relevant genes in a given experiment. One of the most popular hypothesis in microarray data analysis is that coexpression of genes across a given series of experiments is most probably explained through some common functional role (Eisen et al., 1998). Actually, this causal relationship has been used to predict gene function from patterns of co-expression (van Noort, 2003; Mateos et al., 2002).

Here is an example using the data from DeRisi et al., (1997) in which they analyse the complete genome of *Saccharomyces cerevisiae* to carry out a comprehensive study of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. With the aim of finding groups of genes that coexpress across the seven time points measured, gene expression patterns were clustered using the SOTA algorithm (Dopazo and Carazo, 1997; Herrero et al., 2001; see also chapter 10) as implemented in the GEPAS (<http://gepas.bioinfo.cnio.es>) suite of web tools (Herrero et al., 2003). Figure 2 shows the clusters of genes obtained. In the example, a set of genes was selected and analysed with FatiGO. 75% of the genes were annotated as biosynthesis and the differences in proportion with respect to the rest of genes (30%) is clearly

significant. It can be claimed that genes with the described temporal behaviour are involved in biosynthesis biological process. In the event of not performing p-value adjustment, another three processes (sexual reproduction, conjugation and aromatic compound metabolism) would have been considered as important despite the differences in the proportions between the cluster and the rest of genes that can occur simply by chance (the adjusted p-values are too high).

Genes showing significant differential expression when comparing two or more phenotypes, or genes significantly correlated to a trait (e.g. the level of a metabolite) or to survival, can be analysed in the same way. Comparison of distributions of GO terms helps to understand what makes these genes different from the rest.

### **7.9. Future prospects**

The importance of using biological information as an instrument to understand the biological roles played by genes targeted in functional genomics experiments has been highlighted in this Chapter. There are situations in which the existence of noise and/or the weakness of the signal hamper the detection of real inductions or repressions of genes. Improvements in methodologies of data analysis, dealing exclusively with expression values can to some extent help (see chapter 12). Recently, the idea of using biological knowledge as part of the analysis process is gaining in support and popularity. The rationale is similar to the justification of using biological information to understand the biological roles of differentially expressed genes. What differs here is that genes are no longer the units of interest, but groups of genes with a common function. Let us consider a list of genes arranged according their degree of differential expression between two conditions (e.g. patients versus controls). If a given biological process is accounting for the observed phenotypic differences we should then expect to find most genes involved in this process overexpressed in one of the conditions against the other. Contrarily, if the process has nothing to do with the phenotypes, the genes will be randomly distributed amongst both classes (for example if genes account for physiological functions unrelated to the disease studied, they will be active or inactive both in patients and controls). Diaz-Uriarte et al. (2003) proposed the use of a sliding window across the list of genes for comparing the distribution of GO terms corresponding to genes within the window

against genes outside the window. If terms (but not necessarily individual genes) were found differentially represented in the extremes of the list, one can conclude that these biological processes are significantly related to the phenotypes. Al-Shahrour et al. (2003) generalized this approach to other types of arrangements based on other types of experiments. Recently, Mootha et al. (2003) proposed a different statistic with the same goal. This is part of a more general question, which would be the study of differences on prespecified groups of genes, which is discussed in chapter 12.

Different creative uses of information in the gene selection process as well as the availability of more detailed annotations will enhance our capability of translating experimental results into biological knowledge.

## References

Al-Shahrour, F., Díaz-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., Herrero, J., Mateos, Á., Santoyo, J., Díaz-Uriarte, R. & Dopazo, J. (2003) Using Gene Ontology on genome-scale studies to find significant associations of biologically relevant terms to group of genes. *Neural Networks for Signal Processing XIII. IEEE Press (New York)*. Pp. 43-52. (see technical report in: <http://bioinfo.cnio.es/docus/papers/techreports.html#FatiGO-NNSP>).

Ashburner, M., Ball, C.A., Blake, et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29.

Bard JB, Rhee SY. (2004) Ontologies in biology: design, applications and future challenges. *Nat Rev Genet.* **5**,213-322

Bassett DE, Eisen MB, Boguski MS (1999) Gene expression informatics – It's all in your mine *Nat Genet*, **21**,51-55

Benjamini, Y, and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal Royal Statistical Society B* **57**,289-300.

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

Beissbarth T., Speed T. (2004) GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*. In press

Blaschke, C., Hirschman, L., Valencia, A. (2002) Information extraction in molecular biology. *Briefings in Bioinformatics*. **3**,154-165.

Castillo-Davis, C.I. and D.L. Hartl (2003). GeneMerge – post-genomic analysis, data mining and hypothesis testing. *Bioinformatics* **19**,891-892

Camon E, Magrane M, Barrell D, et al., (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **3**:662-672.

DeRisi, J.L., Iyer V.R., and Brown P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686

Díaz-Uriarte, R., Al-Shahrour, F. & Dopazo, J. (2003) Use of GO Terms to Understand the Biological Significance of Microarray Differential Gene Expression Data. *Microarray data analysis III. Kluwer Academic. Eds. K. F. Johnson and S. M. Lin. Pp. 233-247*

Doniger S.W., Salomonis, N., Dahlquist K.D., Vranizan K., Lawlor S.C. and Conklin B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology* **4**,R7.

Dopazo, J. and Carazo J.M. (1997). Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.*, **44**, 226-233

Eisen, M., Spellman, P.L., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**,14863-14868.

Ge H, Walhout AJ, Vidal M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**,551-560.

Herrero, J., Al-Shahrour, F., Díaz-Uriarte, R., Mateos, A., Vaquerizas, J.M., Santoyo, J. and Dopazo, J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research* **31**:3461-3467.

Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**,126-136.

Holloway AJ, van Laar RK, Tothill RW, Bowtell DD. (2002) Options available-from start to finish--for obtaining data from DNA microarrays II. *Nat Genet.* **32 Suppl**,481-489.

Jenssen, T.-K.; Laegreid, A.; Komorowski, J.; Hovig, E. (2000) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21-28.

Jensen LJ, Gupta R, Staerfeldt HH, Brunak S. (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics.* **19**,635-42.

Khatri, P., Draghici, S., Ostermeier, G.C. and Krawetz, S.A. (2002). Profiling gene expression using onto-express. *Genomics* **79**,1-5.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32 Database issue**:D277-280

MacBeath, G. (2002) Protein microarrays and proteomics. *Nat Genet.* 32 Suppl:526-532.

Mateos, Á., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M. and Stolovitzky, G. (2002) Systematic Learning of Gene Functional Classes From DNA Array Expression Data by Using Multilayer Perceptrons. *Genome Research* **12**: 1703-1715

Mootha VK, Lindgren CM, Eriksson KF, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* **34**:267-273

Mulder NJ, Apweiler R, Attwood TK, et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**:315-318.

van Noort V., Snel B. and Huynen M. A. (2003) Predicting gene function by conserved co-expression *Trends in Genetics* **19**:238-242

Oliveros, J.C.; Blaschke, C.; Herrero, J.; Dopazo, J.; Valencia, A. (2000) Expression profiles and biological function. *Genome Informatics* **10**, 106-117.

Pavlidis P., Lewis D.P., and Noble, W.S. (2002) Exploring Gene Expression Data with Class Scores. *Pacific Symposium on Biocomputing* **7**, 474-485.

Raychaudhuri S, Chang JT, Imam F, Altman RB. (2003) The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.* **31**,4553-4560

Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. (2002a) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* **12**,203-214

Raychaudhuri S, Schutze H, Altman RB. (2002b) Using text analysis to identify functionally coherent gene groups. *Genome Res.* **12**, 1582-1590.

Robinson, M.D., Grigull, J., Mohammad, N. and Hughes, T.R. (2002). FunSpect: a web-based cluster interpreter for yeast. *BMC bioinformatics.* **3**,1-5.

Shah NH, Fedoroff NV. (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics.* **20**,1196-1197

Schug J, Diskin S, Mazzarelli J, Brunk BP, Stoeckert CJ Jr. (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.* **12**,648-655.

Tanabe, L.; Smith, L.H.; Lee, J.K.; Scherf, U.; Hunter, L.; Weinstein, J.N. (1999) MedMiner: An internet tool for filtering and organizing bio-medical information, with application to gene expression profiling. *BioTechniques* **27**,1210-1217.

Westfall, P. H. and Young, S. S. (1993) Resampling-based multiple testing. John Wiley & Sons. New York.

Xenarios I, Salwinski L., Duan X.J., Higney .P, Kim S., Eisenberg D. 2002 DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303-305.

Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L. (2002) Large-scale protein annotation through gene ontology. *Genome Res.* **2**,785-794

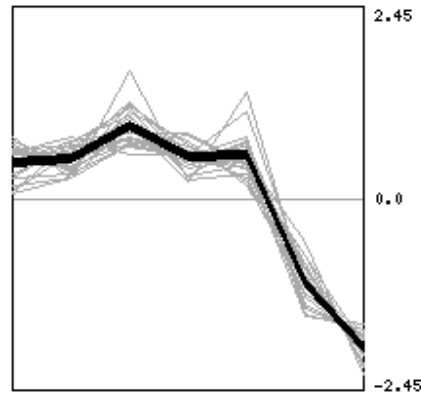
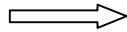
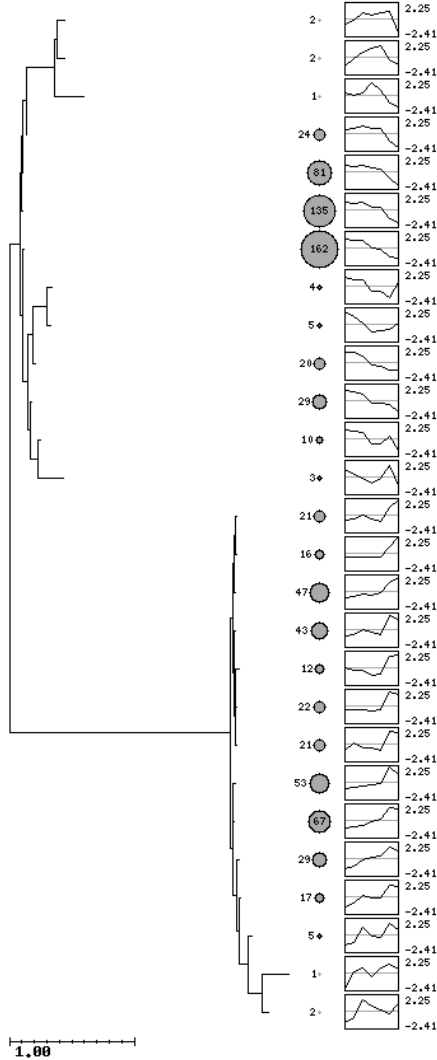
Zeeberg B. R., Feng W., Wang G., et al.. (2003) GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data. *Genome Biology*, 4(4):R28

**Figure 1.** Representation of the inclusive analysis concept. If apoptosis node level is chosen for the analysis, 8 genes, annotated in descendant nodes, will be assigned to the term apoptosis. If inclusive analysis is not used, then four terms: apoptosis (with 2 genes), regulation of apoptosis (3), negative regulation of apoptosis (1) and induction of apoptosis (2) are taken into account with the obvious decrease in the power of the test.

**Figure 2.** Clustering of gene expression patterns from the experiment of diauxic shift of yeast cells (DeRisi et al., 1997) obtained using SOTA algorithm (Herrero et al., 2001). The parameters used were: coefficient of correlation as distance measure and the growth was stopped at 95% of variability (see Herrero et al., 2001 for details on the procedure). The cluster with 21 genes that were initially active and they suffer a late repression was analysed with FatiGO (Al-Shahrour et al., 2004). A 75% of these genes were annotated as biosynthesis, and the differences in proportion with respect to the background (30%) were clearly significant.



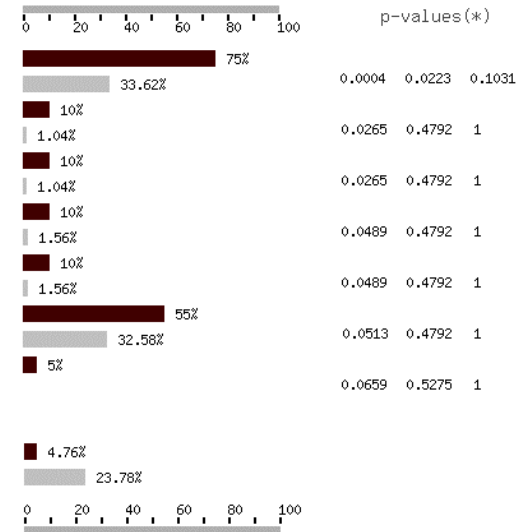
data\_norm - 06/09/2004 11:55:13 GMT



Gene Ontology Term

- biosynthesis**
- sexual reproduction**
- conjugation**
- aromatic compound metabolism**
- heterocycle metabolism**
- protein metabolism**
- one-carbon compound metabolism**

.....



(\*) Unadjusted p-value; FDR(indep.)adjusted p-value; FDR(arbitrary dependence) adjusted p-value