
New Trends in the Analysis of Functional Genomic Data

David Montaner^{1,2}, Fatima Al-Shahrour¹, and Joaquin Dopazo^{1,2}

¹ Bioinformatics Department, Centro de Investigacin Prncipe Felipe (CIPF)
Autopista del Saler 16, E-46013, Valencia, Spain. dmontaner@cipf.es

² Functional Genomics Node, INB, CIPF. Autopista del Saler 16, E-46013,
Valencia, Spain

1 Replications of the same statistical test

Most analyses carried out using high throughput data consist of the repetition of the same statistical test for all genes in the dataset. As a result of such replicated analysis we get, for each gene, several estimates of statistical parameters: statistics, p-values or confidence intervals. Being aware that most statistical methods were developed to test for a single hypothesis, researchers will usually correct p-values for multiple testing before choosing a cut-off that will indicate the rejection of the null hypotheses, whichever it is. Once chosen the genes with alternative pattern (meaning different from the one stated in the null hypothesis) the next step is to biologically interpret such departure from hypothesis. Different repositories of functionally relevant biological information such as Gene Ontology [1], KEGG [2] or Interpro [3] are available and can be used for the functional annotation of genome-scale experiments. Thus the functional properties of the selected genes can be analysed.

The trouble of this approach is that, by discarding genes with p-values above the cut-off, we loose most of our information. Not only we loose the measurements taken over the genes but also the functional annotation that could be linked to them from repositories, making it difficult the biological interpretation of results.

2 Blocks of functional genes

Aiming to prevent such waste of information, some authors have recently proposed to directly analyse the behaviour of blocks of functionally related genes in a whole-genome context. The Gene Set Enrichment Analysis (GSEA) [4,5], the FatiScan [6,7] or the Global Test [8,9] constitute examples of this type of approach inspired from systems biology. This three methodologies address the issue of whether the general expression pattern of a group of genes, for

example a GO term or a KEGG pathway, changes across biological conditions. Here we will discuss just some particular aspects of these methods but a more general view of this and similar methods can be found in Dopazo's revision of 2006 [10].

The Global Test uses generalised linear models to study the relationship between the expression of the genes of the block of interest and a characteristic associated to each biological sample. Such characteristic may be a categorical condition, like the class of the microarray in the context of differential gene expression, or a continuous variable such as a level of a metabolite. In this approach we can see a change in the philosophy of the analysis. The unit of interest is not any more a single gene but a block of genes with a common biological meaning. This new way of looking at the data provides, among others, obvious advantages for the biological interpretation of results and for the p-value adjustment. We just need to correct by the number of blocks, usually smaller than the number of genes.

3 The overall approach

The block of genes is also the unit of interest of the GSEA and the FatiScan. These two methods are similar to the Global Test in that they are also used to discover groups of genes which overall expression pattern changes across biological conditions. Nevertheless, GSEA and FatiScan consider all genes in the data when analysing each of the blocks. They compare the pattern of the genes of one block with the general pattern of the genes in the whole dataset. GSEA is particularly designed for the two class comparison context while FatiScan may be applied in a wider range of studies.

The rationale underlying both methodologies is that, if a property of genes can be described using a continuous index, then the statistical distribution of such index within a functional block of genes can be compared to the general distribution of the index across all genes in the data. We can therefore assess whether the property described by the index depends on the characteristic that defines the block of genes.

As said before GSEA is developed for the two class comparison. In this methodology, a signal-to-noise ratio comparing mean expression across classes is computed for each gene in the dataset. This statistic can be seen as a continuous index that ranks the genes according to their differential expression, from those more expressed in one of the biological conditions to those more expressed the second condition, passing through those genes non differentially expressed. Then, given a block of genes, for instance a functional class that we may be interested in, we can compare the distribution of the signal-to-noise ratio of the genes in the block to the distribution of the same statistic in the remaining genes. If the values of the signal-to-noise ratio are, for instance, systematically higher in the genes of the block compared to the genes in the whole dataset, we will conclude that, as a block, the genes of the functional

class of interest are overexpressed in one of the biological conditions. GSEA uses a modification of the Kolmogorov-Smirnov test to assess differences between the signal-to-noise ratio in the class of interest and in the rest of the genes. Significance of the modified Kolmogorov-Smirnov statistic is computed in GSEA using permutations of the expression data. The original expression data is permuted several times, the signal-to-noise ratios are calculated over each permuted expression dataset and the modified Kolmogorov-Smirnov statistic is computed over each new distribution of the signal-to-noise ratio. Thus GSEA can estimate the random variability of the Kolmogorov-Smirnov statistic and test its significance in the original data.

4 Detaching concepts and algorithms

FatiScan follows the same analytical philosophy than GSEA but with a more general and flexible approach. FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked by any index. The main difference is that FatiScan does not implement a permutation test to assess such asymmetry. Therefore, the algorithm that computes the index and the algorithm that analyses the distribution of the index are completely separated so the calculations can be done in two different steps. This means that FatiScan can be used to study the relationship between biological labels associated to genes and any type of experiment whose outcome is a sorted list of genes or a variable that can be used to rank genes according to some characteristic of interest. Block of genes sorted by differential expression between two experimental conditions can be studied as it would be done using GSEA. But with FatiScan we can also consider many other gene properties or characteristics.

We can easily explore the correlation between gene expression and a clinical continuous variable such as the level of a metabolite. First, for each gene we will compute the correlation between its expression measurements and the levels of the metabolite. Thus we can range the genes from those which expression is more positively correlated to the levels of the metabolite to those inversely correlated, passing by genes which expression does not correlate with the clinical variable. In a second step, FatiScan explores the distribution of such correlation measurements, testing whether the distribution of correlations within a block of genes is different from the overall distribution of correlation in the dataset.

We can fit a Cox proportional hazard model to each gene in our data in order to study the relationship between gene expression and survival times. The estimates of the slope coefficients may be used as an index that ranks genes from those which increased expression is associated with long time survival to those which increased expression is associated to an early death. After computing this rank-index, FatiScan will find those blocks of genes for which the distribution of the slopes differs from the global distribution of the slopes.

The complete separation of the two steps in FatiScan analysis is the key point which provides its flexibility to the method. Such flexibility makes possible to handle many different sources of information, not only microarray gene expression data. Any lists of genes ranked by any other experimental or theoretical criteria can be studied. Genes can be for example arranged by physico-chemical properties, mutability, structural parameters and so on. In order to understand whether there is some biological feature, characterised by the blocks of genes, which is related to the experimental parameter studied.

5 Coda

The three methodologies here mentioned illustrate two of the main new conceptual trends in the analysis of functional genomic data.

The first one is the change of the descriptive unit used to address biological studies, shifting from gene to functional class. Gene still remains the unit of measured information, as what we record at the end is gene expression. But the conceptual entity over which biological interpretation is done, is the functional class of genes. New analytical strategies, like those above mentioned, should consider this fact in order to use the available information in the most efficient an meaningful way

The second one is probably more subtle but not less important. Usual genomic studies follow the classical statistical approach in which one or several hypotheses are stated, estimate statistics and p-values are computed from data and finally, hypotheses are accepted or rejected depending on such estimated values. The analytical approach explicit in FatiScan an implicit in GSEA shows how estimated values provided by one first statistical analysis are not directly interpreted in terms of acceptance or rejection of hypotheses. Instead they are treated as variables quantifying some characteristic of the genes under study. This new variables may then be analysed using statistical methodologies. Thus, statistical results of one step of the analysis become themselves a new dataset which needs to be explored in a second analytical step. As we see, modular implementations of complex data analysis strategies like FatiScan, seem to be both, conceptually useful for the analysis of biological data and computationally advantageous, calling for the development of the theoretical framework within which combinations of statistical methods can be properly done.

References

- [1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, 25, 25-29 (2000)

- [2] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.; The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32, D277-D280 (2004)
- [3] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al.; InterPro, progress and status in 2005. *Nucleic Acids Res.*, 33, D201-D205 (2005)
- [4] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al; PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, 34, 267-273 (2003)
- [5] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al; Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102, 15545-15550 (2005)
- [6] Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J.; Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information *Bioinformatics*, 21, 2988-2993 (2005)
- [7] Al-Shahrour F., Minguez P., Trraga J., Montaner D., Alloza E., Vaquerizas J.M., Conde L., Blaschke C., Vera J. and Dopazo J.; BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucl Acids Res.*, 34, W472-W476 (2006)
- [8] Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C.; A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20, 93-99 (2004)
- [9] Goeman J.J., Oosting J., Cleton-Jansen A.M., Anninga J.K., van Houwelingen H.C.; Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21, 1950-1957 (2005)
- [10] Dopazo, J.; Functional Interpretation of Microarray Experiments. *OMICS: A Journal of Integrative Biology*, 10, 398-410 (2006)