Gene Ontology based classification improves prediction and gene signature interpretability

Tiziana Sanavia[1], Aler Crepaldi[1], Annalisa Barla[2], Barbara Di Camillo[1]

[1] Department of Information Engineering, University of Padova, Italy
[2] Department of Computer and Information Science, University of Genova, Italy

Introduction

In the last decade, large-scale genomic studies have provided a great support to medical research and a high-resolution view of the molecular mechanisms and their possible alterations characterizing different pathologies. Classification of patients based on gene expression measurements of molecular markers is useful for answering to several diagnostic/prognostic questions. A variety of predictive models dealing with high-throughput data have been suggested. However, since the number of available patients is limited and the biology of the samples is extremely complex, this problem cannot be easily solved. In fact, data are characterized by a small number of subjects with respect to the number of variables, leading to more than one possible solution to the classification problem. On the other hand, studies on complex diseases such as cancer have revealed that subjects of the same clinical type are often characterized by heterogeneous genomic alterations, which possibly affect a specific set of biological process, but not the same genes in different patients [1]. These two aspects make the classification a challenging task: results obtained so far from different studies are poorly reproducible and often provide lists of features characterized by a large number of candidate molecular markers which are not easily interpretable. Most of the methods proposed in the literature to improve interpretability of the results are based on *a posteriori* annotation of the selected features in order to describe the main biological processes characterizing the results. Recently, several knowledge-driven methods have been proposed to integrate biological knowledge into the learning process, in order to obtain more easy-to-read lists of candidate genes characterizing the disease and more reliable predictive models. Among others, Tai and Pan [2] proposed a group penalization method that handles the genes within different functional groups with different penalty terms. Lottaz and Spang [3] proposed a structured analysis of microarray data (StAM), which generates a classifier graph according to the Gene Ontology (GO), constructs leaf node classifiers based on selected expression values from shrunken centroid classification, propagates the results through the inner nodes to the root and shrink the classifier graph to obtain a set of molecular symptoms.

Among the functional annotation databases, GO is the most widely used. This controlled vocabulary consists of three independent categories: molecular function, biological process and cellular component [4]. In the GO, the information is structured according to a directed acyclic graph (DAG) in which each node corresponds to a GO term. Each node may have multiple parents: nodes farther from the root (high level nodes) correspond to more specialized terms, nodes closer to the root (low level nodes) to less specialized terms, thus implying that genes annotated with a specific node are also annotated with every ancestor of that node (true path rule). This introduces strong dependencies among the GO terms and redundancy of the information.

Here, we present a method able to integrate classification/feature selection with functional annotations retrieved from the GO in order to improve class prediction and to increase the biological interpretability of the results. The output of the method is organized into subsets of genes both 1) highly correlated and 2) annotated to groups of GO terms with similar meaning.

Methods

The method exploits the direct acyclic graph of the GO to define different sets of genes sharing the same annotation. For each gene set, classification analysis and feature selection are based on l1-l2 regularization with double optimization, as described in [5]. The method is based on an estimator that solves the following optimization problem:

$$\beta = \arg\ \min_{\beta}[\|Y - X\beta\|_2^2 + \tau[\|\beta\|_1 + \varepsilon\|\beta\|_2^2]] \qquad (1)$$

where X is a matrix n x p (p >> n) with each row representing the expression values of a sample (patient, cell line, treatment), Y is a response vector of binary values characterizing each class (e.g. patient or control) and β is a vector of unknown weight coefficients assigned to each gene. The least square term ensures fitting of the data whereas the two penalty terms allow avoiding over-fit. In particular, the l1 term (sum of absolute weights) enforces the solution to be sparse while the l2 term (sum of the squares of the weights) preserves correlation among genes. The two penalty terms are regulated by the parameters τ and ε. The solution β is computed through an iterative soft-thresholding, followed by a second optimization, namely regularized least squares, to estimate the classifier on the selected features.

Starting from the highest level nodes, i.e. the farthest from the root, which are the most specific, the classifier is performed separately on the gene sets annotated to each GO term using a 5-fold cross validation approach. Classification performance is measured in terms of Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (2)$$

If the MCC exceeds a fixed threshold, then the genes selected by the l1-l2 approach are removed from the annotations of all the ancestors of that node. This approach, named “elim”, has been first proposed in [6] to account for the redundancy of GO annotations in the ontology and to preserve the specificity of the biological information associated to the selected genes. A weighted majority vote classifier based on the models built on every single GO term is used for final class prediction.

The output of the proposed method is organized in two ways:

- the ranked list of GO terms selected according to MCC values;

- the ranked list of genes selected for each GO term according to the weights estimated from the l1-l2 classifier.

Results

The method has been applied on three publicly available breast cancer datasets (extracted from Gene Expression Omnibus with identification codes GSE2990, GSE3494, GSE7390) with positive and negative estrogen receptor status, considering separately two category of the Gene Ontology: Biological Process (GOBP) and Molecular Function (GOMF). For each dataset, ten random splits of the data into training and test set have been considered.

In Table 1, the average MCCs and the standard deviations obtained from the test sets of the ten splits in the three datasets are displayed. The MCCs of the ten splits obtained using the new method using Biological Process and Molecular Function categories are significantly higher (p-value always lower than 0.021) with respect to the MCCs obtained with the standard approach. This improvement is probably due to the limited number of genes, restricted to those belonging to a single GO term, used to build each classifier: in this way, the *curse of dimensionality* effect is reduced and a more robust statistical analysis is promoted.

	GSE2990			**GSE3494**			**GSE7390**		
	Standard Method	*GOBP based*	*GOMF based*	*Standard Method*	*GOBP based*	*GOMF based*	*Standard Method*	*GOBP based*	*GOMF based*
Mean	0.445	0.685	0.679	0.396	0.523	0.525	0.661	0.798	0.796
SD	0.137	0.152	0.121	0.069	0.076	0.088	0.117	0.104	0.094

Table 1. Classification performance (MCC) over the ten random splits of the three breast cancer datasets.

The frequency of the GO terms selected across the ten random splits assesses the reproducibility of the results and allows ranking GO terms according to their association to the disease. Considering those GO terms with a frequency equal or higher than 0.8, many of these are common to all the three breast cancer datasets, in particular the GO terms related to “response to oxidative stress”, “developmental process” and “regulation of cell proliferation” in Biological Process and the GO terms related to “oxidoreductase activity” and “metal ion binding” in Molecular Function.

In conclusion, our approach is different with respect to the previous ones proposed in the literature, in the way the GO graph is managed and the final classifier is built. This gives a functional-based characterization of the disease in an easy-to-read way and with a statistically significant improvement of the classification performance with respect to a standard approach analyzing all the features without considering gene annotation.

References

[1] Jones,S. et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science, 321:1801-1806.

[2] Tai,F. and Pan,W. (2007) Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics, 23, 1775–1782.

[3] Lottaz,C. and Spang,R. (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics, 21, 1971–1978.

[4] Ashburner,M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25–29.

[5] De Mol,C. et al. (2009) A Regularized Method for Selecting Nested Group of Genes from Microarray Data. *Journal of Computational Biology*, **16**, 677-690.

[6] Alexa, A. et al. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22: 1600–1607.