Gene Ontology Driven Classification of Gene Expression Patterns

Claudio Lottaz, Stefan Bentink and Rainer Spang
Max-Planck-Institute for Molecular Genetics
Computational Diagnostics Group
Ihnestrasse 73, D-14195 Berlin (Germany)

1 Introduction

From a machine learning point of view, classification of gene expression patterns is a very particular task. Typically, training data consists of few samples (small number of experiments) but contains many variables (expression levels measured in each experiment). In this context classical machine learning methods may cause various difficulties [1]. For instance:

- Statistical models, particularly those with many parameters, may overfit the training data. Thereby, they rather adapt to noise in the data than learn the desired phenomenon.
- Moreover, common machine learning methods do not provide an intuitive and biologically meaningful explanation of their results. However, such explanations help users to trust a computational analysis.

In the research presented here, we try to cope with these two problems in the context of medical diagnosis.

2 Gene Ontology driven classification

We conjecture that the mentioned problems can be tackled by giving the classifier a biologically meaningful structure, i.e., by dividing the classification task into subtasks according to biological criteria. Structuring biological knowledge is one of the central goals of the Gene Ontology database [2]. Biological terms related to molecular functions, biological processes and cellular components are collected into a directed acyclic graph where each node represents a term and child-terms are either members or representatives of their parent-terms. Moreover, genes are attributed to GO-nodes according to their functions, involvement into biological processes and location within the cell. We suggest to use this structure in a classifier as follows.

For each GO-node, one classifier is implemented using classical logistic regression, determining a linear weight for each input variable and providing a probability for each class as classification result. Each of these classifiers is given the same classification task, while input variables are the expression values corresponding to the genes annotated to the classifier's GO-node as well as the classification results of its children. Thus, the classifiers corresponding to the leaf-nodes of the Gene Ontology must be trained first. The overall classification result is provided by the root node's classifier.

In this procedure each classifier bases its decision only on information about the biological function it is attributed to. Therefore, when considering an overall classification result, its rationale can be deduced from the various classifier results. Moreover, the weights determined after training provide information about which biological aspects are deemed important in the classification task. Finally, the partitioning of the input variables among many classifiers, weakens the mentioned overfitting problem.

3 First results and experiences

A first implementation as an R program has been used to evaluate the method on a large dataset from leukemia patients (acute lymphoblastic leukemia) [3]. The recognition of leukemia subtypes has been used as the classification task. This task has been shown to be a rather simple one yielding recognition rates of 96% to 100% using sophisticated feature selection and support vector machines. First tests with our classifier have shown comparable recognition accuracies for five of the six subtypes while the last subtype is more difficult to be recognized. Thereby, many classifiers yield average or weak results and only a few pin-point the important biological aspects for classification.

References

[1] T. Hastie, R. Tibshirani & J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, 2001.

[2] The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, May 2000.

[3] E.-J. Yeoh, M. E. Ross et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–145, March 2002.