# 40. GALGO: Software for building statistical models from microarray data

Victor Trevino and Francesco Falciani

*Biosciences, University of Birmingham, UK*

**Variable selection is an important concept for extracting knowledge from microarray data. Multivariate methods coupled with genetic algorithms (GA) have been used to select significant gene signatures related to cell physiology. We report GALGO, an R package for building multivariate statistical models using GA as variable selection search strategy.**

INTRODUCTION. Microarrays experiments have revolutionized biology and medicine because they enable the inspection of thousands of transcripts in a single experiment. This vast and valuable information needs, however, to be carefully examined to extract reliable biological knowledge.

In the analysis of Functional Genomics data, several statistical methods, including classification and regression modelling techniques, have been used to select genes signatures associated to cell phisiology. These techniques can be applied in a univariate and multivariate fashion, depending on the number of genes considered in the statistical model. Univariate techniques test every gene, separately from the rest, for its capacity to distinguish phenotypic outcomes. Popular univariate approaches includes PAMR (1). Multivariate techniques, on the other hand, consider the effects of several genes combined in the same model. Indeed, multivariate models allow the inclusion of genes interactions that may reflect biological synergy. Nevertheless, the number of possible combinations of models grows exponentially with the number of genes making impossible to explore every single model combination with current computational resources. Therefore, to find reasonable good combinations of genes that together predict the desirable outcome, stochastic search strategies as Markov Chain Monte Carlo and Genetic Algorithms (GA) have been used in the analysis of microarray data (2,3,4). Because of the stochastic behaviour of these methods, and the possibility that different models may be related to the outcome, a large number of good models are analyzed and combined to generate a better model that represents all models found.

GA coupled with classification methods have been used in the analysis of microarray data with promising results (2,3). Current implementations however use a very specific classification method, generate specific and limited output, make different uses of solutions, estimate the error in different ways, and there is no framework to extend these implementations.

To date, no comprehensive package for building multivariate statistical models using efficient variable selection strategies is available. Therefore, to address this issue, we have developed GALGO, an object-oriented R package to perform GA searches coupled to a generic fitness function. The package includes several classification methods that can be used as fitness function. As an example of GALGO benefit, we compare the classification accuracy of different classification methods versus univariate methods.

SYSTEMS AND METHODS. We developed GALGO as an object-oriented implementation of GA under the R language. R is a statistical programming environment that is platform-independent, robust,

free, and is widely used for the analysis of microarray data. The default GA implementation includes the core GA concepts as gene, chromosomes, niches and genetic algorithm as objects, which implement methods for mutation, crossover, migration, elitism, and progeny according to GA principles (5). A more general object encloses and organizes the collected models from thousands of GA runs. All objects within the package enclose user-properties that control the behaviour and organize the whole process. Moreover, users can add their own properties and methods to customize the functionality according to their statistical models or hypothesis.

Current implementation includes several classification methods as nearest centroid, k-nearest-neighbours, maximum likelihood discriminant functions, support vector machines, classification trees, and neural networks. To estimate error in resulted models, the data is split in train and test sets many times and the error distribution is analyzed. Train set is further split to estimate the error inside the GA. Several error estimation methods as random splits, k-fold cross-validation, and resubstitution were implemented and can be used in combination with any fitness function.

The resulted models can be processed in different ways. They can be filtered to study specific models and their pooled effects; or, can be enhanced using any other criteria, as removing unnecessary genes using the backward elimination function given; or can be combined using the most frequent genes to build a more general model using the forward selection function supplied.

Information related to any component of the process can be retrieved using plots and functions provided. For example, the evolution of fitness inside the GA can be plotted in real-time, the gene frequency of the collected models, the classification accuracy, the confusion matrix, and heatmaps and analysis of models can be shown and computed directly.

To confirm the benefits of using GALGO in a well know microarray classification problem, we have compared models developed by GALGO using different classification methods against those resulted with PAMR. The results are that the models found using GALGO outperformed those developed by PAMR.

CONCLUSION. GALGO is a valuable, user-friendly R package designed to develop statistical models using large-scale data. The current version includes methods for supervised classification problems and no further coding is need for their usage. Therefore, the package is also suitable for biologists. The R environment and object-oriented implementation makes GALGO an ideal framework for any method that uses genetic algorithms as search strategy coupled to statistical analysis.

REFERENCES

[1] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A. 2002 May 14;99(10):6567-72.
[2] Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays:Expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001 Jan 2;98(1):31-36.
[3] Ooi CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics. 2003 Jan;19(1):37-44.
[4] Sha N, Vannuci M, Tadesse M, Brown P, Dragoni I, Davies N, Roberts T, Contestabile A, Salmon M, Buckley C, Falciani F. Bayesian Variable Selection in Multinomial Probit Models to Identify

Molecular Signatures of Disease Stage. Biometrics 60, 812-819. September, 2004.
[5] Goldberg, D. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Pub Co, Jan 1989.