A copula-based method for synthetic microarray data generation

Sergio Lew[1], Jordi Solé-Casals[2], Cesar F. Caiafa[3,4] and Josep Bau-Macià[2]

[1] IIBM-FIUBA, Av. Paseo Colón 850 (1063), Buenos Aires, ARGENTINA
[2] University of Vic, Sagrada Família 7, 08500, Vic, SPAIN
[3] IAR-CONICET, C.C.5, (1894) Villa Elisa, Buenos Aires, ARGENTINA
[4] FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA

Abstract: In this work, we propose a copula-based method to generate synthetic gene expression data that account for marginal and joint probability distributions features captured from real data. Our method allows us to implant significant genes in the synthetic dataset in a controlled manner, giving the possibility of testing new detection algorithms under more realistic environments.

Keywords: Copulas; gene expression; microarray data.

1 Introduction

Detection of differentially expressed genes in microarray experiments has been subject of great effort in the bioinformatics community. Optimal detection methods allow to reduce the amount of both, pathological and control experiments and, in consequence, time and costs [Dupuy A. and Simon R. M., 2007]. However, most of the developed algorithms have been tested with synthetic data using simple generative models and assuming incorrect hypothesis about variable statistics and their dependence. The proposed method captures the statistical structure of real datasets allowing us to generate new random samples drawn from a copula-based random generator.

2 Materials and Methods

The proposed method is shown in figure 1. Briefly, we fit real microarray data to a t-copula [Nelsen R. B., 1999] and then we generate random gene expression data sharing marginal and high-order dependence with the original data.
Firstly, original gene expressions dataset (Fig. 1.a) are mapped into a unitary hypercube by means of a monotonically increasing function, i.e. the inverse cumulative distribution function of the marginal distributions.

An approximated Maximum Likelihood (ML) method is used for fitting this transformed data to a t-copula. Once the copula parameters are obtained, random samples are generated according to this copula structure (Fig. 1.b). This new dataset, which have uniform marginal distributions, is then mapped to a $\mathcal{N}(0,1)$ marginal Gaussian distributions (Fig. 1.c). It is important to remark that monotonically increasing transformations do not alter high-order dependence measures like Kendall-τ or Spearman-ρ. At this point, significantly expressed genes are introduced in a controlled manner into the data (Fig. 1.d). By means of the inverse transformations used before, we then back-transform the data to the original space, obtaining a synthetic dataset which preserve the same marginal distributions and high-order variable dependence as the real one (Fig. 1.e-f).

3 Results

When added to the synthetic data, significant genes were recovered by the step-down minP adjusted p-values method [Westfall P. H. and Young S. S., 1993] in all the cases. However, its important to ensure that the algorithm is able to minimize the number of false positive (FP) cases. To prove the robustness of the method against FP generation, we compare the results of our method versus synthetic data generated by multivariable gaussian random process with the same covariance matrices of the original data [Carmona-Saez P. et al, 2006]. We ran 30 experiments for both types of synthetic data with no significant genes added, meaning that the significance test should recover (almost) zero genes differentially expressed between pathological and control groups. Due to the sparseness of significant genes in microarray experiments (less than 1%, under and over-expressed genes) the copula captures the distribution of normally expressed genes. In that sense, our synthetic microarray data produces much less FP genes that the ones generated with a multivariable gaussian process having the same covariance matrix (1.5 ± 0.23 vs 24.76 ± 1.04, $p < 0.0001$, $mean \pm s.e.m.$)

4 Conclusions

In this paper, we propose a new copula-based method for synthetic microarray data generation that allows us to control the number of under and over-expressed genes, preserving the original statistical structure of real data. To our knowledge, this is the first work that overcomes the problem of building synthetic data using simple generative models. Experimental results show the robustness of the method and its usefulness helping researchers to develop new and more powerful algorithms for gene filtering and clustering.

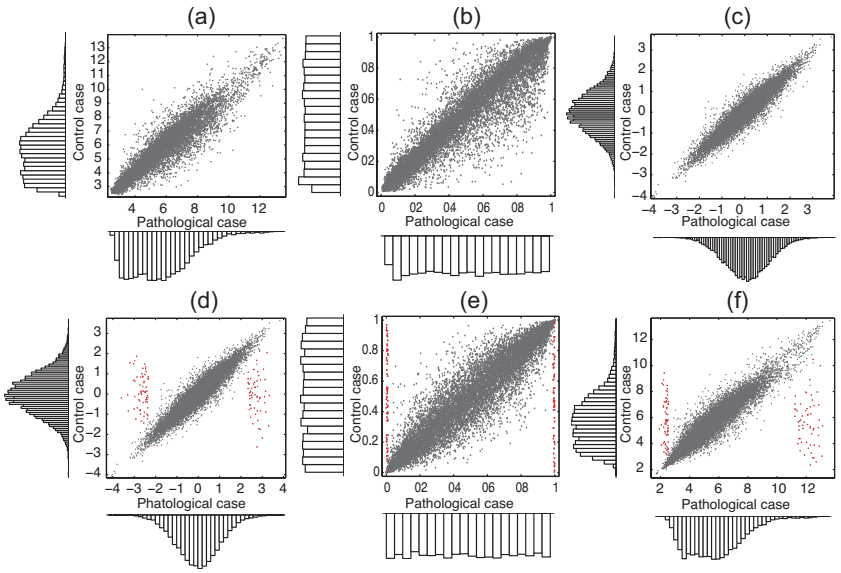

FIGURE 1. Proposed method to generate synthetic data preserving the same marginal distributions and high-order variable dependences of the real data.

Acknowledgments: This work has been in part supported by the MINCYT-MICINN Research Program 2010-2011 (Ref. AR2009-0010) and by the University of Vic under the grants R0904 and R0901.

References

Carmona-Saez P., Pascual-Marqui R. D., Tirado F., Carazo J. M., Pascual-Montano A. (2006). Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. *BMC Bioinformatics*, **7(78)**, 1-18.

Dupuy A., Simon R.M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.*, **99(2)**, 147-157.

Nelsen R. B. (1999). *An Introduction to Copulas*. New York: Springer-Verlag.

Westfall P. H. and Young S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons.