Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

The full Bayesian significance test for mixture models: results in gene expression clustering

M.S. Lauretto, C.A.B. Pereira and J.M. Stern

Instituto de Matemática e Estatística, Universidade de São Paulo,
São Paulo, SP, Brasil

Corresponding author: M.S. Lauretto
E-mail: lauretto@ime.usp.br

Genet. Mol. Res. 7 (3): 883-897 (2008)
Received june 2, 2008
Accepted August 11, 2008
Published September 23, 2008

ABSTRACT. Gene clustering is a useful exploratory technique to group together genes with similar expression levels under distinct cell cycle phases or distinct conditions. It helps the biologist to identify potentially meaningful relationships between genes. In this study, we propose a clustering method based on multivariate normal mixture models, where the number of clusters is predicted via sequential hypothesis tests: at each step, the method considers a mixture model of m components ($m = 2$ in the first step) and tests if in fact it should be $m - 1$. If the hypothesis is rejected, m is increased and a new test is carried out. The method continues (increasing m) until the hypothesis is accepted. The theoretical core of the method is the full Bayesian significance test, an intuitive Bayesian approach, which needs no model complexity penalization nor positive probabilities for sharp hypotheses. Numerical experiments were based on a cDNA microarray dataset consisting of expression levels of 205 genes belonging to four functional categories, for 10 distinct strains of *Saccharomyces cerevisiae*. To analyze the method's sensitivity to data dimension, we performed principal components analysis on the original dataset and predicted the number of classes using 2 to 10 principal components. Compared to Mclust (model-based clustering), our method shows more consistent results.

Key words: Gene clustering; Mixture models; Significance test; Expression data analysis