Mining microarray data for inflammatory stimuli probes selection and classification

Tuana G[1], Volpato V[2], Ricciardi-Castagnoli P[3], Stella F[2], Foti M[4], Zolezzi F[1]

Motivation

Inflammation is part of the pathogenesis of different human diseases. The activation of components of the immune system is recognized as a component of some neurodegenerative diseases like Alzheimer's disease, Parkinson's disease, Huntington's disease and Amyotrophic Lateral Sclerosis. Also in Cystic Fibrosis inflammation is a very early event and can contribute to worsen the patient clinical situation. Identification of marker genes that play an active role in the immune response could be useful to understand disease progression. Microarray experiments can be used to learn a classifier capable of discriminating samples in two or more classes. This approach is challenging since the number of available samples is small with respect to the number of explanatory variables and the data are often noisy.

Methods

In this work, a Microarray dataset was analyzed to solve a binary classification problem in order to discriminate between components that induce or not an inflammatory profile in dendritic cells. All experiments used the D1 murine cell line treated for 2, 4, 8, 12 and 24 hours with inflammatory and not inflammatory stimuli. Most experiments have been done in biological duplicates. Total RNA has been extracted, labeled and hybridized to Affymetrix® GeneChip®. Three different kinds of arrays (Affymetrix® MOE430 2.0, MOE430A 2.0 and MGU74Av2) have been used. Signal summarization was performed for each array using the Affymetrix GeneChip operating Software® (GCOS) and a scaling target intensity of 100 for all probe sets. Thus the learning dataset consists of 155 arrays (15 stimuli, 30 time series) whereas the validation set has 49 (7 stimuli, 10 time series); and 7,829 probe sets (features) representing the intersection of the probe sets belonging to three different microarray platforms. A filtering procedure to remove signal values below the background level has been applied. At the end, signal intensity data has been used to compute Z-scores. The two final dataset consists of 5,802 features

[1] Genopolis Consortium for functional genomics, Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy. [2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy. [3] Singapore Immunology Network (SIgN), Biomedical Sciences Institutes, Agency for Science, Technology and Research (A*STAR), 8A Biomedical Grove, IMMUNOS, 138648, Singapore. [4] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy

and has been used for different phase of analysis. The analysis protocol can be summarized into two steps: 1. Features Selection: features selection has been used to discover which probes are relevant to the classification task and has been implemented by the ADTree algorithm 10-folds cross validation on the training dataset. 2. Supervised Classification: the following models have been compared by using the features selected through the previous step: ZeroR, IB-3, C4.5, Logistic, Multi Layer Perceptron (MLP), Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SMO-puk) and Tree Augmented Naïve bayes (TAN). The average performance measures have been estimated through 10 repeated 10-folds cross validation. 3. Numerical Validation: performance of the classification algorithms were compared on the validation data set.

Results

The estimated performance is satisfactory (>95%) for a subset of supervised classification models in the training phase and in some cases it has been confirmed in the numerical validation phase. The model that achieves the highest accuracy value in the validation phase, 95.9%, is the C4.5 that confirms the performance obtained by cross validation. Both RF and SMO-puk achieve validation accuracy equal to 91.8%, which significantly differs from the expected 98.6% for SMO-puk and 99.1% for RF (the best accuracy computed via cross validation). The learning accuracy of IB-3 and NB drops from 98.1% to 83.7% and from 94.2% to 89.8%, respectively, that differs from what recorded in the validation phase. The same significant drops occur between Logistic and MLP. The Zero-R shows the worst performance in the learning phase (68.4%) and it keeps its behavior in validation phase (75.5 %). The protocol allowed to reduce the number of initial probes while some classification models still achieve good performances compared with training phase performances. The 10 selected probes have been inputted to the Ingenuity Pathway Analysis® software to search biological and functional relationships among genes. It has been found that 6 of them belongs to a network mapping into cellular growth and proliferation and humoral immune response pathway; moreover, 3 genes (Il12b, Cd40 e Socs3) out of 6 are well-known genes related to immune system. In conclusion, the best accuracy performance on the validation set using just the 6 probes is achieved by Tree Augmented Naïve Bayes (93.9%) and Support Vector Machines (91.8%), but Naïve Bayes (89.8%) and Nearest Neighbor (89.8%) achieve good performances also.

Contact e-mail

giacomo.tuana@unimib.it