Identifying the Hypoxia Signature of Neuroblastoma via Regularization

Barla A[(1)] , Fardin P[(2)], Rosasco L[(1,3)] , Mosci S[(1,4)], Verri A[(1)] , Varesio L[(2)]

[(1)]Dipartimento di Informatica e Scienze dell'Informazione, Universita di Genova
[(2)] Laboratorio di Biologia Molecolare, Giannina Gaslini Institute, Genova, Italy
[(3)] Brain & Cognitive Sciences Department McGovern Institute for Brain Research, Massachusetts Institute of Technology U.S.A.
[(4)] Dipartimento di Fisica, Universita di Genova

Motivation

The understanding of tumors nature at a molecular level represents one of the main goals in biology and medicine and opens up possibilities on improvements of clinical protocols. Technologies such as microarrays are able to measure the expression of thousands of genes at once, generating a global picture of the cellular function, hence they are one of the most used technologies to deeply investigate the characteristics of diseases and to interpret their molecular mechanisms. The goal of gene expression analysis is finding a predictive gene signature, that is a panel of genes which are able to discriminate between two given classes. The typical task, from the statistical analysis viewpoint, is to extract information from a small number of samples represented in a very high-dimensional space. The ability of a system to detect genes which are discriminative across different classes is usually constrained by the limited amount of available data, therefore great care is needed to properly design the statistical protocol in order to avoid selection bias. Neuroblastoma is a common pediatric solid tumor and is characterized by heterogeneity with respect to histology and clinical outcome. The hypoxic status of neuroblastoma is a critical factor of its progression. We aim at defining a signature of hypoxia in neuroblastoma by applying our feature selection framework on a dataset of neuroblastoma cell lines in normoxic and hypoxic status.

Methods

The idea is that we want to build an effective discriminative rule for the hypoxic status while selecting the genes that are relevant for such a task. The idea is to design a procedure allowing to control the number of genes in the model while building a powerful discriminative rule. The large number of genes compared to the small number of training data suggests multivariate techniques can incur into overtraining so that statistical tool designed to deal with high dimensional data are needed. The method we considered allows to have ranking and selection within a unique step alleviating the need for data. In many biological studies some of the input variables may be highly correlated with each other. As a consequence, when one variable is considered relevant to the problem, its correlated variables should be considered relevant as well. We use robust multivariate machine learning tools to define a reliable and refined hypoxia signature. The core of our unbiased approach is the elastic net selection criterion, based on Empirical Risk minimization combined with a mixed penalty, that simultaneously enforces the sparsity of the solution by the l1 term, while preserving correlation among input variables with the l2 term. Once the relevant features are selected we use regularized least squares (RLS) to estimate the classifier. This multivariate approach allows us to exploit the linear interaction of many genes at once and the mechanism to avoid overfitting is based on discarding the genes that do not improve classification performance. The regularization parameters control the tradeoff between classification on the validation samples and number of selected genes and are chosen in a double nested loop of cross validation. The free parameter epsilon in the elastic net is fixed a priori and governs the amount of intra-gene correlation we wish to take into account. The dataset we used consists of 9 neuroblatoma (NB) cell lines cultured under normoxic and hypoxic conditions. Their gene expression profile was assessed by Affymetrix GeneChip U133 plus 2.0.

Results

We performed a supervised analysis aimed at detecting if there are genes that are significantly regulated by hypoxia. A classification rule is built able to discriminate the cell lines depending on their hypoxic status, hence the classification error is the parameter which evaluates the goodness of the model. A gene is considered to be relevant if it is contributes in building a multivariate discriminative model for Hypoxia. By tuning the correlation parameter epsilon we are

enabled to define a minimal list of genes that are independent with each other and also one (or more) correlation aware lists comprising intra-correlated genes. We trained the system with two different values for epsilon obtaining a minimal list of probesets and a correlation-aware list. Since the framework works within a leaveone-out cross validation loop every element of the lists is associated with a frequency parameter. When the relative frequency is above roughly 30% (6/18) we select 16 variables for the correlation aware list and 11 for the minimal list. The leave one out error for both is 17% (3 out of 18). In this case standard methods such as the hypothesis testing with a Benjamini and Hochberg correction for controlling the False Discovery Rate do not identify any relevant probeset, while our procedure is able to detect a panel of significant probesets that succeed in discriminating the hypoxic status.

Contact : annalisa.barla@unige.it