Modeling of bioassay data and genes expression in drug discovery experiments: A supervised principal component analysis approach

Theophile Bigirumurame[1], Nolen Perualila-Tan[1] ,Ziv Shkedy[1] , Adetayo Kasim[2]

[1] Interuniversity Institute for Biostatistics, Hasselt University, Belgium, [2]*Wolfson Research Institute, Durham University Queen's Campus, University Boulevard, Thornaby, Stockton-on-Tees, UK*

INTRODUCTION

Nowadays, microarray technology are used to monitor simultaneously the activity of thousands genes and their response to a certain treatment.

However, microarray data imply high dimensionality of the data and relatively small number of observations (Amaratunga, 2014).

Large number of genes, but a small number of them is associated with the response and can be used in further analysis (predictive model).

We present a resampling based feature selection procedure based on Supervised Principal Component Analysis (SPCA),

We combine information from three data source to develop a predictive model for the bioassay based on the gene expression controlling the chemical structure (Figure 1).

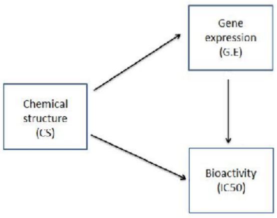

Fig 1:Data structure

SUPERVISED PRINCIPAL COMPONENT ANALYSIS (SPCA)

Bair et al. (2006) proposed SPCA to construct gene profile used in prediction of quantitative response.

SPCA reduce the dimension of the expression matrix X **.** It ensures that genes having strongest relationship with the response are used to build the biomarker.

The analysis consists of four steps:

1. Fit gene-specific models and estimate the association measure.

$$E(Y_i|X_{ij},Z_i) = \mu + \beta Z_i + \alpha_j X_{ij}.$$

where Z_i is a fingerprint, X_{ij} is the jth gene expression (j=1,…,m), on the ith compound (i=1,…,n), Y_i is the bioassay (IC50).

2. Form a reduced expression matrix consisting of genes with specific association measures.

3. For each reduced matrix X_R, compute the first principal component $\mathrm{U}(X_R)_i$ (the biomarker).

4. Fit the predictive model

$$Y_i^* = \beta_0 + \beta_1 \mathrm{U}(X_R)_i + \beta_2 Z_i + \varepsilon_i.$$

where Z_i is the fingerprint, and $\mathrm{U}(X_R)_i$ the biomarker.

GENE SELECTION

The data is split into a training set (2/3) and test set (1/3). The process is done in loop shown in Figure 2 .

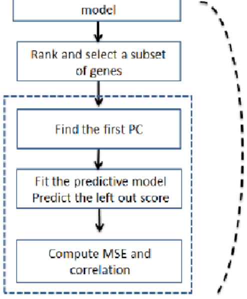
Fit the joint or the marginal model
Rank and select a subset of genes
Find the first PC
Fit the predictive model
Predict the left out score
Compute MSE and correlation

Fig 2:Flowchart of steps involved in the SPCA approach

The inner square in Figure 2 shows the iterative step three, in which at each iteration a new gene is added to build a new PCA. The outer broken line represents the outer loop in which the whole procedure is repeated a 1000 times.

DATASET

Data from a drug development project in oncology is used to illustrate the applicability of the SPCA. The dataset focuses on inhibition of the epidermal growth factor receptor. There were 3595 genes after pre-processing. We used one fingerprint feature across 35 compounds, and one bioassay (IC50).

RESULTS

The loop in figure 1 was run 1000 time, and genes were ranked based on their selection frequencies. Figure 3 shows how the correlation between the observed and predicted bioactivity changes.

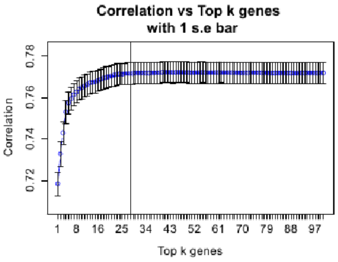

Fig3: Correlation between the observed and predicted bioassay using top k genes in the profile.

When 28 genes were used to build the profile, there was no more improvement in the correlation. The next stage was to identify those genes.

Three different approaches were used to build the joint biomarker:

1. We construct the first principal component using the top k genes. After the inclusion of the fourth gene, the correlation between the bioassay and the joint biomarker decreases (Figure 4 left panel).

2. We include a gene , if it results in the increase of the correlation between the joint biomarker and the bioassay. 12 genes give an determination coefficient equal to 89% before it starts to decrease again.

3. We build the first PC using top k genes, rank them based on their loadings, then build a biomarker based on the top k genes with the largest loadings. Figure 4 right panel shows the resulting correlations. The highest correlation is obtained with top 25 genes in the biomarker.

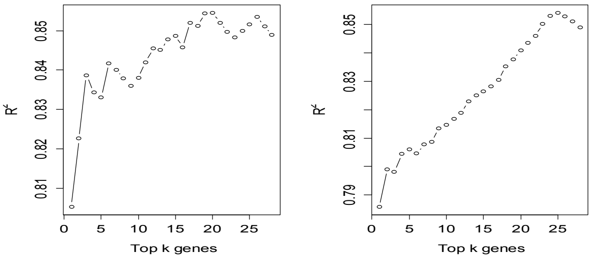

Fig4: Correlation between the bioassay and the joint biomarker made of the top k genes. Left panel, joint biomarker built using the first approach; right panel the joint biomarker built using the third approach

DISCUSSION AND CONCLUSIONS

In this paper, we were interested in combining information from three data sources to construct a predictive model using SPCA.

Three different methods were used. The first gave a biomarker with about 21 genes, the second with 12 genes and the last one with 25 genes.

The obtained association measures between the joint biomarker and the bioassay seem to be high and their significance have to be tested.

A permutation test can be performed to check this significance.

REFERENCES

[1] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, **101**, 119-137.

[2] Chen, X., Wang, L., Smith, J. D., and Zhang, B. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics (Oxford, England), **24**(21), 2474–81..

[3] Amaratunga, D., Cabrera, J., and Shkedy, Z.,. Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, Wiley (2014).