Multilevel Modeling to Estimate the Gene-Drug Activity Matrix

Victor Moreno

Epidemiology Service, Catalan Institute of Oncology

The most interesting objective of the NCI-60 Cancer Cell Lines database is the exploration of significative associations between drug treatments and gene expression. The difficulty in this analysis is to get an adequate estimate of the association, taking into account that each dimension of information (drug activity and gene expression) relate to the same cell lines but have been gathered at different levels. We should remember the authors' recognition that these relationships are correlative, not casual. The authors used the Pearson correlation coefficient calculated for each combination of a gene and a drug. This association measure is interesting because is easy to calculate and symmetric. We propose here the use of a modelling approach to esimate the associations. This method is more elaborated and may be computationally prohibitive for large dataset but accounts better for the structure of the information available.

We propose a multilevel model to study the GI50 as the response of interest to predict. At the first level one coefficient for each drug measures average differences in GI50. At the second level, expressions for each gene are the covariates that act as contextual information and have specific error terms (random effects). Our interest is to estimate the drug-gene expression interaction terms, across levels, thaikng into account the hierarchical structure of the data. The model equations would be:

$$\textit{Level 1:}\quad GI_{50_{ij}} = a_i drug_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} = N(0,\sigma^2)$$

$$\textit{Level 2:}\quad a_i = b_j\, gene_j / I_j + U_j$$

$$U_j = N(0,\tau_j)$$

The parameter estimation of this model needs special software but there are several possibilities available, both comercial (for example SAS proc mixed, S-Plus nlme library, Mlwin) and free (R nlme library and BUGS, for a bayesian approach).

This mixed model approach provides a test of significance for the association parameters that may be used to evaluate and select relevant associations. The main limitation of the procedure is the computational burden needed to estimate the parameters in multivariate models.

After the estimation of the associations, other exploratory tools such as cluster analysis can be used to further classify the drugs and genes. The examples described in the paper (5-FU and L-asparaginase) will be studied with this approach and compared with the simpler correlation method.