Integrated analysis of multi-source data in drug discovery experiments using structural equation models

Integrated analysis of multi-source data in drug discovery experiments using structural equation models

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

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Hasselt, Belgium
[2] Wolfson Research Institute, Durham University Queen's Campus,Durham, U.K

E-mail for correspondence: theophile.bigirumurame@uhasselt.be

Abstract: The drug discovery and development processes are typically costly and time consuming. Hence, it is crucial to identify early failure of candidate compounds and thereby save time and investment in a later stage. We propose structural equation modeling (SEM) based approach for an integrated analysis which combines information from three data sources: (1) bioactivity variables, (2) variables representing the chemical structure of the compounds, and (3) gene expression data. The proposed model allows to estimate the effects of the gene expression on the biological activity variable and furthermore, it allows to decompose the effect of the chemical structure on the biological activity into direct and indirect (i.e. the effect via the gene expression) effects.

Keywords: Structural equation modeling; Microarray data; Bioassays.

1 Introduction

The drug discovery and development processes are typically costly and time consuming. Hence, it is crucial to identify early failure and thereby save time and investment in a later stage. The decision to continue/stop a development process in drug discovery should ideally be based on scientific parameters that are predictive of later outcomes, and which can be determined quickly and at relatively low cost.
Currently, microarray technology (Amaratunga *et al.*, 2014) is used to monitor simultaneously the activity of thousands genes and their response to a certain drug. Microarrays are providing new insights into the molecular

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

pathology of human cancers and are helping to identify many new additional targets for drug discovery. By understanding gene expression patterns, researchers can gain information that can link sites of expression, biochemical pathways, and normal or pathological functions in organs and whole organisms.
However, relevant biologically data are acquired in a late stage of the research process. The use of biomarkers can reduce the costs and increase efficiency, and they should be incorporated early in the development process to gain information that can aid the development. In this paper we propose an approach based on structural equation modeling to combine information from the three data source, i.e., the bioactivity, the chemical structure of the compound, and the gene expression, and select a subset of genes which can be used as biomarkers.

2 Methodology : Structural Equations Modeling (SEM)

Three data sources were obtained from an oncology project, which focused on the inhibition of the fibroblast growth factor receptor (Verbist et al., 2015). The chemical structure (presence or absence of fingerprint feature, FFP, in a compound/molecule), the gene expression data and the bioactivity (IC50) outcome. Let X_{ij} be the j^{th} gene expression ($j = 1, 2, ..., 3595$), of the the i^{th} compound ($i = 1, 2, ..., 35$). The measurement for the bioactivity is denoted by Y_i. Let Z_i be an indicator variable, which takes value 1 if the fingerprint feature (FFP) is present in the i^{th} compound, and 0 otherwise.
The key idea behind structural equation models(SEM) is that the causal relationships among the variables determine the expected pattern of correlation Li *et al.* (2006). For the analysis presented in this paper, SEM with observed variables were considered (Bollen, 1989). The main advantage of our approach is that it allows to decompose the total effect of chemical substructure on the bioassay into the direct (effect of the Z on Y unmediated by X) and indirect effects (effect of the Z on Y, mediated by X). Our primary interest is to select genes which maximize the indirect effects. The indirect and direct effects are shown in Figure 1.
The SEM consists of a structural model (i.e., a path analysis model) which describes the casual relationship between the variables. The model is visualized in Figure 1 (right panel). Formaly the model can be expressed as set of two model given by:

$$\begin{aligned} X &= \gamma_1 Z + \varepsilon_1 \\ Y &= \gamma_2 Z + \beta X + \varepsilon_2 \end{aligned} \tag{1}$$

where: γ_1 and γ_2 are the fingerprint effects on the gene expression and the bioassay respectively, β is the gene specific effects on the bioassay,

ε_1 and ε_2 are the uncorrelated measurement errors. It is assumed that $(\varepsilon_1, \varepsilon_2) \sim N(0, \psi)$, $var(Z) = \phi$, and $Cov(\varepsilon_i, Z) = 0$. The indirect effect of the FFP for a given gene j is equal to $\gamma_{1j} * \beta$, whereas the direct effect is γ_2.

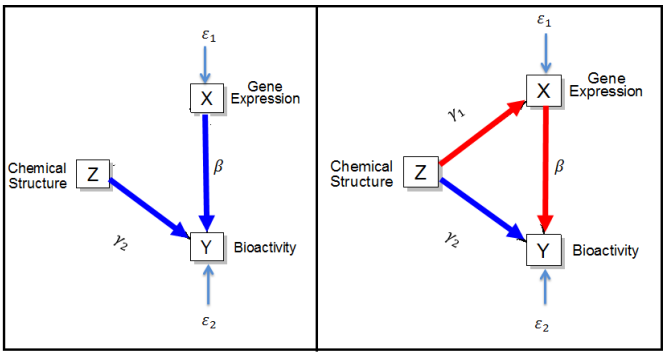

FIGURE 1. Fingerprint feature and gene expression direct effects on the bioassay (blue arrows in the left panel. Fingerprint feature indirect effect on the bioassay through the gene expression (red arrows in the right panel)

The unknown parameters can be estimated using the maximum likelihood estimation method. The model in Equation (1) is fitted gene by gene, and multiple testing adjustment using FDR (Benjamini and Hochberg, 1995) is performed to find significant parameters.

3 Results

In this section we present results for one of the genes that was selected using the SEM. This gene (gene 1) corresponds to the subset of genes which maximize the indirect effects. For these subset of genes , we expect to explain most of the FFP effects on the bioassay through the FFP effects on the gene expression. This type of genes is characterized by high FFP direct effect on the gene expression and high gene direct effect on the bioassay. These genes have relatively high correlation between the gene expression and the bioassay and they are differentially expressed. Figure 2 shows a typical gene. The indirect effect was equal to -0.56 and the direct effect was equal to -0.11. Note that BH-FDR procedure was applied to correct for multiple testing

4 Discussion

There are many challenges in the drug discovery and development. Relevant biological data are acquired too late in the research processes and the use of

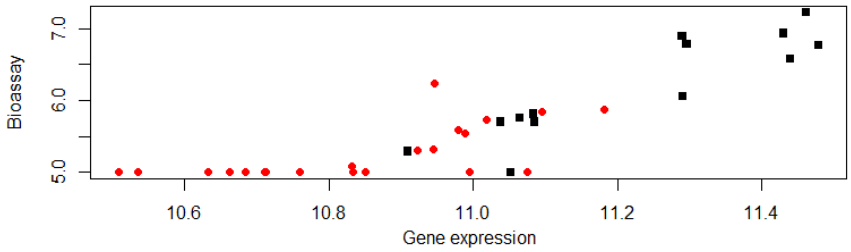

FIGURE 2. Typical gene which maximizes the indirect effect.

biomarkers can reduce the cost and increase efficiency. The SEM presented in this paper can be used to select genetic biomarkers which can help in the development process. Genes maximizing the indirect effect could help in explaining the the effect of the FFP on the bioassay.
After detecting significant genes, one can find to which pathways they belong , in order to have an insight about the mechanism of action of given chemical. The gained information, thus could help in lead optimization. If toxicity related genes are identified, it could help in deciding to continue/or stop the development process with compounds having a given chemical substructure.

References

Amaratunga, D. , Cabrera, J. and Shkedy, Z. (2014),*Exploration and Analysis of DNA Microarray and Other High Dimensional Data*. New York: John Wiley.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289 - 300.

Bollen, K.A. (1998). *Structural equation models*. New York, Wiley Series in Probability and Mathematical Statistics. Wiley.

Li, R. , Tsaih S.W., Shockley, K., Stylianou, I.M., Wergedal, J., Paigen, B., Churchill, G.A. (2006). Structural model analysis of multiple quantitative traits. *PLoS Genet*, **2**(7):e114.

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., QSTAR Consortium, Shkedy, Z., Thas, O., Bender, A., Hinrich, W., Göhlmann, H., Hochreiter, S. (2015). Using Transcriptomics to Guide Lead Optimization in Drug Discovery Projects: Lessons Learned from the QSTAR Project. *Drug Discovery Today, In Press*.