Development of Genetic Biomarkers in Drug Discovery and Early Drug Development Experiments

Nolen Perualila-Tan* Adetayo Kasim† Willem Talloen‡
Hinrich W.H. Göhlmann§ Ziv Shkedy¶

Abstract

Quantitative Structure Transcriptional Activity Relationship (QSTAR) involves relating the three data sources in early drug discovery namely (1) chemical structure (fingerprint features), (2) bioactivity (bio-assay read-outs) data for targets of interest, and (3) transcriptomic (gene expression) data of a set of compounds. In this paper, a gene-specific and fingerprint feature-specific joint model is presented as a tool to model the association between gene expression and biological activity taking into account the chemical structure of the compounds. The model allows to detect genes that are associated to the bio-assay read-out for which some of the associations are mainly induced by certain fingerprint feature(s) of compounds. The joint model is applied to two oncology projects. Results show that a number of compounds' fingerprint features have differential effects on both bio-assay read-outs and a set of correlated genes.

Key Words: Bioactivity, Chemical structures, Joint model, QSTAR, Transcriptomic

1. INTRODUCTION

Selecting candidate molecules for an already defined biomolecular target in the pre-clinical stages of drug development involves the analysis of several data sources to get a better understanding of their chemical properties and biological activities or mechanism of action. A set of compounds with observed activity may still need structural modifications either to fit better to the target or to eliminate undesirable chemical features. Biological and chemical information of compounds can be quantified in various ways. The structural information of the compounds can be encoded using different molecular descriptors. A comprehensive discussion of different molecular descriptors in chemoinformatics is given by Todeschini and Consonni (2009). The Extended Connectivity Fingerprints (ECFPs) were developed specifically for structure-activity modeling (Rogers and Hahn, 2010). This is characterized by a vector of binary values, also known as molecular fingerprint that describes which chemical features are present or absent in the molecule. For the bioactivity data, typically, the efficacy of the candidate compounds can be measured via the dose-response experiments wherein a range of compound concentrations is tested in a target-based assay to assess the concentration or dose dependence of the assay's readout. This is usually expressed as an IC50 or as an EC50 in enzyme-, protein-, antibody-, or cell-based assays.

Another source of biological activity data (including on-target and off-target effects) in early drug discovery is the use of gene expression profiling (Bai et al., 2013). This technique measures multiple biological effects of a compound on a whole genome transcriptional level, and thereby gives an information-rich snapshot of the biological state of

*Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, Belgium

†Wolfson Research Institute for Health and Wellbeing, Durham University, United Kingdom

‡Janssen Pharmaceutica NV, 2340, Beerse, Belgium

§Janssen Pharmaceutica NV, 2340, Beerse, Belgium

¶Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, Belgium

a cell (Göhlmann and Talloen, 2009; Amaratunga, Cabrera, and Shkedy, 2014). Besides being able to provide information on thousands of genes in one experiment, microarray experiments are also fairly rapid, and relatively low-priced in contrast to assay development (Fadiel and Naftolin, 2003). It has also been observed that transcriptomic data mostly detect biologically relevant signals and are often able to help prioritizing compounds beyond conventional target-based assays. Moreover, this enables us to investigate downstream effects of candidate drugs through pathway-associated gene signatures. This offers the chance of finding a biological basis for the disease and biomarkers involved in the disease pathway.

Biomarker identification is a major application for microarray experiments in early drug development which often parallels and facilitates compound selection. Hence, many studies have been devoted to identify genes that are associated to the biological activity of interest, the inhibition of a certain enzyme, for instance. It is also equally important to detect toxicity at the early stages of development. Reliable biomarker for toxicity can be very helpful in this respect as it allows cost-effective testing of other drug candidates and leads in compound series under investigation. For example, Lin et al. (2010) and Tilahun et al. (2010) identified gene-specific biomarkers for continuous outcomes (the distance traveled by a rat under treatment and the HAMD scores for psychiatric patients, respectively). Van Sanden et al. (2012) identified gene specific biomarkers for toxicity data presented as a binary response. This paper exemplifies the usefulness of a joint model, a well established tool in finding genetic biomarkers, applied within the Quantitative Structure Transcriptional Activity Relationship (QSTAR) framework. Figure 1 displays the biomarker setting where the joint model can be applied. It shows how the transciptomic variable X is associated with the clinical outcome of interest Y given that both can be influenced by a condition Z. It is a highly flexible technique to find a biomarker−endpoint pair that are both driven by the same factor. The factor can be any binary variable such as treatment/control, ac-

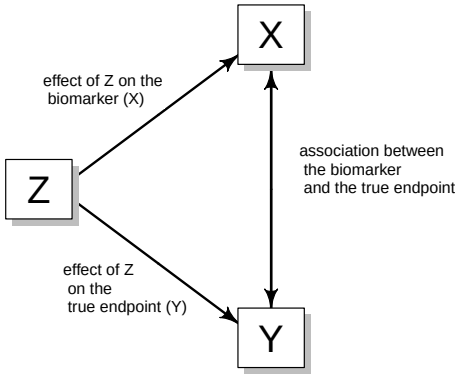


Figure 1: The Biomarker setting. The arrows represent the association of two continuous responses (biomarker X and the endpoint Y) and 1 common explanatory variable (Z) that can be quantified by using a joint model.

tive/inactive compounds, compounds with/without tox effects, binary chemical descriptors, etc. The joint modeling approach fits in the QSTAR framework wherein it investigates the relationship between bio-assay data (Y) and transcriptomic data (X) taking into account the presence/absence of a chemical substructure (Z) of a compound set. The analysis is performed gene-by-gene and fingerprint-by-fingerprint. This approach provides a solution that is very helpful in extracting relevant information from the high dimensional and complex microarray and chemical data.

In this particular application, the joint model provides a list of genes that are associated

with the bioactivity read-outs, but taking into account that both the gene expression levels and the bioactivity read-outs could be influenced by a chemical substructure(s) that is(are) inducing the observed association. In this regard, finding relevant genes that are linearly related to the biological response is already a valuable information per se but noting that this linear relationship is caused by the presence or absence of a particular chemical substructure(s) provides another level of information in designing new molecule, in improving drugs or in prioritizing compounds to carry on in the next phase of drug discovery.

The joint modeling framework that we proposed in this paper allows us to: (1) identify gene signatures of activity for directing chemistry, (2) determine chemical substructures (also termed as fingerprint features, FF) of compounds that are related with effects on the bio-assay data for target(s) of interest and (3) know whether this effect can also be confirmed by the gene expression changes (either on- or off- target related).

1.1 Graphical illustration of the different types of genetic biomarkers for compounds' efficacy

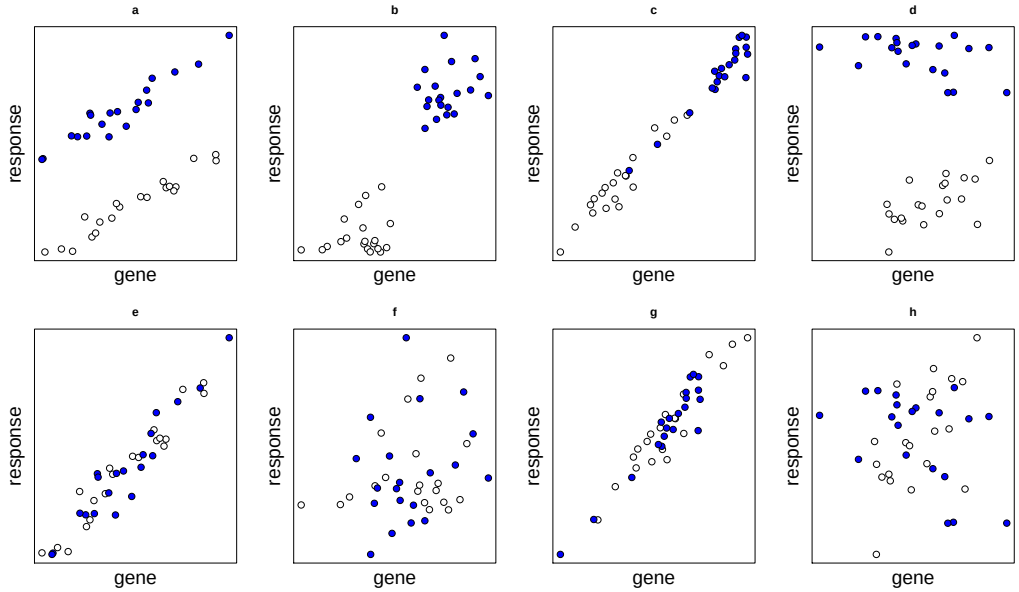


Figure 2: Hypothetical examples of the association between response variable to expression levels when the response is differentially expressed. Each point represents a compound. Solid blue points and black points represent the presence/absence of a fingerprint feature , respectively. Upper row: scatterplots for the response versus the gene expression. Lower row: scatterplots for the residuals after adjusting for fingerprint effects.

Several interesting associations between genes and a response accounting for the effect of a fingerprint feature can be discovered by using a joint model. The different types of association are illustrated in Figure 2 using a hypothetical data. Each point in the plot represents a compound and the solid ones are compounds having the fingerprint feature.

For this application, the interest lies only on the fingerprint feature that shows differential effects on the bioactivity, the response in this case; thus the four possible scenarios between the gene expression and response variable of interest presented in the upper panels of Figure 2(a-d). The lower panels(e-h) display the same data with their respective upper panels adjusted for fingerprint feature effect for both the response and the gene expression.

In panel (a) the gene is not differentially expressed and has a linear association with the response irrespective of the presence/absence of the fingerprint feature. Note that the linear pattern remains after adjusting for the effect of the fingerprint feature as shown in panel

(e). Panel (b) shows an example in which the gene is differentially expressed, the clouds of points are clearly separated in both dimensions. Moreover, it can be observed that within the group, the association between the gene expression and the response does not have a linear pattern, which is evident in panel (f) after the adjustment. Note that for this pattern, the association between the gene expression and the response in induced by the chemical structure. Therefore, when the model is adjusted for the chemical effect the association disappeared (Figure 2 f).

Panel (c) shows a combination of the previous two patterns. Both the gene expression and the response are differentially expressed, that is compounds having the fingerprint feature are inducing higher activity than those that don't have the feature. In this setting, the association between the gene expression and the response can be summarized by a straight line, this can be clearly seen from panel (g) which shows the same example after adjusting for fingerprint effects.

Lastly, most genes are expected to be uncorrelated with the bioassay as depicted by panel (d). Within each group of compounds (with and without the fingerprint feature), linear pattern is not evident; thus, adjusting for this effect also provides a random scattering of points (panel (h)).

2. Materials and Method

2.1 Data

Data from two drug development projects in oncology are used to illustrate the applicability of the joint model. For each project, information about the three data sources, transcriptomic, phenotypic and chemical structure data is available for each compound.

The ROS1 dataset consists of eight-nine (89) compounds tested for target inhibition. The cellular assay provides the inhibitory activity measurements of the compounds given by the IC50, half-maximal inhibitory concentration. In this analysis, the pIC50 scale (-log IC50) is used, in which higher values indicate exponentially greater potency. A total of 1289 genes were retained after the pre-processing steps. Moreover, a total of 312 unique profiles of chemical substructures using Extended-Connectivity Fingerprints with a search depth of 6 (ECFP6) were identified for this compound set. The chemical structure data is given as binary variables which denote the presence/absence in a compound of a certain chemical substructure also termed as fingerprint feature (FF).

The EGFR dataset focuses on inhibition of the epidermal growth factor receptor. Thirty-five compounds with a macrocycle structure were profiled in order to identify compounds with similar biological effects as the current EGFR inhibitors, gefitinib and erlotinib, serving as the reference compounds. Gene expression profiles are available for 3595 genes after pre-processing. For this project, a total of 138 unique profiles of fingerprint features across 35 compounds was generated.

2.2 The Statistical Model

Let $\mathbf{X}$ be the gene expression matrix where X_{ij} is the j^{th} gene expression $j = 1,\ldots,m$, of the i^{th} compound, $i = 1,\ldots,n$, and denote the measurement for the bioassay by Y_i. Both gene expression and bio-assay read-outs are assumed to be normally distributed. Let $\mathbf{Z}$ be the fingerprint feature matrix where Z_{ki} be an indicator variable representing the k^{th} fingerprint feature. Note that the three data sources are connected by compounds. For a given fingerprint feature, the gene-specific joint model that allows testing for which gene is also differentially expressed and which gene is predictive of the response irrespective of the effect of the fingerprint feature is given as follows:

$$\begin{aligned} X_{ij} &= \mu_{jk} + \alpha_{jk} Z_{ki} + \varepsilon_{ijk} \\ Y_i &= \mu_{Yk} + \beta_k Z_{ki} + \varepsilon_{ki} \end{aligned} \tag{1}$$

or equivalently formulated as

$$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_{jk} + \alpha_{jk} Z_{ki} \\ \mu_{Yk} + \beta_k Z_k \end{pmatrix}, \Sigma_{jk} \right] \tag{2}$$

where the error terms have a joint zero-mean normal distribution with FF-specific and gene-specific covariance matrix, Σ_{jk}.

$$\Sigma_{jk} = \begin{pmatrix} \sigma_{jjk} & \sigma_{jYk} \\ \sigma_{jYk} & \sigma_{YYk} \end{pmatrix} \tag{3}$$

The parameters α_{jk} and β_k are the ML estimates of the k^{th} fingerprint feature effects for the j^{th} gene and the response, respectively, and μ_{jk} and μ_{Yk} are gene- and fingerprint-specific and the response-related intercepts, respectively.

Thus, the gene-specific association with the response can be obtained using adjusted association (Buyse and Molenberghs, 1998; Amaratunga, Cabrera, and Shkedy, 2014), a coefficient that is derived from the covariance matrix, Σ_{jk}, of gene-specific joint model (Eqn. 3):

$$\rho_{jk} = \frac{\sigma_{jYk}}{\sqrt{\sigma_{jjk}\sigma_{YYk}}}. \tag{4}$$

Indeed, $\rho_{jk} = 1$ indicates a deterministic relationship between the gene expression and the response after accounting for the effect of the k_{th} fingerprint feature.

2.2.1 Testing for differentially expressed genes

The model allows testing for differentially expressed genes, hence for each gene, we test the hypotheses

$$\begin{aligned} H_{0_{jk}} &: \alpha_{jk} = 0, \\ H_{1_{jk}} &: \alpha_{jk} \neq 0. \end{aligned} \tag{5}$$

For a microarray with m genes, there are m null hypotheses to be tested, which implies that an adjustment for multiple testing should be applied. Throughout this paper, we apply the FDR approach proposed by Benjamini and Hochberg (1995).

2.2.2 Testing for the association between the gene expression and the bioactivity data after accounting for the effect of achemical structure

In order to make inference about ρ_{jk}, there is a need to test whether the expression level of a gene and the bio-assay read-out are correlated, specifically, whether the expression level of a gene can predict the bio-assay read-out. Thus, in addition to the hypotheses in (5), one needs to test the hypotheses

$$\begin{aligned} H_{0_{jk}} &: \rho_{jk} = 0, \\ H_{1_{jk}} &: \rho_{jk} \neq 0. \end{aligned} \qquad \text{or equivalently} \qquad \begin{aligned} H_{0_{jk}} &: \sigma_{jYk} = 0, \\ H_{1_{jk}} &: \sigma_{jYk} \neq 0. \end{aligned} \tag{6}$$

Under the null hypothesis, the joint model in (2) is reduced to

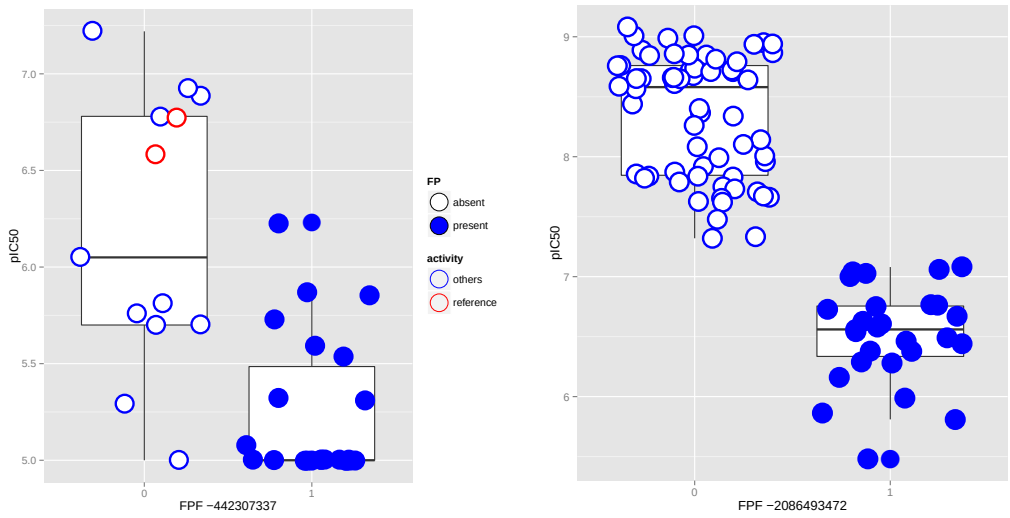
$$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_{jk} + \alpha_{jk} Z_{ik} \\ \mu_{Yk} + \beta_k Z_k \end{pmatrix}, \Sigma_j = \begin{pmatrix} \sigma_{jjk} & 0 \\ 0 & \sigma_{YYk} \end{pmatrix} \right] \tag{7}$$

Consequently, the inference for the adjusted association can be based on a likelihood ratio test by comparing models in (2) and (7). Asymptotically, the likelihood ratio statistic follows a χ^2 distribution with one degree of freedom. Benjamini and Hochberg (1995) procedure is used to adjust for false discovery rate when testing for the null hypotheses of $H_{0_j} : \rho_{jk} = 0$ for all the genes simultaneously by fingerprint feature.

3. Results

Given that the chemical structure effect upon the activity data is present, the different types of genetic biomarkers for compound efficacy presented in Figure 2 can be obtained from the hypothesis testing in (5) and (6). In this early drug development set up, the interest lies on two gene classes where the chemical structure has a significant effect on both the gene ($\alpha \neq 0$) and the response as shown in Figure 2 b and c. For the first group of genes, the association is driven by the FP effect whilst for the second group of genes the association between the gene expression and pIC50 exists regardless of the effect of a chemical substructure of the compound.

This paper covers the results obtained from the joint model using a fingerprint feature that is most responsible for the variation in compound activity. Specifically, FF-442307337 and FF-2086493472 are ranked first based on a feature-by-feature two-sample t-test of bioactivity data for the EGFR and ROS1 projects, respectively. These substructures are prominent on less potent compounds, i.e. those with pIC50 values less than 6.5 (Figure 3). Figure 4a shows the chemical structure of FF-442307337, an oxygen in ortho position of



(a) FF -442307337 for the EGFR project (b) FF-2086493472 for the ROS1 project

Figure 3: Top fingerprint feature differentiating bioactivity.

the aniline (highlighted in red). This structure is not present in two reference compounds gefitinib and erlotinib (Figure 4b,c) as well as in potent compounds. However, there are also some less potent compounds that do not have this feature which could mean that this substructure is probably not the sole reason for compounds' lower activity. For the ROS1 project, most of the differentially expressed genes belong to the first gene class, that is the correlations observed between the pIC50 and gene expression can be attributed to this substructure as the correlation disappears after adjusting for this chemical feature (239 versus 139 genes, see Table 1). The top 5 genes for this gene class are displayed in Figure 5 where it can be clearly seen that the slope of the lines in the upper panels significantly

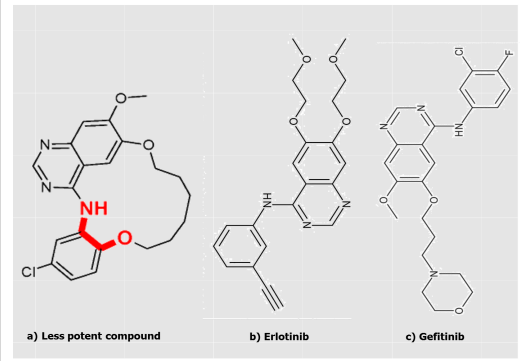


Figure 4: Chemical structures of a) identified less potent compound; and the two reference compounds in EGFR project b) erlotinib and c) gefitinib.

Table 1: Testing the null hypotheses in (5) and (6) for the ROS1 project for FF -2086493472 at 5% FDR.

		ρ_j	
		H_0 is rejected	H_0 is not rejected
α_j	H_0 is rejected	139	239
	H_0 is not rejected	382	529

dropped to around zero after the adjustment of the effect of the chemical structure as shown in their corresponding lower panels. For this group, the joint model indicated very low adjusted correlation between the genes and the activity.

Table 2: Testing the null hypotheses in (5) and (6) for the EGFR project for FF -442307337 at 5% FDR.

		ρ_j	
		H_0 is rejected	H_0 is not rejected
α_j	H_0 is rejected	396	61
	H_0 is not rejected	1099	2039

For the EGFR project, results show that most of the differentially expressed genes are still associated with the bioactivity data upon adjustments of the effect of the chemical structure (396 versus 61, see Table 2). Figure 6 shows the 5 most differentially expressed genes with the adjusted association remaining high after adjustment of the chemical structure. The plots in the lower panels still follow the same linear patters with their respective upper panels. Most of these genes are known to participate in biological processes involving cell proliferation (positive and negative), survival and differentiation.

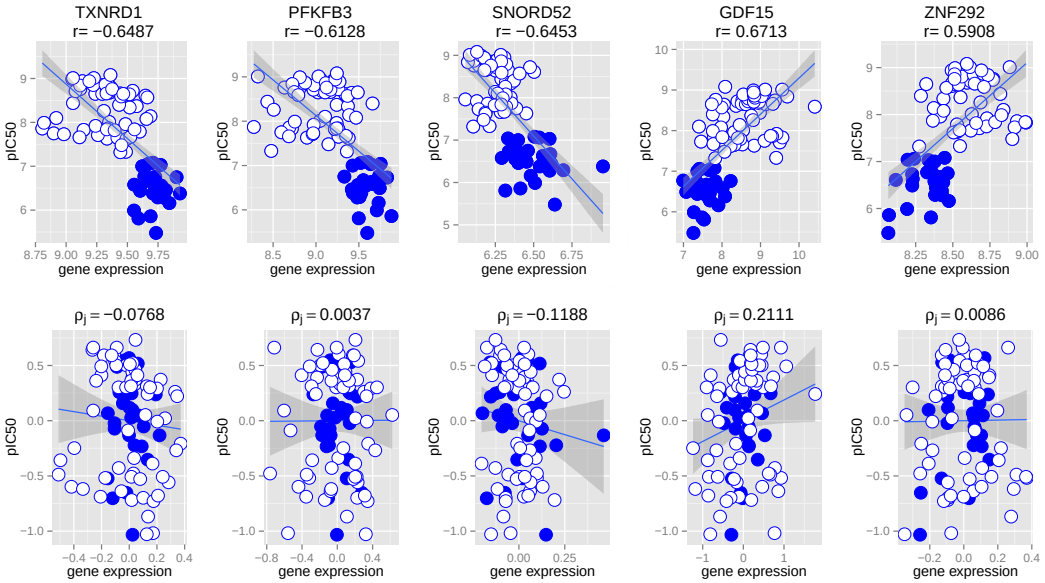


Figure 5: Scatterplot of the expression level and bioactivity data given by the pIC50 (upper panel) and their corresponding residuals from the joint model (lower panel) of the top 5 differentially expressed genes with low adjusted correlation. The correlation between the gene expression and the inhibitory activity against ROS1, of the compounds (represented by points in the plots) can be explained by the substructure FF-2086493472.

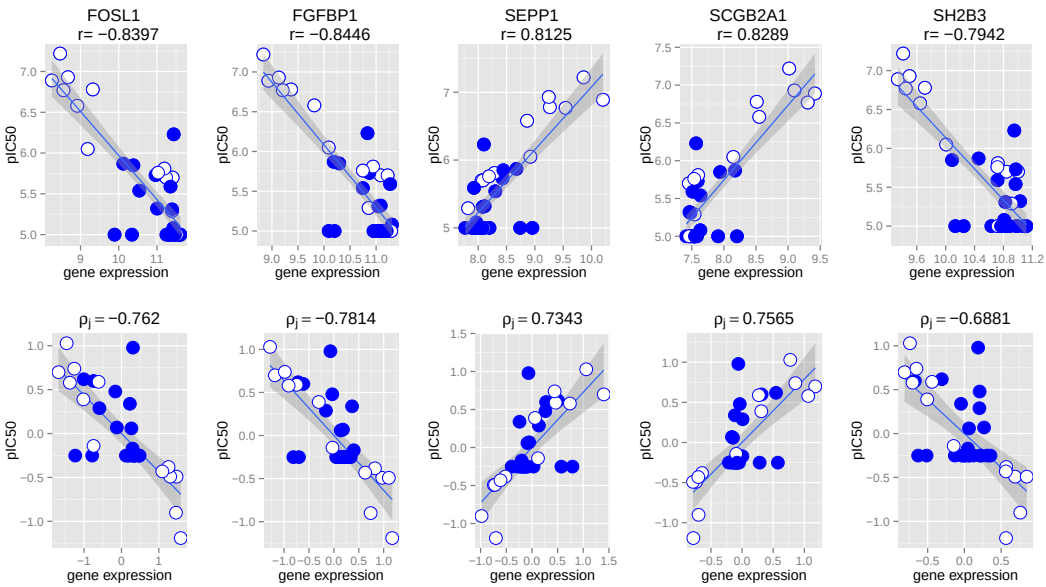


Figure 6: Scatterplot of the expression level and bioactivity data given by the pIC50 (upper panel) and their corresponding residuals from the joint model (lower panel) of the top 5 differentially expressed genes with high adjusted correlation for the EGFR project.

4. Discussion and Conclusion

Identification of transcriptomics biomarkers is not limited only for classification problems but can be applied for prediction problems as well. In this paper, we have shown that the joint model can identify genes that are predictive of compound activity, measured by pIC50, and can therefore serve as genetic biomarkers for compounds efficacy. In addition, we have shown that the effect of a particular chemical substructure on the expression level of each gene and/or its influence on the observed transcriptomic-phenotypic association can be estimated.

The joint modeling approach, although implemented using only one feature at a time for every data source, facilitates the extraction of valuable insights on compounds structural and biological mechanisms. Although, we focused in this paper on one fingerprint feature and on-target assay per project, this method can easily be run in loops. In the pharmaceutical pipeline implementation, this model is applied to all or to a number of interesting chemical substructures, genes and biological assays (efficacy or toxicity related). The large amount of output can then be collated and filtered for vital information that can help the research team, especially, the medicinal chemist and biologist in taking the next step in the drug development process.

The joint modeling of bioactivity and gene expression data not only confirms the underlying biological mechanisms of candidate compounds but also models the association existing between the responses when accounting for the effect of a chemical substructure. It would be interesting from a lead optimization angle, if a structure is actually responsible for driving the association. The effect of a promising fingerprint feature could be experimentally validated to determine whether chemical modification of compounds involving this substructure may improve compounds' activity. In addition, the datasets in early drug development experiments are typically of high dimension and a multivariate approach that integrates all these datasets could be performed. Even then, the joint model proposed in this paper could still be very helpful in extracting relevant information from the high dimensional and complex microarray and chemical data and providing an answer to the relevant research questions posed by drug development teams in the pharmaceutical companies.

References

Amaratunga, D., J. Cabrera, and Z. Shkedy (2014): *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*, Wiley Series in Probability and Statistics.

Bai, J. P. F., A. V. Alekseyenko, A. Statnikov, I.-M. Wang, and P. H. Wong (2013): "Strategic applications of gene expression: from drug discovery/development to bedside." *The AAPS Journal*.

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

Buyse, M. and G. Molenberghs (1998): "The validation of surrogate endpoints in randomized experiments," *Biometrics*, 54, 186–201.

Fadiel, A. and F. Naftolin (2003): "Microarray applications and challenges: a vast array of possibilities," *Reproductive Sciences*, 1, 1111–21.

Göhlmann, H. and W. Talloen (2009): *Gene Expression Studies Using Affymetrix Microarrays*, Chapman & Hall/CRC Mathematical & Computational Biology.

Lin, D., Z. Shkedy, G. Molenberghs, W. Talloen, H. Gohlmann, and L. Bijnens (2010): "Selection and evaluation of gene-specific biomarkers in pre-clinical and clinical microarray experiments," *Online Jounal of Bioinformatics*, 11(1), 106–127.

Rogers, D. and M. Hahn (2010): "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*.

Tilahun, A., D. Lin, Z. Shkedy, H. Geys, A. Alonso, P. Peeters, W. Talloen, W. Drinkenburg, H. Gohlmann, E. Gorden, L. Bijnens, and G. Molenberghs (2010): "Genomic biomarkers for depression: Feature-specific and joint biomarkers," *Statistics in Biopharmaceutical Research*, 2(3), 419–434.

Todeschini, R. and V. Consonni (2009): *Molecular Descriptors for Chemoinformatics*, Wiley.

Van Sanden, S., Z. Shkedy, T. Burzykowski, H. Gohlmann, W. Talloen, and L. Bijnens (2012): "Genomic biomarkers for a binary response in early drug development microarray experiments," *Journal of Biopharmaceutical Statistics*, 22(1), 72–92.