Principal bicorrelation analysis: Unraveling associations between three data sources

Federico Mattiello\textsuperscript{a}, Olivier Thas\textsuperscript{a,b}, and Bie Verbist\textsuperscript{c}

\textsuperscript{a}Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Gent, Belgium; \textsuperscript{b}National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia; \textsuperscript{c}Janssen Pharmaceutica, Beerse, Belgium

ABSTRACT
In this article, we propose a statistical explorative method for data integration. It is developed in the context of early drug development for which it enables the detection of chemical substructures and the identification of genes that mediate their association with the bioactivity (BA). The core of the method is a sparse singular value decomposition for the identification of the gene set and a permutation-based method for the control of the false discovery rate. The method is illustrated using a real dataset, and its properties are empirically evaluated by means of a simulation study. Quantitative Structure Transcriptional Activity Relationship (QSTAR, www.qstar-consortium.org) is a new paradigm in early drug development that extends QSAR by not only considering data on the chemical structure of the compounds and on the compound-induced BA, but by simultaneously using transcriptomics data (gene expression). This approach enables, for example, the detection of chemical substructures that are associated with BA, while at the same time a gene set is correlated with both these substructures and the BA. Although causal associations cannot be formally concluded, these associations may suggest that the compounds act on the BA through a particular genomic pathway.

1. Introduction
In the pharmaceutical industry, efficiency and effectiveness of R&D has been declining over the past years, resulting in fewer new drugs reaching the market. A major cause is the failure of new drug candidates in late-stage development, due to safety issues or previously undiscovered side effects. This is why the choice of which compounds to advance through early phases, particularly during lead optimization, is becoming fundamental. It is believed that relevant biological data on the various desired and undesired compound effects need to be acquired and integrated in the early stages of the research process. For example, compound-induced perturbations in transcriptomics networks have been used to understand the biology and mechanisms of action (Feng et al., 2009; Searfoss et al., 2005) and their utility in helping the decision-making in compound selection has been evaluated in Hochreiter et al. (2015). Instead of looking at one or two data sources at a time, data-integration methods aim at using all available data simultaneously so as to more efficiently extract the relevant information required for the decision-making.

Data integration is becoming an important process in many sciences that can benefit from the simultaneous analysis of several datasets. In this article, we focus on data in the lead optimization step in early drug development. In this stage, chemists typically have synthesized 30–90 molecules...
(compounds) that belong to a limited number of chemotypes and hence show often only small variations in some of their substructures. Chemists face here the problem of identifying substructures that are (i) essential for the desired on-target effects and (ii) responsible for the undesired off-target effects. A specifically developed bioassay is used to measure the effect (potency) of the compounds on the target identified for the disease under study. Other bioassays are used to measure off-target effects, which may be predictive for side effects. This information can guide them in further optimizing the chemical structure (i.e., lead optimization).

QSAR (Quantitative Structure Activity Relationship) is a methodology aiming at the identification of chemical substructures that are related to a particular activity (Nikolova and Jaworska, 2004). In the present context, QSAR requires two datasets: bioactivity (BA) readouts and chemical descriptors for all compounds. We will work within a new paradigm that extends QSAR by including a third data source: gene expression (GE) microarray measurements, as it is believed that many drugs may act on the target indirectly by affecting the gene transcription process (Hochreiter et al., 2015; Feng et al., 2009; Searfoss et al., 2005). These microarray experiments are performed on cell lines relevant for the disease under study. The method presented in this article is only one of several Quantitative Structure Transcriptional Activity Relationship (QSTAR) methods that have been proposed for an integrated analysis of transcriptomics, bioassay and chemometrics data. QSTAR stands for Quantitative Structure Transcriptomics Activity Relationship, a term coined by the QSTAR consortium (www.qstar-consortium.org).

For the case study presented later on, we have data available from 35 compounds that were synthesized aiming at the discovery (lead optimization) of a drug inhibiting the epidermal growth factor receptor (EGFR) protein, which is overexpressed in many types of human cancer (Raymond et al., 2000). After I/NI filtering (Talloen et al., 2007), GE values of 3595 genes were retained for each compound. The information on the chemical structure of the compounds was quantified by means of the Extended-Connectivity Fingerprint (ECFP) algorithm (Rogers & Hahn, 2010). This encoding stems from a two-dimensional representation of the molecule and provides binary descriptors of the presence/absence of the molecule substructures. These substructure indicators are henceforward referred to as Fingerprint Features (FFs). To address tautomeric conversion, a phenomenon which affects the encoding, compounds were normalized prior to the encoding procedure (Sitzmann et al., 2010). The on-target BA is measured through a dose–response bioassay, resulting in an IC50 value for each compound. Although the compounds were screened with many more (off-target) bioassays, our methods deal with one bioassay at a time.

In the next few paragraphs we discuss some data analysis methods that may be used for the purpose of lead optimization. However, none of these methods are suitable for our goals, either because the methods involve meaningless linear combinations of the binary FFs (methods: principal component analysis (PCA), spectral map, and canonical correlation analysis (CCA) or because they do not allow to incorporate the FF data in the analysis of the other two data sources (methods: supervised PCA and integrative factor analysis (iFad)). Details follow.

PCA is a well-known method for dimension reduction and data exploration of multivariate datasets (Joliffe, 2002). It is directly applicable to the GE matrix, but the principal components resulting from the PCA do not necessarily relate well to the BA outcome and the chemical substructure indicators. The spectral map Analysis, which is basically a PCA on the double-centred GE matrix (Wouters et al., 2003), is therefore only recommended to be used in an exploratory data analysis phase. Moreover, the application of PCA to the FF data is meaningless, because linear combinations of the binary presence/absence indicators have no interpretation. Such linear combinations also appear in CCA, which is designed to integrate two multivariate datasets (see for example Härdle and Simar, 2007). CCA looks for linear combinations of the variables of the first and the second dataset that show maximal correlation.

Bair and Hastie (2004) and Bair et al. (2006) proposed supervised PCA, which performs data integration of a multivariate dataset with a single response variable. Applied to our setting, the method consists in: (1) filtering genes by removing those with a small correlation with the BA
response and (2) performing a PCA on the remaining genes. In this way, the principal components contain only genes showing a large correlation with the BA. The method, however, does not allow the simultaneous incorporation of FF data.

Recently, Ma and Zhao (2012) introduced iFad for integrating GE data into the drug-discovery process. iFad is basically a Bayesian sparse factor analysis method aiming to construct sparse linear combinations of GEs that show a large correlation with BA outcomes. The method allows for the use of biological pathway information (e.g., KEGG pathways) to direct the solution so that the pathway explains the correlation between the two data sources. Although we consider this a valuable method in drug discovery, the method does not serve our purposes as it does not integrate the FF data.

The method that we propose is tailored to our objectives, but it shares some of the properties of the classical methods. It aims to find linear combinations of GEs, resulting in \( \text{components} \) that show large association with \( \text{both} \) the BA outcome and a single FF. In particular, we define the \( \text{bicorrelation} \) as the sum of the squared associations between the component and the bioassay outcome, and between the component and the FF. Given the resemblance with PCA and the central role of the bicorrelation, the new method is referred to as \textit{principal bicorrelation analysis} (PBA). So far, the method works on a single FF, because it is extremely difficult to interpret the effect of multiple FFs without a clear understanding of the chemistry. Instead, we propose to apply the PBA to each FF separately, and to use a nonparametric permutation procedure for selecting the important ones. In doing so, the method allows for controlling the false discovery rate (FDR) at a desired level. A further improvement is achieved by enforcing sparseness in the components, for an easier interpretation and a more direct identification of the important genes. Since the core of the PBA is a singular value decomposition (SVD), we make use of sparse SVD methods as proposed by Witten and Tibshirani (2009a,b).

In Section 2, the PBA method is described along with its typical outputs and their interpretation. In Section 3, the method’s performance is empirically investigated by means of a simulation study, while in Section 4, it is demonstrated on a real case study. We conclude with a discussion in Section 5.

2. Principal bicorrelation analysis

Let \( n \) represent the number of tested compounds, \( n_z \) the number of FFs, \( n_x \) the number of genes, and \( n_y \) the number of BAs. Let \( Z \) denote the \( n \times n_z \) FF matrix, \( X \) the \( n \times n_x \) GE matrix, and \( Y \) the \( n \times n_y \) bioassay readouts. In this article, we focus on one single bioassay, i.e. \( n_y = 1 \). We denote by \( Z_j \) the \( j \)th column of \( Z \). The method is outlined for the analysis of one FF \( Z_j \) at a time, but in the discussion (Section 5), we describe how the results may be combined. We also assume that the GE matrix \( X \) and the FF matrix \( Z \) have centred columns.

2.1. The loadings and the scores

A \textit{gene set score} (GSS) is a univariate variable defined as a linear combination of GEs, i.e. \( \text{GSS} = Xw \), where \( w \) is the \( n_x \)-vector containing the loadings. Since we aim to identify genes that are simultaneously strongly correlated with the FF and with the BA, we propose to maximize:

\[
\text{BC}(w) = R^2 \{ Z_j, Xw \} + R^2 \{ Y, Xw \}
\]  

(1)

where \( R\{\cdot\} \) is a generic association measure and \( w \) is subject to the constraint \( w^T w = \|w\|^2 = 1 \). The constraint is the same as in the PCA, and although here it is rather arbitrary, some constraint is required for defining \( w \) uniquely. We refer to Equation (1) as the \textit{bicorrelation} between the GSS and the two other data sources.
The association measure in the definition may be chosen according to the nature of the data at hand; for example in the simulation study we employed the usual correlation whereas in the case study the covariance was used as $R \{ \cdot \}$ (more details on that can be found in the case-study section).

Another possible association measure for a binary FF $Z_j$ may be the Spearman’s rank correlation, which is algebraically equivalent to the Wilcoxon’s rank sum test statistic rescaled to the $[-1, 1]$ interval.

In Appendix A, we demonstrate that the loadings that maximize the bicorrelation, say $\hat{w}$, are given by the first right singular vector of the matrix $M = (R \{ Z_j , X \}, R \{ Y , X \})^T$. The resulting loadings vector is denoted by $\hat{w}_j$ to stress the dependence on FF $j$. In order to gain interpretability of the gene set, we used the sparse SVD method of Witten and Tibshirani (2009b); see section 2.2 for more details. In this way, only a limited number of genes are involved in each gene set. The genes with nonzero loadings are referred to as the gene set (GS).

We borrowed the terms “loadings” and “scores” from PCA because, similar to PCA, PBA will result in more than one vector of loadings. Although here these vectors cannot be orthogonal like in PCA, a certain kind of separation is attained by imposing orthogonality on the $U$ matrix of the SVD: $M = USV^T$ (see Witten and Tibshirani (2009b) for details). The SVD solutions are ordered conventionally by decreasing singular values. We use the notation $\hat{w}_{jk}$ to denote the $k$th solution in this sequence, with $j$ still referring to the $j$th FF. Similarly, $GSS_{jk} = X_{\hat{w}_{jk}}$ gives the $k$th GSS vector for FF $j$. Note also that only two solutions exist (i.e., $k = 1, 2$) for $M$ has rank 2.

The method described in this section can be repeated for all FFs. Hence, for each FF $Z_j$, we have: (1) a $n \times 2$ matrix of gene loadings, (2) the corresponding maximized bicorrelations, and (3) an $n \times 2$ matrix of GSSs.

### 2.2. Sparse loadings

Appendix A explains how a sparse SVD may be used for obtaining a sparse GSS. The level of sparseness is controlled by a tuning parameter $\gamma$. A bad choice of $\gamma$ may result in a poor or suboptimal specification of the GSS, which may eventually lead to failure to identify important FFs and genes.

We propose to use cross-validation to select the best $\gamma$ from a set of candidate values, say $\gamma_1, \ldots, \gamma_m$. More specifically, we make use of a $K$-fold cross-validation in which, for each candidate $\gamma_i$ value, we (i) split the set of compounds into $K$ parts; (ii) perform PBA leaving out one of these parts; (iii) compute the bicorrelation on the data left out; (iv) repeat (ii) and (iii) for each of the $K$ parts so that all data have been used; (v) compute $BC(\hat{w}(\gamma_i))$, which is now the average bicorrelation obtained with $\gamma_i$; (vi) select the $\gamma_i$ for which $BC(\hat{w}(\gamma_i))$ is maximal.

### 2.3. Selecting important FFs

In this section, we outline a procedure for selecting important FFs while controlling the FDR at a desired level. A FF that is selected in this way is said to be a called FF.

An important FF is one that is highly correlated with a GSS which is in turn highly correlated with the bioassay outcome. Inspired by (1), we propose to make a scatter plot of $R^2 \{ Z_j, X_{\hat{w}_1} \}$ against $R^2 \{ Y, X_{\hat{w}_1} \} (j = 1, 2, \ldots, n_z)$. This graph shows one dot for each FF and is referred to as the bicorrelation graph. Figure 1 (top panel) shows an example (details of the data will be given in Section 4). The reference line corresponds to a constant bicorrelation; we expect to find FFs with a large maximized bicorrelation and a good balance between its two composing elements close to the upper right corner. The lower panel of Fig. 1 shows an alternative representation that uses the non-squared association measures $R \{ Z_j , X_{\hat{w}_1} \}$ and $R \{ Y , X_{\hat{w}_1} \}$. The constant bicorrelation line has here
a circular shape and the more interesting FFs are expected to be located further outside of the contour line.

Although informative and helpful for selecting important FFs, bicorrelation graphs do not protect us from false discoveries, particularly when many FFs are involved. Hereafter, we describe two simple methods for controlling the FDR.

The first method consists of (1) calculating $p$-values for each individual FF and (2) transforming them to $q$-values by using, for example, the method of (Benjamini and Hochberg, 1995). All FFs with a $q$-value smaller than the desired FDR level are called. Here, the $p$-value for a single FF refers to the null hypothesis: $H_{0j}: Z_j \perp (Y, X_w)$, i.e. the FF $Z_j$ is jointly independent from GSS$_j$ and BA outcome $Y$ and the alternative hypothesis $H_{1j}: \text{BC}(w_j) > 0$.

**Figure 1.** Bicorrelation plots for the EGFR project. Upper panel: squared covariance version, lower panel: non-squared covariance version. Each dot corresponds to a FF. The 25 called FFs (FDR = 5%) have their code printed. The two FFs that were selected as interesting are indicated with an arrow. Points on the dash-dotted line have constant bicorrelation. Light (dark) dots indicate FFs that possess a positive (negative) correlation with the BA outcome. The size of the dot is proportional to the absolute value of that correlation as well as the brightness, i.e. the lighter the closer to +1, the darker the closer to −1.
Note that the null hypothesis implies the invariance of the estimated bicorrelation distribution under permutations of the compounds in the FF matrix $Z$.

We propose to keep the tuning parameter $\gamma$ for the sparseness fixed throughout the permutations, although it is actually selected in a data-driven fashion. This should be considered as a simplification to reduce the computational cost.

A second method shows similarities with the SAM method (Tusher et al., 2001), where the maximized bicorrelation acts as the test statistic. Whereas the original SAM algorithm looks at the permutation null distributions of the ordered test statistics, our method ignores the order and focuses on the marginal permutation null distribution. Again, we keep the sparseness tuning parameter fixed. Appendix B gives a detailed description of the algorithm.

### 2.4. Interpreting the results of a single FF

For the called FF, the corresponding GSS may now be investigated in detail so as to identify the important genes. For a sparse solution, the gene loadings can be easily visualized; see Table 1 for the data example of Section 4. Note that the sign of each gene loading matches the gene–BA correlation such that a positive (negative) value corresponds to a gene for which an upregulation (down-regulation) contributes to increase (decrease) in the BA outcome.

Figure 2 is very helpful in understanding the associations between the three data sources. The top panel shows a scatter plot of the bioassay outcomes against the GSS on the first dimension (i.e. based on $\tilde{w}_{j1}$), where each dot represents a compound. This graph highlights a strong positive correlation between GSS and BA. The bottom panel shows boxplots of the GSS: one for the compounds with the FF present and one for compounds without the FF. From this graph, we see that the FF discriminates well between lower and higher GSS on the selected gene set, and therefore it also discriminates between low and high BA compounds. This illustrates that the PBA succeeded in selecting a FF and identifying a gene module that are strongly associated, while both also strongly relate to the bioassay outcome. Moreover, the graph in the top panel allows for the selection of compounds that are interesting for further research. For example, the compounds in the top-right corner have a high score on the bioassay (good on-target effect) and a high GSS, whereas from the boxplots we deduce that the selected FF induces inactivity, hence a FF that should be avoided when possible. A more in-depth interpretation will be given in Section 4.

### 3. Simulation study

In this section, we present the results of a simulation study aiming to empirically demonstrate the correct behaviour of the method. In particular, we will demonstrate that PBA succeeds in: (1) identifying the important gene sets; (2) selecting the FFs that are correlated with a gene set that is in turn also highly correlated with the BA outcome; (3) controlling the FDR at the desired level. Note that for this simulation study we chose the usual Pearson’s correlation as the association measure $R \{ \cdot \}$.

### Table 1. Genes and their loadings in GSS$_1$ as identified by the potent FF.

| Gene_Symbol | Weight_Dim.1 | Gene_Name                                      |
|-------------|--------------|-----------------------------------------------|
| 1 FOSL1     | −0.89        | FOS-like antigen 1                            |
| 2 FGFBP1    | −0.39        | Fibroblast growth factor-binding protein 1    |
| 3 SEPP1     | 0.15         | Selenoprotein P, plasma, 1                    |
| 4 SCGB2A1   | 0.13         | Secretoglobin, family 2A, member 1            |
| 5 RRM2      | −0.09        | Ribonucleotide reductase M2                   |
| 6 SLCO4A1   | −0.08        | Solute carrier organic anion transporter family, member 4A1 |
| 7 SH2B3     | −0.03        | SH2B adaptor protein 3                        |
| 8 PHLDA1    | −0.03        | Pleckstrin homology-like domain, family A, member 1 |
| 9 CDC6      | −0.01        | Cell division cycle 6 homolog (S. cerevisiae) |
We will describe the data-generation process in three steps: (1) generation of gene sets, which involves the specification of loadings and the simulation of GE matrices; (2) generation of BA outcomes, possibly correlated with a gene set; (3) generation of FFs, some of which may be correlated with gene sets and/or BA. We always include \( n = 90 \) compounds, \( n_x = 100 \) genes, \( n_z = 80 \) FFs, and a single BA. In total, 100 Monte Carlo replications will be used and 250 permutations in each Monte Carlo run for the FDR control.

The FFs can be divided into four groups that vary in the strengths of the associations between the three data sources. In particular, we will include a setting in which a gene set is correlated with both the BA and a few FFs: our method is designed to identify specifically these FFs. Two other subsets of FFs are only correlated with a gene set or with the BA. The fourth group shows no dependence with the other data sources. Referring to the dependence structures shown in Fig. 3, we name these FF subsets \((X, Y, Z), (X, Z), \) and \((X, Y)\).

![Figure 2. Expression level graph for genes in GSS1 selected by the potent FF in the EGFR project. Gene symbols are listed on the left and compound names at the bottom. Compounds highlighted with a black circle have the FF present. The final two compounds are the reference compounds (17195907 and 19564740). The least potent (left) and the most potent (right) compounds are highlighted with a gray rectangle. Gene profiles are ordered according to their loadings value from high (top) to low (bottom) and the dashed horizontal reference line marks the change of sign in the loadings. In the right panel, violin plots show the distribution of the expression levels for each gene. The vertical reference line corresponds to the overall average (over all genes). The values at the right from the violin plots show the average expression values.](image-url)
Although no genuine competitor to our method is known to us, we included two lasso regression methods (Friedman et al., 2010). They serve as a kind of benchmark. Given that lasso regression does not allow for the integration of all three data sources, we used this method to study the relationships FF–GE and GE–BA separately. In particular, we regressed the BA onto the GE outcomes, and we performed logistic lasso regression for each individual FF with GE as predictors. It is beyond the scope of this article to develop a method to control the FDR for the lasso regressions.

3.1. Data generation

GEs are simulated for all compounds and for all genes, in which the first three sets of five genes are simulated by a zero-mean multivariate normal distribution with the covariance matrix:

$$\begin{pmatrix}
1 & 0.7 & 1 \\
0.7 & 0.5 & 1 \\
-0.7 & -0.7 & -0.7 & 1 \\
-0.7 & -0.7 & -0.7 & 0.7 & 1
\end{pmatrix}$$

while the other 85 genes are generated as i.i.d. $\mathcal{N}(0, 1)$.

Two gene sets are generated: the first only depends on the first five genes, $GSS^*_1 = -0.4X_1 + 0.5X_2 - 0.4X_3 - 0.5X_4 + 0.5X_5$, and the second on the next five genes, $GSS^*_2 = -0.4X_6 + 0.6X_7 - 0.4X_8 + 0.5X_9 - 0.6X_{10}$. We use the * notation to distinguish the true GSS (GSS*) from the GSS that will be extracted with the PBA method. The genes that are not involved in GSS$^*_i$ or GSS$^*_2$ are referred to as noisy genes. The BA outcomes are simulated as $Y = GSS^*_i + U_1$, where $U_1$ i.i.d. $\mathcal{U}[-1, +1]$.

Eighty FFs are simulated, among which the first three sets of 10 FFs have a specific association structure, shown in Fig. 3. The first 10 FFs are simulated to be highly correlated with both GSS$^*_i$ and Y. In particular, $Z_j = \mathbb{I}[GSS^*_1 > U_2]$ with $U_2$ i.i.d. $\mathcal{U}[-0.7, +0.7], j = 1, \ldots, 10$.

The next 10 FFs are simulated to be highly correlated with GSS$^*_2$, but not correlated with the BA: $Z_j = \mathbb{I}[GSS^*_2 > U_3]$, with $U_3$ i.i.d. $\mathcal{U}[-0.4, +0.4], j = 11, \ldots, 20$.

A third type of FFs is constructed to show a strong correlation with the BA while being almost uncorrelated with the gene sets. This is accomplished in two steps: (1) generate $Z'_j = \mathbb{I}[Y > U_4]$ with $U_4$ i.i.d. $\mathcal{U}[-0.7, +0.7], j = 21, \ldots, 30$; (2) regress each $Z'_j$ onto GSS$^*_1$ and extract the residuals, say $R'_j$. These residuals are by construction uncorrelated with regressor GSS$^*_1$, but they are no longer 0/1 valued; (3) set $Z'_j = \mathbb{I}[R'_j > -0.07]$ (the threshold of $-0.07$ results in a good balance between zeroes and ones and also in suitable values for the correlations of interest). A residual correlation between the binary FF and the GSS$^*_1$ is still present, because a zero correlation is not attainable for a binary FF and with a large correlation between BA and GSS$^*_1$.
The remaining 50 FFs (fourth group) for all compounds are independently generated as Bernoulli with probability 0.3.

Supplementary Material includes a table showing the correlations between generated datasets, averaged over the 100 simulations.

### 3.2. Simulation results

The results of the simulation study are based on 100 Monte Carlo runs. If the PBA method works properly, then it must detect the FFs and genes involved in the \((X, Y, Z)\) dependence structure (Fig. 3): FFs 1–10 and genes 1–5. The other FFs are not of interest because they do not show the \((X, Y, Z)\) dependence structure. Since we also want to study from what FF group the PBA method selects FFs, we require both enough true and false positives. For this reason, the FDR is set to 20%. The Pearson correlation is chosen as the association measure \(R_{\cdot \cdot}^\circ\). This allows us to focus the study on how the associations between and within the three data sources affect the performance of the PBA method, without being distracted by differences in variances of GE levels. In the case study of Section 4.1, we elaborate on this issue.

Figure 4 gives a summary of the simulation results with emphasis on the probability to detect the important FFs and the genes involved in the association. The figure shows the number of times 0, 1, 2, …, or 5 (≥ 5 in case of noisy genes) of the genes in the true GSS1*, GSS2*, or noisy genes were detected, counting only genes in GSSs for the FFs called at the FDR level of 20%. Table 2 shows for each of the four FF groups the number of times (and percentages) a FF was called at the FDR level of 20%. In the supplementary material, results are reported at a more detailed level.

![Figure 4](image)

**Figure 4.** Graphical representation of the simulation results, showing the frequencies of the number of genes from GSS1 (left), GSS2 (middle), and the noisy genes (right), which were identified by the PBA method (i.e., nonzero loading in GSS for FFs called at the FDR level of 20%). Frequencies are shown for each group of FFs separately.

| FF group | Count | First | Second | Third | Fourth |
|----------|-------|-------|--------|-------|--------|
| FF calls | %     |       |        |       |        |
| 1st−group| 981   | 56.1  | 4.2    | 25.1  | 14.6   |
| 2nd−group| 73    | 981   | 56.1  | 4.2    | 25.1  |
| 3rd−group| 438   | 981   | 56.1  | 4.2    | 25.1  |
| 4th−group| 256   | 981   | 56.1  | 4.2    | 25.1  |

Numbers are summed for FFs within the same group. The percentages refer to the total number of calls.

From these outputs, we see that the FFs from group 1 show the largest probability of being called (56%), and when they are called, at least one gene from GSS/C3 is selected with a probability of 97.6%, while this probability drops to about 40% for genes from GSS/C3. Note also that when a FF of the third group is called, it will most likely (97% probability) select genes from GSS/C3. This is a consequence of the nonzero correlation between GSS/C3 and the FFs (Fig. 3), making this scenario resemble an attenuated (X, Y, Z) dependence structure. Figure 5 shows the average gene loadings in $\hat{w}_1$ averaged over the simulation runs for which a FF of the first group was called at the FDR level of 20%. For each simulation run, only loadings of FFs called at the FDR level of 20% were considered. The “+” symbols represent the average loadings, and the “×” indicate the true loadings with which GSS/C3 and GSS were built. Note that results are shown only for the first 15 genes: first five are involved in GSS/C3; second five in GSS/C3; and last five are just five noisy genes. Our method aims at identifying the genes in GSS/C3, because these genes are associated with the BA and with FFs 1–10 that are in turn associated with the BA.

Figure 5. Gene loadings in $\hat{w}_1$ averaged over the simulation runs for which a FF of the first group was called at the FDR level of 20%. For each simulation run, only loadings of FFs called at the FDR level of 20% were considered. The “+” symbols represent the average loadings, and the “×” indicate the true loadings with which GSS/C3 and GSS were built. Note that results are shown only for the first 15 genes: first five are involved in GSS/C3; second five in GSS/C3; and last five are just five noisy genes. Our method aims at identifying the genes in GSS/C3, because these genes are associated with the BA and with FFs 1–10 that are in turn associated with the BA.

From these outputs, we see that the FFs from group 1 show the largest probability of being called (56%), and when they are called, at least one gene from GSS/C3 is selected with a probability of 97.6%, while this probability drops to about 40% for genes from GSS/C3. Note also that when a FF of the third group is called, it will most likely (97% probability) select genes from GSS/C3. This is a consequence of the nonzero correlation between GSS/C3 and the FFs (Fig. 3), making this scenario resemble an attenuated (X, Y, Z) dependence structure. Figure 5 shows the average gene loadings in $\hat{w}_1$ for the first group of FFs, averaged over the simulation runs for which the FFs were called, and over the 10 FFs. The graph demonstrates that for genes 1, 3, and 5, the PBA method succeeded quite well in estimating the true loadings, whereas the average loadings of genes 2 and 4 are shrunken toward zero. This phenomenon is caused by the lasso penalty, for which it is generally known that (1) it does often not select all important correlated features (the five genes in GSS/C3 are correlated) and (2) it overshrinks the parameter estimators. The graph also demonstrates that the average loadings of the other genes are close to zero. Thus, although Fig. 5 shows that FFs called from the first group sometimes select wrong genes, their loadings are on average close to zero.

The results of the lasso regressions can be summarized as follows. The BA–GE lasso regression succeeds in finding the genes in GSS/C3. The GE–FF lasso regression analyses correctly discover that FFs in the first and second FF group are related to the GE involved in GSS/C3 and GSS/C3, but the FFs of the second FF group are not related to the BA. The two lasso regressions do not allow for a formal integration of their results and hence the simultaneous identification of important genes and FFs is subject to error without a proper FDR control. More details and the results are reported in the Supplementary Material.
The probabilities of calling FFs of the other groups are controlled at the FDR level of 20% as it can be seen in Table 3, which presents the FDR actually attained in the simulation study. In particular, for each simulation run, the FDR is computed as the number of called FFs not belonging to the first group, divided by the total number of called FFs, and the table shows summary statistics of this distribution. From this table, we conclude that the mean and median FDR are very close to the nominal level of 0.20.

### 4. Case study

In this section, we present an application of the PBA method to a real oncology project from a pharmaceutical company. The aim of the project was to discover compounds that inhibit the protein EGFR. This protein is over-expressed in certain types of human carcinoma, leading to an anti-apoptotic Ras/Erk signalling cascade and eventually to uncontrolled cell proliferation (Woodburn, 1999).

A set of genes was provided by scientists who selected the genes from the relevant literature related to the specific pathways of the cascade, namely Ras/ERK/MAPK. For example, the protein c-Fos is regulated by this pathway. c-Fos is a component of the dimeric transcription factor activator protein-1 (Ap-1), which is composed mainly of Fos (c-Fos, FosB, Fra-1, and Fra-2) and Jun proteins (c-Jun, JunB, and JunD). Among these, Fra-1 is encoded by the FOSL1 gene, which is present in the data. This explains why FOSL1 (FOS-like antigen 1) was considered to be a very important gene within the experiment.

The compounds synthesized in this project were macrocycles derived from the drugs gefitinib and erlotinib (Raymond et al., 2000). 35 of these compounds were profiled against EGF-stimulated cells for which cell growth (BA) and GEis were measured. All 35 compounds are known to induce activity. 3595 genes survived I/NI-call filtering (Talloen et al., 2007) and were used in the data analysis. As for the chemical descriptors, 101 unique FFs were retained from a total of 16,725. In particular, the rules for keeping a specific FF are: (i) it has to be present either in more than 1 compound (singleton is not useful) or absent in maximum 33 out of 35 compounds; (ii) if two or more FFs are present in exactly the same compounds, only one of those is kept.

Here as association measure \( R \{ \cdot \} \), we employed the usual covariance in place of the correlation of the simulation study. In fact, the variance of a gene is important in microarray measurements as it basically contains the information about how much that gene is affected by the different samples: if the variance of a gene is really small, the gene is not interesting, especially in the lead optimization stage. This information would be lost upon using the correlation because all genes would be individually scaled.

#### 4.1. Principal bicorrelation analysis

Genes differ in their expression level variances and genes showing high variance are considered more informative than stable (low-variance) genes. For this reason, we have chosen to use here
the covariance as the association measure $\mathcal{R}\{\cdot\}$. A FDR of 5% was chosen as this appeared to give a list of 25 called FFs, which was considered rich enough by the chemists and biologists working on the project. From the list of called FFs and the bicorrelation graphs (Fig. 1), and with the help of the chemists, we selected two FFs for further investigation. In particular, one seems to induce high activity because it is present almost exclusively in compounds with high BA, while the other is present almost exclusively in compounds with lower BA readouts. These FFs, which we refer to as the potent and the less potent FFs, will be analyzed in more detail in the following sections.

4.1.1. Potent fingerprint feature

Figure 6 shows a strong positive correlation between GSS$_1$ and the BA. Compounds with a high GSS$_1$ are more potent and grouped closely to the reference compounds, gefitinib and erlotinib. The boxplot of Fig. 6 shows that the FF is present in most of these highly active macrocycles. In the next step, the genes forming GSS$_1$ were explored so as to investigate the biology behind the association (Table 1). Note that, due to the induced sparseness, only few genes are retrieved.

Figure 2 shows that all selected genes have similar expression profiles, especially in the most potent compounds. We also learn that, for all selected genes, most of the compounds with the FF present have GEs similar to those of the reference compounds. This may suggest that they act along the same biological pathway.

Although the PBA only selected few genes, more genes can be found in the data by looking for highly correlated genes and/or genes with a similar expression profile on the compounds of interest. We consider this post-analysis step important, because it is generally known that the induced sparseness causes the selection of only one or at most a few of a set of highly correlated genes. Nonetheless, it is worth noting that FOSL1 has been selected by the FF together with other genes with similar profiles.

Our analysis thus resulted in the discovery of a FF that, when present, causes a high pIC50 in the BA of interest and an associated large GSS value. Through the loadings, we know that this coincides

![Figure 6](image-url)
with a downregulation of several genes, some of which are known to play a role in the targeted ERK pathway.

### 4.1.2. Less potent fingerprint feature

For the less potent FF, a strong correlation is observed between the bioassay and GSS$_1$ (Fig. 7). The PBA now selected the same genes as the potent FF with a slightly changed ranking among them (Table 4).

From Fig. 7, we also see that most of the less potent compounds have the FF present, which corresponds to compounds with negative GSS$_1$ scores. By looking at the gene loadings (Table 4), we deduce that negative GSS$_1$ scores are associated with downregulation of genes FOSL1 and FGFBP1 and upregulation of SEPP1 among others. Figure 8 shows the same gene profiles as the potent FF, as both potent and less potent FFs selected the same set of genes, but here the black-bordered dots highlight compounds where the less potent FF is present.

Since the role of FOSL1 in the ERK pathway is known, we conclude that compounds with the FF present do not show the desired EGRF inhibition because the ERK pathway is not suppressed. The

---

**Table 4.** Genes and their loadings in GSS$_1$ as identified by the less potent FF.

| Gene_Symbol | Weight_Dim.1 | Gene_Name |
|-------------|--------------|-----------|
| 1 FOSL1     | -0.90        | FOS-like antigen 1 |
| 2 FGFBP1    | -0.38        | Fibroblast growth factor binding protein 1 |
| 3 SEPP1     | 0.15         | Selenoprotein P, plasma, 1 |
| 4 SCGB2A1   | 0.13         | Secretoglobin, family 2A, member 1 |
| 5 SLCO4A1   | -0.08        | Solute carrier organic anion transporter family, member 4A1 |
| 6 RRM2      | -0.08        | Ribonucleotide reductase M2 |
| 7 SH2B3     | -0.05        | SH2B adaptor protein 3 |
| 8 PHLDA1    | -0.03        | Pleckstrin homology-like domain, family A, member 1 |
| 9 CDC6      | -0.00        | Cell division cycle 6 homolog (S. cerevisiae) |
PBA method thus succeeded in identifying a FF that is less desired so that chemists may be warned not to synthesize molecules with this FF.

4.2. Conclusions

Figure 9 shows the chemical structure of the least potent (left) and the most potent (right) compounds while highlighting the discovered FFs. The difference between the potent and less potent FFs is the group in ortho of the aniline which is a CH$_2$ or an oxygen, respectively. Since all compounds with relatively smaller activity have the less potent FF present, we can conclude that the oxygen is the major cause. The potent FF, however, is not present in all highly active compounds, indicating that only part of the solution has been found.

Among the other FFs called by the PBA, we discovered FFs that refer to the same substructures to which the selected potent FF refers. This is a consequence of the
hierarchical nature of the ECFP-encoding algorithm that generates the FFs. For instance, the dotted (thinner) circle in Fig. 9 highlights the first FF in the PBA ranking, but given the bioisosterism of Cl and Br, it points also to the difference in the oxygen or CH\textsubscript{2} attached linker.

The final conclusion of the case study is that inhibition of the EGFR downregulates certain genes of the ERK pathway and that compounds having an oxygen in ortho of the aniline, in the series of compounds studied, have lower EGFR inhibitory activity.

5. Discussion

Lead selection and optimization in early drug development are time-consuming and very complex processes that heavily rely on the expertise and insights of the chemists. Nowadays, however, high throughput technologies can quickly generate large amounts of data to support the decision-making process and to direct the next steps to be taken in lead optimization. Although many data-mining and statistical methods can be directly applied to the supporting data, more informative analyses would become possible if more specific statistical methods for data integration of multiple data sources were available.

In this article, we developed a method for data exploration and for genes and FF discovery, using three data sources simultaneously: bioassay outcomes (BA), GEs, and chemical substructure indicators (FFs). The method, which is named PBA, resembles PCA in the sense that linear combinations of variables (GEs) are constructed, but instead of aiming to maximize the variance of the component, our method maximizes its bicorrelation with the BA and a single FF, resulting in a component showing a large association with both of them. This allows the identification of genes that play a role in the pathway from the compound (chemistry) to the BA (bioassay). Moreover, by imposing sparseness on the loadings of the components, their interpretation in terms of only a limited number of genes becomes feasible. The method is applied to a single FF at-a-time, but by repeating the process for all FFs we are able to select only the most important substructures while controlling the FDR at a desired level by means of a permutation testing procedure.

In a simulation study, we have empirically demonstrated that PBA succeeds in identifying the important genes and chemical substructures while controlling the FDR quite well. Moreover, the PBA method has been applied to a dataset from a real project, for which it succeeded not only in identifying biologically relevant genes which are associated with the bioassay of interest but also enabled the discovery of important chemical features which might drive this association.

We implemented PBA as an R-package (R Core Team, 2013), named PBA, which is available from the website of the QSTAR consortium: www.qstar-consortium.org. PBA is a method that
belongs to the QSTAR paradigm, which extends QSAR by adding transcriptomics data as a third data source to help the understanding of the relationship between the chemical structure and its biological effect.

PBA should be considered as one of the methods in the QSTAR toolbox. We therefore advise the additional use of other methods for data mining and data exploration, and the consultation of publicly available data sources. For example, the called FFs are not necessarily easily interpretable or can be used directly by the chemists to construct new compounds with better BA characteristics. Thus, if the FF is in an area of the molecule that cannot be changed without compromising the docking properties, this FF cannot be altered easily by the chemists. Moreover, most of the available chemical encodings are based on two-dimensional models that do not account for the spatial structure of the molecule.

However, the selected FFs from the PBA method in combination with the public ChEMBL database can be used for the identification of interesting chemical analogues (see Gaulton et al., 2012). A chemical analogue is a molecular compound that is similar to an existing molecule, but with some of its substructures changed so as to alter some of its effects on BA or GE. Similarly, since our method does not only select FFs, but also identifies interesting compounds (see Section 2.4), the ChEMBL database can be screened for compounds that show a large similarity with these selected compounds. We refer to Bender and Glen (2004) and Nikolova and Jaworska (2004) for discussions on molecular similarity measures.

If, before the starting the data analysis, one has decided to use the results to look for analogues in the ChEMBL database, the interpretation of the FF is no longer very important. This opens doors to use less interpretable chemical encodings as a Z matrix. One may, for example, first transform the \( n_z \) columns of \( Z \) to the scores on all \( n_z \) principal components, and then use the first few as input. When only similarities between the compounds are available, another approach can be to transform the similarities to a Z matrix by using a multidimensional scaling technique.

The identified gene sets may also be used for searching analogues in the Connectivity Map (cMAP) project (Lamb et al., 2006), which provides a public dataset with GE profiles from many chemical compounds in many different cell lines. Selecting compounds in cMAP that show a large similarity with the gene sets that are identified with our method may be a first step toward discovering new chemical analogues.

The PBA method has been described as being applied to each FF separately, resulting in \( n_z \) sparse loading martices \( \hat{w}_1 \) of dimension \( n_x \times 2 \). It may be expected, however, that some FFs act on the bioassay via the same gene set, or through different sets of genes that have similar impacts on the activity. A simple way to group FFs into genomically similar groups is by applying a cluster analysis to the \( n_z \times n_z \) distance matrix \( J_{n_z} = (G^T G)^2 \), where \( (G^T G)_{ij} = \hat{w}_i^T X^T X \hat{w}_j \), \( J_{n_z} \) is the square matrix with all elements equal to 1, and the square operator is applied entry-wise. Hence the jth column of \( G \) contains \( X \hat{w}_{j1}, \) i.e., the GSS for FF \( Z_j \). In this way, the correlation between genes is taken into account, and two FFs that selected different gene sets may be still clustered together because of the high correlation between the two sets.

### 5.1. Possible extension

So far, the method looks for FFs that have an indirect effect on the BA through the gene transcription process but it is not directly concerned with the direct effect (hence ignoring the GE measurements as in QSAR). A possible extension of the method is the inclusion of a penalty in the definition of the bicorrelation, penalizing this direct effect. This adaptation will force the method to look for FFs that cannot be easily found within the QSAR framework, such as FFs which are not directly affecting the BA, but which affect a set of genes that in turn has a strong impact on the BA.
represent, respectively, the GE matrix, the FF matrix, and the BA vector. Let also $Z$ be the column of the matrix, i.e., the FF. Moreover, let $R$ be an association measure such that, if $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{m \times q}$ then $R \{ A, B \} = C \in \mathbb{R}^{p \times q}$. The estimated bicorrelation can then be written as

\[ R \{ Z, Y \} = C. \]

**Appendix A: Loadings of the gene modules**

Let the matrices $X$, $Z$, and $Y$ represent, respectively, the GE matrix, the FF matrix, and the BA vector. Let also $Z_j$ be the $j$th column of the matrix, i.e., the $j$th FF. Moreover, let $R \{ \cdot \}$ be an association measure such that, if $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{m \times q}$ then $R \{ A, B \} = C \in \mathbb{R}^{p \times q}$. The estimated bicorrelation can then be written as

\[ R \{ Z_j, Y \} = Cj. \]
The maximization problem may now be formulated as finding \( \mathbf{w} \) that maximizes:

\[
\mathbf{w}^T \mathbf{M}^T \mathbf{M} \mathbf{w} - \lambda (\mathbf{w}^T \mathbf{w} - 1),
\]

where \( \lambda \) is a Lagrange multiplier to enforce the restriction \( \mathbf{w}^T \mathbf{w} = 1 \). Upon equating the partial derivative of (A2) to zero, we find

\[
\frac{\partial}{\partial \mathbf{w}} \left( \mathbf{w}^T \mathbf{M}^T \mathbf{M} \mathbf{w} - \lambda (\mathbf{w}^T \mathbf{w} - 1) \right) = 0 \iff \mathbf{M}^T \mathbf{M} \mathbf{w} = \lambda \mathbf{w}.
\]

Hence, \( \mathbf{w} \) and \( \lambda \) are the first eigenvector and eigenvalue of the \( n_z \times n_z \) matrix \( \mathbf{M}^T \mathbf{M} \), or equivalently, \( \mathbf{w} \) is the first right singular vector of the SVD of \( \mathbf{M} \). Sparse loadings may be obtained by replacing the ordinary SVD of the matrix \( \mathbf{M} \) with the penalized matrix decomposition of Witten and Tibshirani (2009b) with an \( L_1 \)-norm (lasso) penalty on \( \mathbf{w} \). The level of sparseness is determined by a tuning parameter.

**Appendix B: Sam-like method for FDR control**

In this section, we describe the nonparametric algorithm we used for calling FFs while controlling the FDR. The procedure is based on \( N \) permutations.

1. Based on the observed dataset, the \( n_z \) maximized bicorrelations are denoted by \( B_1, \ldots, B_{n_z} \).
2. For each \( k = 1, \ldots, N \) (counterindex for permutations), do the following: randomly permute the rows of the FF matrix \( \mathbf{Z} \) and calculate for each FF the maximized bicorrelation, using the same sparseness tuning parameter \( \gamma \) as for the original dataset. Let \( B_1^{(k)}, \ldots, B_{n_z}^{(k)} \) denote these maximized bicorrelations.
3. Set the constant \( \Delta \) to some arbitrary value so that not too many FFs result in \( B_j > \Delta \). Let \( N_s \) denote this number of FFs, i.e., \( N_s \) is the number of selected FFs when \( \Delta \) is used as the threshold. In formula: \( N_s = \sum_j I(B_j > \Delta) \).
4. Let \( N_{f,k} \) denote the number of FFs for which \( B_j^{(k)} > \Delta \), in formula \( N_{f,k} = \sum_j I(B_j^{(k)} > \Delta) \), and let \( N_f \) denote the median of these \( N_{f,k}, k = 1, \ldots, N \). \( N_f \) is interpreted as an estimate of the number of false discoveries.
5. Compute \( \text{FDR}(\Delta) = N_f/N_s \), i.e., the ratio of the estimate of the expected number of false discoveries, over the number of discoveries. It is considered as a estimate of the FDR when \( \Delta \) is used as a threshold for calling discoveries.
6. Select a \( \Delta \) so that \( \text{FDR}(\Delta) \) is controlled at the desired rate, let this final threshold be \( \Delta^* \). All FFs \( j \) for which \( B_j > \Delta^* \) are considered to be FFs that affect the bioassay outcome through a gene set.