Integrative Methods for the Analysis of Structure-Transcription-Assay Relationships in Drug Discovery and Early Development

Nolen Joy Perualila

Promotor: Prof. dr. Ziv Shkedy
Co-Promotor: Dr. Adetayo Kasim
Co-Promotor: Prof. dr. Willem Talloen

Acknowledgements

This thesis is the culmination of my four-year epic journey across the space of statistical bioinformatics carrying out the QSTAR mission in the field of drug discovery and development. I couldn't be an accomplished STAR (i.e. **S**hkedy-**T**alloen-**A**detayo **R**esearcher) without other stars, near and far, that lend their light to direct me to the right path. I would like to thank them for journeying with me.

I have my utmost respect and unreserved gratitude to my promoter, Prof. dr. Ziv Shkedy. Thank you for your wisdom, unending support, guidance and encouragement throughout my Ph.D. I will always cherish every meeting, trip and humorous conversation we shared. Ziv, your kindness and patience to listen and address my concerns (be it personal or academic) will forever stay with me. Your trust and belief in me have made all this possible. I am looking forward to more insightful discussions with you!

I am privileged to have both Prof. dr. Willem Talloen and Dr. Adetayo Kasim as my co-promoters. Willem, our weekly visits in Janssen were always rewarding, not just because of free lunch, but especially because of your conscious effort to make time for our meetings despite your hectic schedule. Your calm and perceptive personality is admirable. Kasim, you always have the right words to uplift my spirit and motivate me to keep going. You have always extended your helping hand for me and my family! I only have positive words to describe the hospitality I experienced during my short stay in Durham. Thanks, Willem and Kasim, for sharing your knowledge, expertise and invaluable advice!

I am also grateful to Dr. Hinrich Goehlmann of the Functional Genomics department and Dr. Luc Bijnens of the Non-Clinical Statistics department of Janssen Pharmaceuticals. Hinrich, thank you for spearheading the IWT QSTAR project that became the framework of my Ph.D. The bi-yearly La Calestienne meetings, Kasterlee workshops and series of teleconferences are only a few of the pleasant memories to reminisce from this initiative. Luc, the nice working atmosphere in NCS is a reflection of your respectable leadership.

Thank you for extending it to us. It was a pleasure to be part of your team!

I would like to express my sincere gratitude to Prof. Javier Cabrera of Rutgers University, Dr. John Kostis and staff of the Cardio-Vascular Institute, New Jersey for their warm welcome and hospitality during my research stay. Javier, thank you for your guidance and for showing me a different perspective that summed up to a beautiful experience! The same goes to Dr. Dhammika Amaratunga for his expertise and input in my work, thanks!

I am also thankful to Dr. Andreas Bender and Aakash for a fruitful collaboration and for the opportunity to visit them in Cambridge. I would like to extend my appreciation to all members of the QSTAR Consortium for their contributions. Special thanks to Bie Verbist and Liesbet Vervoort for their expertise in chemistry and molecular biology, respectively. Also to Martin Otava, a wonderful colleague, dependable friend, careful driver, protective bodyguard, and trustworthy confidant rolled into one, thank you!

Moreover, I would like to thank Prof. dr. Tomasz Burzyowski, Prof. dr. Marc Aerts and many of the above-mentioned people for being part of my defense committee and for taking the time to read my thesis!

I am also glad to be surrounded by nice colleagues in CenStat. Your smiles and hellos are appreciated, thanks! To my research circle Marijke, Ewoud, Rudradev and Theophile and my travel buddies to Janssen, Sammy, Doreen, Leacky, Martin and Tanya, you've been a source of intellectual discussions, witty conversations and lasting friendship, thanks!

All these remarkable experiences in Hasselt University started six years ago, from the financial support of VLIR-UOS, which I gratefully acknowledge. I would like to extend my appreciation to Prof. Geraldine Garcia, who told me about the Masters Programme in Biostatistics. I am also thankful to Dr. Sonia Formacion, Prof. Vicente Balinas, Prof. Leah Araneta and Prof. Nilo Araneta for their support in my application.

I owe a great debt of gratitude to my extended family here in Belgium. Nang Gemma and Guido for treating me like your own daughter; Ma'am Tina Sotto for your time, help and advice; Ate Chella, John, Cris, Jedelyn, Pia, Ate Rochelle, Kian and Kuya Johan and Nang Gemma Sulit, for being always there, Salamat!

I am eternally indebted to my parents, Noel and Josephine, who supported me in all my pursuits. I am happy to share all that I have accomplished with my brothers and sisters and my in-laws. Thank you for your continuing love and support!

Finally, I dedicate this thesis to my husband, Cesar, who put his professional life on hold so I could achieve this dream. I cannot thank you enough for all your sacrifices, encouragements, unconditional love and unwavering support. I am blessed to have you and our daughter, Naya, my favorite star of all! Salamat Langga!

Nolen Joy Perualila
Diepenbeek, 6 November 2015

Publications

The materials presented here, are based on the following publications and reports:

Manuscripts

Ravindranath, A.C., **Perualila-Tan, N.**, Kasim, A., Drakakis, G., Liggi, S., Brewerton, S., Mason, D., Bodkin, M., Bhagwat, A., Talloen, W., Gohlmann, H.W. QSTAR Consortium, and Shkedy, Z., Bender, A. (2015). Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism of action analysis. *Molecular Biosystems* 11(1), 86–96.

Perualila-Tan, N., Kasim, A., Talloen,W., Verbist, B., Goehlmann, H., QSTAR Consortium, and Shkedy, Z. (2015) A joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. *Under revision in Statistical Application of Genetics and Molecular Biology.*

Perualila-Tan, N., Shkedy, Z, Talloen,W., Goehlmann, H., QSTAR Consortium, Van Moerbeke, M., and Kasim, A. (2015) Weighted similarity-based clustering of chemical Structures and bioactivity data in early drug discovery. *Under revision in Journal of Bioinformatics and Computational Biology.*

Perualila-Tan, N., Kasim, A., Talloen,W., Goehlmann, H., QSTAR Consortium, Shkedy, Z. (2016) Sparse multiple factor analysis for high-content screening and gene expression data integration to guide drug development. *To be submitted for publication.*

Book Chapters

Perualila-Tan, N., Shkedy, Z., Adetayo, K. and the QSTAR consortium (2016). High Dimensional Biomarkers in Drug Discovery. *In*: Molenberghs, G., Alonso, A., Van der Elst, W., Bigirumurame, T., Buyse, M., Burzykowski, T. Shkedy, Z. (eds.), *Applied Surrogate Endpoint Evaluation Methods with SAS and R.* Chapman & Hall/CRC (to be published in July 2016).

Perualila-Tan, N., Shkedy, Z., Ravindranath, A., Drakkis, G., Liggi, S., Bender, A., Kasim, A., Talloen, W. and Göhlmann, H.W. (2016). Biclustering Methods in Chemoinformatics and Molecular Modelling in Drug Discovery Experiments. *In*: Kasim, A., Shkedy Z., Kaiser, S., Hochreiter, S. and Talloen, W. (eds.), *Applied Biclustering Methods for High Dimensional Data Using R.* Chapman & Hall/CRC.

Perualila-Tan, N., Shkedy, Z., Amaratunga, D., Cabrera, J. and Kasim, A. (2016) Enrichment of Gene Expression Modules using Multiple Factor Analysis and Biclustering. *In*: Kasim, A., Shkedy Z., Kaiser, S., Hochreiter, S. and Talloen, W. (eds.), *Applied Biclustering Methods for High Dimensional Data Using R.* Chapman & Hall/CRC.

Perualila-Tan, N., Shkedy, Z., Clevert, D.A., Klambauer, G., Hochreiter, S. (2016) Ranking of Biclusters in Drug Discovery Experiments. *In*: Kasim, A., Shkedy Z., Kaiser, S., Hochreiter, S. and Talloen, W. (eds.), *Applied Biclustering Methods for High Dimensional Data Using R.* Chapman & Hall/CRC.

Shkedy, Z., Sengupta, R. and **Perualila-Tan, N.** (2016) Identification of Local Patterns in the NBA Performance Indicators. *In*: Kasim, A., Shkedy Z., Kaiser, S., Hochreiter, S. and Talloen, W. (eds.), *Applied Biclustering Methods for High Dimensional Data Using R.* Chapman & Hall/CRC.

Amaratunga, D., Cabrera, J., **Perualila-Tan, N.**, Kasim, A. and Shkedy, Z.(2016) from Clustering to Biclustering. *In*: Kasim, A., Shkedy Z., Kaiser, S., Hochreiter, S. and Talloen, W. (eds.), *Applied Biclustering Methods for High Dimensional Data Using R.* Chapman & Hall/CRC.

Conference Proceedings

Perualila-Tan, N., Kasim, A., Talloen,W., Goehlmann, H. and Shkedy, Z. (2014) Development of Genetic Biomarkers in drug discovery and early drug development experiments. In JSM Proceedings, Biopharmaceutical Section. Alexandria, VA: American Statistical Association. 3021-3030.

Software Products

Perualila-Tan, N. (2015). `biclustRank`: Ranking of Biclusters. R package. `https://r-forge.r-project.org/projects/biclustrank/`.

Perualila-Tan, N., et al. (2015). `IntBC`: Integrative Biclustering. R package. `https://r-forge.r-project.org/projects/intbiclust/`.

Van Moerbeke, M., **Perualila-Tan, N.** (2015). `IntClust`: Integrative Clustering. R package. `https://r-forge.r-project.org/projects/IntClust/`.

Contents

List of abbreviations

BA	:	Bioactivity
BH	:	Benjamini-Hochberg
BIC	:	Bayesian Information Criterion
biMFA	:	biclustering with Multiple Factor Analysis
ECFP	:	Extended Connectivity Fingerprint
FABIA	:	Factor Analysis Bicluster Acquisition
FARMS	:	Factor Analysis for Robust Microarray Summarization
FDR	:	False Discovery Rate
FF	:	Fingerprint Feature
GE	:	Gene Expression
GO	:	Gene Ontology
HCS	:	High Content Screening
HTS	:	High Throughput Screening
IC50	:	Inhibitory Concentration 50
I/NI	:	Informative/Non-informative
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
MFA	:	Multiple Factor Analysis
MoA	:	Mechanism of Action
NB	:	Naive Bayes
QSAR	:	Quantitative Structure Activity Relationship
QSTAR	:	Quantitative Structure Transcription Assay Relationship
SEM	:	Structural Equations Modelings
SMFA	:	Sparse Multiple Factor Analysis
SPCA	:	Supervised Principal Component analysis
sPCA	:	sparse Principal Component Analysis
SVD	:	Singular Value Decomposition
SSVD	:	Sparse Singular Value Decomposition

Chapter 1

Introduction

1.1 Integrated Analysis of Multi-source Drug Discovery Data

Early drug discovery research and development process involves a range of technologies for measuring the chemical and biological effects of compounds at the molecular level in order to make a decision about the development of a new drug. Consequently, this process generates multiple sources of high-dimensional data which include high-throughput screening (HTS), chemical structures, gene expression, image-based high-content screening (HCS), among others. An integrated analysis of these data sources is the central theme of this thesis. High-dimensional data are characterized as having an enormous number of features (variables) and relatively few compounds (samples). This leads us to the problem of data integration and opens up a challenging venue for methodological development and application to extract vital information from the intersection of biology and chemistry. An integrative method that allows to detect the relationship of all these features can be very relevant to evaluate compound efficacy and safety as lead compounds progress through lead optimization.

In drug discovery, scientists work together and start to identify a potential biomolecular “*target*" which is usually a single molecule, typically a protein, that is involved in a particular disease. This target should be drugable, that is, it can interact with and be affected by a molecule. Upon the identification and validation of the target follows the process of discovering promising compounds which could ultimately turn into a medicine for a particular disease. The discovery, therefore, starts on either creating a new molecule or repurposing an existing molecule. At this point, thousands of candidate molecules could be screened against the target for activity using HTS assays and then optimize it by modifying its structure for better activity.

Over several decades, Quantitative Structure-Activity Relationship (QSAR) modeling techniques (Nantasenamat *et al.*, 2009) have been extensively used to quantify the relationship between chemical structure and activity to gain understanding on how the chemical substructures affect the biological activity of a compound and then use this understanding to design compounds with improved activity either relating to greater efficacy or lesser toxicity (Dearden, 2003, Martin *et al.*, 2002, Bruce *et al.*, 2008). The fundamental principle underlying the QSAR approach is based on the observation that chemicals of similar structures frequently share similar physiochemical properties and biological activities (Johnson and Maggiora, 1990, Verma *et al.*, 2010).

The Quantitative Structure-Transcriptional-Assay Relationship (QSTAR, Ravindranath and Perualila-Tan *et al.*, 2015, Verbist *et al.*, 2015, Perualila-Tan *et al.*, 2016) modeling framework is an extension of the QSAR approach (Figure 1.1). Here, transcriptional data are integrated with structural compound information as well as experimental bioactivity data in order to analyze compound effects in biological systems from different angles to elucidate the mechanism of action of compounds (MoA). This could provide an insight into inadvertent phenotypic effects which can greatly help in early-stage pharmaceutical decision-making.

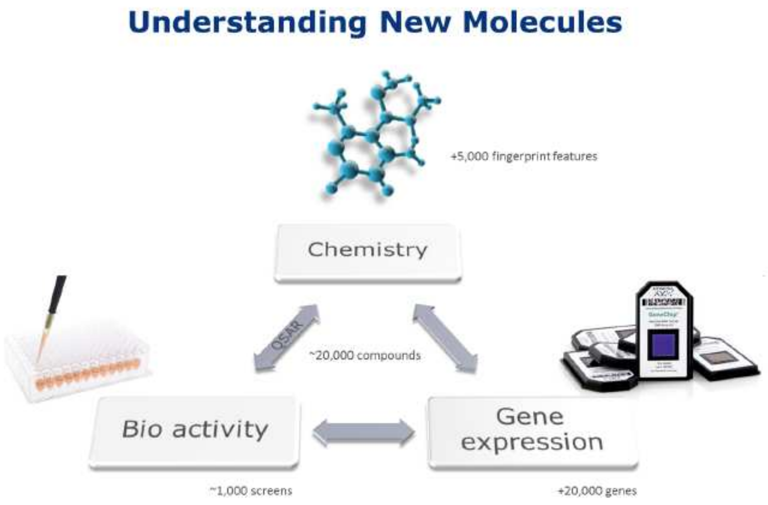


Figure 1.1: The QSTAR framework. The integration of 3 high-dimensional datatypes; gene-expression, fingerprints features (FFs representing the chemical structures) and bioassay data (phenotype).

Although the bioactivity data which is typically measured per target assay is key in the optimization process of chemically designing compounds, it does not provide much valuable insight information on the underlying biological mechanisms. In contrast to the bioassay data that captures single biological effects, the gene expression data, as a multi-dimensional assay, measures a wide diversity of biological effects of a compound on a whole genome transcriptional level, and thereby gives an information-rich snapshot of the biological state of a cell (Göhlmann and Talloen, 2009, Amaratunga *et al.*, 2014). Transcriptomic changes following compound administration can also be measured in high throughput, allowing screening of many compounds in multiple cell lines at low cost. 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 (Verbist *et al.*, 2015). Applications using gene expression profiles to observe several genes and signalling pathways concurrently enrich the understanding of underlying mechanisms. 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. Within the QSTAR framework, mRNA biomarkers may be discovered by compounds that cause disease-related variation of the gene expression. Analysis of the transcription profiles allows identifying new biomarkers related to certain biological effects induced by these compounds. With this approach, a significant amount of resources can be saved with identification of undesired compound effects avoiding failures in the late-stage pharmaceutical drug development.

Due to the advances in the genome studies, there is a wealth of microarray data that has been deposited in public databases such as Expression Atlas, which is a subset of ArrayExpress (Kapushesky *et al.*, 2011, Brazma *et al.*, 2003). In recent years, the new transcriptomics databases “Connectivity Map (CMap)" and “Library of Integrated Network-Based Cellular Signatures (LINCS Consortium (2013))" become publicly available and allow researchers to explore and characterize biological effects of small molecules in a large scale. Several applications ranging from pathway elucidation (Bai *et al.*, 2013), toxicity models (van der Veen *et al.*, 2013, Magkoufopoulou *et al.*, 2012) and toxicogenomics classifications (Jiang *et al.*, 2007), to tool discovery and drug repurposing (Iorio *et al.*, 2010, Dudley *et al.*, 2011, Sirota *et al.*, 2011, Pacini *et al.*, 2013), have been developed based on drug-induced gene expression profiling (Bol and Ebner, 2006). Verbist *et al.* (2015) demonstrated the utility of transcriptomics to guide lead optimization in various QSTAR drug discovery projects.

The data analysis approach of QSTAR provides the opportunity to explore the application or development of methods for data integration which is the main topic that we discuss in this thesis. Our aim is to build a knowledge platform to assist decision-making in drug discovery and early development projects, i.e. prioritizing of chemotypes during hit to lead, lead optimization and identifying analog structures.

This thesis constitutes several analysis workflows to integrate three high-dimensional datatypes; gene-expression, fingerprint features (FFs representing the chemical structures) and bioassay data (phenotype). The methodologies presented in this thesis are divided into three types: the QSTAR modeling framework, semi-supervised methods, from clustering to biclustering analysis, and unsupervised multivariate methods for data exploration and integration. The last part of the thesis discusses the statistical software developed alongside with the methodologies.

The first part of the thesis is devoted to statistical models that are suitable for the QSTAR setting. **Chapter 2** covers the joint modeling framework which allows 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 bioassay 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). **Chapter 3** covers the relationship of joint modeling to path analysis modeling and conditional modeling.

The second part of the thesis includes sequential integration of datasets in order to explain the mechanism of action of a subset of compounds using clustering and biclustering techniques (Kasim *et al.*, 2016). Clustering algorithms use the similarity data in order to group objects and are typically performed on one data source. In **Chapter 4**, a clustering solution that handles multiple data sources is presented in the context of drug discovery. A typical strategy in compound selection involves the clustering of compounds based upon their chemical structure. This idea is extended wherein an integrative clustering approach that makes use of both data sources for the purpose of discovering a subset of compounds with aligned structural and biological properties is presented. This method combines bioactivity-based and chemical structure-based similarity matrices, assigned with complementary weights, producing a weighted similarity matrix, which is the generic input in any clustering algorithm. A secondary analysis is performed wherein each biologically and structurally driven compound cluster is further linked to a set of transcriptomic features. A new subset of compounds that are structurally and biologically similar to the reference compounds are discovered using the proposed clustering approach. **Chapter 5** deals with integrating gene expression profiles with certain proteins to improve our understanding of the fundamental mechanisms in protein-ligand binding. This chapter spotlights the integration of gene expression data and in silico target prediction scores, providing insight into Mechanism of Action (MoA). Compounds are clustered based upon the similarity of their predicted protein targets and each cluster is linked to gene sets using

Linear Models for Microarray Data. Pathway analysis is used to identify gene sets based upon their biological processes and a qualitative search is performed on the homogeneous target-based compound clusters to identify pathways. **Chapter 6** presents a workflow on ranking gene expression data-based biclusters using another source of information, in this case, the chemical structure.

The third part of the thesis is comprised of three chapters starting with **Chapter 7** that introduces the use of Multiple Factor Analysis (MFA) for datasets normalization and integration. Here, the three QSTAR datasets are used as input in the analysis. **Chapter 8** illustrates the use of MFA as a gene-module enrichment technique. In **Chapter 9**, two variants of MFA are presented, SMFA and biMFA. In this chapter, the gene expression (GE) data is integrated with high content screening (HCS) data to relate the compounds' transcriptional effects with image-based bio-activity measures in a cell following administration. Identifying phenotypic subclasses (genes and HCS) that are co-regulated across a subset of compounds can be used as a "biology" screening tool to assess compounds' potential for efficacy/toxicity. This is in line with the basic concept of biclustering but accounting for multiple data sources. Hence, for this purpose, sparse Multiple Factor Analysis (SMFA) and biclustering with MFA (biMFA) are developed to simultaneously search for features-compounds association. These integrative methods combine the ideas of MFA and penalized singular value decomposition techniques. The results highlight a group of potentially geno-toxic lead compounds, a Tubulin-linked compound set along with respective HCS features indicators.

The last two chapters of the thesis present the R products which were developed for methodologies presented in the thesis. The first R package biclustRank is presented in **Chapter 10** while **Chapter 11** discusses the R package biMFA developed for methodologies in the third part of the thesis.

1.2 Case Studies and Datasets

In this section we discuss the structure of the three types of dataset, transcriptomic, bioassay and chemical structure data. The QSTAR framework presented in Figure 1.1 can be translated into the data structure presented in Figure 1.2. Note that compounds are the common dimension and form the link between the three different data sources. To be able to link these data sources, a standard way to encode the molecular information in a textual identifier was needed, and to this end InChiKey's were generated for all datasets as unique identifiers (McNaught, 2006). Additionally, the target prediction data can be calculated with the bioactivity and chemical structure data as input. In what follows we describe each dataset in more details.

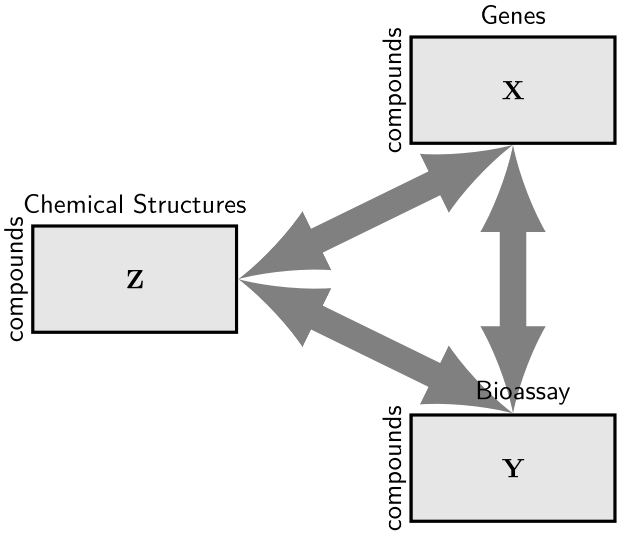


Figure 1.2: The QSTAR data structure.

1.2.1 Transcriptomic Data

For microarray data, pre-processing steps attempt to remove the technical sources of variation originating from different levels during the process, from manufacturing of the microarrays to the biological and the microarray experiment. Gene expression raw intensities were log-transformed, quantile normalized and subsequently summarized using Factor Analysis for Robust Microarray Summarization (FARMS) using Entrez gene customCDF annotations (Hochreiter *et al.*, 2006). Probe sets were filtered using I/NI filtering (Talloen *et al.*, 2007, Kasim *et al.*, 2010, Amaratunga *et al.*, 2014).

FARMS is a model based approach for summarizing microarray data (Hochreiter *et al.*, 2006). The main idea of the FARMS algorithm for expression arrays is to detect a common hidden cause of the observed measurements that cannot arise from noise which is uncorrelated for different measurements. The hidden cause is the true but unobserved mRNA concentration in the tissue sample which leads to a simultaneous decrease or increase in probe intensities measuring this mRNA. The hidden cause is called “signal” since it indicates the mRNA concentration. The core of the FARMS algorithm is a factor analysis – a multivariate technique to detect a common structure in the data of multiple probes that measure the same target. The informative/non-informative (I/NI) calls were proposed as an extension of the FARMS algorithm (Talloen *et al.*, 2007, 2010) for gene filtering.

The pre-processed gene expression matrix is given by $\mathbf{X}$ where the element x_{ji} denotes the expression level of the jth gene for the ith compound, $j = 1, \ldots, m$ and $i = 1, \ldots, n$,

$$\mathbf{X}_{m\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

1.2.2 Bioassay Data

The experimental bioactivity data are part of an internal database at Janssen Pharmaceutica which covers both on- and off-target effects. However, these data are rather sparse since only the most promising compounds are screened to a broader number of assays to assess potential efficacy and safety issues. Per target stimulated assay, compounds were tested related to the inhibition using several doses. From the dose-response experiment, the IC50 (Lin *et al.*, 2012), half-maximal inhibitory concentration, is extracted as the compound's measurement of efficacy (Figure 1.3). In all analysis, the pIC50 scale (-log IC50) is used, in which higher values indicate exponentially greater potency.

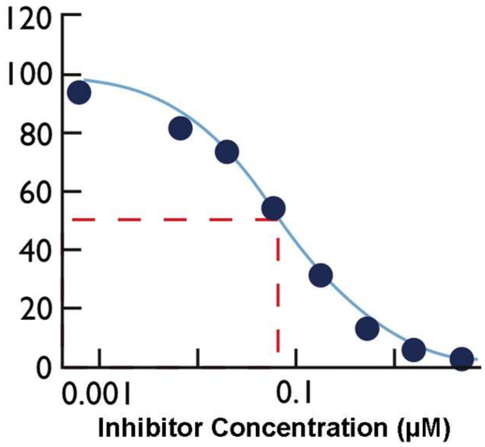


Figure 1.3: The IC50 (vertical dashed line) is the concentration of an inhibitor where the response (or binding) is reduced by half.

The bioactivity matrix, $\mathbf{Y}_{B\times n}$, with element y_{bi} representing the activity value of the ith compound measured on the bth assay is given by

$$\mathbf{Y}_{B\times n} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_{B1} & y_{B2} & \cdots & y_{Bn} \end{pmatrix}.$$

1.2.3 Chemistry Data (Chemical Encoding)

The whole molecule can be described as a list of local substructure patterns (fingerprint features). The structural information of the compounds is encoded using a jCompoundMapper-based pipeline (Hinselmann *et al.*, 2011). The Extended-Connectivity Fingerprints (Bender *et al.*, 2004) with a search depth of 6 (ECFP6) algorithm for calculating fingerprint features (FFs) for compounds is implemented. The ECFPs were developed specifically for structure-activity modeling. This is characterized by a vector of binary values, also termed as fingerprint feature (FF), denoting the presence/absence of a certain chemical substructure in a compound.

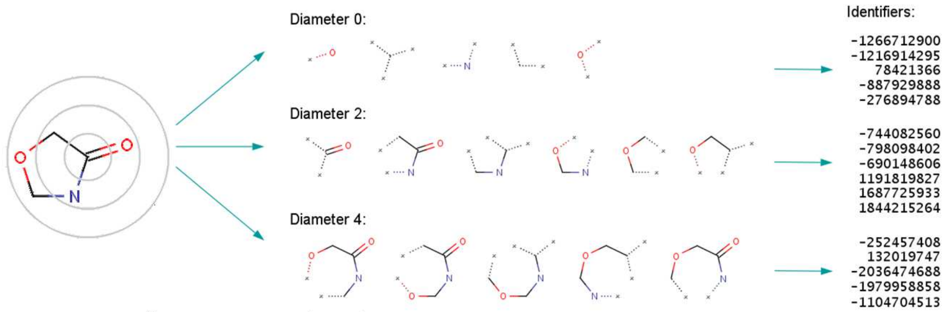


Image from : https://docs.chemaxon.com/pages/viewpage.action?pageId=14483752

Figure 1.4: The ECFPx generation process of a molecule starts at a single atom and expands iteratively to the next topological level of connected atoms till a maximum threshold x is reached.The fingerprint of a whole molecule is a combination of all FFs executing this method from every single atom. Note that these are all the substructures (fingerprint features) present in this molecule.

Figure 1.4 illustrates the encoding process. The ECFP fingerprint algorithm starts from a single atom and expands in circles (topological in the 2D graph) to the next level of connected atoms. It captures the bond orders of the bonds between atoms and ensures substructures are normalized, which assigns each substructure a unique number. The expansion of the spheres stops by default after six expansion steps, so we are calculating ECFP6 FFs. This process is then started from every single atom in the compound. A detailed description of the algorithm can be found in Rogers and Hahn (2010).

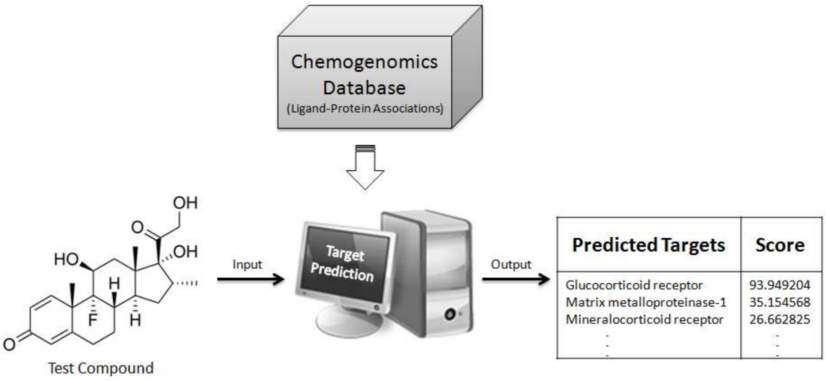
For the analysis presented in this thesis we use the FF matrix. Let $\mathbf{Z}_{K\times n}$ be the chemical structure or fingerprint feature matrix in which the kith element, $z_{ki} = 0$ representing absence or $z_{ki} = 1$ denoting presence of the kth fingerprint feature in the ith compound. The kth row vector refers to the kth fingerprint feature profile while the ith column vector refers to the ith molecule's fingerprint.

$$\mathbf{Z}_{K\times n} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ z_{K1} & z_{K2} & \cdots & z_{Kn} \end{pmatrix}.$$

1.2.4 Target Prediction Data

Although a compound is designed to hit a single target, compound molecules often interact with multiple targets, coined as polypharmacology (Reddy and Zhang, 2013). This unintended compound-target interactions may give rise to undesired side-effects. Experiments using phenotypic based assays can be performed to quantify the activity of every molecule against a biological molecular target. However, this process is often time- and resource-intensive, making it unfeasible to test all possible compound-target interactions, especially when little or no previous knowledge of potential molecular targets is available (Terstappen *et al.*, 2007). Thus, in silico target prediction data was developed to describe the likelihood of compounds to be active on a certain target. In this section we describe the target prediction algorithm developed by Koutsoukas *et al.* (2013). This is a probabilistic machine learning algorithm for predicting protein targets of bioactive molecules, which employs the Laplacian-modified Naive Bayes classifier (NB). Chemical similarity is the underlying principle of the method which is built on the approach that, if compounds are similar in structural space they trigger similar targets. Compounds structural features (Extended Connectivity Fingerprints 4) are used as molecular descriptors.

The target prediction algorithm was employed to predict probable protein targets for compounds without target information (Klabunde, 2007, Koutsoukas *et al.*, 2011). The resulting prediction provides each test compound with probable protein targets and their respective scores, representing the likelihood of binding to 477 protein targets (Koutsoukas *et al.*, 2011, 2013). Figure 1.5 visualizes the target prediction process.



Predicted Targets	Score
Glucocorticoid receptor	93.949204
Matrix metalloproteinase-1	35.154568
Mineralocorticoid receptor	26.662825
⋮	⋮

Figure 1.5: Target prediction overview. The orphan compound fingerprint information is fed into the algorithm, which predicts the likelihood (score) of binding to proteins based upon prior knowledge. This method establishes the link between the compound and protein targets, further linking it to the MoA.

The NB classifier is defined using the following equation (Koutsoukas *et al.*, 2013),

$$P(C=\omega|D=f)=\frac{P(D=f|C=\omega)P(C=\omega)}{P(D=f)}.$$

Here, the probability of a compound, C, belonging to feature class ω given the chemical feature f is calculated. $P(C=\omega)$ is the priori probability of C belonging to feature class ω and $P(D=f)$ is a priori probability of the features, f. $P(D=f|C=\omega)$, is the key value in this equation, which is the likelihood of the feature f given the class ω. This probability is estimated by the NB classifier from a training set, which assumes that the features are independent of each other for a given class. It has been observed before that the NB classifier is still an effective classifier in cases where features are correlated. In machine learning practices, a training set is employed for the classifier to learn from the examples and make predictions for the unseen dataset; the test set. The classifier is trained on a large benchmark dataset of bioactive compounds retrieved from the publicly available ChEMBL database, which is a repository of small bio-active molecules extracted from scientific literature. The training dataset covers 477 human protein targets with around 190,000 protein-ligand associations, based upon the reported bioactivities (Ki/Kd/IC50/EC50) being equal or better than 10 μM with a confidence score of 8 or 9. These rules for extracting compounds ensured reliable compound-target associations for training the model. The target prediction algorithm performance was evaluated by 5-fold cross validation (Koutsoukas *et al.*, 2013).

At the end of the process, for each compound, the posterior probability to bind to a target (target prediction score) in the target list is calculated. In the next step, the target prediction score matrix is dichotomized using target-specific confidence score cut-offs calculated internally, in order to increase the prediction accuracy (Paolini *et al.*, 2006). The target prediction score matrix is given by

$$\mathbf{T}_{M\times n} = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ t_{M1} & t_{M2} & \dots & t_{Mn} \end{pmatrix},$$

with entries,

$$t_{mi} = \begin{cases} 1 & \text{compund } i \text{ is predicted to hit target } m, \quad i = 1, \ \dots, \ n \text{ and } m = 1, \dots, M, \\ 0 & \text{otherwise.} \end{cases}$$

1.2.5 High Content Imaging Data

For high content image analysis, 661 features per compound were extracted using DCILabs and Phaedra images (Figure 1.6). The features provides measurements on the granularity, shape, intensity, texture, etc. from three genomic loci LaminB1 (nuclear pore), TUBA1B (microtubules) and ACTB (actin filament) labelled with respectively blue, green and red fluorescent protein genes.

We define the high content imaging matrix, $\mathbf{H}_{S\times n}$, for which the sith entry is the sth image feature measurement of the ith compound where $i = 1, ..., n$ compounds and $s = 1, ..., S$ HCS features given by,

$$\mathbf{H}_{S\times n} = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ h_{S1} & h_{S2} & \dots & h_{Sn} \end{pmatrix}.$$

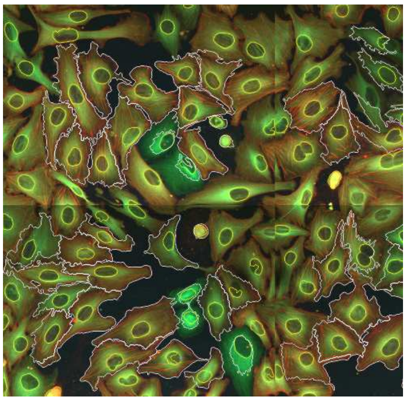


Figure 1.6: High content imaging picture of flourescently labeled cell culture cells that have been analyzed to identify the cell boundaries and the cell nuclei.

1.3 QSTAR Drug Discovery Projects

The datasets mentioned in Section 1.1 were collected for several drug development projects. Each project is focused on a different compound family and contains different number of compounds. In this section we present the overview of the different projects used for the analysis presented in the thesis. Table 1.1 presents the number of compounds, genes after I/NI, fold change and variance filtering, bioassays and unique fingerprint features per development project and a short description of the projects is given below.

Table 1.1: Overview of QSTAR datasets.

Project	Compounds	Genes	Primary Bioassay	Unique FF	Targets	Therapeutic area	Chapters
mGluR2PAM	62	566	4	300	NA	neuroscience	6,8
ROS1	89	1289	1	312	NA	oncology	2,3
EGFR	35	3595	4	138	NA	oncology	2,3,4,7
PDE10	16	75	661(HCS)	NA	NA	neuroscience	9
CMap	176	2434	NA	NA	477	oncology	5,8

1.3.1 mGluR2PAM

Glutamate is the major excitatory neurotransmitter in the human brain and different metabotropic glutamate receptors (mGluR) function to regulate glutamate release. mGluR2 is an overspill-receptor: if too much glutamate is present, it will bind to mGluR2 and this will decrease further glutamate-release. In several anxiety and stress disorders (eg. schizophrenia, anxiety, mood disorders and epilepsy) an excess of glutamate is present, leading to an over-excitement of the neurons. Targeting the mGluR2 with positive allosteric modulators (PAM) will reduce glutamate release in the presence of glutamate, which binds to the orthosteric site PAM (Soudijn *et al.*, 2004). These mGluR2 PAMs could be interesting compounds for the treatment of anxiety disorders. In this project, gene expression profiling was done for 62 compounds and 566 differentially expressed genes remained after the filtering steps. There were 300 unique fingerprint features for this compound set.

1.3.2 ROS1

ROS1 (reactive oxygen species) is a proto-oncogene which is highly-expressed in a variety of tumor cell lines and belongs to the sevenless subfamily of tyrosine kinase insulin receptor genes. Aberrant expression and oncogenic fusions resulting from chromosomal rearrangement occur in lung cancer, cholangiocarcinoma and glioblastoma. Aberrant expression is also detected in a variety of other cancer types. ROS1 inhibition is expected to have anti-tumoral effects in cells where ROS1 is activated (Acquaviva *et al.*, 2009, Charest *et al.*, 2006). Note that currently, the endogenous ligand is still unknown.

This project sought to develop compounds that inhibit ROS1. The ROS1 dataset consists of eight-nine (89) compounds tested for target inhibition. A total of 1289 differentially expressed genes were retained after the pre-processing steps. For this project, a total of 312 unique profiles of fingerprint features was generated from the 89 compounds.

1.3.3 EGFR

The EGFR project focuses on inhibition of the epidermal growth factor receptor (Woodburn, 1999). 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 all the filtering steps. Moreover, a total of 138 unique profiles of chemical substructures were identified for this compound set.

1.3.4 PDE10

PDE10A (phosphodiesterase 10) is an intracellular enzyme that is present in the brain. The high level of expression of PDE10A in the brain suggests that inhibition of this enzyme will result in changes in behaviors. PDE10A inhibitors may represent a novel approach to the treatment of behavioral disorders like psychosis (Menniti *et al.*, 2006). In this project, the efficacy of the investigated compounds was high and the focus is on adverse effects

(Verbist *et al.*, 2015). Therefore, a follow-up experiment was performed wherein sixteen (16) compounds targeting PDE10 were subjected to microarray experiment and high content imaging to explore potential unwarranted toxic effects. Only 75 genes were retained after pre-processing. A total of 661 HCS features are measured for each compound image.

1.4 The Connectivity Map (CMap)

In addition to the datasets mentioned in Section 1.3, the Connectivity Map (CMap, Lamb *et al.* (2006)) dataset is used for the analyses presented in this thesis as well. The CMap dataset contains information about 2434 differentialy expressed genes after the filtering steps measured for 1309 compounds and is available online. The CMap study aims to construct large libraries of drug and gene signatures and provides a pattern-matching tool that detects signature similarities in order to establish a relationship between disease and therapeutic MoA. The libraries were used to design the method that compares gene signatures to diseases in the database and predict the connection; the MoA. Due to the ability of finding connections and similarities between the genes, disease and drugs, the results are termed connectivity maps. The database consists of 1309 diverse bioactive compounds on four different cell lines, where nearly 800 of the compounds are currently available in the market (Lamb *et al.*, 2006, Lamb, 2007).

The CMap dataset was extracted from the Connectivity Map server and consisted of 1309 drug-like compounds with their respective genome-wide expression profiles. For the analyses presented in this thesis we used data from MCF7 (breast cancer epithelial cell) and PC3 (human prostate cancer) cell lines, containing 75 and 101 compounds, respectively. The compounds were retained after filtering for compounds administered for a duration of 6 hours and a maximum concentration of 10μM. When multiple instances of compounds were found, the average gene expression level was used.

For this data, target prediction scores of compounds from the 2 cell lines (MCF7 and PC3) were generated for the 477 Chembl protein targets.

Part I

QSTAR Modeling Framework for Multi-source Data Integration

Chapter 2

A Joint Modeling Approach to Guide Lead Selection and Genomic Biomarker Development

2.1 Introduction

In this chapter, we present the joint modeling of transcriptomic and phenotypic data, conditioned on the chemical structure as a fundamental modeling tool for data integration within this QSTAR modeling framework. This modeling approach can be used to uncover, for a given set of compounds, the association between gene expression and biological activity taking into account the influence of the chemical structure of the compound on both variables. The model allows to detect genes that are associated to the bioactivity data facilitating the identification of potential genomic biomarkers for compounds efficacy. In addition, the effect of every structural feature on both genes and pIC50 and their associations can be simultaneously investigated.

Biomarker identification is a major application of microarrays in early drug development which often parallels and facilitates compound selection. Many studies have been devoted to identify genes that are correlated to a biological activity of interest, for instance, the inhibition of a certain enzyme. 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 the rats under treatment and the HAMD scores for psychiatric patients). Van Sanden *et al.* (2012) identified gene specific biomarkers for toxicity data presented as a binary response.

In this chapter we follow the modeling approach proposed by Lin *et al.* (2010) and Tilahun *et al.* (2010) and use a joint model for the QSTAR framework. The joint modeling framework 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 associated with the bioassay data from a biological target(s) of interest and (3) investigate whether the association between the compounds and the bioassay can be confirmed by the gene expression changes (either on- or off- target related).

Identifying relevant genes that are associated with biological response is already a valuable information, but showing that this association is caused by the presence or absence of a particular chemical substructure(s) provides additional information that is particularly useful in drug design to improve or prioritize compounds.

2.2 Method

2.2.1 The Joint Modeling Framework

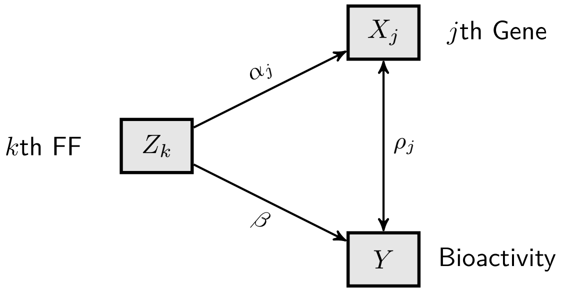


Figure 2.1: Gene-specific and fingerprint-specific joint model for the QSTAR data.

Let $\mathbf{X}$ be the gene expression matrix and x_{ji} be the jth gene expression of the ith compound, where$j = 1, \ldots, m$ and $i = 1, \ldots, n$. Let y_i denote the corresponding bioassay data for the ith compound. Both gene expression and bioassay read-outs are assumed to be normally distributed. Let $\mathbf{Z}_{K \times n}$ be the binary chemical structure or fingerprint feature matrix in which the kith element takes a value of one $(z_{ki} = 1)$ or zero $(z_{ki} = 0)$, if the kth fingerprint feature is respectively present or absent in the ith compound. Schematically, the gene-by-gene, fingerprint-by-fingerprint joint model is given in Figure 2.1. 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:

$$\begin{aligned} X_{ji} &= \mu_j + \alpha_j Z_i + \varepsilon_{ij}, \\ Y_i &= \mu_Y + \beta Z_i + \varepsilon_i, \end{aligned} \tag{2.1}$$

or equivalently formulated as

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_j + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma_j \right], \tag{2.2}$$

where the error terms have a joint zero-mean normal distribution with gene-specific covariance matrix, Σ_j.

$$\Sigma_j = \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix}. \tag{2.3}$$

The parameters α_j and β represent fingerprint feature effects for the jth gene and the bioassay data, respectively, and μ_j and μ_Y are gene-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), a coefficient that is derived from the covariance matrix, Σ_j,

$$\rho_j = \frac{\sigma_{jY}}{\sqrt{\sigma_{jj}\sigma_{YY}}}. \tag{2.4}$$

Indeed, $\rho_j = 1$ indicates a deterministic relationship between the gene expression and the response after adjusting for fingerprint feature.

2.2.2 Inference

As mentioned in section 2.2.1, the model tests for differentially expressed genes based on the hypotheses,

$$\begin{aligned} H_{0_j} &: \alpha_j = 0, \\ H_{1_j} &: \alpha_j \neq 0. \end{aligned} \tag{2.5}$$

For a microarray with m genes, there are m null hypotheses to be tested, which requires multiple testing correction. Throughout this chapter, we apply the FDR approach proposed by Benjamini and Hochberg (1995).

Moreover, in order to make inference about ρ_j, there is a need to test whether the expression level of a gene and the bioassay read-out are correlated, specifically, whether the expression level of a gene can predict the bioassay read-out. Thus, in addition to the hypotheses in (2.5), one needs to test the hypotheses

$$\begin{array}{lcl} H_{0_j}: \rho_j = 0, & \text{or equivalently} & H_{0_j}: \sigma_{jY} = 0, \\ H_{1_j}: \rho_j \neq 0. & & H_{1_j}: \sigma_{jY} \neq 0. \end{array} \tag{2.6}$$

Under the null hypothesis, the joint model in (3.2) is reduced to

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_j + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma_j = \begin{pmatrix} \sigma_{jj} & 0 \\ 0 & \sigma_{YY} \end{pmatrix} \right]. \tag{2.7}$$

Consequently, the inference for the adjusted association was based on a likelihood ratio test by comparing models (2.2) and (2.7). 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 $H_{0j}: \rho_j = 0$ for all the genes simultaneously per fingerprint feature.

2.3 Graphical Interpretation of Association Between a Gene and Bioactivity Accounting for the Effect of a Fingerprint Feature

Several interesting associations between genes and a response accounting for the effect of a fingerprint feature can be discovered by using the joint model. The different types of association are presented in Figure 2.2 using 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 and response presented in the upper panels of Figure 2.2 (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 or absence of the fingerprint feature. Note that the linear pattern remains after adjusting for 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.

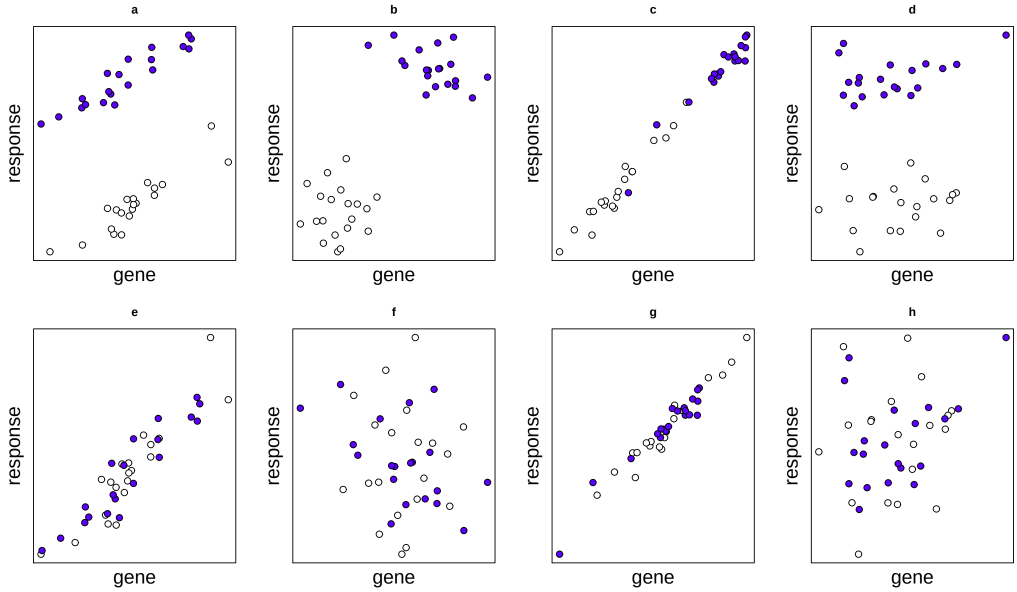


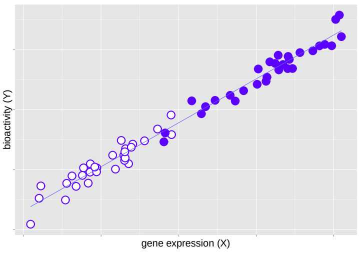
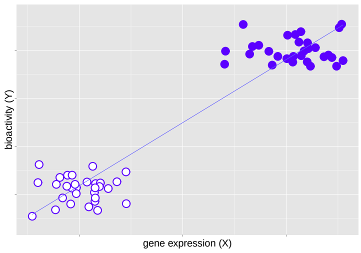
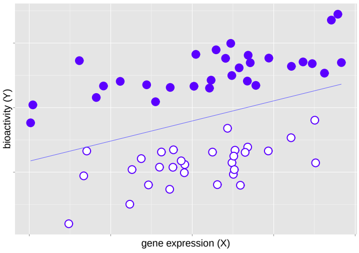
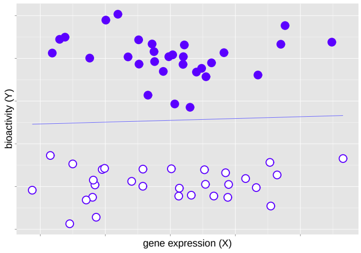
Figure 2.2: Hypothetical illustrations of associations between a response variable, gene expression levels and fingerprint features. Each point represents a compound. Solid blue points and black points represent the presence/absence of a fingerprint feature, respectively. Upper row panels: scatterplots for the response versus the gene expression. Lower row panels: scatterplots for the residuals after adjusting for fingerprint effects.

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 induce higher activity than those that don't have the fingerprint 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 feature.

Lastly, most genes are expected to be uncorrelated with the bioassay data 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)).

The joint modeling framework is useful for identifying genes that can predict compound activity, measured by pIC50, and can therefore serve as genetic biomarkers for compounds' efficacy. On top of this, 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. Consequently, as shown in Table 2.1, genes can be classified into subgroups according to the results obtained from the hypothesis testing in (2.5) and (2.6).

Table 2.1: Subclasses of genes using a hypothetical example. If both α and β are significantly different from zero, the correlation between the gene expression and bioactivity is present but in contrast with scenario (a), the gene expression in scenario (b) is correlated with the bioactivity variable only due to the effect of the fingerprint feature, hence its adjusted association is zero, $\rho_j = 0$. From the point of view of the structural optimization in the early drug development, the association observed in (b) is desirable while the one observed in (a) is an ideal genetic biomarker for bioactivity.

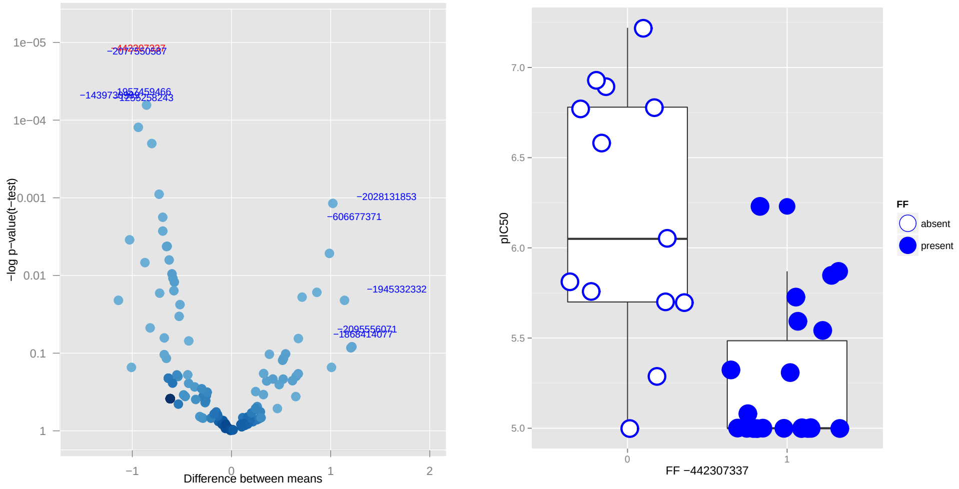
	$\rho_j \neq 0$	$\rho_j = 0$
$\beta \neq 0$ & $\alpha_j \neq 0$	 **(a)** X and Y are correlated, the gene is differentially expressed.	 **(b)** X and Y are correlated but are **conditionally independent.**
$\beta \neq 0$ & $\alpha_j = 0$	 **(c)** X and Y are correlated, the gene is not differentially expressed.	 **(d)** X and Y are uncorrelated.

For the first group of genes, the association between the gene expression and pIC50 exists regardless of the effect of a chemical substructure of the compound while the association from the second group of genes is driven by the fingerprint feature. This association can also further expand our knowledge about the biological mechanisms of compounds to guide decision-making in lead selection. Ideally, results from the joint modeling of every fingerprint feature, gene and activity data are generated. In this chapter, we only present the results of applying the joint model using a fingerprint feature that is mostly associated with the variation in compound activity.

2.4 Analysis of the EGFR and the ROS1 projects

2.4.1 Application to the EGFR Project

This oncology project focuses on the inhibition of the epidermal growth factor receptor (EGFR) which has been identified in many human epithelial cancers, colorectal, breast, pancreatic, non-small cell lung and brain cancer (Shaib *et al.*, 2013).



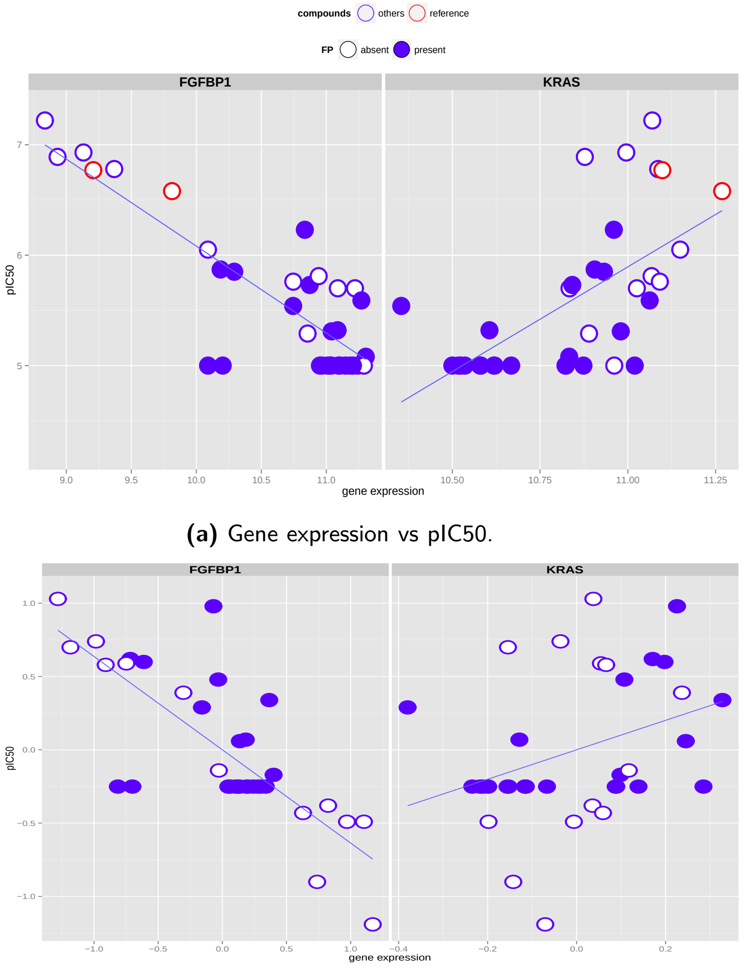
(a) Volcano plot to evaluate the capacity of the FFs to differentiate the bioactivity.

(b) The top FF for the EGFR project: -442307337.

Figure 2.3: FF -442307337 for the EGFR project.

For this project, of the 55 FFs that demonstrated differential effects on the primary bioassay, FF-442307337 came out first based on a feature-by-feature two-sample t-test of bioactivity data (Figure 2.3a). This substructure is prominent on less potent compounds, i.e. those with pIC50 values less than 6.5 (Figure 2.3b).

Several genes correlate with the inhibitory activity against the target. Figure 2.4 highlights the linear association between pIC50 from the anti-proliferation assay and gene expression changes of two on-target cancer-related genes: FGFBP1 and KRAS.



(a) Gene expression vs pIC50.

(b) Gene expression vs pIC50 after adjustment.

Figure 2.4: Two on-target genes that correlate with EGFR-inhibition: FGFBP1 and KRAS. Each point is a compound with the two reference compounds highlighted in red circle. The solid blue points indicate the presence of FF -442307337.

The gene FGFBP1 encodes for the fibroblast growth factor carrier protein (FGF-BP1) whose overexpression is noted in cell lines, from lung (Brattström *et al.*, 2002, Pardo *et al.*, 2003), prostate (Tassi *et al.*, 2006), pancreas (Kuwahara *et al.*, 2003), and colon cancer (Hauptmann *et al.*, 2003). By using the joint model, it has been shown that the expression is down regulated via the MAPK/ERK pathways after EGF-stimulated inhibition of EGFR (Harris *et al.*, 2000). Figure 2.4a shows that more potent compounds down-regulate FGFBP1 but upregulate KRAS.

KRAS protein has a pivotal role in the transduction of EGFR signaling (Shaib *et al.*, 2013), it encodes a small GTP binding protein that transmits the original signal from EGFR downstream to activate important cell functions, in particular, proliferation and survival (van Krieken *et al.*, 2008). Upregulation of the KRAS gene in response to EGFR inhibition could be a negative feedback mechanism of the cell to trigger cell survival. Several authors have indicated KRAS as part of a potential mechanism of resistance to EGFR inhibition which makes KRAS a key target oncogene (Zimmermann *et al.*, 2013, Collins and di Magliano, 2014). This gene participates in a large number of signaling pathways including MAPK, ErbB, VEGF and a number of biological processes.

On the structure-activity side, the chemical feature, FF-442307337, is also linked with differential expression of numerous genes. In addition, some of the correlations observed between the pIC50 and gene expression can be attributed to this substructure as the correlation changes after adjusting for this chemical feature (Figure 2.5).

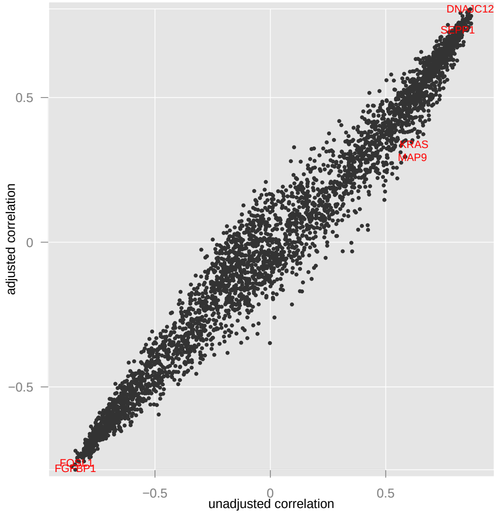


Figure 2.5: Unadjusted vs. Adjusted Correlations. Each point is a gene. Genes that have high correlation but very low adjusted correlation indicates that the fingeprint feature is creating the association.

Next, genes are classified into subgroups based on whether their expression changes are linked with the structure and/or the association remains linear after adjustments for the chemical structure. The number of genes for each subgroups are presented in Table 2.2.

Table 2.2: Results for FF -442307337 (EGFR) at 5% FDR.

		ρ	
		$\neq 0$	0
α	$\neq 0$	396	61
	0	1099	2039

KRAS and FGFBP1 seem to belong to different gene classes. Figure 2.6 shows the 5 most differentially expressed genes with the adjusted association remaining high after adjusting for the chemical structure including the gene FGFBP1 while the estimates for the top 10 genes are given in Table 2.3. The association observed between the gene FGFBP1 and pIC50 is still evident after adjusting for chemical structure (Figure 2.4b). Most of these genes are known to participate in biological processes involving cell proliferation (positive and negative), survival and differentiation. Another set of differentially expressed genes following similar pattern to gene KRAS is presented in Table 2.4 with the visualisation of the top 5 genes in Figure 2.7. For this group, the joint model resulted in very low adjusted correlation (p-adj(ρ)>0.05) between the genes and the activity. Unlike for FGFBP1, the adjustment has a considerable effect in the observed association (from unadjusted correlation, r=0.62 to adjusted correlation, ρ =0.34, see Figure 2.4b).

Table 2.3: List of top 10 differentially expressed genes with high adjusted association (adj-p < 0.05) after adjusting for FF -442307337 (EGFR).

Genes	Effect	p-adj(Effect)	r	ρ	p-adj(ρ)
FOSL1	1.19	0.01	-0.84	-0.76	0.00
FGFBP1	0.79	0.01	-0.84	-0.78	0.00
SEPP1	-0.64	0.01	0.81	0.73	0.00
SCGB2A1	-0.61	0.01	0.83	0.76	0.00
SH2B3	0.61	0.01	-0.79	-0.69	0.00
SLCO4A1	0.60	0.01	-0.79	-0.70	0.00
PHLDA1	0.58	0.01	-0.85	-0.77	0.00
RRM2	0.56	0.02	-0.77	-0.70	0.00
TXNIP	-0.53	0.00	0.75	0.58	0.00
CDC6	0.52	0.01	-0.80	-0.73	0.00

The substructure FF-442307337 is present on majority of the compounds that inhibit cell growth to a lesser extent and FGFBP1. Figure 2.8a shows the chemical structure of

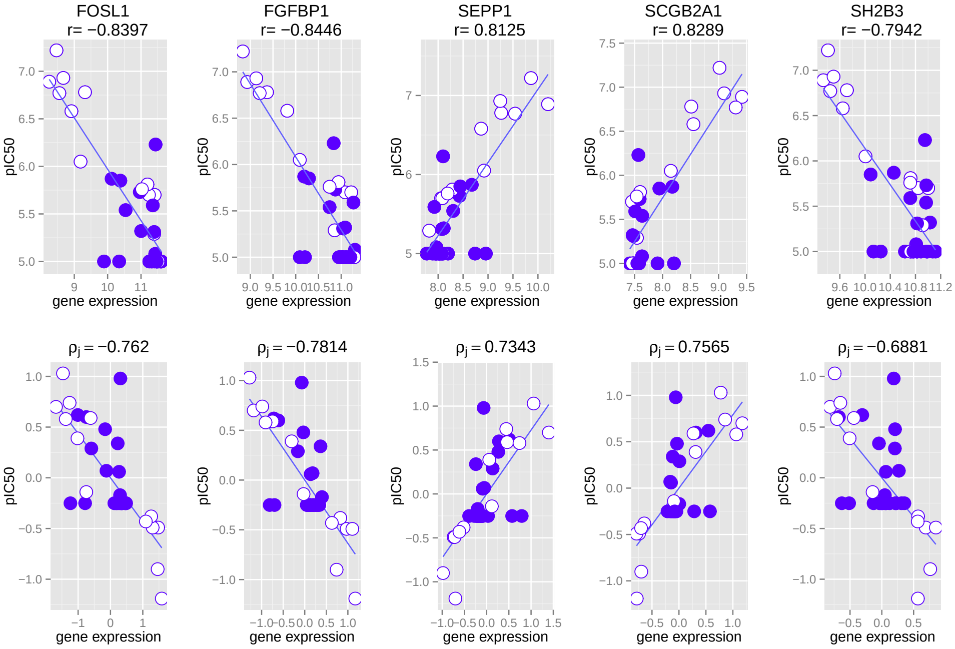


Figure 2.6: Top 5 differentially expressed genes with high adjusted correlation. The correlation between the gene expression and the inhibitory activity against EGFR, given by the pIC50, of the compounds (represented by points in the plots) can be explained by the substructure FF -442307337.

Table 2.4: List of top 10 differentially expressed genes with low adjusted association (adj-p > 0.05) after adjusting for FF -442307337 (EGFR).

Genes	Effect	p-adj(Effect)	r	ρ	p-adj(ρ)
KRAS	-0.30	0.00	0.62	0.34	0.07
MAP9	-0.13	0.00	0.62	0.29	0.13
SMG1	-0.10	0.00	0.62	0.35	0.06
PTER	-0.10	0.00	0.61	0.35	0.06
ODZ3	-0.14	0.01	0.59	0.35	0.06
SCAF11	-0.16	0.00	0.59	0.30	0.12
PCYOX1	-0.23	0.00	0.58	0.30	0.12
PHACTR2	-0.13	0.01	0.58	0.35	0.06
USP3	-0.07	0.01	0.57	0.35	0.06
FBXO21	-0.12	0.00	0.57	0.27	0.16

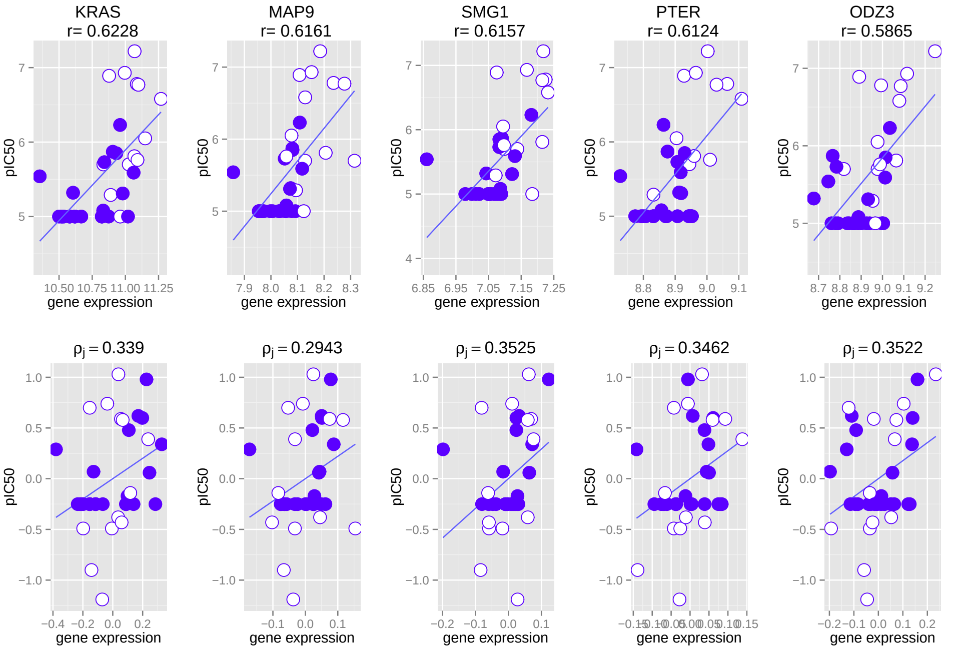


Figure 2.7: Top 5 differentially expressed genes with low adjusted correlation. The correlation between the gene expression and the inhibitory activity against EGFR, given by the pIC50, of the compounds (represented by points in the plots) can be explained by the substructure FF -442307337.

FF-442307337, an oxygen in ortho position of the aniline (highlighted in red). The next compound is very similar to the less potent compound but without FF-442307337 and it is one of the highly potent compounds in this experiment (Figure 2.8b) along with the two reference compounds gefitinib and erlotinib (Figure 2.8c-d). However, other less potent compounds do not have this feature, this substructure is probably not the sole reason for compounds' lower activity.

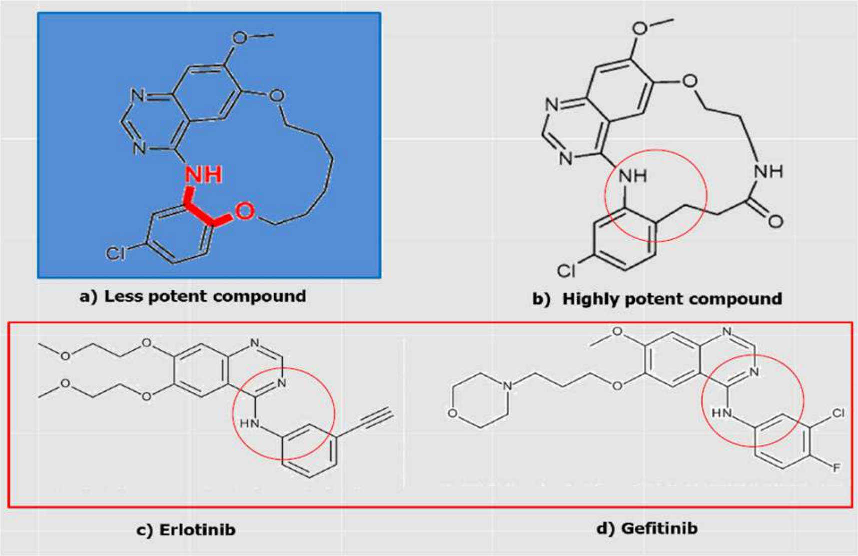
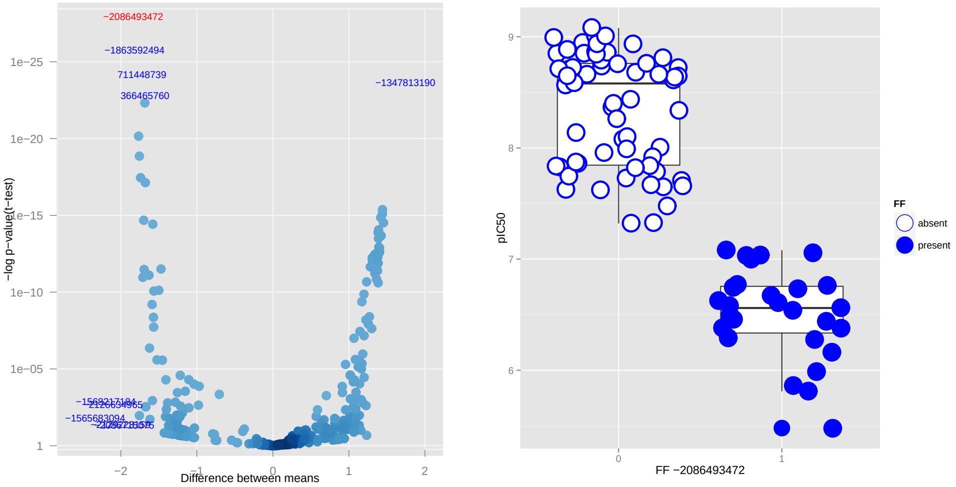


Figure 2.8: Chemical structures of **(a)** identified less potent compound; **(b)** highly potent compound in the EGFR project; and the two reference compounds in this experiment **(c)** erlotinib and **(d)** gefitinib.

2.4.2 Application to the ROS1 Project

This project sought to develop compounds that inhibit ROS1 (reactive oxygen species), known to be overexpressed in several cancer types. Excessive quantities of ROS1 causes oxidative stress that is generally detrimental to cells (Gorrini *et al.*, 2013). Cellular assay for target inhibition showed several compounds with high inhibitory activity. FF -2086493472 came out to be the top fingerprint feature that can well separate the bioactivity of the compounds (Figure 2.9b). Here, the feature can be linked with lower potency since all compounds having the fingerprint feature have lower pIC50 values than those compounds not having the feature.

The joint modeling resulted in identification of genes that are associated with the pIC50. A number of genes showed positive correlation like FNIP1 while TXNRD1 along with other genes showed negative correlation (Figure 2.10a).



(a) Volcano plots to evaluate the capacity of the FFs to differentiate the bioactivity.

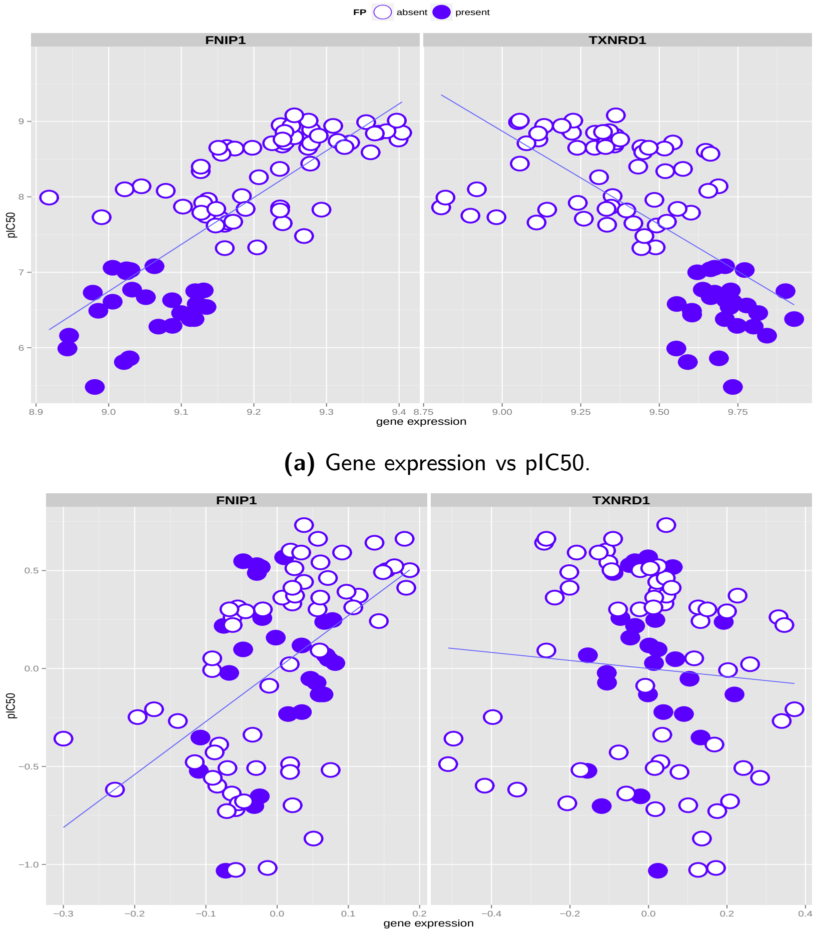
(b) Top chemical substructure for the ROS1 project: FF-2086493472.

Figure 2.9: Plots highlighting the most relevant FF for the ROS1 project.

Interestingly, TXNRD1, a key player in oxidative stress control, is also evaluated as a cancer drug target associated with aggressive tumor growth (Powis and Kirkpatrick, 2007, Eriksson *et al.*, 2009). Elevated levels of this gene in many human cancers contributes to increased proliferation, resistance to cell death and increased angiogenesis. Dai *et al.* (2013) shows that simultaneous inhibition of TXNRD1 and AKT pathways (activated by ROS1) induced robust ROS1 production. Discovering potential inhibitors of this gene could contribute to cancer therapy (Urig and Becker, 2006). In this experiment, compounds with high ROS1 inhibitory activity also shows inhibition of TXNRD1.

The joint model furthermore revealed that potent compounds with lower gene expression effects on TXNRD1 lack FF-2086493472. Moreover, the association between the pIC50 and the expression of gene TXRND1 can be fully explained by the absence/presence of this feature (Figure 2.10b). Table 2.5 shows a set of differentially expressed genes with the same type of association observed between TXNRD1 and pIC50 that disappears after adjusting for FF-2086493472 (Figure 2.12). Figure 2.11 shows how the correlation between the pIC50 and all genes changes after accounting for FF-2086493472.

Little is known about the biology of FNIP1 gene, particularly relating to cancer.



(a) Gene expression vs pIC50.

(b) Gene expression vs pIC50 after adjustment.

Figure 2.10: Two cancer-related genes that correlate with ROS1-inhibition: FNIP1 and TXNRD1. The solid blue points indicate the presence of FF -2086493472.

Table 2.5: List of top 10 differentially expressed genes with low adjusted association (adj-p > 0.05) after adjusting for FF -2086493472 (ROS1).

Genes	Effect	p-adj(Effect)	r	ρ	p-adj(ρ)
TXNRD1	0.39	0.00	-0.65	-0.08	0.54
PFKFB3	0.57	0.00	-0.61	0.00	0.97
SNORD52	0.23	0.00	-0.65	-0.12	0.33
GDF15	-1.09	0.00	0.67	0.21	0.07
ZNF292	-0.30	0.00	0.59	0.01	0.95
CTPS	0.30	0.00	-0.63	-0.16	0.19
KIRREL	0.34	0.00	-0.64	-0.19	0.11
HMGCS1	0.77	0.00	-0.58	-0.04	0.77
TFPI	-0.46	0.00	0.60	0.11	0.37
HIST1H1A	0.49	0.00	-0.52	0.09	0.49

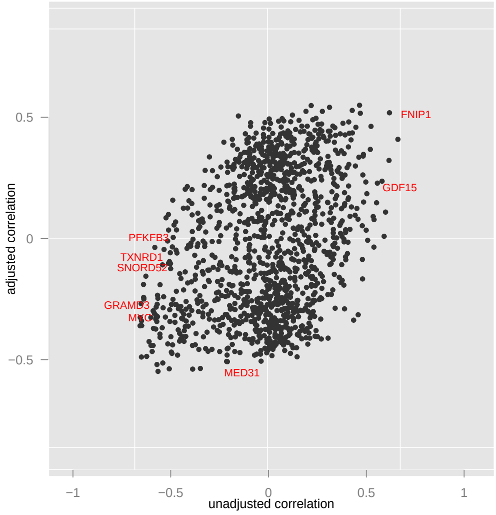


Figure 2.11: Unadjusted vs. Adjusted Correlations. Each point is a gene. Genes that have high correlation but very low adjusted correlation indicates that the fingeprint feature is creating the association.

Hasumi *et al.* (2008) indicated that FNIP1 mRNA was significantly higher in renal cell carcinoma compared to normal kidney. Unlike TXNRD1, the correlation between the bioassay and the gene remains moderately strong after adjusting for fingerprint feature. This implies that FNIP1 remains to be linearly associated with the efficacy data independent of this structural feature. Table 2.6 presents other 9 genes showing the same type of asocciation with pIC50 as FNIP1 (Figure 2.13). The number of genes in each subclass is given in Table 2.7.

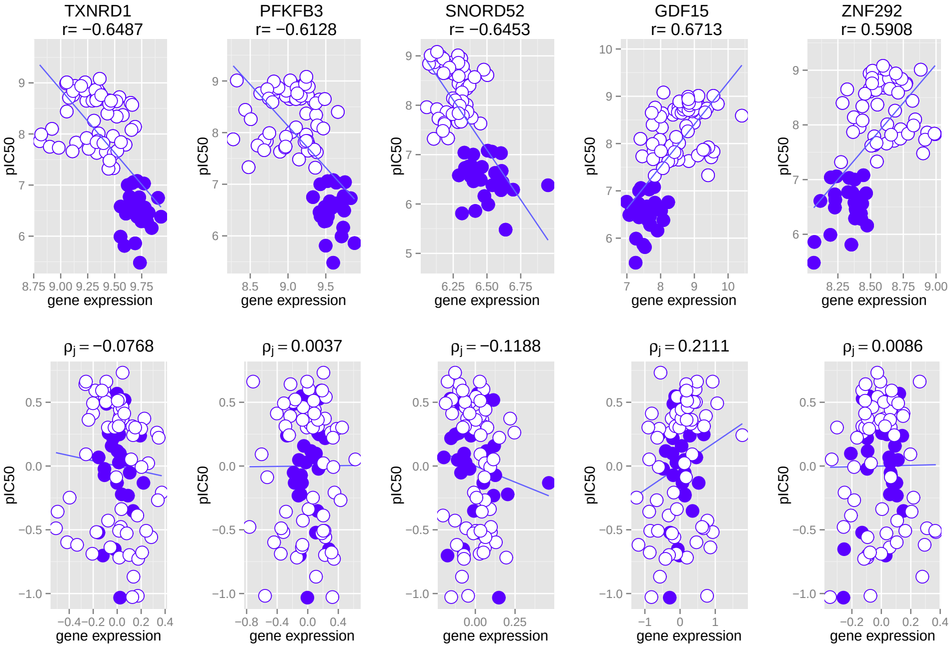


Figure 2.12: Top 5 differentially expressed genes with low adjusted correlation. The correlation between the gene expression and the inhibitory activity against ROS1, given by the pIC50, of the compounds (represented by points in the plots) can be explained by the substructure FF -2086493472.

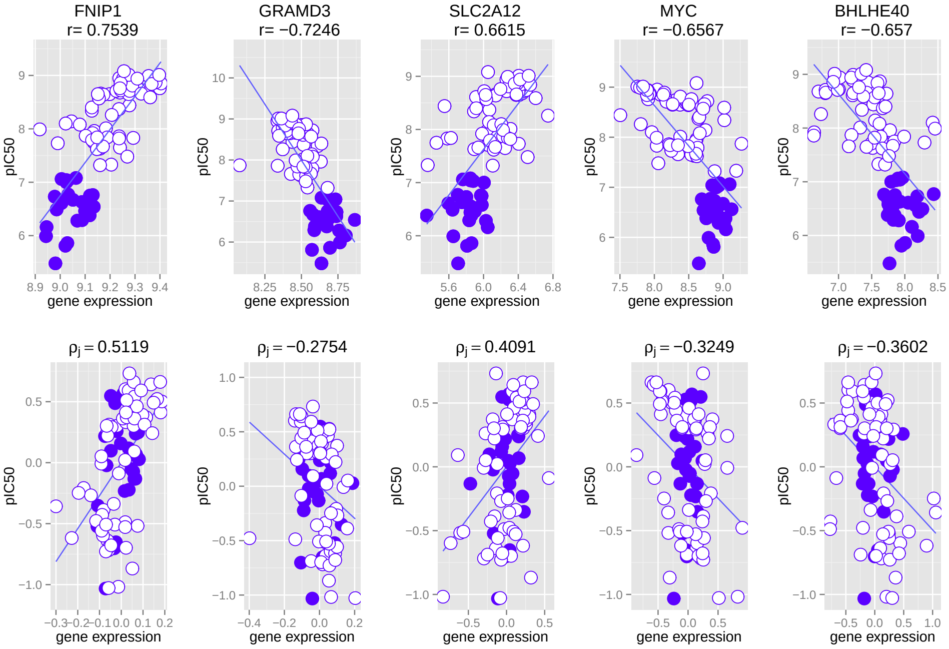


Figure 2.13: Top 5 significantly differentially expressed genes with significant adjusted correlation.

Table 2.6: List of top 10 differentially expressed genes with high adjusted association (adj-p < 0.05) after adjusting for FF -2086493472 (ROS1).

Genes	Effect	p-adj(Effect)	r	ρ	p-adj(ρ)
FNIP1	-0.16	0.00	0.75	0.51	0.00
GRAMD3	0.20	0.00	-0.72	-0.27	0.02
SLC2A12	-0.37	0.00	0.66	0.41	0.00
MYC	0.52	0.00	-0.66	-0.33	0.00
BHLHE40	0.55	0.00	-0.66	-0.36	0.00
TGFB2	0.57	0.00	-0.66	-0.33	0.00
TMEM177	0.13	0.00	-0.65	-0.27	0.02
SNORD4B	0.25	0.00	-0.65	-0.49	0.00
TNFRSF12A	0.82	0.00	-0.65	-0.34	0.00
SNORD44	0.19	0.00	-0.65	-0.36	0.00

Table 2.7: Results for FF -2086493472 (ROS) at 5% FDR.

		ρ	
		$\neq 0$	0
α	$\neq 0$	139	239
	0	382	529

2.5 Discussion

The joint modeling framework facilitates an integration of multi-source data in early drug development phase, particularly, the associations between chemical structures, biological activity and gene expressions in order to identify potential leads in early phase of drug discovery alongside with the development of genomic biomarkers for efficacy of compounds. Selecting and evaluating biomarkers in drug discovery and early drug development can substantially shorten development time or the time to reach a critical decision point, such as candidate selection, in exploratory drug development.

The joint modeling approach, although implemented using only one feature at a time for every data source, facilitates the extraction of valuable insights into the associations between chemical structures and mechanism of actions. Although, we focused in this chapter on one fingerprint feature and on-target assay per project, this method can easily be run in loops. In a pharmaceutical pipeline implementation, this model can be applied to all or a defined set 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.

The joint model presented in this chapter is not restricted to normal distribution. In many applications, linear models are frequently used for microarray data analysis with the convenience of model parameters interpretability despite its parametric assumptions (Scholtens and von Heydebreck, 2005). It has been observed that the use of log-transformed and normalized gene expression data are fairly robust to violations of normality assumption (Trabzuni and Thomson, 2014, Smyth, 2004, Smyth *et al.*, 2005). The top genes ranked by their resulting p-values from the test for differential expression should be cautiously interpreted. It is recommended to plot the top genes versus the bioassay per subclass to further evaluate whether differential expression is only due to outlying observations.

If the interest is on identifying a set of multiple genes that could best predict the response, other linear modeling techniques such as lasso and elastic net among others may be considered. Further, each group of genes identified by the model can be summarized following the techniques presented by Tilahun *et al.* (2010) to come up with a gene signature that can predict efficacy of compounds to inhibit cancer signaling pathway.

Possible extensions within the joint modeling framework accounting for the dependency among genes will be outlined in the next chapter. It incorporates the idea of penalized likelihood in the selection of genes and estimation of model parameters. Accounting, however, for the dependency among the fingerprint features is not meaningful. The binary representation of the chemical structure is a very simplified representation of the molecules and interpretation is very challenging for medicinal chemists. Chemical modification of compounds for improved activity may involve only adjustment of one substructure.

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. Also, 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 could still be very helpful in extracting relevant information from the high dimensional and complex microarray and chemical data with the hope of providing an answer to the relevant research questions posed by pharmaceutical companies.

In conclusion, the gene-specific joint model is a simple approach that is easy to interpret and to integrate within a drug development pipeline.

Chapter 3

From Joint Modeling to Path Analysis and Conditional Modeling

3.1 Introduction

The modeling approach behind the joint model presented in Chapter 2 is focused on modeling the correlation between gene expression and bioactivity data accounting for the effect of the chemical structure using gene-specific models. In this chapter, we discuss related models within the QSTAR modeling framework that can be used to reveal the relationship between the three data sources. Here, we formulate the gene-specific QSTAR setting in the context of path analysis and conditional models. The modeling approaches discussed in this chapter is new within the drug discovery framework. The aim of this chapter is to provide a new set of modeling tools that facilitates better understanding of the mechanism of action of a candidate compound(s). Our aim in this chapter is to establish the relationship between the chemical and the biological dimensions taking into account that the biological dimension is measured by two components: bioactivity data and gene expression data. In Section 3.2 we present the path analysis model followed by a discussion on its relationship with the joint model in Section 3.3. In Section 3.4, we apply the path analysis model using the same datasets presented in Chapter 2. The discussion then follows in Section 3.6.

3.2 The Path Analysis Model

The structural equations modeling (SEM) technique involves multiple-equation models that represent a putative causal relationship among various variables. The path analysis model corresponds to the aspect of SEM that concerns the relations among independent (exogenous) and dependent (endogenous) variables, either observed or latent. In this chapter we include only observed data in the model. In the joint model, we model the adjusted covariances or correlations between the gene expression and bioactivity which represent relationships without an explicitly defined causal direction. The SEM approach allows to test the researcher's hypotheses of causality within a system. Two important aspects of the procedures involved are: (1) that the causal processes under study are represented by a series of structural (i.e. regression) equations, and (2) that these structural relations can be modelled pictorially and tested statistically (Byrne, 1994). In contrast with multiple regressions, ANOVA or MANOVA, the model may include directional relations among dependent variables. However, SEM cannot test the directionality of the relationships. The directions of arrows in a structural equation model represent the researcher's choice of variables and pathways which may not always represent the patterns that have been observed in nature.

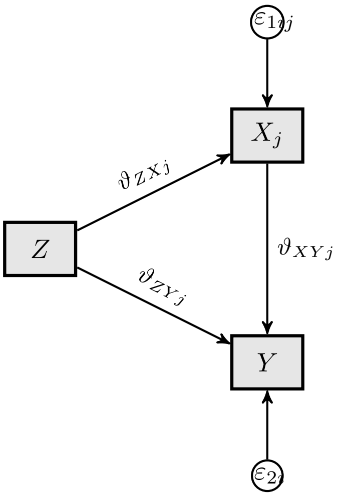


Figure 3.1: The QSTAR path diagram.

The path diagram in Figure 3.1 corresponds to the QSTAR setting which shows the structural relations of the triplet (X, Y, Z) denoting gene expression, bioassay read-out and chemical structure, respectively. These structural relations formulated as a set of two equations given by

$$\begin{aligned} X_{ji} &= \vartheta_{ZXj} Z_i + \varepsilon_{1ij}, \\ Y_i &= \vartheta_{XYj} X_{ji} + \vartheta_{ZYj} Z_i + \varepsilon_{2ij}. \end{aligned} \tag{3.1}$$

Here, i is the index for the compound $i = 1, \ldots, n$ and X_{ji} is the expression level of the jth gene for the ith compound, $j = 1, \ldots, m$. An important property of the model specified in (3.1) is that all paths are estimated and evaluated simultaneously (Hoyle, 1995). Given that all these relations are significant, then it states that the observed variable Y is causally determined by X_j and Z but Y is not perfectly explained by X_j and Z, as reflected by the error $\varepsilon_2 j$. Similarly, X_j is causally determined by the observed variable Z with an error of $\varepsilon_1 j$. Note that, in contrast with the joint model specified in (2.1), the error terms in (3.1) are assumed to be uncorrelated since the association between X_j and Y is modeled via the parameter ϑ_{XYj}. Note that the path analysis model in (3.1) and the joint model specified in (2.1) are closely related, this point is further discussed in Section 3.3. In SEM terminology, we say that Z has a *direct* effect on Y quantified by the parameter ϑ_{ZYj} and also has an *indirect* causal relationship on Y through X_j which can be estimated by $\vartheta_{ZXj} \times \vartheta_{XYj}$. The sum of the direct and indirect effects gives the *total* effect of Z on Y.

The statistical theory underlying SEM is grounded in covariance structure analysis, and the study of covariance matrices is preferred when using this technique (Cudeck, 1989). Three key requirements for SEM are as follows: thorough knowledge of the theory; adequate assessment of statistical criteria; and parsimony (ability to predict the greatest amount of variance in the outcome variable or variables using the smallest number of predictor variables).

Estimation involves finding estimates of the parameters in the model that generate an estimated covariance matrix Σ (model-based covariance matrix) that is as close as possible to the sample (observed) covariance matrix S. Various methods can be used to generate Σ. Jöreskog (1973) proposed the use of ML to test structural equation models and this estimator, which assumes multivariate normality and thus normally distributed errors, remains the most widely used (Bollen, 1989, Chou and Bentler, 1995). A Monte Carlo simulation study suggests that the ML estimator is not biased when small samples are considered, provided that distributions are multivariate normal (Curan *et al.*, 1996).

3.3 Path Analysis and Joint Modeling

The model formulated in (3.1) can be fitted per gene for a given fingerprint feature and bioactivity variable. However, not all genes have the same structural relationship with the bioactivity and fingerprint feature. Similar to the joint model, where the genes can be classified into 4 types based on inference for ρ and α, the path analysis model allows to test the significance of each of the paths link between the triplets (X, Y, Z). However, in contrast to the joint model that treats the association as symmetric, the path analysis model assumes a direction in causality.

While the joint model provides the association between genes and bioactivity accounting for the chemical structure, the path analysis model can provide extra information by decomposing the total effect of Z on Y into direct and indirect effects. The direct effect, ϑ_{ZYj}, measures the direct impact of the FF on the bioactivity when not mediated through a gene. The effect is said to be indirect when the FF has an effect on gene expression and this consequently affects bioactivity. This effect can be estimated using the parameters ϑ_{ZXj} and ϑ_{XYj}. Different structural relationships are visualized in Table 3.1. Within the context of the QSTAR framework, the last two structural relationships, d and e, are not of interest in drug discovery studies since they represent a scenario in which the gene expression is correlated with the bioactivity variable regardless of the chemical structure (scenario d) or is not correlated with the bioactivity variable (scenario e). Note that both scenarios can be considered as relevant if the QSAR framework is used since under both scenarios the FF has an influence on the bioactivity variable. The first three scenarios are of primary interest in drug discovery studies since all three can be used to established the relationship between chemical and biological dimensions. The first scenario illustrates a structural relationship in which both direct and indirect effects are present. Under scenario b, the direct effect is absorbed completely by the indirect effect due to the high correlation between the gene expression and the bioactivity variable. The conditional independence relationship, illustrated in scenario c, represents a scenario in which the association between the gene expression and the bioactivity variables is only due to the chemical structure and conditioned on the FF level, the association disappears.

Of course, the patterns mentioned above are closely related to the association patterns discussed in Chapter 2. In Table 3.2 we presented several hypothetical numerical examples to compare between the two models. Notice that the column of ϑ_{ZXj} is identical to the column of α_j of the joint model since they both estimate the fingerprint feature effect on the jth gene, X_j. Hence, in terms of differentially expressed genes, the two models provide the same results. The total effect is the sum of the direct effect (ϑ_{ZYj}) and indirect effect $(\vartheta_{ZXYj} = \vartheta_{ZXj} \times \vartheta_{XYj})$ of Z on Y. The indirect effect is insignificant whenever one of its components, ϑ_{XYj} or ϑ_{ZXj} is not significant. Here, using path analysis, we decompose the total effect of Z on Y to take into account the effect of gene as an intervening factor.

The adjusted association and the indirect effect can be used to identify genes that are associated to the bioactivity. A high indirect effect implies a high adjusted association but

Table 3.1: The QSTAR path diagram.

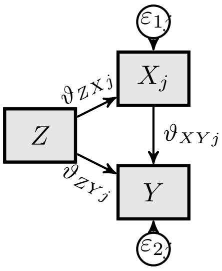
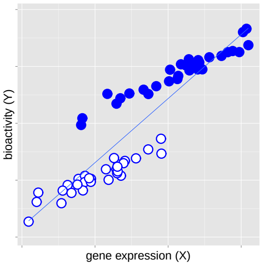
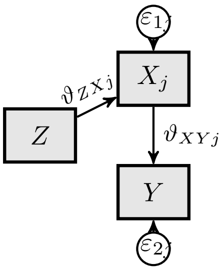
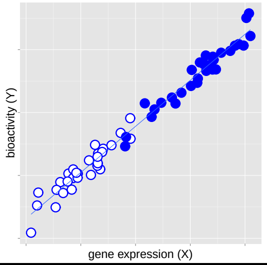
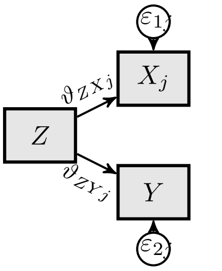
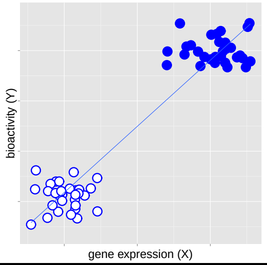
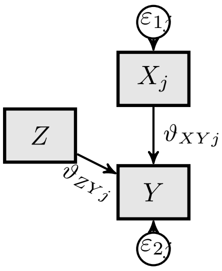
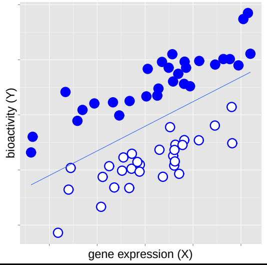
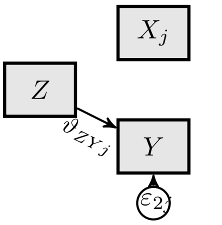
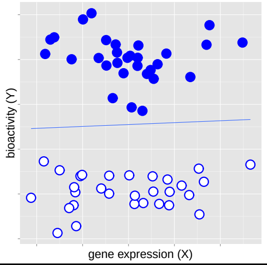
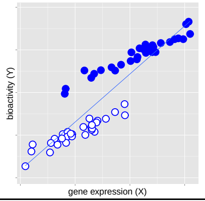
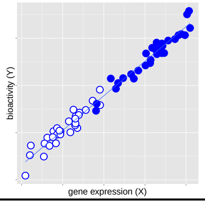
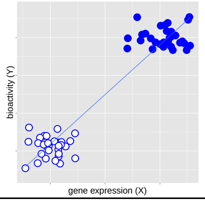
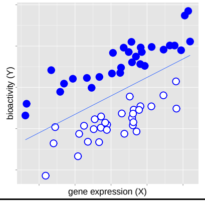
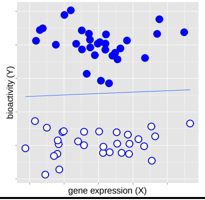
Type	Path Diagram	Gene Profile
(a) The total effect of Z on Y can be estimated via its direct and indirect effects.	 ε_{1j} X_j ϑ_{ZXj} Z ϑ_{XYj} ϑ_{ZYj} Y ε_{2j}	
(b) Z is indirectly related to Y via X.	 ε_{1j} X_j ϑ_{ZXj} Z ϑ_{XYj} Y ε_{2j}	
(c) Conditional independence. Z has direct effects on X and Y. X is not directly related to Y given Z.	 ε_{1j} X_j ϑ_{ZXj} Z ϑ_{ZYj} Y ε_{2j}	
(d) Z is directly related to Y but not to X. X and Y are directly related.	 ε_{1j} X_j Z ϑ_{XYj} ϑ_{ZYj} Y ε_{2j}	
(e) Z is directly related to Y.	 X_j Z ϑ_{ZYj} Y ε_{2j}	

Table 3.2: Table of path analysis model and joint model standardised parameter estimates with their corresponding standard error inside the parenthesis. ns means not significant path link at 0.05 level of significance.

Gene Profile	Path Analysis Model					Joint Model		
	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	$total_j$	α_j	ρ_j	r_j
(a) 	0.513 (0.021)	0.825 (0.073)	0.529 (0.021)	0.436 (0.042)	0.950 (0.042)	0.825 (0.073)	0.955	0.953
(b) 	-0.037^{ns} (0.041)	0.886 (0.060)	1.021 (0.041)	0.905 (0.071)	0.869 (0.064)	0.886 (0.060)	0.954	0.989
(c) 	0.966 (0.088)	0.984 (0.023)	0.019^{ns} (0.088)	0.018^{ns} (0.085)	0.966 (0.034)	0.984 (0.023)	0.0272	0.952
(d) 	0.877 (0.043)	0.105^{ns} (0.128)	0.271 (0.043)	0.028^{ns} (0.035)	0.906 (0.055)	0.105^{ns} (0.128)	0.636	0.363
(e) 	0.982 (0.021)	0.105^{ns} (0.128)	-0.039^{ns} (0.021)	0.001^{ns} (0.002)	0.984 (0.023)	0.105^{ns} (0.131)	0.023	0.108

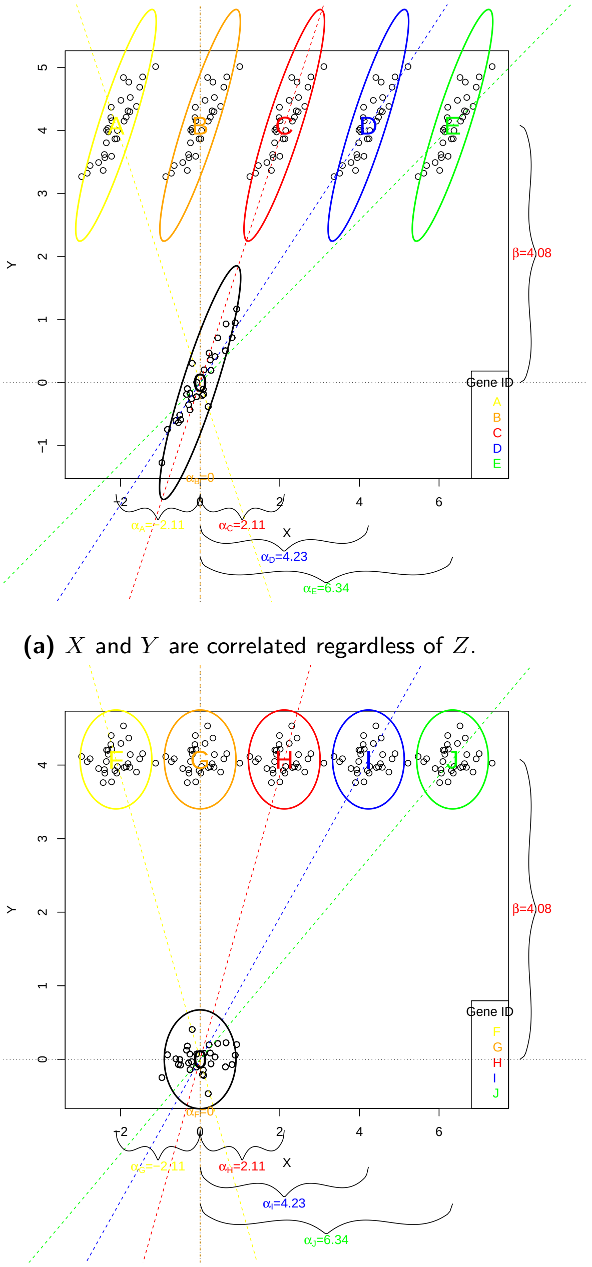
the reverse does not hold. A gene with high adjusted association may have low indirect effect. Since the indirect effect, $\vartheta_{ZXYj} = \vartheta_{ZXj} * \vartheta_{XYj}$, is composed of two components, it will only be maximised if both ϑ_{ZXj} and ϑ_{XYj} are high. The gene in Table 3.2d is not differentially expressed having $\alpha_j \simeq 0$ and $\vartheta_{ZXj} \simeq 0$ thus getting a low indirect effect estimate yet a high adjusted association.

3.3.1 Relationship Between Path Analysis and Joint Model

As mentioned previously, within the context of the QSTAR framework, the first three structural relationships presented in Table 3.1 are of interest in drug discovery. Note that all three of them show the following patterns: (1) X and Y are linearly associated; and (2) Z has an effect on both X and Y. In this section, we focus on these three scenarios by illustrating their differences when varying the effect of Z on X holding the effect of Z on Y constant as shown in Figure 3.2. In addition, Figure 3.3 shows the settings when we fix the effect of Z on X and changing the effect of Z on Y. There are a total of 20 settings (represented by 20 letters in Figures 3.2 and 3.3). For each of these settings, we obtain the parameter estimates for the joint model and the path analysis model given in Table 3.3 with their corresponding standard errors in Table A1. The slopes of the lines represent their respective relative effects (Buyse and Molenberghs, 1998) given by β/α.

Inspecting, Table 3.3, it can be readily seen that for all settings the columns of the total effect is equal to the column of β. This is expected since they both quantify the full effect of the chemical structure upon the bioactivity data. The same holds for the estimates of α and ϑ_{ZX} since they both quantify the full effect of the chemical structure upon the gene expression. Moreover, for every group of 5 settings (as divided by the horizonal line), the columns ϑ_{XY} and ρ are constant. This is because within each group, the same conditional relationship exists between X and Y given Z. A non-significant ϑ_{XY} would imply conditional independence of X and Y and a very low adjusted association ($\rho \approx 0$) whether or not the gene/bioassay is differentially expressed (settings F-O).

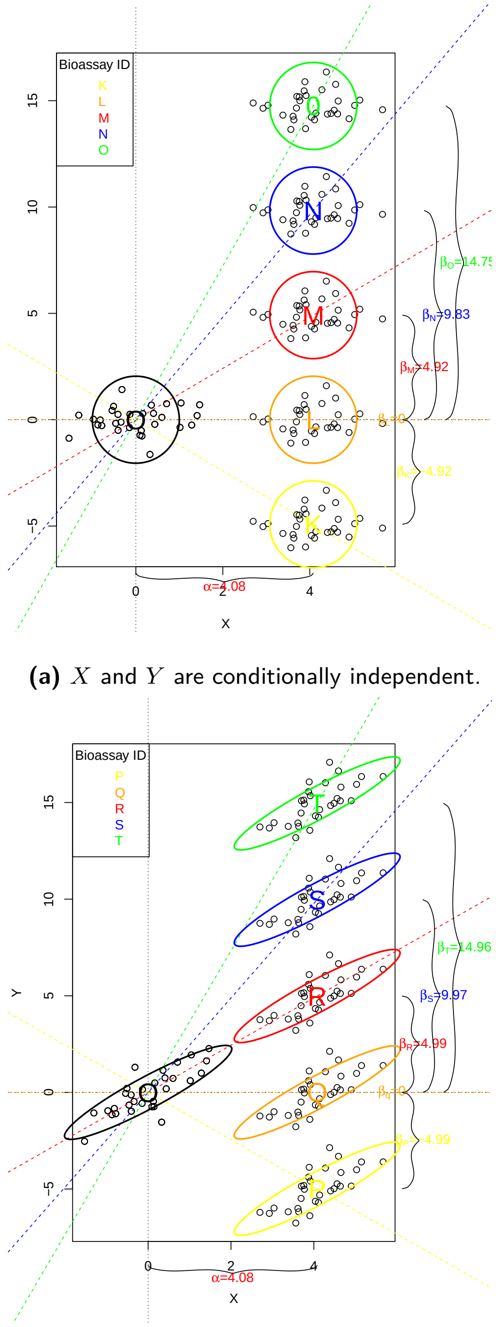
Although settings C, D, and E have all significant estimates for the joint model, setting D has a non-significant direct effect ϑ_{ZY}. This is because the total effect of Z on Y is fully captured by X ($\vartheta_{ZXY} \approx$ total effect). In setting D, we have an indirect effect, $\vartheta_{ZXY} = 4.48$, and the total effect$= 4.08$. Recall that the total effect is simply the sum of the direct and indirect effect. Thus, in this case, we get a very low and non-significant direct effect. The same relationship holds for setting R. Note that, settings C, D and E will be classified in the same group by the joint model which are all differentially expressed genes having high adjusted association, setting D will be highlighted by the path analysis model to be the most desirable one. Using the joint model parameters, D, I, M and R are the settings where α and β are roughly equal, thereby giving a relative effect close to



(a) X and Y are correlated regardless of Z.

(b) X and Y are conditionally independent.

Figure 3.2: Illustrative example I. Gene expression versus a response. The genes (indicated by a letter in the plot) have varying levels of FF effect.



(a) X and Y are conditionally independent.

(b) X and Y are correlated regardless of Z.

Figure 3.3: Illustrative example II. Gene expression versus a response. The bioassays (indicated by a letter in the plot) have varying level of fingerprint feature effect.

Table 3.3: Parameter estimates for the illustrative example presented in Figures 3.2 and 3.3. Standard errors are presented in Table A1 in the appendix. ns means not significant path link at 0.05 level of significance.

Setting	α	β	ρ	r	ϑ_{ZY}	ϑ_{ZX}	ϑ_{XY}	ϑ_{ZXY}	Total
A	-2.11	4.08	0.91	-0.80	6.32	-2.11	1.06	-2.24	4.08
B	0.00^{ns}	4.08	0.91	0.23	4.08	0.00^{ns}	1.06	0.00^{ns}	4.08
C	2.11	4.08	0.91	0.98	1.84	2.11	1.06	2.24	4.08
D	4.23	4.08	0.91	0.99	-0.39^{ns}	4.22	1.06	4.48	4.08
E	6.34	4.08	0.91	0.99	-2.63	6.34	1.06	6.71	4.08
F	-2.11	4.08	0.12	-0.91	4.18	-2.11	0.05^{ns}	-0.10^{ns}	4.08
G	-0.00^{ns}	4.08	0.12	0.01	4.08	0.00^{ns}	0.05^{ns}	0.00^{ns}	4.08
H	2.11	4.08	0.12	0.92	3.98	2.11	0.05^{ns}	0.10^{ns}	4.08
I	4.23	4.08	0.12	0.98	3.88	4.22	0.05^{ns}	0.20^{ns}	4.08
J	6.34	4.08	0.12	0.99	3.77	6.34	0.05^{ns}	0.30^{ns}	4.08
K	4.08	-4.92	0.12	-0.91	-5.34	4.08	0.10^{ns}	0.42^{ns}	-4.92
L	4.08	-0.00^{ns}	0.12	0.04	-0.42^{ns}	4.08	0.10^{ns}	0.42^{ns}	-0.00^{ns}
M	4.08	4.92	0.12	0.93	4.49	4.08	0.10^{ns}	0.42^{ns}	4.92
N	4.08	9.83	0.12	0.94	9.41	4.08	0.10^{ns}	0.42^{ns}	9.83
O	4.08	14.75	0.12	0.94	14.33	4.08	0.10^{ns}	0.42^{ns}	14.75
P	4.08	-4.99	0.76	-0.77	-9.55	4.08	1.12	4.57	-4.99
Q	4.08	0.00^{ns}	0.76	0.25	-4.57	4.08	1.12	4.57	0.00^{ns}
R	4.08	4.99	0.76	0.97	0.42^{ns}	4.08	1.12	4.57	4.99
S	4.08	9.97	0.76	0.97	5.41	4.08	1.12	4.57	9.97
T	4.08	14.96	0.76	0.97	10.39	4.08	1.12	4.57	14.96

one, that is : $RE = \beta/\alpha = 1.$

For the last 10 settings (K-T), within the group of 5, they only vary with respect to the total effect (or β) and the direct effect since the effect of Z on X and the conditional association between X and Y are fixed to be the same within the group, hence, a constant indirect effect.

3.4 Application to the Data

3.4.1 The EGFR Project

As mentioned in Section 2.4.1, two genes were identified based on inference for ρ and the FF effect on the gene, α. The two genes are highlighted in Figures 3.4a and 3.4b showing that they belong to different gene classes.

Although, the joint model and path analysis model are specified differently and are used in different context, their parameter estimates are expected to be related. In fact, the models would give the same set of differentially expressed genes since the estimates of

α and ϑ_{ZX} are equal (Figure 3.5a). In this project, many of the genes with high adjusted association has also high indirect effect (Figure 3.5b). Figure 3.6 highlights one gene per group along with the path parameter estimates. The top 10 genes for each group are given in Table 3.4.

Table 3.4: The EGFR Project. Classification of genes based on the significance of path(s).

(a) Top genes having all paths significant.

Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
FOSL1	**-0.42**	1.19	-0.43	**-0.52**	-0.94
FGFBP1	**-0.44**	0.79	-0.64	**-0.50**	-0.94
SEPP1	**-0.48**	-0.64	0.73	**-0.46**	-0.94
SCGB2A1	**-0.46**	-0.61	0.79	**-0.48**	-0.94
SH2B3	**-0.47**	0.61	-0.77	**-0.47**	-0.94
SLCO4A1	**-0.50**	0.60	-0.73	**-0.44**	-0.94
PHLDA1	**-0.40**	0.57	-0.94	**-0.54**	-0.94
RRM2	**-0.55**	0.56	-0.68	**-0.39**	-0.94
TXNIP	**-0.49**	-0.53	0.85	**-0.45**	-0.94
CDC6	**-0.50**	0.52	-0.84	**-0.44**	-0.94

(b) Top genes with high indirect effects.

Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
LIMS1	-0.14	-0.18	4.34	**-0.80**	-0.94
LOC1005	-0.15	-0.38	2.09	**-0.79**	-0.94
PNISR	-0.15	-0.24	3.32	**-0.79**	-0.94
NAA35	-0.19	-0.32	2.30	**-0.74**	-0.94
OXR1	-0.20	-0.16	4.58	**-0.74**	-0.94
BPTF	-0.21	-0.18	4.03	**-0.73**	-0.94
WSB1	-0.23	-0.29	2.44	**-0.70**	-0.94
MYO6	-0.24	-0.14	5.03	**-0.70**	-0.94
CEP68	-0.25	-0.13	5.30	**-0.69**	-0.94
KCNN4	-0.27	0.29	-2.31	**-0.67**	-0.94

(c) Top differentially expressed genes with direct effect only.

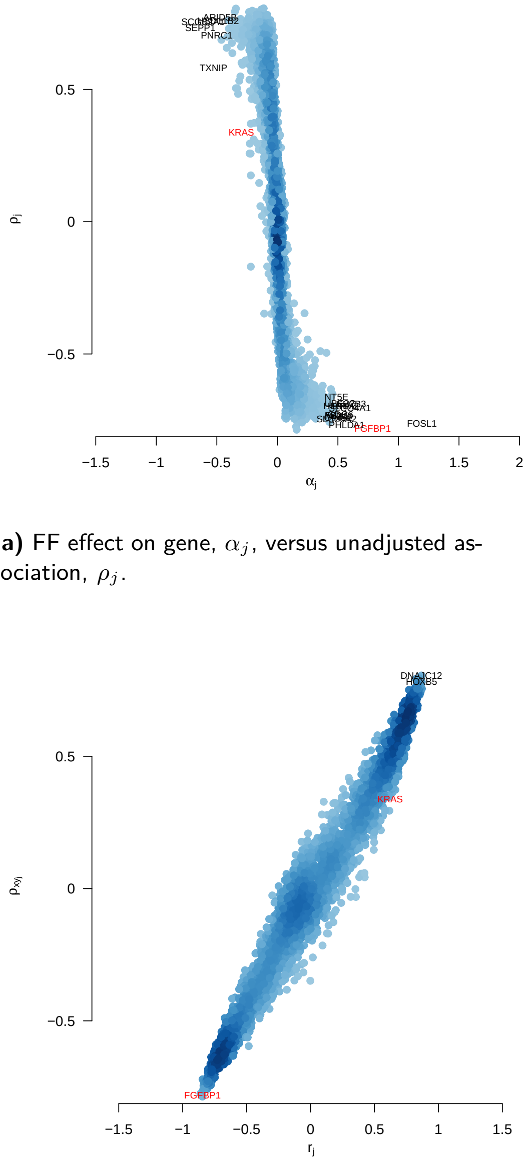
Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
KRAS	**-0.64**	**-0.30**	1.00	-0.30	-0.94
ANKRD11	**-0.76**	**-0.23**	0.80	-0.18	-0.94
PCYOX1	**-0.70**	**-0.23**	1.05	-0.24	-0.94
BRD2	**-0.74**	**-0.22**	0.89	-0.20	-0.94
EIF3A	**-0.81**	**-0.22**	0.61	-0.13	-0.94
CUEDC1	**-0.72**	**-0.20**	1.12	-0.22	-0.94
SCAF11	**-0.70**	**-0.16**	1.48	-0.24	-0.94
SPIN1	**-0.76**	**-0.15**	1.15	-0.18	-0.94
COL4A3BP	**-0.68**	**-0.15**	1.72	-0.26	-0.94
PDP1	**-0.60**	**-0.15**	2.25	-0.34	-0.94

(d) Genes with high direct effect and related to bioactivity.

Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
FLRT3	**-0.95**	-0.01	**-0.83**	0.01	-0.94
KIAA0754	**-0.93**	0.00	**-4.05**	-0.01	-0.94
CKAP2	**-0.92**	0.00	**-4.17**	-0.02	-0.94
TFAM	**-0.92**	0.00	**-3.69**	-0.02	-0.94
PLK2	**-0.92**	0.01	**-2.48**	-0.02	-0.94
EIF4G1	**-0.91**	0.01	**-2.35**	-0.03	-0.94
TMEM38B	**-0.91**	0.01	**-4.66**	-0.03	-0.94
NUDT1	**-0.91**	0.01	**-3.67**	-0.03	-0.94
LRRC16A	**-0.90**	-0.01	**4.46**	-0.04	-0.94
DHRS2	**-0.90**	0.02	**-1.58**	-0.04	-0.94

(e) Genes with high direct effect and not related to bioactivity.

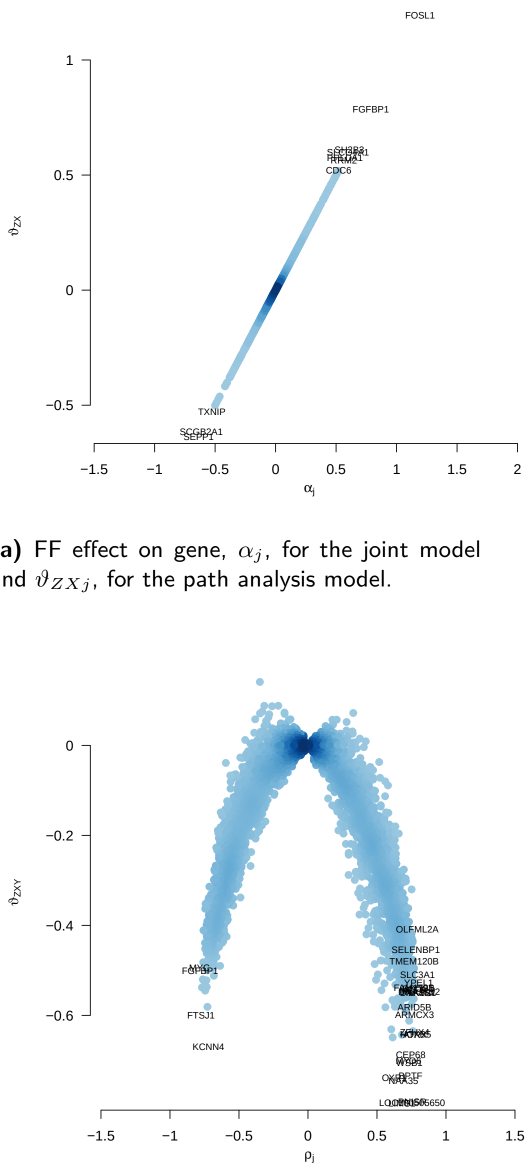
Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
PSMD1	**-1.08**	-0.05	-2.84	0.14	-0.94
HNRNPL	**-1.03**	-0.02	-4.89	0.09	-0.94
MFHAS1	**-1.03**	-0.05	-1.78	0.09	-0.94
UBE2G2	**-1.03**	-0.06	-1.55	0.09	-0.94
ABHD2	**-1.01**	-0.02	-3.16	0.07	-0.94
NAA15	**-1.01**	-0.22	-0.33	0.07	-0.94
ZNF835	**-1.01**	0.02	3.31	0.07	-0.94
TOP2A	**-1.01**	-0.11	-0.65	0.07	-0.94
RSRC1	**-1.01**	-0.08	-0.83	0.07	-0.94
MAP3K7	**-1.01**	-0.02	-2.91	0.07	-0.94



(a) FF effect on gene, α_j, versus unadjusted association, ρ_j.

(b) Unadjusted correlation, r_j versus unadjusted association, ρ_j.

Figure 3.4: Plots highlighting the most relevant genes for the EGFR project as discussed in Chapter 2.



(a) FF effect on gene, α_j, for the joint model and ϑ_{ZXj}, for the path analysis model.

(b) Adjusted association, ρ_j versus the indirect effect, $\vartheta_{ZXYj} = \vartheta_{ZXj} * \vartheta_{XYj}$.

Figure 3.5: The EGFR project. Parameter estimates of the joint model and path analysis model.

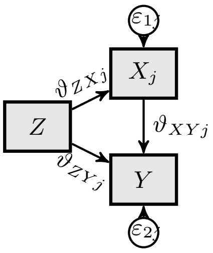
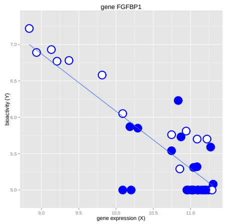
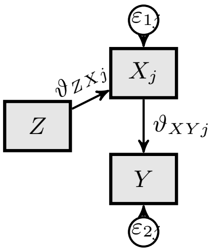
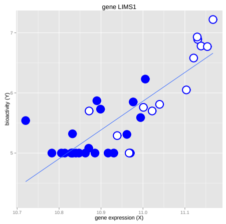
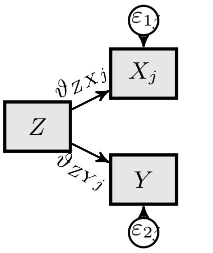
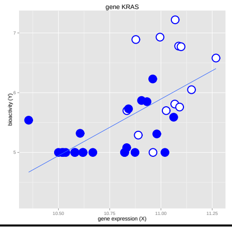
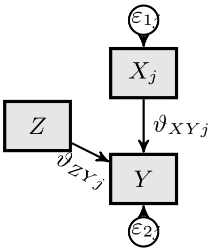
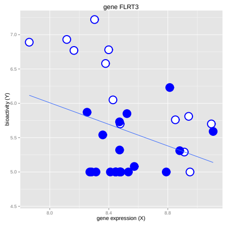
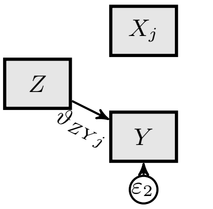
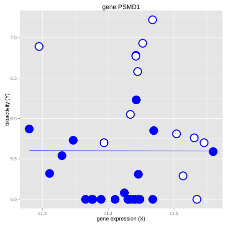
Type	Path Diagram	EGFR genes	Estimates
(a) The effect of Z on Y can be estimated via its direct and indirect effects.	 ε_1 X_j ϑ_{ZXj} Z ϑ_{XYj} ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.44$, $\vartheta_{ZXj} = 0.79$, $\vartheta_{XYj} = -0.64$, indirect$= -0.50$, total$= -0.94$
(b) Z is indirectly related to Y via Z.	 ε_1 X_j ϑ_{ZXj} Z ϑ_{XYj} Y ε_2		$\vartheta_{ZYj} = -0.14$, $\vartheta_{ZXj} = -0.18$, $\vartheta_{XYj} = 4.34$, indirect$= -0.80$, total$= -0.94$
(c) Z has direct effects on X and Y. X is not directly related to Y given Z.	 ε_1 X_j ϑ_{ZXj} Z ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.64$, $\vartheta_{ZXj} = -0.30$, $\vartheta_{XYj} = 1.00$, indirect$= -0.30$, total$= -0.94$
(d) Z is directly related to Y but not to X. X and Y are directly related given Z.	 ε_1 X_j Z ϑ_{XYj} ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.95$, $\vartheta_{ZXj} = -0.01$, $\vartheta_{XYj} = -0.83$, indirect$= 0.01$, total$= -0.94$
(e) Z is directly related to Y.			$\vartheta_{ZYj} = -1.408$, $\vartheta_{ZXj} = 0.05$, $\vartheta_{XYj} = -2.84$, indirect$= -0.14$, total$= -0.94$

Figure 3.6: The EGFR Project. Genes discovered from testing path links and classified according to the the corresponding diagram.

3.4.2 The ROS1 Project

Recall that in Chapter 2, the genes TXNRD1 and FNIP1 from the ROS1 project are highlighted to be exhibiting different types of association with the response when adjusting for chemical structure. Both genes are differentially expressed and correlated to bioactivity having a high unadjusted association, r_j (Figures 3.7a and 3.7b). However adjusting for the effect of a fingerprint feature, only FNIP1 remains to be correlated with the bioactivity. In contrast with the EGFR project, for the ROS1 project we see that many genes with high adjusted association have low indirect effect (Figure 3.8b). The results of the path analysis for this project reveals gene classes depending on which path is significant. Figure 3.9 highlights one gene per group along with the path parameter estimates. The top 10 genes for each group are given in Table 3.5.

Table 3.5: The ROS1 project: Classification of genes based on the significance of path(s).

(a) Top genes having all paths significant.

Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
FNIP1	**-1.39**	-0.16	2.70	**-0.45**	-1.84
MRPL17	**-1.42**	0.10	-4.34	**-0.41**	-1.84
AP1S3	**-1.47**	0.15	-2.38	**-0.37**	-1.84
LANCL1	**-2.17**	0.18	1.86	**0.34**	-1.84
GULP1	**-1.53**	-0.20	1.52	**-0.30**	-1.84
C1orf38	**-1.53**	-0.17	1.82	**-0.30**	-1.84
SNORD4B	**-1.54**	0.25	-1.19	**-0.30**	-1.84
CLK3	**-2.13**	-0.64	-0.46	**0.29**	-1.84
SLC2A12	**-1.54**	-0.37	0.79	**-0.29**	-1.84
SREK1	**-1.54**	0.10	-2.96	**-0.29**	-1.84

(b) Top differentially expressed genes with direct effect only.

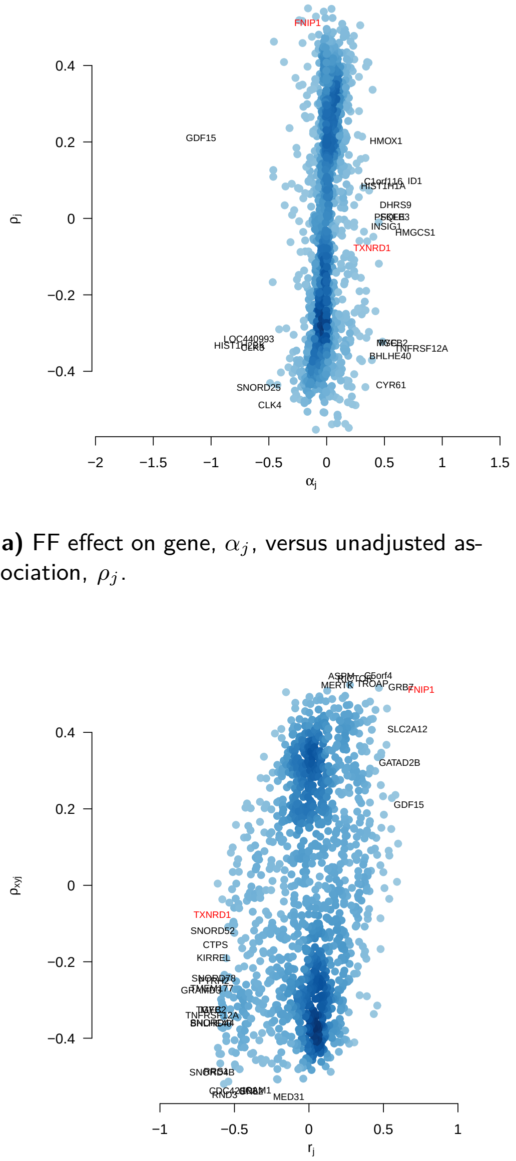
Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
GDF15	**-1.63**	**-1.09**	0.19	-0.20	-1.84
HMGCS1	**-1.80**	**0.77**	-0.04	-0.03	-1.84
ID1	**-1.92**	**0.76**	0.11	0.08	-1.84
DHRS9	**-1.86**	**0.60**	0.04	0.02	-1.84
SQLE	**-1.84**	**0.57**	0.00	0.00	-1.84
PFKFB3	**-1.84**	**0.57**	0.01	0.00	-1.84
INSIG1	**-1.83**	**0.52**	-0.02	-0.01	-1.84
HMOX1	**-1.99**	**0.52**	0.29	0.15	-1.84
MSMO1	**-1.76**	**0.50**	-0.15	-0.07	-1.84
HIST1H1A	**-1.91**	**0.49**	0.15	0.07	-1.84

(c) Genes with high direct effect and related to bioactivity.

Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
CDC42EP3	**-1.66**	0.16	**-1.06**	-0.17	-1.84
GAK	**-2.00**	0.06	**2.83**	0.16	-1.84
TARS2	**-1.99**	0.13	**1.19**	0.16	-1.84
UBA3	**-1.99**	0.08	**1.99**	0.15	-1.84
DOCK7	**-1.99**	0.08	**1.94**	0.15	-1.84
YWHAG	**-1.68**	0.05	**-3.23**	-0.15	-1.84
PNO1	**-1.69**	0.07	**-2.14**	-0.15	-1.84
SRSF5	**-1.99**	0.13	**1.16**	0.15	-1.84
RUFY1	**-1.98**	0.08	**1.75**	0.15	-1.84
GAB2	**-1.69**	-0.09	**1.68**	-0.15	-1.84

(d) Genes with high direct effect and not related to bioactivity.

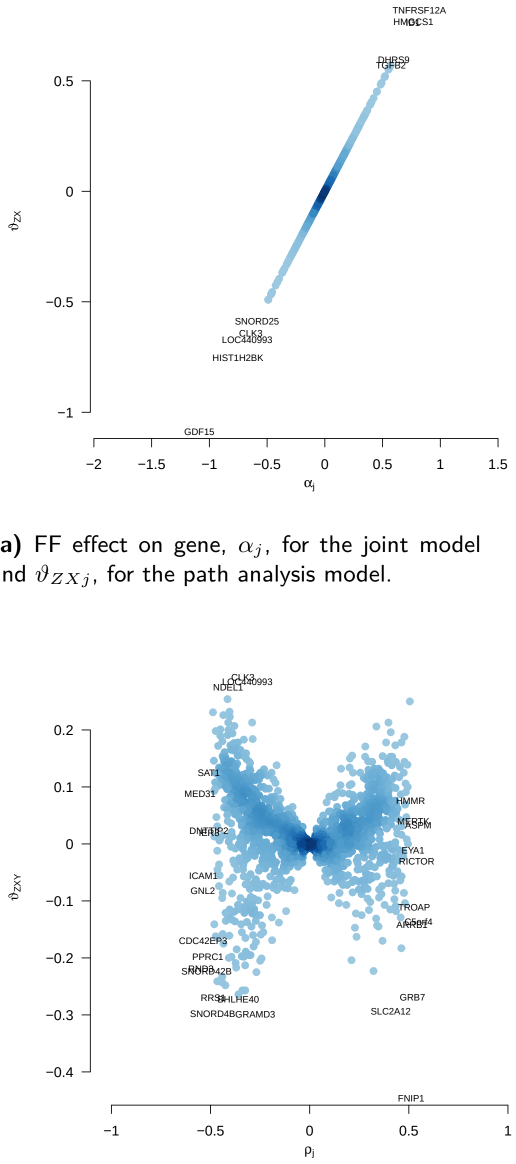
Genes	ϑ_{ZYj}	ϑ_{ZXj}	ϑ_{XYj}	ϑ_{ZXYj}	Total
FAM113A	**-1.93**	-0.07	-1.32	0.10	-1.84
TMED10	**-1.93**	-0.10	-0.93	0.10	-1.84
NEB	**-1.93**	-0.08	-1.14	0.09	-1.84
FKBP7	**-1.93**	-0.08	-1.22	0.09	-1.84
ATP13A1	**-1.93**	-0.04	-2.10	0.09	-1.84
MAPK14	**-1.93**	0.06	1.56	0.09	-1.84
AP1G1	**-1.93**	0.06	1.69	0.09	-1.84
LRRC4	**-1.93**	0.05	1.96	0.09	-1.84
GNE	**-1.93**	0.08	1.12	0.09	-1.84
PI4KB	**-1.93**	-0.03	-3.09	0.09	-1.84



(a) FF effect on gene, α_j, versus unadjusted association, ρ_j.

(b) Unadjusted correlation, r_j versus unadjusted association, ρ_j.

Figure 3.7: Plots highlighting the most relevant genes for the ROS1 project as discussed in Chapter 2.



(a) FF effect on gene, α_j, for the joint model and ϑ_{ZXj}, for the path analysis model.

(b) Adjusted association, ρ_j versus the indirect effect, $\vartheta_{ZXYj} = \vartheta_{ZXj} * \vartheta_{XYj}$.

Figure 3.8: The ROS1 project. Parameter estimates of the joint model and path analysis model.

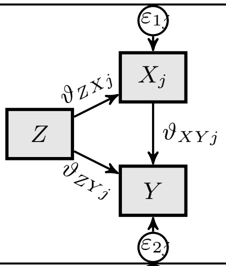
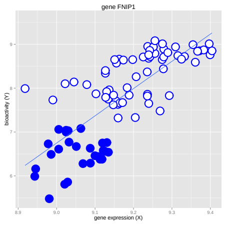
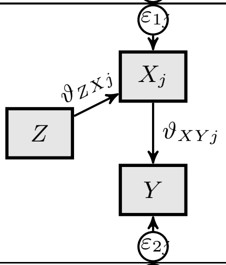
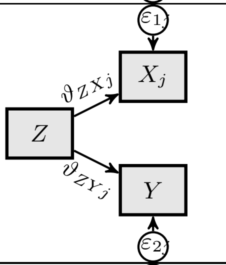
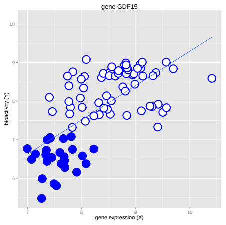
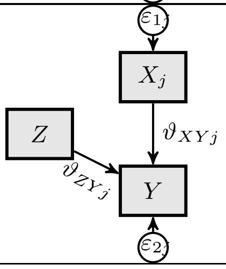
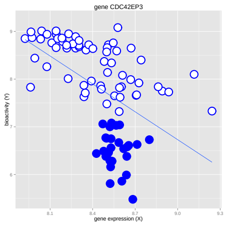
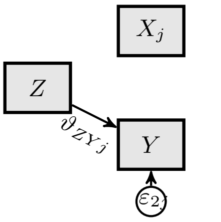
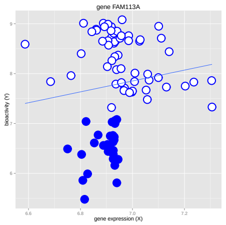
Type	Path Diagram	ROS genes	Estimates
(a) The effect of Z on Y can be estimated via its direct and indirect effects.	 ε_1 X_j Z ϑ_{ZXj} ϑ_{XYj} ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.44$, $\vartheta_{ZXj} = 0.79$, $\vartheta_{XYj} = -0.64$, indirect= -0.50, total= -0.94
(b) Z is indirectly related to Y via Z.	 ε_1 X_j Z ϑ_{ZXj} ϑ_{XYj} Y ε_2		None
(c) Z has direct effects on X and Y. X is not directly related to Y given Z.	 ε_1 X_j Z ϑ_{ZXj} ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.64$, $\vartheta_{ZXj} = -0.30$, $\vartheta_{XYj} = 1.00$, indirect= -0.30, total= -0.94
(d) Z is directly related to Y but not to X. X and Y are directly related given Z.	 ε_1 X_j Z ϑ_{XYj} ϑ_{ZYj} Y ε_2		$\vartheta_{ZYj} = -0.95$, $\vartheta_{ZXj} = -0.01$, $\vartheta_{XYj} = -0.83$, indirect= 0.01, total= -0.94
(e) Z is directly related to Y.			$\vartheta_{ZYj} = -1.408$, $\vartheta_{ZXj} = 0.05$, $\vartheta_{XYj} = -2.84$, indirect= -0.14, total= -0.94

Figure 3.9: The ROS1 Project. Genes discovered from testing path links and classified according to the the corresponding diagram.

3.5 The Conditional Model

Recall that a gene-by-gene joint model is given by

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_j + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma_j \right], \tag{3.2}$$

where the error terms follow a bivariate normal distribution with zero mean and gene-specific covariance matrix, Σ_j, given by

$$\Sigma_j = \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix}. \tag{3.3}$$

Using the joint model, we were able to identify potential genetic biomarkers for compound efficacy as measured by pIC50 on the basis of the adjusted association. In this case with normally distributed bivariate responses, the joint model implies the following conditional distribution of Y given X and Z (Burzykowski *et al.*, 2005),

$$Y_i | X_{ji}, Z_i = \gamma_{0j} + \gamma_{1j} Z_i + \gamma_{2j} X_{ji} + \epsilon_{3ij}, \tag{3.4}$$

or

$$Y_i | Z_i, X_{ji} \sim N \left(\gamma_{0j} + \gamma_{1j} Z_i + \gamma_{2j} X_{ji}, \sigma_j^2 \right),$$

with the following relationships:

$$\gamma_{0j} = \mu_Y - \sigma_{jY} \left({\sigma_{jj}}^{-1} \mu_j \right),$$

$$\gamma_{1j} = \beta - \sigma_{jY} \left({\sigma_{jj}}^{-1} \right) \alpha_j,$$

$$\gamma_{2j} = \sigma_{jY} \left({\sigma_{jj}}^{-1} \right), \text{and}$$

$$\sigma_j^2 = \sigma_{YY} - \sigma_{jY} \left({\sigma_{jj}}^{-1} \right).$$

Moreover, using the information theory approach (Alonso and Molenberghs, 2006), we can compare the following linear models to determine whether the variation in the bioactivity data can be better explained when the gene is known:

$$E(Y_i | Z_i) = \delta_0 + \delta_1 Z_i, \tag{3.5}$$

$$E(Y_i | Z_i, X_{ji}) = \gamma_{0j} + \gamma_{1j} Z_i + \gamma_{2j} X_{ji}. \tag{3.6}$$

Model (3.5) relates the expected value of the bioactivity read-out to the fingerprint feature only while (3.6) is the conditional distribution in (3.4) that relates the bioactivity read-out to the gene expression as well. Upon fitting models (3.5) and (3.6), the association can be measured by:

$$R^2_{hj} = 1 - exp\left(\frac{-G^2}{n}\right), \tag{3.7}$$

where G^2 denotes the likelihood ratio statistics to compare models (3.5) and (3.6), and n is the sample size. Note that for continuous outcomes, R^2_{hj} and the squared adjusted association $R^2_{hj} = {\rho_j}^2$ give identical results.

We can extend model (3.4) to account for an interaction term between X and Z or to account for multiple genes in the model. That is, we use the 'gene combination' consisting of g genes, $\sum\limits_{j=1}^{g} \gamma_{2j} X_j$, instead of a gene X.

3.6 Discussion

In this chapter, we discussed the relationship between the joint model and path analysis model in relation to the QSTAR paradigm. Also, we presented the implied conditional distribution of the joint model which can be extended to account for gene dependencies. With the conditional approach, it is possible to test whether a gene or group of genes can further explain the variability of the bioactivity read-outs given the effect of the fingerprint feature. With this model specification, we can move from a gene-by-gene approach to

Table 3.6: The QSTAR modeling framework.

Model	Specification
Joint model	$X_{ji} = \mu_j + \alpha_j Z_i + \varepsilon_{ij},$ $Y_i = \mu_Y + \beta Z_i + \varepsilon_i.$
Path analysis	$X_{ji} = \vartheta_{ZXj} Z_i + \varepsilon_{1ij},$ $Y_i = \vartheta_{XYj} X_{ji} + \vartheta_{ZYj} Z_i + \varepsilon_{2ij}.$
Conditional model	$Y_i = \delta_0 + \delta_1 Z_i + \varepsilon_{1i},$ $Y_i = \gamma_{0j} + \gamma_{1j} Z_i + \gamma_{2j} X_{ji} + \varepsilon_{2ij}.$

accounting for all genes in the model and simultaneously provide a gene combination that is related to the bioactivity given a fingerprint feature. This topic is further discussed in Chapter 12. We cannot however identify differentially expressed genes using this model in contrast to the joint model and path analysis model.

The ability of path analysis modelling to test models with multiple dependent variables amongst the predictors makes it ideal for the QSTAR setting. Here, the interest is on the total effect of the chemical substructure on the bioactivity, but a portion (or all) of this effect is maybe due to the effect of the structure via a gene. This decomposition of the effect can only be estimated using this model. However, it is important to note that exploration of relationships among variables without a priori specification may result in statistical significance but has little biological significance.

Part II

Semi-supervised Multi-source Data Integration

Chapter 4

Weighted Similarity-based Clustering of Chemical Structures and Bioactivity Data

4.1 Introduction

During the lead selection phase of the early drug discovery process, two key pieces of information about the compounds' property are gathered: structural and phenotypic profiles. Following the structure-based virtual screening and hit identification using single-dose screening, the resulting chemically diverse molecules can be classified into structural classes via cluster analysis (Olah *et al.*, 2004). Note that chemical diversity and novelty are subsequently verified in biological assays. 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 dose dependence of the assay's readout. This is usually expressed as an IC50 in enzyme-, protein-, antibody-, or cell-based assays.

Several statistical approaches have been proposed to quantify structure-activity relationships. Harrison (1968) applied cluster analysis in drug discovery asserting that clusters exist within the chemical space which favor biological activity. In practice, however, structurally similar compounds are not necessarily biologically similar (Drakakaki *et al.*, 2011).

One of the few attempts to cluster compounds on joint information of chemistry and biology is presented in Drakakaki *et al.* (2011) who used both structural and phenotypic information of compounds. However, they started with independent hierarchical clustering of each data source. They concluded that diverse structures could result in similar phenotypes and that a simple structure-based analysis resulted in very weak co-clustering of related phenotypic profiles.

In line with this context, we present two approaches of drug discovery clustering in this chapter. The first strategy is compound clustering based on either chemical similarity to obtain representative chemically diverse compounds or based on potency information to identify compound clusters with desirable activity. Note that clustering algorithms use similarity data in order to group objects and are typically performed on one data source. Compounds that are consistently clustered together on separate clustering results are of great interest. However, separately clustering multi-source data does not allow instant inference about compound clusters that are jointly supported or consistent with the different data sources. In this regard, the second strategy involves a clustering technique that can group compounds according to the similarity of both the biological and chemical profiles can be a very relevant exploratory tool for the discovery or prioritization of lead compounds. In general, the more information is used for grouping a set of objects, the more defined and meaningful the derived clusters are. Here, the biological and chemical information is integrated at the level of the similarity measures. In addition to clustering, a secondary analysis is performed wherein each biologically and structurally driven compound cluster is further linked to a set of transcriptomic features to better understand their mechanism of action. Compounds that are structurally and biologically similar to the reference compounds are discovered using the proposed clustering approach.

4.2 Cluster Analysis Based on Single Data Source

Cluster analysis relies on partitioning cases into a number of meaningful subgroups (clusters) on the basis of a set of measured features. Clusters should exhibit high intra-cluster homogeneity and high inter-cluster heterogeneity (Hartigan, 1975). Liu and Johnson (2009) discussed in details the three main steps for compound clustering based on chemical structures: the computation of structural features, the selection of a difference metric, and the application of the clustering algorithm.

Let $\mathbf{Z}_{JXN}$ be the chemical structure or fingerprint feature matrix in which the jith element, z_{ji} is a binary indicator representing absence or presence of the jth fingerprint feature in the ith compound. For each pair of compounds i and i', the structural similarity $0 \leq s_{i,i'} \leq 1$ can be quantified based on a set of J fingerprint features using the Tanimoto statistic (Willett *et al.*, 1998). Using the similarity matrix, $\mathbf{S}_N^Z$ in which the ii' entry is equal to $s_{i,i'}$, as input to Ward's linkage-based agglomerative hierarchical clustering, each compound is absorbed into increasingly large clusters until the dataset can be expressed as a single cluster composed of all compounds. The structure-based clustering solution, denoted by $\mathbf{C}(Z, k)$, that allows for the grouping of compounds into k clusters, can be represented by the so-called dendrogram to visualize which compounds are joined together at each step. The number of clusters, k, is chosen using the gap statistic proposed by Tibshirani *et al.* (2001).

Another meaningful way to group compounds is based upon their bioactivity profiles, $\mathbf{Y}_{BXN}$ with elements y_{bi} representing the activity value of the ith compound measured on the bth assay. The activity data are standardized and the Euclidean distance is used to obtain the similarity matrix of N compounds based on B assays, $\mathbf{S}_N^Y$. This bioactivity similarity data can be used to hierarchically cluster compounds to provide another clustering solution denoted by $\mathbf{C}(Y, k)$.

Clustering in drug discovery is mostly based either on chemical similarity, (e.g. Li *et al.* (2011)) or on similarity of bioactivity data (e.g. Cheng *et al.* (2010)). One of the few attempts to cluster compounds on joint information of chemistry and biology is presented in Drakakaki *et al.* (2011) who used both structural and phenotypic information of a set of compounds. However, they started with separate hierarchical clustering of each data source. They concluded that diverse structures could result in similar phenotypes and that a simple structure-based analysis resulted in very weak co-clustering of related phenotypic profiles.

4.3 Weighted Similarity-based Clustering

When structure-activity relation principle holds, the individual clustering results based on chemical structure and biological activity are expected to be similar. If not, it is difficult to identify a set of compounds with aligned structural and biological properties. A clustering method that combines both data sources can help to gain instant access to structurally and biologically similar compounds and will be presented in the next section.

Our proposed approach is an adaptation of the clustering technique based on combining multiple data sources at the similarity-level (Xu *et al.*, 2012, Hu, 2011, Lange and Buhmann, 2005, Liu *et al.*, 2013). This technique allows the combination of multiple similarity matrices via a weighting scheme. In general, for M data sources, the weighted similarity matrix, $\mathbf{S}_N^W$ is given by

$$\mathbf{S}_N^W = \sum_{m=1}^{M} \omega_m \cdot \mathbf{S}_N^m, \qquad 0 \leq \omega_m \leq 1, \qquad \sum_{m=1}^{M} \omega_m = 1. \tag{4.1}$$

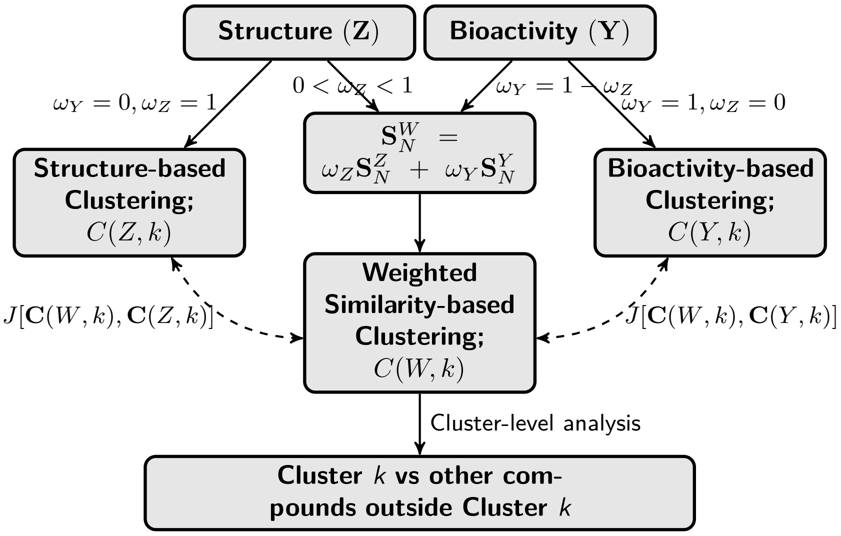

Structure $(\mathbf{Z})$
Bioactivity $(\mathbf{Y})$
$\omega_Y = 0, \omega_Z = 1$
$0 < \omega_Z < 1$
$\omega_Y = 1 - \omega_Z$
$\omega_Y = 1, \omega_Z = 0$
Structure-based Clustering; $C(Z,k)$
$\mathbf{S}_N^W = \omega_Z \mathbf{S}_N^Z + \omega_Y \mathbf{S}_N^Y$
Bioactivity-based Clustering; $C(Y,k)$
$J[\mathbf{C}(W,k), \mathbf{C}(Z,k)]$
Weighted Similarity-based Clustering; $C(W,k)$
$J[\mathbf{C}(W,k), \mathbf{C}(Y,k)]$
Cluster-level analysis
Cluster k vs other compounds outside Cluster k

Figure 4.1: Diagram summarizing the steps involved in the compound clustering for k clusters. A weighted similarity matrix is derived from two initial compound similarity matrices from two data sets, bioactivity and chemical structure of a compound set. $J[\mathbf{C}(W,k), \mathbf{C}(\cdot,k)]$ similarity index represents the similarity of the clustering structure between the weighted- and single-source clusterings, for pre-determined k clusters. The clusters can be explored in greater detail by looking at common and distinctive properties in chemistry, phenotype or other information such as gene expression.

Here, $\mathbf{S}_N^m$ is the similarity matrix for the mth data source and ω_m is the weight associated with the mth source. The possibility to assign a weight on each similarity matrix provides flexibility on the respective influence from each data source in the combined clustering solution. For the case of two data sources, i.e. $m = 2$, a weight of 0 assigned to the first data source, implies that the clustering will be based on the second data source alone and 1 means the first data source alone.

The weighted similarity matrix, $\mathbf{S}_N^W$ presented in (4.1), will be used as input for the agglomerative hierarchical clustering producing the weighted clustering solution, denoted by $\mathbf{C}(\mathbf{W}, k)$. Note that, compound pairs with the same similarity scores on both chemical and bioactivity spaces will be clustered together regardless of the weights imposed on both similarity matrices. Analogously, no similarity in any of the data sources will result to low summarized similarity. The weights will, however, affect the clustering of compound pairs having high similarity score in one space and low to moderate in the other. The choice of weights can bias the merging of similarity information towards a particular data source.

An overview of the proposed weighted similarity clustering approach integrating bioactivity and chemical structure data sources is presented in Figure 4.1. In this chapter, the following weighting scheme is considered. The impact of the weights on the clustering results is based on the method of co-membership proposed by Tibshirani and Walther (2005). The co-membership is defined as compound pairs that are clustered together,

$$c_{i,i'} = \begin{cases} 1 & \text{compounds } i \text{ and } i' \text{ are clusterd together,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

For any clustering solution with k clusters, $\mathbf{C}(\cdot, k)$, each compound pair $(i, i', \quad i = 1, \ldots, N)$ is evaluated whether they are co-members or not, thereby producing a binary vector with $\left(\begin{smallmatrix} N \\ 2 \end{smallmatrix}\right)$ possible compound pairs. For two independent clustering solutions, denoted by $\mathbf{C}(\mathbf{Z}, k)$ and $\mathbf{C}(\mathbf{Y}, k)$, we can determine whether compounds that are co-members in structure-based clustering are also co-members in bioactivity-based clustering, i.e., for any two compounds, i and i', $c_{i,i'}[\mathbf{C}(Z,k)] = 1$ and $c_{i,i'}[\mathbf{C}(Y,k)] = 1$. Eventually, the similarity of the two clustering solutions based on compounds' co-membership can be measured using the Jaccard statistic given by

$$J[C(Z,k), C(Y,k)] = \frac{N_{1,1}c(i,i')}{N_{1,0}c(i,i') + N_{0,1}c(i,i') + N_{1,1}c(i,i')}. \tag{4.3}$$

Here, $N_{1,1}c(i,i')$ is the number of compound pairs that are co-members in both clustering solution; $N_{1,0}c(i,i')$ and $N_{0,1}c(i,i')$ is the number of compound pairs that are only co-members in the first and second clustering solutions, respectively.

A Jaccard index of 1 means that the clustering based on structure and activity are identical. A high Jaccard index would mean an ideal structure-activity relationship setting. A low Jaccard index however, would benefit from the combination of the data sources to recluster the compounds by leveraging the effects of both data sources via the weighted-based similarity clustering approach.

The weighted similarity matrix of bioactivity and chemical structure data is given by,

$$\mathbf{S}_N^W = \omega_Z \mathbf{S}_N^Z + \omega_Y \mathbf{S}_N^Y, \quad \omega_Z = 1 - \omega_Y. \tag{4.4}$$

Starting from $\omega_Y = 0$ and increasing its value up to 1 gradually integrates the bioactivity data to the structure-based clustering solution. Using different set of weights may result in a different integrated clustering solution, $\mathbf{C}(W(\omega_Y), k)$. The Jaccard similarity index for comparing two clustering solutions based on co-membership scores of compounds will be used to guide the choice of a fair ω_Y, given by ω_Y^*, reflecting a balanced contribution of the two data sources to the integrative solution.

Let $J[\mathbf{C}(W(\omega_Y), k), \mathbf{C}(Y, k)]$ be the Jaccard index representing the clustering similarity solution of the weighted clustering and the bioactivity-based clustering. Similarly,

$J[\mathbf{C}(W(\omega_Y),k),\mathbf{C}(Z,k)]$ is the similarity between the weighted clustering and the chemical structure-based clustering. A value of $J[\mathbf{C}(W(\omega_Y),k),\mathbf{C}(\cdot,k)]$ equal to 1 indicates a matching of weighted similarity and single source clustering approaches, while a value of 0 signals the complete randomness of the two clustering solutions. The desirable set of weights (ω_Y^*,ω_Z^*) to be used in integrative clustering is one with equal contributions of the various single source solutions. Therefore, a clustering commonality index, R, given by the ratio of the two similarity indices,

$$R=\frac{J[\mathbf{C}(W(\omega_Y^*),k),\mathbf{C}(Z,k)]}{J[\mathbf{C}(W(\omega_Y^*),k),\mathbf{C}(Y,k)]}\simeq 1,$$

is proposed to guide the choice of weights.

4.3.1 Cluster-related Genes and Structural Features

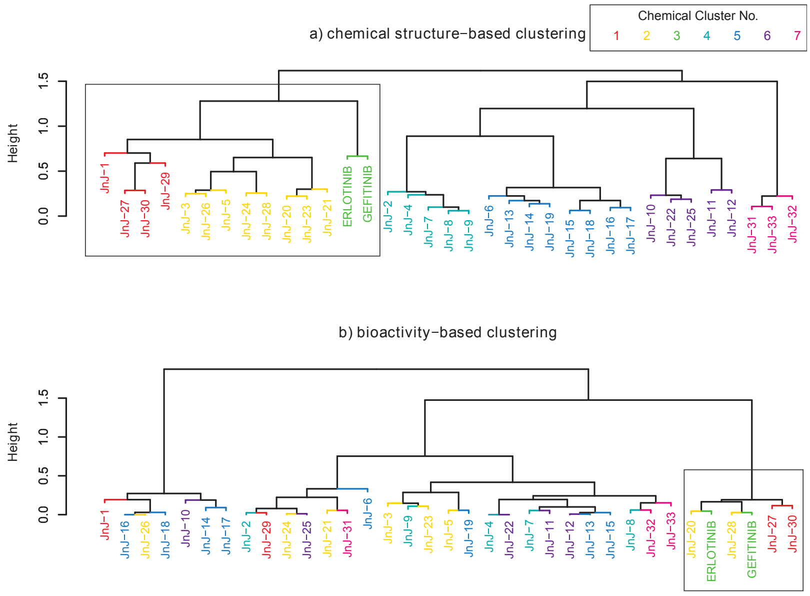
Once the weighed clustering solution is obtained, we follow the data analysis approach proposed by Ravindranath *et al.* (2015) and identified genes and structural features that could be linked to a specific cluster of compounds. This is done by employing Fisher's exact test (Fisher, 1922) on each fingerprint feature and by running LIMMA (Smyth *et al.*, 2005, Smyth, 2004) on the gene expression set with the chosen compound cluster as one group and all other compounds forming another group.

4.4 Application to the EGFR Project

Based on chemical structure, seven compound clusters were identified by hierarchical clustering using the gap statistic (Figure 4.2a). The two reference compounds, Erlotinib and Gefitinib, representing the third (green) cluster are, as anticipated, structurally very similar. The first two structural clusters are similar to EGFR-inhibitors, as denoted by the black rectangle in Figure 4.2a. The analysis based on bioactivity similarity matrix, shown in Figure 4.2b, results in a different cluster structure.

For example, compounds marked in yellow in Figure 4.2a are grouped with compounds from different structural clusters in Figure 4.2b. This implies that similarity on chemical space does not always translate to similarity in biological space. Note, however, that the reference compounds are still clustered together along with a subset of compounds marked in red and yellow. This indicates that these four compounds are similar, with respect to both structure and activity, to the reference compounds.

In practice, however, when reference compounds are absent or when many compounds of interest are available, inspecting the two separate clustering results to find compounds sharing similar structures as well as inducing similar activities is not straightforward. To


a) chemical structure−based clustering

b) bioactivity−based clustering

Figure 4.2: Two clustering solutions: **(a)** chemical-structure based and **(b)** bioactivity-based. The color of the compounds indicates the chemical-structural cluster where they belong.

this end, we implemented the weighted similarity clustering combining both similarities to yield integrated clusters. This approach generates compound clusters with similar chemistry and similar bioactivity.

Figure 4.3 shows how the membership of the compounds under investigation changes from the chemical structure-based clustering when the weights of the bioactivity data are gradually increased by 0.05. Interestingly, the two reference compounds remain together in the same cluster in all weighted clustering solutions. This is because both compounds are highly similar in both structure and bioactivity so that the weighted average remains high regardless of the set of weights used. In this case study, the bioactivity similarity information starts to influence the clustering at $\omega_Y \approx 0.1$. Assigning equal weight of 0.5 to both data sources yields a clustering solution that is slightly more influenced by chemical structure similarity ($J[\mathbf{C}(W,7), \mathbf{C}(Y,7)] = 0.23$ versus $J[\mathbf{C}(W,7), \mathbf{C}(Z,7)] = 0.38$).

The set of candidate weights, starting from $\omega_Y = 0$, that gradually integrates bioactivity information in the chemical structure-based clustering and their corresponding similarity indices are presented Table 4.2.

Figure 4.4 shows that the weight combination $\omega_Y = 0.55$ and $\omega_Z = 0.45$ leads to $R = 0.85$. This set of weights is therefore selected to obtain the weighted clustering

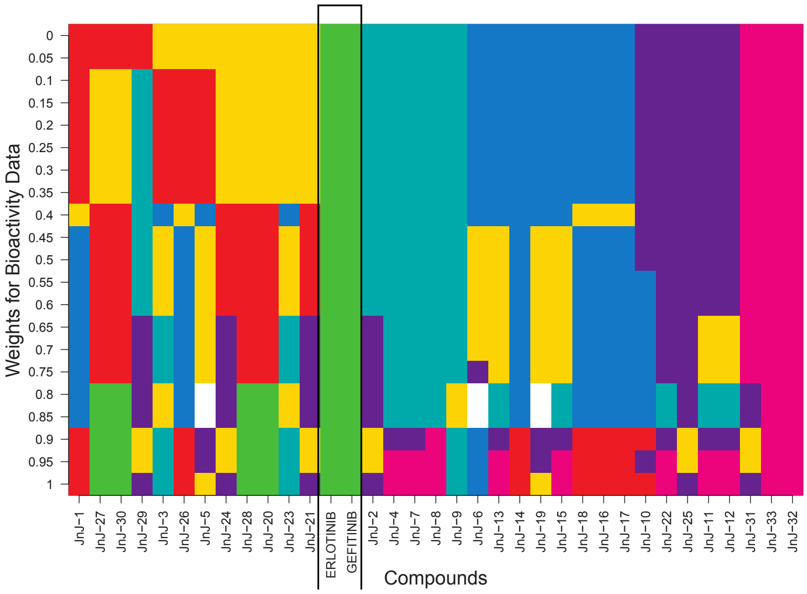


Figure 4.3: Changes in the cluster structures when increasing the weights of the bioactivity similarity data in the weighted similarity clustering approach. The colors denote the cluster membership of the compounds per clustering solution (row).

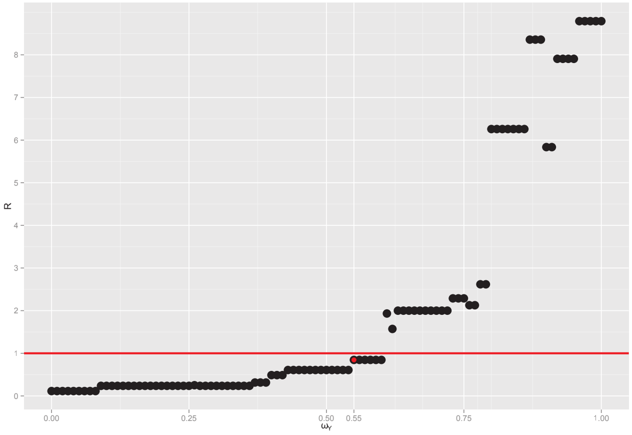


Figure 4.4: Clustering commonality index, R where the point closest to the red line indicated the weight for the bioactivity data that would give a $R \simeq 1$. In this case, $\omega_Y = 0.55$ which implies $\omega_Z = 0.45$.

Table 4.1: Table of Similarity Indices when varying weight sets by 0.01.

ω_Y	$J[\mathbf{C}(W,7),\mathbf{C}(Y,7)]$	$J[\mathbf{C}(W,7),\mathbf{C}(Z,7)]$	R
0.00	0.11	1.00	0.11
0.01	0.11	1.00	0.11
0.02	0.11	1.00	0.11
0.03	0.11	1.00	0.11
0.04	0.11	1.00	0.11
0.05	0.11	1.00	0.11
0.06	0.11	1.00	0.11
0.07	0.11	1.00	0.11
0.08	0.11	1.00	0.11
0.09	0.15	0.63	0.24
0.10	0.15	0.63	0.24
0.11	0.15	0.63	0.24
0.12	0.15	0.63	0.24
0.13	0.15	0.63	0.24
0.14	0.15	0.63	0.24
0.15	0.15	0.63	0.24
0.16	0.15	0.63	0.24
0.17	0.15	0.63	0.24
0.18	0.15	0.63	0.24
0.19	0.15	0.63	0.24
0.20	0.15	0.63	0.24
0.21	0.15	0.63	0.24
0.22	0.15	0.63	0.24
0.23	0.15	0.63	0.24
0.24	0.15	0.63	0.24
0.25	0.15	0.63	0.24
0.26	0.14	0.56	0.25
0.27	0.15	0.63	0.24
0.28	0.15	0.63	0.24
0.29	0.15	0.63	0.24
0.30	0.15	0.63	0.24
0.31	0.15	0.63	0.24
0.32	0.15	0.63	0.24
0.33	0.15	0.63	0.24

Table 4.2: (*cont.*) Table of Similarity Indices when varying weight sets by 0.01.

ω_Y	$J[\mathbf{C}(W,7),\mathbf{C}(Y,7)]$	$J[\mathbf{C}(W,7),\mathbf{C}(Z,7)]$	R
0.34	0.15	0.63	0.24
0.35	0.15	0.63	0.24
0.36	0.15	0.63	0.24
0.37	0.18	0.56	0.31
0.38	0.18	0.56	0.31
0.39	0.18	0.56	0.31
0.40	0.19	0.39	0.49
0.41	0.19	0.39	0.49
0.42	0.19	0.39	0.49
0.43	0.23	0.38	0.61
0.44	0.23	0.38	0.61
0.45	0.23	0.38	0.61
0.46	0.23	0.38	0.61
0.47	0.23	0.38	0.61
0.48	0.23	0.38	0.61
0.49	0.23	0.38	0.61
0.50	0.23	0.38	0.61
0.51	0.23	0.38	0.61
0.52	0.23	0.38	0.61
0.53	0.23	0.38	0.61
0.54	0.23	0.38	0.61
0.55	0.28	0.33	0.85
0.56	0.28	0.33	0.85
0.57	0.28	0.33	0.85
0.58	0.28	0.33	0.85
0.59	0.28	0.33	0.85
0.60	0.28	0.33	0.85
0.61	0.35	0.18	1.93
0.62	0.36	0.23	1.57
0.63	0.40	0.20	2.00
0.64	0.40	0.20	2.00
0.65	0.40	0.20	2.00
0.66	0.40	0.20	2.00

solution presented in Figure 4.5. The impact of the bioactivity data in the structural clusters is mostly observed in clusters marked in red and yellow where less compounds are clustered with the reference compounds. The other structural classes seem to be stable despite the contribution of the bioactivity data in the clustering. This indicates that these compounds are similar in terms of bioactivity profiles hence, there is not much regrouping. Furthermore, this shows that these compounds are not inducing the same activity level as the reference EGFR-inhibitors. The weighted-similarity based clustering solution presented in Figure 4.5 identified a subset of 6 compounds similar to the EGFR-inhibitors (see the outer box in Figure 4.5). These compounds are not just structurally similar but are also potentially potent EGFR-inhibitors. In the analysis based on bioactivity alone, two of these six compounds (inner black box in Figure 4.5) would have been missed (see Figure 4.2b). These two compounds may not be as highly potent as the reference and the four other compounds, but given their similar chemical structure, these compounds can still be interesting and warrant further investigation together with the other compounds in the cluster.

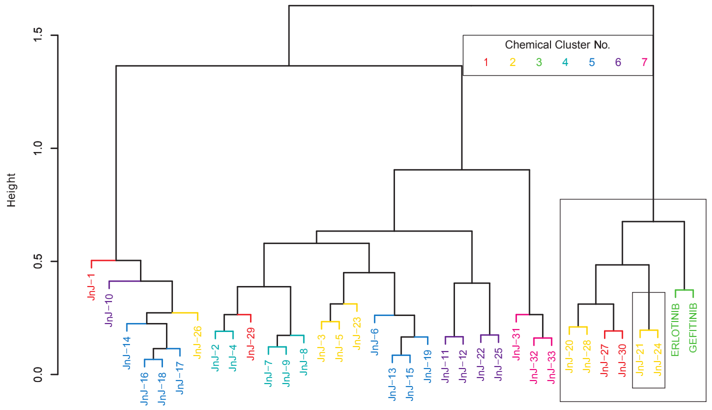


Figure 4.5: Weighted-similarity based clustering solution with weights $\omega_Y = 0.55$ and $\omega_Z = 0.45$.

4.4.1 Lead Compound Cluster

As illustrated in Ravindranath and Perualila-Tan *et al.* (2015), based on the weighted similarity clustering, compound clusters could be further linked to transcriptomics data to gain more biological insight. The compound cluster containing the two reference compounds along with the 6 candidate compounds discussed above was selected for further investigation (Figure 4.5). The set of structural features that makes these compounds a chemical cluster is shown in Figure 4.6. Notice that this set of fingerprint features

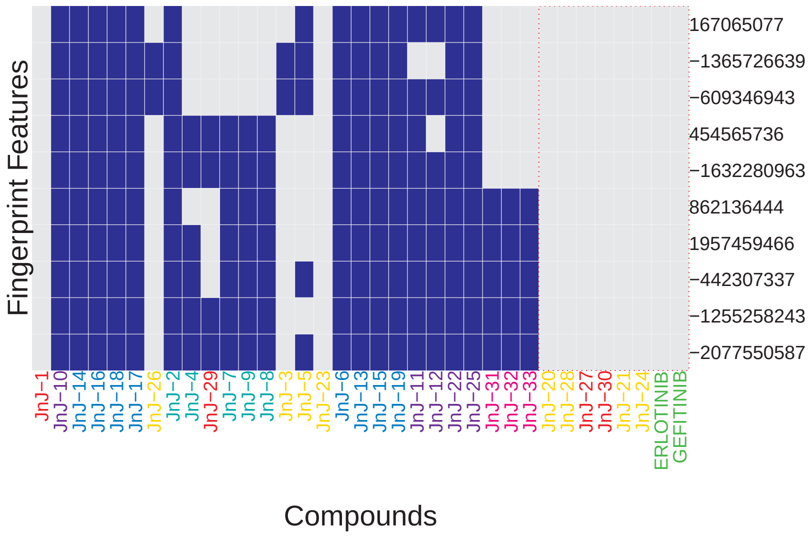


Figure 4.6: A subset of structural features that characterizes the chosen compound cluster in Figure 4.5. It is clear that the absence of almost all of these identified features drives the formation of this cluster.

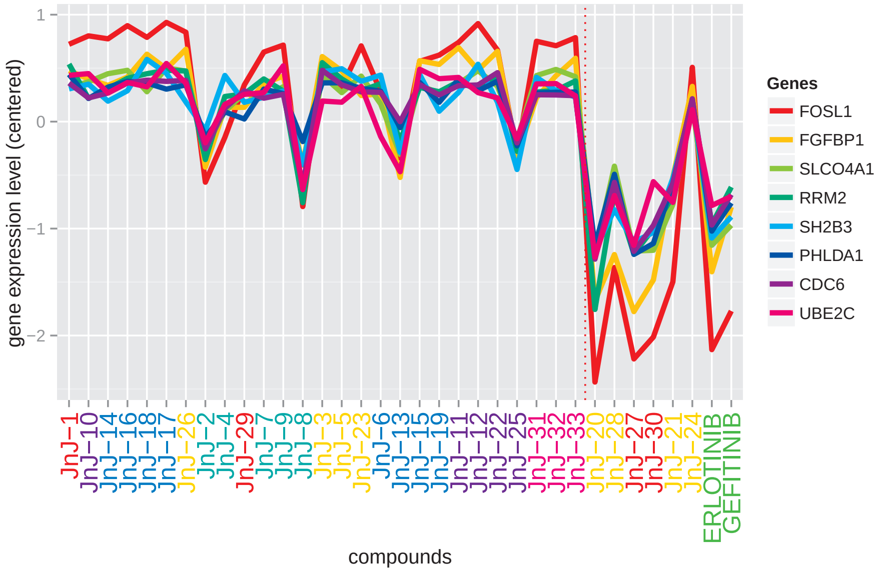


Figure 4.7: Profiles plot of genes that can be linked to the chosen compound cluster.

are absent in this compound cluster and are mostly present on the other compounds except for the first compound (JNJ-1) and some yellow compounds, which were dropped from the initial structural class when integrating activity similarity. The gene expression analysis for this compound cluster shows that all compounds, except one (JNJ-24), have inhibitory activity on multiple cancer-related genes like FGFBP1 and FOSL1 similar to the two EGFR-inhibitors, Gefitinib and Erlotinib (Figure 4.7). The gene FGFBP1 encodes for the fibroblast growth factor carrier protein (FGF-BP1) whose overexpression is noted in cell lines, from lung (Brattström *et al.*, 2002, Pardo *et al.*, 2003), prostate(Tassi *et al.*, 2006), pancreas (Kuwahara *et al.*, 2003), and colon cancer (Hauptmann *et al.*, 2003). Most of these differentially expressed genes are known to participate in biological processes involving cell proliferation (positive and negative), survival and differentiation. Here, the gene expression results support that the cluster of compounds discovered with weighted similarity based clustering seem to combine compounds with similar mechanism of action.

4.5 Discussion

In pharmaceutical drug development process, the availability of multiple data sources that can be used to jointly describe a compound set requires adjustment of classical approaches to cluster compounds. This is typically based solely on chemical-structure. In this chapter, we proposed a weighted similarity clustering approach, that assigns complementary weights to the similarity matrices and can be used to integrate information from both chemical structure and bioactivity data. We have shown that the weighted similarity clustering approach led to the identification of a subset of compounds which share similar structural and biological profiles.

We have shown that the weighted similarity-based clustering approach was able to identify more compounds that are structurally and biologically similar to the reference EGFR-inhibitors, a finding that is further supported by the gene expression analysis. In the absence of reference compounds, a subset of compounds belonging to the same structural class and yet are still discovered to be clustered together in weighted-similarity based clustering indicates that either the structure-based similarity scores are extremely high complemented by very weak bioactivity similarity profiles or both structure and bioactivity similarity scores have the same level,i.e., both low or both high.

Cluster analysis is an exploratory tool from both statistical and pharmaceutical point of view. However, we have shown that its output can be used to generate hypotheses and can contribute to enhance the discovering process of a subset of candidate compounds with highly desired properties, both structurally and biologically.

The integration step presented in this chapter makes use of the similarity matrix from each data source. Although, the Tanimoto coefficient (which is also equivalent to Jaccard coefficient in the case of binary data) has been commonly used for calculating chemical similarity (Willett *et al.*, 1998), other (dis)similarity measures maybe also used. The Euclidean distance is most popular for continuous data and the Hamming distance which is equivalent to the squared Euclidean distance for binary data is also advocated. Assessing the effect of different(dis)similarity measures by the extent to which the results differ would be an interesting extension of the work presented in this chapter. Moreover, aside from the Jaccard index that is presented in this chapter as a “measure" of similarity between the two clustering solutions, other indices like the Rand index (Rand, 1971), Fowles-Mallows Index (Fowlkes and Mallows, 1983), etc. could be also used. Assessing the sensitivity of these measures on the number of clusters and cluster sizes should be also investigated since the index is derived for a specified number of clusters.

In this chapter, the utility of this unsupervised technique to discover multi-source-driven clusters is illustrated in early drug discovery setting but this can be also applied in other research areas where different types of high dimensional data are being collected on the same samples. Moreover, following this work, within the QSTAR setting, other integrative clustering methodologies have been explored and investigated. An R package `IntClust` was developed to serve this purpose (Van Moerbeke, 2015).

Chapter 5

Biclustering Methods in Chemoinformatics and Molecular Modelling in Drug Discovery Experiments: Connecting Gene Expression and Target Prediction Data

5.1 Introduction

In this chapter, we discuss two approaches for the analysis of multi-source *drug discovery* data in order to gain insights into the compounds mechanism-of-action (MoA). The first approach is based on a two-step integrative analysis and the second is a biclustering analysis based on FABIA. In contrast to biclustering methods that find a subset of genes with similar expression profiles across a subset of compounds, the first approach first finds subsets of compounds that share similar predicted protein targets (via clustering) and then link them to a subset of genes by testing differential expression. This analysis workflow is similar to the previous chapter. However, instead of combining the information on the bioactivity and chemical structure of compounds via clustering, the bioactivity and chemical structure are used as input in calculating in silico target prediction data. Hence, we have another type of data coming from the two data sources. The first approach is discussed in Ravindranath *et al.* (2015).

5.1.1 Connecting Target Prediction and Gene Expression Data

Integrating gene expression profiles with certain proteins can improve our insight of the fundamental mechanisms in protein-ligand binding. Understanding protein target and off-target effects of bioactive compounds is a critical challenge in the field of drug discovery. These effects are of great importance as bioactive compounds that indicate a certain therapeutic effect could cause inadvertent phenotypic effects by binding to unexpected protein targets, thus resulting in disruption of compound efficacy (MacDonald *et al.*, 2006). Public chemogenomics databases such as ChEMBL and PubChem contain large amounts of bioactivity data that aid using machine learning techniques to classify new and orphan ligands for potential protein targets, or off-targets, based upon the similarity of the chemical structures. The target prediction algorithm was employed to predict probable protein targets for compounds without target information (Klabunde, 2007, Koutsoukas *et al.*, 2011). Target prediction approaches have been recently applied in a variety of areas (Chen *et al.*, 2011), such as the elucidation of MoA of compounds used in traditional medicine (including ayurvedic and Chinese medicine (Mohd Fauzi *et al.*, 2013)) and are also used in examining ADR (Takarabe *et al.*, 2012).

The use of drug-induced gene expression profiles to observe several co-regulated genes and signalling pathways concurrently enriches the understanding of underlying mechanisms. Due to the advances in the genome studies, large amount of microarray data has been deposited in public databases such as Connectivity Map (CMap) that was introduced in Chapter 1. The CMap dataset consists of drug-like compounds tested for gene expression in four cell lines. However, it is largely unknown how a compound exactly modulates gene expression and only a few data analysis approaches exist.

In addition, protein targets do not influence gene expression changes directly; they work through signalling cascades. Pathway databases provide information for linking genes and protein targets. Databases such as KEGG (Ogata *et al.*, 1999) and GO (Consortium, 2004) have been used in the study to rationalise the findings. The repositories KEGG and GO, combine information across all organisms which makes it flexible to integrate the information from different databases and thus to study functionality of recently discovered genes (du Plessis *et al.*, 2011). As shown in Figure 5.1, the links between a compound of interest and gene expression could be established by microarray data, the link between compound and protein target was established by employing a target prediction algorithm (see Section 1.2.4) while the link between protein target and gene expression (CMap) was

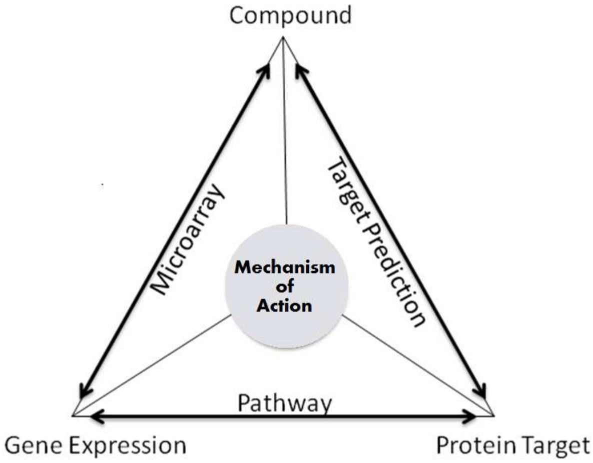


Figure 5.1: Mechanism of Action of compound. The compound to protein target information is derived from the target prediction algorithm and the phenotypic gene expression information is curated from experimental annotated data. To complete the triangle KEGG and GO pathways information is annotated for the genes and proteins and were overlapped to find similar pathways.

achieved by information on the pathways.

The analysis presented in this chapter consists of two input datasets: gene expression and target prediction data. Both datasets were introduced in Chapter 1. In this chapter, the analysis is focused on the 35 compounds of MCF7 (breast cancer epithelial cell) treated for duration of 6 hours with concentration of 10μM. For compounds having multiple instances, the average gene expression level was used. The target prediction algorithm provides each test compound with probable protein targets and their respective scores, representing the likelihood of binding to ChEMBL protein targets. With this data, the target prediction matrix scores of the 35 MCF7 compounds were extracted across all 477 available protein targets. For the analysis presented in this chapter, the target prediction score matrix was dichotomized using target-specific confidence score cut-offs (Paolini *et al.*, 2006).

The target prediction score matrix is given by

$$\mathbf{T} = \begin{pmatrix} T_{11} & T_{12} & \ldots & T_{1m} \\ T_{21} & T_{22} & \ldots & T_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ T_{n1} & T_{n2} & \ldots & T_{nm} \end{pmatrix}, \tag{5.1}$$

with entries,

$$T_{ij} = \begin{cases} 1 & \text{compound } i \text{ hit target } j, \quad i = 1, \ldots, n \text{ and } j = 1, \ldots, m, \\ 0 & \text{otherwise.} \end{cases}$$

5.2 Integrative Data Analysis Steps

The data analysis process consists of two main steps. In the first step, we cluster the compounds based on similarity in target prediction. Once the clustering is completed, feature selection (for both genes and target predictions) is performed. In the second step of the analysis, based on both genes and targets that were selected in the first step, pathway analysis was conducted in order to find biological pathways related to a cluster(s) of interest.

5.2.1 Clustering of Compounds

Similarity Matrix

We calculate a score for each pair of compounds which represents the degree of similarity based on their target prediction profiles. The distance between each pair of compounds is based on the Tanimoto coefficient (Tc) (Willett *et al.*, 1998), given by

$$Tc_{ab} = \frac{N_{ba}}{N_a + N_b - N_{ab}}. \tag{5.2}$$

Here, N_a is the number of targets with score 1 in compound a, N_b hit by compound b and N_{ab} is the number of targets common for both compounds. This gives us a similarity matrix, denoted by $\mathbf{S}_n$, containing pairwise Tanimoto similarity scores of n compounds.

Target Prediction-Based Clustering of Compounds

Compounds are clustered into groups that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity according to the similarity of their predicted targets. There are several existing clustering algorithms. The clustering procedure was based on agglomerative hierarchical clustering approach (Sokal and Michener, 1958), which begins with assumption that each entity in the similarity or distance matrix is a cluster. Thereafter, each compound is absorbed into larger and larger clusters until the dataset is expressed as a single cluster containing all compounds.

The hierarchical clustering of compounds according to the similarity of their target prediction profiles based on the 477 ChEMBL targets is presented in Figure 5.2 for the

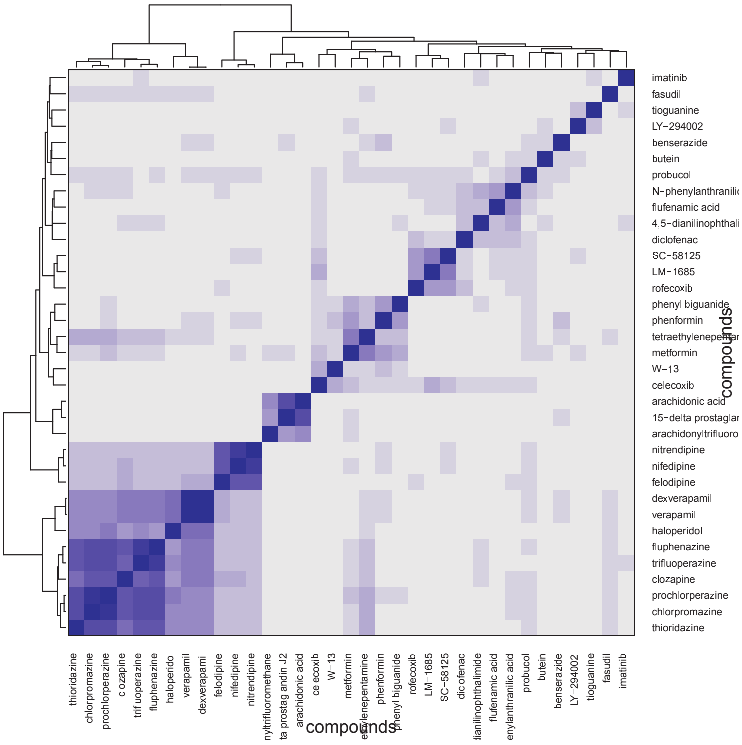


Figure 5.2: Target prediction-based clustering.

MCF7 cell line. From here, any compound cluster of interest can be linked to a set of features (both targets and differentially expressed genes). In the next sections, the feature selection is illustrated using the first target-driven cluster composed of 6 antipsychotic drugs (clozapine, thioridazine, chlorpromazine, trifluoperazine, prochlorperazine and fluphenazine). Other clusters can be analyzed in the same way.

5.2.2 Feature Selection

For every cluster-target combination, a 2x2 frequency table of compounds is formed. For example, in Table 5.1 we have the tabulated compound frequency for target Cytochrome P450 2D6 and cluster 1 membership.

Table 5.1: Frequency table for cluster - target combination.

Cluster 1	CytochromeP2D6	
	Active	Inactive
In	5	1
Out	0	29

The Fisher's exact test is used to determine whether the proportions of compounds that predicts Cytochrome P450 2D6 are the same inside and outside of cluster 1. This test is done for every target. Consequently, the top ranked targets based on p-value were identified. Benjamini-Hochberg false discovery rate (BH-FDR) method was used to adjust for multiplicity. Protein targets and genes were ranked based on their adjusted p-values (Benjamini and Hochberg, 1995). The top predicted protein targets of antipsychotic drugs are muscarinic, histamine, dopamine and adrenergic receptors , and cytochrome P450 2D6 whose profiles are displayed in Figure 5.3a.

In order to identify genes with differential expression in a cluster of compounds, we use the Linear Models for Microarray Data (Limma) method (Smyth *et al.*, 2005, Smyth, 2004). The top 8 differentially expressed genes by the antipsychotic drugs includes IDI1, INSIG1, MSMO1, LPIN1, SQLE, HMGCS1, NPC2, and BHLHE40. From Figure 5.3b, we can see that majority of the six compounds in the cluster induce a relatively higher expression than the other compounds for these 8 genes.

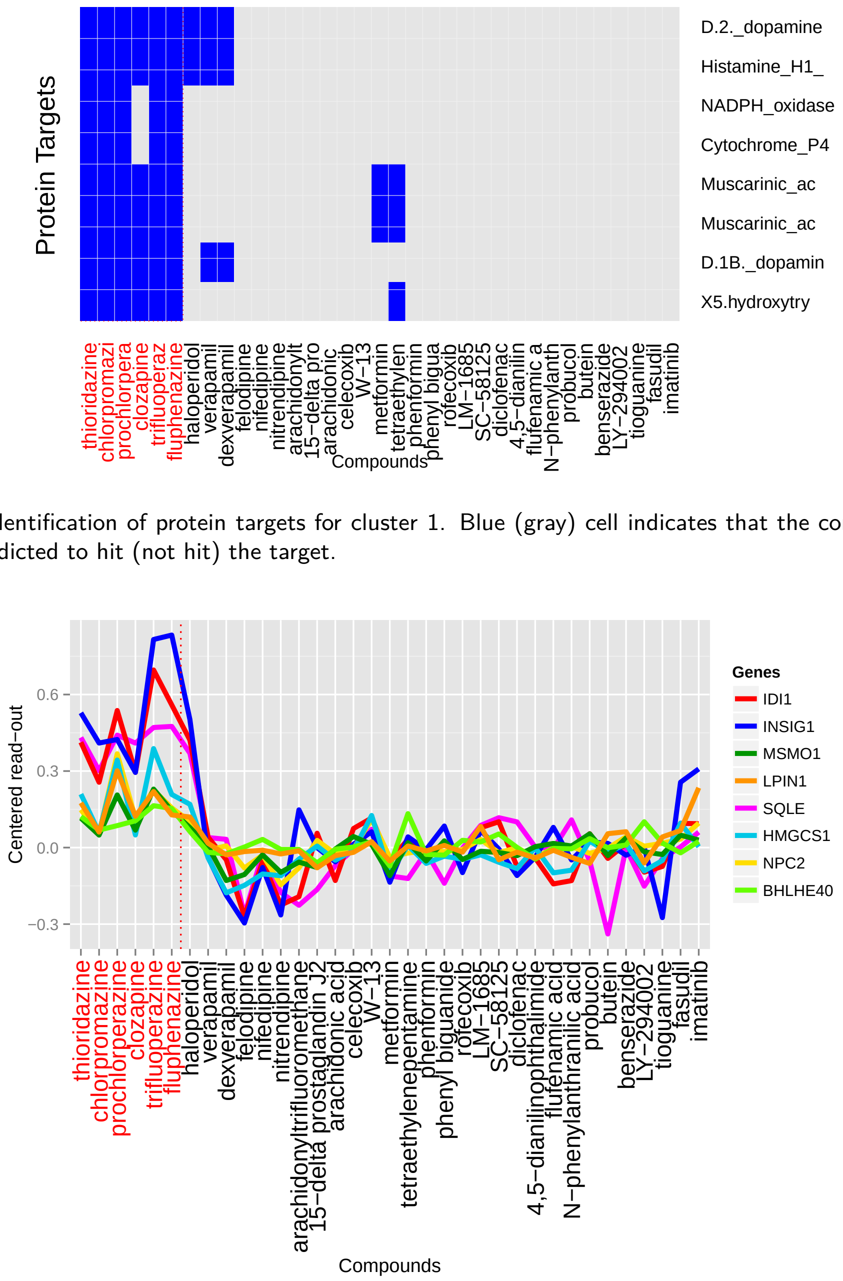
5.2.3 Pathway Analysis

In this stage, a group of genes and protein targets were linked to a group of compounds. Pathway analysis is performed to facilitate more intuitive interpretation of the biological function of the selected subset of genes/protein targets. In what follows, we present two different approaches, the first is based on pathway search based on both genes and protein targets and the second is based only on gene expression data. The latter should be seen as a biological enrichment of the gene set that was identified in the previous section.

Overlapping Pathway Search Using KEGG and GO Databases

Pathway information was extracted from the KEGG and GO databases for the gene sets and protein targets involved in this study. Each gene and target protein could be linked to several biological pathways. In this pathway-oriented approach, we get two sets of pathways, one for the top genes and another for the top protein targets. We then look for the intersection of these two sets of pathways and identify which of the input genes and targets are involved in each of the identified common pathways. This qualitative search of common pathways between targets and genes is dependent upon the completeness of the KEGG and GO pathway databases.

A search for overlapping pathway was executed on the antipsychotic cluster, where genes INSIG-1, LDLR and protein target CYP2D6 were observed to overlap with “steroid metabolic process pathway". This observation complies with the study by Polymeropoulos *et al.* (2009), in which it was shown that genes INSIG-1 and LDLR were



(a) Identification of protein targets for cluster 1. Blue (gray) cell indicates that the compound is predicted to hit (not hit) the target.

(b) Profiles plot of 8 differentially expressed genes linked to the 6 compounds (marked in red) in cluster 1.

Figure 5.3: Features linked to cluster 1 compounds.

up-regulated by antipsychotic drugs that also influenced the steroid biosynthesis.

Gene Set Analysis Using Mean Log P-value (MLP) Analysis

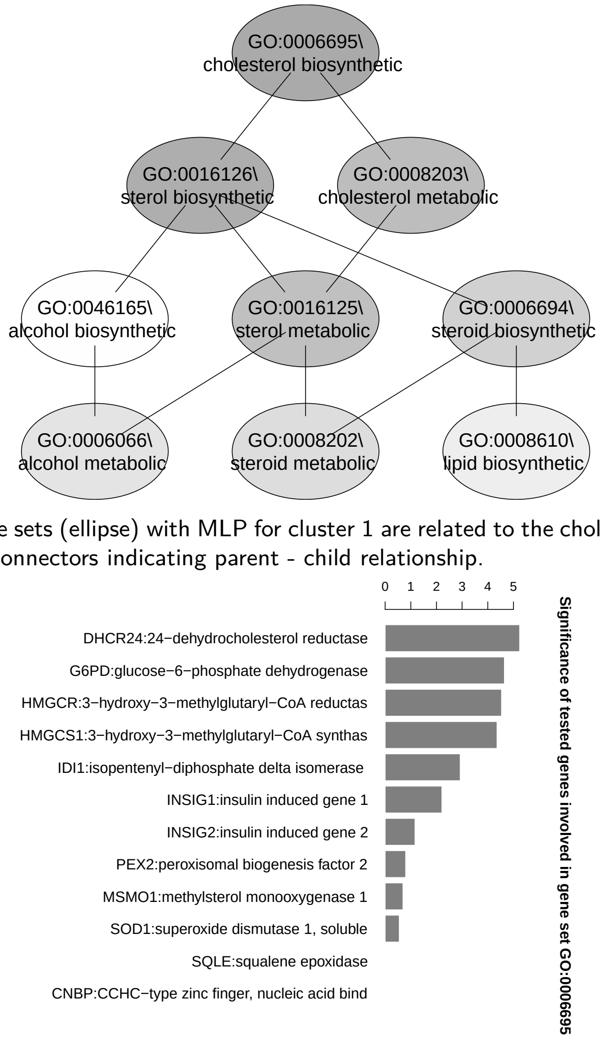
MLP analysis (Amaratunga *et al.*, 2014), in contrast to the pathway-oriented approach, does not involve pre-selection of genes prior to the analysis. Genes are categorized into gene sets according to their functional relationship. A gene set is most likely significant if many of the genes comprising that set have small p-values.

For the antipsychotic cluster, MLP indicated that the “steroid metabolic process" pathway was significantly enriched. Enrichment was also observed for the related pathways “cholesterol biosynthesis process", “sterol biosynthesis process", “cholesterol metabolic process", “sterol metabolic process" and “steroid biosynthesis process" with their interconnections visualized by using the GO graph in Figure 5.4a. The gene Dhcr24, as shown in Figure 5.4b, is predicted to be highly significant on the ”cholesterol biosynthetic process” and is known to code for the protein cholesterol-synthesizing enzyme seladin-1, which agrees with the study by Crameri *et al.* (2006), Wechsler *et al.* (2003). Another gene in the list, G6PD, was also known to regulate the pathway through protein sterol regulatory element-binding proteins (SREBP) (Horton *et al.*, 2002). Studies by Iskar *et al.* (2013) have shown that the genes LDLR, INSIG1, IDI1, SQLE and HMGCS1 are responsible for the “cholesterol metabolic process", which is in accordance with our results. As stated by Polymeropoulos *et al.* (2009), “activation of antipsychotics by genes associated with lipid homeostasis is not just a common off-target effect of these drugs but rather the common central mechanism by which they achieve their antipsychotic activity."

5.3 Biclustering with FABIA

The data analysis process, discussed in the previous section, was based on cluster analysis of the compounds of interest according to target prediction profiles and then identification of differentially expressed genes for a cluster(s) of interest. This means that we identified a subset of genes for which a similar expression profiles were detected across a subset of compounds. In other words, we identified a bicluster in the expression matrix $\mathbf{X}$, although, we did not use it so far in this chapter.

In this section, we apply the FABIA method to the expression matrix and show the similarity and the difference between the results obtained from the biclustering and the integrative approach in the previous section. Let $\mathbf{X}$ be the expression matrix given by



(a) Three top gene sets (ellipse) with MLP for cluster 1 are related to the cholesterol biosynthetic process with the connectors indicating parent - child relationship.

(b) Significance plot of the top genes contributing to cholesterol biosynthetic process. The height of a bar represents -$\log_{10}$(geneStatistic) of the gene indicated on the y-axis. Unlike the overlap method for pathways search, which uses a short list of annotated genes and targets, MLP makes use of all the p-values obtained from LIMMA analysis to identify gene sets enriched in small p-value.

Figure 5.4: MLP analysis for cluster 1.

$$
\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{G1} & X_{G2} & \dots & X_{Gn} \end{pmatrix}. \tag{5.3}
$$

Here, the number of genes $G = 2434$ and the number of compounds $n = 35$. The FABIA model assumes that the expression matrix can be expressed as

$$
\mathbf{X} = \sum_{i=1}^{p} \lambda_i {\gamma_i}^T + \Upsilon,
$$

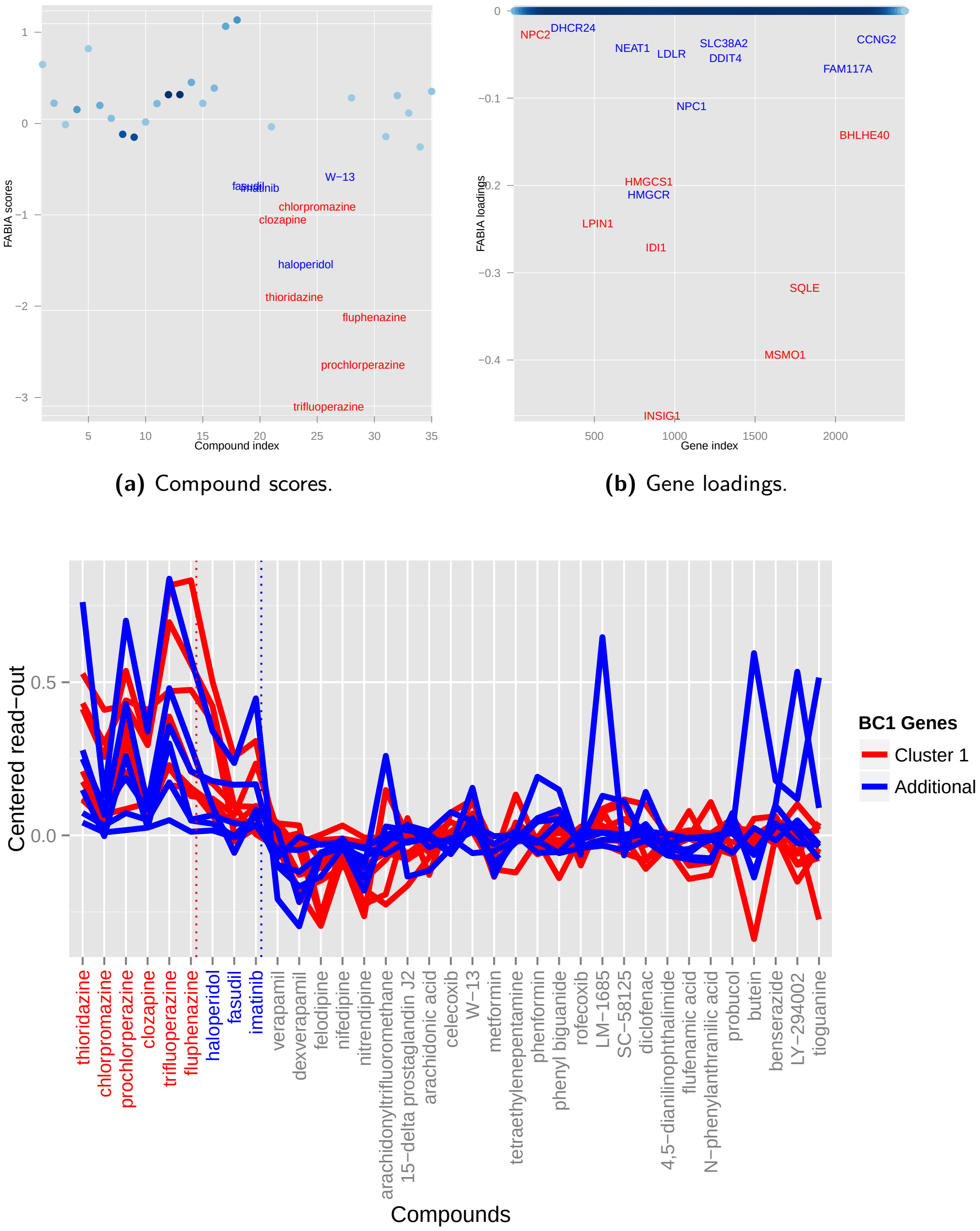
where p is the number of biclusters, the additive noise is Υ, λ_i and γ_i is a sparse vector of factor loadings and factor scores, respectively, for the ith bicluster.

From the FABIA output, we search for the bicluster that contains compounds similar to cluster 1 in the previous section. The plot of compound scores and gene loadings of FABIA bicluster 1 are presented in Figure 5.5. In addition to the 6 antipsychotic drugs marked in red in Figure 5.5a, 3 more compounds (marked in blue) are part of bicluster 1.

Moreover, all cluster 1-related genes (marked in red in Figure 5.5b) along with 6 additional genes (marked in blue) belong to the first bicluster discovered by FABIA. The expression profiles of all these genes are presented in Figure 5.5c where the additional 6 genes are indicated by the blue line. The compounds that are part of cluster 1 are marked in red while the 3 additional compounds are marked in blue. These 3 extra compounds are part of other clusters indicating that they have different target prediction profiles than cluster 1. While FABIA provides a simultaneous local search of a subset of compounds defined by a subset of genes, the integrated approach in the previous section provides a subset of genes that are linked to an independently derived compound cluster using another source of information, in this case, the similarity of their target profiles.

5.4 Integrated Analysis Results for the Other Clusters

The two previous sections illustrated the similarities of the results of the two approaches. In this section, we present more results using the integrated workflow for gene expression and target prediction data. Here, we included more compounds in the analysis by considering other doses around 10 μM. The MoA for 6 of the 8 MCF7 clusters and 6 of the 11 PC3 clusters, are established using this integrated approach.



(a) Compound scores.

(b) Gene loadings.

(c) Profiles plot of the genes belonging to the first bicluster discovered by FABIA. Eight genes (red lines) and six compounds (red text) that are members of this bicluster are also identified to be linked in the integrated approach.

Figure 5.5: FABIA bicluster 1.

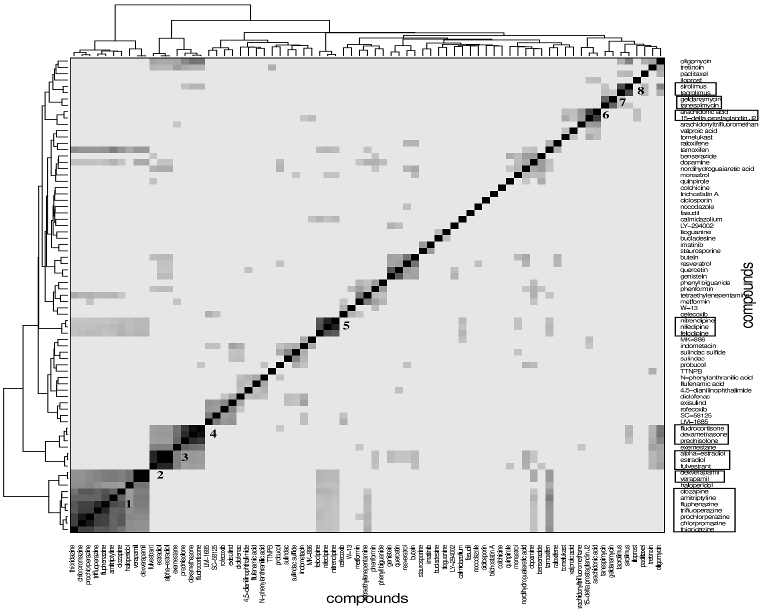
The hierarchical clustering of compounds according to the similarity of their target prediction profiles, based upon the 477 ChEMBL targets, is presented in Figure 5.6 for the MCF7 and PC3 cell lines. Several interesting target-based compound clusters (with Tanimoto coefficient > 0.5) are identified in each cell line; 8 from MCF7 and 11 from PC3 (numbered in their respective heatmaps). The target prediction data depends upon the structural make-up of the compounds; hence a compound cluster observed in one cell line will also hold for another cell line, given that all member compounds are present in both cell lines. This is the case for a cluster common to MCF7 and PC3, which contains compounds estradiol, alpha-estradiol and fulvestraat.

5.4.1 Benzoquinone Antineoplastic Antibiotic

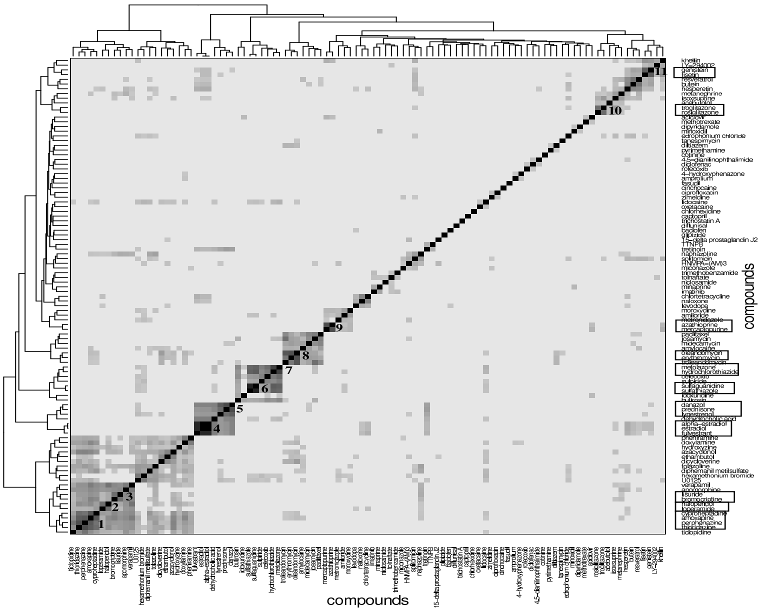
One compound cluster consists of the compounds geldanamycin and tanespimycin from cluster 7 of the MCF7 cell line. Both compounds are benzoquinone antineoplastic antibiotics, which are used to inhibit the function of heat shock protein 90 (HSP90) (Modi *et al.*, 2011, Taldone *et al.*, 2008). In Figure 5.7 the top differentially expressed genes between these 2 compounds and other compounds in the set are displayed. Gene HSPA1B shows a perturbation of an above 2-fold change in both the compounds and also shows a -log(p-value) greater than 50 in the volcano plot (Figure 5.7). The HSP90AB1 gene shows a fold change above 1.2 in both the compounds and was found to be statistically significant. Other genes such as HSPA6, HSPA4L, DNAJB4, HSPH1 were all found to be statistically significant with a fold change above 1. Literature shows that protein HSP90 is encoded by the HSP90AB1 gene (Chen *et al.*, 2005). The compounds are seen to perturb similar type of genes, thus showing that clustering compounds based upon targets is useful in bringing compounds of similar therapeutic class together.

Clustering of compounds based on protein target similarity is presented in Figure 5.8a, highlighting the cluster of geldanamycin and tanespimycin. This clustering is identical to that presented in Figure 5.6a, with geldanamycin and tanespimycin as cluster 7. Figure 5.8b represents the set of protein targets that are likely to bind to these two compounds, based upon the results from protein-target prediction. The expression profile plot for the top differentially expressed genes of this compound cluster clearly shows these two compounds induce a relatively higher expression than the rest (Figure 5.8c). The ordering of compounds in the x-axis is the same for all plots. The top 5 protein targets are Transcription factor AP-1 (AP-1), Transient receptor potential cation channel subfamily V member 1 (TPCC), Tyrosine protein kinase BTK (BTK), Heat shock protein HSP90 alpha (HSP90), Protein kinase C zeta type (PKCZ) and G-protein coupled receptor 55 (GPCR).

In the studied cluster, the pathway “response to the unfolded protein" (GO:006986)



(a) MCF7 cell line (75 compounds).



(b) PC3 cell line (101 compounds).

Figure 5.6: Heatmaps with dendrograms showing compound similarity scores based upon protein target prediction data for the (a) MCF7 and (b) PC3 cell lines. The colour is scaled such that darker colours represent increased similarity among the compounds, while similarities below the 90th percentile are represented in white. Compound clusters with high Tanimoto coefficient (> 0.5) are identified and numbered in the heatmaps, leading to 8 subclusters for MCF7 and 11 clusters for PC3. Subcluster 3 of MCF7 and subcluster 4 of PC3 cell line represent the same set of 22 compounds present in both the datasets. Given that these compounds share similar predicted protein targets, they form defined therapeutic type sub-clusters.

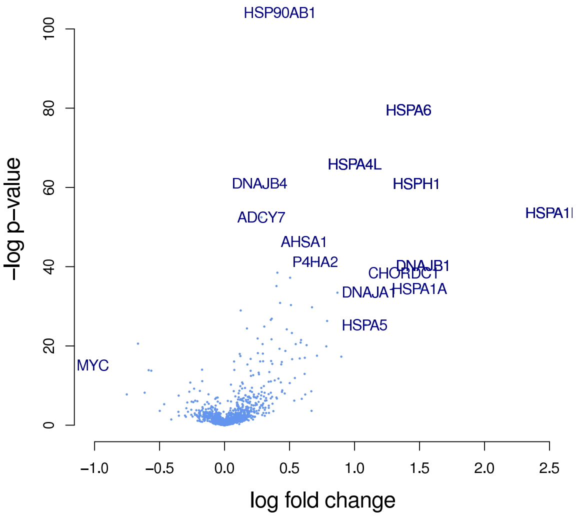
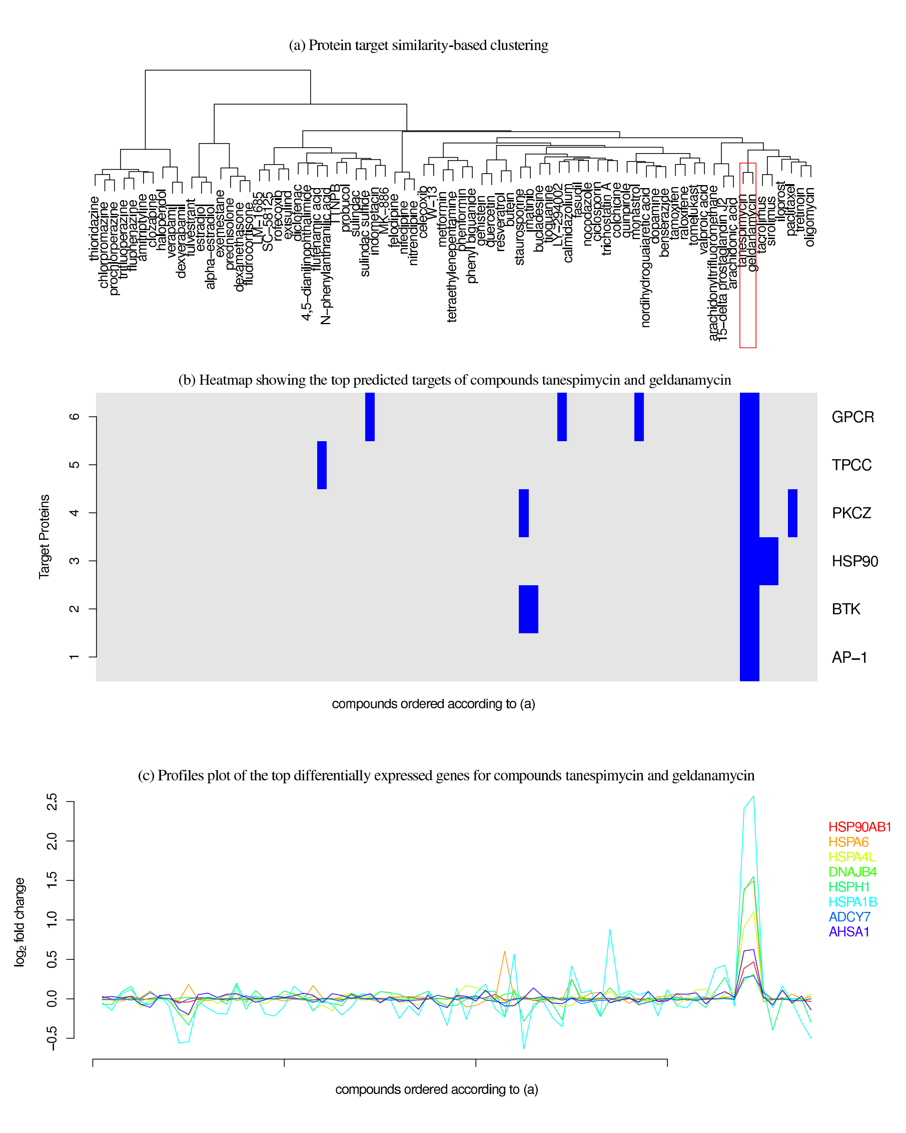


Figure 5.7: Volcano plot. -log(p-value) vs. fold change. Every gene is represented by a dot in the graph. Genes HSPA6, HSPA4L, DNAJB4 and HSPH1 at the top have the smallest P-values (*i.e.* the highest evidence for statistical significance) when testing for differentially expressed genes between the cluster of interest and other compounds in the set. Genes at the left and right-hand sides of the graph have the largest effect size (fold-change). The HSP-related genes for sub-cluster containing benzoquinone antineoplastic antibiotic compounds in the plot are seen to be highly significant, thus suggesting their role in the MoA of these compounds.

was found to be an overlapping pathway involving the predicted protein target heat shock protein HSP90 alpha and the genes HSP90AB1, HSPA6, HSPA4L, DNAJB4, HSPA1B and DNAJB1. Literature has also shown that HSP90 inhibition is associated with the activation of unfolded protein response. Moreover, the compound geldanamycin is a known inhibitor of HSP90, thus modulating the unfolded protein response (Davenport *et al.*, 2007).

Similarly, the overlap between HSP protein and the genes HSP90B1, HSPA1B, HSPA1A and HSP90AA1 shows response to the KEGG pathway “antigen processing and presentation". The genes and proteins in the overlap are known to be involved in these pathways as shown in Table 5.2. A study carried out by Albert (2004) also supports our finding that HSP plays a role in antigen processing and presentation, where these proteins are released during cell death in order to bind to cell surface receptors of the antigen-presenting cells.

The top 5 set of significant GO terms according to their structure in the ontology are displayed in Figure 5.9a. The MLP results agree with those from the pathway search and literature on the pathway “response to unfolded protein", which is on the top gene



(a) Protein target similarity-based clustering

(b) Heatmap showing the top predicted targets of compounds tanespimycin and geldanamycin

(c) Profiles plot of the top differentially expressed genes for compounds tanespimycin and geldanamycin

Figure 5.8: Genes and Protein targets regulated by compounds geldenamycin and tanespimycin. (a) Protein-target similarity-based hierarchical clustering of compounds; (b) heatmap of the proteins target (rows) and compounds(columns) coloured according to activation/inactivation of protein targets; (c) the profile plot of the top differentially expressed genes with compounds ordered according to (a) in the x-axis and fold-change in the y-axis. The selected compound sub-cluster contains the only compounds that predicted the targets represented in blue. Thus, some genes (HSP90AB1, HSPA6, HSPA4L, DNAJB4, HSPH1, HSPA1B, ADCY7 and AHSA1) are particularly perturbed with respect to the sub cluster selected.

Table 5.2: Overlapping Pathways. Pathway Search involving the top protein targets and genes regulated by the compounds geldenamycin and tanespimycin.

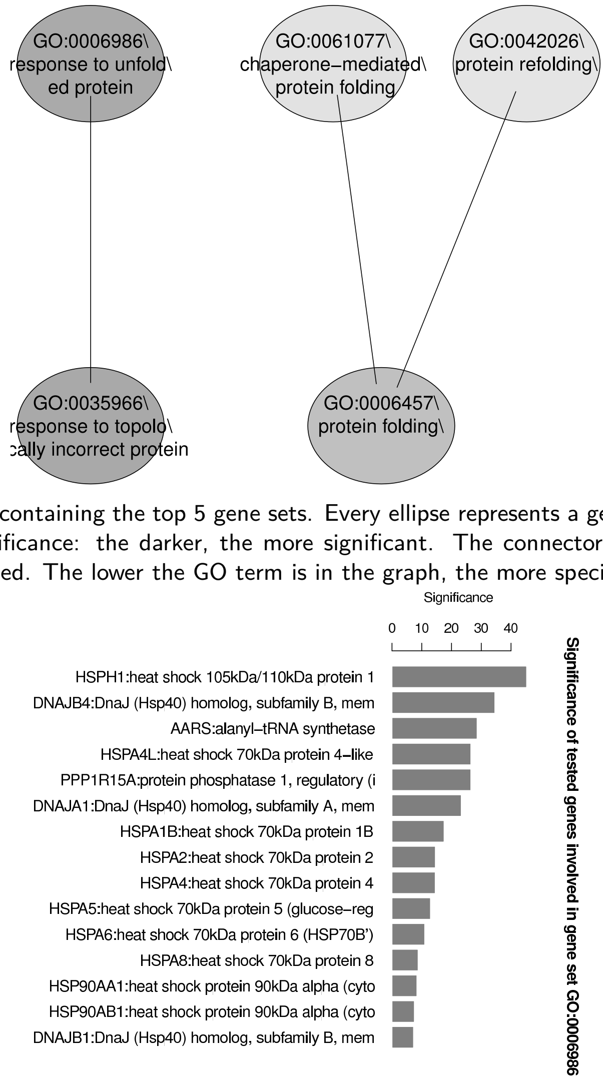
Pathway	Target	Genes
response to unfolded protein	Heat shock protein 90 alpha	HSP90B1
		HSPA6
		HSPA4L
		DNAJB4
		HSPA1B
antigen processing and presentation	Heat shock protein 90 alpha	HSP90B1
		HSPA1B
		HSPA1A
		HSP90AA1

set in the analysis (Davenport *et al.*, 2007). The pathway search provides information on known existing gene-pathway links, whereas MLP analysis shows statistically enriched pathways that are significant (with or without available literature evidence). While the pathway search makes use of top differentially expressed genes, providing 5 genes linked to this pathway, the MLP analysis can provide an enriched set of genes biologically linked through the "response to unfolded protein" pathway. Using the LIMMA p-values as the input, the HSP and DNAJ-related genes are shown to dominate this gene set (Figure 5.9b).

The MLP method therefore provides statistically significant genes and also the significance of each gene in the pathway of interest. The gene set enrichment analysis is a good start when there is limited pathway information, in understanding the MoA of compounds.

5.4.2 Antidiabetic and Anti-inflammatory Drugs

A PC3 cell line cluster (Table 5.4) comprising of thiazolidinediones (rosiglitazone and troglitazone drugs) was found to have both antidiabetic and anti-inflammatory effects(Mahindroo *et al.*, 2005, Fryer *et al.*, 2002). In silico target prediction algorithm indicated that these compounds were likely to bind to the peroxisome proliferator activated receptor gamma (PPAR-gamma), peroxisome proliferator activated receptor alpha (PPAR-alpha) and acyl CoA desaturase. Spiegelman (1998) has shown the MoA of antidiabetic thiazolidinediones to induce activation of PPAR gamma and thus regulate genes involved in glucose and lipid metabolism. Gene expression profiles of genes FABP4 and ANGPTL4 have fold changes of 3 and 1 respectively. Studies have shown that antidiabetic thiazolidinediones are ligands for the nuclear receptor PPAR, which exert their anti-hyperglycaemic effects by regulation of the PPAR responsive genes and also that gene FABP4 is rapidly up-regulated upon PPAR gamma ligand administration; this confirms our



(a) GO pathways containing the top 5 gene sets. Every ellipse represents a gene set. The colour indicates the significance: the darker, the more significant. The connectors indicate that the gene sets are related. The lower the GO term is in the graph, the more specific is the gene set.

(b) Significance plot of top functionally related genes contributing in the pathways “response to unfolded proteins". The plot represents the top 15 genes contributing with the level of significance in the bar for the respective pathways in the MLP analysis for geldanamycin and tanespimycin compound cluster.

Figure 5.9: MLP analysis for benzoquinone antineoplastic antibiotic compounds.

finding of this gene showing high fold change (Kapushesky *et al.*, 2011, Szatmari *et al.*, 2006). A study by Pal *et al.* (2011) showed that gene ANGPTL4 is responsible for epidermal differentiation mediated via the PPAR protein.

During overlap pathway analysis, genes FABP4 and ANGPTL4 were found to share pathway “PPAR signalling" with proteins PPAR-gamma, PPAR-alpha and acyl CoA desaturase. Confirming our observation, antidiabetic thiazolidinediones in pathway hsa03320 of the KEGG database induce “PPAR signaling pathway" by perturbing genes FABP4 and ANGPTL4 and PPAR proteins. This indicates that the MoA of antidiabetic thiazolidinediones involves PPAR signalling.

There was overlap of the pathway “induction of apoptosis" with gene PRKCD and protein target PPAR-gamma. In the study by Elrod and Sun (2008), thiazolidinediones were shown to have potential for inducing apoptosis in cancer cells by binding to protein PPAR-gamma. In our study on thiazolidinediones, we also observed that gene PRKCD is down-regulated substantially when compared to other compounds in the dataset showing selectivity for this particular gene. Hence suggesting gene PRKCD to be involved in the MoA for thiazolidinediones. Some of the links (compound-genes, compound-target and genes-pathway-target), however, lacked literature support (Table 5.3 for MCF7 and Table5.4 for PC3). The target prediction similarity data also produces many singletons, which are compounds that do not share any targets with remaining compounds in the set, thus providing a limited number of clusters to be investigated.

5.5 Discussion

Combining target-based compound similarity with corresponding gene expression information provides a better understanding of compound cluster behavior, both on the bioactivity level and on the transcriptional level. In this chapter, we present an integrated statistical framework to link two data sources: the gene expression data and target prediction data of a set of compounds. Strictly speaking, the analysis presented in Section 5.2 is not a biclustering analysis but a two-stage analysis in which, the first stage consists of cluster analysis and the second stage consists of identification of differentially expressed genes (for a given cluster). However, the output of the analysis in Section 5.2 consists of a subset of compounds (which form the cluster of interest) and a subset of genes (which were found to be differentially expressed). Using the terminology of this book, the two subsets of genes and compounds form a bicluster in the expression matrix. In Section 5.3, we have shown that we can use FABIA, as an alternative approach, to identify the bicluster.

The extra information provided by the target prediction data in clustering compounds restricts the overlapping of compounds that is observed in the biclustering approach. That is, compounds inducing the same level of expression in a set of genes but predicting different targets cannot be part of the target-driven ‘bicluster’. This is the case for the 3 extra compounds obtained in FABIA that are not clustered with the antipsychotic drugs. Haloperidol, one of the three compounds, however, is clustered next to the antipsychotic drugs, indicating that they still share some targets.

Table 5.3: Overlapping Pathways for MCF7 cell line. Pathway Search involving the top protein targets and genes regulated by distinct compound set. MoA of the compound cluster is comprehended through pathway overlap between the significant genes and predicted protein targets, as shown in the table. The compound clustering based upon similar targets lead to sub-clustering of compounds with similar therapeutic classes. Cluster 1 containing antipsychotic drugs and cluster 7 containing compounds geldanamycin and tanespimycin, were studied in detail. As an example, in cluster 1 our method suggests genes INSIG1 and LDLR and target CYP450, share pathway "steroid metabolic process" for the listed antipsychotic compounds. This study was also supported by the findings of Polymeropoulos *et al.* (2009)

No.	Compounds	Pathway	Target	Genes
1	amitriptyline clozapine thioridazine chlorpromazine trifluoperazine prochlorperazine fluphenazine	Steroid metabolic process	Cytochrome P450 2D6	INSIG1 LDLR
2	verapamil dexverapamil	MAPK signaling pathway	Voltage gated T type calcium channel alpha 1G subunit	DUSP9
3	estradiol alphaestradiol fulvestrant	**no overlapping pathways found**		
4	dexamethasone prednisolone fludrocortisone	**no overlapping pathways found**		
5	nifedipine nitrendipine felodipine	Aging	Induced myeloid leukemia cell differentiation protein Mcl.1	IFIT1
6	15-delta prostaglandin J2 arachidonic acid	Arachidonic acid metabolism	Thromboxane A synthase	AKR1C3
7	Tanespimycin geldanamycin	Response to Unfolded Protein	Heat shock protein HSP 90 alpha	HSP90AB1 HSPA6 HSPA4L DNAJB4 HSPA1B DNAJB1 DNAJA1 HSP90AA1
		Antigen processing and presentation	Heat shock protein HSP 90 alpha	HSP90AB1 HSPA1B HSPA1A HSP90AA1
8	tacrolimus	Activation of cysteine-type endopeptidase activity involved in apoptotic process	Proteinase activated receptor 1	MOAP1
		Negative regulation of cell proliferation	Proteinase activated receptor 1	S100A11
	sirolimus	Regulation of actin cytoskeleton	Proteinase activated receptor 1	PFN1
		Prostate cancer	Heat shock protein HSP 90 alpha	GSK3B

Table 5.4: Overlapping Pathways for PC3 cell line. Pathway Search involving the top protein targets and genes regulated by compound set. MoA of the compound cluster is comprehended through pathway overlap between the significant genes and predicted protein targets as mentioned. The compounds clustering based on similar targets lead to sub-clustering of compounds with similar therapeutic class. For example in cluster 10, our method suggests genes FABP4 and ANGPTL4 and targets PPAR-gamma, PPAR-alpha and acyl CoA desaturase, share pathway "PPAR signalling" for the listed thiazolidinediones and findings confirmed by KEGG database pathway(hsa03320).

No.	Compounds	Pathways	Target	Genes
1	thioridazine amoxapine cyproheptadine perphenazine	Protein modification process	Muscarinic acetylcholine receptor M3	HERPUD1
2	loperamide haloperidol	**no overlapping pathways found**		
3	bromocriptine lisuride	**no overlapping pathways found**		
4	estradiol alpha-estradiol fulvestrant	**no overlapping pathways found**		
5	prednisone	Neuroactive ligand-receptor interaction	Glucocorticoid receptor	GHR
	lynestrenol	Focal adhesion	Protein kinase C alpha	BIRC3
	danazol	MAPK signaling pathway	Protein kinase C alpha	IL1R2
6	sulfathiazole sulfaguanidine	Small cell lung cancer	Cyclin dependent kinase2	LAMB3
7	hydrochlorothiazide metolazone	**no overlapping pathways found**		
8	erythromycin oleandomycin	**no overlapping pathways found**		
		Cytokine-cytokine receptor interaction	Vascular endothelial growth factor receptor2	IL17RA
9	mercaptopurine azathioprine	ErbB signaling pathway Wnt signaling pathway Colorectal Cancer Endometrial Cancer	Glycogen synthase kinase.3 beta	MYC
		p53 signaling pathway	Cyclin dependent kinase 1	THBS1
		Focal Adhesion	Glycogen synthase kinase.3 beta Vascular endothelial growth factor receptor2	THBS1
		Induction of apoptosis	Peroxisome proliferator.activated receptor gamma	PRKCD
10	rosiglitazone	Positive regulation of transcription from RNA polymerase II promoter	Peroxisome proliferator.activated receptor gamma	TNFRSF1A
	troglitazone	PPAR signaling pathway	Peroxisome proliferator.activated receptor gamma Peroxisome proliferator.activated receptor alpha Acyl.CoA desaturase	FABP4 ANGPTL4
		Adipocytokine signaling pathway	Peroxisome proliferator.activated receptor alpha	TNFRSF1A
11	fisetin	Wnt signaling pathway	Glycogen synthase kinase.3 beta	VANGL1 DKK1
	genistein	Axon guidance	Glycogen synthase kinase.3 beta	EPHA2

Chapter 6

Ranking of Biclusters in Drug Discovery Experiments

6.1 Introduction

In Chapter 5, the biclustering of gene expression data was introduced. Ideally, we would like to examine all biclusters that were discovered by an algorithm. However, in many cases, a large number of biclusters are reported in a bicluster solution. This implies that a procedure to prioritize biclusters, irrespective of biclustering algorithm is needed. Clearly, one open question is how to determine which biclusters are most informative and rank them on the basis of their importance. In many studies, biclusters are empirically evaluated based on different statistical measures (Koyuturk *et al.*, 2004) or biologically validated based on gene ontology annotations or other literature-based enrichment analysis (Bagyamani *et al.*, 2013). In FABIA, for example, biclusters are ranked according to the information they contain. Pio *et al.* (2013) proposed to rank biclusters based on the p-values of a statistical test, which compares functional similarity within and outside the bicluster using a similarity measure computed according to the genes' annotations in GO. Kidane *et al.* (2013), on the other hand, computed bicluster enrichment score in drug targets in order to prioritize biclusters.

In early drug discovery studies, biclustering of gene expression data can be routinely applied to guide compound prioritization and gene module detection. Typically, early drug discovery data involves not only gene expression data but also other information related to the chemical structures and bioactivities properties of the set of compounds under development. These data can be combined and mined in order to prioritize biclusters.

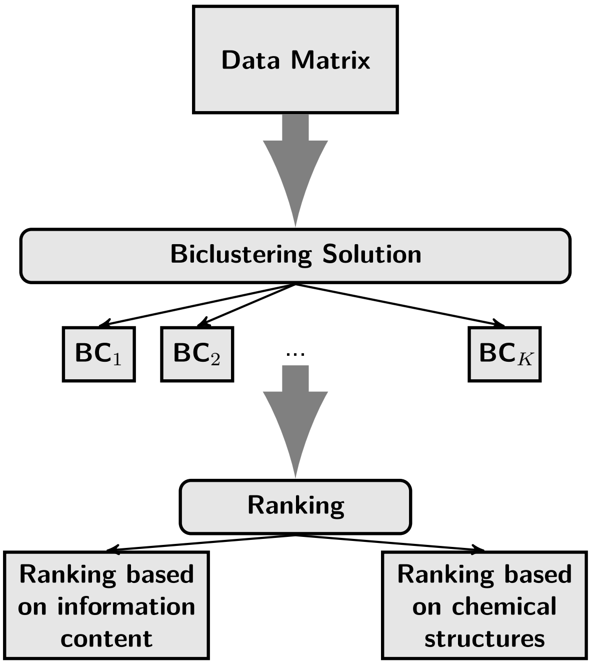

Data Matrix
Biclustering Solution
BC$_1$
BC$_2$
...
BC$_K$
Ranking
Ranking based on information content
Ranking based on chemical structures

Figure 6.1: Ranking of biclusters.

In this chapter, we present two approaches to evaluate and prioritize biclusters as illustrated in Figure 6.1. In the first approach, presented in Section 6.2, we rank biclusters based on information content. In Section 6.3 we discuss a second ranking approach in which the biclusters are ranked based on chemical structure. The R package `biclustRank`, discussed in Chapter 10, is used to rank the biclusters and to visualize the results.

6.2 The Information Content of Biclusters

6.2.1 Theoretical Background

We consider a FABIA model (for a solution with K biclusters) of the form

$$X \;=\; \sum_{k=1}^{K} \lambda_k \, \gamma_k^T \;+\; \epsilon \;=\; \Lambda \, \Gamma \;+\; \epsilon \;, \tag{6.1}$$

where $\epsilon \in \mathbb{R}^{N \times M}$ is additive noise; $\lambda_k \in \mathbb{R}^N$ is the vector of feature memberships to the k-th bicluster and $\gamma_k \in \mathbb{R}^M$ is the vector of sample memberships to the k-th bicluster (Hochreiter *et al.*, 2006).

According to Eq. 6.1, the j-th sample x_j, i.e., the j-th column of X, is

$$x_j \ = \ \sum_{k=1}^{K} \lambda_k \ \gamma_{kj} \ + \ \epsilon_j \ = \ \Lambda \ \tilde{\gamma}_j \ + \ \epsilon_j \ , \tag{6.2}$$

where ϵ_j is the j-th column of the error matrix ϵ and $\tilde{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{Kj})^T$ denotes the j-th column of the matrix Γ. In this model, ϵ is $N(0, \Psi)$-distributed, $\tilde{\gamma}_j \sim N(0, \Xi_j)$ where the covariance matrix Ξ_j is diagonal, $x_j \sim N(0, \Psi + \Lambda \Xi_j \Lambda^T)$ and $x_j | \tilde{\gamma}_j \sim N(\Lambda \ \tilde{\gamma}_j, \Psi)$.

FABIA allows to rank the extracted biclusters analogously to principal components which are ranked according to the data variance they explain. Biclusters are ranked according to the information they contain about the data. As shown in (Hochreiter *et al.*, 2006), the information content of $\tilde{\gamma}_j$ for the j-th observation x_j is the mutual information between $\tilde{\gamma}_j$ and x_j as

$$\mathrm{I}(x_j; \tilde{\gamma}_j) \ = \ \mathrm{H}(\tilde{\gamma}_j) \ - \ \mathrm{H}(\tilde{\gamma}_j \mid x_j) \ = \tfrac{1}{2} \ln \left| I_K \ + \ \Xi_j \ \Lambda^T \ \Psi^{-1} \ \Lambda \right| \ , \tag{6.3}$$

where H is the entropy. The independence of x_j and $\tilde{\gamma}_j$ across j gives

$$\mathrm{I}(X; \gamma) \ = \ \tfrac{1}{2} \sum_{j=1}^{M} \ln \left| I_K \ + \ \Xi_j \ \Lambda^T \ \Psi^{-1} \ \Lambda \right| \ . \tag{6.4}$$

To assess the information content of one factor, factor $\tilde{\gamma}_k$ is removed from the final model and, consequently, the explained covariance $\xi_{kj} \ \lambda_k \ \lambda_k^T$, where ξ_{kj} is the jth column in the covariance matrix Ξ_j, must be considered as noise:

$$x_j \mid (\tilde{\gamma}_j \setminus \gamma_{kj}) \ \sim \ \mathcal{N}\big(\Lambda \ \tilde{\gamma}_j|_{\gamma_{kj}=0} \ , \ \Psi \ + \ \xi_{kj} \ \lambda_k \ \lambda_k^T\big) \tag{6.5}$$

The information of γ_{kj} given the other factors is

$$\begin{aligned} \mathrm{I}\big(x_j; \gamma_{kj} \mid (\tilde{\gamma}_j \setminus \gamma_{kj})\big) \ &= \ \mathrm{H}(\gamma_{kj} \mid (\tilde{\gamma}_j \setminus \gamma_{kj})) - \mathrm{H}(\gamma_{kj} \mid (\tilde{\gamma}_j \setminus \gamma_{kj}), x_j) && (6.6) \\ &= \ \tfrac{1}{2} \ln \big(1 \ + \ \xi_{kj} \ \lambda_k^T \Psi^{-1} \lambda_k\big) \ . && (6.7) \end{aligned}$$

Again independence across j gives

$$\mathrm{I}\big(X; \gamma_k^T \mid (\gamma \setminus \gamma_k^T)\big) \;=\; \tfrac{1}{2} \sum_{j=1}^{M} \, \ln\big(1 \;+\; \xi_{kj} \; \lambda_k^T \Psi^{-1} \lambda_k\big) \; . \tag{6.8}$$

This information content gives that part of the information in x that γ_k^T conveys across all examples. Note that the information content grows with the number of nonzero λ_k's (size of the bicluster).

6.2.2 Application to mGluR2PAM Project

For illustration, we use the mGluR2PAM dataset. The gene expression matrix $\mathbf{X}$ consists of $J = 566$ genes and $n = 62$ compounds,

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,62} \\ X_{2,1} & X_{2,2} & \dots & X_{2,62} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{566,1} & X_{566,2} & \dots & X_{566,62} \end{pmatrix} .$$

Using the mGluR2PAM data matrix as input, we analyze the data using the `fabia` package with `p=10` biclusters. For example, the first panel in Fig. 6.2 shows the information content of 10 biclusters. We notice that the information content of the first bicluster in the solution (361.15) is bigger than the information content of the other biclusters.

6.3 Ranking of Biclusters Based on their Chemical Structures

6.3.1 Incorporating Information about Chemical Structures Similarity

In the previous section, we ranked the biclusters according to their information content. In this section, we further evaluate how homogeneous are the biclusters according to the chemical structure similarity scores of the compounds belonging to the biclusters. Hence, a bicluster in which the compounds are similar in terms of their chemical structure receives a higher rank than those with low chemical similarity.

Let $\mathbf{Z}$ be the chemical structure matrix containing F binary features, representing

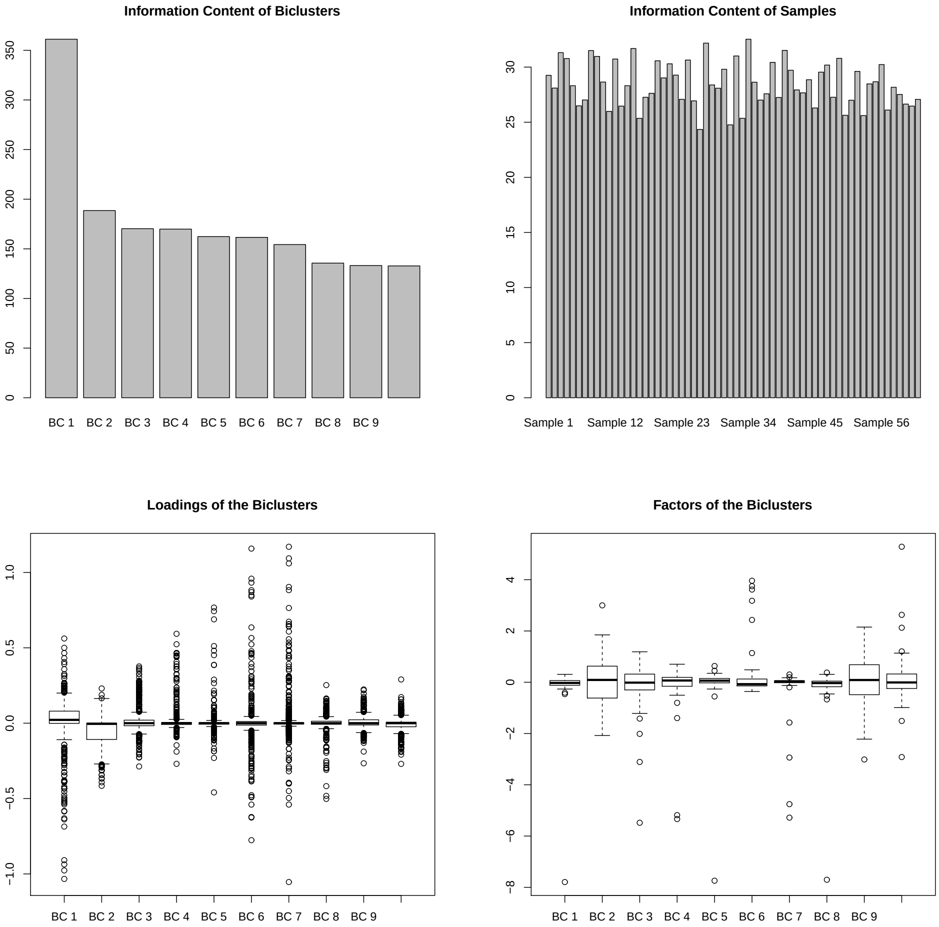


Figure 6.2: Top left: the information content of biclusters. Top right: the information content of samples. Lower left: the loadings of biclusters. Lower right: the factors of biclusters.

the chemical structure of n compounds. We can calculate a similarity score, S_{ij} for pair compounds i and j, for all $(i,j) \leq n$ using the Tanimoto statistic based on all F binary features. The similarity matrix, $\mathbf{S}_n$, is given by

$$\mathbf{Z}_{F\times n} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{21} & Z_{22} & \dots & Z_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ Z_{F1} & Z_{F2} & \dots & Z_{Fn} \end{pmatrix}, \quad \mathbf{S_n} = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{21} & S_{22} & \dots & S_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ S_{n1} & S_{n2} & \dots & S_{nn} \end{pmatrix}.$$

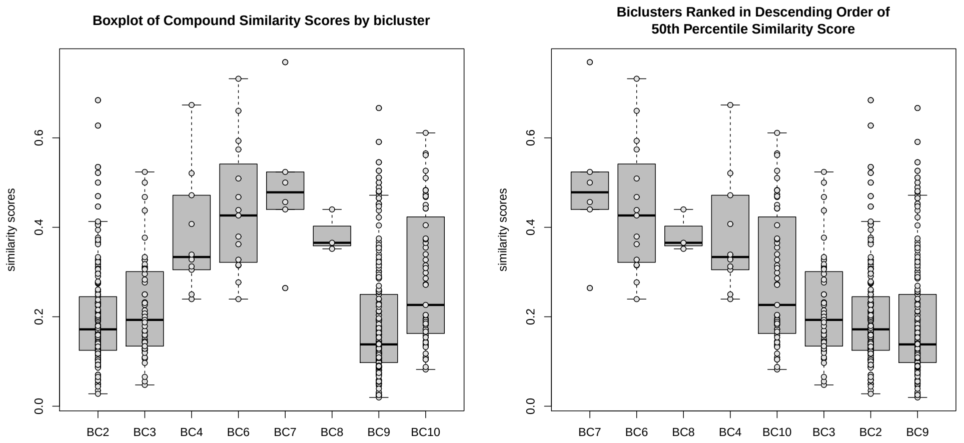
For each bicluster, we can derive from $\mathbf{S}_n$ a submatrix of similarity scores ($\mathbf{S}_k$) of the compounds belonging to the kth bicluster. In the next step, we compare the distribution of the compound's similarity scores between the biclusters. Biclusters of interest are characterized by homogeneous and relatively high similarity scores. Note that biclusters with either 1 gene or sample are excluded from the analysis.

Table 6.1: Summary statistics of the structural similarity scores by bicluster. Here, the biclusters are ordered according to the coefficient of variation (CV). The higher the CV, the greater the dispersion in the bicluster. Biclusters with 1 sample or gene are excluded.

BC	mean	median	SD	CV	MAD	Range	IQR
BC8	0.39	0.37	0.05	12.31	0.02	0.09	0.04
BC7	0.49	0.48	0.16	33.23	0.06	0.51	0.07
BC6	0.44	0.43	0.15	33.25	0.16	0.49	0.22
BC4	0.38	0.33	0.14	35.31	0.12	0.43	0.15
BC3	0.22	0.19	0.11	50.48	0.10	0.48	0.17
BC10	0.28	0.23	0.15	52.66	0.14	0.53	0.26
BC2	0.20	0.17	0.11	57.19	0.08	0.66	0.12
BC9	0.18	0.14	0.13	70.96	0.08	0.65	0.15

Table 6.1 presents the summary statistics of the similarity scores by bicluster. Biclusters with less variability is preferred. For those with comparable level of variability, biclusters with higher mean/median are of interest. The coefficient of variation which describes the amount of variability relative to the mean is used to rank the biclusters as presented in Table 6.1. Biclusters 8,7,6 and 4 have similarity scores that are less dispersed. Although compounds in biclusters 2,3,9 and 10 are regulating similar subset of genes, their structural similarity scores are less homogeneous compare to the other biclusters. Biclusters 1 and 5 are ignored since they contain only 1 compound and 1 gene, respectively.

6.3.2 Similarity Scores Plot



Boxplot of Compound Similarity Scores by bicluster

Biclusters Ranked in Descending Order of 50th Percentile Similarity Score

(a) Boxplot of similarity scores by bicluster.

(b) Biclusters are ordered according to the median value.

Figure 6.3: Priority will be given to biclusters showing consistently high similarity scores.

Several visualisation tools can be used to aid in prioritizing biclusters according to other sources of information. This is highly beneficial when there are a large number of biclusters to investigate.

Figures 6.3 (left panel) shows the distribution of the structural similarity scores of compounds by bicluster. From the perspective of early drug discovery, a group of compounds that are found to be biologically and structurally similar are of primary interest for further development. Therefore, biclusters with less variability within the group and with relatively high similarity scores (for example, BC7) will prioritized over other biclusters.

For direct visualization of the ordering of biclusters, the boxplots can be reordered according to the median similarity score as displayed in Figure 6.3b. Using the median similarity score to rank biclusters, biclusters 7,6,8 and 4 consist of the most similar subset of compounds.

Figures 6.4 shows the cumulative distribution of the similarity scores within the biclusters of interest. The distribution in biclusters 7,6,4 and 8 are shifted to the right compared to the other biclusters indicating higher similarity scores in these biclusters.

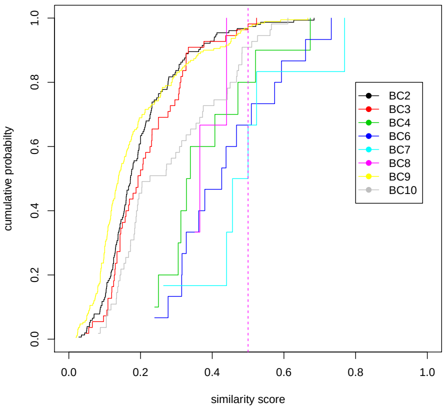
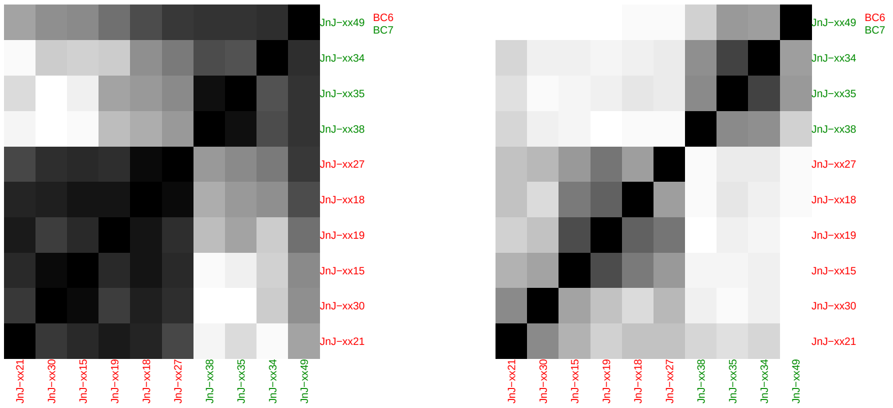


Figure 6.4: Cumulative distribution plot of the similarity scores by Bicluster. Interesting biclusters have curves that are consistently shifted to the right, i.e, with high scores.



(a) Gene expression similarities of compounds.

(b) Structural similarities of compounds.

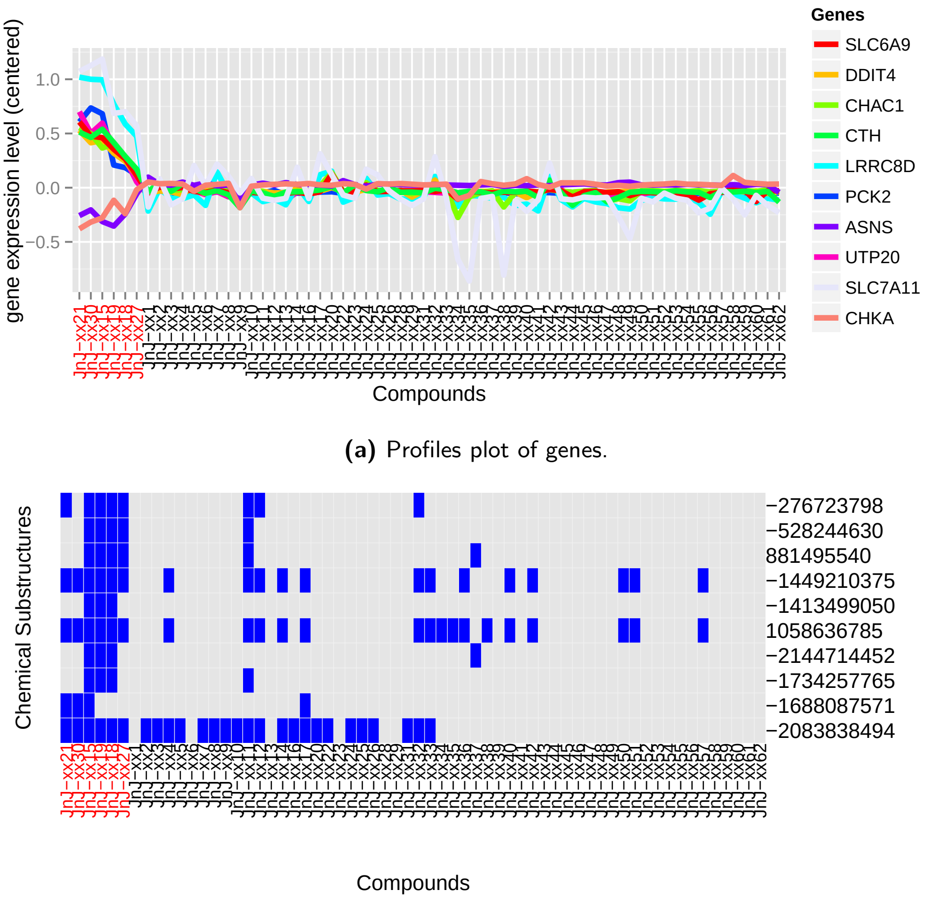
Figure 6.5: Similarity matrix of compounds from biclusters 6 and 7.

6.3.2.1 Heatmap of Similarity Scores

In Figure 6.5a, the blocks of compounds based on gene expresssion similarity for bicluster 6 (marked in red) and 7 (marked in green) are clearly reflected as well on the heatmap of the similarity scores based on chemical structure (Figure 6.5b). These two biclusters have no overlapping compounds and are interesting biclusters taking into account both data sources.

6.3.3 Profiles Plot of Genes and Heatmap of Chemical Structures for a Given Bicluster

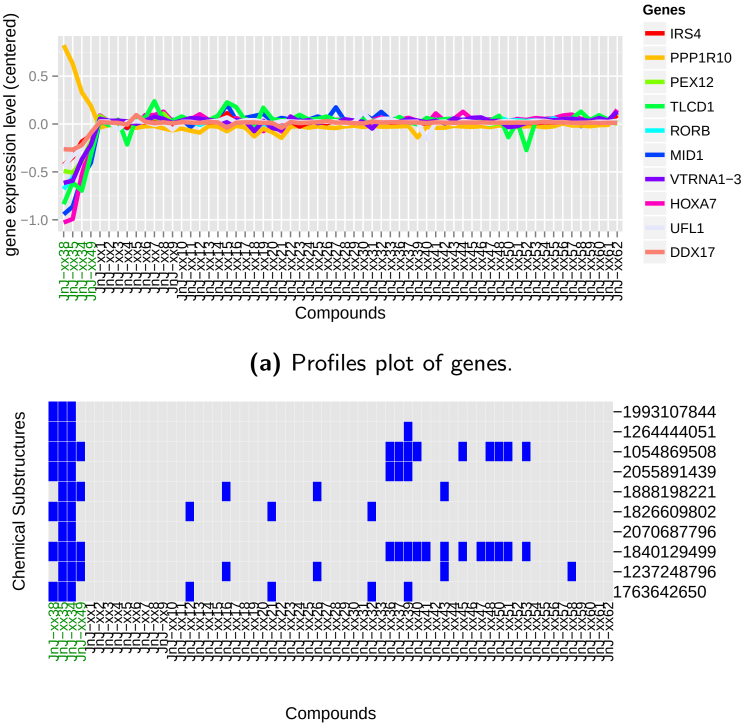
We can visualize the transcriptional profiles and the chemical structures that differentiates compounds in a bicluster from the rest (see Figure 6.6 for bicluster 6 and Figure 6.7 for bicluster 7).



(a) Profiles plot of genes.

(b) Heatmap of chemical structures.

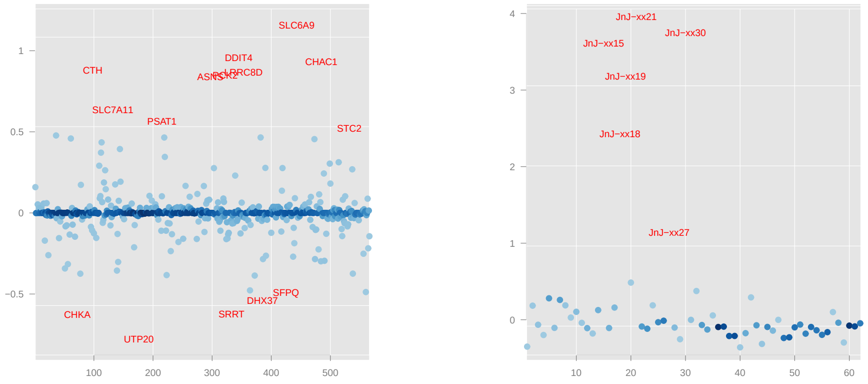
Figure 6.6: Profiles plot of genes and heatmap of chemical structures associated to bicluster 6.



(a) Profiles plot of genes.

(b) Heatmap of chemical structures.

Figure 6.7: Profiles plot of genes and heatmap of chemical structures associated to bicluster 7.



(a) Gene loadings.

(b) Compound Scores.

Figure 6.8: FABIA bicluster 6.

6.3.4 Loadings and Scores

We can further examine the factor loadings and the factor scores for a bicluster of interest. These plots can help to identify genes/compound associated with the bicluster (Figure 6.8).

6.4 Discussion

Biclustering has been extensively used in extracting relevant subsets of genes and samples from microarray experiments. In drug discovery studies, more information about compounds (such as chemical structure, bioactivity properties, toxicity etc.) are available.
In this chapter, we show two approaches of ranking biclusters: (1) how FABIA ranks biclusters according to their information content, which is an information-theory based statistic and (2) how to incorporate other sources of information to prioritize biclusters. For the latter, other biclustering methods, such as the Plaid model and ISA can be used.

The integration of additional information to rank gene-expression based biclusters mainly focus on the use of similarity scores. Here, we used the Tanimoto coefficient which is commonly applied for chemical structure similarity (Willett *et al.*, 1998). However, other similarity measures could be also used which may give slightly different scores and in extreme cases may affect the ranking. This instability of the ranking is of minor concern since we do not aim to choose one bicluster but rather to prioritize interesting biclusters. In fact, we showed that different descriptive statistics may lead to a different ranks but a similar set of prioritized biclusters. Although in this application, we only generated ten biclusters to illustrate the analysis workflow, in other cases a large number of biclusters can be extracted and this exploratory approach would be useful.

In Section 6.3 we focus on ranking based on chemical structure. Of course, other sources of data can be used. The analysis presented in this chapter helps us not just to identify local patterns in the data but interpret these patterns in terms of the dimensions of interest (for example chemical structures, toxicity etc.).

Part III

Unsupervised Multivariate Data Integration

Chapter 7

Multiple Factor Analysis (MFA) for QSTAR Data Integration

7.1 Introduction

In the second part of the thesis we discussed semi-supervised methods for data integration of the QSTAR datasets. In this part of the thesis we focus on unsupervised approaches and in particular, in this chapter, we focus on Multiple Factor Analysis (MFA, Escofier and Pagés (1988)). Multiple Factor Analysis makes it possible to analyze several sets of variables simultaneously and to discover the underlying patterns which are common across the different data sets. Multiple Factor Analysis makes it possible to analyse several sets of variables simultaneously or globally with the focus on making the structure of the samples induced by these sets of variables balanced or comparable. Since the analysis consists of several datasets, measured on the same group of samples (i.e., compound), the first step of a MFA is a normalization step in which the different datasets are normalized so the variability across the different datasets becomes comparable. Without this step a single dataset can dominate the analysis. The number of variables in each dataset may differ and the nature of the variables (nominal or quantitative) can vary from one dataset to the other but the variables should be of the same nature in a given group. In the context of the QSTAR framework, the samples are the compounds and the variables in the different datasets represents gene expression levels, bioactivity read-outs and chemical structures.

MFA was first introduced by Escofier and Pagés (1988) with application on chemometrics data. It then gained popularity among researchers working on multiple datasets measured on the same observations. Recent fields of application include environmental chemistry (Stanimirova *et al.*, 2005), fisheries (Petitgas and Poulard, 2009), geology, broadcasting, neuroimaging, and also in genetics (Franco *et al.*, 2010, de Tayrac *et al.*, 2009, Pavoine and Bailly, 2007). In this chapter, we introduce a new data analysis approach for drug discovery studies using MFA. In Section 7.2 we present different aspects of the MFA method. Section 7.3 is devoted to the application of the MFA within the QSTAR context and the proposed method is applied to the EGFR data, as discussed in the previous parts of the thesis. We conclude the chapter with a discussion presented in Section 7.4.

7.2 Multiple Factor Analysis

Principal component analysis (PCA, Jolliffe (2002)) has been used to find correlated features in an unsupervised manner. In the context of ordinary PCA, only one set of variables is analysed. Performing PCA on multiple sets of variables simultaneously was introduced by Escofier and Pagés (1988) who suggested to use Multiple Factor Analysis (MFA) (Escofier and Pagés, 1990, 1988, 1983) for this setting. MFA can be seen as a weighted PCA. It is characterized by two steps: (1) dataset normalization via weighting of the inverse of the first singular value; and (2) ordinary Principal Component Analysis (PCA) analysis (frequently implemented as SVD) on the combined normalized datasets.

7.2.1 Normalization Step

Quantitative Datasets

Let $\mathbf{X}_1, \dots, \mathbf{X}_D$ be a set of D data matrices where each matrix has a dimension of $n \times m_d$, $d = 1, \dots, D$ for which the number of columns of the dth matrix is equal to m_d variables and the number of rows equal to n samples. The main idea behind MFA is the weighting of the D datasets to achieve a 'fair' data integration. Following the idea of a z-score normalization, where each variable is centered and divided by its standard deviation making all variables comparable, each dataset, $\mathbf{X}_d$, is divided by its first singular value (which can be seen as the standard deviation) so that their first principal component has the same length prior to the integrated analysis.

Recall that the singular value decomposition (SVD) of an $n \times m$ data matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I},$$

where r is the rank of $\mathbf{X}$ and Λ is a diagonal matrix with a rank-ordered set of positive singular values, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$, as elements. The matrices $\mathbf{U}$ and $\mathbf{V}$ are $n \times r$ and $m \times r$ matrices, respectively containing the orthonormal left-singular vectors $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ and the orthonormal right-singular vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_r)$. SVD decomposes $\mathbf{X}$ into a sum of r rank-one matrices $\sigma_i \mathbf{u}_i \mathbf{v}_i^T$, termed as SVD layer. Typically, only the first K layers with large σ_i values are retained to represent the data and the remaining $(r - K)$ layers are considered as less useful.

Then the size of the dth matrix, $\mathbf{X}_d$, can be measured by $\sum\limits_i \sigma_{di}^2$ where σ_{di}^2 is the eigenvalue of the ith component. In addition, the redundancy of information in the submatrix can be measured by the proportion of variance accounted for by the first principal component, given by $\sigma_{d1}^2 / \sum\limits_i \sigma_{di}^2$. Hence, matrix $\mathbf{X_d}$ can be corrected for size and redundancy using the inverse of the first singular value as weight as shown (Van Deun *et al.*, 2009):

$$\frac{1}{\sqrt{\sum\limits_i \sigma_{di}^2}} \frac{1}{\sqrt{\frac{\sigma_{d1}^2}{\sum\limits_i \sigma_{di}^2}}} X_d = \frac{1}{\sigma_{d1}} X_d.$$

Escofier and Pagés (1988) proposed to use the matrix-specific weight in order to correct for possible unwanted dominance of large matrices and to avoid that the solution is dominated by the matrix with homogeneous information. Note that this weighting does not balance the total variance of the different datasets. Thus, a set with more features will contribute to more dimensions but will not particularly contribute to the first dimension. Moreover, a matrix $\mathbf{X}_d$ that contains only a number of correlated variables can strongly contribute to only one dimension which will be the first dimension (Escofier and Pagés, 1988).

Qualitative Dataset

To integrate categorical dataset into MFA, Bécue-Bertauta and Pagés (2007) showed the equivalence between Multiple Correspondence Analysis (MCA) and PCA. MCA is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. This is done by representing data as points in a low-dimensional space. The procedure thus appears to be the counterpart of PCA for categorical data (Greenacre, 2007). In MCA, each category is associated with an indicator variable. In the case of a set of F binary variables, each variable will have two indicators giving a total of $K = 2 \cdot F$ indicator variables. Let $\mathbf{Z}$ be a data matrix of categorical variables with K

categories for which the ikth entry is given by

$$z_{ik} = \begin{cases} 1 & \text{sample } i \text{ belongs to the category } k, \\ 0 & \text{otherwise.} \end{cases}$$

The results of MCA applied on matrix $\mathbf{Z}$ can be obtained by performing a PCA on

$$\frac{(z_{ik} - w_k)}{w_k}$$

where $w_k = \sum_i^n p_i \cdot z_{ik}$. Here, p_i is the weight assigned for the ith sample which is usually uniform with $p_i = 1/n$. Hence, w_k is simply the mean of the kth column. Note that the $\sum_k^K w_k = F$ where F is the total number of binary variables. Finally, the kth column is weighted by w_k/F to have a PCA-MCA equivalent results. Hence, following the discussion in the previous subsection, we use $1/\sigma_{d1}$ for the weight of $\mathbf{X}_d$ in the case of quantitative set and $(w_{k_d}/F_d)/\sigma_{d1}$ in the case of a categorical dataset.

7.2.2 Simultaneous Analysis

Once the data matrices are normalized following the procedure discussed in the previous section, MFA can be applied for the combined normalized data.

Let $\mathbf{X} = [\mathbf{X}_1|\dots|\mathbf{X}_d|\dots|\mathbf{X}_D]$. The number of samples (rows) is n, which is the common dimension among the D matrices, and the total number of variables (columns) in $\mathbf{X}$ is $m = \sum_{d=1}^{D} m_d$ where m_d is the number of variables in $\mathbf{X}_d$, then $\mathbf{X}$ is an $n \times m$ matrix. Using the SVD decomposition, $\mathbf{X}$ can be expressed in the form of

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T = \mathbf{T}\mathbf{V}^T, \tag{7.1}$$

where $\mathbf{T} = \mathbf{U}\Lambda$ are the principal components (associated with the observations) and the columns of $\mathbf{V}$ store the corresponding loadings associated to the principal components (associated with the variables).

Note that $\mathbf{T}$ denotes one matrix of common component scores the same across all D data sources representing a compromise score for all datasets. The matrix $\mathbf{V}$ can be partitioned in the same way as $\mathbf{X}$, representing the matrix of combined feature loadings.

Specifically, $\mathbf{V}$ can be expressed as a column block matrix as:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_d \\ \vdots \\ \mathbf{V}_D \end{bmatrix} = [\mathbf{V}_1^T | \ldots | \mathbf{V}_d^T | \ldots \mathbf{V}_D^T]^T,$$

where $\mathbf{V}_d$ is a $m_d \times r$ (r is the rank of $\mathbf{X}$) matrix storing the right singular vectors corresponding to the variables of the matrix $\mathbf{X}_d$. Following the dataset weighting of MFA, we define matrix $\tilde{\mathbf{X}}$,

$$\tilde{\mathbf{X}} = [\sqrt{\alpha_1}\mathbf{X}_1 | \ldots | \sqrt{\alpha_d}\mathbf{X}_d | \ldots | \sqrt{\alpha_D}\mathbf{X}_D],$$

where $\alpha_d = 1/\sigma_{d1}^2$. Hence, the SVD of $\tilde{\mathbf{X}}$ can be expressed as

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\Lambda}\tilde{\mathbf{V}}^T.$$

It can be shown (i.e. Abdi *et al.* (2013)) that the factor (dimension) scores for the observations can be obtained by

$$T = \tilde{\mathbf{U}}\tilde{\Lambda},$$

and the loadings for the dth dataset are obtained as

$$\mathbf{V}_d = \frac{1}{\sqrt{\alpha_d}}\tilde{\mathbf{V}}_d^T.$$

7.2.3 Identification of Important Features, Observations per Factor

MFA highlights the important features and observations that best describes each component. As in the standard PCA, the MFA factor scores, which is seen as composite scores from all datasets, can be used to identify observations with similar structure across all the datasets. As the components are obtained by combining the original variables, each variable 'contributes' in a different way to each component in the solution. The contribution of each variable can be measured by the loading of a variable on a component. The loading reflects the importance of that variable for this component. In MFA, the loadings are the correlations between the original variables and the factor scores. The higher is the loading the better this variable is 'explained' by the components.

Alternatively, we can also evaluate the importance of each observation (variable) to a specific component as the proportion of the explained variance of the component by the observation (variable). This statistics is called *contributions* (Abdi and Williams, 2010, Husson *et al.*, 2011).

For the observation, the contribution of the ith observation to the rth component, denoted by $ctr_{i,r}$ is given by the ratio of the squared weighted factor score $(t_{i,r})$ and the component's variance, its eigenvalue σ_r^2. For observations with equal weights of $1/n$, it is equal to,

$$ctr_{i,r} = \frac{\frac{1}{n} \cdot t_{i,r}^2}{\sigma_r^2}, \quad 0 \leq ctr_{i,r} \leq 1 \text{ and } \sum_{i=1}^{n} ctr_{i,r} = 1.$$

The larger the value of $ctr_{i,r}$, the higher the contribution of the ith observation to the rth component. Observations with high contributions and whose factor scores have different signs can then be contrasted to help interpret the component (Abdi *et al.*, 2013).

Similarly, we can find the important variables for a given component by computing variable contributions given by the ratio of its squared weighted loading $(v_{j,r}^2)$ for this component and its variance. The variance of the loadings for the variables is equal to one when the weights $1/\sigma_j$ are used for the jth variable, that is

$$1 = \sum_{j=1}^{m} \frac{1}{\sigma_j} \cdot v_{j,r}^2.$$

Note that variables from the same dataset have the same weight, $1/\sigma_j$. Thus, the variable contribution, denoted by $ctr_{j,r}$, is given by

$$ctr_{j,r} = \frac{1}{\sigma_j} \cdot v_{j,r}^2 \text{ where } 0 \leq ctr_{j,r} \leq 1 \text{ and } \sum_{j=1}^{m} ctr_{j,r} = 1.$$

The higher the value of $ctr_{j,r}$ the higher the contribution of the variable for the rth component. Variables with high contributions and whose loadings have different signs can then be contrasted to help interpreting the component.

The total contribution of the $\mathbf{X}_d$ to the rth component, denoted by $ctr_{d,r}$ is given by

$$ctr_{d,r} = \sum_{j=1}^{m_d} ctr_{j,r}.$$

Between Dataset Correlation

The Escoufier's R_V coefficient (Robert and Escofier, 1976), a non-centered squared coefficient of correlation between two matrices, can be used to evaluate the similarity between two datasets. Specifically, the R_V coefficient between datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ is computed as

$$R_V = \frac{tr(\mathbf{X}_1\mathbf{X}_1^T \times \mathbf{X}_2\mathbf{X}_2^T)}{\sqrt{tr(\mathbf{X}_1\mathbf{X}_1^T \times \mathbf{X}_1\mathbf{X}_1^T) \times tr(\mathbf{X}_2\mathbf{X}_2^T \times \mathbf{X}_2\mathbf{X}_2^T)}},$$

where $0 \leq R_V \leq 1$.

If the R_V coefficient is very low then it is not useful to study these datasets simultaneously since there are no common structures shared by the two datasets. We can also evaluate the canonical correlation coefficient between component r and dataset $\mathbf{X}_d$ as well. This would indicate whether the structure defined by component r is induced by the variables of dataset $\mathbf{X}_d$.

7.3 Application to the EGFR Project

The QSTAR data structure consists of three data sources, shown in Figure 7.1. Our aim is to find a subset of genes, bioactivity and chemical structures which share a common dimension, i.e. they share similar patterns across a subset of compounds.

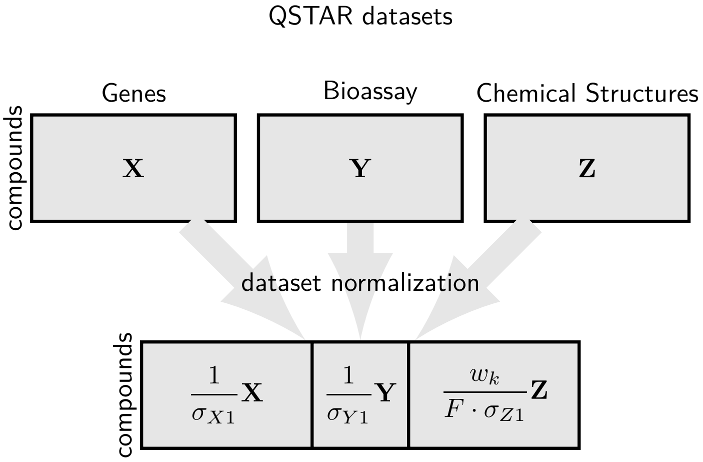

QSTAR datasets
Genes
Bioassay
Chemical Structures
compounds
X
Y
Z
dataset normalization
compounds
$\frac{1}{\sigma_{X1}}\mathbf{X}$
$\frac{1}{\sigma_{Y1}}\mathbf{Y}$
$\frac{w_k}{F \cdot \sigma_{Z1}}\mathbf{Z}$

Figure 7.1: QSTAR concatenated data for MFA.

Using the EGFR datasets, we compute the between dataset correlation matrix (shown in Table 7.1). R_V ranges between 0.34 to 0.56 indicating that a common structure is present between the three datasets.

The highest eigenvalue obtained from the separate PCA of $\mathbf{X}$ and $\mathbf{Y}$ and MCA for $\mathbf{Z}$ are equal to 1316.48, 2.898 and 0.15177, respectively. The variable loadings are presented in Figure 7.2. From these plots, we can identify the variables with high contribution to

Table 7.1: R_V coefficients for the 3 EGFR datasets.

Datasets	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{Z}$
Genes ($\mathbf{X}$)	1.00	0.56	0.36
Bioassays ($\mathbf{Y}$)	0.56	1.00	0.34
Chemical Structures ($\mathbf{Z}$)	0.36	0.34	1.00

the first 2 components. Notice that the fingerprint feature in Figure 7.2b marked in red is the fingerprint feature we highlighted in Chapter 2.

The compound scores for the first 2 factors are also shown in Figure 7.3a. Moreover, the contribution of each compound to the construction of a factor allows us to detect which compounds have the highest contribution to the factor. Figure 7.3b,the compounds scores and contributions for the first factor where the two reference compounds, gefitinib and erlotinib, are part of the group having a higher contribution.

The contribution of each dataset reflected by the proportion of the variance of a component that can be attributed to this dataset is given in Table 7.2. The larger the contribution of a dataset to a component, the more important this dataset is for this component. We can see that the three data sources have comparable contributions on the first factor. Alternatively, we can inspect the canonical correlation between the MFA factor and each dataset. We can see in Table 7.3 that the first MFA factor is highly correlated with the three datasets while the second factor is mostly correlated to the chemical structure dataset.

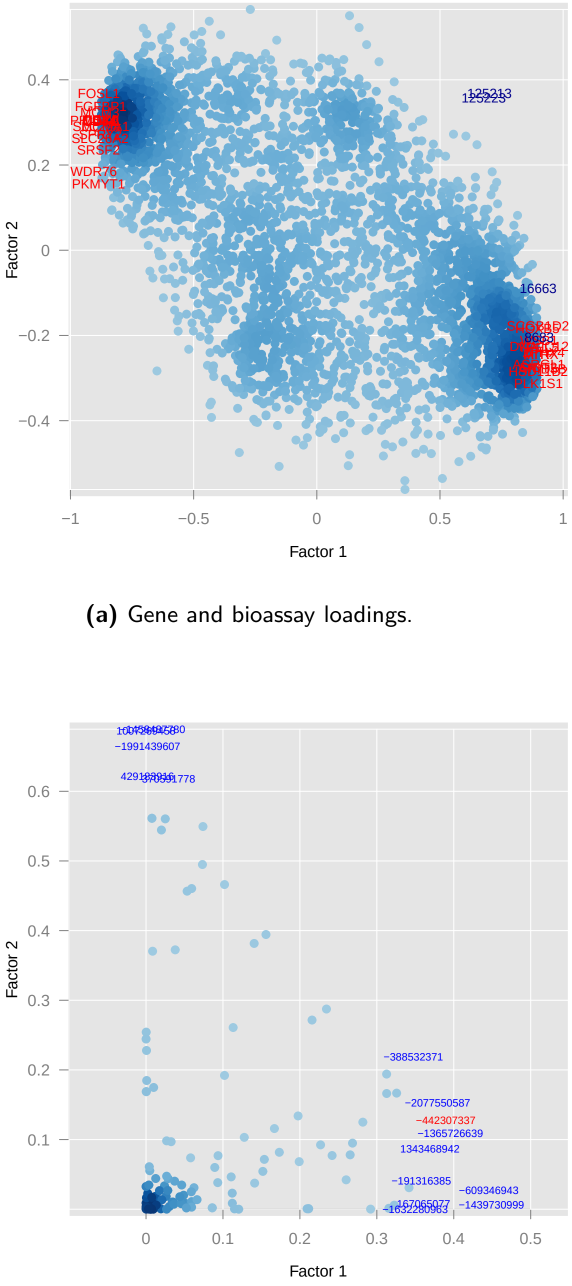
Table 7.2: Contributions of each dataset to the first 2 factors. Note that the sum of the contributions by factor is 1.

Datasets	Factor 1	Factor 2
Genes ($\mathbf{X}$)	35.28	16.40
Bioassays ($\mathbf{Y}$)	37.59	10.48
Chemical Structures ($\mathbf{Z}$)	27.13	73.11

Table 7.3: Canonical correlation of each dataset to the first 2 factors.

Datasets	Factor 1	Factor 2
Genes ($\mathbf{X}$)	0.92	0.51
Bioassays ($\mathbf{Y}$)	0.95	0.51
Chemical Structures ($\mathbf{Z}$)	0.86	0.95

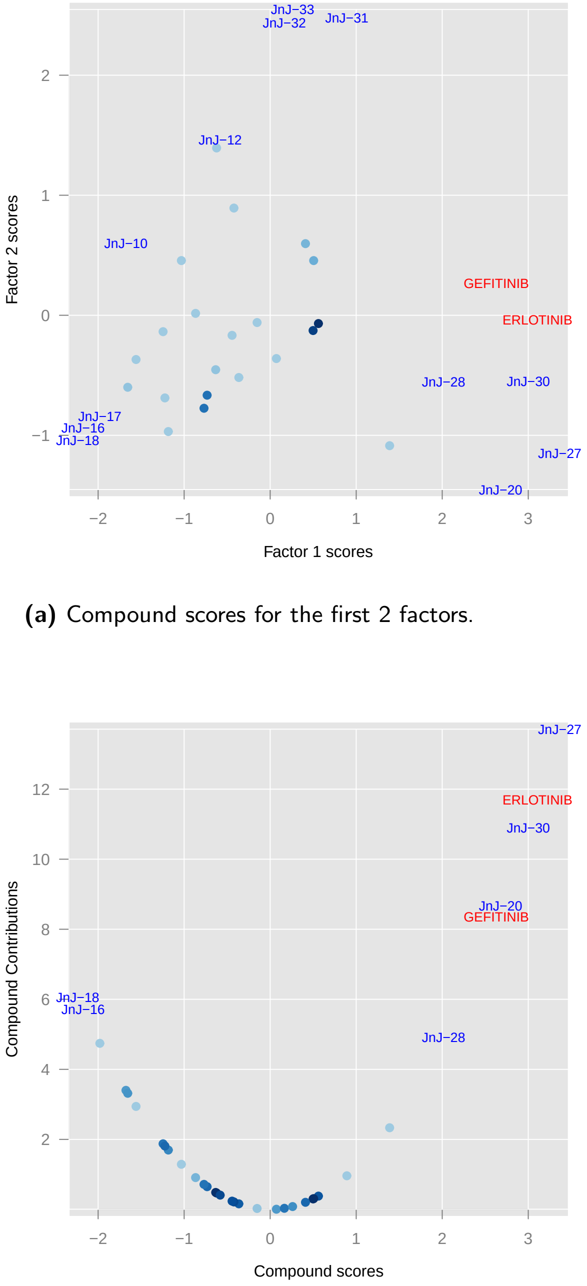
Once a subset of features, in each data source, with high contribution to the first factor were identified we inspect their profiles plots which is presented in Figure 7.4. Notice that the quantitative features are differentially expressed among the compounds



(a) Gene and bioassay loadings.

(b) Plot of Fingerprint feature loadings.

Figure 7.2: Variable loadings for the first two dimensions.



(a) Compound scores for the first 2 factors.

(b) Compound scores and contribution for the first factor.

Figure 7.3: Compound scores and contributions.

(first 7 compounds) that are close to the 2 reference compounds in Figure 7.4. Moreover, those compounds are characterized by the absence (in gray) of the fingerprint features that are highlighted in Figure 7.2b. This compound set contributing most to MFA factor 1 is the same compound cluster without JNJ-24 identified using the weighted similarity-based clustering presented in Chapter 4.

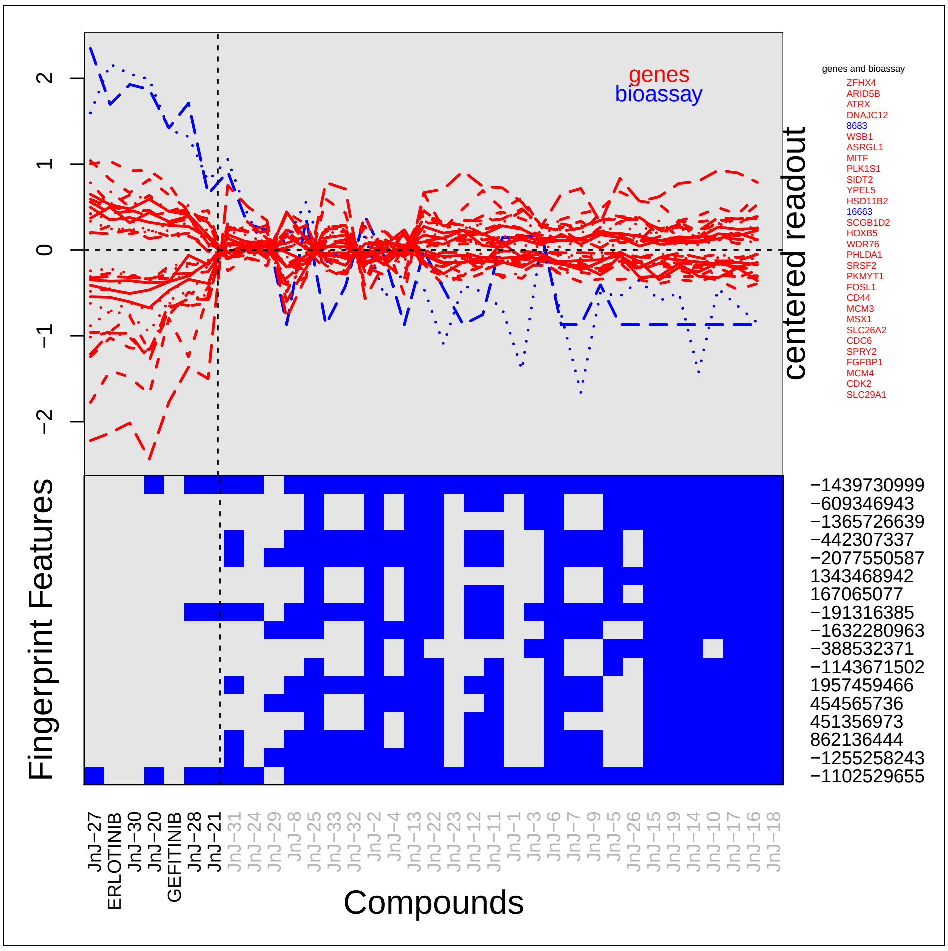


Figure 7.4: Profiles of features with the highest contribution to the first MFA factor. The plot highlights the profiles of the features that best describes the structure of the first factor. For the fingerprint features, blue indicates the presence and gray the absence of that feature in the compound.

7.4 Discussion

In this chapter, we apply the MFA method in order to integrate the three QSTAR datasets. This allows us to identify the main structures shared by these datasets. In this chapter a structure is defined as a subset of genes, bioactivity data and chemical structure which share similar patterns across a subset of compounds.

We have shown that, for the EGFR datasets, it is possible to identify, using the MFA method, a subset of features in each data source that have the same expression profile across a subset of compounds. Hence, we were able to identify the patterns in the chemical structures that initiate the biological pattern that was observed in the bioactivity data and gene expression data. MFA, however, is an exploratory multivariate technique. The inferential aspects particularly on the reliability and robustness of the results should be further investigated. Moreover, MFA uses a weighting scheme to achieve a fair integration of datasets. It gives less weight to bigger and/or redundant datasets. In extreme cases, this would mean giving more weight to noisy data which is undesirable. Also, in the presence of multiple datasets, it would be reasonable to give more weight to datasets that share more information. Hence, the normalization step in MFA could be modified or improved to account for other properties inherent in high-dimensional datasets.

In this chapter we focused on the first factor since all 3 datasets have comparable influence on this structure. Typically, this is the case in early drug discovery projects when the focus is placed on the primary assays, which are (in most cases) correlated, and the bioassay data reflects one dominant structure. The dominant common structure in the MFA solution is further exploited in the context of gene module enrichment. Moreover, for each dataset, only the top contributing features to the factor were selected to facilitate interpretation of the latent structure. This can be done by setting a threshold value on the univariate association between the feature and the latent factor scores. Another way is to simply select the top $k << m_d$ features, with k chosen arbitrarily. In Chapter 9, we will present an alternative method of component-wise feature selection approach.

In the next two chapters, we continue to focus on MFA. In Chapter 8 we show how MFA can be used in order to enrich a gene module. In this setting, we use only one data source and our aim is to find, in a given expression matrix, a subset of genes with similar expression profiles across a subset of compounds. In Chapter 9 we present the sparse MFA method that penalizes for the number of variables/samples related to a given factor and explore the connection between MFA and biclustering as presented in Figure 7.5 .

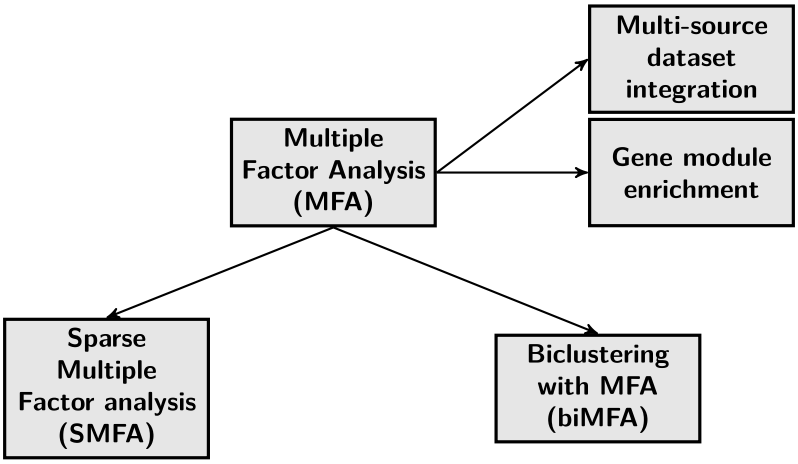

Multi-source dataset integration
Multiple Factor Analysis (MFA)
Gene module enrichment
Sparse Multiple Factor analysis (SMFA)
Biclustering with MFA (biMFA)

Figure 7.5: Modified MFA for feature selection and bicluster extraction. In SMFA, the sparsity is only introduced in the dimension of variables whereas in biMFA, the sparsity is imposed on both dimensions: samples and variables.

Chapter 8

Enrichment of Gene Expression Modules using Multiple Factor Analysis and Biclustering

8.1 Introduction

In Chapter 7, we discussed the use of MFA as an integrative method for the analysis of several data sources. In this chapter, we present the use of Multiple Factor Analysis (MFA) and biclustering methods for gene set enrichment when the subset of the lead genes is known a priori. Recall from Chapter 6 that local patterns were found by applying biclustering methods to the data matrix. A bicluster can be of interest in case it contains genes that are not only regulated under a subset of conditions but are also mostly functionally coherent. The summarized expression profiles of these genes that act in concert to carry out a specific function will then be presented as a *gene module* . Hence, a gene module is a subset of known genes for which a local pattern is observed across a subset of compounds. The aim of the analysis presented in this chapter is to find new genes which share the same local pattern as the genes belonging to the gene module. We term this process, a gene module enrichment.

8.2 Data Setting

Let $\mathbf{X}$ be an expression matrix of G genes (rows) and n compounds (columns). We assume that $\mathbf{X}$ can be partitioned into two submatrices sharing a common dimension. In the case where a set of lead genes are known, then we can extract it from $\mathbf{X}$ to form the submatrix $\mathbf{X}_1$ while the remaining genes form the other submatrix $\mathbf{X}_2$. The matrix $\mathbf{X}$ partitioned into two submatrices sharing the same columns is shown in (8.1).

$$\mathbf{X} = \left(\frac{\mathbf{X}_1}{\mathbf{X}_2} \right). \tag{8.1}$$

Alternatively, the submatrices could also share the same rows in common. For instance, a matrix $\mathbf{X}_1$, contains gene expression measurement for a set of compounds under a given condition. In addition, there is another set of compounds that have been profiled on the same set of genes for which the expression matrix is denoted by $\mathbf{X}_2$. In this case, the two matrices can be merged by rows to give one gene expression matrix $\mathbf{X}$,

$$\mathbf{X} = \left(\mathbf{X}_1 \,|\, \mathbf{X}_2 \right). \tag{8.2}$$

In general, for the data matrix given in (8.1) we are looking for a subset of rows in $\mathbf{X}_1$ and $\mathbf{X}_2$ sharing similar profiles across a subset of the columns. For the data matrix specified in (8.2), we aim to identify a subset of conditions from both submatrices having similar profiles across the rows. In this chapter, the focus is placed on the first setting with two subsets of gene expression data measured on the same set of observation. Note that we assume that local pattern(s) are observed for the genes belonging to $\mathbf{X}_1$ (i.e., the gene module).

8.3 Gene Module

Consider an experiment in which gene expression data of G genes is available for a set of n compounds. M of this G $(M << G)$ genes are a priori identified as a gene module of interest. Let $\mathbf{X}_M$ be the expression matrix of the gene module given by

$$\mathbf{X}_M = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{M1} & X_{M2} & \dots & X_{Mn} \end{pmatrix}.$$

Note that $\mathbf{X}_M$, containing information about M genes, is a submatrix of the expression matrix $\mathbf{X}$.

8.3.1 Examples of Gene Module

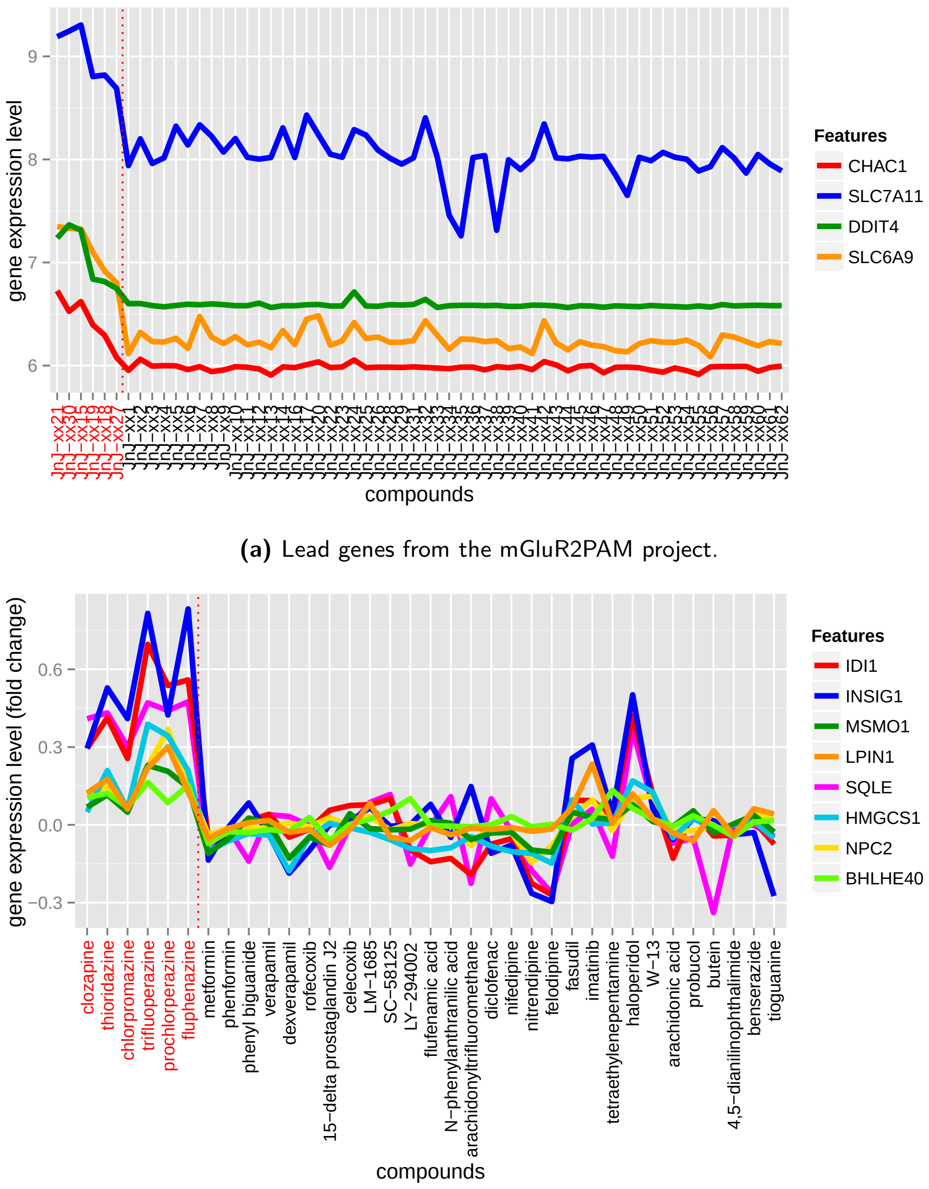
The first example of a gene module is related to the drug discovery project mGluR2PAM introduced in Chapter 1. The mGluR2PAM project consists of a set of $n = 62$ compounds described by the expression level of $G = 566$ genes. The research question is related to a gene module comprising of four $(M = 4)$ genes that are known to be biologically related and are linked to the phenotype of interest. Figure 8.1a displays the profiles of these 4 genes where similar expression pattern can be detected across a subset of 6 compounds.

The second example of gene module that we consider in this chapter is related to the CMap data. In this example, the gene module is a subset of genes identified to be differentially expressed on a certain condition. For example, the top 8 differentially expressed genes for cluster 1, presented in Chapter 5, of the CMap data whose profiles are presented in Figure 8.1b. Note that, in this example, $M = 8$ and $G = 2434$ for $n = 35$ compounds.

8.3.2 Gene Module Summarization

Genes that belong to the gene module are expected to be correlated since they share the same local pattern in the expression matrix. For example, in Figure 8.1 we can clearly observe the local patterns among the genes belonging to the two gene modules. Hence, we expect that a one factor solution in a factor analysis model will capture a substantial proportion of the total variability from these genes. The factor, is the true, but unobserved, gene module. Therefore, classical variable reduction methods such as factor analysis model or the principal component analysis or their variants can be used to estimate the latent factor that summarizes the information present in all the genes of this module.

In Figure 8.2, we present three different methods to summarize the gene module:



(a) Lead genes from the mGluR2PAM project.

(b) Lead Genes from the CMap data.

Figure 8.1: Profiles plot of genes in a module.

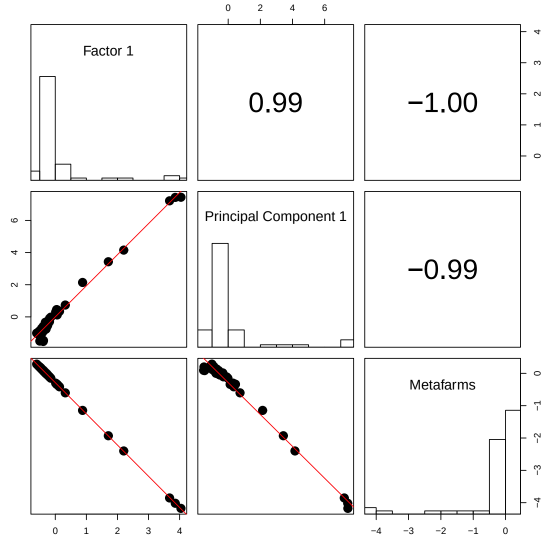


Figure 8.2: Gene module summarization using the first factor from Factor Analysis, PC1 from PCA and metafarms (mGluR2PAM data).

(1) factor analysis with one factor solution (Factor 1); (2) the first principal component in a principal component analysis (PC1) and (3) by using the summarization method metaFarms (Verbist *et al.*, 2015). Note that for the mGluR2PAM data, the summarization obtained by the three methods are highly correlated. Factor 1 explains about 88.7% of the variance while PC1 captures about 91.4% of the total variance in the data.

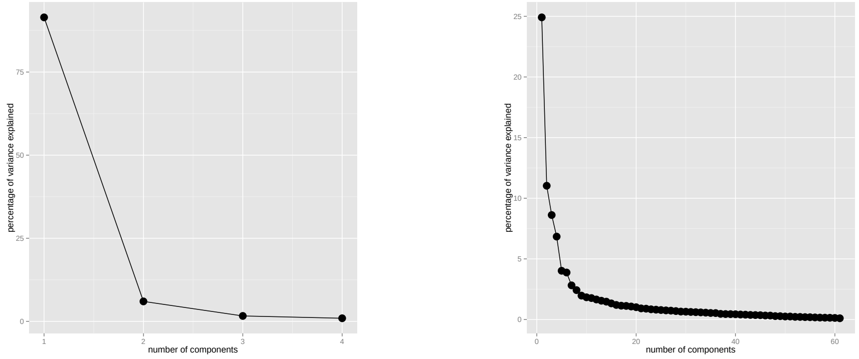
8.3.3 Enrichment of Gene Module

Let $\mathbf{X}_{\bar{M}}$ be a $(G-M) \times n$ submatrix of $\mathbf{X}$ without the module of interest. Note that $\mathbf{X}_M$ and $\mathbf{X}_{\bar{M}}$ have the same set of compounds (the same column dimension), i.e.,

$$\mathbf{X} = \left(\frac{\mathbf{X}_M}{\mathbf{X}_{\bar{M}}} \right),$$

In contrast to $\mathbf{X}_M$ that contains highly correlated genes and can be summarized using the first factor/component, $\mathbf{X}_{\bar{M}}$ needs at least 15 factors to retain 80% of the variability (Figure 8.3), an indication that there is no dominant structure present in the second set.

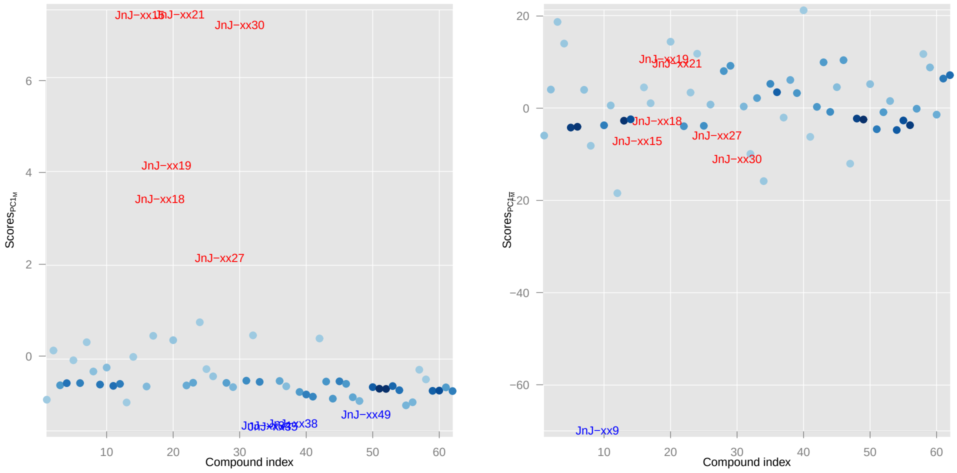
Plotting the scores of the first principal components of the two submatrices in Figure 8.4 a shows that the first PC is dominated by the 6 compounds (marked in red) while only one compound contributed to the first PC in $\mathbf{X}_{\bar{M}}$ (Figure 8.4b). Given that we have one



(a) Proportion of variance explained for $\mathbf{X}_M$. **(b)** Proportion of variance explained for $\mathbf{X}_{\bar{M}}$.

Figure 8.3: Scree plots of the 2 submatrices.

known structure in $\mathbf{X}_M$, the first PC, we wish to find if that structure is also present in $\mathbf{X}_{\bar{M}}$, i.e., our aim is to identify a subset of genes in $\mathbf{X}_{\bar{M}}$ having similar expression profiles to those genes in $\mathbf{X}_M$. The result presented in Figure 8.4 is expected since we do not assume that there is only one dominant pattern in $\mathbf{X}_{\bar{M}}$. In the next section we illustrate how the multiple factor analysis can be used in order to identify the subset of genes in $\mathbf{X}_{\bar{M}}$ that share the same local pattern as the genes belonging to the gene module in $\mathbf{X}_M$.



(a) First factor for $\mathbf{X}_M$. **(b)** First factor for $\mathbf{X}_{\bar{M}}$.

Figure 8.4: PC1 compound scores for each submatrix.

8.4 MFA for Gene Module Enrichment

As introduced in the previous chapter, Multiple Factor Analysis (MFA, Escofier and Pages, 1994) allows to analyse several sets of quantitative variables which were measured on the same units, simultaneously. The units are the common dimension in the submatrices. The aim of the analysis is to study the relationship between the observations, the variables, and the datasets. In particular, we want to find links between datasets (presence of common structure) and to quantify the contributions of each dataset to the common structure. The basis of MFA can be viewed as a weighted PCA applied to multiple datasets. It begins with a normalization of each dataset to ensure that no particular dataset can dominate the common structure. In the second step, a PCA of concatenated normalized datasets is performed.

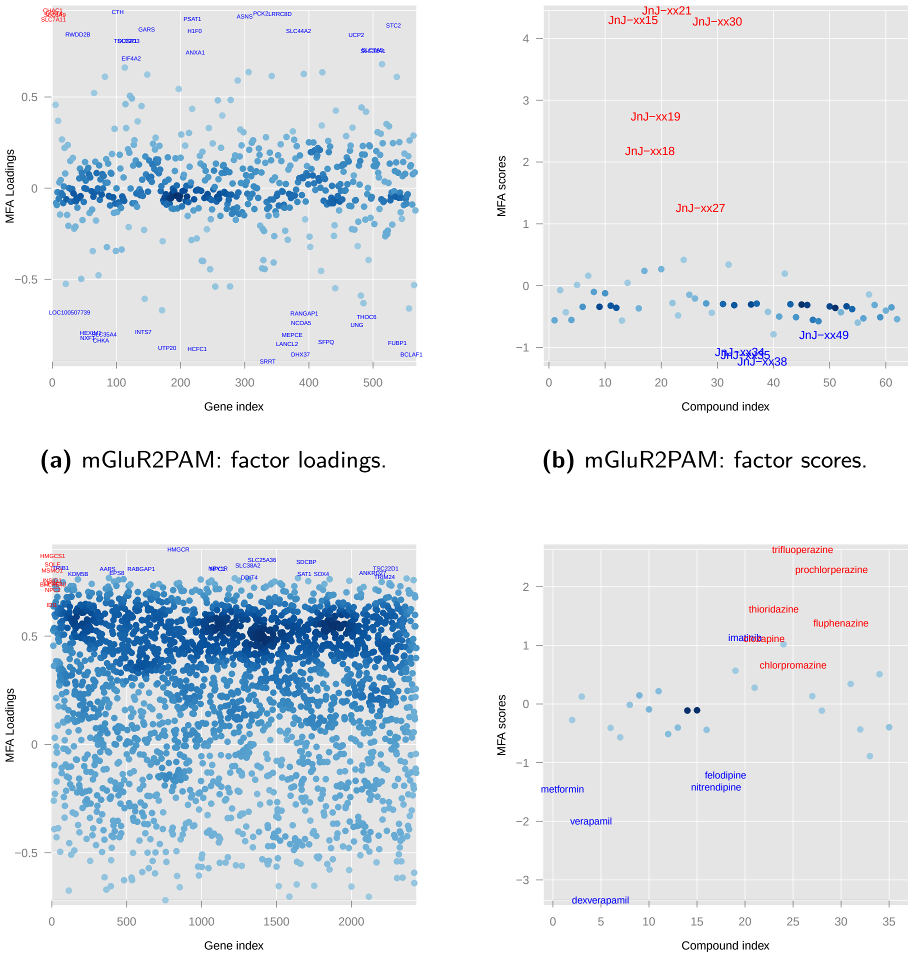
In this particular application of MFA on gene module enrichment, two submatrices sharing the same samples are involved. Each submatrix is normalized and combined to form matrix $\mathbf{X}$. Here we define the combined matrix $\mathbf{X}$ with compounds (the common dimension) in the rows and genes in the columns. Hence, the concatenated gene expression matrix is written as

$$\mathbf{X} = \left(\frac{1}{\sigma_{1_M}} \mathbf{X}_M^T \,\middle|\, \frac{1}{\sigma_{1_{\bar{M}}}} \mathbf{X}_{\bar{M}}^T \right). \tag{8.3}$$

For the mGluR2PAM data, the largest eigenvalues for $\mathbf{X}_M$ and $\mathbf{X}_{\bar{M}}$ are respectively 3.6579 and 140.0129 indicating a need to normalize these two matrices prior to combining them. We use 0.5228 as weighting factor for $\mathbf{X}_M$ and 0.0845 for $\mathbf{X}_{\bar{M}}$. The same holds for the CMap data where the largest eigenvalues for $\mathbf{X}_M$ and $\mathbf{X}_{\bar{M}}$ are respectively 6.4806 and 812.9709.

We apply MFA to the mGluR2PAM gene expression dataset. The results are presented in Figures 8.5a and 8.5b and Table 8.1. The gene loadings are presented in Figure 8.5a. Genes with relatively high and low loadings are related to the first factor. Note that the four genes comprising the gene module are marked in red. Figure 8.5b shows the compounds scores. In addition to the six lead compounds marked in red, we can observe three compounds with relatively low scores in relation to the first factor. As we can see in Table 8.1, for the mGluR2PAM data, about 24% of the variance of the first factor can be attributed to the second gene set and the majority of the contributions, as expected, come from the gene module. The second factor can be totally attributed to the second gene set which should be the case since the first factor already accounts for the one structure present in the gene module.

The expression profiles of genes that are highly correlated with the first MFA factor are shown in Figure 8.6. The enriched gene module consists of 42 genes instead of 4 genes that originally belong to the gene module. In addition to the 6 compounds (marked in



(a) mGluR2PAM: factor loadings.

(b) mGluR2PAM: factor scores.

(c) CMap: factor loadings.

(d) CMap: factor scores.

Figure 8.5: Genes and compounds that contribute to the first factor of MFA.

Table 8.1: Dataset contribution to each factor.

Datasets	Factor 1	Factor 2
$\mathbf{X}_M$	76.48	0.68
$\mathbf{X}_{\bar{M}}$	23.52	99.32

red) that characterize the gene module, the enriched set identifies also 3 extra compounds (marked in blue) that are also active on most of the member genes of the enriched gene module. The first principal component can be used to summarize these genes into one set of compromised scores, and in this example, it can explain 72% of the variability in the dataset (Figure 8.7).

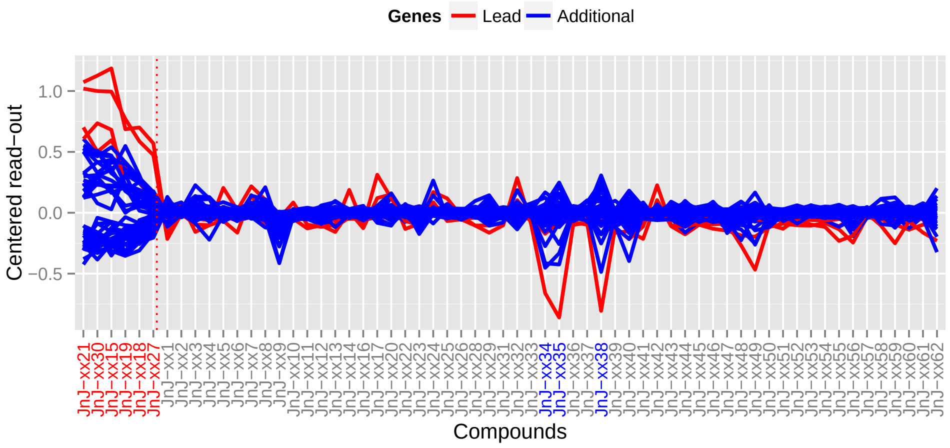


Figure 8.6: mGluR2PAM: Profiles plot of enriched gene module.

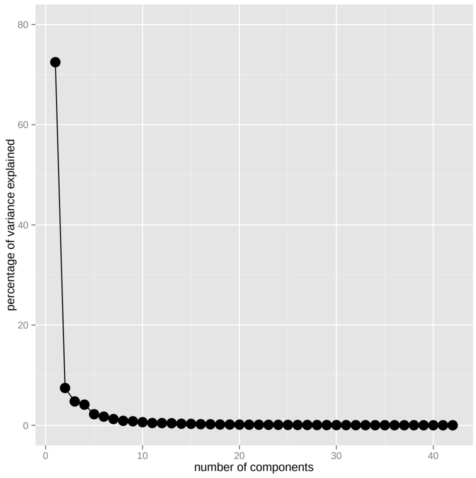


Figure 8.7: mGluR2PAM: Percentage of variance explained by component for the enriched gene module.

8.5 Biclustering and Multiple Factor Analysis to find Gene Modules

The results of the MFA presented Figure 8.6 reveal a subset of genes (the enriched gene module) that are regulated by a subset of compounds. Using the terminology of this book, these subsets of genes and the compounds form a bicluster. In this section we explore the similarity between MFA and FABIA. We focus on factor loadings and factor scores and investigate how the local patterns in $\mathbf{X}$ can be identified using MFA or FABIA.

Figure 8.8 shows the factor loading and factor scores obtained for MFA and FABIA. We notice that both factor scores (of the compounds) and factor loadings (of the genes) are highly correlated indicating that both MFA and FABIA identified the same enriched gene module. Consequently, the expression profiles presented in Figure 8.8c are very similar to the profiles presented in Figure 8.6 (for the MFA solution).

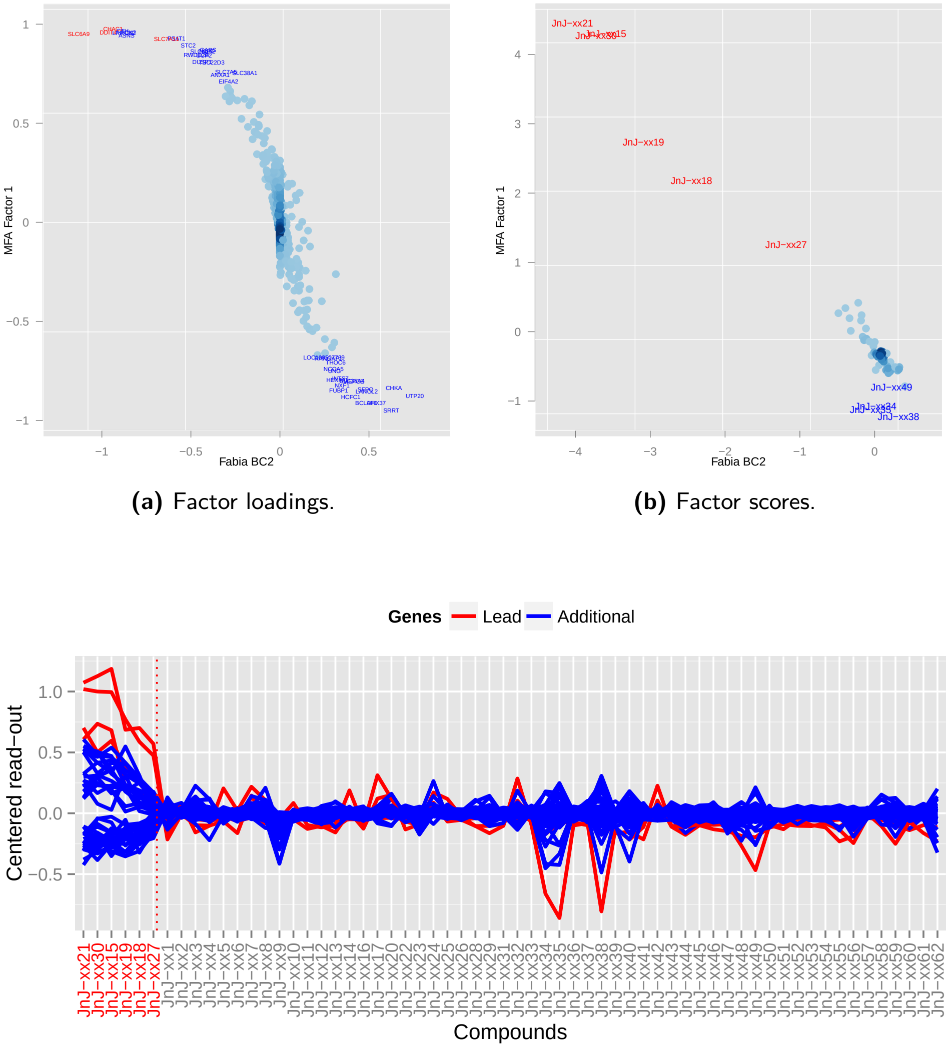
Similar patterns were obtained for the CMap dataset (shown in Figure 8.9), but in addition, it clearly highlights the effect of the sparsity factor imposed on the Fabia loadings which is not available under MFA (Figure 8.9a). While FABIA searches only for correlated profiles across a subset of samples, MFA uses the similarity of gene profiles across all compounds. As a result, some genes discovered by MFA are not part of the FABIA solution and vice versa. This is the case since the size of the bicluster depends on the sparseness parameters specified for the FABIA algorithm. The enriched CMap gene module is the first bicluster discovered by FABIA consisting of 12 genes and 8 of which are the genes from the gene module (Figure 8.9c).

For the mGluR2PAM project, the second bicluster obtained by FABIA that contains the 4 lead genes is composed of 42 genes which were also identified by the MFA (as the top 42 most correlated genes with the first factor). Interestingly, although the two sets of genes identified by the two methods are not identical, the estimated latent structure (i.e. summarized gene module) underlying them is almost identical as shown in Figure 8.10a. On the other hand, for the CMap data, the gene modules from FABIA and MFA, each consisting of 12 genes, is also correlated (Figure 8.10b).

8.6 Discussion

The application of Multiple Factor Analysis and biclustering technique in the context of gene expression module enrichment is summarized in Figue 8.11.

MFA is a descriptive multivariate technique for integrating multi-source datasets or multiple datasets from the same source. This approach helps us to find common structures within and between the different datasets. It is typically used to discover common

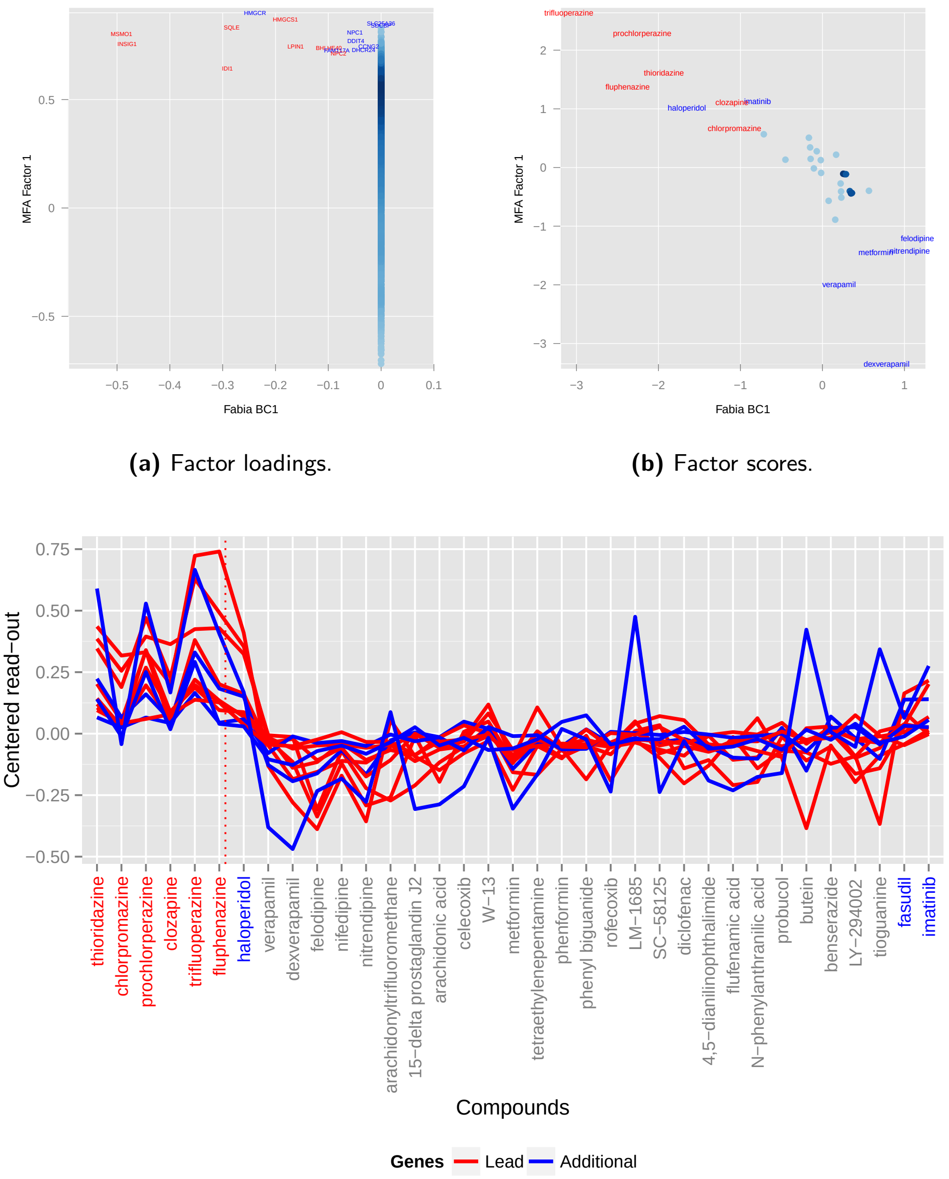


(a) Factor loadings.

(b) Factor scores.

(c) Profiles plot of genes in FABIA bicluster 2.

Figure 8.8: The mGluR2PAM Dataset: FABIA bicluster 2 versus MFA factor 1.

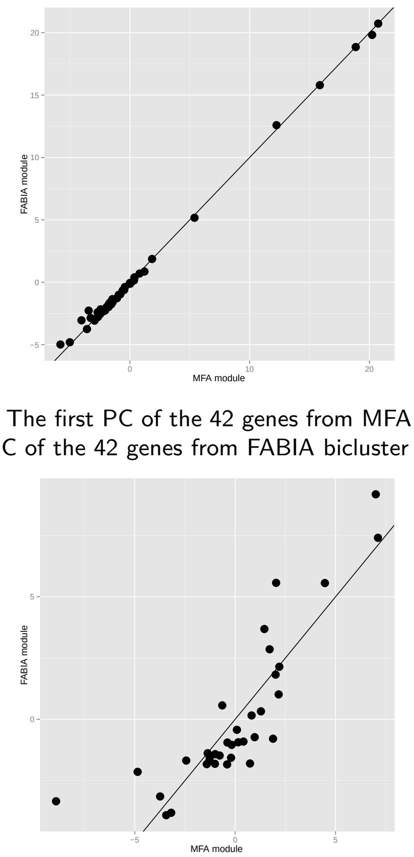


(a) Factor loadings.

(b) Factor scores.

(c) Profiles plot of genes in the bicluster.

Figure 8.9: The CMap Dataset: FABIA bicluster 1 versus MFA factor 1.



(a) mGluR2PAM: The first PC of the 42 genes from MFA factor 1 is plotted against the first PC of the 42 genes from FABIA bicluster 2.

(b) CMap: The first PC of the 12 genes from MFA factor 1 is plotted against the first PC of the 12 genes from FABIA bicluster 1.

Figure 8.10: Summarized gene module obtained for MFA and FABIA.

structures shared by several sets of variables describing the same set of observations. In this chapter, we have multiple datasets from the same source, the gene expression data. The dataset is split into two groups: (1) lead genes (termed as “gene module") and (2) the other genes in the dataset. Here, we propose to use the MFA as a gene module enrichment technique wherein additional genes from the second group that are co-regulated with the lead genes are discovered. Using MFA, the first factor will always be described by the two groups of genes. Without the splitting and normalization step, the structure of the lead genes may not be identified as the main factor of variability in the data due to the noisy nature of microarray datasets.

Biclustering analysis of gene expression data using FABIA produces several biclusters. A bicluster containing the gene module can be identified as an enriched gene module.

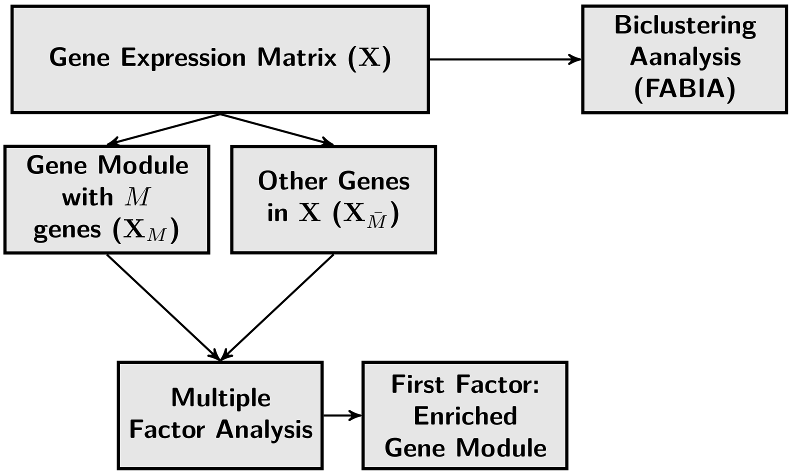

Gene Expression Matrix ($\mathbf{X}$)
Biclustering Aanalysis (FABIA)
Gene Module with M genes ($\mathbf{X}_M$)
Other Genes in $\mathbf{X}$ ($\mathbf{X}_{\bar{M}}$)
Multiple Factor Analysis
First Factor: Enriched Gene Module

Figure 8.11: Gene module enrichment technique using MFA and biclustering.

However, in contrast to MFA that does not depend on any tuning parameters, the biclustering results are not stable and may depend on the input parameters. Although in MFA, the interest lies only on the first factor, the other structures present in the dataset characterizes the remaining factors. Here, MFA can be viewed as a guided biclustering method as supported by the similarity of the results of MFA and FABIA for gene module enrichment purpose.

In FABIA, we take a bicluster as the enriched gene module where the number of member genes per bicluster depends on the specified tuning parameters. In MFA, as mentioned in Chapter 7, we can highlight important features, i.e. genes with high loadings or contributions to factor 1 according to some cut-off value for the loadings. In the next chapter, we propose to modify the MFA method by incorporating component-wise feature selection which facilitates bicluster extraction for multiple data sources.

Chapter 9

Unsupervised Integrative Methods for High Content Screening and Gene Expression Data Analysis: Sparse MFA and biMFA

9.1 Introduction

In this chapter, we focus on two extensions of the methodology presented in Chapters 7 and 8. The first extension is related to a new data source, high-content imaging data (HCS) and the second is an extension of the MFA methodology which includes a penalty on the number of variables and observations related to a given factor in the solution.

The underlying idea is that the inclusion of high-content imaging data in the data analysis pipeline and the integration of this data source with gene expression data is expected to improve our understanding of the on- and off-target effects of compounds already identified as potential leads by traditional high-throughput screening (HTS) in early drug discovery. During lead optimization, a very narrow chemical space is being considered and transcriptional profiling can help elucidate the mechanism of action of these compounds. Verbist *et al.* (2015) highlighted the utility of gene expression profiling during the lead optimization phase, particularly on the detection of off-target effects. On the other hand, HCS provides information-rich data sets containing readouts on multiple cellular parameters. Image-based screen cellular phenotypes can suggest both selectivity and toxicity of lead compounds.

Independent analysis of these two data sources can already reveal valuable information about the compounds' mechanism of action. For instance, extracting subsets of functionally coherent genes from the microarray data can suggest possible biological pathways of lead compounds. Similarly, identifying a number of HCS parameters indicating cellular phenotypic changes upon compound treatment relative to the control can direct us to identify phenotypic subclasses without the need to inspect every compound image. Connecting gene expression to the phenotypic changes in HCS can serve as a "biology" screening tool of lead compounds. In fact, HCS and gene expression data complement each other to provide a more mechanistic insight into compounds' biological activity, most especially genotoxicity. The image-based phenotypic changes may be rationalized by identifying the affected genes. For example, compounds exhibiting phenotypic features indicating toxic events can be explained by compound perturbations on toxicity-related genes. Likewise the biological relevance of gene expression can be confirmed by HCS images. Hence, an integrated analysis of these two datasets can provide a deeper understanding of compound effects both on transcriptional and phenotypic levels.

The Multiple Factor Analysis discussed in Chapter 7 can be applied to integrate these two datasets. Recall that MFA can be used to find co-regulated profiles between or among the coupled matrices, that is multiple datasets defined on the same set of samples. As mentioned in Chapter 7, the analysis consists of two steps: (1) dataset normalization via weighting based on the inverse of the first singular value; and (2) Principal Component Analysis (PCA) analysis (implemented as SVD) on the combined normalized datasets.

One difficulty related to the MFA presented in Chapters 7 and 8 is the interpretation of a given factor or a selection procedure of variables and observations related to a given factor. In this chapter, we address this problem by including, per factor, penalties for variables, observations or both. Using these penalties, some of the loadings and scores will be shrunk to zero and will not be related to the factor. This motivated the use of sparse Principal Component Analysis (SPCA) (Zou *et al.*, 2006, Witten *et al.*, 2009) which penalizes the estimation of loadings by integrating sparsity inducing penalty terms in the regression criterion to shrink many of the loadings to zero (since most genes do not contribute much or are just noise). Moreover, in early drug discovery microarray experiments, biclustering techniques have been successfully applied to find local patterns in the data, i.e. genes that are coregulated under a subset of conditions. The major drawback is its reliance on random starting seeds that could produce inconsistent results

(Sill *et al.*, 2011). SPCA, however, is not suitable for biclustering since sparsity is induced only in the estimation of gene loadings and not on the compound scores. A recent SVD-based methods to detect block structures is the so called sparse SVD (SSVD), proposed by Lee *et al.* (2010) and improved by Sill *et al.* (2011) using a robust version of the SSVD. The SSVD method applies sparsity on both the left and right singular vectors, i.e., on both observations and variables.

In this chapter, we continue to focus on integrated analysis and introduce two sparse methods. The sparse Multiple Factor Analysis (SMFA) is presented in Section 9.3 and a new biclustering algorithm for multi-source data, based on SSVD (biMFA, biclustering using Multiple Factor Analysis) is presented in Section 9.4. In Section 9.2 we present the data structure that will be considered in this chapter. The proposed methods are applied to the PDE10 data and presented in Section 9.5.

9.2 Data Structure

The analysis presented in this chapter consists of two data matrices, a $m \times n$ gene expression matrix, $\mathbf{X}$ containing m genes and n compounds, and a $s \times n$ HCS data matrix, $\mathbf{H}$, given by

$$H_{s\times n} = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ h_{s1} & h_{s2} & \dots & h_{sn} \end{pmatrix}.$$

Note that the h_{ji} is the measurement of the jth HCS feature for the ith compound. We express the combined matrix with variables in the columns and their common dimension, the samples, in the rows as a $n \times p$ matrix $\mathbf{Q} = \left[\mathbf{X}^T | \mathbf{H}^T\right]$ where $p = m + s$ is the total number of features.

9.3 Sparse Multiple Factor Analysis (SMFA) for Simultaneous Feature Selection and Data Integration

MFA, as presented in the previous chapters, is simply a weighted PCA. This is very useful especially if the PCs can be readily interpreted. However, this is not the case when simultaneously dealing with gene expression and high content screening data, with the derived PCs being a linear function of a large number of variables. Hence, as pointed out earlier, the attractive way from the interpretability point of view, is to impose sparsity on either the variable loadings, component scores or both. This can be done by penalizing the right singular vector ($\mathbf{V}$) to get sparse variable loadings or the left singular vectors ($\mathbf{U}$) to get sparse component scores, or both vectors to reveal biclusters within a least-squares approach.

Sparse Multiple Factor Analysis is similar to the MFA algorithm but it includes a Sparse PCA (SPCA) in the simultaneous analysis instead of PCA. As pointed out by Zou *et al.* (2006), it is desirable not only to achieve the dimensionality reduction but also to reduce the number of used variables to facilitate ease of interpretation. One direct way to achieve this is by setting some variable loadings in PCA with absolute value smaller than an (arbitrary) threshold to zero. Cadima and Jolliffe (1995) shows that this technique, however, can be misleading and can misidentify important variables.

Shen and Huang (2008) presented an SPCA approach using the low rank approximation of SVD with penalized loadings. Let $\mathbf{X}$ be any data matrix of size $n \times m$ with elements x_{ij} where $i = 1, ..., n$ samples and $j = 1, ..., m$ variables. Then the SVD of $\mathbf{X}$, is given by

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T \quad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I},$$

where r is the rank of $\mathbf{X}$ and Λ is a diagonal matrix with a rank-ordered set of positive singular values, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$, as elements. The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. Then, the columns of an $n \times r$ matrix $\mathbf{T} = \mathbf{U}\Lambda$ are the PCs, and the columns of an $m \times r$ matrix $\mathbf{V}$ are the corresponding loadings.

Note that, for the analysis of multiple data sources, the data matrix $\mathbf{X}$ is the normalized combined matrix $\mathbf{Q}$ defined in the previous section. The closets rank-l approximation to the data matrix $\mathbf{X}$ is given by

$$\mathbf{X}^{(l)} \approx \mathbf{U}\Lambda\mathbf{V}^T = \sum_{i=1}^{l} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Note that any $n \times m$ rank-one matrix can be written as $\tilde{\mathbf{u}}\tilde{\mathbf{v}}^T$, where $\tilde{\mathbf{u}}$ is a n-vector and $\tilde{v}$ is a m-vector. According to Eckart and Young (1936), the first SVD-layer gives us the best rank-one approximation of $\mathbf{X}$ with respect to the squared Frobenius norm, i.e.

$$(\sigma_1, \mathbf{u}_1, \mathbf{v}_1) = \text{argmin}_{\tilde{\mathbf{u}}\tilde{\mathbf{v}}} \|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^T\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \tilde{u}_i\tilde{v}_j)^2,$$

where $\|\cdot\|_F^2$ indicates the squared Frobenius norm, which is the sum of squared elements of the matrix. The vector $\mathbf{u}_1$ and $\mathbf{v}_1$ are the right and left singular vectors of the first layer, respectively and σ_1 is the first singular value.

Shen and Huang (2008) presented the sparse PCA via regularized SVD which imposes regularization penalties that promote shrinkage and sparsity on the $\tilde{\mathbf{v}}$ thereby achieving sparse loadings. Specifically, their method searches for an n-vector $\tilde{\mathbf{u}}$ subject to $\|\tilde{\mathbf{u}}\| = 1$ and a m-vector $\tilde{\mathbf{v}}$ that minimize the following penalized sum-of-squares criterion,

$$\|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^T\|_F^2 + P_\lambda(\tilde{\mathbf{v}}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \tilde{u}_i\tilde{v}_j)^2 + \sum_{j=1}^{m} p_\lambda(|\tilde{v}_j|). \tag{9.1}$$

The second term $P_\lambda(\tilde{\mathbf{v}})$ is a sparsity penalty on the vector of loadings, $\tilde{\mathbf{v}}$, with $\lambda \geq 0$ as a tuning parameter.

The iterative algorithm to minimize (9.1) proceeds as follows to estimate the best rank-one approximation, $(\mathbf{u}_1, \mathbf{v}_1)$, of $\mathbf{X}$.

1. Initialize: Apply the standard SVD to $\mathbf{X}$ and obtain the best rank-one approximation of $\mathbf{X}$ as $\sigma_1\mathbf{u}^*\mathbf{v}^{*T}$ where $\mathbf{u}^*$ and $\mathbf{v}^*$ are unit vectors. Set $\tilde{\mathbf{v}}_{old} = \sigma_1\mathbf{v}^*$ and $\tilde{\mathbf{u}}_{old} = \mathbf{u}^*$.

2. Update:

   (a) For a fixed $\tilde{\mathbf{u}}$, subject to $\sum\limits_{i=1}^{n} \tilde{u}_i = 1$, it is shown that by expanding the squares in (9.1), the optimal $\tilde{\mathbf{v}}$ minimizes $\tilde{v}_j^2 - 2(\mathbf{X}^T\tilde{\mathbf{u}})_j\tilde{v}_j + p_\lambda(|\tilde{v}_j|)$ (Shen and Huang, 2008). It follows that the optimal $\tilde{v}_j$ can be obtained by applying a thresholding function, h_λ, to the vector $\mathbf{X}^T\tilde{\mathbf{u}}_{old}$ which depends on the form of the penalty function $p_\lambda(\tilde{v}_j)$.

   Applying the soft-thresholding rule with penalty function $p_\lambda(|\tilde{v}_j|) = 2\lambda|\tilde{v}_j|$ (Tibshirani, 1996), hard thresholding rule (Donoho,1994) with penalty function $p_\lambda(|\tilde{v}_j|) = \lambda^2 I(|\tilde{v}_j| \neq 0)$ and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), we respectively get

   $$\begin{aligned}
   \tilde{v}_{j,new}^{soft} &= \text{sign}(\mathbf{X}^T\tilde{\mathbf{u}}_{old})(|\mathbf{X}^T\tilde{\mathbf{u}}_{old}| - \lambda)_+,\\
   \tilde{v}_{j,new}^{hard} &= I(|\mathbf{X}^T\tilde{\mathbf{u}}_{old}| > \lambda)(\mathbf{X}^T\tilde{\mathbf{u}}_{old}),\\
   \tilde{v}_{j,new}^{SCAD} &= \begin{cases} \text{sign}(\mathbf{X}^T\tilde{\mathbf{u}}_{old})(\mathbf{X}^T\tilde{\mathbf{u}}_{old} - \lambda)_+ & \text{if } |\mathbf{X}^T\tilde{\mathbf{u}}_{old}| \leq 2\lambda,\\ (a-1)\mathbf{X}^T\tilde{\mathbf{u}}_{old} - \text{sign}(\mathbf{X}^T\tilde{\mathbf{u}}_{old})a\lambda/(a-2) & \text{if } 2\lambda < |\mathbf{X}^T\tilde{\mathbf{u}}_{old}| \leq a\lambda,\\ \mathbf{X}^T\tilde{\mathbf{u}}_{old} & \text{if } |\mathbf{X}^T\tilde{\mathbf{u}}_{old}| > a\lambda. \end{cases}
   \end{aligned}$$

   (b) After fixing $\tilde{\mathbf{v}}$, we can get the minimizer of (9.1) given by $\tilde{\mathbf{u}}_{new} = \mathbf{X}^T\tilde{\mathbf{v}}_{new}/\|\mathbf{X}^T\tilde{\mathbf{v}}_{new}\|$.

3. Repeat Step 2 replacing $\tilde{\mathbf{u}}_{old}$ and $\tilde{\mathbf{v}}_{old}$ by $\tilde{\mathbf{u}}_{new}$ and $\tilde{\mathbf{v}}_{new}$ until convergence.

4. Standardize the final $\tilde{\mathbf{v}}_{new}$ as $\mathbf{v}_1 = \tilde{\mathbf{v}}_{new}/\|\tilde{\mathbf{v}}_{new}\|$, and $\mathbf{u}_1 = \tilde{\mathbf{u}}_{new}$.

We can apply the same procedure to the residual matrix, $\mathbf{X} - \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$, to estimate the next SMFA factor. Note that for $\lambda = 0$ we get $\tilde{\mathbf{v}}_{new} = \mathbf{X}^T \tilde{\mathbf{u}}_{old}$. Then the above algorithm simplifies to the calculation of SVD using the alternating least squares (Gabriel and Zamir, 1979).

Penalty Parameter Selection

The estimation of sparse vector loadings depends on the penalty function with λ as the tuning parameter since a is usually fixed at $a = 3.7$ for the SCAD thresholding rule (Fan and Li, 2001). The choice of λ is still an open question. Here, we investigated, three ways of tuning the parameter λ: (1) the use of the cumulative percentage of explained variance proposed by Shen and Huang (2008), (2) the k-fold cross validation technique (2008) and (3) the use of Bayesian Information Criterion (Schwarz, 1978) (Lee *et al.*, 2010).

These algorithms makes use of the degree of sparsity ($df(\lambda)$) as the tuning parameter instead of λ. The degree of sparsity is defined as the number of variables with non-zero loadings in $\tilde{v}_{j,new}$. That is, setting the degree of sparsity to be $j,\ j \in (0, m-1)$ is the same as setting $|\mathbf{X}^T \tilde{\mathbf{u}}_{old}|_{(j)} \leq \lambda \leq |\mathbf{X}^T \tilde{\mathbf{u}}_{old}|_{(j+1)}$ where $|\mathbf{X}^T \tilde{\mathbf{u}}_{old}|_{(j)}$ is the jth order statistic of $|\mathbf{X}^T \tilde{\mathbf{u}}_{old}|$.

(1) The Variance Explained by PC

Let $\mathbf{V}_r = [\mathbf{v}_1, ..., \mathbf{v}_r]$ be a $n \times r$ matrix of the first r sparse loading vectors and $\mathbf{T}_r = \mathbf{X}\mathbf{V}_r$ denote the matrix of the first r principal components. When the PCs, $\mathbf{T}$, are uncorrelated and their loadings are orthogonal then $\mathbf{X}_r = \mathbf{T}_r \mathbf{V}_r^T$ and the total explained variance is simply $tr(\mathbf{T}_r^T \mathbf{T}_r)$. However, in sparse PCA, these properties are lost (Zou *et al.*, 2006). Shen and Huang (2008) proposed that to deal with the correlation among PCs, in calculating the added variance explained by an additional PC, the variance attributable to the previous PCs should be adjusted for and the *adjusted* variance of the rth PC, denoted by σ_r^2, is given by

$$\sigma_r^2 = tr(\mathbf{X}_r^T \mathbf{X}_r) - tr(\mathbf{X}_{r-1}^T \mathbf{X}_{r-1}),$$

where $\mathbf{X}_r = \mathbf{X}\mathbf{V}_r(\mathbf{V}_r^T \mathbf{V}_r)^{-1} \mathbf{V}_r^T$.

The cumulative percentage of variance explained (CPEV) by the first r PCs is $tr(\mathbf{X}_r^T \mathbf{X}_r)/tr(\mathbf{X}^T \mathbf{X})$. Therefore, for each component, the corresponding λ or $df(\lambda) = j$ is chosen sequentially such that the CPEV is still close to the CPEV when $\lambda = 0$ and $j = 0$. For example, a drop of at most 10%. This procedure, therefore, involves personal judgement on deciding the value of λ.

(2) K-fold Cross-validation (CV)

Another way of assessing the degree of sparsity is by performing a K-fold cross-validation. We follow the suggestion of Shen and Huang to calculate a cross-validation score (CV).

Their algorithm proceeds as follows:

1. Randomly group the rows of an $n \times m$ matrix $\mathbf{X}$ into K almost uniformly-sized groups, denoted as $\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^K$.

2. For each $j = 0, 1, \ldots, m$, do the following;

   (a) For $k = 1, \ldots .K$, let $\mathbf{X}^{-k}$ be the data matrix $\mathbf{X}$ leaving out $\mathbf{X}^k$. Perform the spca algorithm, presented in section 9.3, on $\mathbf{X}^{-k}$ to extract its corresponding loadings $v^{-k}(j)$. Then project the left-out dataset, $\mathbf{X}^K$, onto $v^{-k}(j)$, giving us $u^k(j) = \mathbf{X}^k v^{-k}(j)$.

   (b) Calculate the K-fold CV score which is defined by

$$CV(j) = \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} {}_k \sum_{l=1}^{m} \{x_{il}^k - u_i^k(j) v_l^{-k}(j)\}^2}{n_k m},$$

   where n_k is the number of rows of $\mathbf{X}^k$, and u_i^k and v_l^{-k} are respectively the ith and lth elements of u^k and v^{-k}

3. The degree of sparsity is given by $\hat{j} = \text{argmin}_j \{CV(j)\}$.

(3) Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC, Schwarz, 1978) has been successfully applied to select the optimal number of non-zero coefficients in a penalized regression (Zou, Hastie and Tibshirani, 2007). In the penalized regression in (9.1) with fixed $(\tilde{u}_i)$, the BIC is defined as

$$BIC(\lambda) = \frac{\|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^T\|^2}{nm \cdot \hat{\sigma}^2} + \frac{log(nm)}{nm} df(\lambda),$$

where $df(\lambda)$ is the degree of sparsity with λ as the penalty parameter and σ^2 is the OLS estimate of the error variance of the model. We then choose λ or $df(\lambda)$ that minimizes the $BIC(\lambda)$.

Illustration of SMFA for Variable Selection

Suppose, we have a hypothetical data matrix, $\mathbf{Q}$, from the combined normalized datasets with $n = 70$ rows (samples) and $m = 100$ columns (variables). The heatmap of the dataset is shown in Figure 9.1. We can see that there is a subset of rows (20 samples) and a subset of columns (30 variables) that are associated.

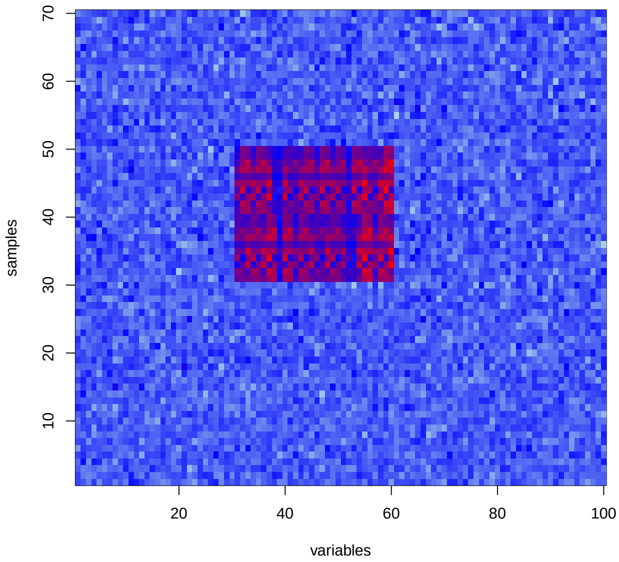
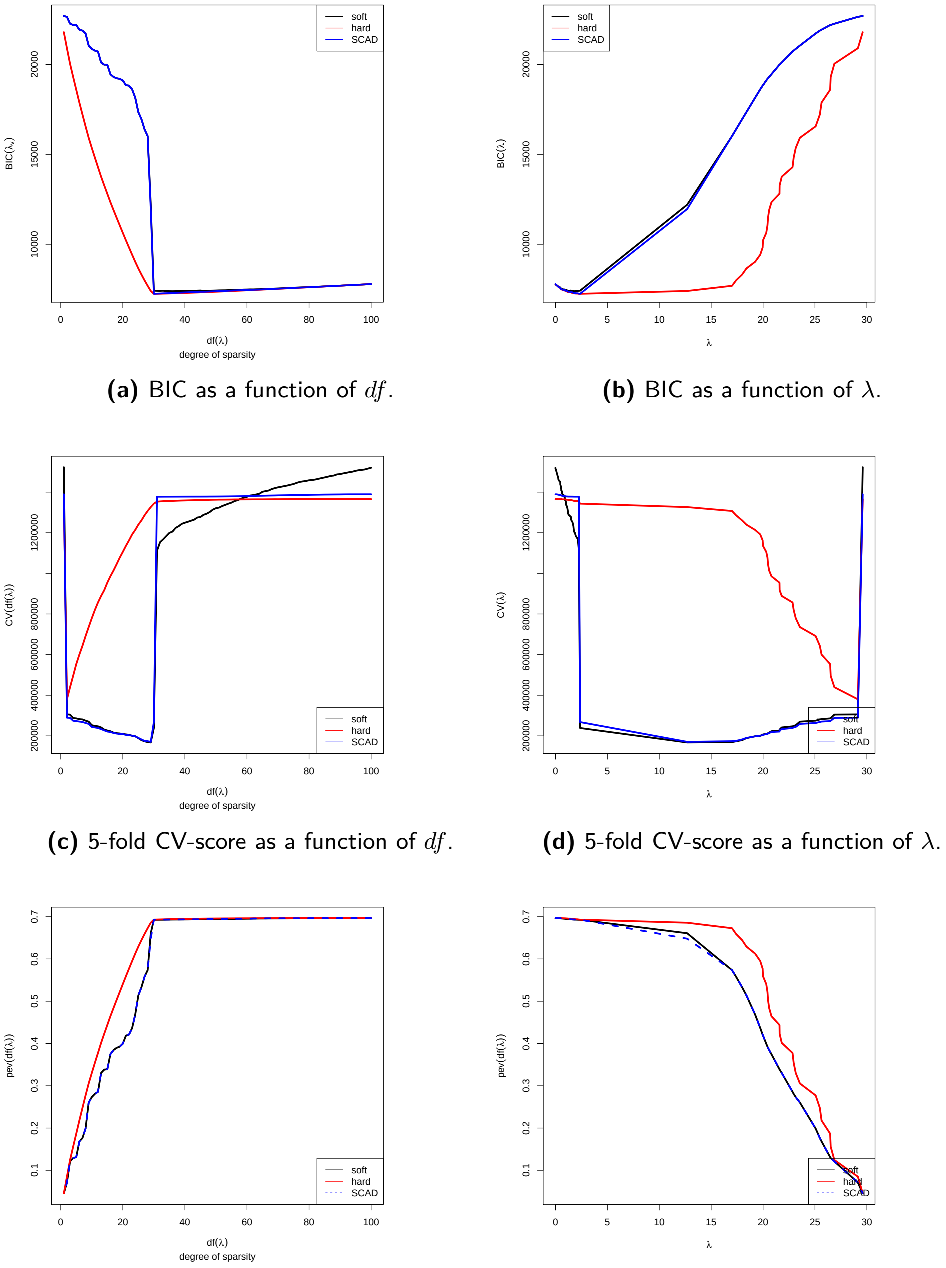


Figure 9.1: Hypothetical data: Heatmap of the 70 samples $\times 100$ variables data matrix, $\mathbf{Q}$.

We want to apply SMFA on $\mathbf{Q}$ to select variables. Using the 3 criteria to choose the optimal value of λ which is essentially equivalent to finding the optimal number of non-zeroes in the vector loadings, $df(\lambda)$. The three types of thresholding rules are accounted for in identifying the parameters. The hard-thresholding generally deviates from SCAD and soft-thresholding which behave similarly across the 3 criteria as shown in Figure 9.2. Figures 9.2b, 9.2d, 9.2f indicate that the 3 criteria point to similar choice for the $df(\lambda)$. Similarly, Figures 9.2a, 9.2c, 9.2e shows the effect of varying λ on the BIC, CV-score and PEV.

In Figure 9.3, we show the variable loadings and sample scores when using the three thresholding rules for SMFA. Here, we only induce the penalty on the loading vector. Hence, the greater the λ, the more variable loadings are shrunk to zero in Figures 9.3a, 9.3c, and 9.3e. The red lines are the 30 variables that are part of the red block in Figure 9.1. The corresponding scores do not vary across lambda since the sparsity is only imposed on the variable loadings allowing for variable selection.



(a) BIC as a function of df.

(b) BIC as a function of λ.

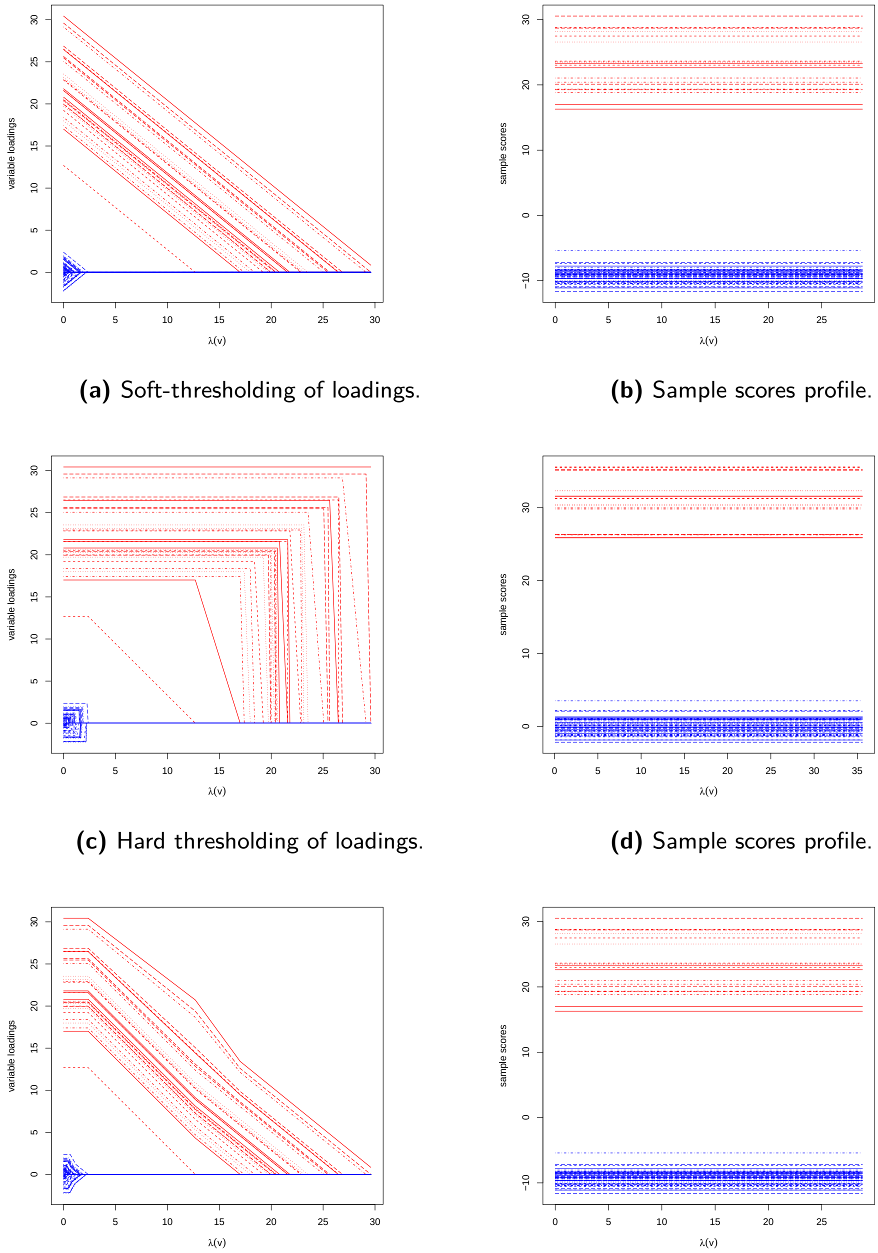
(c) 5-fold CV-score as a function of df.

(d) 5-fold CV-score as a function of λ.

(e) PEV as a function of df.

(f) PEV as a function of λ.

Figure 9.2: Penalty parameter selection using BIC, 5-fold CV and PEV.



(a) Soft-thresholding of loadings.

(b) Sample scores profile.

(c) Hard thresholding of loadings.

(d) Sample scores profile.

(e) Variable loadings using SCAD.

(f) Sample scores profile.

Figure 9.3: Variable loadings (left panels) and sample scores (right panels) across finite values of λ by thresholding rule (row panels). The red lines represent the 30 variables and 20 compounds that are part of the block structure present in Figure 9.1.

9.4 Biclustering of Multi-Source Data Using the biMFA Algorithm

SMFA serves as a dimension reduction or feature selection technique producing modified PCs with sparse loadings that can explain most of the variation in a dataset. However, if the aim is to simultaneously identify subset of rows and subset of columns of $\mathbf{X}$ that are associated, then a biclustering method is more appropriate. While SPCA imposes sparsity on the loadings of the principal components, the Sparse SVD (SSVD) proposed by Lee *et al.* (2010) imposes sparsity on both the left and right singular vectors in order to extract sub-matrices showing row-column interactions when forming a low-ranked approximation to the data matrix. Combining the ideas of MFA for data integration and the SSVD as biclustering leads to biclustering using MFA (biMFA)..

In this case, decomposing $\mathbf{X}$ by the SVD, each submatrix $\mathbf{X}_i$ will be associated with a singular vector pair$(\mathbf{u}_i, \mathbf{v}_i)$ such that non-zero coefficients in $\mathbf{u}_i$ represent the rows that belong to $\mathbf{X}_i$ and the non-zero coefficients in $\mathbf{v}_i$ represent the columns that belong to $\mathbf{X}_i$.

Following Lee's SSVD approach, we can obtain the sparse vectors $\mathbf{u}$ and $\mathbf{v}$ by minimizing the following penalized sum of squares criterion,

$$||\mathbf{X} - \mathbf{u}\tilde{\mathbf{v}}^T||_F^2 + \lambda_{\mathbf{u}} P_1(\tilde{\mathbf{u}}) + \lambda_{\mathbf{v}} P_2(\tilde{\mathbf{v}}), \quad \text{with} \quad \tilde{\mathbf{u}} = \sigma\mathbf{u}, \tilde{\mathbf{v}} = \sigma\mathbf{v}, \tag{9.2}$$

where $P_1(\tilde{\mathbf{u}})$ and $P_2(\tilde{\mathbf{v}})$ are sparsity inducing penalty terms and $\lambda_{\mathbf{u}}$ and $\lambda_{\mathbf{v}}$ are 2 non-negative penalty parameters that balance the goodness-of-fit measure of $||\mathbf{X} - \mathbf{u}\tilde{\mathbf{v}}^T||_F^2$ and the penalty terms. Here, it allows for varying levels of sparsity for $\mathbf{u}$ and $\mathbf{v}$. When $\lambda_{\mathbf{u}} = \lambda_{\mathbf{v}} = 0$, the criterion in (9.2) reduces to SVD. Moreover, by setting $\tilde{\mathbf{v}} = \sigma\mathbf{v}$ and $\lambda_{\mathbf{u}} = 0, \lambda_{\mathbf{v}} = \lambda$ and $P_2(\cdot) = P(\cdot)$ in (9.2), the SPCA criterion in (9.1) is obtained.

Recall that in Section 9.3, we have used the lasso penalty given by $P(\tilde{v}) = \sum_{j=1}^{m} p_\lambda(|\tilde{v}_j|)$ and the estimation of the sparse loading vector $\tilde{v}_j$ depends on the choice of the thresholding rule.

In biMFA, we use the adaptive lasso penalties given by

$$P_2(\tilde{v}) = \sigma \sum_{j=1}^{m} w_{2,j}(|\tilde{v}_j|) \quad \text{and} \quad P_1(\tilde{u}) = \sigma \sum_{i=1}^{n} w_{1,i}(|\tilde{u}_i|),$$

where $w_{1,i}$ and $w_{2,j}$ are possibly data-driven weights. When $w_{1,i} = w_{1,j} = 1$, we obtain the lasso penalty. In this biclustering algorithm, the weights are set to $w_{1,i} \equiv (w_{1,1}, \ldots, w_{1,n}) = |\tilde{u}|_1^\gamma$ and $w_{2,j} \equiv (w_{2,1}, , w_{2,m}) = |\tilde{v}|_2^\gamma$, following the suggestion of

Zou *et al.* (2006). Some suggested values for γ_1 or γ_2 are 0 (corresponds to lasso fit), 1 (similar to nonnegative garrote (Breiman, 1995)) and 0.5 and 0.2 by Zou *et al.* (2006). The iterative algorithm presented in Section (9.3) to estimate sparse loading vectors is extended to also estimate the sparse scores vectors to extract a bicluster.

1. Initialize: Apply the standard SVD to $\mathbf{X}$ and obtain the best rank-one approximation of $\mathbf{X}$ as $\sigma_1 \mathbf{u}^* \mathbf{v}^{*T}$ where $\mathbf{u}^*$ and $\mathbf{v}^*$ are unit vectors. Set $\tilde{\mathbf{v}}_{old} = \sigma_1 \mathbf{v}^*$ and $\tilde{\mathbf{u}}_{old} = \sigma_1 \mathbf{u}^*$.

2. Update:
   (a) For a fixed $\tilde{\mathbf{u}}$ the optimal $\tilde{v}_j$ can be obtained by applying a thresholding rule to the vector $\mathbf{X}^T \tilde{\mathbf{u}}_{old}$ which depends on the form of the penalty function and the choice of γ_2.

   Similar to Section 9.3, we can use the soft-thresholding penalty (Tibshirani, 1996), hard thresholding (Donoho,1994) and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) to solve for the closed-form solution of (9.2) and we respectively get

   $$\tilde{v}_{j,new}^{soft} = \text{sign}(\mathbf{X}^T \tilde{\mathbf{u}}_{old})(|\mathbf{X}^T \tilde{\mathbf{u}}_{old}| - \lambda_v w_{2,j}/2)_+,$$
   $$\tilde{v}_{j,new}^{hard} = I(|\mathbf{X}^T \tilde{\mathbf{u}}_{old}| > \lambda_v w_{2,j}/2)(\mathbf{X}^T \tilde{\mathbf{u}}_{old}),$$
   $$\tilde{v}_{j,new}^{SCAD} = \begin{cases} \text{sign}(\mathbf{X}^T \tilde{\mathbf{u}}_{old})(\mathbf{X}^T \tilde{\mathbf{u}}_{old} - \lambda_v w_{2,j}/2)_+ & \text{if } |\mathbf{X}^T \tilde{\mathbf{u}}_{old}| \le 2\lambda_v w_{2,j}/2, \\ (a-1)\mathbf{X}^T \tilde{\mathbf{u}}_{old} - \text{ sign } (\mathbf{X}^T \tilde{\mathbf{u}}_{old}) a(\lambda_v w_{2,j}/2)/(a-2) & \text{if } 2\lambda_v w2, j/2 < |\mathbf{X}^T \tilde{\mathbf{u}}_{old}| \le a\lambda_v w_{2,j}/2, \\ \mathbf{X}^T \tilde{\mathbf{u}}_{old} & \text{if } |\mathbf{X}^T \tilde{\mathbf{u}}_{old}| > a\lambda_v w_{2,j}/2. \end{cases}$$

   The λ_v that minimizes the $BIC(\lambda_v)$ and the K-fold CV-score as well as following the selection based on the adjusted variance as previously discussed is selected. Standardize $\tilde{\mathbf{v}} = \tilde{\mathbf{v}}_{new}/\|\tilde{\mathbf{v}}_{new}\|$.

   (b) Similarly, for a fixed $\tilde{\mathbf{v}}$, we can obtain the optimal $\tilde{u}_i$ by applying a thresholding rule, to the vector $\mathbf{X}\tilde{\mathbf{v}}$ instead of $\mathbf{X}^T \tilde{\mathbf{u}}_{old}$ as shown in 2a. The λ_u that minimizes the $BIC(\lambda_u)$ and the K-fold CV-score as well as on the bases of the adjusted variance is chosen. Standardize the $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}_{new}/\|\tilde{\mathbf{u}}\|$.

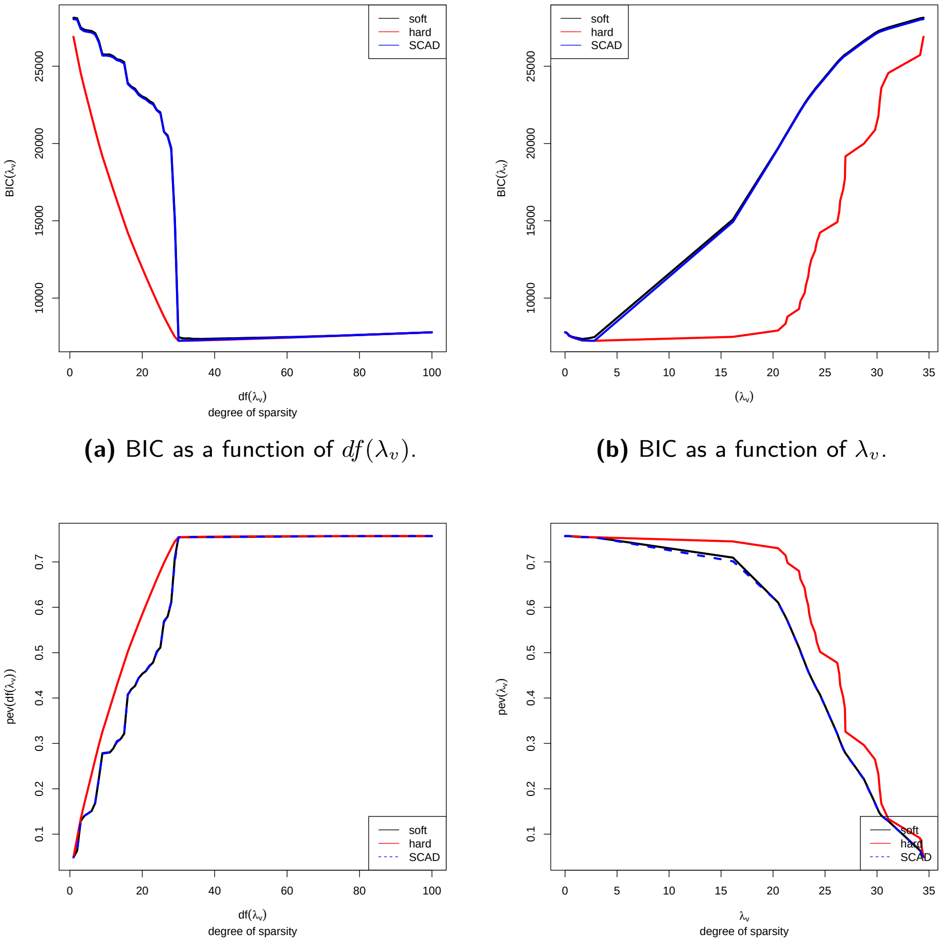
3. Repeat Step 2 replacing $\tilde{\mathbf{u}}_{old}$ and $\tilde{\mathbf{v}}_{old}$ by $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$, respectively, until convergence.

4. Set $\mathbf{u} = \tilde{\mathbf{u}}, \mathbf{v} = \tilde{\mathbf{v}}$ and $\sigma = \mathbf{u}^T \mathbf{X} \mathbf{v}$ at convergence.

In order to search for more biclusters, we use the residual matrix, $\mathbf{X} - \sigma \mathbf{u}\mathbf{v}^T$ as input to the iterative algorithm.

Illustration of biMFA for Bicluster Extraction

When the interest is on detecting a bicluster, then biMFA can be used which imposes sparsity on both dimensions, the scores and the loadings. We can use the same set of criteria to assess the values of the tuning parameters for both dimensions. For illustration, we apply biMFA on the same dataset presented in Figure 9.1. We have the BIC and PEV for the variable loadings across a set of λ_v and presented in Figure 9.4. Similarly, in Figure 9.5, we use the BIC and PEV for the sample scores across λ_u and $df(\lambda_u)$.



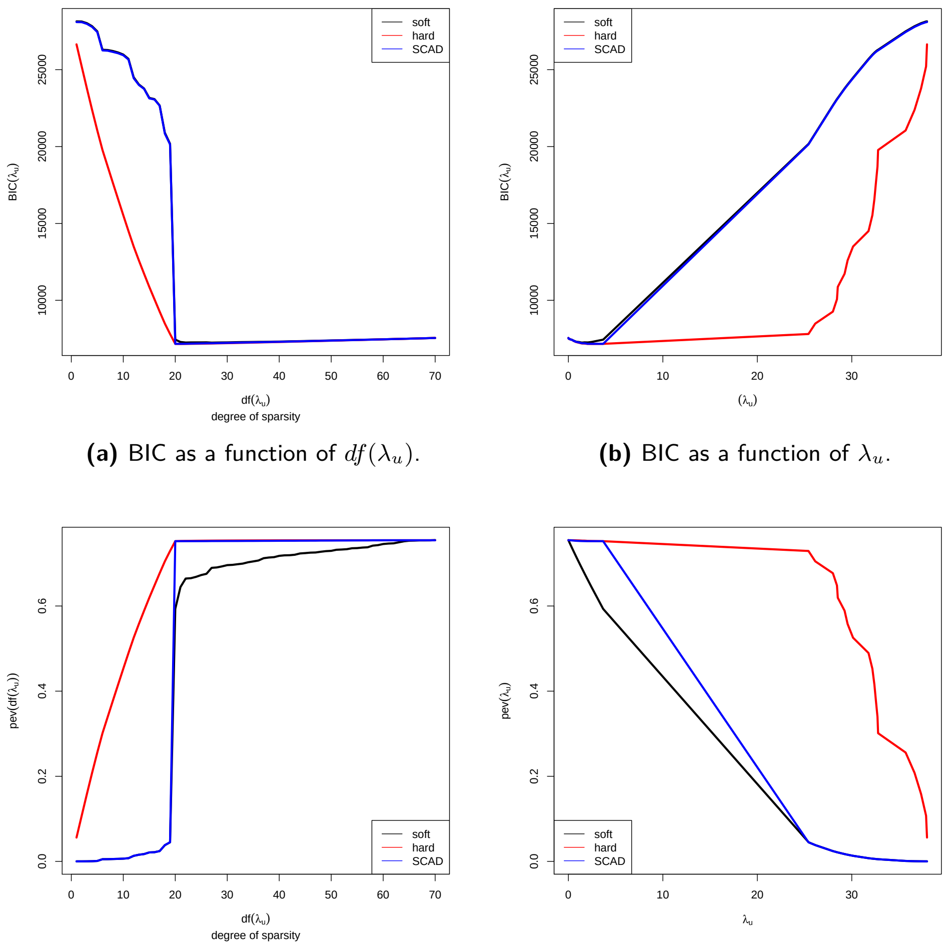
(a) BIC as a function of $df(\lambda_v)$.

(b) BIC as a function of λ_v.

(c) PEV as a function of $df(\lambda_v)$.

(d) PEV as a function of λ_v.

Figure 9.4: biMFA: Selection of λ_v using BIC and PEV.



(a) BIC as a function of $df(\lambda_u)$.

(b) BIC as a function of λ_u.

(c) PEV as a function of $df(\lambda_u)$.

(d) PEV as a function of λ_u.

Figure 9.5: biMFA: Selection of λ_u using BIC and PEV.

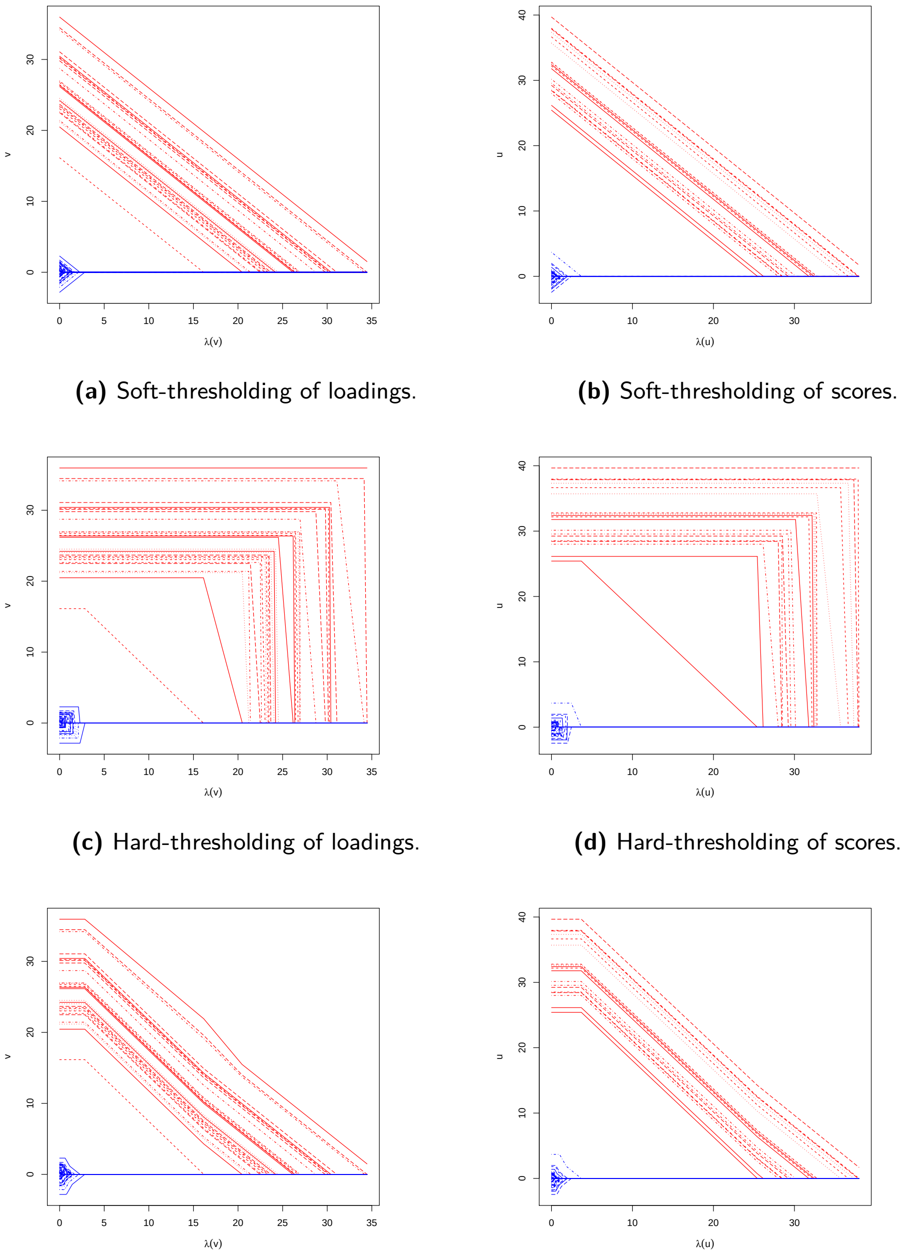
Figure 9.6 displays the loadings (panels a,c,e) and scores (panels b,d,f) when varying the penalty parameter λ_v and λ_u, respectively. The double penalization gives a subset of variables and a subset of samples subject to the choice of λ_v and λ_u, where the choice can be guided by the BIC and PEV presented in Figures 9.4 and 9.5. After fixing the tuning parameters, the members of the biclusters may depend on the thresholding rule used. Note that, for biMFA, we can use the adaptive lasso penalties which require the specification of γ_v and γ_u. So far, what we have presented uses the lasso penalty, that is setting $\gamma_v = 0$ and $\gamma_u = 0$. Figure 9.7a shows the BIC values across λ_v by $\gamma_v = 0, 1, 2$ while Figure 9.7b displays the BIC values across λ_u by $\gamma_u = 0, 1, 2$. We can see that $\gamma_u = \gamma_v = 2$ would generally give a lower BIC for any λ_v and λ_u. Figure 9.8 gives the loadings and scores profiles across different choices of λ fixing $\gamma_v = \gamma_u = 2$. It is easy to see that at $df(\lambda_v) = 30$ and $df(\lambda_u) = 20$ and their corresponding λ_v and λ_u that gives the minimum BIC. Using the biMFA algorithm with this tuning parameters, we can partition the observed data matrix into 2 matrices presented in Figure 9.9: the predicted matrix (9.9b) and the residual matrix (9.9c).

9.5 Application to the PDE10 Dataset

For the integration of HCS and gene expression data, we get a combined normalized matrix that contains 16 compounds × 736 genes and HCS features combined.

If the aim is limited to feature selection we can apply SMFA which is just equivalent to a biMFA analysis without penalizing the scores. Also, by getting rid of the penalty terms in biMFA and SMFA, the analysis reduces to that of MFA. Using the combined normalized HCS and gene expression datasets, we run MFA, sMFA and biMFA. Figure 9.10 compares the feature loadings when using MFA versus SMFA and biMFA for one factor. Both induces sparsity on the loadings to select important features. The difference on the number of non-zero loadings between SMFA and biMFA mainly depends on the tuning parameters. In addition, biMFA identifies genes that are co-regulated on a subset of compounds whereas SMFA groups features that exhibits similar profiles across all compounds. The two methods, however, are quite consistent in selecting features with relatively high loadings. With respect to the compound scores, the idea is to automatically associate a subset of compounds to a subset of features. The biMFA induces sparsity on the compound scores which is not done in SMFA. Each factor derived by biMFA now corresponds to a bicluster consisting of subset of features and compounds with non-zero loadings and scores, respectively.

Now, we will focus on the results of applying biMFA where we identified several biclusters in the integrated datasets. Focusing on one-bicluster, we can see from Figures 9.11a



(a) Soft-thresholding of loadings.

(b) Soft-thresholding of scores.

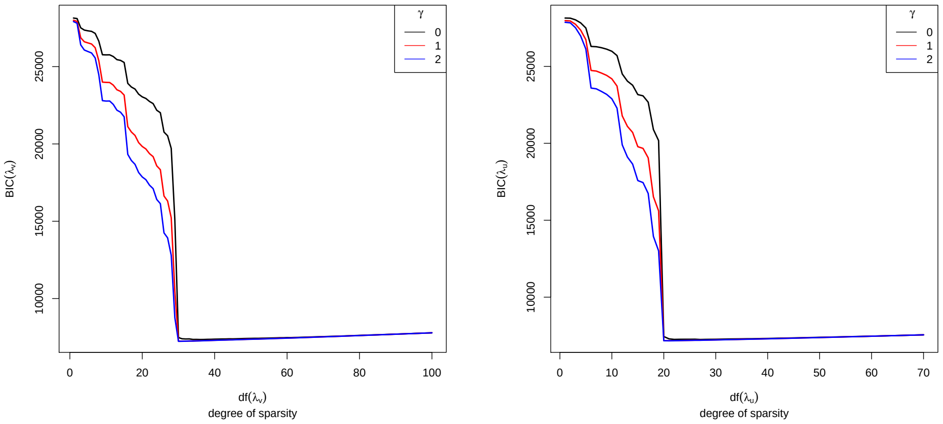
(c) Hard-thresholding of loadings.

(d) Hard-thresholding of scores.

(e) Variable loadings using SCAD.

(f) Sample scores using SCAD.

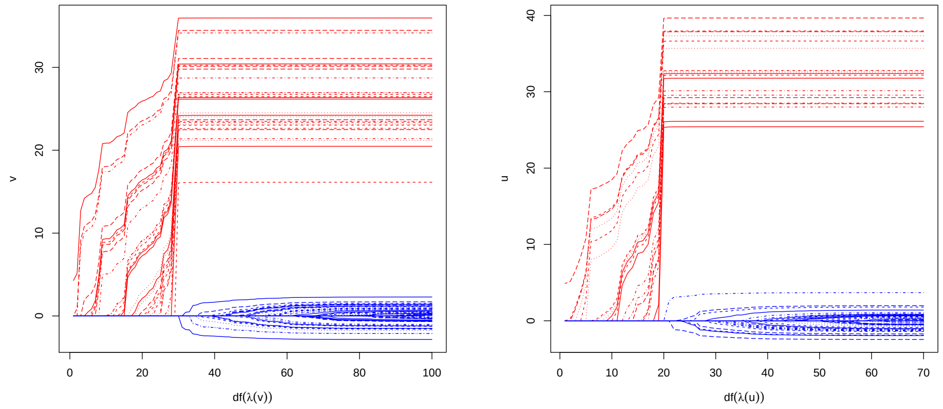
Figure 9.6: Variable loadings (left panels) and sample scores (right panels) across finite values of λ_v and λ_u by thresholding rule (row panels). The red lines represent the 30 variables and 20 compounds that are part of the block structure present in Figure 9.1.



(a) BIC as a function of λ_v by γ_v.

(b) BIC as a function of λ_u by γ_u.

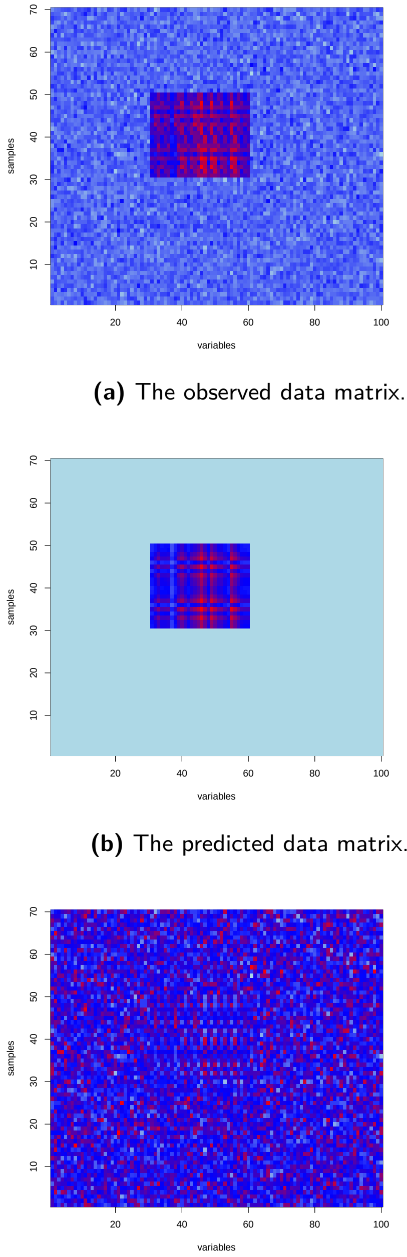
Figure 9.7: biMFA: Selection of γ_v and γ_u using BIC with soft-thresholding rule applied on varying values of λ_v and λ_u, respectively.



(a) Variable loadings across $df(\lambda_v)$.

(b) Sample scores across $df(\lambda_u)$.

Figure 9.8: Variable loadings (left panel) and sample scores (right panel) across a set of finite values of the degree of sparsity using the adaptive lasso penalty weights fixed at $\gamma_v = 2$ and $\gamma_u = 2$ applying the soft-thresholding. The red lines represent the 30 variables and 20 compounds that are part of the block structure present in Figure 9.1.



(a) The observed data matrix.

(b) The predicted data matrix.

(c) The residual matrix.

Figure 9.9: Hypothetical data: The observed, predicted and the residual matrices when running a biMFA analysis fixing the degree of sparsity to $df(\lambda_v) = 30$ and $df(\lambda_u) = 20$.

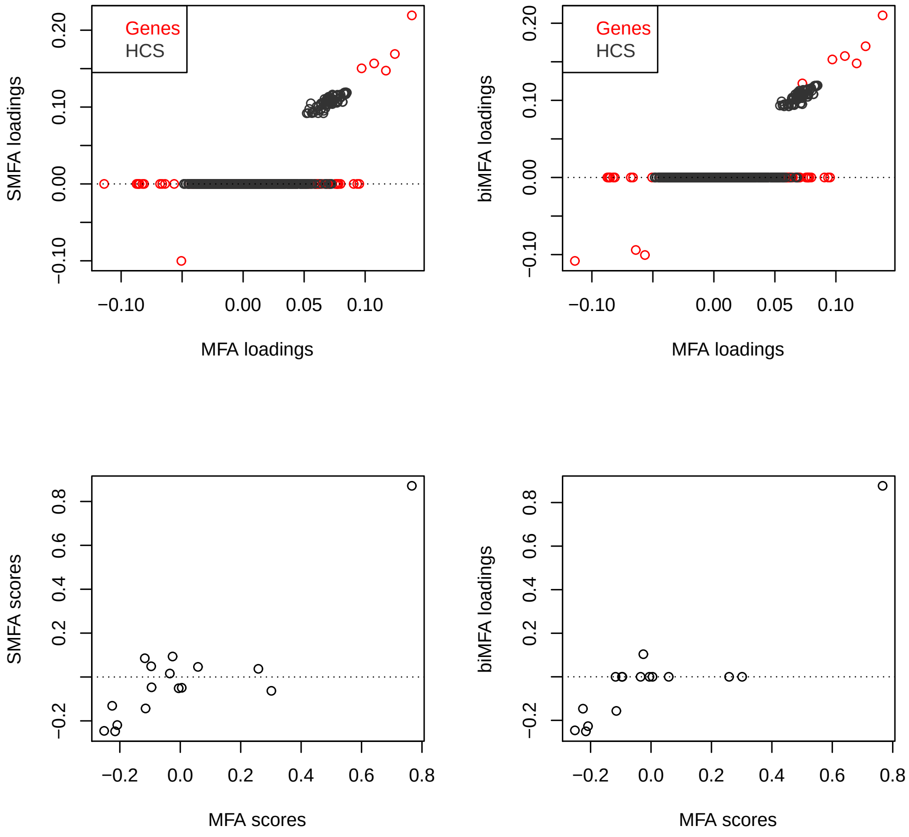
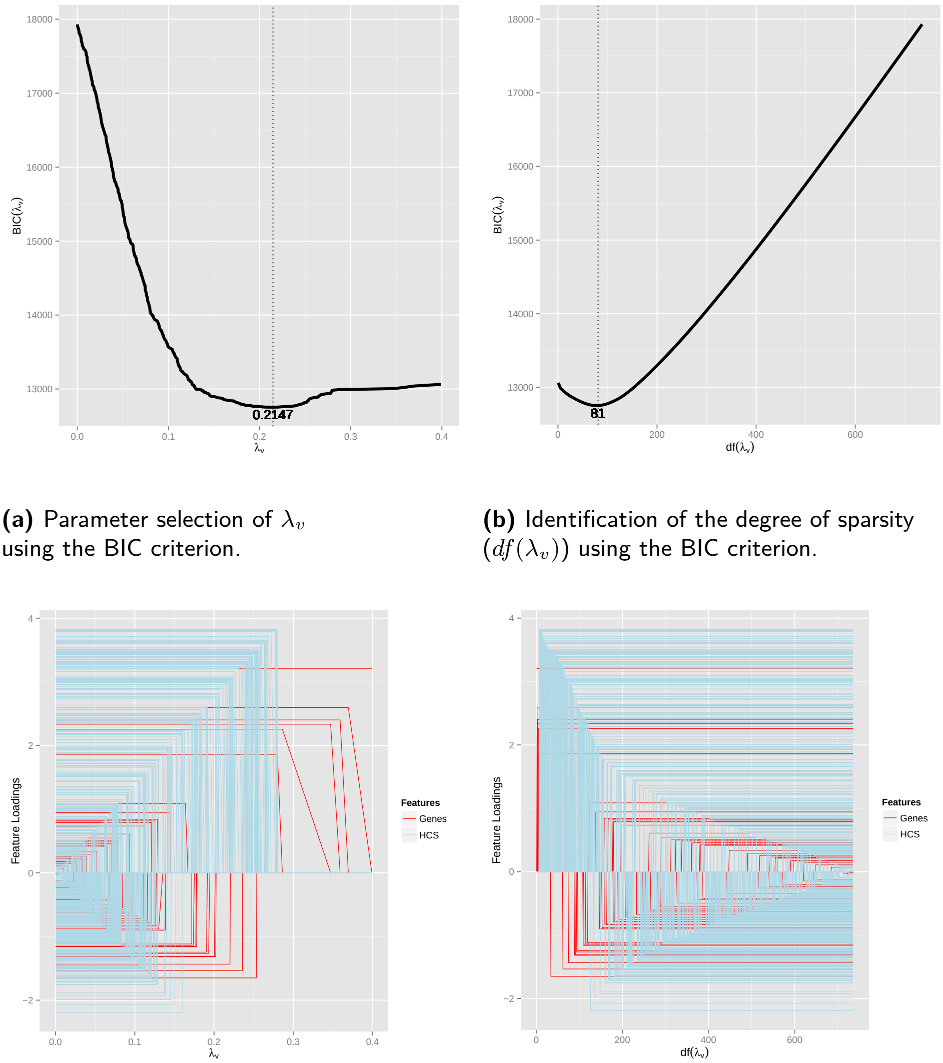


Figure 9.10: Comparison of loadings from MFA and SMFA and biMFA.

and 9.11b that we have 81 features for this bicluster with corresponding $\lambda_v = 0.21$. Figures 9.11c and 9.11d shows the feature loading profiles across the values of the tuning parameters. The same set of plots are presented in Figure 9.12 for the compound scores. Figure 9.12b shows that 7 compounds are part of this bicluster. The compound score profiles across λ_u and $df(\lambda_u)$ are given in Figures 9.12c and 9.12d.

Figures 9.13 and 9.14 shows the loadings(scores) per bicluster which is helpful in exploring the number of features (compounds) with non-zero coefficients. The first and third bicluster are mainly characterized by similar set of genes. Bicluster 2 and 5 are interesting since they share 2 common compounds, Taxotere and compound JnJ-h1 (Figure



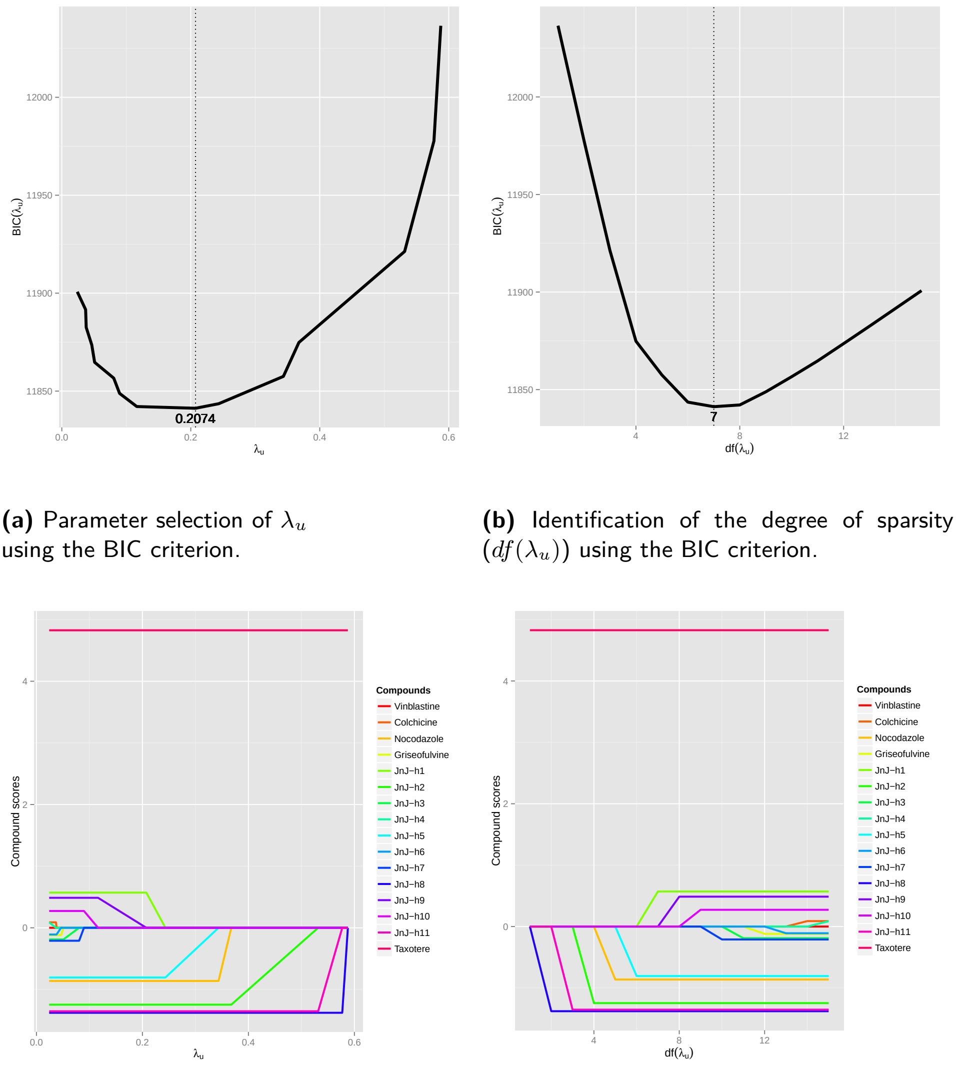
(a) Parameter selection of λ_v using the BIC criterion.

(b) Identification of the degree of sparsity $(df(\lambda_v))$ using the BIC criterion.

(c) Profiles of feature loadings across the values of λ_v.

(d) Profiles of feature loadings across the values of $df(\lambda_v)$.

Figure 9.11: The effect of tuning parameters λ_v and $df(\lambda_v)$ on the loadings using hard thresholding.



(a) Parameter selection of λ_u using the BIC criterion.

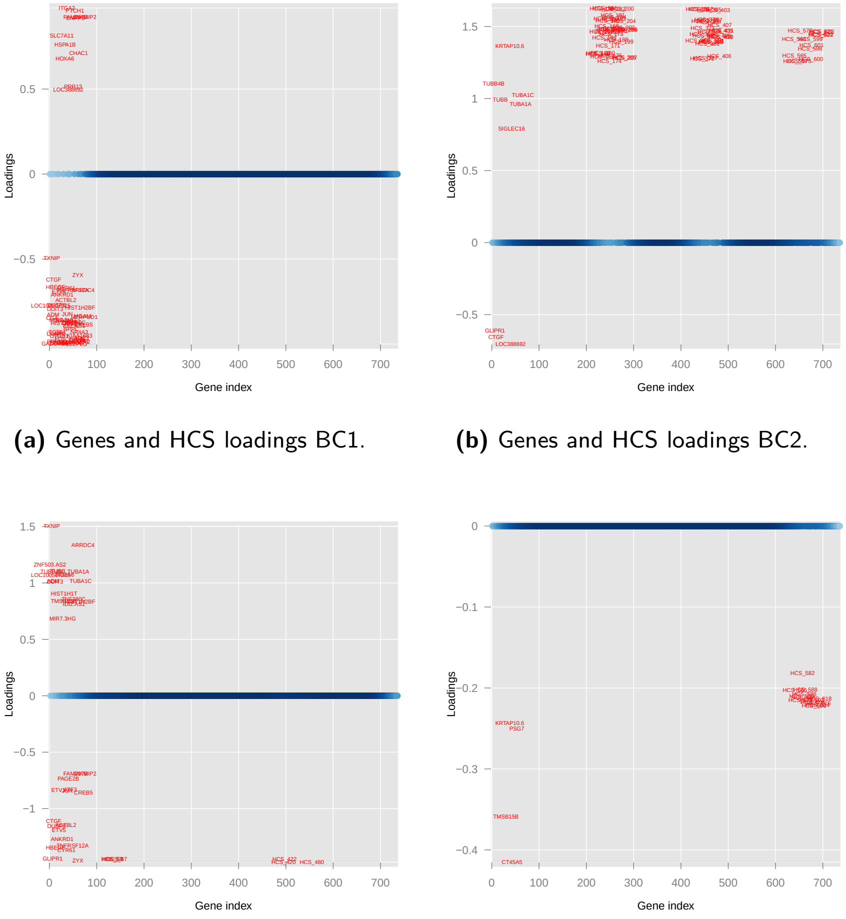
(b) Identification of the degree of sparsity ($df(\lambda_u)$) using the BIC criterion.

(c) Profiles of compound scores across the values of λ_u.

(d) Profiles of compound scores across the values of $df(\lambda_u)$.

Figure 9.12: The effect of tuning parameters λ_u and $df(\lambda_u)$ on the compound score using hard thresholding.

9.15b) but only one feature, the gene KRTAP10-6 (Figure 9.15a). Moreover, the genes that dominates bicluster 2 are Tubulin genes. The profiles plot of the member features are presented in Figure 9.16 where the first compounds marked in red are part of the bicluster. The features are only co-regulated on the compounds within the bicluster.



(a) Genes and HCS loadings BC1.

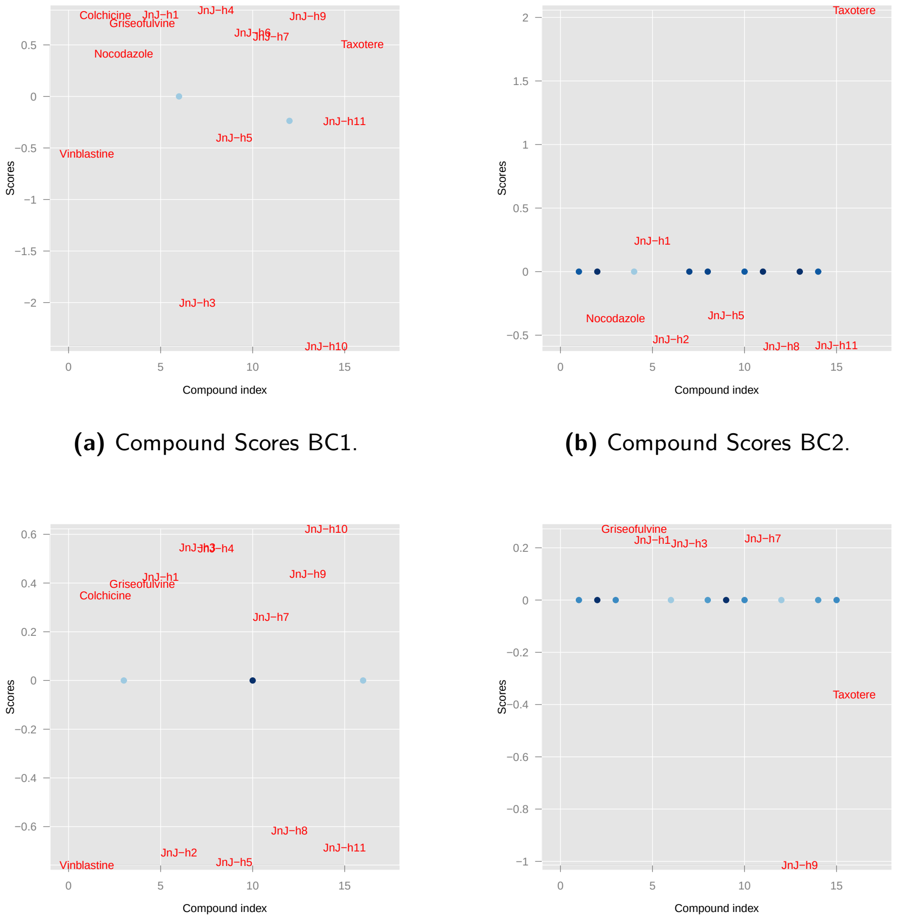
(b) Genes and HCS loadings BC2.

(c) Genes and HCS loadings BC3.

(d) Genes and HCS loadings BC5.

Figure 9.13: Variable loadings of Biclusters with eigenvalue>1.

For the interpretation of the results, we focus on the second bicluster which is characterized by the downregulation of Tubulin genes. Down regulation of Tubulin genes indicates also a possible genotoxic effect on the microtubule-based chromosome segregation. This is exhibited by 5 compounds since Taxotere and JnJ-h1 shows upregulation.



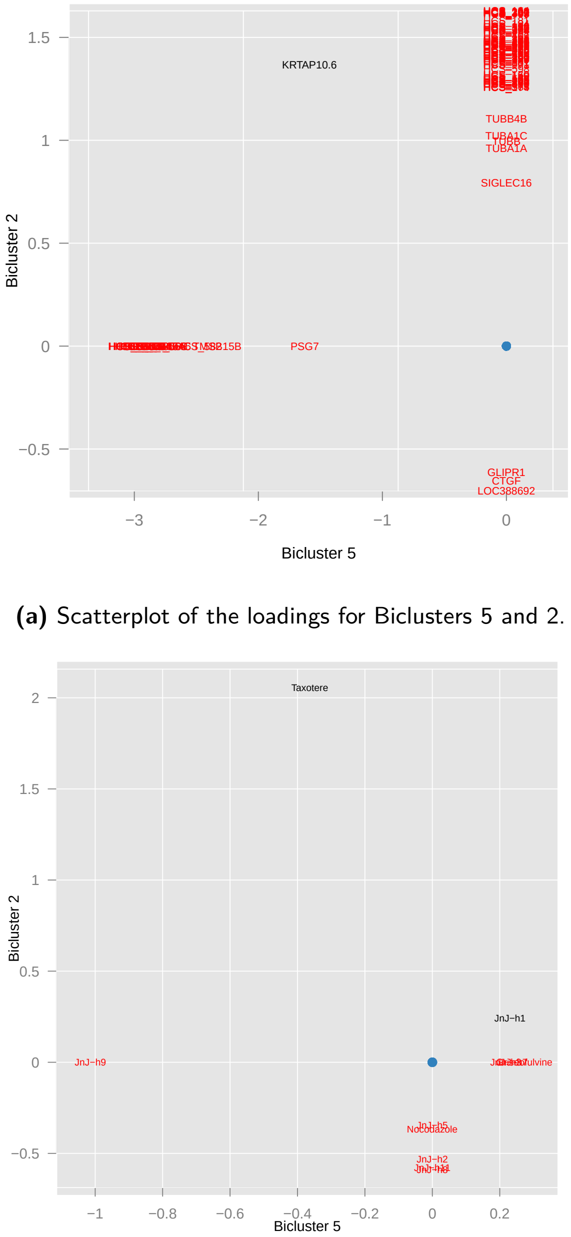
(a) Compound Scores BC1.

(b) Compound Scores BC2.

(c) Compound scores BC3.

(d) Compound scores BC5.

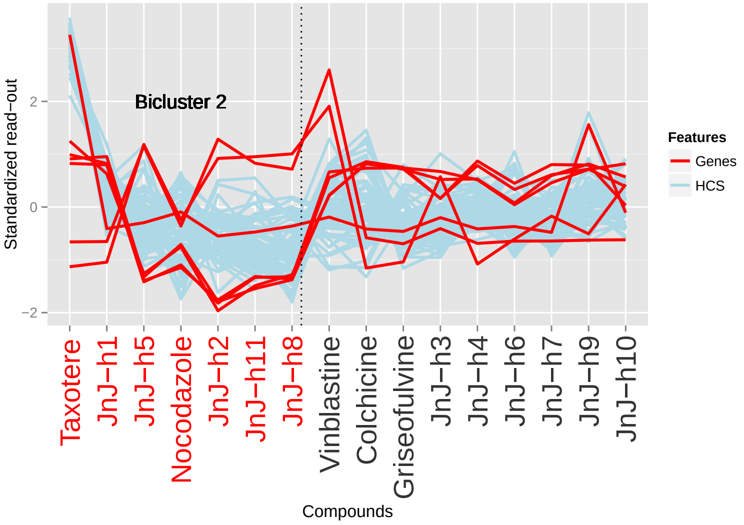
Figure 9.14: Compound scores of Biclusters with eigenvalue>1.



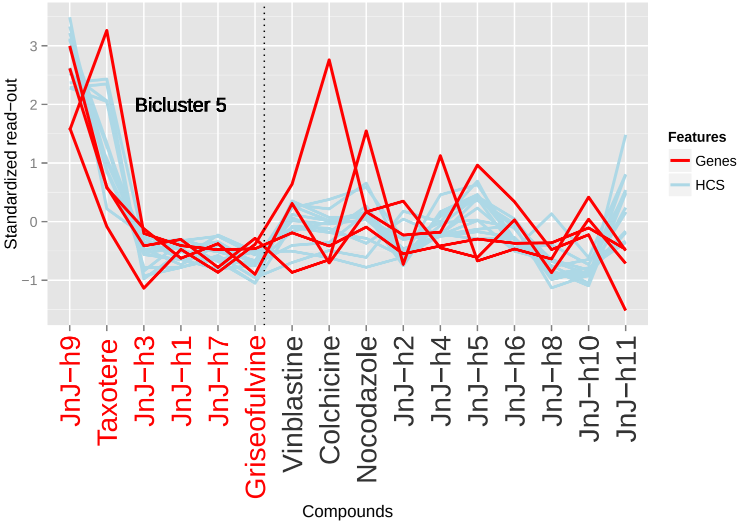
(a) Scatterplot of the loadings for Biclusters 5 and 2.

(b) Scatterplot of the scores for Biclusters 5 and 2.

Figure 9.15: Scatterplot of loadings and scores of BC2 and BC5 highlighting the member features and compounds.



(a) Profiles plot of features in BC2.



(b) Profiles plot of features in BC5.

Figure 9.16: Profiles plot of features in BC2 and BC5.

In fact, in this experiment, Nocadozole and Taxotere are used as positive and negative control, respectively for the downregulation of tubulin. Moreover, the HCS features identified in the Tubulin-linked bicluster are microtubule related features that are located in the cytoplasm and cells whereas those in the bicluster 5 are mostly Tubulin-linked features in the nucleus.

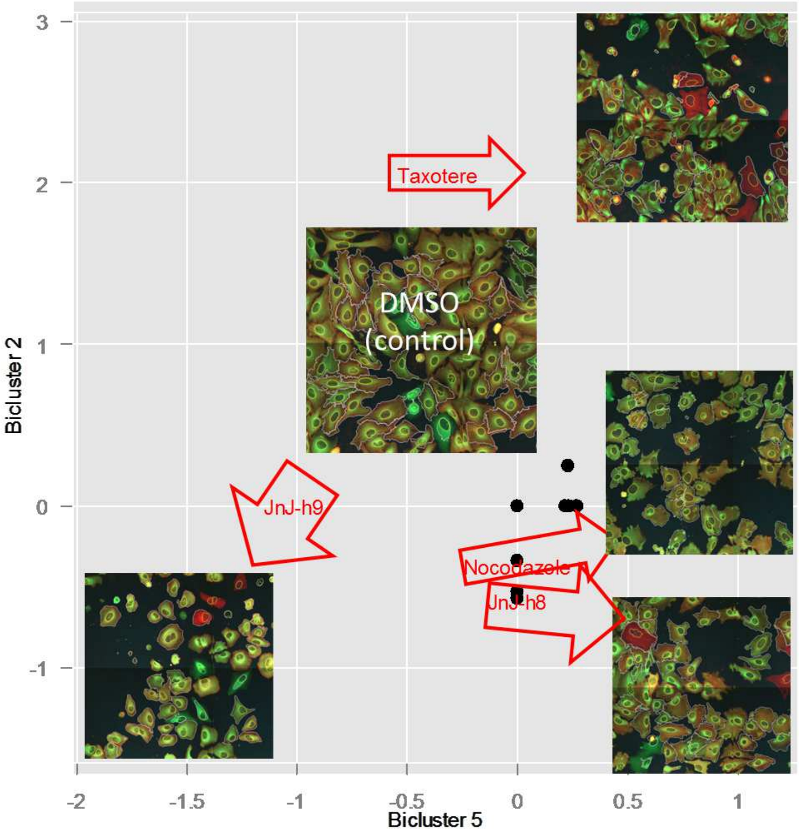


Figure 9.17: Compound scores of Bicluster 5 and Bicluster 2. Compounds that are outlying with respect to Bicluster 2 are identified with their respective HCS Images after compound administration.

Inspecting the high content images of some compounds (Figure 9.17), compared to the DMSO, compound JnJ-h8, identified to be down-regulating Tubulin genes has little green dots present at the border of the cell while long green lines in the cells are observed in Taxotere. This is an indication of microtubule aggregation which is an indication of genotoxicity. Therefore compounds that downregulate Tubulin genes are indicating toxic effects and thus should be carefully examined in order to make a go/no go decision in the lead optimization process.

9.6 Discussion

In this chapter, we present a new biclustering approach which include feature selection on a properly integrated datasets. The usefulness of Multiple Factor Analysis for the integration of several datasets has been explored in different applications. In the drug development setting, however, applying this method to multiple sources of high-dimensional data leads to difficulty in identifying and interpreting few relevant features that can well characterize the activities of the lead compounds. Hence, a modification of MFA is proposed in this chapter by making use of the sparse extension of PCA, such as sparse PCA for dimension reduction and feature selection and SSVD for finding biclusters. This leads to the two unsupervised integrative methods which we term SMFA and biMFA.

The results show that combining gene expression data and high content screening can jointly uncover biological insights about the compounds. Identifying potentially genotoxic compounds early in the drug development pipeline can prevent the discovery of toxic effects of lead compounds at a later stages. This means that this strategy can save considerable amount of investments, both in time and money. Failed compounds may still be repurposed or altered to exhibit desirable characteristics. SMFA and biMFA are useful methods to explore structures present in several data sets. The output of these methods allow scientists to make informed decisions that can guide medicinal chemists on which lead compounds to prioritise early on in the development.

Part IV

Software Products

Chapter 10

The biclustRank R Package

In Chapter 6 we presented an analysis in which biclusters were ranked based on the information content of the bicluster or the chemical similarity of the compounds that belong to the bicluster. The analysis was conducted using a new R package developed for this purpose, the `biclustRank` R package (Perualila, *et al., 2016*).

Consider a data matrix $\mathbf{X}$ and let assume that our aim is to find local patterns in the $\mathbf{X}$. The R package `biclustRank` allows to use the output of FABIA (Hochreiter *et al.*, 2006), Plaid or ISA for ranking. Hence, in the first stage, biclustering can be done using in the following way:

```
> resFabia <- fabia(X, p=10, cyc=1000, alpha=0.1,  random=0)

> resIsa <-   isa(X)
> resPlaid <- biclust(X, method = BCPlaid(), back.fit = 2,
shuffle = 3, fit.model = ~m + a + b, iter.startup = 5,
iter.layer = 30, verbose = F)
```

Let us focus on the result obtained for FABIA which are stored in the R object `resFabia`. The biclustering result can be viewed using the function `summary`. The first list is the information content of biclusters, the second list is the information content of samples, the third statistics is the factors per bicluster, and the last statistics is the loading per bicluster.

```
> summary(resFabia)

An object of class Factorization

call:
"fabia"

Number of rows:  566
Number of columns:  62
Number of clusters:  10

 Information content of the clusters:
   BC 1     BC 2     BC 3     BC 4     BC 5
 361.15   188.57   170.31   169.85   162.32
   BC 6     BC 7     BC 8     BC 9    BC 10
 161.57   154.31   135.61   133.17   132.82
 BC sum
1764.04

 Information content of the samples:
  Sample 1    Sample 2    Sample 3    Sample 4
     29.26       28.10       31.32       30.78
  ....
 Sample 61   Sample 62  Sample sum
     26.46       27.08     1764.04

 Column clusters / Factors:
 BC 1              BC 2
 Min.   :-7.7939   Min.   :-2.07772
 ...

 Row clusters / Loadings:
 BC 1               BC 2
 Min.   :-1.03309   Min.   :-0.415639
 1st Qu.:-0.00109   1st Qu.:-0.107525
 ...
```

In the analysis presented above, the biclusters were ranked based on the information content of the biclusters. In many cases, there is a need to rank the biclusters based on a different criterion which was NOT used for biclustering. For example, ranking the biclusters based on the biological functionality of the genes belonging to the biclusters or based on the chemical similarity of the compounds belonging to the biclusters. For the latter, the R package `biclustRank` can be used. In the first step we need to identify the rows and columns belong to each one of the biclusters. This can be done using the function `extractBicList`.

```
> library(biclustRank)
> bicF <- extractBicList(data = X,
          biclustRes = resFabia, p=10, bcMethod="fabia")

> str(bicF)
List of 10
 $ BC1 :List of 2
  ..$ samples: chr "JnJ-xx9"
  ..$ genes  : chr [1:63] "LOC100288637" "KRTAP5-3"  ...
 $ BC2 :List of 2
  ..$ samples: chr [1:18] "JnJ-xx3" "JnJ-xx20"  ...
  ..$ genes  : chr [1:42] "PSMB6" "RPS13" "NUDT5"  ...
 ...
 $ BC10:List of 2
  ..$ samples: chr [1:11] "JnJ-xx12" "JnJ-xx9" ...
  ..$ genes  : chr [1:2] "TBC1D3B" "LOC100506667"
```

We can use the function `Distance` to compute for the distance matrix of **Z**, given by $\mathbf{D}_n$. We subtract $\mathbf{D}_n$ from 1 to get the similarity scores $\mathbf{S_n}$.

```
Dn <- Distance(dataB = t(Z), distmeasure = "tanimoto")
Sn <- 1-Dn #similarity matrix for n compounds
```

The similarity score per biclusters can be calculated by

```
> Sk <- extractSimBC(biclustList = bicF, simMat = Sn, p=10 )
> SkScores <- getLowerSim(simMat)
> str(SkScores)
List of 8
 $ BC2 : num [1:153] 0.317 0.278 0.106 0.41 0.308 ...
 $ BC3 : num [1:55] 0.0476 0.1061 0.2319 0.2321 0.25 ...
 $ BC4 : num [1:10] 0.239 0.305 0.673 0.407 0.328 ...
 $ BC6 : num [1:15] 0.509 0.379 0.277 0.315 0.328 ...
 $ BC7 : num [1:6] 0.524 0.5 0.264 0.769 0.457 ...
 $ BC8 : num [1:3] 0.44 0.352 0.365
 $ BC9 : num [1:190] 0.4894 0.0909 0.3571 0.1525 0.0845 ...
 $ BC10: num [1:55] 0.273 0.203 0.138 0.611 0.271 ...
```

This results are summarized using the function `statTab`. The `ordering` arguments takes a value of 0 to 7 with 0 as default. This allows to reorder the table's rows (i.e., rank the biclusters) according to several statistics with 0=no reordering, 1= similarity mean, 2= similarity median, 3=SD(standard deviation), 4= CoefVar(coefficient of variation), 5= Range, 6 = MAD(mean absolute deviation), and 7= IQR (interquartile range).

```
> statTab(SkScores,ordering=4)
```

Figure 6.3 was produced with the function`boxplotBC`

```
> boxplotBC(simVecBC,qVal=0.5,rank=FALSE)
 BC2  BC3  BC4  BC6  BC7  BC8  BC9 BC10
0.17 0.19 0.33 0.43 0.48 0.37 0.14 0.23

> boxplotBC(SkScores,qVal=0.5,rank=TRUE)
 BC7  BC6  BC8  BC4 BC10  BC3  BC2  BC9
0.48 0.43 0.37 0.33 0.23 0.19 0.17 0.14
```

The arguments `qVal=0.5` and rank=TRUE implies that the biclusters should be according to the quantile value with default set at 0.5 (equivalent to ordering by median similarity score). The function cumBC is used to produce the cumulative probability distribution plot, as shown in Figure 6.4. This function displays the probability of observing a similarity score greater than the `refScore`, with default value set to 0.5. Based on this probability, the top biclusters are 6,7, and 4. While bicluster 8 shows desirable summary statistics, it includes individual similarity scores less than the reference score, hence a probability value of 0.

```
> cumBC(SkScores,prob=TRUE, refScore=0.5)
   BC6    BC7    BC4   BC10    BC9    BC2    BC3    BC8
0.3333 0.3333 0.2000 0.0909 0.0263 0.0261 0.0182 0.0000
```

The following code can be used to display the heatmap of similarity matrix:

```
> bcNum=c(6,7) # can be more than 2 biclusters
> heatmapBC(simMat=Sn, bicRes=bicF, bcNum=bcNum,
            main = "Structural Profiles Similarity",...)
```

The functions `ppBC` and `heatmapBC2` are used to respectively output the two plots in Figure 6.6. For `heatmapBC2`, the number of top features, N, should be specified.

```
> par(mfrow=c(2,1))
> ppBC(bicF,eMat=X, bcNum=6)
> heatmapBC2(fingerprints,bicF,bcNum=6, N=10)
```

We can use the function `plotFabia` to output the plots of loadings (`plot=1`) and scores (`plot=2`) for a given bicluster.

```
> plotFabia(resFabia, bicF, bcNum=6, plot=1)
```

Chapter 11

The `biMFA` R Package

In Chapter 9, we discussed the SMFA and biMFA methods for the integration of multi-source data. In this chapter, we present the `biMFA` R package containing the functions used to conduct the analysis presented in Chapter 9.

We consider D data matrices, $\mathbf{X}_1, \ldots, \mathbf{X}_D$ with rows as the common dimension and let $\mathbf{X}$ be the combined data matrix. This matrix is used as input to the function `biMFA`, an adaptation of Shen and Huang's `ssvd` method.

The `biMFA` package allows to perform three types of multi-source integration: MFA, SMFA or biMFA. The choice of the integration method depends on the specification of the parameters `threu` (represents the type of thresholding rule with 0 = none, 1=soft, 2=hard, and 3=SCAD for the **u** vector) and `threv` (represents the type of thresholding rule with 0 = none, 1=soft, 2=hard, and 3=SCAD for the **v** vector). MFA can be conducted using threu=0 and `threv=0`, in this case sparsity will not be applied for both loadings and scores.

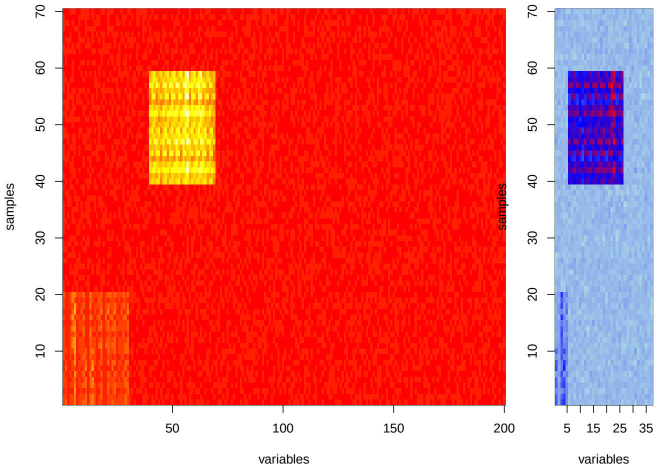
In order to conduct a SMFA we need to specify threu= 0 and `threv=1` or threu= 1 and `threv=0`. In this case, sparsity will be applied to the dimension for which the parameter is not zero. A biMFA requires a non-zero value for both threu=2 and `threv=2` for example to apply the hard thresholding rule. For example, a biMFA with two biclusters (`K=2`) and soft-threshold can be done by using the following code

```
> X <- list(X1, X2,...,XD)
> res <- biMFA(X ,K=2, threu = 1, threv = 1, gamu = 0, gamv = 0)
```

Note that `gamu` and `gamv` corresponds to γ_1 and γ_2, respectively, discussed in Section 9.3.2.

The object `res` is a list of length `K` factors containing a list of several values that are presented in Chapter 9.

The data matrices used for illustration are shown in Figure 11.1. The output is shown in the panel below.



(a) X1, a 70×200 matrix with 2 biclusters each containing 20 samples and 30 variables.

(b) X2, a 70×37 matrix, with 2 biclusters of sizes 20 samples × 5 variables and 20 samples × 21 variables.

Figure 11.1: Two datasets to integrate.

```
> str(res)
List of 2
 $ :List of 21
  ..$ u         : num [1:70] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ v         : num [1:237] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ iter      : num 3
  ..$ lambdau   : num [1:71] 0 0.000869 0.00102 0.001104 ...
  ..$ BIClambdau: num [1:70] 17200 17190 17181 17171 17162 ...
  ..$ lambdauSel: num 0.0632
  ..$ lambdav   : num [1:238] 0.00 5.64e-06 1.95e-05  ...
  ..$ BIClambdav: num [1:237] 18656 18646 18636 18627  ...
  ..$ lambdavSel: num 0.0391
  ..$ SSEudelta : num [1:70] 17.8 17.8 17.8 17.8 17.8 ...
  ..$ SSEvdelta : num [1:237] 17.8 17.8 17.8 17.8 17.8 ...
  ..$ vdelta    : num [1:237, 1:237] -0.0134 -0.0134  ...
  ..$ udelta    : num [1:70, 1:70] -0.00422 -0.00335  ...
  ..$ pevv      : num [1:237] 0.916 0.916 0.916 0.916 ...
  ..$ pevu      : num [1:70] 0.916 0.915 0.915 0.915 ...
  ..$ dfu       : num [1:70] 70 69 68 67 66 65 64 63  ...
  ..$ dfv       : num [1:237] 237 236 235 234 233 232  ...
  ..$ dfuSel    : num 22
  ..$ dfvSel    : num 55
...
 $ :List of 21
  ..$ u         : num [1:70] -0.263 -0.24 -0.182 -0.256 ...
  ..$ v         : num [1:237] -0.1377 -0.0979 -0.0422 ...
...
```

The following code can be used to plot the BIC values versus λ_v and λ_u and $df(\lambda_v)$ and $df(\lambda_u)$.

```
> k=1
> plot(res[[k]]$lambdav[-1],res[[k]]$BIClambdav,
type="p", xlab=expression((lambda[v])),
 ylab=expression(BIC(lambda[v])))
> plot(res[[k]]$dfv,res[[k]]$BIClambdav,
type="p", xlab=expression((df[v])),
 ylab=expression(BIC(lambda[v])))

> plot(res[[k]]$lambdau[-1],res[[k]]$BIClambdau,
 type="p", xlab=expression((lambda[u])),
  ylab=expression(BIC(lambda[u])))
> plot(res[[k]]$dfu,res[[k]]$BIClambdau,
type="p", xlab=expression((df[u])),
 ylab=expression(BIC(lambda[u])))
```

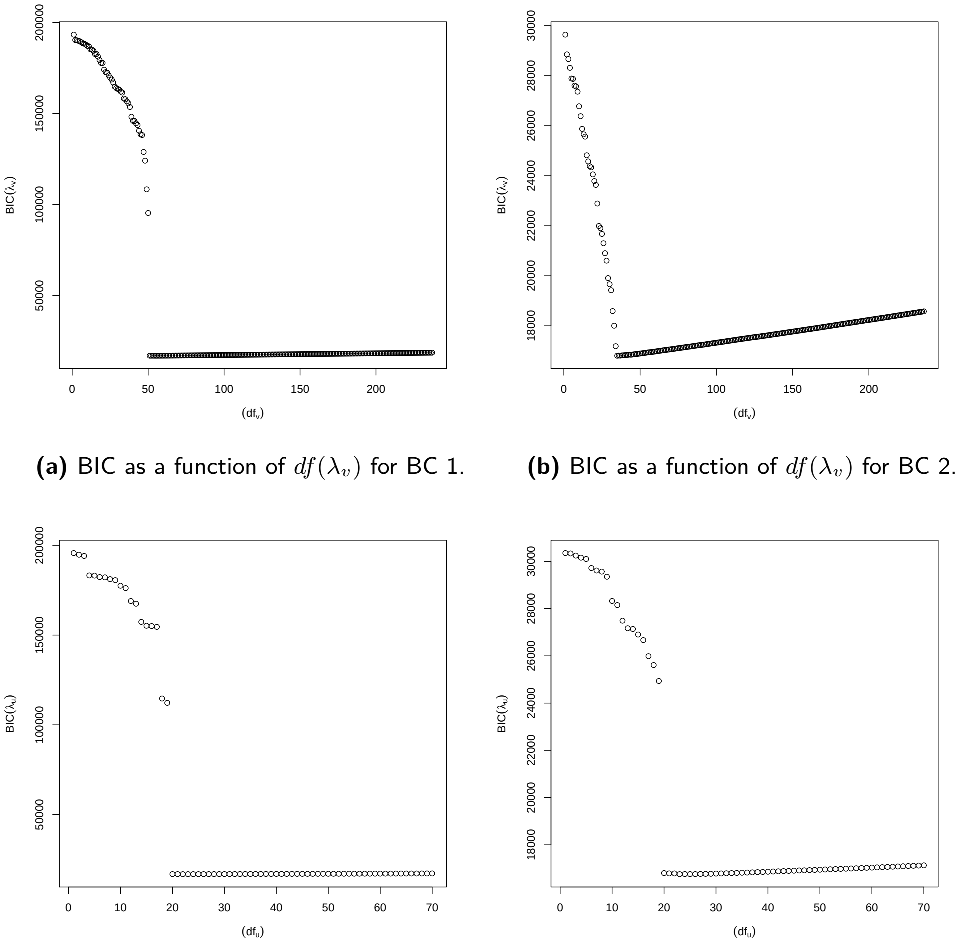
The function `getWeightedDat` can be used to get the normalized data of X1 given by

```
> X1w <- getWeightedDat(X1, scale.unit=FALSE, res=FALSE)
```

Set `res=TRUE` if the used weight and the results of the PCA is desired.

The `cvScore` function can be used to perform K-fold cross-validation as discussed in Section 9.3.1 where Q is the combined normalized data.

```
> cvScore(Q, K=10)
```



(a) BIC as a function of $df(\lambda_v)$ for BC 1.

(b) BIC as a function of $df(\lambda_v)$ for BC 2.

(c) BIC as a function of $df(\lambda_v)$ for BC 1.

(d) BIC as a function of $df(\lambda_v)$ for BC 2.

Figure 11.2: biMFA: Selection of $df(\lambda_v)$ using BIC. We have two biclusters each containing, 20 samples × 51 variables and 20 samples × 35 variables.

Chapter 12

Discussion and Future Research

The research presented in this thesis is focused on integrative data analysis methods to address the pharmaceutical research questions arising from the QSTAR approach in drug discovery and early development studies. The initiative of integrating gene expression data to the conventional structure-activity analysis in order to understand new molecules is analytically challenging, for three reasons: (1) it involves the integration of three high-dimensional datasets, (2) data analysis work flow does not exist for this framework and (3) the results should be interpretable. Various methodologies were covered: joint modeling, path analysis, penalized regression, integrative clustering, ranking of biclusters, multiple factor analysis for data integration and gene module enrichment, sparse multiple factor analysis and biclustering using MFA. For the remainder of this chapter we discuss several research lines that can be further developed based on the research presented in this thesis.

12.1 QSTAR Statistical Modeling Framework

The first method presented in the thesis is the joint modeling of bioactivity and gene expression accounting for the chemical structure of the compounds. This is a simple approach in the sense that it models one feature per data source at a time. This approach has been earlier demonstrated to be useful in identifying genetic biomarkers for efficacy. In this thesis, the model was applied on two case studies to demonstrate its utility, but in practice, this model is implemented in the pharmaceutical pipeline to different number of interesting chemical substructures, genes and biological assays (efficacy or toxicity related). The large amount of output are collated and filtered for vital information that can help the research team, especially, the medicinal chemist and biologist in taking the next step. The same idea is also reformulated in the context of path analysis modeling. Instead of modeling the association between genes and bioactivity adjusting for the structure effect, in path analysis, it is presented as a decomposition problem of the total effect of the fingerprint feature on the bioactivity via gene expression. It allows us to estimate the direct and indirect effects which is not possible with the joint modeling. Although both models are specified differently, congruent information can be retrieved as illustrated by the results of a case study.

Two main issues related to the joint model can be further investigated. The first is related to computation time. Taking into account that the joint and path analysis models presented in Chapters 2 and 3 were fitted to a specific finger print, we should expect that a complete analysis including all possible fingerprint features and bioactivity data will require relatively long computation time. Therefore a solution for this problem, like parallel programming using worker framework in a computer cluster, will allow to implement the analysis presented in Chapters 2 and 3 as a pipeline analysis for any size of discovery project.

The second issue is related directly to the "feature by feature" analysis presented in Chapters 2 and 3. An alternative approach is to construct a biomarker using all information available in the expression matrix. This can be done in several ways discussed below.

Construction of a Joint Biomarker

The gene-specific joint model can be also implemented when using a gene signature, $U(X)$, which is a latent score that represents a genetic score of the sample. Let us assume that $U(X)$ is known, in this case we can re-formulate the joint model as

$$\begin{aligned} U(X)_i &= \varphi 0 + \varphi_2 Z_i + \varepsilon_{1i}, \\ Y_i &= \phi_1 + \phi_2 Z_i + \varepsilon_{2i}. \end{aligned} \tag{12.1}$$

where $U(X) = \sum_{j=1}^{g} \eta_j X_j, \quad g \leq m$ genes, η_j is a gene-specific weights .

Similarly, the model formulated in (3.5 and 3.6) can be expressed as

$$E(Y_i|Z_i) = \delta_0 + \delta_1 Z_i, \tag{12.2}$$

and

$$E(Y_i|Z_i, U(X)_i) = \lambda_0 + \lambda_1 Z_i + \lambda_2 U(X)_i. \tag{12.3}$$

Note that $R_h^2(U(X))$ can be calculated by comparing the likelihood of models (12.2) and (12.3) following (3.7). When $U(X)$ is unknown, it can be estimated using either supervised PCA (SPCA, Bair *et al.* (2004)) or penalized regression methods such as

LASSO or elastic net.

Hence, this is a two-stage joint modeling approach involving (1) gene signature construction and (2) joint modeling of bioactivity and gene signature. For the first step, several methods can be used. Once, we have identified different classes of genes, then supervised PCA (SPCA) can be used to construct $U(X)$. On the other hand, we can use penalized regression (lasso, elastic net), random forest, among others to select genes to be summarized to $U(X)$.

Supervised PCA

The principal components (PCA) method can be used to construct a gene profile that can be used to predict a quantitative response. However, as mentioned by Bövelstad (2007), a drawback of PCA is that there is no guarantee that the principal component is associated with the response. Bair *et al.* (2004) then proposed the supervised PCA which only applies PCA on genes that are associated with the response. This supervised gene screening step reduces the dimension of the expression matrix ($\mathbf{X}$) and ensures that the resulting principal components are associated with the outcome of interest. The SPCA, therefore, relies on the underlying assumption that there is a latent variable $U(X)$ (the gene signature), which is maximally associated with the response variable Y. In this case, the candidate genetic biomarker is a latent score and not observed but can be estimated using the SPCA method. From this point, the joint model in (12.1) can be fitted as the second step to estimate the association between $U(X)$ and bioactivity accounting for the effect of a fingerprint feature.

Penalized Likelihoods

A second modeling approach that can be used to estimate $U(X)$ is the penalized regression approach. Similar to (3.2) and (3.4), the model formulated in (12.1) implies that conditioning on the gene signature, Y follows the following regression model

$$Y_i|U(X), Z = \theta_1 + \theta_2 U(X)_i + \theta_3 Z_i + \varepsilon_{3i}. \tag{12.4}$$

Expanding the model in (12.4), we get

$$Y_i|X, Z = \theta_1 + \sum_{j=1}^{g} \theta_{2j} X_{ji} + \theta_3 Z_i + \varepsilon_{3i}. \tag{12.5}$$

Here, we fit a regression model with all genes and one fingerprint feature as explanatory variables. Since $g >> n$, ordinary least squares is not feasible. However, penalized regression models that simultaneously selects genes and estimates the model parameters can be fitted. Note that for model (12.5) the penalty is only imposed on the genes but other models in which penalties are imposed on the fingerprint feature are possible as well.

12.2 Semi-supervised Integration

Integrative clustering was presented in the second part of the thesis. Clustering technique is usually applied in drug discovery to group compounds into different chemical classes. The central idea for this part of the thesis is to incorporate extra information into the clustering procedure of the compounds.

Li *et al.* (2011) proposed an approach to associate structural differences between compounds with the expression level of a defined set of genes by performing clustering on chemical structures to find differentially expressed genes between adjacent clusters of compounds from the same node. Perualila-Tan *et al.* (2015) extended this approach and instead of using only the chemical structure, proposed to use the bioactivity dataset as well. We have shown that independent clustering of these two matrices does not necessarily results in a similar compound clusters. Compounds belonging to the same structural clusters does not necessarily induce the same level of bioactivity. Hence, an integrated clustering of these two datasets is needed to provide a more meaningful characterization of compound clusters, having similar biological and structural properties.

Following the same framework, another drug discovery clustering approach presented in this thesis made use of the in silico target prediction data to cluster compounds. The target prediction score is computed based on bioactivity data and chemical structure data. Here, compounds are predicted to hit protein targets given their chemical structure.

Clustering is an unsupervised approach and the analysis flow presented in this thesis is composed of two levels. The first part is (bi)cluster analysis and the second part is the identification of features that are associated to each (or some) cluster(s). Hence, the approaches used are semi-supervised.

The choice of weights for the similarity matrices is still an open topic for research. Here, the integration was only presented using two datasets but this can be extended to more than two datasets. Other integrative clustering techniques also exist. The R package `IntClust` (Van Moerbeke *et al.* (2015)) provides some of the approaches for integrative clustering but it has not been fully explored yet for the application of drug discovery research.

Furthermore, it was shown that genes-compound cluster association where the clusters are derived using similarities of target prediction are also discovered when applying biclustering analysis on the gene expression alone. This only means that the target prediction data and the gene expression data contain similar information.

Moreover, in this thesis we proposed a way to prioritize biclusters in the presence of multiple sources of compound information. This was done by first identifying the biclusters based on one information and rank them according to similarity of the member observations based on another data source.

12.3 Unsupervised Methods

In the third part of the thesis, we discussed the use of an unsupervised approach to integrate multiple data types which can handle the QSTAR datasets. Here, we first introduced the use of MFA as an integrative method and then as a gene module enrichment teachnique. We have shown that MFA can be used as a biclustering technique by extracting a subset of variables and a subset of samples. Further, we explored the simplicity of MFA as a weighted PCA to extend this method as a biclustering technique which penalizes the loadings and/or scores to get sparse loadings and scores per component. We propose a modification of the MFA method by making use of the sparse extension of PCA, such as sparse PCA for dimension reduction and feature selection and SSVD for finding biclusters. This leads to the two unsupervised integrative methods which we term SMFA and biMFA. The approaches are new biclustering methods of multiple datasets. A thorough investigation of these techniques would be an interesting topic for further research.

12.4 Concluding Remarks

This thesis, however, is not about highlighting which method is the best to use within the QSTAR framework. The aim rather is to layout the potential of each method in generating interpretable results. Based on the results obtained by the methods discussed in this thesis, further investigation (typically experimental validation) will be carried out with respect to alterations in the chemistry and their transcriptional effects on a cell. Hence, consistency of the results across different methods would prove helpful in making concrete decisions during the lead selection to optimization phase.

Moreover, although we used various methods for data integration that led to interesting discoveries, we provided limited biological insights and structural interpretations of the identified features. We can view the output of the analyses in this thesis as a starting point for providing answers to research questions. It would be interesting to have an extensive interpretation of the chemical and biological features as a next step. In this stage, the medicinal chemists and molecular biologists are indispensable to provide feedback regarding the results. It is worth noting that statistical significance may have little biological or chemical significance. Hence, linking statistical findings back to biology and chemistry is a challenging process.

In conclusion, the QSTAR framework was developed in order to provide data analysis tool box for the decision-making in drug discovery studies. By interrelating chemistry, phenotype, and 'omics' data, functional manifestations (on-target and off-target effects) of drug actions on living cells can be explored and predicted for a set of candidate compounds in the compound optimization step. This can lead to a more efficient pipeline working procedures in which only the most promising compounds need to undergo experimental validation. A step which will reduce the development time of new drugs and reduce the overall production cost due to early detection of on-target and off-target effects.

Bibliography

Abdi, H. and Williams, L. (2010) Principal component cnalysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 433–459.

Abdi, H., Williams, L. and Valentin, D. (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Computational Statistics*, **5**, 149–79.

Acquaviva, J., Wong, R. and Charest, A. (2009) The multifaceted roles of the receptor tyrosine kinase ros in development and cancer. *Biochimica et Biophysica Acta*, **1795**, 37–52.

Albert, M. L. (2004) Death-defying immunity: do apoptotic cells influence antigen processing and presentation? *Nat Rev Immunol*, **4**, 223–231.

Alonso, A. and Molenberghs, G. (2006) Surrogate marker evaluation from an information theory perspective. *Biometrics*.

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

Bagyamani, J., Thangavel, K. and Rathipriya, R. (2013) Comparison of biological significance of biclusters of simbic and simbic+ biclustering models. *ACEEE Int. J. on Information Technology*, **3**.

Bai, J. P. F., Alekseyenko, A. V., Statnikov, A., Wang, I.-M. and Wong, P. H. (2013) Strategic applications of gene expression: from drug discovery/development to bedside. *The AAPS Journal*, **15**, 427–37.

Bécue-Bertauta, M. and Pagés, J. (2007) Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis*, **52**, 3255–3268.

Bender, A., Mussa, H. and Glen, R. (2004) Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *Journal of Chemical Information and Modeling*, **44**, 170–178.

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 (Methodological)*, **57**, 289–300.

Bol, D. and Ebner, R. (2006) Gene expression profiling in the discovery, optimization and development of novel drugs: one universal screening platform. *Pharmacogenomics*, **7**, 227–35.

Bollen, K. A. (1989) *Structural equations with latent variables.* New York: Wiley.

Brattström, D., Bergqvist, M., Hesselius, P., Larsson, A., Lamberg, K., Wernlund, J., Brodin, O. and Wagenius, G. (2002) Elevated preoperative serum levels of angiogenic cytokines correlate to larger primary tumours and poorer survival in non-small cell lung cancer patients. *Lung Cancer*, **7**, 57–63.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. and Sansone, S.-A. (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, **31**, 68–71.

Bruce, E., Autenrieth, R., Burghardt, R., Donnelly, K. and McDonald, T. (2008) Using quantitative structure-activity relationships (qsar) to predict toxic endpoints for polycyclic aromatic hydrocarbons (pah). *Journal of Toxicology and Environmental Health. Part A.*, **71**, 1073–84.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2005) *The Evaluation of Surrogate Endpoints*. Springer.

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

Byrne, B. M. (1994) *Structural equation modeling with EQS and EQS/Windows: basic concepts, applications and programming.* Thousand Oaks.

Charest, A., Wilker, E., McLaughlin, M., Lane, K., Gowda, R., Coven, S., McMahon, K., Kovach, S., Feng, Y., Yaffe, M., Jacks, T. and Housman, D. (2006) Ros fusion tyrosine kinase activates a sh2 domain-containing phosphatase-2/phosphatidylinositol 3-kinase/mammalian target of rapamycin signaling axis to form glioblastoma in mice. *Cancer Research*, **66**, 7473–81.

Chen, B., McConnell, K. J., Wale, N., Wild, D. J. and Gifford, E. M. (2011) Comparing bioassay response and similarity ensemble approaches to probing protein pharmacology. *Bioinformatics*, **27**, 3044–3049.

Chen, B., Piel, W. H., Gui, L., Bruford, E. and Monteiro, A. (2005) The HSP90 family of genes in the human genome: insights into their divergence and evolution. *Genomics*, **86**, 627–637.

Cheng, T., Wang, Y. and Bryant, S. (2010) Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics*, **26**, 2881–2888.

Chou, C. P. and Bentler, P. M. (1995) *Structural equation modeling: Concepts, issues, and applications*, chap. Estimates and tests in structural equation modeling., 37–55. Thousand Oaks, CA: Sage Publications.

Collins, M. and di Magliano, M. (2014) Kras as a key oncogene and therapeutic target in pancreatic cancer. *Frontiers in Physiology*, **4**, 407.

Consortium, G. O. (2004) The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, **32**, D258–D261.

Consortium, L. (2013) URL `http://www.lincsproject.org/`.

Crameri, A., Biondi, E., Kuehnle, K., Lütjohann, D., Thelen, K., Perga, S., Dotti, C., Nitsch, R., Ledesma, M. and Mohajeri, M. H. (2006) The role of seladin-1/dhcr24 in cholesterol biosynthesis, app processing and abeta generation in vivo. *EMBO J.*, **25**, 432–443.

Cudeck, R. (1989) Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, **105**, 317–327.

Curan, P., West, S. and J., F. (1996) The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, **1**, 16–29.

Dai, B., Yoo, S., Bartholomeusz, G., Graham, R., Majidi, M., Yan, S., Meng, J., Ji, L., Coombes, K., Minna, J., Fang, B. and Roth, J. (2013) Keap1-dependent synthetic lethality induced by akt and txnrd1 inhibitors in lung cancer. *Cancer Research*.

Davenport, E. L., Moore, H. E., Dunlop, A. S., Sharp, S. Y., Workman, P., Morgan, G. J. and Davies, F. E. (2007) Heat shock protein inhibition is associated with activation of the unfolded protein response pathway in myeloma plasma cells. *Blood*, **110**, 2641–2649. URL `http://www.ncbi.nlm.nih.gov/pubmed/17525289`.

Dearden, J. C. (2003) In silico prediction of drug toxicity. *Journal of Computer-Aided Molecular Design*, **17**, 119–127.

Drakakaki, G., Robert, S., Szatmarie, A., Brown, M., Nagawa, S., van Damme, D., Leonard, M., Yang, Z., Girke, T., Schmid, S., Russinova, E., Frimi, J., Raikhel, N. and Hicks, G. (2011) Clusters of bioactive compounds target dynamic endomembrane networks in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 17850–17855.

Dudley, J., Sirota, M., Shenoy, M., Pai, R., Roedder, S., Chiang, A., Morgan, A., Sarwal, M., Pasricha, P. and Butte, A. (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science Translational Medicine*, **3**, 96ra76.

Elrod, H. and Sun, S. (2008) PPAR? and apoptosis in cancer. *PPAR Res*.

Eriksson, S.E.and Prast-Nielsen, S., Flaberg, E., Szekely, L. and Arner, E. (2009) High levels of thioredoxin reductase 1 modulate drug-specific cytotoxic efficacy. *Free Radical Biology & Medicine*.

Escofier, B. and Pagés, J. (1983) Method for the analysis of several groups of variables. application to the characterization of red wines of the loire valley. *Journal of Applied Statistics*, **31**, 43–59. URL `http://eudml.org/doc/106148`.

Escofier, B. and Pagés, J. (1988) *Analyses factorielles simples et multiples*. Paris: Dunod.

Escofier, B. and Pagés, J. (1990) Multiple factor analysis. *Computational Statistics and Data Analysis*, **18**, 121–140.

Fisher, R. A. (1922) On the Interpretation of 2 from Contingency Tables, and the Calculation of P. *J. Roy. Statist. Soc.*, **85**, 87–94.

Fowlkes, E. and Mallows, C. L. (1983) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**, 553–569.

Franco, J., Crossa, J. and Desphande, S. (2010) Hierarchical multiple-factor analysis for classifying genotypes based on phenotypic and genetic data. *Crop Science*, **50**, 105–117.

Fryer, L. G. D., Parbu-Patel, A. and Carling, D. (2002) The Anti-diabetic drugs rosiglitazone and metformin stimulate AMP-activated protein kinase through distinct signaling pathways. *J. Biol. Chem.*, **277**, 25226–25232.

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

Gorrini, C., Harris, I. S. and Mak, T. W. (2013) Modulation of oxidative stress as an anticancer strategy. *Nature Reviews Drug Discovery*, **12**, 931–47.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. London: Chapmann & Hall.

Harris, V., Coticchia, C., Kagan, B., Ahmad, S., Wellstein, A. and Riegel, A. (2000) Induction of the angiogenic modulator fibroblast growth factor-binding protein by epidermal growth factor is mediated through both mek/erk and p38 signal transduction pathways. *The Journal of biological chemistry*, **275**, 10802–11.

Harrison, P. (1968) A method of cluster analysis and some applications. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **13**, 226–236.

Hartigan, J. (1975) *Clustering algorithms*. Wiley, New York.

Hasumi, H., Baba, M., Hong, S., Hasumi, Y., Huang, Y., Yao, M., Valera, V., Linehan, W. and Schmidt, L. (2008) Identification and characterization of a novel folliculin-interacting protein fnip2. *Gene*, **415**, 60–7.

Hauptmann, S., Siegert, A., Berger, S., Denkert, C., M., K., Ott, S., Siri, A. and Borsi, L. (2003) Regulation of cell growth and the expression of extracellular matrix proteins in colorectal adenocarcinoma: a fibroblast-tumor cell coculture model to study tumor-host interactions in vitro. *European Journal of Cell Biology*, **82**, 1–8.

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N. and A., Z. (2011) jcompoundmapper: An open source java library and command-line tool for chemical fingerprints. *Journal of Cheminformatics*, **3**, 3.

Hochreiter, S., Clevert, D.-A. and Obermayer, K. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22**, 943–949. URL `http://dx.doi.org/10.1093/bioinformatics/btl033`.

Horton, J. D., Goldstein, J. L. and Brown, M. S. (2002) Srebps: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J. Clin. Invest.*, **109**, 1125–1131.

Hoyle, R. (1995) *Structural equation modeling: Concepts, issues, and applications*, chap. The structural equation modeling approach: Basic concepts and fundamental issues., 1–15. Thousand Oaks, CA: Sage Publications, Inc.

Hu, J. (2011) *Data fusion: A first step in decision informatics.* ProQuest, UMI Dissertation Publishing.

Husson, F., Lê, S. and Pagés, J. (2011) *Exploratory Multivariate Analysis by Example Using R.* Boca Raton: CRC Press.

Iorio, F., Bosottic, R., Scacheric, E., Belcastroa, V., Mithbaokara, P., Ferrieroa, R., Murinob, L., Tagliaferrib, R., Brunetti-Pierria, N., Isacchic, A. and di Bernardoa, D.

(2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 14621–6.

Iskar, M., Zeller, G., Blattmann, P., Campillos, M., Kuhn, M., Kaminska, K. H., Runz, H., Gavin, A.-C., Pepperkok, R., van Noort, V. and Bork, P. (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.*, **9**, 662.

Jiang, Y., Gerhold, D., Holder, H., Figueroa, D., Bailey, W., Guan, P., Skopek, T., Sistare, F. and Sina, J. (2007) Diagnosis of drug-induced renal tubular toxicity using global gene expression profiles. *Journal of Translational Medicine*, **5**, 47.

Johnson, M. and Maggiora, G. (1990) *Concepts and Applications of Molecular Similarity*. New York: Wiley.

Jolliffe, I. (2002) *Principal Component Analysis*. Springer-Verlag New York Inc.

Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N. and et al. (2011) Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.

Kasim, A., Lin, D., Van Sanden, S., Clevert, D., Bijnens, L., Göhlmann, H., Amaratunga, D., Hochreiter, S., Shkedy, Z. and Talloen, W. (2010) Informative or noninformative calls for gene expression: A latent variable approach. *Statistical Applications in Genetics and Molecular Biology*, **9**.

Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S. and Talloen, W. (Eds.) (2016) *Applied Biclustering Methods for Big and High Dinemsional Data Using R*. Chapmann and Hall/CRC.

Kidane, Y., Lawrence, C. and Murali, T. (2013) The landscape of host transcriptional response programs commonly perturbed by bacterial pathogens: Towards host-oriented broad-spectrum drug targets. *PloS one*, **8**, e58553.

Klabunde, T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.

Koutsoukas, A., Lowe, R., Kalantarmotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B. O., Glen, R. C. and Bender, A. (2013) In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass naïve bayes and parzen-rosenblatt window. *J. Chem. Inf. Model.*, **53**, 1957–1966.

Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P. J., Whitmore, A. V., Zimmer, S., Young, M. P., Jenkins, J. L., Glick, M., Glen, R. C. and Bender, A. (2011) From in silico target prediction to multi-target drug design: current databases, methods and applications. *Journal of proteomics*, **74**, 2554–2574.

Koyuturk, M., Szpankowski, W. and Grama, A. (2004) Biclustering gene-feature matrices for statistically significant dense patterns. In: *IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004*, no. 480-484.

van Krieken, J., Jung, A., Kirchner, T., Carneiro, F., Seruca, R., Bosman, F., Quirke, P., Fléjou, J., Plato, H. T., de Hertogh, G., Jares, P., Langner, C., Hoefler, G., Ligtenberg, M., Tiniakos, D., Tejpar, S., Bevilacqua, G. and Ensari, A. (2008) Kras mutation testing for predicting response to anti-egfr therapy for colorectal carcinoma: proposal for an european quality assurance program. *Virchows Archiv*, **435**, 417–31.

Kuwahara, K., Sasaki, T., Kuwada, Y., Murakami, M., Yamasaki, S. and Chayama, K. (2003) Expressions of angiogenic factors in pancreatic ductal carcinoma: a correlative study with clinicopathologic parameters and patient survival. *Pancreas*, **26**, 344–9.

Lamb, J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60. URL `http://www.ncbi.nlm.nih.gov/pubmed/17186018`.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. D., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S. and Golub, T. R. (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, **313**, 1929–1935.

Lange, T. and Buhmann, J. M. (2005) Fusion of similarity data in clustering. In: *Neural Information Processing Systems*.

Lee, M., Shen, H., Huang, J. and Marron, J. S. (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 10871095.

Li, Y., Tu, K., Zheng, S., Wang, J., Li, Y., Hao, P. and Li, X. (2011) Association of feature gene expression with structural fingerprints of chemical compounds. *Journal of Bioinformatics and Computational Biology*, **9**, 503–19.

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

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. and Bijnens, L. (Eds.) (2012) *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R, Order Restricted Analysis of Microarray Data*. Springer.

Liu, W. and Johnson, D. E. (2009) Clustering and its application in multi-target prediction. *Current Opinion in Drug Discovery & Development*, **12**, 98–107.

Liu, X., Ji, S., Glänzel, W. and De Moor, B. (2013) Multi-view partitioning via tensor methods. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 1056–1069.

MacDonald, M. L., Lamerdin, J., Owens, S., Keon, B. H., Bilter, G. K., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T. and et al. (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.*, **2**, 329–337.

Magkoufopoulou, C., Claessen, S., Tsamou, M., Jennen, D., Kleinjans, J. and van Delft, J. (2012) A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. *Carcinogenesis*, **33**, 1421–9.

Mahindroo, N., Huang, C. F., Peng, Y. H., Wang, C. C., Liao, C. C., Lien, T. W., Chittimalla, S. K., Huang, W. J., Chai, C. H., Prakash, E., Chen, C. P., Hsu, T. A., Peng, C. H., Lu, I. L., Lee, L. H., Chang, Y. W., Chen, W. C., Chou, Y. C., Chen, C. T., Goparaju, C. M. V., Chen, Y. S., Lan, S. J., Yu, M. C., Chen, X., Chao, Y. S., Wu, S. Y. and Hsieh, H. P. (2005) Novel indole-based peroxisome proliferator-activated receptor agonists: design, SAR, structural biology, and biological activities. *J. Med. Chem.*, **48**, 8194–8208.

Martin, Y. C., Kofron, J. L. and Traphagen, L. M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–8. URL `http://dx.doi.org/10.1021/jm020155c`.

McNaught, A. (2006) The iupac international chemical identifier: Inchi - a new standard for molecular informatics. *Chemistry International*, **28**, 12–14.

Menniti, F., Faraci, W. and Schmidt, C. (2006) Phosphodiesterases in the cns: targets for drug development. *Nature Reviews Drug Discovery*, **5**, 660–70.

Modi, S., Stopeck, A., Linden, H., Solit, D., Chandarlapaty, S., Rosen, N., D'Andrea, G., Dickler, M., Moynahan, M. E., Sugarman, S., Ma, W., Patil, S., Norton, L., Hannah, A. L. and Hudis, C. (2011) HSP90 inhibition is effective in breast cancer: a phase II trial of tanespimycin (17-AAG) plus trastuzumab in patients with HER2-positive metastatic breast cancer progressing on trastuzumab. *Clinical Cancer Research*, **17**, 5132–5139.

Mohd Fauzi, F., Koutsoukas, A., Lowe, R., Joshi, K., Fan, T.-P., Glen, R. C. and Bender, A. (2013) Chemogenomics Approaches to Rationalizing the Mode-of-Action of Traditional Chinese and Ayurvedic Medicines. *J. Chem. Inf. Model.*, **53**, 661–673.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T. and Prachayasittikul, V. (2009) A practical overview of quantitative structure-activity relationship. *EXCLI J.*, **8**, 74–78.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **27**, 29–34.

Olah, M., Bologa, C. and Oprea, T. (2004) Strategies for compound selection. *Current Drug Discovery Technologies*, **1**, 211–20.

Pacini, C., Iorio, F., Gonves, E., Iskar, M., Klabunde, T., Bork, P. and Saez-Rodriguez, J. (2013) Dvd: An r/cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics*, **29**, 132–4.

Pal, M., Tan, M. J., Huang, R.-L., Goh, Y. Y., Wang, X. L., Tang, M. B. Y. and Tan, N. S. (2011) Angiopoietin-like 4 regulates epidermal differentiation. *PloS one*, **6**, e25377.

Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. and Hopkins, A. L. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.

Pardo, O. E., Lesay, A., Arcaro, A., Lopes, R., Ng, B. L., Warne, P. H., McNeish, I. A., Tetley, T. D., Lemoine, N. R., Mehmet, H., Seckl, M. J. and Downward2, J. (2003) Fibroblast growth factor 2-mediated translational control of iaps blocks mitochondrial release of smac/diablo and apoptosis in small cell lung cancer cells. *Molecular and Cellular Biology*, **23**, 7600–10.

Pavoine, S. and Bailly, X. (2007) New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. *BMC Evolutionary Biology*, **7**, 156.

Petitgas, P. and Poulard, J. (2009) A multivariate indicator to monitor changes in spatial patterns of age-structured fish populations. *Aquatic Living Resources*, **22**, 165–171.

Pio, G., Cece, M., DElia, D., Loglisci, C. and Malerba, D. (2013) A novel biclustering algorithm for the discovery of meaningful biological correlations between micrornas and their target genes. *BMC Bioinformatics*, **14**.

du Plessis, L., Skunca, N. and Dessimoz, C. (2011) The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings Bioinf.*, **12**, 723–735.

Polymeropoulos, M. H., Licamele, L., Volpi, S., Mack, K., Mitkus, S. N., Carstea, E. D., Getoor, L., Thompson, A. and Lavedan, C. (2009) Common effect of antipsychotics on the biosynthesis and regulation of fatty acids and cholesterol supports a key role of lipid homeostasis in schizophrenia. *Schizophr. Res.*, **108**, 134–142.

Powis, G. and Kirkpatrick, D. (2007) Thioredoxin signaling as a target for cancer therapy. *Current Opinion in Pharmacology*, **7**, 392–7.

Rand, W. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Ravindranath, A., Perualila-Tan, N., Kasim, A., Drakakis, G., Liggi, S., Brewerton, S., Mason, D., Bodkin, M., Evans, D., Bhagwat, A., Talloen, W., Göhlmann, H., Consortium, Q., Shkedy, Z. and Bender, A. (2015) Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Molecular BioSystems*, **11**, 86–96.

Reddy, A. and Zhang, C. (2013) Polypharmacology: drug discovery for the future. *Expert Review of Clinical Pharmacology.*, **6**, 41–7.

Robert, P. and Escofier, Y. (1976) A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **25**, 257–265.

Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, **50**, 742–54.

Scholtens, D. and von Heydebreck, A. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chap. Analysis of Differential Gene Expression Studies, 229–248. Springer New York.

Shaib, W., Mahajan, R. and El-Rayes, B. (2013) Markers of resistance to anti-egfr therapy in colorectal cancer. *Journal of Gastrointestinal Oncology*, **4**, 303–18.

Sill, M., Kaiser, S., Benner, A. and Kopp-Schneider, A. (2011) Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics*, **27**, 20892097.

Sirota, M., Dudley, J., J., K., Chiang, A., Morgan, A., Sweet-Cordero, A., Sage, J. and Butte, A. (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine*, **3**, 96ra77.

Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 397–420.

Smyth, G. K., Michaud, J. and Scott, H. S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

Sokal, R. and Michener, C. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**, 1409–1438.

Soudijn, W., Van Wijngaarden, I. and IJzerman, A. (2004) Allosteric modulation of g protein-coupled receptors: perspectives and recent developments. *Drug Discovery Today*, **9**, 752–8.

Spiegelman, B. M. (1998) PPAR-y : Adipogenic Regulator and Thiazolidinedione Receptor B.M. Spiegelman. *Diabetes*, **47**, 507–514.

Stanimirova, I., Walczak, B. and Massart, D. (2005) Multiple factor analysis in environmental chemistry. *Analytica Chimica Acta*, **545**, 1–12.

Szatmari, I., Pap, A., Ruhl, R., Ma, J.-X., Illarionov, P. A., Besra, G. S., Rajnavolgyi, E., Dezso, B. and Nagy, L. (2006) PPAR y controls {CD}1d expression by turning on retinoic acid synthesis in developing human dendritic cells. *J. Exp. Med.*, **203**, 2351–2362.

Takarabe, M., Kotera, M., Nishimura, Y., Goto, S. and Yamanishi, Y. (2012) Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, **28**, i611–i618.

Taldone, T., Gozman, A., Maharaj, R. and Chiosis, G. (2008) Targeting Hsp90: small-molecule inhibitors and their clinical development. *Curr. Opin. Pharmacol.*, **8**, 370–374.

Talloen, W., Clevert, D., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S. and Göhlmann, H. (2007) I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–902.

Talloen, W., Hochreiter, S., Bijnens, L., Kasim, A., Shkedy, Z., Amaratunga, D. and Göhlmann, H. (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, E173–4.

Tassi, E., Henke, R. T., Bowden, E. T., Swift, M. R., Kodack, D. P., Kuo, A. H., Maitra, A. and Wellstein, A. (2006) Expression of a fibroblast growth factorbinding protein during the development of adenocarcinoma of the pancreas and colon. *Cancer Research*, **66**, 1191–8.

de Tayrac, M., Lê, S., Aubry, M., Mosser, J. and Husson, F. (2009) Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach. *BMC Genomics*, **10**, 32.

Terstappen, G. C., Schlupen, C., Raggiaschi, R. and Gaviraghi, G. (2007) Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery*, **6**, 891–903.

Tibshirani, R. and Walther, G. (2005) Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, **14**, 511–528.

Tibshirani, R., Walther, G. and Trevor, H. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society B*, **63**, 411–423.

Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen, W., Drinkenburg, W., Gohlmann, H., Gorden, E., Bijnens, L. and Molenberghs, G. (2010) Genomic biomarkers for depression: Feature-specific and joint biomarkers. *Statistics in Biopharmaceutical Research*, **23**, 419–434.

Trabzuni, D. and Thomson, P. (2014) Analysis of gene expression data using a linear mixed model/finite mixture model approach: application to regional differences in the human brain. *Bioinformatics*, **30**, 1555–1561.

Urig, S. and Becker, K. (2006) On the potential of thioredoxin reductase inhibitors for cancer therapy. *Seminars in Cancer Biology*, **16**, 452–65.

Van Deun, K., Smilde, A., Van der Werf, M., Kiers, H. and Van Mechelen, I. (2009) A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, **10**, 246.

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

van der Veen, J., Pronk, T., van Loveren, H. and Ezendam, J. (2013) Applicability of a keratinocyte gene signature to predict skin sensitizing potential. *Toxicol In Vitro*, **27**, 314–22.

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., Consortium, Q., Shkedy, Z., Thas, O., Bender, A., Göhlmann, H. and Hochreiter, S. (2015) Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the qstar project. *Drug Discovery Today*, **20**, 505–513.

Verma, J., Khedkhar, V. and Coutinho, E. (2010) 3d-qsar in drug design-a review. *Current Topics in Medicinal Chemistry*, **10**, 95–115.

Wechsler, A., Brafman, A., Faerman, A., Björkhem, I. and Feinstein, E. (2003) Generation of viable cholesterol-free mice. *Science*, **302**, 2087.

Willett, P., Barnard, J. and Downs, G. (1998) Chemical Similarity Searching. *J. Chem. Inf. Model.*, **38**, 983–996.

Witten, D., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515 – 534.

Woodburn, J. (1999) The epidermal growth factor receptor and its inhibition in cancer therapy. *Pharmacology & Therapeutics*, **822**, 241–50.

Xu, T., Zhu, R., Liu, Q. and Cao, Z. (2012) Quantitatively integrating molecular structure and bioactivity profile evidence into drug-target relationship analysis. *BMC Bioinformatics*, **13**, 186.

Zimmermann, G., Papke, B., Ismail, S., Vartak, N., Chandra, A., Hoffmann, M., Hahn, S. A., Triola, G., Wittinghofer, A., Bastiaens, P. I. H. and Waldmann, H. (2013) Small molecule inhibition of the kraspdeδ interaction impairs oncogenic kras signalling. *Nature*, **497**, 638–42.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265–286.

Appendix A

Supplementary Results for Chapter 3

Table A1: Estimated standard errors corresponding to the estimates presented in Table 3.3.

Setting ID	se(α)	se(β)	se(ϑ_{ZY})	se(ϑ_{ZX})	se(ϑ_{XY})	se(ϑ_{ZXY})	se(Total)
A	0.12	0.14	0.14	0.12	0.06	0.18	0.14
B	0.12	0.14	0.06	0.12	0.06	0.12	0.14
C	0.12	0.14	0.14	0.12	0.06	0.18	0.14
D	0.12	0.14	0.27	0.12	0.06	0.29	0.14
E	0.12	0.14	0.40	0.12	0.06	0.41	0.14
F	0.12	0.05	0.12	0.12	0.05	0.11	0.04
G	0.12	0.05	0.04	0.12	0.05	0.00	0.04
H	0.12	0.05	0.12	0.12	0.05	0.11	0.04
I	0.12	0.05	0.22	0.12	0.05	0.21	0.04
J	0.12	0.05	0.32	0.12	0.05	0.32	0.04
K	0.19	0.16	0.47	0.18	0.11	0.45	0.16
L	0.19	0.16	0.47	0.18	0.11	0.45	0.16
M	0.19	0.16	0.47	0.18	0.11	0.45	0.16
N	0.19	0.16	0.47	0.18	0.11	0.45	0.16
O	0.19	0.16	0.47	0.18	0.11	0.45	0.16
P	0.19	0.28	0.54	0.18	0.12	0.55	0.27
Q	0.19	0.28	0.54	0.18	0.12	0.55	0.27
R	0.19	0.28	0.54	0.18	0.12	0.55	0.27
S	0.19	0.28	0.54	0.18	0.12	0.55	0.27
T	0.19	0.28	0.54	0.18	0.12	0.55	0.27

Samenvatting

Vroeg geneesmiddel onderzoek en de bijhorende ontwikkelprocessen omvatten verschillende technologiem de chemische en biologische effecten van chemische samenstellingen op een moleculair niveau te meten en vormen de basis om beslissingen te maken tijdens de ontwikkeling voor een nieuw geneesmiddel. Bijgevolg genereert dit proces meerdere bronnen aan hoogdimensionale gegevens die onder andere high throughput screening (HTS) omvatten, chemische structuren, genexpressie en image based high content screening (HCS). Een geïntegreerde analyse van deze bronnen is het centrale thema van deze scriptie. Hoogdimensionale gegevens worden gekarakteriseerd door het hebben van een enorm aantal kenmerken (variabelen) en relatief weinig chemische samenstellingen (samples). Dit brengt ons bij het probleem van data integratie en vormt een uitdagend platform voor het ontwikkelen van een methodiek en het toepassen ervan om essentiële informatie af te leiden van zowel de biologie als de chemie. Een geïntegreerde methode die toelaat om de relatie tussen al deze kenmerken in kaart te brengen kan zeer relevant zijn om het nut en de veiligheid van samenstellingen te evalueren als eventuele leidende samenstellingen dooruit de optimalisatie. In het onderzoek naar nieuwe geneesmiddelen werken wetenschappers samen om een mogelijk biomoleculair "target"te identificeren. Deze bestaat meestal uit een enkel molecule, gewoonlijk een eiwit, die betrokken is in een bepaalde ziekte en moet kunnen interageren met en beïnvloed worden door een molecuul.

Na de identificatie en validatie van het target volgt het proces om veelbelovende samenstellingen te ontdekken die uiteindelijk zouden kunnen uitmonden in een geneesmiddel voor een bepaalde ziekte. Het ontdekken begint daarom met de creatie van een nieuw molecule of het hergebruiken van een bestaand molecule. Op dit punt in het onderzoek kunnen duizenden kandidaat samenstellingen gescreend worden tegen het target voor interactie gebruik maken van HTS reeksen om het vervolgens te optimaliseren door de structuur aan te passen voor een betere interactie.

Sinds enkele decennia worden Quantitative Strcutural-Activity Relationships (QSAR) modellerings technieken (Nantasenamat *et al.*, 2009) uitgebreid gebruikt om de relatie tussen de chemische structuur en de activiteit te kwantificeren en om hierdoor meer begrip te winnen over hoe chemische substructuren invloed hebben op de biologische activiteit van een verbinding en vervolgens deze kennis te gebruiken om samenstellingen te ontwerpen met een verbeterde activiteit ofwel gerelateerd aan een grotere werkzaamheid ofwel aan een lagere toxiciteit (Dearden, 2003, Martin *et al.*, 2002, Bruce *et al.*, 2008). Het fundamentele uitgangspunt voor de QSAR aanpak is gebaseerd op de waarneming dat chemicaliën met soortgelijke structuren vaak vergelijkbare fysische eigenschappen and biological activiteiten delen.

De Quantitative Structure-Transcriptionele-Assay Relationship (QSTAR) modellerings framework is een uitbreiding van de QSAR. Hier worden transcriptie data gïntegreerd met de structurele informatie van de samenstelling alsook met experimentele biologische gegevens om de effecten van de samenstellingen in biologische systemen te analyseren vanuit verschillende hoeken om een licht te werpen op het werkingsmechanisme (Mechanism of Action, MoA) van de verbindingen. Dit kan inzicht bieden in onbedoelde fenotypische effecten die van grote waarde kunnen zijn in een vroeg stadium van de farmaceutische besluitvorming.

Hoewel de bioactieve data, die typisch gemeten wordt per target assay, belangrijk is in het optimalisatieproces van het chemisch ontwerpen van samenstelling, biedt het niet veel waardevol inzicht over de onderliggende biologische mechanismen. In contrast met de biologische gegevens die enkel biologische effecten beschrijven, is genexpressie data, als een multidimensionale assay, in staat om informatie te geven over een brede verscheidenheid van biologische effecten van een samenstelling op het transcriptionele niveau van het gehele genoom, en geeft daarmee een informatie-rijke snapshot van de biologische toestand van een cel (Gölmann en Talloen, 2009, Amaratunga *et al.*, 2014). Transcriptomische veranderingen die optreden na toediening van een samenstelling kunnen ook worden gemeten in high throughput, waardoor screening van veel stoffen in meerdere cellijnen tegen een lage kost kan gebeuren. Ook is het geobserveerd dat transcriptie data veelal biologisch relevante signalen detecteren en beter in staat is om samenstellingen te prioriseren dan de conventionele target assays (Verbist *et al.*, 2015). Toepassingen die met behulp van genexpressie profielen meerdere genen en biologische reactiepaden tegelijkertijd observeren verrijken het inzicht in de onderliggende mechanismen. Binnen het QSTAR kader, kunnen mRNA biomarkers ontdekt worden door samenstelling die ziektegerelateerde variatie van de genexpressie veroorzaken. Analyse van de transcriptie profielen maakt het mogelijk om nieuwe biomarkers gerelateerd aan bepaalde biologische effecten veroorzaakt door deze samenstellingen te identificeren. Met deze aanpak kan een aanzienlijke hoeveelheid middelen worden uitgespaard vanwege de vroege identificatie van gevaren en zo fouten te vermijden in de latere stadium van de farmaceutische ontwikkeling van geneesmiddelen.

Dit proefschrift bestaat uit een aantal analyse workflows om de drie hoog-dimensionale datatypes te integreren; gen-expressie, fingerprint eigenschappen (FFS die de chemische structuren voorstellen) en bioassay data (fenotype). De methoden in dit proefschrift zijn verdeeld in drie soorten: het QSTAR modelkader, semi-supervised methoden, van een clustering tot een biclustering analyse en unsupervised multivariate methoden voor data exploratie en integratie. Het laatste deel van het proefschrift behandelt de statistische software ontwikkeld samen met de methoden.

Het eerste deel van het proefschrift is toegewijd aan statische modellen die toepasselijk zijn in de context van QSTAR. **Hoofstuk 2** omvat het kader van gezamenlijk modelleren (joint modeling framework) wat ons toestaat om (1) actieve genhandtekeningen die de chemie sturen te identificeren, (2) chemische substructuren (oftewel ‘fingerprint features‘, FF) van samenstellingen te bepalen die gerelateerd zijn met de effecten op bio-assay data voor specifieke "targets"van interesse en (3) om na te gaan of dit effect ook kan bevestigd worden door veranderingen in genexpressies (zowel on- of off-target gerelateerd). **Hoofdstuk 3** beschrijft de relatie tussen gezamenlijk modelleren, "path analysis"modelleren, en voorwaardelijk modelleren.

Het tweede deel van het proefschrift bevat de sequentiële integratie van meerdere datasets met het doel het werkingsmechanisme te verklaren van een deelgroep van samenstellingen (chemische stoffen) met behulp van clustering en biclustering technieken (Kasim *et al.*, 2016). Clustering algoritmes gebruiken de similariteit data om objecten te groeperen en worden voornamelijk gebruikt op één databron. In **Hoofdstuk 4** wordt een clustering oplossing voorgelegd die meerdere databronnen kan hanteren in de context van ontwikkeling van nieuwe geneesmiddelen. Een typische strategie voor de selectieprocedure van chemische stoffen omvat het clusteren van deze stoffen gebaseerd op hun chemische structuur. Dit idee wordt verder uitgebreid met een geïntegreerde clustering benadering die gebruik maakt van beide databronnen met het oog op de ontdekking van een deelgroep van chemische stoffen met gelijkaardige structuur- en biologische eigenschappen. Deze methode voegt bioactiviteit en structuur gebaseerde similariteitsmatrices met behulp van complementaire gewichten samen waardoor er een gewogen similariteitsmatrix, de standaard invoer in elk clustering algoritme, wordt gevormd. Hierna wordt een tweede analyse uitgevoerd waarin elke biologische en structuur gestuurde cluster van chemische stoffen verder gelinkt wordt aan een set van transcriptoom kenmerken.

Een nieuwe deelgroep van chemische samenstellingen die op vlak van structuur en biologische eigenschappen gelijken op de referentie stof worden zo ontdekt met de voorgestelde clustering benadering. **Hoofdstuk 5** behandelt het integreren van genexpressie profielen van specifieke proteïnen met als doel onze kennis van de fundamentele mechanismen in proteïne-ligand bindingen te verrijken. Dit hoofdstuk focust op de integratie van genexpressie data en in-silico target predictie scores, om inzicht te verschaffen over het werkingsmechanisme (Mechanism of Action, MoA). De chemische samenstellingen worden geclusterd op basis van de similariteit van hun voorspelde proteïne targets waarna elke cluster gelinkt wordt aan "gene sets"door middel van Lineaire Modellen voor Microarray Data. Pathway analysis wordt vervolgens gebruikt om de gene sets te identificeren op basis van hun biologische processen. Verder wordt er een kwalitatief onderzoek uitgevoerd op the homogene clusters van de chemische stoffen gebaseerd op hun targets om de pathways te identificeren. **Hoofdstuk 6** stelt een workflow voor om de genexpressie biclusters te ranken met behulp van een andere bron van informatie wat in dit geval de chemische structuur zal zijn.

Het derde deel van het proefschrift bestaat uit 3 hoofdstukken, beginnende met **Hoofdstuk 7** waarin de Multiple Factor Analysis (MFA) voor normalisatie en integratie van datasets wordt geïntroduceerd. Voor deze analyse zullen de 3 QSTAR datasets gebruikt worden. **Hoofdstuk 8** illustreert het gebruik van MFA als een gen-module verrijkingstechniek. In **Hoofdstuk 9** worden 2 varianten van MFA gepresenteerd, namelijk SMFA and biMFA. In dit hoofdstuk wordt de genexpressie (GE) data geïntegreerd met "high content screening"(HCS) data. Het doel hiervan is de transcriptionele effecten van de chemische samenstellingen te relateren met bioactiviteitsmetingen in een cel na toediening gebaseerd op beelden. Deze idenficatie van fenotypische subklasses (GE en HCS) die co-gereguleerd zijn over een deelgroep van chemische stoffen, kan worden toegepast als een biologisch screening tool om het potentieel voor doeltreffendheid en toxiciteit van deze stoffen te schatten. Dit is in lijn met het basis concept van biclustering, rekening houdende met meerdere databronnen. Vandaar, met het oog op deze doelstelling, zijn Sparse Multiple Factor Analysis (SMFA) en biclustering met MFA (biMFA) ontwikkeld om simultaan te zoeken naar associatie tussen kenmerken en chemische stoffen. Deze integratie methodes combineren de ideeën van MFA en singulierewaardeontbinding technieken met een strafterm. De resultaten brengen een groep van potentiële geno-toxische leidende stoffen en een Tubulin-gelinkte groep van stoffen, samen met hun respectievelijke HCS kenmerken indicators aan het licht.

De laatste 2 hoofdstukken van het proefschrift bevatten de ontwikkelde R producten voor de voorgestelde methodologie in de verhandeling. Het eerste R pakket biclustRank, wordt behandeld in **Hoofdstuk 10** en **Hoofdstuk 11** bespreek het R pakket biMFA welke ontwikkeld is voor de methodologie in het derde deel van het proefschrift.