2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Exploring local patterns between gene expression profiles and chemical structures (fingerprints) of compounds

Promotor :
Prof. dr. Ziv SHKEDY

Supervisor :
Mevrouw Nolen Joy PERUALILA

Emmanuel Abatih

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2014•2015
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Exploring local patterns between gene expression profiles and chemical structures (fingerprints) of compounds

Promotor :
Prof. dr. Ziv SHKEDY

Supervisor :
Mevrouw Nolen Joy PERUALILA

Emmanuel Abatih
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

ACKNOWLEDGEMENTS

I only wanted to acquire some skills on how to analyse genomic data to satisfy consultancy needs at the Unit of Epidemiology and Biostatistics at the Institute of Tropical Medicine. I started with just one course, and then later a few and finally before I knew, I did all the modules necessary to permit me present this thesis. What a journey it has been!

I wish to sincerely thank Prof Dr. Dirk Berkvens for encouraging me to follow this program and for covering all expenses related to this undertaking. When I think of the sacrifices you were ready to make to let me achieve this goal, I studied even harder. Many thanks for taking up all my responsibilities while I was away.

I am indebted to my Supervisor Prof Dr. Ziv Shkedy for agreeing to supervise my thesis even under such short notice. Thanks also for the guidance the flexibility and for sharing so much about analysis of gene expressions. I wish to equally thank Nolen Joy Perualila (PhD student) for her kind assistance and for the many new ideas towards the realization of this work.

Furthermore, I wish to thank all my class mates especially Ewoud, Marijke, Susan, Peter and Bernard for the encouragements. I wish to also express my gratitude to my colleagues at the department of Biomedical sciences especially Joule, Nele, Kim, Lynn and Lilian for listening when I needed to complain about how much work I had to do and for all the encouragements.

I wish to thank all my very good friends who still don't understand why I decided to take up another Masters program. Whenever I said I had classes, the question was always : "you going to give lessons or to take lessons (with a grimace)". Well, when it comes to knowledge, I don't mind being greedy. Thanks to my naughty cousin Elvis for expressing lots of surprises when ever I claimed not to understand something about Bioinformatics and for laughing out loud when I failed in some courses.

Finally, my warmest appreciation goes to my ever so wonderful wife Carinne Joelle and "my boy" Nathan for sustaining the greatest consequences associated with my absence. Without you the journey would have been meaningless.

IN LOVING MEMORY OF MY MUM: *JOHANNA LUM*

(7th APRIL 2014)

SUMMARY:

The availability of high throughput technologies such as microarrays and next generation sequencing have made it possible to cheaply collect large amounts of drug-gene expression data sets. Combining compounds and their characteristics with gene expression data is called connectivity mapping and holds promise for in-depth analysis and understanding of biological processes, discovery of new drug targets and new drugs and prediction of toxic potential of unknown compounds. These goals can be achieved using the connectivity map data base and using appropriate methods.

For studying relationships between gene expression profiles of human cells following the introduction of chemical compounds and the fingerprints of the compounds, the recently developed Multiple factor analysis (MFA) which seeks patterns in data consisting of quantitative as well as qualitative variables can be applied. The results of the MFA can often be made more robust by applying hierarchical clustering analysis. In addition, ignoring the gene expression profiles and working only with fingerprints of compounds, it was determined whether groups of compounds are associated with groups of fingerprints using five different methods: Multiple correspondence analysis (MCA), Binary inclusion-maximal biclustering (Bimax) algorithm, Factor analysis for Bi-cluster acquisition (FABIA), Iterative Binary biclustering of gene sets and Factor analysis for binary data. These biclustering approaches simultaneously cluster rows and columns of the data matrix.

The performance of the biclcustering methods applied in this study appeared to be similar except for BiMax and FABIA. MCA, IBBiGS and Factor analysis for binary data yielded very similarly results on most of the biclusters. Given the different approaches to biclustering the methods all identified the compounds: 4,5-dianilinophthalimide, N-phenylanthranilic acid, flufenamic acid, phenyl and biguanide except for BiMax which only identified 4,5-dianilinophthalimide, N-phenylanthranilic acid and flufenamic acid. In addition, these compounds were found to exhibit the fingerprints: FP47, FP105, FP140 and FP215. These were consistently present in these compounds across the different methods except for BiMax where none of the fingerprints featured. The poor performance of BiMax could be attributed to the sparsity in the data whereas FABIA is not intended for binary data and should be cautiously interpreted.

In conclusion, for exploring local patterns, no one method could be judged superior over the others as evidenced in the literature. However, for sparse binary data like the one we presented in this study, a combination of the results from the three methods: factor analysis for binary data, multiple correspondence analysis (MCA) and Iterative Binary biclustering of gene sets (IBBiGs) will be the most optimal approach as they appear to be robust especially to sparse binary data. In addition, for a combination of groups of variables (quantitative and qualitative), the multiple factor analysis (MFA) combined with hierarchical clustering should be used. Finally the compounds 4,5-dianilinophthalimide, N-phenylanthranilic acid, flufenamic acid, phenyl and biguanide with their fingerprints FP47, FP105, FP140, FP215 should be further investigated.

Contents

List of Tables

List of Figures

1 Introduction

The most recent advances in biotechnology have lead to the generation of huge amounts of data such as drug-treatment gene expression data which could be used to understand biological processes, discover new targets and new drugs and predict toxic potential of unknown compounds. In modern drug discovery pipelines, the data collected, includes three important sources:

1. Chemical properties of the compounds being investigated (Fingerprints)
2. Bio-assay data for targets of interest (targets), and more recently
3. Micro-array gene expression data.

These data sets are hosted within the Connectivity Map (CMap) database (Lamb et al., 2006) which provides the basis for a data-driven study of drug-effect relationships at a genome- wide level. In effect, CMap is host to the largest collection of high-dimensional gene expression profiles derived from treatment of three different cultured human cells with over one thousand bioactive small molecules (Lamb et al., 2006). The idea is that any perturbation to the gene expression profiles can be summarized by a gene signature. These gene signatures are often obtained using microarray technologies and used as proxies of disease phenotypes and drug effects. The matching of different diseases to chemical compounds based on gene signatures is known as connectivity mapping (Parkkinen and Kaski, 2014).

Connectivity mapping has been used in several different studies revealing new biological links between drugs and between drugs and diseases. In addition, genome-wide gene expression responses from the CMap have been used to discover clusters of drugs having similar mechanisms of action in turn resulting in novel findings, such as effects of heat shock protein inhibitors and identification of modulators of autophagy (Iorio et al., 2010). Finally the CMap data have also been successfully used in large scale integrative studies including the analysis of regulation of drug targets (Iskar et al., 2010) and interactions of drugs with protein networks (Laenen et al., 2013).

A particular challenge is the separate or integrated statistical analysis of these multiple data sources from the CMap database in order to uncover local patterns. To this regard, several methods have been used. The particular methods to be used depend on the particular data type to be analyzed and on the goals of the analysis.

A traditional, popular and very successful method that has been used to discover patterns in data of such nature is clustering. Most clustering algorithms such as the unsupervised hierarchical clustering (EisenMB et al. 1998) or partitioning methods such as Partitioning around Medoids (PAM) (Kaufman,1990) try to group data (compounds) into classes in which within-group similarity is maximized, or for which between-group similarity is minimized, all based on a well defined similarity measure. In addition, most of these clustering methods are one-dimensional: they can only cluster the rows or columns of a matrix separately or for one then followed by the other (Eren et al. 2012). This aspect imposes a limitation to traditional clustering methods on microarray data especially in drug design, where researchers want to reveal how compounds affect gene expression. For example many chemical structures or drug targets are common to a given set of compounds and vice versa. In addition, certain genes can

be co-expressed only under the influence of a given set of chemical compounds. These same genes might be expected to behave differently in the presence of other sets of compounds. The discovery of local patterns of this nature holds promise for mining and possibly discovering new pathways or biomarkers.

A recent and more complete and promising approach is called biclustering which allows for the simultaneous clustering of both rows and columns in the data matrix. Thus whereas clustering seeks for global patterns, biclustering is after local patterns. The idea of biclustering was first introduced to gene expression analysis by Cheng and Churh (2000). Following this, several algorithms have been proposed for finding biclusters in data tables.

Such clusters are biologically interesting since they not only allow us to capture the correlated genes, but also enable the identification of genes that do not behave similar in all conditions (Eren et al. 2012). Hence, biclustering is more likely to yield the discovery of biological clusters that a clustering algorithm might fail to recover.

A rather different approach to biclustering is to find clusters of compounds for example and then determine genes and chemical structures that are associated with these clusters. This is the case with the recently redeveloped approach that can handle different data types such as gene expressions and fingerprints called Multiple Factor Analysis (MFA) (Escofier, B. and Pagès, J. ,1990; Bécue-Bertaut M and Pagès J, 2008). This method enables us to study a combination of quantitative and qualitative variables in order to determine the association between chemical structure of compounds and gene expression profiles obtained on the same compounds. This has the tendency to increase the power for detecting compound fingerprint features that are associated with groups of compounds which in turn up or down-regulate groups of genes. MFA also provides a balanced group representation that makes it possible to identify specific and common structures.

Motivated by an appetite for local patterns in the CMap database, this thesis aimed at applying different data reduction, clustering and visualisation techniques to identify local patterns.

The main objectives of this study were to identify groups of compounds with defined fingerprints that co-regulate groups of genes. More specifically,

- the gene expression profiles of cells following the introduction of chemical compounds with defined fingerprints will be explored to determine which compounds with which chemical structures co-regulate which genes.
- In addition using only the fingerprint matrix, several biclustering methods will be used to determine if they identify similar groups of compounds associated with similar fingerprint structures.
- Account for sparseness in the binary data matrix

2 Material and Methods

2.1 Data description and pre-processing

2.1.1 The CMAP database

The CMAP database supplied is a subset of the larger cmap database and consisted of a gene expression matrix with 2434 genes, a target matrix with 477 targets and a fingerprint matrix with 250 chemical structures on 56 chemical compounds. A target was coded on the 0-1 binary scale where 1 reflected affinity of the chemical compound to the target and zero for lack of affinity of the compound to the target. We note that many compounds can have affinity to the same target. Since the names of these targets were too long, an alternative renaming scheme was applied so that the new names were bio1 up until bio477 to represent the 477 targets. On the other hand, a fingerprint was coded on a 0-1 binary scale to give the representation of the molecular structure of the compounds where 1 coded for the presence of the sub-structure (referred to as a fingerprint structure which can be an atom, or an atom with several bonds) and 0 to code for the absence of the fingerprint feature. In a similar manner, since the names of the fingerprints were rather too long, we renamed to FP1 to FP250 to represent the 250 fingerprints.

Table 1: Gene and fingerprint matrix showing the gene expression levels and the molecular structure of the compounds where 1 codes for the presence of the sub-structure and 0 otherwise

	LOC100129361	PDCD6IP	SH2B3	SAE1	FP2	FP3	FP4	...
metformin	-0.07	-0.06	-0.06	-0.10	0.00	0.00	0.00	...
phenformin	-0.02	-0.06	-0.01	-0.04	0.00	0.00	0.00	...
phenyl biguanide	0.06	-0.02	0.01	0.02	0.00	0.00	0.00	...
estradiol	-0.00	0.02	-0.02	-0.00	0.00	0.00	0.00	...
dexamethasone	-0.04	-0.06	-0.06	-0.03	0.00	0.00	0.00	...
verapamil	-0.09	-0.11	0.27	-0.45	0.00	0.00	0.00	...
exemestane	-0.04	-0.12	0.19	-0.10	0.00	0.00	0.00	...
rofecoxib	-0.00	-0.04	0.01	0.02	0.00	0.00	0.00	...
amitriptyline	-0.12	-0.08	0.06	-0.07	0.00	0.00	0.00	...
15-delta prostaglandin J2	0.07	-0.02	0.13	-0.21	0.00	0.00	0.00	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

Table 1 is a juxtaposed table containing in its first 2434 columns the genes and in the next 250 columns, the fingerprints. Each row on the table represents a chemical compound.

Table 2 represents the fingerprint matrix for the 56 compounds under consideration. Since these are all binary matrices, the same biclustering methods were applied to them.

In this study, focus will be on Tables 1 and 2. The motivation for studying Table 1 is to explore the MFA method which combines quantitative and qualitative data. On the other hand, for Table 2, the motivation was to compare several methods for biclustering binary data.

2.1.2 Gene filtering and processing of the fingerprint matrix

Filtering was applied to reduce the number of genes prior to the MFA analysis. The idea of the filtering was to remove genes where the intensities are consistently very low across all the compounds (say

Table 2: Fingerprint matrix showing the representation of the molecular structure of the compounds where 1 coded for the presence of the sub-structure and 0 otherwise

Compound	FP1	FP2	FP3	FP4	FP5	...
metformin	0.00	0.00	0.00	0.00	0.00	...
phenformin	0.00	0.00	0.00	0.00	0.00	...
phenyl biguanide	0.00	0.00	0.00	0.00	0.00	...
estradiol	0.00	0.00	0.00	0.00	0.00	...
dexamethasone	0.00	0.00	0.00	0.00	0.00	...
verapamil	0.00	0.00	0.00	0.00	0.00	...
exemestane	0.00	0.00	0.00	0.00	1.00	...
rofecoxib	0.00	0.00	0.00	0.00	0.00	...
amitriptyline	0.00	0.00	0.00	0.00	0.00	...
15-delta prostaglandin J2	0.00	0.00	0.00	0.00	0.00	...
⋮	⋮	⋮	⋮	⋮	⋮	...

more than 25% of the compounds). A filtering based on variance and intensity was applied using the **filterVarInt()** function under the **a4base** package. Typically an inter-quartile range (IQR) filtering was applied with the option to cut off genes for which their intensity scores have IQR of less than 0.05 on the log base 2 scale. Here the normal IQR range of 0.5 was modified to 0.05 because at 0.5 no genes were retained since the intensities were very low. In fact, the IQRs per gene for our data set were found to be around 0.05.

Many of the fingerprints had very low frequencies for a given category or even the same value i.e values were either only 0 or 1, so there was no variability in these variables. The MFA method does not work when there is a categorical variable which has only one category because this causes what is known as "separation" in the data matrix. Therefore, all fingerprint variables with only one level for the response (1 or 0) were removed from the data base.

2.2 Methodology

In this subsection, the various methods that were applied across the two different data tables are briefly described. Additional analysis steps that were performed to render the methods more comparable across the different data tables are also discussed.

2.2.1 Multiple Factor Analysis

MFA is a multivariate ordination method that permits examination of common structures in data sets with many variables that can be separated into different groups (Escofier and Pagès 1990) such as gene expression profiles and fingerprints for chemical compounds. In other words, MFA is a principal component/Multiple correspondence analysis method which allows to explore and visualise blocks of data like those of Table 1 where the chemical compounds are described by their chemical structures and gene expression profiles. The aim of the MFA is to study the similarities between compounds as viewed by their chemical structures and also their corresponding gene expression profiles. In addition, MFA also studies the links between groups of variables and compares the information brought by each group.

MFA proceeds in three steps:

- First it performs a principal component analysis (PCA) on the gene expression matrix, and a multiple correspondence analysis (MCA) on the fingerprint matrix separately and "normalizes" each data table by dividing all its elements by the square root of the first eigenvalue obtained from its PCA. This brings out the information that is common to each data table. The different data tables are then juxtaposed as shown on Figure 1.

- Secondly, all the normalized data tables are aggregated into a grand data table that is analyzed via a global (non-normalized) PCA that gives a set of factor scores for the compounds and loadings for the different groups of variables. The global PCA amounts to computing a Singular Value Decomposition (SVD) of the grand data table.

- Finally the individual datasets are then projected onto the global analysis. In this way, variables in each group are permitted to maintain free covariances amongst themselves, and the relationships between groups of variables can be examined without the influence of within-group covariance.

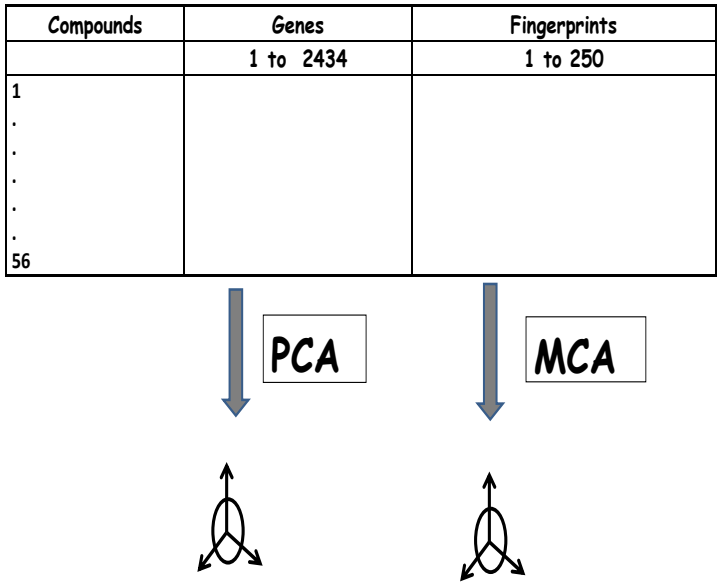

Figure 1: Juxtaposed matrix showing the quantitative (gene expression) and qualitative (fingerprint) variables for the 56 compounds

MFA provides for each data table a set of partial factor scores for the observations that reflect the specific "view-point" of this data table. Interestingly, the common factor scores could be obtained by replacing the original normalized data tables by the normalized factor scores obtained from the PCA/MCA of each of these tables.

MFA was implemented using the mfa() function in the **FactoMineR** package in R (R Core Team, 2014) which takes as input, the data matrix in Table 1 with the fingerprints converted into characters. Following the MFA, clustering was performed to take into account most of the information considered

as relevant in the separate analyses (Bécue-Bertaut & Pages, 2007). In fact, as suggested by Lebart (1994), MFA and MCA can be used as a pre-processing step for hierarchical cluster analysis.

The **FactoMineR** package contains a hierarchical clustering analysis function called **HCPC** (Hierarchical Clustering on Principal Components) that performs clustering to visualize and highlight the main features of a dataset. It was used to draw a hierarchical cluster tree, select the clusters of compounds and provide a description of the clusters as viewed by the genes and by the fingerprints.

2.2.2 Factor Analysis for Bicluster Acquisition (FABIA)

Factor Analysis for Bicluster Acquisition (FABIA) (Hochreiter et al., 2010) is a biclustering method that biclusters gene expression profiles or fingerprints and compounds using Factor analysis. It is based on a multiplicative model, which accounts for linear dependencies between gene expression or fingerprints and compounds, and also captures heavy-tailed distributions as observed in real-world gene expression data. In this framework, two vectors are similar if one is a multiple of the other and the angle between them is zero. The algorithm selects the model parameters that best explain the data using a variational expectation maximization (EM) algorithm (Hochreiter et al., 2006; Talloen et al., 2007).

The biclusters are found by sparse factor analysis where both the factors and the loadings are sparse (Vectors containing many zeros or values close to zero are called sparse vectors) and ranked based on their mutual information content. Weaker members of a bicluster are optionally pruned with a threshold.

This method was applied to Table 2. It is worth mentioning that FABIA was developed for continuous data even though it is being used here on the binary (0,1) data. This was motivated by the fact the method has been claimed to work well on binary data (Personal communication with Prof. Ziv Shkedy) so we wanted to evaluate once more its performance.

Biclusters were generated using the **fabia()** function in the **fabia** package in R Bioconductor.

2.2.3 Binary inclusion-maximal biclustering algorithm: BiMax

Binary inclusion-maximal biclustering algorithm (BiMax) is a simple reference algorithm that seeks biclusters of 1's in a binary matrix. It was introduced as a reference algorithm for comparing other methods by Prelic et al. (2006). It biclusters the data matrix using a divide-and-conquer approach, recursively dividing it into a checker board format. The algorithm works only on binary data and thus is only applicable to Table 2.

Essentially, considering Table 2=M as an example with rows from 1 to 56 and labelled Cpd1,...,Cpd56 and columns from 1:250 denoted FP=(FP1,...,FP250), the biclustering according to BiMax proceeds as follows:

1. The algorithm starts by randomly choosing any row labelled Cpd* containing a mixture of 0's and 1's. If the chosen row has only 1's then it is a bicluster otherwise there is no bicluster
2. The selected row (Cpd*) is used to cleave the matrix into two submatrices each of which can be separately analyzed in turn.

3. The submatrices are found by diving the FP1,..,FP250 columns into two different groups: those for which row Cp* is 1 and those for which it is 0.
    - $FP_a = \{fp : M[Cpd*, fp] = 1\}$
    - $FP_b = FP - FP_a$
4. Next, the 56 rows of M are split into three sets
    - Cpd_a : Rows with 1's only in FP_a
    - Cpd_b : Rows with 1's only in FP_b
    - Cpd_c : Rows with 1's in both FP_a and FP_b
5. The submatrix formed by (Cpd_a, FP_b) is empty and cannot contain any biclusters.
6. The submatrix U $(Cpd_a \cup Cpd_c, Cpd_a)$ and the submatrix V=$(Cpd_b \cup Cpd_c, FP_a \cup FP_b)$ contain all possible biclusters in M
7. The procedure continues by recursively processing U and V and ends if the current matrix represents a bicluster (contains only 1s).

The biclust package (Kaiser and Leisch, 2008) consists of biclustering algorithms including Bimax. To run BiMax, the exact size of the expected biclusters has to be provided, because otherwise it would halt prematurely, recovering only a small portion of the expected biclusters.

A bicluster in this study based on Table 2 was considered as a cluster of compounds with a strong co-expression (association) across given fingerprints. The differential co-expression score for each bicluster SB() was used to rank the biclusters. The score was computes using Chia and Karuturi Function (Chia and Karuturi, 2010). The function computes row (T) and column (B) effects for a chosen bicluster. The scores for Strong positive SB scores indicated strong co-expression in the bicluster and weaker or no co-expression out of the bicluster and vice versa.

2.2.4 The Iterative Binary Biclustering of Genesets (iBBiG)

The iBBiG algorithm (Gusenleitner et al. 2012) identifies bi-clusters (or modules) in a matrix of binary data. iBBiG is optimized for discovering clusters in sparse noisy binary genomics data. iBBiG uses an iterative approach which enables it to discover weak signals, even if they are potentially masked by stronger ones. An advantage of iBBiG relative to other methods is that it does not require a priori knowledge of the true number of clusters. Although iBBiG includes several parameters, it has been shown that most impact is only on computation time and not on effective cluster discovery.

When applied to Table 2, iBBiG will extract clusters or 'modules' of groups of compounds with similar fingerprints. The clusters of compounds from iBBiG are ranked by an information score, and within each cluster, compounds are ranked by a fitness score that measures its weight in the cluster. It uses a genetic algorithm to maximize the size and entropy of each bi-cluster producing a small number of bi-clusters. Finds overlapping clusters and has been shown to perform better than FABIA, bimax

The number of true clusters can be estimated from the weighted cluster scores and RowScorexNumber of the extracted clusters. RowScorexNumber gives the score for each compound (row) in each cluster and can be used to select the top compounds in each cluster. A threshold of importance for each compound was set as 5. Thus compounds with RowScorexNumber$\geq$5 were considered as contributing highly to the cluster in question. In addition, NumberxCol gives the number of fingerprints (columns) in each module or cluster. The score for each module, the fingerprints and the compounds linked to the cluster were presented in a table.

The iBBiG algorithm was applied using the iBBiG() function in the **iBBiG** package in R.

2.2.5 Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA)(Greenacre, 1984) is an exploratory technique to identify and visualize the relation(s) between fingerprints and compounds in Table 2.

A useful information provided by a MCA is that the original table is reduced to latent dimensions. These dimensions are computed according to their contribution to the global χ^2-statistic for the complete data matrix. Only a limited number of latent dimensions are used for interpretation based on the inertia they explain. Next, the fingerprints and compounds values are assigned a position with respect to these latent dimensions which are the axes of two-dimensional (or three-dimensional) plots. In those plots, variable values assigned to the same quadrant are associated. Taking into account our aim, viz. explaining the origin of four instances of confounding, the relative positions of the variables values rather than the interpretation of the latent dimensions are our main concern. The MCA was performed by means of the of the **MCA()** function in the **FactoMineR** library in R. The input for the MCA is Table 2 and analysis is based on an indicator matrix created from the table.

Methodologically, the interest in using MCA on the CMap data for uncovering local patterns can be summarized as follows: First, it allows identifying the association between the different compounds and their fingerprints. In addition, it allows creating a typology of the compounds based on the fingerprints, Finally, the MCA points out which compounds are more similar amongst themselves, and what types of fingerprints are predominant in their characterization.

For MFA and MCA, to search for variables that characterise each of the dimensions, the function **dimdesc()** was used to observe which variables are highly correlated to a certain principal component. It returns not only the correlation coefficients, but also performs a test if the variables are significant. For each dimension for all genes with a p-value smaller than 0.05 the results are returned. For our study we selected only genes that had an absolute correlation coefficient greater than 0.65 with each principal component. For compounds, a useful heuristic is to base the interpretation of each dimension on the compounds that have contributions larger than the average contribution.

For qualitative variables, a one-way ANOVA model is used where the response is each time the principal component (dim or factor) and the categorical explanatory variable is the fingerprint or target. A list is returned with the p-values based on an F-test for the overall significance of each variable. The list is sorted so that at the top, we have the most characteristic variables and the least characteristic are at the bottom.

2.2.6 Exploratory Factor analysis for binary data

Factor analysis (Bartholomew ,1980) is a method to find similarities between variables and use these similarities to group the variables into a smaller amount of factors. Factor analysis uses a correlation matrix to compare if variables are similar to each other or not.

Standard methods of performing factor analysis assume that the variables are continuous and follow a multivariate normal distribution. When the data includes variables that are binary such as those of Table 2 a factor analysis can be performed using a polychoric correlation matrix.

The polychoric correlation matrix was computed using the function **fa.poly()** which computes a heterogenous correlation matrix, consisting of polychoric correlations between the binary variables.

The polychoric correlations were used to determine the number of interpretable factors based on the verse simple structure criterion (VSS) of Revelle and Rocklins (1978). The VSS criterion compares the fit of the simplified model to the original correlations and for a given complexity will tend to peak at the optimal (most interpretable) number of factors (Revelle and Rocklin, 1979).

The compounds and fingerprints related to the selected factors were obtained using the **fa.poly()** function with a varimax rotation in the **psych** package (Revelle, 2012) in R. The varimax rotation is an orthogonal rotation method, which tries to reach a minimum correlation between the different factors and a high variance for each factor.

Studying the compounds means understanding the similarities between them in terms of their chemical structures. In other words, to provide a typology of the compounds: which are the most similar (and most dissimilar) compounds ? Are there groups of compounds which are consistent in terms of their chemical structure similarities?. Two compounds will for instance be considered similar if they have the same chemical structure. Compounds are compared on a basis of presence or absence of the chemical structure in question. From this perspective alone, the distance between two compounds depends entirely on their characteristics and not on those of the other compounds. However, it is important to account for the characteristics of the other compounds when calculating this distance.

For computing the distance between two compounds consider the following:

- If two compounds respond positive to the same chemical structure, the distance which separates them should be zero.
- If two compounds both respond positive to the same chemical structures, they should be close together.
- If two compounds respond to all of the same chemical structures except for one which is selected by one of the compounds and only rarely by all of the other compounds, they should be distanced to account for the uniqueness of one of the two.
- If two compounds share a rare chemical structure, they should be close together despite their differences elsewhere in order to account for their common distinctiveness.

These different examples can be used to show that the compounds must be compared category by

category whilst at the same time taking into account the rarity or the universal nature of that category.

3 Results

Based on the filtering by variance and intensity, 629 (25.8%) genes were retained and following the check for separation in the data and removal of variables with only 0's or 1's, 249 out of the 250 fingerprints were retained. Only a summary of the main results will be presented for each method. The R codes are attached on the appendix.

3.1 MFA on Gene expression profiles and Fingerprints

The MFA() function under the **FactoMineR** package was used to perform a multiple factor analysis on the two groups of variables. The fingerprints and gene expression profiles were all considered as active sets of variables.

Prior to interpreting the results, the first step was to determine the number of factors to retain. Table 3 shows the proportion of variance explained based on the first few components. It was decided to keep only components which explain at least 5% of the total variance. Besides only the first two eigen values are greater than 1. Therefore only dimensions 1 and 2 (Factor 1 and Factor 2) were interpreted. It is worth noting that the very small proportion of variances explained could be due to the sparseness of the binary matrix for the fingerprints.

Table 3: Eigenvalues of the MFA and % of variance explained

	eigenvalue	% of variance	cumulative % of variance
dim 1	1.37	7.11	7.11
dim 2	1.04	5.41	12.52
dim 3	0.92	4.77	17.28
dim 4	0.76	3.94	21.23
dim 5	0.73	3.79	25.02
dim 6	0.68	3.56	28.57
dim 7	0.66	3.41	31.98
dim 8	0.62	3.21	35.19
dim 9	0.57	2.94	38.13
dim 10	0.56	2.91	41.05

3.1.1 Group and partial axes representation

Figure 2 shows that the first dimension of the MFA is closely linked to the gene expression profiles for the compounds whereas the second dimension in more closely linked to the fingerprints. In addition, the gene expression profiles loaded very low on the second dimension whereas the fingerprints loaded fairly high on the first dimension. The first two dimensions together explain only 12.5% of the total variance.

Figure 3 shows the projection of the PCA of the gene expression levels, and the MCA results of the fingerprint data on the global PCA analysis. The circle of radius 1 represents the maximum length of a partial standardized axes. The first two axes together explain only 12.5% of the total variance. The first dimension from each group of variables are well represented on the global PCA. The first dimensions of

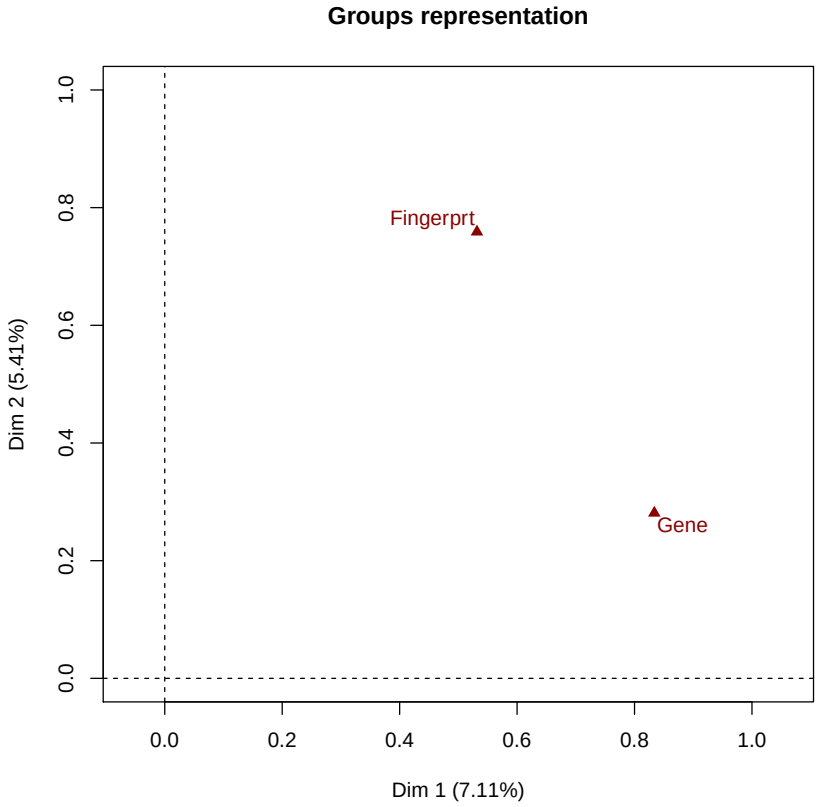

Figure 2: Relationship square: representation of the different groups of variables indicating the gene expression profiles of the compounds were more closely linked to dimension 1 whereas the fingerprints were more closely linked to dimension 2

the gene expressions is highly linked to the first dimension of the MFA whereas the first dimension of fingerprint appears to be highly linked to the second dimension of the MFA.

Based on Table 4, the fingerprints appear to have the highest influence on the overall results of the MFA analysis followed by the gene expression. The RV coefficients (generalized correlation measure) between groups of variables and the MFA consensus plot were > 0.5 implying that the consensus are sufficient. In addition, the fingerprints appear to be only partially linked to the genes.

Table 4: RV-Coefficients showing the correlation between the overall MFA and each data set and between datasets

	Gene	Fingerprt	MFA
Gene	1.00	0.39	0.72
Fingerprint	0.39	1.00	0.92
MFA	0.72	0.92	1.00

Table 5: Contributions of gene expression profiles and fingerprints in each of the dimensions of the MF

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Gene	61.06	27.03	10.59	47.86	57.06
Fingerprint	38.94	72.97	89.41	52.14	42.94

Table 5 gives the contributions of the gene expression and the chemical structures of the compounds

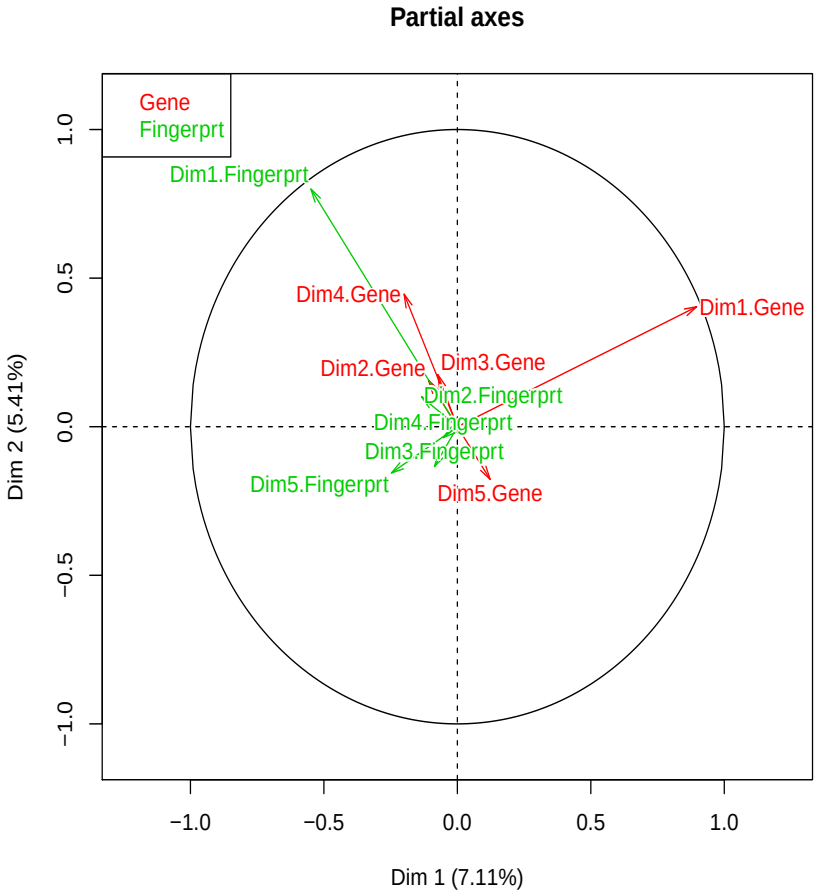

Figure 3: Representation of the partial axes (principal components of the PCA/MCA) on the first plane of the MFA

to each component of the MFA. Dim2 and Dim3 appears to be due mainly to the fingerprint matrix. Dim1 was mainly dominated by the gene expression whereas both have a weak association with Dim4 and Dim5.

Table 6 shows that overall, the first few dimensions were highly linked to the gene expression and the fingerprints with the fingerprint being consistently highly linked to all dimensions.

Table 6: Correlations of gene expression profiles and fingerprints in each of the dimensions of the MFA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Gene	0.92	0.64	0.66	0.84	0.93
Fingerprt	0.84	0.91	0.99	0.86	0.88

3.1.2 Lead compounds and genes for Factor 1 and Factor 2

For Factors 1 and 2 (dim 1 and 2), the average contribution was 1.785714 thus lead compounds were considered as those with contributions larger than 1.785714. The most important genes and lead compounds for Factor 1 and Factor 2 are as shown on Figures 4 and 5 respectively.

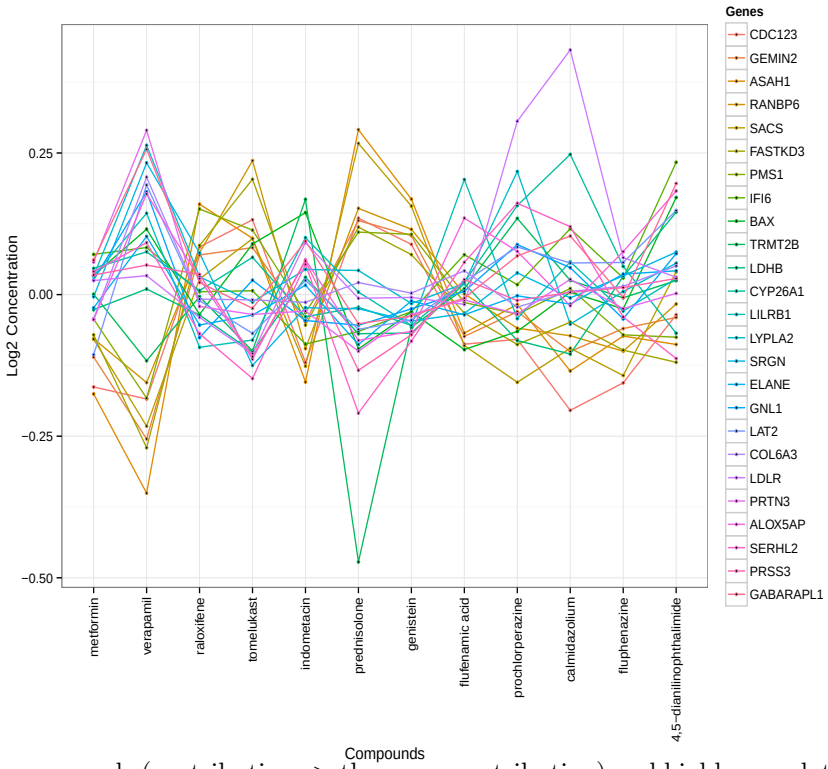

Figure 4: Lead compounds (contributions > the mean contribution) and highly correlated genes for Dim 1 (Factor 1)

From Figure 4 it can clearly be seen that all the compounds with high contributions to factor 1 importantly regulated (up or down) all the genes highly correlated with factor 1. Compounds such as verapamil, prednisolone and calmidazolium appeared to exert higher regulation activities compared to the other compounds linked to factor 1. Similarly, from Figure 5, it appears that trifluoperazine and LY-294002 exerted higher gene regulations on average compared to the other compounds linked to factor 2

3.1.3 Qualitative variables highly characteristic of Factor 1 and Factor 2

Among all the fingerprint categories that were significant (p-value < 0.05) based on the t-test comparing the average of the category with the general mean, the top 10 for factor 1 and 2 are as presented on Table 7 and 8 respectively. For factor 1, it can be concluded that the compounds which are linked to factor 1, lacked fingerprints such as FP184, FP188 and FP19 and possessed the fingerprints FP211 and FP218.

On the other hand, all the compounds with high contributions to factor 2 possessed fingerprints such as FP240, FP151, FP92 and FP129 and did not contain fingerprints such as FP111, FP24 and FP196.

3.1.4 Hierarchical Cluster analysis

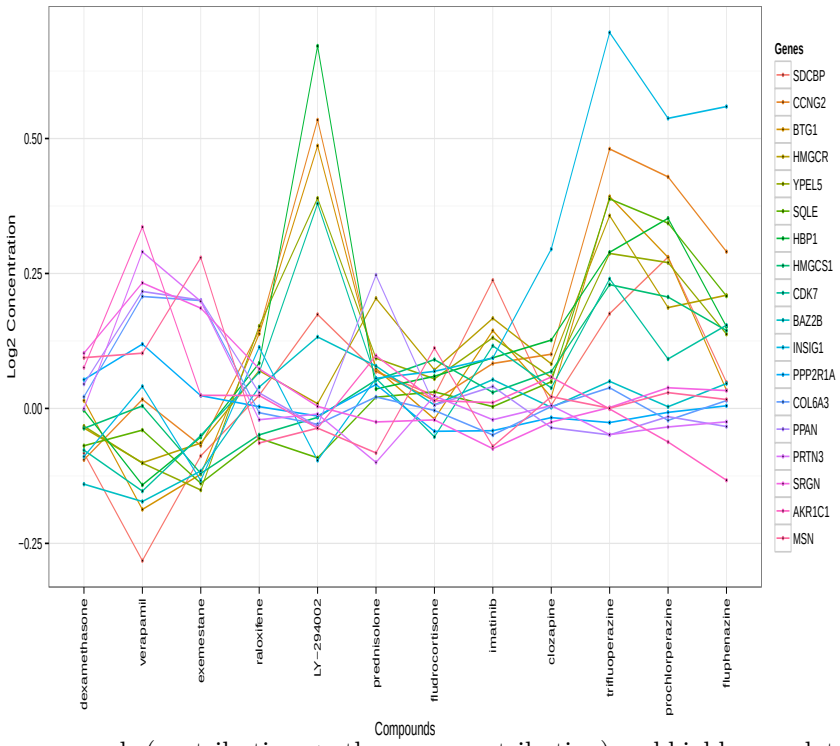

Figure 5: Lead compounds (contributions > the mean contribution) and highly correlated genes for Dim 2 (Factor 2)

Table 7: Top 10 fingerprint categories that highly characterise Dim 1. These are ranked using p-values of a t-test to compare the average of the category with the general mean.

Fingerprint Code	Estimate	p.value
FP184_0	1.15	<0.001
FP188_0	1.29	<0.001
FP19_0	2.08	<0.001
FP218_1	0.54	<0.001
FP66_0	1.42	<0.001
FP167_0	0.50	<0.001
FP239_0	0.63	<0.001
FP61_0	0.91	<0.001
FP162_0	0.82	<0.001
FP211_1	0.48	<0.001

Table 8: Top 10 fingerprint categories that highly characterise Dim 2. These are ranked using p-values of a t-test to compare the average of the category with the general mean.

Fingerprint Code	Estimate	p.value
FP240_1	0.81	<0.001
FP151_1	1.19	<0.001
FP92_1	0.98	<0.001
FP129_1	1.30	<0.001
FP8_1	1.13	<0.001
FP111_0	0.94	<0.001
FP24_0	1.28	<0.001
FP91_0	0.99	<0.001
FP196_0	0.72	<0.001
FP46_1	0.95	<0.001

Following the MFA, a HCPC (Hierarchical Clustering on Principal Components) analysis was performed and it suggested 3 clusters of compounds as shown on Figure 6.

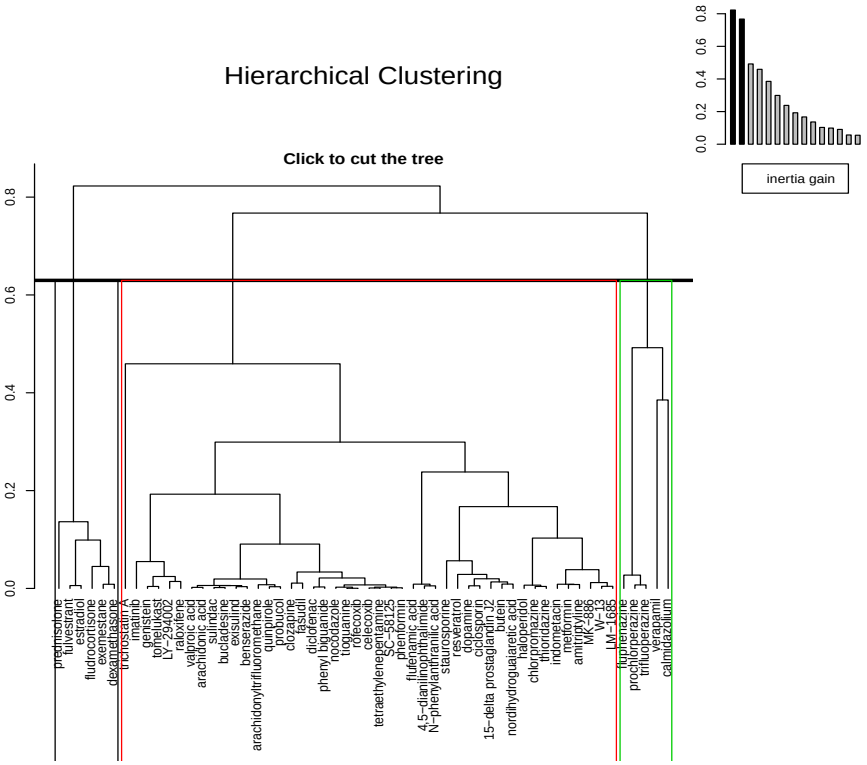

Figure 6: Results of the clusters made by the cluster analysis using the Ward method. Cutting point indicates the minimum possible groups that maximized the explained variance which was 3 in this case

The partitioning into 3 clusters is represented on the map produced by the first two principal components and the compounds are coloured according to their cluster (Figure 7). The barycentre of each cluster is also represented by a square. The graph shows that the 3 clusters are well-separated on the first two principal components.

Table 9: Paragons: Compounds closest to the center of each cluster

CLUSTER					
CLUSTER 1	chlorpromazine	haloperidol	thioridazine	fluphenazine	prochlorperazine
	0.7996771	1.2681776	1.2796898	1.5208936	2.0682367
CLUSTER 2	dopamine	tetraethylenepentamine	celecoxib	SC-58125	phenformin
	0.4373451	0.4712053	0.5208013	0.5369278	0.5981553
CLUSTER 3	estradiol	dexamethasone	fludrocortisone	fulvestrant	exemestane
	1.001026	1.154125	1.339470	1.569099	1.881981

Table 9 shows the compounds that are closest to the center of each cluster. As indicated, estradiol was closest the center of cluster 3, dopamine to the center of cluster 2 and chlorpromazine to the center of cluster 1. Therefore these compounds each best represent the various clusters to which they are closest to the center.

On the other hand, compounds furthest from the barycenter of the other clusters from the one considered are as shown on Table 10. As indicated, prednisolone is specific to cluster 3 but it is furthest from the

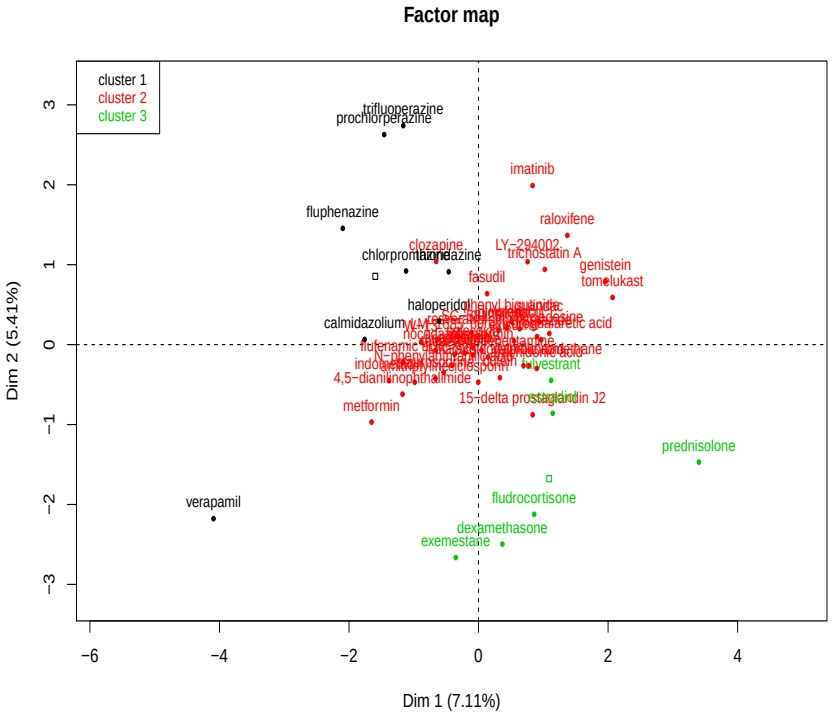

Figure 7: Representation of the clusters on the map induced by the first two principal components

Table 10: Distance: Compounds furthest from the barycenter of the other clusters from the one considered

CLUSTER				
CLUSTER: 1				
calmidazolium	verapamil	trifluoperazine	prochlorperazine	fluphenazine
5.208793	5.114140	4.352992	3.994919	3.728681
CLUSTER: 2				
trichostatin A	N-phenylanthranilic acid	4,5-dianilinophthalimide	imatinib	genistein
5.664240	3.667819	3.635855	3.509624	3.503030
CLUSTER: 3				
prednisolone	fludrocortisone	dexamethasone	exemestane	estradiol
4.692733	3.621239	3.533220	3.413393	2.113561

center of clusters 1 and 2. Similarly, calmidazolium is specific to cluster 1 and is furthest from the centers of cluster 2 and 3 whereas trichostatin A , specific to cluster 2 is furthest from the centers of cluster 1 and 3.

The three clusters obtained are characterised by quantitative (gene expression) as well as categorical variables (fingerprints). When the clusters are characterized by categorical variables, such as ours, an association test is performed between the categorical variables and the cluster variable and only the categorical variables with a p-value <=0.02 are presented. The ranking is based on the values of the test statistic with negative values indicating negative associations.

Description of cluster 1: Table 11 is a shortened version of the the full table showing the fingerprints significantly associated with cluster 1. We have selected only the top 10 fingerprints according to the significance of the correlation test (v.test) between the fingerprint and the cluster. Cla/Mod gives the proportion of compounds which indicate presence of the chemical structure (FP) indicated and which are present in cluster 1 whereas Mod/Cla gives the proportion of compounds in the cluster which indicate presence of indicated chemical structure. For example, considering the fingerprint FP46, the value of Cla/Mod is 100%. This says that of all the compounds in the study which indicate presence of the chemical structure, all 100% of them are in cluster 1. On the other hand, Mod/Cla has a value of 62.5% stating that 62.5% of compounds in the cluster indicate presence of the chemical structure. The percentage is different because the cluster is made of compounds which indicate presence of the chemical structure and others which do not indicate presence of the chemical structure. From a global point of view, the compounds in cluster 1 turn not to have the fingerprints listed (FP=0) as they were found to be more significantly and negatively correlated with cluster 1 (global scores were all ≥ 55).

Table 11: Top 10 fingerprints highly positively or negatively correlated (based on v.test) with cluster 1

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP239=FP239_1	70.00	87.50	17.86	<0.001	4.61
FP46=FP46_1	100.00	62.50	8.93	<0.001	4.33
FP200=FP200_1	83.33	62.50	10.71	<0.001	3.93
FP139=FP139_1	100.00	50.00	7.14	<0.001	3.73
FP67=FP67_1	100.00	50.00	7.14	<0.001	3.73
FP201=FP201_1	43.75	87.50	28.57	<0.001	3.58
FP167=FP167_1	32.00	100.00	44.64	<0.001	3.37
FP227=FP227_1	80.00	50.00	8.93	<0.001	3.32
FP107=FP107_1	55.56	62.50	16.07	<0.001	3.16
FP188=FP188_1	100.00	37.50	5.36	<0.001	3.09
FP99=FP99_0	9.43	62.50	94.64	<0.001	-3.09
FP107=FP107_0	6.38	37.50	83.93	<0.001	-3.16
FP227=FP227_0	7.84	50.00	91.07	<0.001	-3.32
FP167=FP167_0	0.00	0.00	55.36	<0.001	-3.37
FP201=FP201_0	2.50	12.50	71.43	<0.001	-3.58
FP139=FP139_0	7.69	50.00	92.86	<0.001	-3.73
FP67=FP67_0	7.69	50.00	92.86	<0.001	-3.73
FP200=FP200_0	6.00	37.50	89.29	<0.001	-3.93
FP46=FP46_0	5.88	37.50	91.07	<0.001	-4.33
FP239=FP239_0	2.17	12.50	82.14	<0.001	-4.61

Table 12 shows the top 10 up-regulated genes associated with cluster 1. These genes are mostly up-regulated by the compounds in cluster 1.

Table 13 shows the top 10 down-regulated genes associated with cluster 1. These genes are mostly down-regulated by the compounds that make up cluster 1

Table 12: Top 10 up-regulated genes that characterise cluster 1

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
INSIG1	5.09	0.49	0.10	0.26	0.23	<0.001
LDLR	4.37	0.14	0.01	0.14	0.08	<0.001
BHLHE40	4.35	0.23	0.04	0.25	0.14	<0.001
CYP26A1	4.32	0.08	0.00	0.08	0.05	<0.001
SQLE	4.29	0.19	0.03	0.13	0.11	<0.001
HMGCS1	4.22	0.12	0.03	0.07	0.07	<0.001
LPIN1	4.10	0.21	0.05	0.19	0.12	<0.001
PLAU	4.02	0.10	0.01	0.09	0.06	<0.001
CCNG2	3.79	0.26	0.08	0.17	0.14	<0.001
HMGCR	3.78	0.18	0.05	0.13	0.10	<0.001

Table 13: Top 10 down-regulated gene that characterise cluster 1

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
NAT10	-3.48	-0.11	-0.03	0.07	0.06	<0.001
CDC123	-3.56	-0.09	-0.01	0.07	0.07	<0.001
PTDSS1	-3.67	-0.09	-0.01	0.07	0.06	<0.001
PTER	-3.67	-0.09	-0.00	0.05	0.07	<0.001
ZMPSTE24	-3.69	-0.09	-0.00	0.09	0.07	<0.001
FPGT	-3.85	-0.06	0.01	0.05	0.06	<0.001
ZNHIT6	-3.87	-0.14	-0.01	0.15	0.10	<0.001
ILF2	-4.02	-0.12	-0.01	0.15	0.08	<0.001
TM2D3	-4.25	-0.11	-0.01	0.08	0.07	<0.001
PNO1	-4.62	-0.12	-0.01	0.07	0.07	<0.001

Description of cluster 2

The compounds closest to the center of cluster 2 are: dopamine, tetraethylenepentamine, celecoxib,SC-58125 and phenformin . Table 14 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 2.

Table 14: Description of Cluster 2 by the top 10 fingerprint categories

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP176=FP176_0	84.00	100.00	89.29	0.00	3.91
FP111=FP111_0	84.00	100.00	89.29	0.00	3.91
FP89=FP89_0	89.74	83.33	69.64	0.00	3.60
FP185=FP185_0	85.11	95.24	83.93	0.00	3.52
FP46=FP46_0	82.35	100.00	91.07	0.00	3.47
FP239=FP239_0	84.78	92.86	82.14	0.00	3.23
FP231=FP231_0	80.77	100.00	92.86	0.00	3.00
FP230=FP230_0	80.77	100.00	92.86	0.00	3.00
FP139=FP139_0	80.77	100.00	92.86	0.00	3.00
FP67=FP67_0	80.77	100.00	92.86	0.00	3.00
FP219=FP219_1	33.33	7.14	16.07	0.01	-2.78
FP232=FP232_1	16.67	2.38	10.71	0.00	-2.99
FP200=FP200_1	16.67	2.38	10.71	0.00	-2.99
FP231=FP231_1	0.00	0.00	7.14	0.00	-3.00
FP230=FP230_1	0.00	0.00	7.14	0.00	-3.00
FP139=FP139_1	0.00	0.00	7.14	0.00	-3.00
FP67=FP67_1	0.00	0.00	7.14	0.00	-3.00
FP239=FP239_1	30.00	7.14	17.86	0.00	-3.23
FP46=FP46_1	0.00	0.00	8.93	0.00	-3.47
FP185=FP185_1	22.22	4.76	16.07	0.00	-3.52

From a global point of view, the compounds in cluster 2 turn not to possess the fingerprints listed (FP=0) as they were found to be more significantly and positively correlated with cluster 2 (global scores were all ≥ 83).

Table 15 shows the top 10 up-regulated genes associated with cluster 2 whereas Table 16 shows the top 10 down-regulated genes associated with cluster 2. These genes are up-regulated and down-regulated respectively following the introduction of the compounds in cluster 2.

Table 15: Top 10 up-regulated genes that characterise cluster 2

	v.test	Mean in category	Overall mean		Overall sd	p.value
FPGT	3.54	0.03	0.01	0.05	0.06	<0.001
MTERFD1	3.25	0.00	-0.02	0.07	0.08	<0.001
VAMP3	3.23	0.00	-0.01	0.05	0.07	<0.001
TM2D3	2.99	0.00	-0.01	0.05	0.07	<0.001
ILF2	2.94	0.01	-0.01	0.04	0.08	<0.001
PNO1	2.80	0.00	-0.01	0.05	0.07	0.01
PCYOX1	2.80	0.05	0.03	0.09	0.10	0.01
TTC19	2.75	0.01	-0.01	0.04	0.05	0.01
PTER	2.68	0.01	-0.00	0.07	0.07	0.01
NUP37	2.60	0.00	-0.01	0.05	0.06	0.01

Table 16: Top 10 down-regulated genes that characterise cluster 2

	v.test	Mean in category	Overall mean		Overall sd	p.value
SQLE	-2.62	0.00	0.03	0.09	0.11	0.01
LPIN1	-2.73	0.02	0.05	0.08	0.12	0.01
HMGCR	-2.74	0.03	0.05	0.08	0.10	0.01
HMGCS1	-2.82	0.01	0.03	0.05	0.07	<0.001
PLAU	-3.09	-0.00	0.01	0.04	0.06	<0.001
LDLR	-3.22	-0.01	0.01	0.05	0.08	<0.001
INSIG1	-3.28	0.04	0.10	0.15	0.23	<0.001
BHLHE40	-3.33	0.00	0.04	0.07	0.14	<0.001
CYP26A1	-3.48	-0.01	0.00	0.04	0.05	<0.001
DNMBP	-4.07	-0.01	0.01	0.04	0.05	<0.001

Description of cluster 3

The compounds that closest to the center of cluster 3 are: estradiol , dexamethasone, fludrocortisone, fulvestrant and exemestane. Table 17 is a shortened version of the the full table showing the fingerprints significantly associated with compounds in cluster 3.

Table 17: Description of Cluster 3 by the categories of the fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP111=FP111_1	100.00	100.00	10.71	<0.001	5.54
FP219=FP219_1	66.67	100.00	16.07	<0.001	4.70
FP185=FP185_1	66.67	100.00	16.07	<0.001	4.70
FP232=FP232_1	83.33	83.33	10.71	<0.001	4.43
FP176=FP176_1	83.33	83.33	10.71	<0.001	4.43
FP231=FP231_1	100.00	66.67	7.14	<0.001	4.10
FP126=FP126_1	80.00	66.67	8.93	<0.001	3.72
FP91=FP91_1	80.00	66.67	8.93	<0.001	3.72
FP196=FP196_1	50.00	83.33	17.86	<0.001	3.56
FP89=FP89_1	35.29	100.00	30.36	<0.001	3.55
FP89=FP89_0	0.00	0.00	69.64	<0.001	-3.55
FP196=FP196_0	2.17	16.67	82.14	<0.001	-3.56
FP126=FP126_0	3.92	33.33	91.07	<0.001	-3.72
FP91=FP91_0	3.92	33.33	91.07	<0.001	-3.72
FP231=FP231_0	3.85	33.33	92.86	<0.001	-4.10
FP232=FP232_0	2.00	16.67	89.29	<0.001	-4.43
FP176=FP176_0	2.00	16.67	89.29	<0.001	-4.43
FP219=FP219_0	0.00	0.00	83.93	<0.001	-4.70
FP185=FP185_0	0.00	0.00	83.93	<0.001	-4.70
FP111=FP111_0	0.00	0.00	89.29	<0.001	-5.54

From a global point of view, the compounds in cluster 3 possess the fingerprints listed (FP=1) as they were found to be more significantly and positively correlated with cluster 3 (global scores were all ≥ 66)

Table 18 shows the top 10 up-regulated genes associated with cluster 3 one down-regulated gene associated with cluster 3. In the presence of the compounds in cluster 3, these genes are regulated.

Table 18: Top 10 Up-regulated genes and one down regulated gene that characterise cluster 3

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
SNAI2	2.79	0.10	0.02	0.09	0.08	0.01
CLPP	2.78	0.06	0.01	0.05	0.04	0.01
PPAN	2.52	0.10	0.01	0.09	0.09	0.01
COMMD3	2.41	0.12	0.04	0.10	0.09	0.02
DNMBP	2.38	0.05	0.01	0.04	0.05	0.02
RDBP	2.36	0.05	-0.00	0.05	0.06	0.02
ACTG1	2.33	0.06	-0.01	0.11	0.08	0.02
ZIC1	2.23	0.08	-0.01	0.13	0.11	0.03
NAV2	2.19	0.07	0.02	0.07	0.07	0.03
NAT1	2.17	0.04	-0.01	0.04	0.06	0.03
APOBEC3C	-2.64	-0.06	0.01	0.03	0.07	0.01

3.2 MCA on Fingerprints (Table 2)

The MCA was applied to Table 2 which does not include gene expression, and the results were used to perform a hierarchical clustering analysis. For the first two factors (dimensions), the lead compounds and the fingerprints were obtained and are represented in the subsections that follow:

Prior to interpreting the results from the MCA, the first step was to determine the number of factors to retain. Table 19 shows the proportion of variance explained based on the first few components. It was decided to keep only components which explain at least 5% of the total variance. Besides only the first two eigen values are greater than 1. Therefore only dimensions 1 and 2 (Factor 1 and Factor 2) were interpreted. It is worth noting that the very small proportion of variances explained could be due to the sparseness of the binary matrix for the fingerprints.

Table 19: Eigenvalues of the MCA and % of variance explained

	eigenvalue	% of variance	cumulative % of variance
dim 1	0.07	6.94	6.94
dim 2	0.06	5.80	12.74
dim 3	0.04	4.39	17.13
dim 4	0.04	4.19	21.32
dim 5	0.04	3.96	25.28
dim 6	0.04	3.74	29.02
dim 7	0.03	3.39	32.41
dim 8	0.03	3.35	35.76
dim 9	0.03	3.10	38.87
dim 10	0.03	2.94	41.81

3.2.1 Lead compounds highly contributing to factor 1 and 2 based on the MCA

For Factors 1 and 2 (dim 1 and 2) based on the MCA, the average contribution for all compounds was 1.785714 thus lead compounds were considered as those with contributions larger than 1.785714. The lead compounds for Factor 1 and Factor 2 are as presented on Table 20.

Table 20

Table 20: Compounds with high contributions to factors 1 and 2

Factor	% Var explained	Compounds
Factor 1	6.94	estradiol, dexamethasone, exemestane , prednisolone, fludrocortisone, chlorpromazine , trifluoperazine, prochlorperazine , fluphenazine, fulvestrant, imatinib , clozapine
Factor 2	5.80	estradiol , dexamethasone , prednisolone fludrocortisone, fulvestrant, N-phenylanthranilic acid , diclofenac, chlorpromazine , trifluoperazine , prochlorperazine , fluphenazine, 4,5-dianilinophthalimide, exemestane

3.2.2 Qualitative variables highly characteristic of Factor 1 and Factor 2

Among all the fingerprint categories that were significant (p-value < 0.05) based on the t-test comparing the average of the category with the general mean, the top 10 for factor 1 and 2 are as presented on Table21 and 22 respectively. For factor 1, it can be concluded that the compounds which are highly linked to factor 1, did not contain the listed fingerprints with the exception of FP150.

Table 21: Top 10 fingerprint categories that highly characterise Dim 1. These are ranked using p-values of a t-test to compare the average of the category with the general mean.

Fingerprint Code	Estimate	p.value
FP111_0	0.32	<0.001
FP91_0	0.31	<0.001
FP232_0	0.28	<0.001
FP219_0	0.23	<0.001
FP18_0	0.37	<0.001
FP185_0	0.22	<0.001
FP126_0	0.28	<0.001
FP77_0	0.35	<0.001
FP150_1	0.16	<0.001
FP237_0	0.35	<0.001

Table 22: Top 10 fingerprint categories that highly characterise Dim 2. These are ranked using p-values of a t-test to compare the average of the category with the general mean.

Fingerprint Code	Estimate	p.value
FP99_1	0.36	<0.001
FP67_1	0.31	<0.001
FP46_1	0.27	<0.001
FP139_1	0.30	<0.001
FP227_1	0.25	<0.001
FP151_1	0.27	<0.001
FP239_1	0.18	<0.001
FP96_1	0.37	<0.001
FP200_1	0.22	<0.001
FP89_1	0.15	<0.001

The HCPC (Hierarchical Clustering on Principal Components) analysis on the coordinates of the compounds resulting from the MCA suggested 6 clusters as shown on Figure 8.

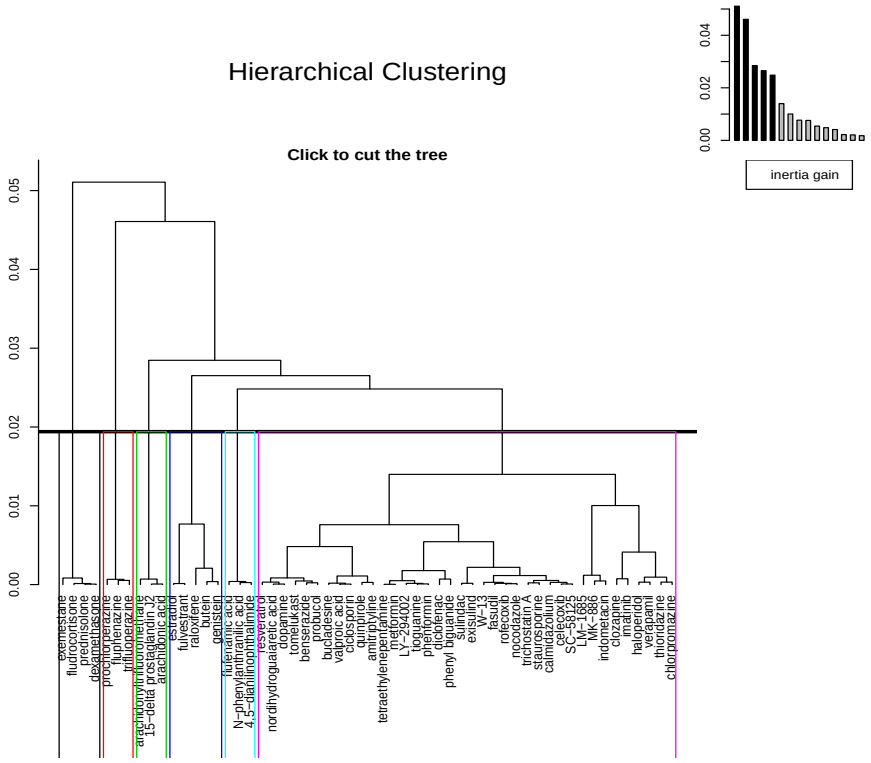

Figure 8: Result of the clusters made by the cluster analysis using the Ward method. Cutting point indicates the minimum possible groups that maximized the explained variance. There were 6 clusters in total

The partitioning into six clusters is represented on the map produced by the first two principal components and the compounds are coloured according to their cluster (Figure 9). The barycentre of each cluster is also represented by a square. The graph shows that the six clusters are fairly well-separated on the first two principal components.

The first cluster is made up of: dexamethasone , fludrocortisone , prednisolone and exemestane which are closest to its centre. Table 23 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 1.

Table 23: Description of Cluster 1 by fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP237=FP237_1	100.00	75.00	5.36	< 0.001	3.80
FP77=FP77_1	100.00	75.00	5.36	< 0.001	3.80
FP24=FP24_1	100.00	75.00	5.36	< 0.001	3.80
FP18=FP18_1	100.00	75.00	5.36	< 0.001	3.80
FP148=FP148_1	100.00	50.00	3.57	< 0.001	2.89
FP121=FP121_1	100.00	50.00	3.57	< 0.001	2.89
FP120=FP120_1	100.00	50.00	3.57	< 0.001	2.89
FP31=FP31_1	100.00	50.00	3.57	< 0.001	2.89

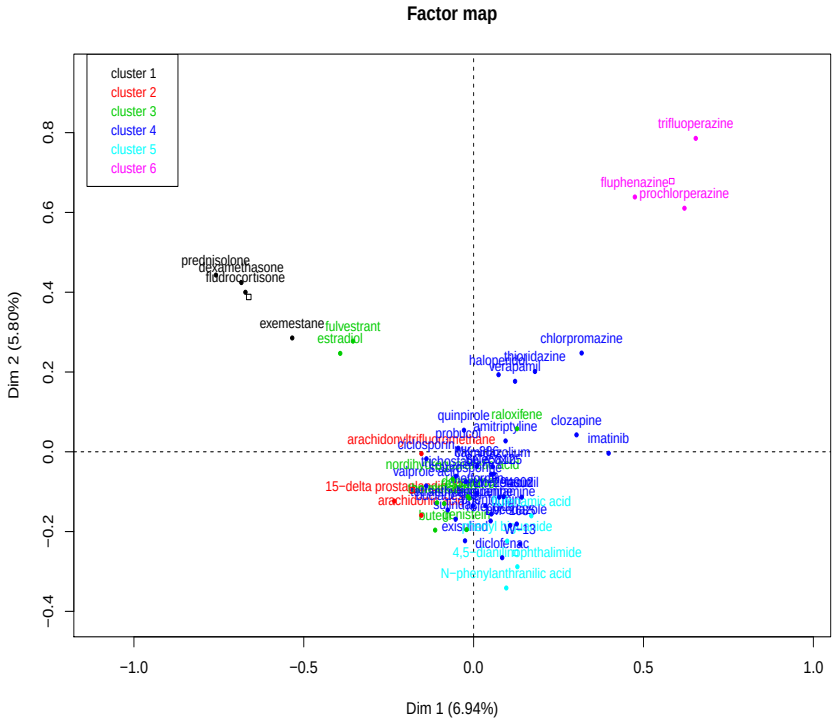

Figure 9: Representation of the clusters on the map induced by the first two principal components

The second cluster is made up of: 15-delta prostaglandin J2 , arachidonic acid and arachidonyltrifluoromethane. Table 24 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 2.

The third cluster is made up of: estradiol, raloxifene, nordihydroguaiaretic acid, tomelukast, genistein, fulvestrant, dopamine, resveratrol, butein, and benserazide. However, those that are closest to its barycenter are:resveratrol, nordihydroguaiaretic acid , dopamine, butein and benserazide. Table 25 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 3.

The fourth cluster is made up of metformin, phenformin, verapamil, rofecoxib, amitriptyline, celecoxib, LM-1685, SC-58125, LY-294002, ciclosporin, indometacin, MK-886, sulindac, exisulind, staurosporine, trichostatin A, diclofenac, fasudil,valproic acid, imatinib, tetraethylenepentamine, clozapine, thioridazine, haloperidol, chlorpromazine, W-13, quinpirole, calmidazolium, bucladesine, probucol, nocodazole, tioguanine. However those that are closest to its center are: SC-58125, celecoxib, fasudil, staurosporine and trichostatin A . Table 26 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 4.

The fifth cluster is made up of: 4,5-dianilinophthalimide, N-phenylanthranilic acid, flufenamic acid, phenyl, biguanide. Table 27 is a shortened version of the the full table showing the fingerprints significantly associated with cluster 5. Based on the global scores (>75), it can be concluded that the cluster is made up of compounds without the listed fingerprints.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP126=FP126_1	80.00	100.00	8.93	< 0.001	4.35
FP91=FP91_1	80.00	100.00	8.93	< 0.001	4.35
FP111=FP111_1	66.67	100.00	10.71	< 0.001	4.10
FP237=FP237_1	100.00	75.00	5.36	< 0.001	3.80
FP77=FP77_1	100.00	75.00	5.36	< 0.001	3.80
FP24=FP24_1	100.00	75.00	5.36	< 0.001	3.80
FP18=FP18_1	100.00	75.00	5.36	< 0.001	3.80
FP219=FP219_1	44.44	100.00	16.07	< 0.001	3.58
FP185=FP185_1	44.44	100.00	16.07	< 0.001	3.58
FP230=FP230_1	75.00	75.00	7.14	< 0.001	3.44
FP230=FP230_0	1.92	25.00	92.86	< 0.001	-3.44
FP219=FP219_0	0.00	0.00	83.93	< 0.001	-3.58
FP185=FP185_0	0.00	0.00	83.93	< 0.001	-3.58
FP237=FP237_0	1.89	25.00	94.64	< 0.001	-3.80
FP77=FP77_0	1.89	25.00	94.64	< 0.001	-3.80
FP24=FP24_0	1.89	25.00	94.64	< 0.001	-3.80
FP18=FP18_0	1.89	25.00	94.64	< 0.001	-3.80
FP111=FP111_0	0.00	0.00	89.29	< 0.001	-4.10
FP126=FP126_0	0.00	0.00	91.07	< 0.001	-4.35
FP91=FP91_0	0.00	0.00	91.07	< 0.001	-4.35

Table 24: Description of Cluster 2 by fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP123=FP123_1	100.00	100.00	5.36	< 0.001	4.13
FP207=FP207_1	75.00	100.00	7.14	< 0.001	3.80
FP146=FP146_1	75.00	100.00	7.14	< 0.001	3.80
FP195=FP195_1	60.00	100.00	8.93	< 0.001	3.57
FP203=FP203_1	42.86	100.00	12.50	< 0.001	3.22
FP143=FP143_1	100.00	66.67	3.57	< 0.001	3.10
FP68=FP68_1	100.00	66.67	3.57	< 0.001	3.10
FP223=FP223_1	37.50	100.00	14.29	< 0.001	3.09
FP234=FP234_1	66.67	66.67	5.36	0.01	2.76
FP132=FP132_0	20.00	100.00	26.79	0.02	2.40
FP132=FP132_1	0.00	0.00	73.21	0.02	-2.40
FP234=FP234_0	1.89	33.33	94.64	0.01	-2.76
FP223=FP223_0	0.00	0.00	85.71	< 0.001	-3.09
FP143=FP143_0	1.85	33.33	96.43	< 0.001	-3.10
FP68=FP68_0	1.85	33.33	96.43	< 0.001	-3.10
FP203=FP203_0	0.00	0.00	87.50	< 0.001	-3.22
FP195=FP195_0	0.00	0.00	91.07	< 0.001	-3.57
FP207=FP207_0	0.00	0.00	92.86	< 0.001	-3.80
FP146=FP146_0	0.00	0.00	92.86	< 0.001	-3.80
FP123=FP123_0	0.00	0.00	94.64	< 0.001	-4.13

The sixth cluster is made up of:trifluoperazine , fluphenazine and prochlorperazine. Table 28 is a shortened version of the the full table showing the fingerprint categories significantly associated with cluster 6. Based on the global scores, it can be concluded that the cluster is made up of compounds without the listed fingerprints.

3.3 BiMax on Fingerprints (Table 2)

Table 29 shows the results of 5 biclusters obtained using the BiMax algorithm. The clusters were ranked using the SB score. The resulting biclusters contain large groups of compounds which exhibit

Table 25: Description of Cluster 3 by fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP226=FP226_1	90.91	100.00	19.64	< 0.001	6.29
FP153=FP153_1	100.00	60.00	10.71	< 0.001	4.51
FP136=FP136_1	100.00	50.00	8.93	< 0.001	3.99
FP218=FP218_1	34.48	100.00	51.79	< 0.001	3.45
FP141=FP141_1	100.00	40.00	7.14	< 0.001	3.44
FP117=FP117_1	100.00	40.00	7.14	< 0.001	3.44
FP147=FP147_1	100.00	30.00	5.36	< 0.001	2.85
FP115=FP115_1	100.00	30.00	5.36	< 0.001	2.85
FP150=FP150_0	27.78	100.00	64.29	0.01	2.69
FP167=FP167_0	29.03	90.00	55.36	0.02	2.39
FP167=FP167_1	4.00	10.00	44.64	0.02	-2.39
FP150=FP150_1	0.00	0.00	35.71	0.01	-2.69
FP147=FP147_0	13.21	70.00	94.64	< 0.001	-2.85
FP115=FP115_0	13.21	70.00	94.64	< 0.001	-2.85
FP141=FP141_0	11.54	60.00	92.86	< 0.001	-3.44
FP117=FP117_0	11.54	60.00	92.86	< 0.001	-3.44
FP218=FP218_0	0.00	0.00	48.21	< 0.001	-3.45
FP136=FP136_0	9.80	50.00	91.07	< 0.001	-3.99
FP153=FP153_0	8.00	40.00	89.29	< 0.001	-4.51
FP226=FP226_0	0.00	0.00	80.36	< 0.001	-6.29

Table 26: Description of Cluster 4 by fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP167=FP167_1	84.00	65.62	44.64	< 0.001	3.62
FP226=FP226_0	68.89	96.88	80.36	< 0.001	3.51
FP218=FP218_0	77.78	65.62	48.21	< 0.001	2.95
FP153=FP153_0	64.00	100.00	89.29	< 0.001	2.87
FP111=FP111_0	64.00	100.00	89.29	< 0.001	2.87
FP136=FP136_0	62.75	100.00	91.07	0.01	2.54
FP91=FP91_0	62.75	100.00	91.07	0.01	2.54
FP56=FP56_1	100.00	21.88	12.50	0.01	2.44
FP209=FP209_1	90.00	28.12	17.86	0.02	2.28
FP161=FP161_1	90.00	28.12	17.86	0.02	2.28
FP209=FP209_0	50.00	71.88	82.14	0.02	-2.28
FP161=FP161_0	50.00	71.88	82.14	0.02	-2.28
FP56=FP56_0	51.02	78.12	87.50	0.01	-2.44
FP136=FP136_1	0.00	0.00	8.93	0.01	-2.54
FP91=FP91_1	0.00	0.00	8.93	0.01	-2.54
FP153=FP153_1	0.00	0.00	10.71	< 0.001	-2.87
FP111=FP111_1	0.00	0.00	10.71	< 0.001	-2.87
FP218=FP218_1	37.93	34.38	51.79	< 0.001	-2.95
FP226=FP226_1	9.09	3.12	19.64	< 0.001	-3.51
FP167=FP167_0	35.48	34.38	55.36	< 0.001	-3.62

small groups of fingerprints. An additional observed characteristic of the biclusters is that there is an overlap with respect to the compounds and the finger prints. The compounds raloxifene, MK-888, butein, genistein, sulindac and N-phenylanthranilic acid were present in all the 5 biclusters reported by BiMax algorithm indicating that a strong grouping tendency for these compounds. The corresponding fingerprint that were common to all top 4 clusters were FP76, FP83, and FP132 with FP83 present all the compounds in all the biclusters.

In comparison with the results of MCA, it can be seen that the first two compounds closest to the center of cluster 4 of the MCA: SC-58125 and celecoxib were all present in the top 3 biclusters based on BiMax. However the corresponding fingerprints related to these compounds were different. Those based on BiMax FP76; FP83 and FP132 were not among those based on MCA for cluster 4. In addition, three top compounds for cluster 5 of the MCA: 4,5-dianilinophthalimide, N-phenylanthranilic acid and

Table 27: Description of Cluster 5 by fingerprints

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP215=FP215_1	100.00	100.00	7.14	< 0.001	4.69
FP105=FP105_1	80.00	100.00	8.93	< 0.001	4.35
FP245=FP245_1	100.00	75.00	5.36	< 0.001	3.80
FP157=FP157_1	100.00	75.00	5.36	< 0.001	3.80
FP140=FP140_1	40.00	100.00	17.86	< 0.001	3.44
FP171=FP171_1	28.57	100.00	25.00	< 0.001	3.00
FP102=FP102_0	28.57	100.00	25.00	< 0.001	3.00
FP81=FP81_1	50.00	75.00	10.71	< 0.001	2.99
FP47=FP47_1	50.00	75.00	10.71	< 0.001	2.99
FP72=FP72_1	100.00	50.00	3.57	< 0.001	2.89
FP44=FP44_0	3.70	50.00	96.43	< 0.001	-2.89
FP81=FP81_0	2.00	25.00	89.29	< 0.001	-2.99
FP47=FP47_0	2.00	25.00	89.29	< 0.001	-2.99
FP171=FP171_0	0.00	0.00	75.00	< 0.001	-3.00
FP102=FP102_1	0.00	0.00	75.00	< 0.001	-3.00
FP140=FP140_0	0.00	0.00	82.14	< 0.001	-3.44
FP245=FP245_0	1.89	25.00	94.64	< 0.001	-3.80
FP157=FP157_0	1.89	25.00	94.64	< 0.001	-3.80
FP105=FP105_0	0.00	0.00	91.07	< 0.001	-4.35
FP215=FP215_0	0.00	0.00	92.86	< 0.001	-4.69

Table 28: Description of Cluster 6 by fingerprints.

	Cla/Mod	Mod/Cla	Global	p.value	v.test
FP99=FP99_1	100.00	100.00	5.36	< 0.001	4.13
FP151=FP151_1	75.00	100.00	7.14	< 0.001	3.80
FP139=FP139_1	75.00	100.00	7.14	< 0.001	3.80
FP67=FP67_1	75.00	100.00	7.14	< 0.001	3.80
FP227=FP227_1	60.00	100.00	8.93	< 0.001	3.57
FP46=FP46_1	60.00	100.00	8.93	< 0.001	3.57
FP200=FP200_1	50.00	100.00	10.71	< 0.001	3.38
FP92=FP92_1	50.00	100.00	10.71	< 0.001	3.38
FP101=FP101_1	100.00	66.67	3.57	< 0.001	3.10
FP96=FP96_1	100.00	66.67	3.57	< 0.001	3.10
FP101=FP101_0	1.85	33.33	96.43	< 0.001	-3.10
FP96=FP96_0	1.85	33.33	96.43	< 0.001	-3.10
FP200=FP200_0	0.00	0.00	89.29	< 0.001	-3.38
FP92=FP92_0	0.00	0.00	89.29	< 0.001	-3.38
FP227=FP227_0	0.00	0.00	91.07	< 0.001	-3.57
FP46=FP46_0	0.00	0.00	91.07	< 0.001	-3.57
FP151=FP151_0	0.00	0.00	92.86	< 0.001	-3.80
FP139=FP139_0	0.00	0.00	92.86	< 0.001	-3.80
FP67=FP67_0	0.00	0.00	92.86	< 0.001	-3.80
FP99=FP99_0	0.00	0.00	94.64	< 0.001	-4.13

flufenamic acid, were subsets in bicluster 4 based on BiMax

Overall, based the chemical compounds, some clusters for MCA and some biclusters BiMax were similar. However none of the fingerprints exhibited by the compounds were similar across the two methods for neither of the clusters nor biclusters.

3.4 Results of iBBiGs on Fingerprints (Table 2)

Table 30 shows all the biclusters obtained using the iBBiG method. The results indicated that the compounds in bicluster 4: nocodazole,4,5-dianilinophthalimide , imatinib, diclofenac, phenyl, biguanide exhibiting the fingerprints: FP47, FP75, FP105 ,FP138, FP140, FP150, FP171 and FP215 constitute a

Table 29: Biclusters obtained using the BiMax algorithm

Bicluster	Score	Fingerprints	Compounds
1	3.53	FP76, FP83 ,FP127, FP132	estradiol, verapamil, raloxifene, nordihydroguaiaretic acid, celecoxib, LM-1685, SC-58125 indometacin, MK-886, genistein, sulindac, exisulind, fulvestrant, flufenamic acid, imatinib, clozapine, thioridazine, chlorpromazine, trifluoperazine, W-13, prochlorperazine, calmidazolium, fluphenazine, dopamine, resveratrol, butein, nocodazole, 4,5-dianilinophthalimide
2	3.47	FP76, FP83, FP132, FP211	rofecoxib, raloxifene, celecoxib, LM-1685, SC-58125, tomelukast, LY-294002, indometacin, MK-886, genistein, sulindac, exisulind, fulvestrant, staurosporine, flufenamic acid, N-phenylanthranilic acid, trichostatin A, diclofenac, fasudil, imatinib, haloperidol, W-13, butein, nocodazole, 4,5-dianilinophthalimide, benserazide
3	3.46	FP76, FP83, FP127, FP132, FP211	raloxifene, celecoxib, LM-1685 ,SC-58125, indometacin, MK-886 ,genistein, sulindac, exisulind, fulvestrant, flufenamic acid, imatinib, W-13 , butein, nocodazole, 4,5-dianilinophthalimide
4	3.45	FP76 ,FP83, FP211, FP229	dexamethasone, exemestane, rofecoxib , 15-delta prostaglandin J2, raloxifene, LM-1685, tomelukast ,LY-294002 ,ciclosporin , indometacin ,MK-886 ,prednisolone, genistein, fludrocortisone, sulindac ,exisulind ,staurosporine, flufenamic acid, N-phenylanthranilic acid , trichostatin A, diclofenac, imatinib , haloperidol, bucladesine ,butein, nocodazole, 4,5-dianilinophthalimide ,benserazide
5	3.38	FP83, FP211, FP218, FP229	dexamethasone, 15-delta prostaglandin J2, raloxifene, tomelukast, ciclosporin , indometacin, MK-886 ,prednisolone , genistein, fludrocortisone ,sulindac exisulind, flufenamic acid, N-phenylanthranilic acid, trichostatin A ,diclofenac haloperidol, arachidonic acid , bucladesine, butein, benserazide

bicluster which is similar to the bicluster obtained by cluster 5 of the the MCA method. In fact all the compounds obtained in cluster 5 of the MCA form a subset of those obtained by the IBBiG algorithm. However, only a subset (FP215, FP105, FP140, FP47) of the fingerprints exhibited by cluster 4 of iBBiG were found based on cluster 5 of the MCA. In addition, cluster 1 of the MCA and cluster 2 of IBBiG were similar with respect to the compounds in each bicluster but only fairly agreed on the fingerprints (only FP18 and FP4 were found across the two methods). Finally cluster 1 and 3 of iBBiGs has similar bicluster contents with respect to the compounds present as those of biclusters 4 and 6 of the MCA respectively. Their fingerprints did not match.

3.5 FABIA on Fingerprints (Table 2)

For interpreting the biclusters and ranking their components, two different scores were used: bixv and biypv. bixv gives the values of the extracted compounds that have absolute values above a threshold whereas biypv gives the values of the extracted fingerprints that have absolute values above a threshold. Table 31 shows the results of 4 biclusters obtained using the FABIA method.

Table 30: IBBiG results

Factor	Score	Fingerprints	Compounds
1	203.9	FP76 , FP83, FP102, FP108 FP127 FP229, FP132 FP211 FP218	raloxifene, LM-1685 , tomelukast, indometacin, MK-886, sulindac, exisulind, fulvestrant, imatinib
2	47.3	FP18 , FP24, FP75 , FP89 FP111 ,FP126 ,FP144 , FP91 FP95 , FP178, FP185, FP176 FP232, FP196, FP219, FP230 ,FP231	dexamethasone, prednisolone ,fludrocortisone
3	47.2	FP76 , FP79 , FP83, FP102, FP132, FP239, FP240 , FP138, FP150, FP167, FP201	thioridazine, trifluoperazine, prochlorperazine
4	36.3	FP47, FP75, FP105 ,FP138 , FP140 ,FP150 ,FP171, FP215	flufenamic acid N-phenylanthranilic acid , nocodazole ,4,5-dianilinophthalimide , imatinib, diclofenac, phenyl, biguanide
5	33.3	FP8 , FP46, FP67 , FP92 , FP96, FP98 , FP99 ,FP101 ,FP104, FP107 ,FP129 ,FP139, FP151, FP184, FP188 ,FP200 ,FP227	trifluoperazine, prochlorperazine, fluphenazine

Table 31: FABIA results

Factor	Score	Fingerprints	Compounds
1	670	FP102 FP171 FP228 FP243 FP177 FP103 FP20	tetraethylenepentamine, W-13
2	521	FP161 FP172 FP140 FP76 FP83 FP241 FP244 FP55 FP220 FP177	tioguanine
3	501	FP171 FP132 FP150 FP138 FP140 FP83 FP105 FP76 FP47 FP81 FP229 FP74 FP75 FP157 FP245 FP215 FP211 FP145 FP72 FP218 FP192 FP44 FP225	N-phenylanthranilic acid, 4,5-dianilinophthalimide, phenyl, biguanide , diclofenac, and flufenamic acid
4	488	FP165 FP75 FP160 FP79 FP167 FP220 FP108 FP78 FP177	metformin

3.6 Factor analysis on Fingerprints (Table 2)

The very simple structure (VSS) procedure indicated that 8 factors should be retained for this analysis. The results of the first four factors are as presented on Table 32. All compounds with absolute scores $\geq$10 and all fingerprints with weights$\geq$0.7 were considered as the most linked to the factors. The results obtained appear to be quite similar to those of MCA and iBBiG in terms of the content of the biclusters. In addition cluster 4 was similar to cluster 5 of MCA method, cluster 4 of iBBiG and cluster 3 of FABIA.

Table 32: Factor analysis results

Factor	Score	Fingerprints	Compounds
1	52.74	FP10, FP12 ,FP14 , FP25 ,FP39 ,FP40 ,FP93	phenyl, biguanide , raloxifene , LM-1685 , MK-886 , genistein , trichostatin A , valproic acid , imatinib W-13 butein, tioguanine,
2	23.61	FP18, FP24 , FP77, FP91 FP120, FP121, FP126, FP219, FP230 ,FP231, FP176, FP185, FP237, FP111, FP232	phenyl, biguanide , estradiol, dexamethasone, exemestane ,15-delta prostaglandin J2 , LM-1685, indometacin , prednisolone , fludrocortison fulvestrant , fasudil , imatini clozapine , chlorpromazine , W-13, nocodazole , 4,5-dianilinophthalimide
3	19.87	FP8 , FP46 , FP67 ,FP139 , FP151 ,FP188 ,FP200 ,FP227 , FP98 , FP99 ,FP101, FP129, FP92 , FP239 ,FP96	verapamil, 15-delta prostaglandin J2 , sulindac exisulind , valproic acid , imatinib clozapine , thioridazine , chlorpromazine trifluoperazine , arachidonic acid , prochlorperazine fluphenazine , bucladesine , butein,benserazide
4	19.04	FP47 FP105 FP140 FP145 ,FP157 ,FP215, FP245	phenyl biguanide , estradiol , verapamil, raloxifene , fulvestrant , flufenamic acid, N-phenylanthranilic acid , diclofenac , imatinib 4,5-dianilinophthalimide

4 Discussion and Conclusions

The availability of high throughput technologies such as microarrays and next generation sequencing is making it possible to cheaply and rapidly collect large amounts of drug-gene expression datasets. Combining compounds and their characteristics with gene expression data is called connectivity mapping and holds promise for in-depth analysis and understanding of biological processes, discovery of new drug targets and new drugs and prediction of the toxic potential of unknown compounds. These goals can be achieved using the connectivity map data base and using appropriate statistical analysis methods.

The first objective of this study was to identify groups of compounds with defined fingerprints that co-regulate groups of genes using the CMap database. This was achieved using the MFA method developed by Escofier and Pagès (1990) and used in many different studies (Abdi et al. 2013, de Tayrac et al. 2009). The MFA revealed groups of compounds with or without particular fingerprints that co-regulated groups of genes and which are all linked to latent factors (dimensions). The results of the MFA were beefed with those of hierarchical clustering analysis which also indicated clusters of compounds with particular chemical structures inducing activity in groups of genes. Regarding the MFA, it is worth noting that the low proportions of variance explained by the first few principal components of the global MFA analyses is related to the heterogeneity of the matrices for the fingerprints that were used, in which the proportions of some of the fingerprints were too low. This is why the MFA results should be combined with those of a hierarchical cluster analysis to improve performance and viability of the MFA (Cardillo Alberti, 2013). A limitation of stopping the analysis at this stage is that the pathways associated with the groups of genes were not identified and thus the interpretation of the results is limiting. It is recommended to combine the results of this method with those of a pathway analysis (Subramanian et al. 2005, de Tayrac, 2009) to identify biological processes involved.

The second objective of this study was to use only the fingerprint matrix and apply several biclustering methods and determine if they identify similar groups of compounds associated with similar fingerprint structures. This goal was achieved using different techniques under the theme of biclustering. The

five different methods applied are : Multiple correspondence analysis (MCA), Binary inclusion-maximal biclustering (Bimax) algorithm, Factor analysis for Bi-cluster acquisition (FABIA), Iterative Binary biclustering of gene sets and Factor analysis for binary data.

The MCA identified 6 clusters based on the hierarchical clustering analysis containing compounds with or without a group of fingerprints. A subset of the large clusters were identified as those closest to the center of the clusters and thus more representative of the clusters. Based on the results of BiMAx, several compounds and fingerprints were found to be common to the top 4 biclusters namely: raloxifene, MK-888, butein, genistein, sulindac and N-phenylanthranilic acid and for the fingerprints: FP76, FP83 and FP132. It can thus be concluded that these compounds and fingerprints form a true bicluster. Overall, both MCA and BiMax reported similar compounds in their clusters and biclusters respectively but there was a complete mismatch with respect to the fingerprints. The disparities are pronounced probably because of the sparse nature of the fingerprint data set that was used. Since MCA has a more robust option to spareness, the results were considered to be more reliable compared to those of BiMax.

All the compounds obtained in cluster 5 of the MCA form a subset of those obtained by cluster 4 of the the iBBiG algorithm. However, only a subset (FP215, FP105, FP140, FP47) of the fingerprints exhibited by cluster 4 of iBBiG were found based on cluster 5 of the MCA. In addition, cluster 1 of the MCA and cluster 2 of IBBiG were similar with respect to the compounds in each bicluster but only fairly agreed on the fingerprints (only FP18 and FP4 were found across the two methods). Finally cluster 1 and 3 of iBBiGs have similar bicluster contents with respect to the compounds present as those of biclusters 4 and 6 of the MCA respectively. However, their fingerprints did not match. These two methods are the most robust in terms of their ability to work on sparse binary data and thus the similarities in their results. In all the clusters obtained with BiMAx, subsets were also found by cluster 4 of iBBiG. However, the fingerprints did not match at all.

FABIA performed fairly well in terms of the biclusters identified as the results for cluster 3 matched those of cluster 5 using MCA and biclusters 4 using Factor analysis for binary data and bicluster 4 using iBBiG. However, the results of FABIA were not over interpreted given that the method was initially designed for continues data.

The results of the factor analysis for binary data were quite similar to those of MCA and iBBiG in terms of the content of the biclusters. The method is robust to sparse data as it works with polychoric correlations instead of the raw data.

The performance of the biclustering methods applied in this study appeared to be similar except for BiMax and FABIA. MCA, iBBiGS and Factor analysis for binary data yielded very similarly results on most of the biclusters. Given the different approaches to biclustering, the methods all identified the compounds: 4,5-dianilinophthalimide, N-phenylanthranilic acid, flufenamic acid, phenyl and biguanide except for BiMax which only identified 4,5-dianilinophthalimide, N-phenylanthranilic acid and flufenamic acid. In addition, these group of compounds were found to exhibit the fingerprints: FP47, FP105, FP140 and FP215. These were consistently present in these compounds across the different methods except for BiMax where none of the fingerprints featured. The poor performance of BiMax could be attributed to the sparsity in the data whereas FABIA is not intended for binary data and should be cautiously interpreted.

Since only a subset of the CMAP data was analyzed, it is obvious that the results presented in this study do not cover the complete picture of patterns that exist in the CMAP database. Future similar studies based on the full CMAP database are therefore needed to fully appreciate the findings of this study.

In conclusion, for exploring local patterns, no one method could be judged superior over the others as evidenced in the literature (Prelić et al. 2006, Oghabian et al. 2014). However, for sparse binary data like the one we presented in this study, a combination of the results from the three methods: factor analysis for binary data, multiple correspondence analysis (MCA) and Iterative Binary biclustering of gene sets (IBBiGs) will be the most optimal approach. In addition, for a combination of groups of variables (quantitative and qualitative), the multiple factor analysis (MFA) combined with hierarchical clustering should be used. Finally, even though the different data sources can be explored independently, an integrated analysis is encouraged since it can reveal patterns or features of interest that may not be detected when analyzing the data sources one-by-one.

5 References

- Justin Lamb, Emily D. Crawford, David Peck, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006 Sep 29;313(5795):1929-3
- Juuso A Parkkinen and Samuel Kaski. Probabilistic drug connectivity mapping. BMC Bioinformatics 2014 15:113
- Iorio F, Bosotti R, Scacheri E, et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. PNAS 107(33): 14621-14626.
- Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol. 2000;8:93-103.
- Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006 Volume 22, Issue 9Pp. 1122-1129
- de Tayrac M, Lê S, Aubry M, Mosser J, Husson F. BMC Genomics.Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. 2009 Jan 20;10:32. doi: 10.1186/1471-2164-10-32.
- Iskar, M., Campillos, M., Kuhn, M., Jensen, L. J., van Noort, V., & Bork, P.(2010). Drug-Induced Regulation of Target Expression. PLoS Computational Biology 6:9
- Laenen G, Thorrez L, Börnigen D, Moreau Y. (2013). Finding the targets of a drug by integration of gene expression data with a protein interaction network. Mol BioSyst 9: 1676-1685.
- Escofier, B. and Pagès, J. (1990). Analyses factorielles simples et multiples: objectifs, méthodes, interprétation. Dunod, Paris.
- Kemal Eren, Mehmet Deveci, Onur Küçüktunç, and Ümit V. Çatalyürek. A comparative analysis of biclustering algorithms for gene expression data Brief Bioinform (2013) 14 (3): 279-292
- Bécue-Bertaut M, Pagès J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis Volume 52, Issue 6, 20 February 2008, Pages 3255–3268*
- Lebart, L. (1994). Complementary use of correspondence analysis and cluster analysis. In Greenacre, M. and Blasius, J., editors, Correspondence Analysis in the Social Sciences, pages 162–178. Academic Press.
- Greenacre, Michael and Pardo, Rafael, Multiple Correspondence Analysis of a Subset of Response Categories (November 2005). Available at SSRN: http://ssrn.com/abstract=847647 or http://dx.doi.org/10.2139/ssrn.847647
- Sebastian Kaiser and Friedrich Leisch (2008). A toolbox for bicluster analysis in R. In Paula Brito, editor, Compstat 2008–Proceedings in Computational Statistics, pages 201-208. Physica Verlag, Heidelberg, Germany.
- Brief Bioinform. 2013 May;14(3):279-92. doi: 10.1093/bib/bbs032. Epub 2012 Jul 6. A comparative analysis of biclustering algorithms for gene expression data. Eren K1, Deveci M, Küçüktunç O, Çatalyürek ÜV.
- Gusenleitner D, Howe EA, Bentink S, Quackenbush J, Culhane AC. iBBiG:Iterative binary bi-clustering of gene sets. Bioinformatics 2012; epub Jul 12

- Daniel Gusenleitner and Aedin Culhane (2011). iBBiG: Iterative Binary Biclustering of Genesets. R package version 1.8.0. http://bcb.dfci.harvard.edu/ aedin/publications/
- Kaufman L, Rousseeuw: Finding groups in data: An introduction to cluster analysis. Wiley, New York, 1990.
- EisenMB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998, 95:14863-14868.
- Talloen W1, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Göhlmann HW. I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. Bioinformatics. 2007;23:2897–2902.
- Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., ... & Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 22(9), 1122-1129.
- Bartholomew D. J. Factor Analysis for Categorical Data. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 42, No. 3 (1980), pp. 293-321
- Revelle, W. (2012). psych: Procedures for Personality and Psychological Research. North-western University, Evanston. R package version 1.2.8
- Revelle, W., & Rocklin, T. (1979). Very simple structure-alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research, 14(4), pp. 403-414
- Chia, B. K. H. and Karuturi, R. K. M. (2010) Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. Algorithms for Molecular Biology, 5, 23
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
- Abdi, H., Williams, L. J. and Valentin, D. (2013), Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comp Stat, 5: 149–179. doi: 10.1002/wics.1246
- Marcelo Cardillo and Jimena Alberti, "Stone Tool Manufacture Strategies and Lithic Raw Material Exploitation in Coastal Patagonia, Argentina: A Multivariate Approach," Journal of Archaeology, vol. 2013, Article ID 128470, 12 pages, 2013. doi:10.1155/2013/128470
- Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–15550
- Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E (2014) Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. PLoS ONE 9(3): e90801. doi:10.1371/journal.pone.0090801

Appendix

```
#########################MFA
setwd("C:\\Users\\enjiabatih\\Documents\\MAsters in BioInformatics\\2014 Second\\AGE")
cmap <- load(file = "Emmanuel1.Rdata")
cmap

##fingerprintMat<-as.matrix(fingerprintMat)
##fingerprintMat<- ifelse(fingerprintMat=="TRUE",1,0)



head(geneMat)
head(fingerprintMat)

###"Filtering Genes
gene.exprs <- new("ExpressionSet", exprs=geneMat)
selEset <-filterVarInt(object = gene.exprs,IntCutOff = log(0.9), IntPropSamples = 0.9, VarCutOff = 0.05)
propSelGenes <- round((dim(selEset)[1]/dim(gene.exprs )[1])*100,1)
propSelGenes

Sel.geneMat<-exprs(selEset)
head(Sel.geneMat)



######Fingerprint Matrix
namesfinger<-as.vector(paste("FP", 1:250, sep = ""))
fingerprintMat1<-as.data.frame(fingerprintMat)
names(fingerprintMat1)<-namesfinger
##names(fingerprintMat1)<-namesfinger
fingerprintMat3<- ifelse(fingerprintMat1=="TRUE",1,0)

####check which of the tegets is present or absent for all compounds
for (i in 1:250){
print(sum(fingerprintMat3[,i]))
}

vectorp<-matrix(1:250,250,2)
  for (j in 1:250){
    vectorp[j,2]<-sum(fingerprintMat3[,j])
  }
```

```
vectorp3<-subset(vectorp,vectorp[,2]>5)

for (i in 1:ncol(fingerprintMat3))
{fingerprintMat3[1,i]<-as.character(fingerprintMat3[1,i])}
str(fingerprintMat3)
dim(mergeddata2)
xtable(mergeddata2[1:10,c(1:4,630:632)])

mergeddata2<-data.frame(t(Sel.geneMat),fingerprintMat3[,-1])
head(mergeddata2)
names(mergeddata2)

resmfa <- MFA(mergeddata2, group=c(nrow(Sel.geneMat),
ncol(fingerprintMat3[,-1])),type=c("s","n"),name.group=c("Gene","Fingerprt"), ncp=5)
summary(resmfa )

resultsdesc<-dimdesc(resmfa )

###Find lead compounds and lead genes

leadcpds3<-which(resmfa$ind$contrib[,1]>mean(resmfa$ind$contrib[,1]))
leadcpds3
mean(resmfa$ind$contrib[,3])

leadcpds3<-which(resmfa$ind$contrib[,1]>3)
leadcpds3

summary(round(resmfa$ind$contrib[,1],3))



data0<-as.matrix(geneMat[,leadcpds3])
data1<-data0[genelist1, ]

rownames(data1)
names(data1)
library(ggplot2)
library("reshape2")
data2 <- melt(t(data1), id.vars=rownames(t(data1)), value.name="value")
names(data2)<-c("Var1","Genes","value")
ggplot(data=data2 , aes(x=Var1, y=value, group = Genes,colour=Genes)) +
    geom_line() +
```

```
   xlab("Compounds") +
    ylab("Log2 Concentration") +
    geom_point( size=1, shape=21, fill="black")+theme_bw() +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))



##Find genes highly correlated with factor 2
hicorrgenes3<-which(resultsdesc$Dim.2$quanti[,1]>=0.51|resultsdesc$Dim.2$quanti[,1]<=-0.34)
hicorrgenes3
genelist2<-as.character(names(hicorrgenes3))
c("PPP2R1A",  "COL6A3",    "PPAN",   "PRTN3",    "SRGN",  "AKR1C1",     "MSN")

###Find lead compounds and lead genes

leadcpds3<-which(resmfa$ind$contrib[,2]>mean(resmfa$ind$contrib[,2]))
leadcpds3

data0<-as.matrix(geneMat[,leadcpds3])
data1<-data0[genelist2, ]

library(ggplot2)
library("reshape2")
data2 <- melt(t(data1), id.vars=rownames(t(data1)), value.name="value")
names(data2)<-c("Var1","Genes","value")
ggplot(data=data2 , aes(x=Var1, y=value, group = Genes,colour=Genes)) +
    geom_line() +
   xlab("Compounds") +
    ylab("Log2 Concentration") +
    geom_point( size=1, shape=21, fill="black")+theme_bw() +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))


#########################MCA
cmap <- load(file = "Emmanuel1.Rdata")
cmap
##fingerprintMat<-as.matrix(fingerprintMat)
##fingerprintMat<- ifelse(fingerprintMat=="TRUE",1,0)
head(geneMat)
head(fingerprintMat)
######Fingerprint Matrix
namesfinger<-as.vector(paste("FP", 1:250, sep = ""))
```

```
fingerprintMat1<-as.data.frame(fingerprintMat)
names(fingerprintMat1)<-namesfinger
##names(fingerprintMat1)<-namesfinger
fingerprintMat3<- ifelse(fingerprintMat1=="TRUE",1,0)
####check which of the tegets is present or absent for all compounds
for (i in 1:250){
print(sum(fingerprintMat3[,i]))
}
for (i in 1:ncol(fingerprintMat3))
{fingerprintMat3[1,i]<-as.character(fingerprintMat3[1,i])}
str(fingerprintMat3)
####MCA
###Binary data must first be categorized
for (i in 1:ncol(fingerprintMat3))
{fingerprintMat3[1,i]<-as.character(fingerprintMat3[1,i])}
str(fingerprintMat3)
res.mcafg<- MCA(fingerprintMat3[,-1])
desresults<-dimdesc(res.mcafg)
res.mcafg$var
leadcpds1<-which(res.mcafg$ind$contrib[,1]>mean(res.mcafg$ind$contrib[,1]))
leadcpds1
leadcpds2<-which(res.mcafg$ind$contrib[,2]>mean(res.mcafg$ind$contrib[,2]))
leadcpds2
resultsfin<-dimdesc(res.mcafg)
clustermod<-HCPC(res.mcafg)
###Cat variables for cluster 1
indices1<-names(which(clustermod$desc.var$category$`1`[,1]>=100))
catvar1<-clustermod$desc.var$category$`1`[indices1,]
xtable(clustermod$desc.var$category$`1`)
###Cat variables for cluster2
indices1<-names(which(clustermod$desc.var$category$`2`[,1]>=100))
catvar1<-clustermod$desc.var$category$`2`[indices1,]
xtable(clustermod$desc.var$category$`2`)
###Cat variables for cluster 3
indices1<-names(which(clustermod$desc.var$category$`3`[,1]>=100))
catvar1<-clustermod$desc.var$category$`3`[indices1,]
xtable(clustermod$desc.var$category$`3`)
###Cat variables for cluster 4
indices1<-names(which(clustermod$desc.var$category$`4`[,1]>=100))
catvar1<-clustermod$desc.var$category$`4`[indices1,]
xtable(clustermod$desc.var$category$`4`)
###Cat variables for cluster 5
```

```
indices1<-names(which(clustermod$desc.var$category$`5`[,1]>=100))
catvar1<-clustermod$desc.var$category$`5`[indices1,]
xtable(clustermod$desc.var$category$`5`)
###Cat variables for cluster 6
indices1<-names(which(clustermod$desc.var$category$`6`[,1]>=100))
catvar1<-clustermod$desc.var$category$`6`[indices1,]
xtable(clustermod$desc.var$category$`6`)

########################BIMAX
##install.packages("biclust")
library(biclust)
res.biclustfg <- biclust(x=fingerprintMat3[,-1], method=BCBimax(), minr=4, minc=4, number=10)
res.biclustfg
##biclustbarchart(fingerprintMat3[,-1],res.biclustfg, col="#A3E0D8")
###"heatmapBC(x = fingerprintMat3[,-1], res.biclustfg)
bubbleplot(fingerprintMat3[,-1], res.biclustfg, showLabels=TRUE)
#################SBSCORE to rank Clusters
sb1<-ChiaKaruturi(fingerprintMat3[,-1], res.biclustfg,1)
csb1<-c(1,sb1$SBscore)
sb2<-ChiaKaruturi(fingerprintMat3[,-1], res.biclustfg, 2)
csb2<-c(2,sb2$SBscore)
sb3<-ChiaKaruturi(fingerprintMat3[,-1], res.biclustfg, 3)
csb3<-c(3,sb3$SBscore)
sb4<-ChiaKaruturi(fingerprintMat3[,-1], res.biclustfg,4)
csb4<-c(4,sb4$SBscore)
sb5<-ChiaKaruturi(fingerprintMat3[,-1], res.biclustfg, 5)
csb5<-c(5,sb5$SBscore)
###ranking of the biclusters, the higher the sb score the better
rankbs<-rbind(csb1,csb2,csb3,csb4,csb5)
rankbs[order(rankbs[,2]),]
####save and open biclust results
writeBiclusterResults("results.txt", res.biclustfg,"Bimax", dimnames(fingerprintMat3[,-1])[1][[1]],
dimnames(fingerprintMat3[,-1])[2][[1]])

##################IBBIgs
library(iBBiG)
install.packages("clValid")
resfgm<- iBBiG(fingerprintMat3[,-1],nModules=10)
plot(resfgm)
statClust(resfgm)
summary(resfgm)
```

```
order(resfgm@Clusterscores)

par(mfrow=c(1,1))
drawHeatmap(fingerprintMat3[,-1], resfgm, number=1)
par(mfrow=c(1,1))
drawHeatmap(fingerprintMat3[,-1], resfgm, number=2)
par(mfrow=c(1,1))
drawHeatmap(fingerprintMat3[,-1], resfgm, number=9)
which(NumberxCol(resfgm)[1,]==TRUE)
which(RowScorexNumber(resfgm)[,1]>0)
sort(RowScorexNumber(resfgm)[,1])
which(NumberxCol(resfgm)[2,]==TRUE)
which(RowScorexNumber(resfgm)[,2]>0)
###bicluster 3
which(NumberxCol(resfgm)[3,]==TRUE)
which(RowScorexNumber(resfgm)[,3]>0)
sort(RowScorexNumber(resfgm[,3]))
###bicluster 4
which(NumberxCol(resfgm)[4,]==TRUE)
which(RowScorexNumber(resfgm)[,4]>0)
sort(RowScorexNumber(resfgm)[,4])
###bicluster 9
which(NumberxCol(resfgm)[9,]==TRUE)
which(RowScorexNumber(resfgm)[,9]>0)
sort(RowScorexNumber(resfgm)[,9])
par(mfrow=c(2,1))
drawHeatmap(fingerprintMat3[,-1], resfgm, number=2)
drawHeatmap(fingerprintMat3[,-1], resfgm, number=3)
#drawHeatmap(fingerprintMat3[,-1], resfgm, number=3)
#drawHeatmap(fingerprintMat3[,-1], resfgm, number=4)
par(mfrow=c(1,1))


#####FABIA
library(fabia)
resgenefing <- fabia(fingerprintMat3[,-1],25,0.01,5000)
show(resgenefing)
myextractPlot(resgenefing ,ti='FABIA',mergeddata2,which=2)
myextractPlot(resgenefing ,ti='FABIA',mergeddata2,which=3)
myextractPlot(resgenefing ,ti='FABIA',mergeddata2,which=4)
myextractPlot(resgenefing ,ti='FABIA',mergeddata2,which=5)
myextractPlot(resgenefing ,ti='FABIA',fmergeddata2,which=6)
```

```
####Fabia plot with subgroups
plot(resgenefing ,dim=c(2,4),label.tol=0.03,lab.size=0.8)

head(mergeddata2)
extractPlot(resgenefing )
resgenefing1<- extractBic(resgenefing)
###showw info content of billusters
resgenefing@avini
#####First Bicluste
resgenefing1$bic[1,]
plot(1:12,abs(resgenefing1$bic[1,]$bixv),xlab="" )
axis(1, at=1:12, labels=resgenefing1$bic[1,]$bixn)

xtable(data.frame(resgenefing1$bic[1,]$bixv))
xtable(data.frame(resgenefing1$bic[1,]$biypv))
#####Second Bicluster
resgenefing1$bic[2,]
xtable(data.frame(resgenefing1$bic[2,]$bixv))
xtable(data.frame(resgenefing1$bic[2,]$biypv))

#####Third Bicluster
resgenefing1$bic[3,]
xtable(data.frame(resgenefing1$bic[3,]$bixv))
xtable(data.frame(resgenefing1$bic[3,]$biypv))

#####Third Bicluster
resgenefing1$bic[4,]
xtable(data.frame(resgenefing1$bic[4,]$bixv))
xtable(data.frame(resgenefing1$bic[4,]$biypv))
##########Plotr bicuster 1
plotBicluster(resgenefing1,1,opp=TRUE)

####Factor Analysis
library(polycor)
library(psych)

######Fingerprint Matrix
namesfinger<-as.vector(paste("FP", 1:250, sep = ""))
fingerprintMat1<-as.data.frame(fingerprintMat)
names(fingerprintMat1)<-namesfinger
##names(fingerprintMat1)<-namesfinger
fingerprintMat3<- ifelse(fingerprintMat1=="TRUE",1,0)
```

```
faPCdirect <- fa.poly(fingerprintMat3[,-1], nfactors=8, rotate="varimax")


####determinie number of factors:
polycorr<-polychoric(fingerprintMat3[,-1])
vss(polycorr$rho)


########Fingerprints tied to factor 1:4

which(abs(as.matrix(faPCdirect$scores$weights)[,1])>0.7)
which(abs(as.matrix(faPCdirect$scores$weights)[,2])>0.7)
which(abs(as.matrix(faPCdirect$scores$weights)[,3])>0.7)
which(abs(as.matrix(faPCdirect$scores$weights)[,4])>0.7)


########Compounds tied to factors 1:4
compds<-faPCdirect$scores$scores
compds2<-as.data.frame(compds)
which(abs(compds[,1])>1)
which(abs(compds[,2])>1)
which(abs(compds[,3])>1)
which(abs(compds[,4])>1)


#####Plots
factor.plot(faPCdirect$fa,cut=0.7)
fa.diagram(faPCdirect,simple=TRUE,cut=0.6 )
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Exploring local patterns between gene expression profiles and chemical structures (fingerprints) of compounds

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Abatih, Emmanuel

Datum: **25/11/2014**