2012•2013
FACULTY OF SCIENCES
Master of Statistics: Bioinformatics

Masterproef

BFRank: An R package of fingerprints based ranking of gene expression biclusters in early drug development

Promotor :
Prof. dr. Ziv SHKEDY
De heer Martin OTAVA

Silenou Chawo

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

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

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Bioinformatics

Masterproef

BFRank: An R package of fingerprints based ranking of gene expression biclusters in early drug development

Promotor :
Prof. dr. Ziv SHKEDY
De heer Martin OTAVA

Silenou Chawo
Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

universiteit hasselt | Maastricht University

BF-RANK: AN R PACKAGE FOR FINGERPRINT BASED RANKING OF GENE EXPRESSION BICLUSTERS IN EARLY DRUG DEVELOPMENT.

Thesis submitted to the Center for Statistics, Hasselt University, Campus Diepenbeek, Belgium; in partial fulfilment of the requirement for the award of Master of Science in Statistical Bioinformatics.

September 2013

By

SILENOU BERNARD CHAWO

The University of Hasselt

Supervisor: Prof. Dr. Ziv SHKEDY

R-forg

2013

THE PUBLISHER

Certification

This is to certify that this project was carried out by Silenou Bernard Chawo under our thorough supervision and reflects his true research ability.

Silenou Bernard Chawo
Student

.....
Signature

Prof. Dr. Ziv SHKEDY
First Supervisor

.....
Signature

Mr Martin OTAVA
Second Supervisor

.....
Signature

Dedication

I dedicate my dissertation work to my family and many friends. A pecial feeling of gratitude to my loving parents, louis CHAWO and Rose KAMENI whose words of encouragement and fear for the creator always rings in my heart. My uncle Bernard SILENOU and my sisters Mary and Josephine who have never left my side and are very special to me.

I also dedicate this dissertation to my many friends who have supported me throughout the process. I will always appreciate all they have done, especially Geraldine AGBOR for not getting tired of directing and raising me up when I feel discouraged and downtrodden and Blaise NGANOU for the many hours of proofreading.

I dedicate this work and give special thanks to my best friend and spiritual director Christopher SUH, Jean FILBERT and my astounding sponsor Peter EGBE for being there for me throughout the entire masters program. I love you all.

Abstract

Background: Examining the relationship between a chemical structure and its biological function is of great importance for drug discovery. Researchers traditionally focused on the molecular structures in the context of interactions with targets in order to better understand the mechanisms of drug action. The invention of microarray has been of great help in this direction of research. So far studies have been conducted to introduce transcriptomic data into functional investigation but little effort has been made in looking at the relationship between structural fingerprints of compounds with defined intracellular functions.

Objective: In this article, we present **BF-Rank** an R package dedicated to biclustering of microarray data analysis. The main features of this package are the possibilities to use different algorithms to obtain biclusters and finally rank them based on similarity of the compounds in each bicluster. Moreover, ten different measures of similarity, calculated from the fingerprint information of the compounds can be implemented by this package. Finally, numerous graphics are also available with various options.

Methods: Factor Analysis for Bicluster Acquisition (FABIA), ISA and Plaid model where implemented in this package to identify biclusters. Measures of similarity of binary data such as Tanimoto, Ochiai, Gower & Legendre and Czekanowski were also implemented in this package to measure similarity between compounds. In addition, these measures of similarity were used to cluster compounds based on their fingerprints. For a practical application of this study, we present how this package can be used to rank biclusters and finally identify differentially expressed genes using SAM and LIMMA.

Index Terms : Biclustering, microarray data analysis, structural fingerprint, differential expression, False Discovery Rate (FDR), Microarray, Multiple testing.

Contents

1 Introduction

1.1 Background

Examining the relationship between a chemical structure and its biological function is of great priority for drug discovery. It has been reported that for every drug that was pulled from the market because of ADME/T (absorption, distribution, metabolism, elimination and toxicity) difficulties, 10 remained on the market with potential for drug-drug interactions, and for every one of those on the market there were another 10 to 50 that failed before they made it to the market place. Hence, intense interest has been focused in the development of computer based (*in silico*) methods to predict toxicity, efficacy, and adverse effect of drug candidates[1].

Past researches realized almost 150 years ago that relation exists between the physiological action of a substance and its chemical composition and constitution[1]. The challenge facing Chemists and Bioinformaticians today is how to accurately predict the biological activities of chemical compounds. This has been a mind-boggling situation but the invention of microarray has been of great help in this direction of research. So far studies have been conducted to introduce transcriptomic data into functional investigation but little effort has been made in looking at the relationship between structural fingerprints of compounds with defined intracellular functions. Researchers traditionally focused on the molecular structures in the context of interactions with targets in order to better understand the mechanisms of drug action. Questions often demanded are; which set of genes are associated with a condition (tumor) ? and which tumor class does a particular sample belongs ?.

Microarray technology is often used to measure expression levels associated with thousands of genes in a single experiment. The expression level recorded by the microarray is a characterization of the intracellular activities of chemical compounds and their mechanisms of action. The interesting issue about the data set from microarray is that the number of samples (conditions) are far smaller than the number of covariates (genes). As such, special statistical methodologies (like clustering) for microarray data together with some data mining technologies have been brought together to automatically tackle these large data sets. They do provide interesting results when applied to relatively "small" data sets, typically containing tens of experimental conditions and at most several hundred genes, but are of limited use for the analysis of large data sets [2]. In particular, a well-recognized drawback of commonly used clustering algorithms is the fact that they assign each gene to a single cluster, while in fact genes that participate in several functions should be included in multiple clusters. Moreover, genes are analyzed based on their expression under all experimental conditions in clustering methods. This is problematic, because cellular processes are usually affected only by a small subset of these conditions, such that most conditions do not contribute relevant information but rather increase the level of background noise in the data. Thus we need a further look at statistical methods that takes all of these draw backs into account. This calls for an extensive need for dimensional reduction methods like biclustering. Biclustering simultaneously cluster both the rows and columns in a matrix based on the definition of 'similarity'. There are a handful

of biclustering algorithms and they all differ in only how similarity is defined. Three of the most efficient biclustring algorithms; Factor Analysis For Bicluster Acquisition (FABIA), Plaid model and Iterative Signature Algorithm (ISA) were used in this report. Even though many studies have been conducted on biclustering methods, little effort has been made in associating this with the fingerprints of compounds found in biclusters. We found it interesting to carry out such a study by designing a package that will identify biclusters and rank them based on the fingerprints of the compounds.

1.2 Objectives

A gene expression data set (label B) and corresponding fingerprints data set were used in this project. In order to address the research interest, the following scientific questions were posed: Can we confidently rank biclusters based on the fingerprints of the compounds? It is possible to select a reasonable number of genes that are differentially expressed between two clusters of compounds predetermined by their fingerprints? To answer these scientific questions, the objectives of the study were formulated to first of all find biclusters base on the gene expression and device a method that can be used to rank them based on the fingerprints of the compounds using different measures of similarity. In addition, we are to cluster the compounds in say k groups based on their fingerprints information using different measures of association and finally use different statistical methods to identify top T genes that are differentially expressed between a chosen group and the others.

The remainder of this paper is organized as follows. Details of the data set used in this report is given in section 2. We present our methodology implemented in the BFRank package in section 3. We begin section 4 with a brief description of the clustering methods while section 5 presents how the package can be used. We draw conclusion in section 6, present our list of references in section 7 and attach the the manual of the package in section 8.

2 Data structure

A gene expression data (label B) and its corresponding fingerprint data set (label F) were used in this project. The expression data was got by applying 96 different compounds to cultured specimens and the expression levels of the genes measured. This resulted in an expression data of 7103 genes (rows) and 96 compounds (columns). Each chemical on the other hand has its fingerprint information which has a binary value (0 or 1) if the compound has a particular characteristic or not, there were in total 354 characteristics for each compound. figure 1 demonstrate how the fingerprints of a compound are represented from its chemical structure. The fingerprint information was organized in a data set where the fingerprints represent the rows and the columns, the compounds. The data structures are represented physically as shown by equations 1 and 2.

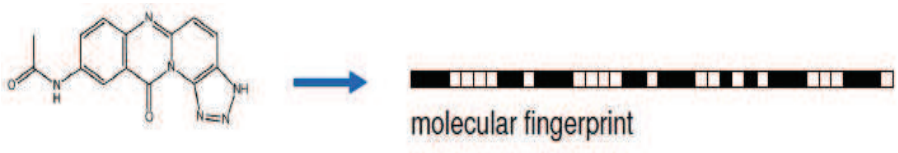

Figure 1: *Representation of molecular fingerprint* [14].

$$B = \begin{pmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,96} \\ \vdots & \vdots & \vdots & \vdots \\ g_{7103,1} & g_{7103,2} & \cdots & g_{7103,96} \end{pmatrix} \tag{1}$$

$$F = \begin{pmatrix} fp_{1,1} & fp_{1,2} & \cdots & fp_{1,96} \\ \vdots & \vdots & \vdots & \vdots \\ fp_{354,1} & fp_{354,2} & \cdots & fp_{354,96} \end{pmatrix} \tag{2}$$

g_{ij} stands for the i^{th} gene in the j^{th} compound while fp_{ij} represents the i^{th} chemacal structure in the j^{th} compound.

.

3 Ranking Based on Fingerprints

3.1 Biclustering

Biclustering, co-clustering, or two-mode clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix. Given a set of m rows in n columns (i.e. $m * n$ matrix), the biclustering algorithm generates biclusters, a subset of rows which exhibit similar behavior across a subset of columns, or vice versa. There are many biclustering algorithms existing and what is peculiar about these algorithms is the way biclusters are defined. These definitions fall basically into three main groups as shown in figure 2 :

- Bicluster with constant values (a),
- Bicluster with constant values on rows (b) or columns (c),
- Bicluster with coherent values (d, e).

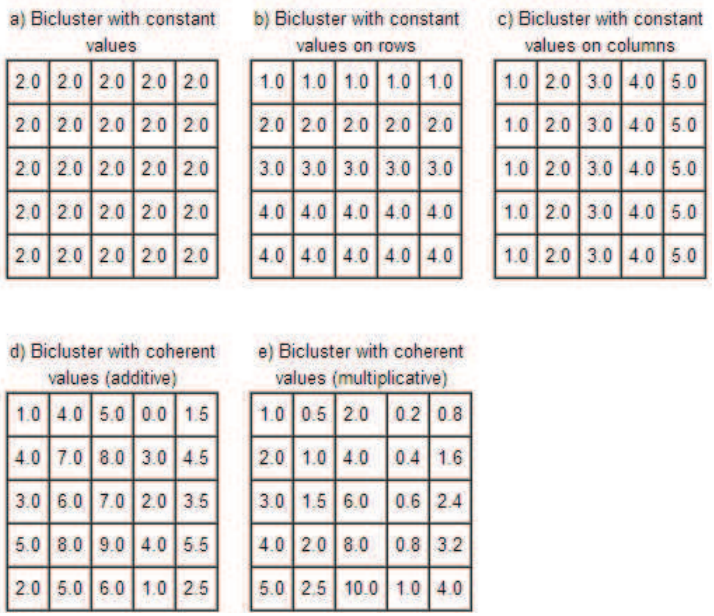

Figure 2: *Types of Biclusters* [10].

Three different biclustering algorithms were used in this study: Plaid Model, ISA (Iterative Signature Algorithm) and FABIA (Factor Analysis for Bicluster Acquisition).

3.1.1 FABIA Model

Factor Analysis for Bicluster Acquisition (FABIA) is a model-based technique for biclustering. It is based on a multiplicative model which accounts for linear dependencies between gene expression and condition, and also captures heavy-tail distributions as observed in real-world transcriptomic data. The model can be describe as follows.

It is a mathematical fact that two vectors are similar if one is a multiple of the other, that is the angle between them is zero or as realization of random variables their correlation coefficient is one. It is clear that such a linear dependency on subsets of rows and columns can be represented as an outer product λz^T of two vectors λ and z. The vector λ corresponds to a prototype column vector that contains zeros for genes not participating in the bicluster, whereas z is a vector of factors with which the prototype column vector is scaled for each sample; clearly z contains zeros for samples not participating in the bicluster. Vectors containing many zeros or values close to zero are called sparse vectors [13]. Figure 3 visualizes this representation by sparse vectors schematically.
The overall model for p biclusters and additive noise is:

$$X = \sum_{i=1}^{p} \lambda_i Z_i^T + \Upsilon = \Lambda Z + \Upsilon \tag{3}$$

Where $\Upsilon \in R^{n*l}$ is additive noise and $\lambda_i \in R^n$ and $z_i \in R^l$ are the sparse prototype vector and the sparse vector of factors of the i^{th} bicluster, respectively. The second formulation above holds if $\lambda_i \in R^l$ is the sparse prototype matrix containing the prototype vectors λ_i as columns and $Z \in R^{p*l}$ is the sparse factor matrix containing the transposed factors Z_i^T as rows.

3.1.2 Plaid model

Plaid model clustering does not suffer from the drawbacks of conventional clustering techniques such as preventing overlapping biclusters, thus making it a particularly attractive method for clustering microarray data. As a biclustering method, each gene cluster is associated with a sample cluster over which the genes are co-regulated, allowing for limited co-regulation. The biclusters represent unusual patterns of expression, so that uninterest-

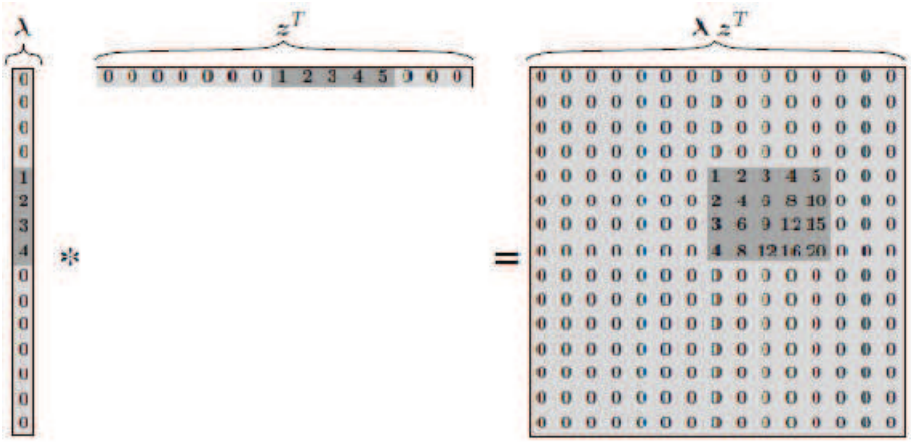

Figure 3: *The outer product λZ^T of two sparse vectors results in a matrix with a bicluster. Note that the non-zero entries in the vectors are adjacent to each other for visualization purposes only* [13].

ing expression profiles are not clustered. Genes involved in more than one active biological process can be accommodated through overlapping biclusters.
The plaid models for biclustering fit layers κ to the model

$$a_{ij} = (\mu_0 + \alpha_{ik} + \beta_{jk}) + \sum_{k=1}^{\kappa} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} + \epsilon_{ij} \tag{4}$$

using ordinary least squares (OLS), where α_{ij} is the value in the data matrix B corresponding to the row i and column j, μ, α, and β represent mean, row and column effects and ρ and κ identify if a row or column is member of the layer, respectively [12].

3.1.3 Iterative Signature Algorithm (ISA)

The iterative signature algorithm for bicluster is based on what is known as the transcription module (TM). In the context of gene expression data these modules consist of subsets of genes that exhibit a coherent expression profile only over a subset of microarray experiments. Genes and arrays may be attributed to multiple modules and the level of required coherence can be varied resulting in different "resolutions" of the modular mapping. The degree of coherence is determined by a pair of threshold parameters for the genes and conditions. The gene threshold contains the gene set, while the condition threshold contains the condition set. The goal of ISA algorithm is to search for modules incorporated in the data. Since the ISA does not rely on the computation of correlation matrices, it is extremely fast even for very large datasets [2].

3.1.4 Diagnostic of Biclusters

The Bicluster Diagnostics plots (BcDiag) package is a visualization technique, for profiling and summarizing Bicluster data, particularly for gene expression level data. Target data matrix is the bicluster genes (rows) and conditions (columns) versus clustered genes or conditions. Line plot were used to visualize the biclusters where genes or conditions in a bicluster are grouped and separated from the others.

After obtaining the biclusters from the expression matrix using the above algorithms, their corresponding fingerprint matrices were got for each bicluster. This gives an $m' * n'$ matrix say, where $m' = 354$ and n' corresponds to the number of compounds per bicluster. Based on these binary matrices, similarity between compounds from the same bicluster were calculated using 10 different measures of association.

3.2 Chemical structure similarity measure

The binary feature vector is one of the most common representations of patterns and measuring similarity and distance measures play an important role in many statistical problems such as clustering, classification, etc. Numerous binary similarity and distance measures have been proposed in various fields. Some measures have been identified as the same mathematically. We present ten different similarity methods that are included in

this package. These similarity coefficients are only applicable for a binary variable, and ranges from 0 to +1 (where +1 is the highest similarity).

Assuming there are two chemicals A and B say, with $a = n_{11}$ the number of structure features in both chemical A and B, $b = n_{10}$ the number of structure features in chemical A but not in B, $c = n_{01}$ the number of structure features in chemical B but not in A and $d = n_{00}$ the number of structure features neither in A nor in B. Then the following similarity measures can be defined: [11]

1. **Jaccard index (1901)**:
   S3 coefficient of Gower & Legendre $s_1 = a/(a + b + c)$
2. **Sockal & Michener index (1958)**:
   S4 coefficient of Gower & Legendre $s_2 = (a + d)/(a + b + c + d)$
3. **Sockal & Sneath(1963)**:
   S5 coefficient of Gower & Legendre $s_3 = a/(a + 2(b + c))$
4. **Rogers & Tanimoto (1960)**:
   S6 coefficient of Gower & Legendre $s_4 = (a + d)/(a + 2(b + c) + d)$
5. **Czekanowski (1913) or Sorensen (1948)**:
   S7 coefficient of Gower & Legendre $s_5 = 2a/(2a + b + c)$
6. **S9 index of Gower & Legendre (1986)**:
   $s_6 = (a - (b + c) + d)/(a + b + c + d)$
7. **Ochiai (1957)**:
   S12 coefficient of Gower & Legendre $s_7 = a/\sqrt{(a + b)(a + c)}$
8. **Sockal & Sneath (1963)**:
   S13 coefficient of Gower & Legendre $s_8 = ad/\sqrt{(a + b)(a + c)(d + b)(d + c)}$
9. **Phi of Pearson**:
   S14 coefficient of Gower & Legendre $s_9 = (ad - bc)/\sqrt{(a + b)(a + c)(d + b)(d + c)}$
10. **S2 coefficient of Gower & Legendre**:
    $s_{10} = a/(a + b + c + d)$

The more common features A and B shares, the greater each of these coefficients and on this account the more similar they are. Finally, a symmetric similarity matrix of say $u' * u'$ was generated, where every grid contains a coefficient indicating the similarity score of the two chemicals. The magnitude of u' depends on the number of compounds in the given bicluster. Based on this symmetric matrix of each biclusters, they can then be ranked by looking at informal measures like their cumulative distribution function, densities and heat maps. The more similar the compounds from a bicluster are, the lower the rank of the bicluster will be.

3.3 Ranking

The goal of this package was not only limited at looking for biclusters but to further rank the biclusters based on their fingerprint information. To achieve this, we found it important to judge from the cumulative distribution of each of the similarity matrices.

3.3.1 Cumulative Distribution Function CDF

The CDF describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x. It can be expressed analytically as the integral of its probability density function f_X as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

It can also be represented by: $F_X(x) = P(X \leq x)$, where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x. After obtaining the similarity scores for all the different biclusters, their CDF were derived and a cut off limit for x was then used to get their probabilies. The biclusters were then ranked starting from the one that gave the highest probability. High probability is an intuition that the compounds are more similar. We did not narrow our attention in looking only at the CDF but also considered the density and heat map for each similarity matrix.

.

4 Cluster Analysis and Differential Expression

This section is quite different from the earlier sections. We start by first clustering the compounds (in F data) using hierarchical clustering algorithm and based on the ten different measures of association defined in section 2.4.2. In the second part, we select a cluster of interest (selecting of this cluster is subjective and depends on the researcher) and find the top N differential expressed genes using LIMMA and SAM.

4.1 Cluster Analysis

Each similarity score s_i was converted to distance d_i using the equation $d_i = \sqrt{1 - s_i}$ [11]. The distance matrices together with hierarchical clustering can then be used to cluster the compounds in to k clusters for each of the methods mentioned above. k is to be specified by the user. Out of these ten different methods, a consistent cluster of size 20 compounds identified by all the ten methods was isolated and denoted as group A while the rest of the compounds were grouped as group B. The next step of the analysis was to identify and report the top N differential expressed genes between group A and B using different statistical methods.

4.2 Differential Expression

Let μ_{Ai} and μ_{Bi}, for $i \in (1, 2, ..., 1354)$ be the mean expressions for gene i in cluster A and B respectively. In order to distinguish genes that differ between both groups, we test the following null hypothesis for each gene:

$$H_{0i} : \mu_{Ai} = \mu_{Bi} \qquad versus \qquad H_{1i} : \mu_{Ai} \neq \mu_{Bi} \tag{5}$$

The statistical techniques used to justify our hypothesis are presented below.

4.2.1 Independent filtering

With high-dimensional data, variable-by-variable statistical testing is often used to select variables whose behavior differs across conditions. Such an approach requires adjustment for multiple testing, which can result in low statistical power. A two-stage approach that first filters variables by a criterion independent of the test statistic, and tests only variables which pass the filter, can provide higher power. In an application to microarray data, it was found that gene-by-gene filtering by overall variance followed by a t-test increased the number of discoveries by 50% [4].

This two-stage approach, the use of which need not be restricted to gene expression applications, assesses each variable on the basis of both a filter statistic (U^I) and a test statistic (U^{II}). Both statistics are required to exceed their respective cutoffs. Filtering can be specific or non specific. Specific filtering is based on a statistical comparison like t-test, ANOVA, Cox Model, while non specific filtering does not. Nonspecific filter statistics (U^I) include, for example, the overall variance and overall mean computed across all arrays, ignoring class label. Advantages of this dimension reduction is that the

clustering algorithms run faster, reduce the problems with multiple testing by restricting attention to active genes, and increase statistical power. Another obvious advantage is clarity of interpretation: the biological meaning of a cluster or gene pathway is more easily discerned if the results of an analysis do not include irrelevant and distracting genes. Non specific filtering was implemented thus:

Select g_i for which $variance(g_i) > \delta$, $i \in (1, 2, ..., 7103)$, where δ = overall variance computed across all arrays, ignoring class label. It is of interest to note that most pre-processing methods return expression estimates on the log-scale so before applying non-specific filtering, expression level estimates need to be un-logged.

4.2.2 Independent t-test

Unlike the fold change, the t-test takes in to account the variance of each gene. It provides a standardized estimate of differential expression according to the following formula:

$$t^* = \frac{\bar{x_1} - \bar{x_2}}{\hat{\sigma}_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6}$$

Where

$$\hat{\sigma}^2_{pool} = \frac{(n_1 - 1)\hat{\sigma}^2_1 + (n_2 - 1)\hat{\sigma}^2_2}{n_1 + n_2 - 2} \tag{7}$$

where $\hat{\sigma}^2_i$ and n_i are the variance and number of replicates for the i^{th} group, respectively. $\hat{\sigma}^2_{pool}$ is the weighted average of the two groups gene-specific variance. The associated probability under the null hypothesis is calculated by reference to the t distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus for a significance level α, we reject H_0 if $t^* > t(n_1 + n_2 - 2)$ [7].

4.2.3 Moderated t-test (LIMMA)

The moderated t-test use an Empirical Bayes method to estimate error associated with differential expression and use a statistical test identical in form to the independent t-statistic shown in equation 6. It uses a posterior variance $\hat{\sigma}^2$, in place of the pooled variance, σ^2_{pool}, in the t-test. Using Bayes rule, the posterior variance, $\hat{\sigma}^2$, for each gene becomes a combination of the observed gene-specific error and an estimate obtained from the prior distribution according to the following formula.

$$\tilde{\sigma}^2 = \frac{(d_g)\hat{\sigma}^2_g + (d_o)\hat{\sigma}^2_o}{d_g + d_o} \tag{8}$$

Where d_g are the degrees of freedom and $\hat{\sigma}^2_g$ are the gene specific variances obtained from the experimental data. d_o and $\hat{\sigma}^2_o$ are the prior degrees of freedom and variance estimates respectively. We are now going to look at how LIMMA estimate the parameters d_o and $\hat{\sigma}^2_o$.

LIMMA t-statistic

LIMMA is a general linear model that is more easily applied to microarray data. This method is based on a model where the variances of the residuals vary from gene to gene and are assumed to be drawn from a chi-square distribution.
The linear model is as follows:

$$E(Y_g = xa_g) \tag{9}$$

where Y_g is the expression summary values for each gene across all arrays, X is a design matrix of full column rank and a_g is the coefficient vector. The contrasts of coefficients that are of interest for a particular gene g are defined as $B_g = C^T a_g$. Although this approach is able to analyse any number of contrasts, we examine two sample comparisons only so B_g can be defined as the log fold change.
The contrast estimators, B_g, are assumed to be normal and the residual variances, s_g^2, are assumed to follow a scaled chi-square distribution. Under this hierarchical model the posterior variance, $\tilde{s}_g^2$, takes the form:

$$\tilde{s}_g^2 = \frac{(d_g)s_g^2 + (d_o)s_o^2}{d_g + d_o} \tag{10}$$

Where d_o and s_o are the prior degrees of freedom and variance and d_g and s_g are the experimental degrees of freedom and the sample variance for a particular gene g, respectively. Because we examine only two sample comparisons (A and B), d_g will always be equal to $n - 2$ where n is the total number of replicates. The two prior parameters, d_o and s_o, can be interpreted as the degrees of freedom and variance of the prior distribution respectively. The prior parameters are estimated by fitting the logged sample variances to a scaled F distribution.

The moderated t-statistic is defined by:

$$\tilde{t} = \frac{\hat{B}}{\tilde{s}_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{11}$$

The associated probability of the moderated t-statistic for the two sample case under the null hypothesis is calculated by reference to the t-distribution with $d_o + d_g$ degrees of freedom [7].

4.2.4 Significance Analysis of Microarrays (SAM)

In a microarray setting, resampling methods are often used in making inference. The main motivation is to avoid inference based on the asymptotic distribution of the test statistics, which, within the microarray setting, can be problematic because of either typically small sample sizes or because of departure from the assumption about the distribution of the response. Also, in some cases the asymptotic distribution of the test statistic may be unknown [9].

SAM is a resampling-based procedure, which uses permutations to approximate the null distribution of the test statistics. Thus, the choice of the test statistic does not have an effect on the SAM procedure. SAM accomplishes its task in three main steps: (1) the adjusted test statistic, (2) approximation of the distribution of the test statistic based on permutations, and (3) the control of the FDR.

The t-test statistic is modified in the SAM procedure as follows:

$$T^* = \frac{\bar{x_1} - \bar{x_2}}{\hat{\sigma}_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + s_0}$$

The constant s_0 is called the fudge factor and it is estimated as the percentile of the gene-wise standard errors that minimizes the coefficient of variation of the SAM test statistics. This modification is used to overcome bias for genes with expression difference (Between clusters variability) close to zero, which have large values of the test statistics due to small within clusters variances. Supposed the observed value of the test statistic is

$$T_0^{SAM} = \begin{pmatrix} T_1^0 \\ T_2^0 \\ \vdots \\ T_m^0 \end{pmatrix}$$

If B permutations are made, the SAM matrix of statistics will be

$$T^{SAM} = \begin{pmatrix} T_{11}^* & T_{12}^* & \dots & T_{1B}^* \\ T_{21}^* & T_{22}^* & \dots & T_{2B}^* \\ \vdots & \vdots & \vdots & \vdots \\ T_{m1}^* & T_{m2}^* & \dots & T_{mB}^* \end{pmatrix}$$

Where $1, 2, ..., m$ rows are the genes and $1, 2, ..., B$ columns are the permutations. Once this matrix is obtained, the columns are sorted and the mean of each row is computed to yield the expected statistics

$$\overline{T^{SAM}} = \begin{pmatrix} \bar{T_1} \\ \bar{T_2} \\ \vdots \\ \bar{T_m} \end{pmatrix}$$

To call a gene significant, the difference between the observed and expected values of the test statistic needs to be larger than a certain cut-off value λ. That is

$$|T_i^0 - \bar{T_i}| > \lambda \quad for \quad i \in (1, ..., m)$$

For a grid of λ values, the corresponding number of significant genes can be listed; at the same time, the number of false positives arising from any permutation matrix T^{SAM} is estimated. Under the null hypotheses, we expect that no differentially expressed genes

are present for each permutation. Consequently the median or 90 percentile number of false positives corresponding to λ can be obtained from permutation matrix. In this way, the FDR can be calculated for each value of λ and an acceptable value of λ can be chosen to control the FDR at the desired level [3]. The SAM Procedure also controls for false discovery rate (FDR) once the permutation matrix is obtained. Apart from automatically controlling for multiple testing, SAM has the strength that the null distribution is generated for all the genes at once by permuting the group labels, so that the correlation between test statistics of all the genes is preserved [3].

4.2.5 Controlling the FDR for multiple testing

The prevalent issue related to gene expression profiling is the adjustment for the large number of comparisons that need to be made. Multiple testing procedures controlling for the Family-wise Error Rate (FWER), that is the probability to reject erroneously at least one true null hypothesis, such as Bonferroni or Holm procedure, are conservative and lead to a small number of rejections. Thus recently intensive research on False Discovery Rate has been conducted, in which the FDR is defined as the expected proportion of false rejections among all rejections. Controlling the FDR has gained its popularity in the microarray setting. Compared with multiple testing procedures that control the FWER, the FDR procedures are less stringent and lead to a larger number of rejections [3]. In this paper, Benjamini and Hochberg (1995) step-up procedure for controlling the FDR was considered.

Benjamini and Hochberg procedure (BH)

Consider the case in which m hypotheses, from which m_0 are true null hypotheses and m_1 are false null hypotheses, need to be tested. Let V be the number of true null hypotheses that we wrongly reject and R be the total number of rejected hypotheses. Benjamini and Hochberg (1995), define the FDR as the expected proportion of false rejection among the rejected hypotheses, $FDR = E(Q)$ where $Q = V/R$ when $R > 0$ and $Q = 0$ otherwise [9]. The BH procedure controls FDR as follows.

Let $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ be the ordered p-values and let $H_{(1)}, H_{(2)}, ..., H_{(m)}$ be the corresponding hypotheses. The procedure rejects $H_{(1)}, H_{(2)}, ..., H_{(l)}$ where l is the largest value of i, for which $P_{(i)} \leq \frac{i}{m}\alpha$. The adjusted p-values are then given by

$$\tilde{p}_i = min_{k=i,...,m}[min(\frac{m}{i}p_{(i)}, 1)] \tag{12}$$

Thus, the null hypothesis $H_{(i)}$ is rejected if $\tilde{p}_i \leq \alpha$.

.

5 Application and Results

We begin this section by presenting an overview of the methods in the BF-Rank package, and lastly, we show some examples of how to use this package on the data sets found in it to achieve what we mentioned in sections 3 and 4.

5.1 Description of the Package BFRank

This package was built by using some methods from other packages while some where designed from scratch. The goal was not only to help us identify biclusters but to rank them based on the fingerprints of the compounds. Similar studies have been conducted to study the association between fingerprint of compounds and gene expression of samples treated with the same compounds, for example Li et al (2011) [1] considered only Tanimoto coefficient as a measure of association. Furthermore most packages in biclustering are so specific for some particular group of algorithms. We found it interesting to come up with a package that considers different measures of association and biclustering algorithms. The main dependencies of this package include; isa2, biclust, ade4, BcDiag, fabia, and heatmap.plus. We describe some of the methods below.

5.2 How to use the BF-Rank Package

In using this package we need to download and install it from CRAN together with all the dependencies. To illustrate some of the results that can be got from BFRank, we use the two examplary data sets found in the package called expressionMatrixB and fingerprintsB. Before getting to a practical application of the functions in this package, we presents a summmary of all the functions in table 1 on page 18.

Loading the data set

```
> library(BFRank)
> data(expressionMatrixB)
> data(fingerprintsB)
```

The functions to begin with after loading the data set is any of the ComInBic functions. The choice depending if the user needs to obtain biclusters based on FABIA, Plaid model or ISA. Each of these functions has as parameters two data sets; the expression data and the fingerprints data. These algorithms starts by obtaining biclusters from the expression data, after this is done, the corresponding fingerprint data is then extracted for compounds in each bicluster. It then returns a list structure where each element in the list is the fingerprint data set for compounds in that bicluster.

Functions	Description
ComInBicPlaid	Obtained biclusters using the Plaid model and returns the fingerprints data only for compounds in biclusters.
ComInBicFabia	Obtained biclusters using the FABIA model and returns the fingerprints data only for compounds in biclusters.
ComInBicIsa	Obtained biclusters using the ISA model and returns the fingerprints data only for compounds in biclusters.
Similarity	Produce the similarity matrix from the fingerprint matrix returned by any of the above ComInBic functions. Ten different methods of defining simmilarity are implemented in this function.
SimDiag	Delete duplicate values from the symmetric similarity matrix and returns just the lower off diagonal elements of it.
Denplot	Create a density plot of each of the similarity matrices.
heat_ map	Create a heat map of each of the similarity matrices.
cumdc	Produces a cumulative distribution plot for each similarity matrix on the same figure.
linePlotBic	Takes an expression matrix and bicluster results got from any of the ComInBic functions and provides profile plots for biclustered and clustered data parallely.
ClustCpd	Cluster compounds based on their fingerprints using hierarchical clustering and ward linkage method. Ten different methods of similarity or distance can be used to cluster the compounds.
filterCvPoverA	Takes a gene expression matrix and threshold to be used for filtering and returns only the genes that satisfy the filter. Filtering is based on coefficient of variation (CV= sd/mean) and pOverA (used in the genefilter package).

Table 1: *Summary of functions in BFRank*

ComInBicFabia function

- *Description*
  This function takes an expression set and its fingerprint data set and apply the fabia algorithm to obtain 5 biclusters using the default parameters. It further extracts and returns the fingerprint data for each bicluster obtained. This function uses the code implemented in the fabia package to obtain biclusters.

- *Usage*

```
> help(ComInBicFabia)
> r = ComInBicFabia(expressionMatrixB,fingerprintsB)
> FpLstCpds = r$FpLstCpdsF # A list where each element in it is the
# fingerprint data (binary) for each bicluster.
> bres = r$bresF # The bicluster result of class factorization.
```

ComInBicPlaid function

- *Description*
  Takes an expression set and its fingerprint data set. Biclusters are extracted using the default Plaid model $fit.model = \ m + a + b$ and the corresponding fingerprint data for each bicluster is returned. This function uses the code implemented in the biclust package to obtain biclusters.

- *Usage*

```
> help(ComInBicPlaid)
> r = ComInBicPlaid(expressionMatrixB,fingerprintsB)
> FpLstCpds = r$FpLstCpdsP # A list where each element in it is the
# fingerprint data (binary) for each bicluster.
> bres = r$bresP # The bicluster result which is an object of class biclust.
> length(FpLstCpds) # number of biclusters found
> colnames(FpLstCpds[[1]]) # names of compounds in bicluster 1
> head(FpLstCpds[[1]])
```

ComInBicIsa function

- *Description*
  This function obtained biclusters using the Iterative Signature Algorithm (Isa) and rerturn the corresponding fingerprints data set of compounds in the biclusters. This function uses the code implemented in the isa2 package to obtain biclusters.

- *Usage*

```
> help(ComInBicIsa)
> r = ComInBicIsa(expressionMatrixB,fingerprintsB)
> FpLstCpds = r$FpLstCpdsI # A list where each element in it is the
# fingerprint data (binary) for each bicluster.
> bres = r$bresI # The bicluster result which is an object of class biclust.
```

In order to rank these biclusters found, we need to determine how similar the compounds are based on the fingerprints information. The function named similarity does this.

Similarity and SimDiag functions

- *Description*
  Takes a list of datasets returned by any of the ComInBic functions and returns a similarity matrix based on the method specified, method = 4 ⇒ Tanimoto and method = 1 ⇒ Jaccard index, see section 3.2 on page 7 for the definition of various methods of similarity measures.

- *Usage*

```
> help(Similarity)
> simtan <- Similarity(mat = FpLstCpds, method = 4)
> View(simtan[[1]]) # similarity matrix for the first bicluster.
```

  The simtan matrix is symmetric along the diagonal so we can select just the lower diagonal section of this matrix using the SimDiag function.

```
> help(SimDiag)
> diag <- SimDiag(simtan)
```

Two functions were implemented to have a view of the similarity values. These are the Denplot and heat_ map functions which are used to get the densities and heat map of the similarity scores respectively.

Denplot and heat_ map functions

- *Description*
  Denplot: Takes a list of similarity scores returns by the SimDiag function and produces a density plot for each bicluster. This plot shows how similar the compounds in a bicluster are based on their fingerprints.
  heat_ map: Takes a list of similarity matrices returns by the Similarity function and draws a heat map of each of them.

- *Usage*

```
> help(Denplot)
> par(mfrow = c(3, 2))
> Denplot(diag)
> heat_map(simtan)
```

Both the density plots and heat maps can give us a visual view on how we can rank the biclusters, but this is just an exploratory tool so we cannot do ranking at this stage but need to go further to look at the CDF for each of the data sets. Only at this stage that ranking can then be done.

cumdc function

- *Description*
  Takes a list of similarity scores returns by the SimDiag function and produces a cumulative distribution plot for each bicluster so that comparison and ranking can be done.

- *Usage*

```
> help(cumdc)
> par(mfrow = c(1, 1))
> cumdc(diag)
```

linePlotBic function

- *Description*
  Takes an expression matrix and bicluster results got from ComInBic function and provides profile plots for biclusterd and clusterd data parallely. This function uses the original code from the BcDiag package.

- *Usage*

```
> help(linePlotBic)
> linePlotBic(expressionMatrixB, bres, mname='fabia')
> linePlotBic(expressionMatrixB, bres, mname='isa2')
```

ClustCpd function

- *Description*
  The ClustCpd functions provide a convenient way to cluster compounds based on their fingerprints using hierarchical clustering and ward linkage method. The ten different methods of similarity in section 3.2 were converted to distance and used to cluster the compounds.

- *Usage*

```
> help(ClustCpd)
> par(mfrow = c(1, 1))
> ClustCpd(fingerprintsB, method = 4, k = 8)
> title("Cluster Dendrogram-Rogers & Tanimoto")
> # Changing the value of k changes the method used to calculate
> # ditance during clustering.
```

filterCvPoverA function

- *Description*
  This function takes a gene expression matrix and threshold to be used for filtering and returns only the gens that satisfy the filter. Two conditions are applied to each gene; Coefficient of variation (CV= sd/mean) and A filter function to filter according to the proportion of elements larger than A (pOverA). Original codes were got from genefilter package.

- *Usage*

```
> help(filterCvPoverA)
> res = filterCvPover(eset=expressionMatrixB ,P=0.10,A=100,a=0.2,b=10)
> small.eset = res$small.eset # reduced expression data to be used for
 # differential expression.
> res$propSelGenes # proportion of genes that do pass the filter.
```

5.3 Results

This section provides the results obtained from the analysis of the two datasets named expresionMatrixB and fingerprintsB found in this package. The first subsection presents the biclusters found using FABIA and how they are ranked based on two different methods. In the second section we present the result of clustering and differential expression using SAM and LIMMA.

5.3.1 Ranking of biclusters

Table 2 list the number of compounds found in each of the 5 biclusters. Some compounds occurred in more than one bicluster since FABIA allows for overlapping of biclusters.

Bicluster	1	2	3	4	5
Number of compounds	14	13	15	12	22

Table 2: *Number of compounds per bicluster*

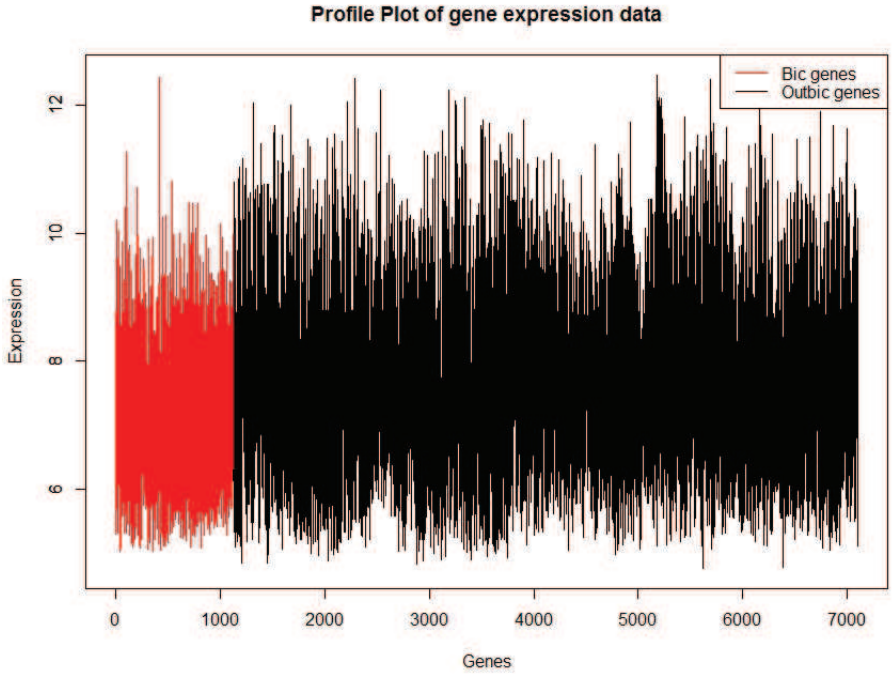

Figure 4: *Profile plot for genes in biclusters*

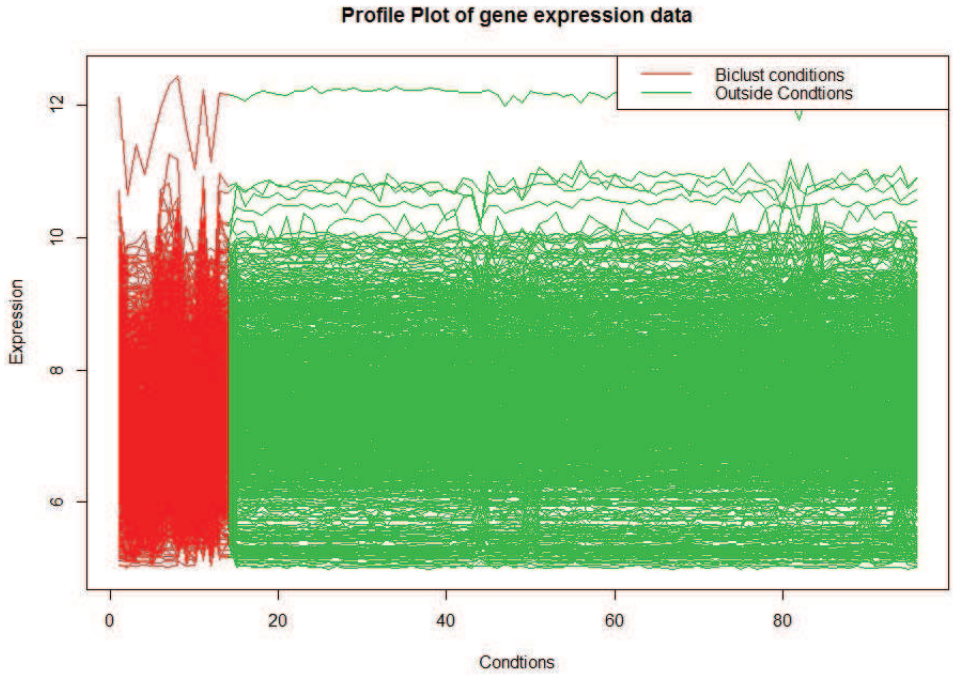

Figure 5: *Profile plot for conditions in biclusters*

The line plot of the genes and conditions in the biclusters are presented by figure 5 and 4. Genes in biclusters appears to have a lower profile than those not in any bicluster, but this is not the same with the compounds (conditions).

The similarity scores for each of the biclusters in table 2 were got using Tanimoto coefficient and Jaccard index in order to compare the different measures of similarity. Figure 6 and 7 presents the densities of the similarity scores for each bicluster based on these mentioned methods. The densities are quite different. The shape of the density (bell shape) can ring a bell on how similar the compounds are in terms of their fingerprints. From figure 7 on page 25 the compounds in bicluster 3 and 5 seem to be more similar than the others, but this will need to be confirmed by the cumulative distribution curve. The heat map (not shown) also reflects this same order of ranking.

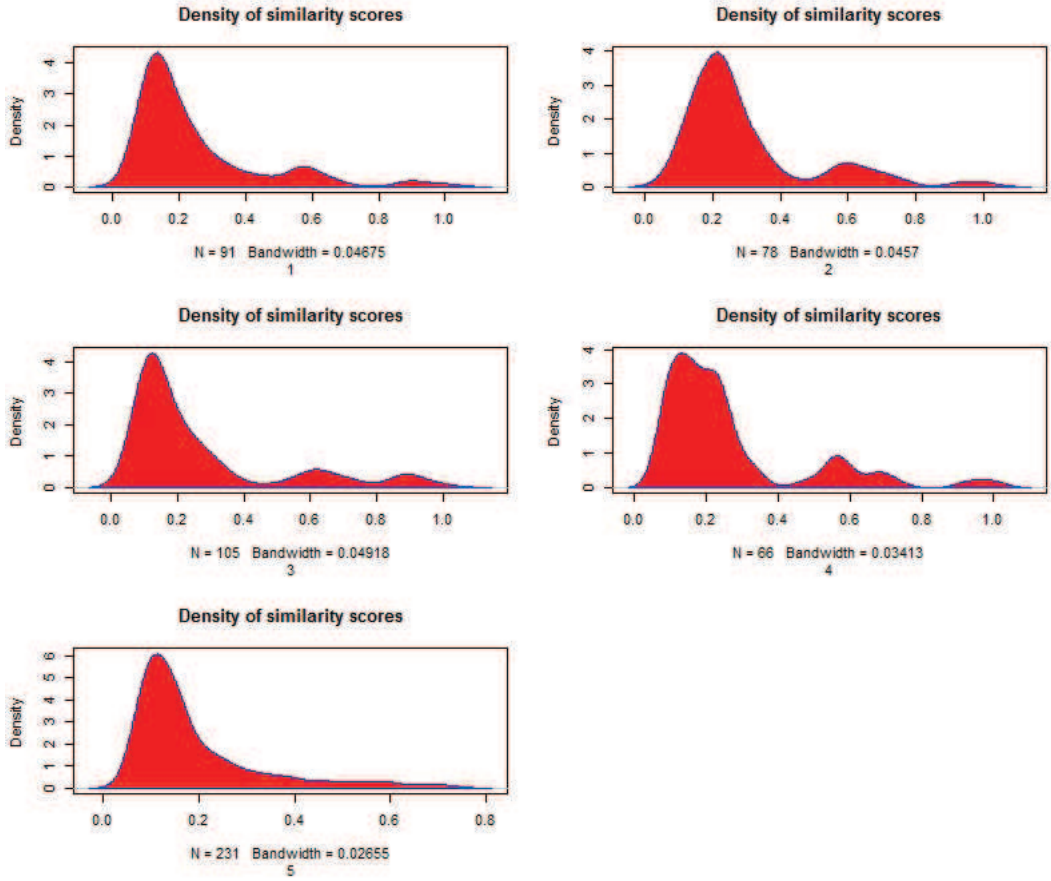

Figure 6: *Density of similarity scores based on Jaccard index*

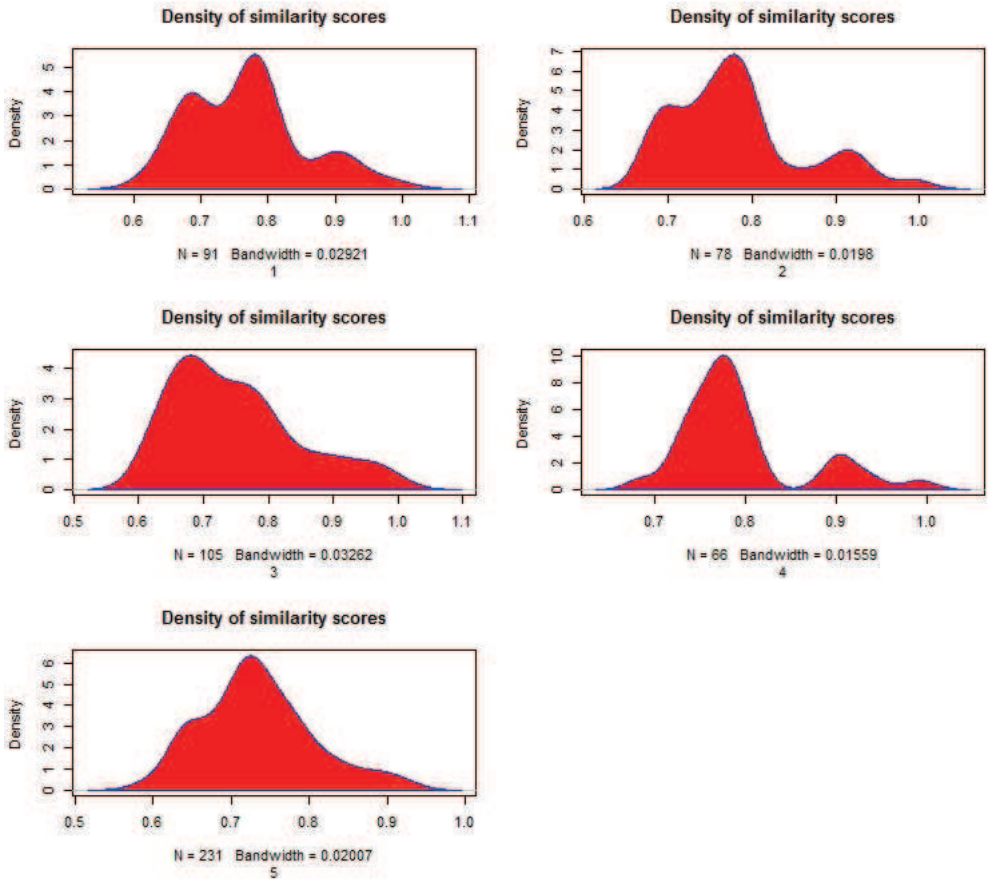

Figure 7: *Density of similarity scores based on Tanimoto coefficient.*

The ranking of the biclusters was done based on the cumulative distribution of the similarity scores. Figure 8 and figure 9 present the cumulative distribution curve based on Tanimoto and Jaccard index respectively. Both methods do rank the biclusters in the same order with the 5^{th} as the first.

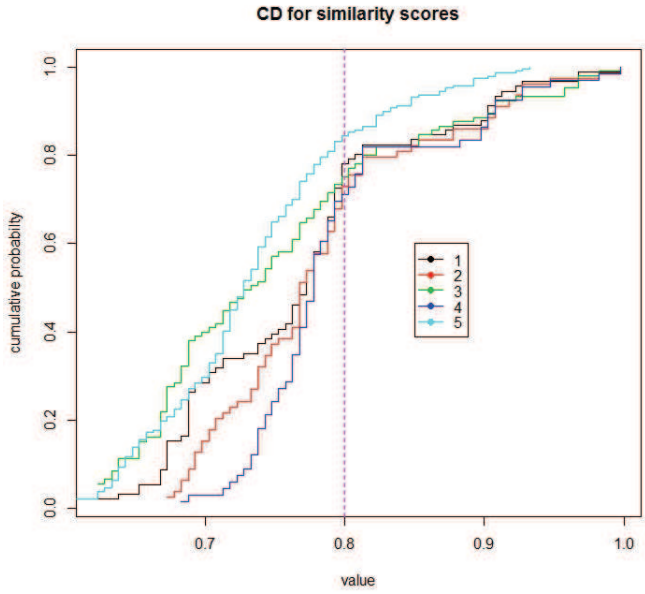

Figure 8: *Cumulative distribution curve based on Tanimoto coefficient, BC 5 appears to be the first followed by 3.*

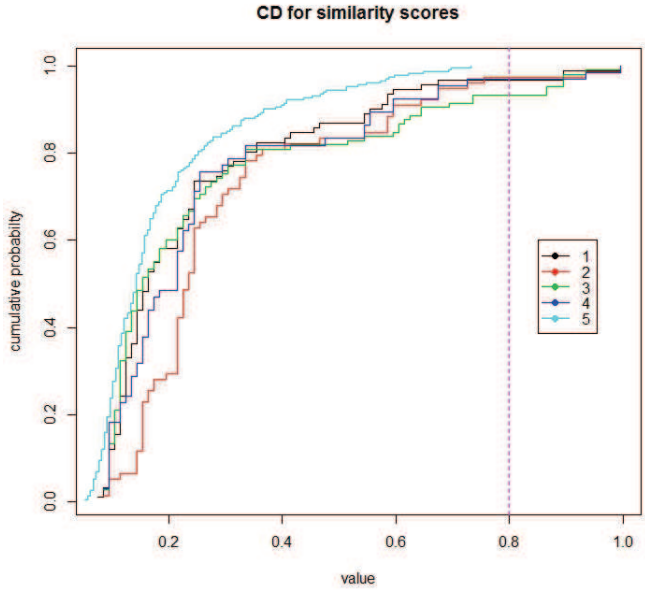

Figure 9: *Cumulative distribution curve based on Jaccard index, BC 5 appears to be the first but no clear separation between the others.*

In the next section, we present some of the results obtained from clustering and differential expression.

5.3.2 Clustering and differential expression

The results unveiled in this section were got from functions in the BFRank package and others from LIMMA and SAM. This part of the analysis commence by clustering the compounds based on their fingerprints using the 10 different methods displayed in section 3.2. It is the duty of the researcher to select the groups from these clusters for which differential expressed genes are to be identified. Figure 10 and 11 illustrate the dendrogram when the fingerprints data was clustered based on Jaccard and Tanimoto respectively as measure of distance. Eight clusters were demanded with one of them denoted as A while the rest as B.

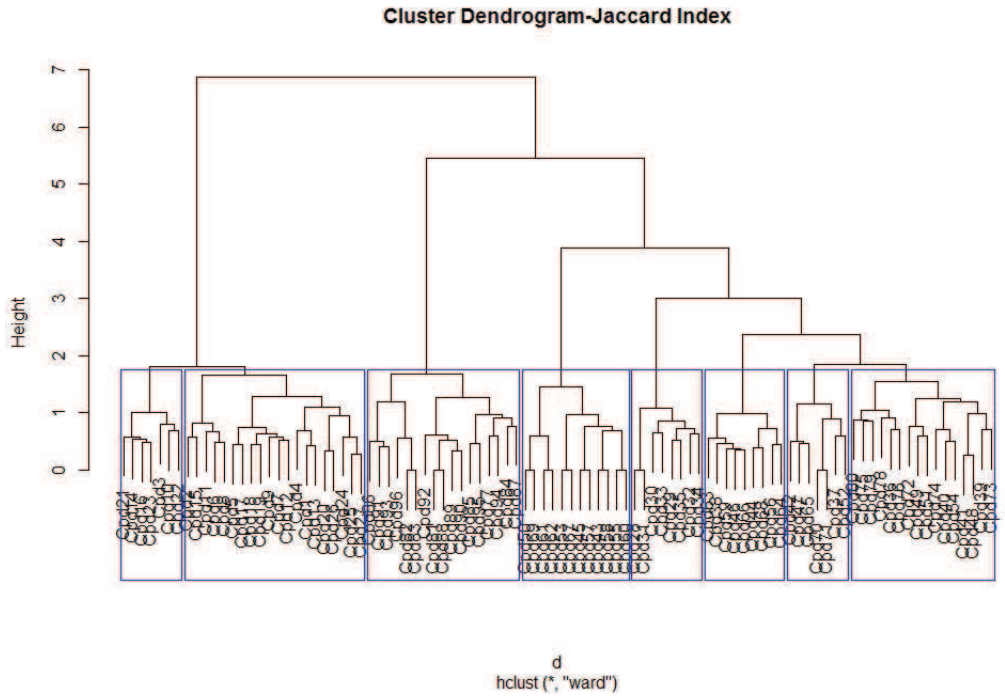

Figure 10: *Cluster dendrogram based on Jaccard index. The selected cluster denoted as A contains 20 compounds and it is the* 2^{nd} *cluster from the left while B has 76 compounds.*

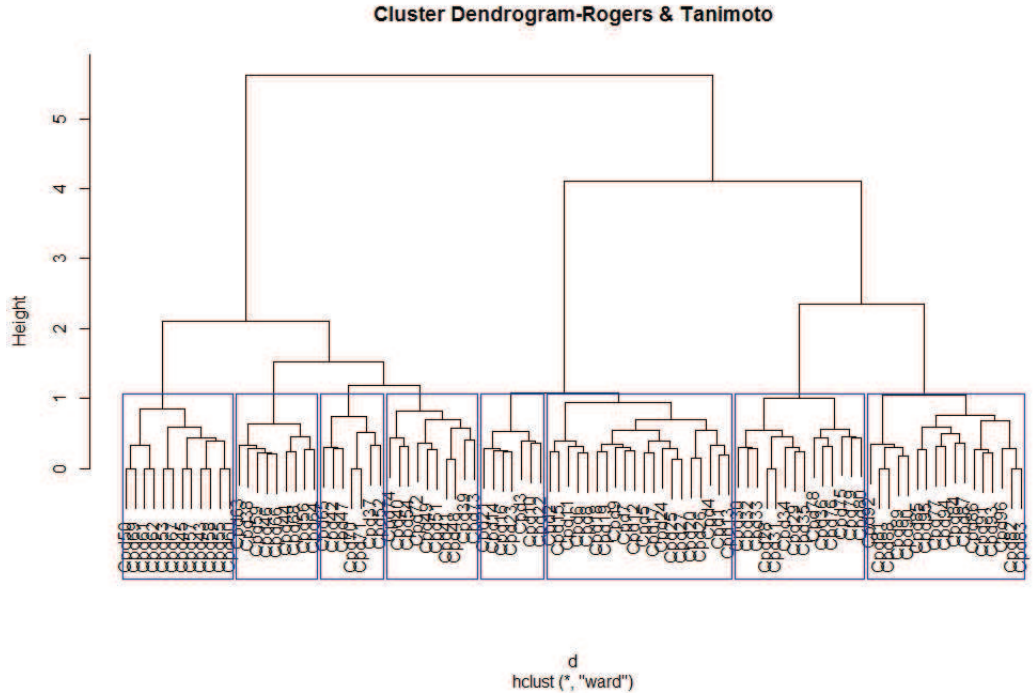

Figure 11: *Cluster dendrogram based on Tanimoto coefficient. The result is the same as that of figure 10, the cluster A appears to be the* 6^{th} *cluster from the left.*

Independent filtering was done before applying LIMMA and SAM. The function filterCvPoverA in the BFRank package was used with the options $P = 0.10$, $A = 100$, $a = 0.2$, $b = 10$ to filter out genes. 5.87% that is 417 out of 7103 of the genes was retained for downstream analysis.

Table 3 shows the top ten most significant genes at a significance level of 5% after adjusting for multiplicity by BH procedure. Figure 12 presents the gene profile of the top 3 genes from table 3, the first 20 conditions from group A have a higher mean expression value than the other thus confirming the fact that they are differentially expressed.

ID	adj.P.Val
Gene2587	2.379226e-10
Gene3604	5.324615e-07
Gene5187	5.567387e-07
Gene5340	5.567387e-07
Gene2538	5.857870e-07
Gene3106	2.323584e-06
Gene1232	2.323584e-06
Gene4873	7.002015e-06
Gene2668	1.051423e-05
Gene2586	2.327600e-05

Table 3: *Top 10 significant genes between groups A and B using LIMMA.*

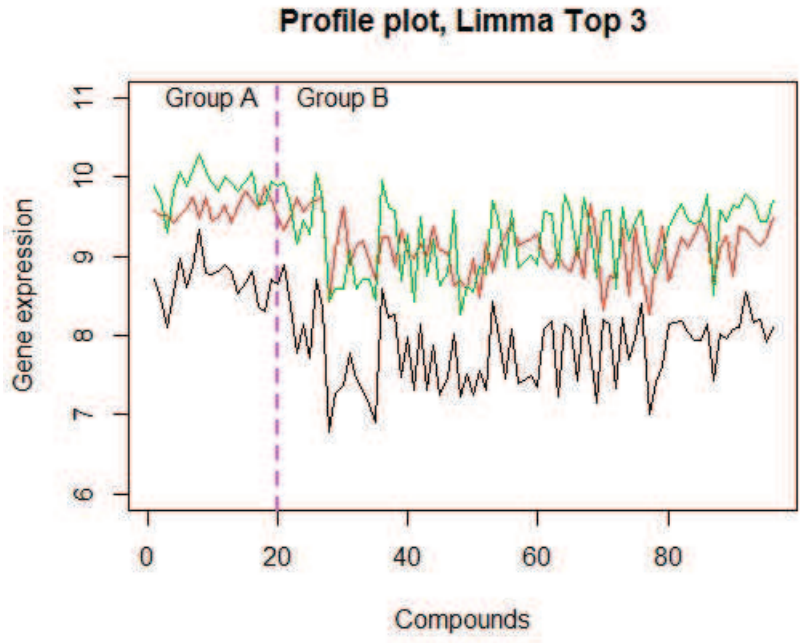

Figure 12: *Profile plot of top 3 based on LIMMA test statistic.*

Analysis by SAM reported 202 up regulated genes and 176 down regulated genes at 5% FDR and $\delta = 0.5$. Table 4 unveiled both up and down top 10 regulated genes. Figure 13 is a SAM plot that shows the up and down regulated genes for various values of δ and FDR.

Up regulated	Down regulated
Gene199	Gene153
Gene332	Gene315
Gene408	Gene164
Gene54	Gene224
Gene380	Gene220
Gene73	Gene297
Gene400	Gene381
Gene367	Gene150
Gene411	Gene207
Gene172	Gene100

Table 4: *Top 10 up and down regulated genes by SAM.*

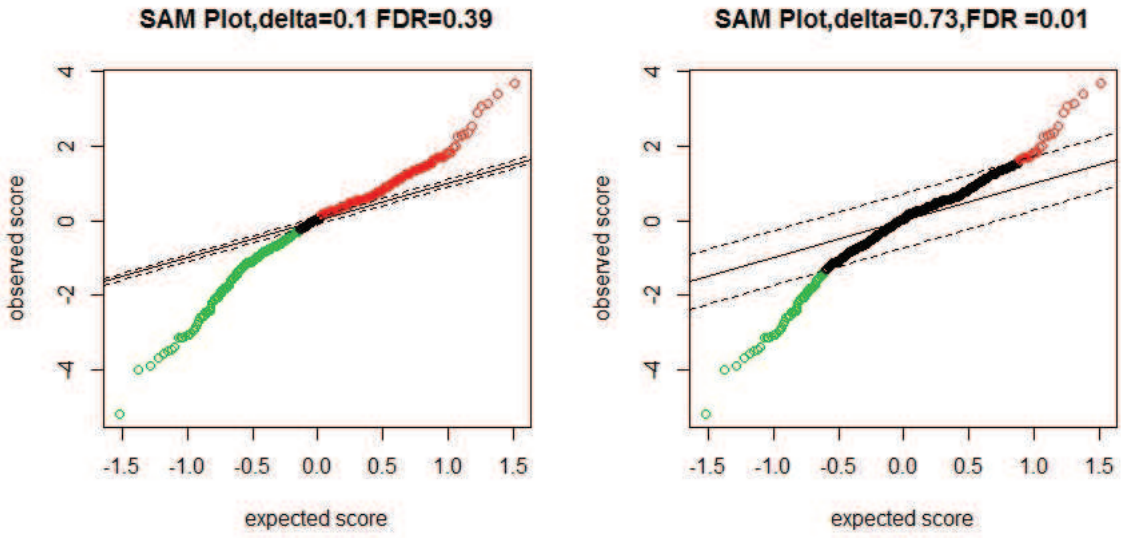

Figure 13: *SAM plot for various values of δ and FDR.*

6 Discussion and Conclusion

Bicluster analysis is specifically developed to identify which subsets of attributes are associated with which subsets of samples. One peculiar property of biclusters analysis is that an attribute or sample can be part of more than one bicluster or of no bicluster. The BFRank package presented in this report implements FABIA, Plaid model and ISA algorithms in obtaining biclusters from a data set based on their default parameters. The package also contains methods than can calculate the similarity scores, density plot, heat map and cumulative distribution of the similarity scores of compounds in biclusters. In addition, ten different methods can be implemented in calculating similarity scores. These measures of similarity can then be used to cluster the compounds with the help of this package.

In the application of this package to the data set, we were able to identify five biclusters using fabia model and rank them based on two different measures of similarity (Tanimoto and Jaccard). Both methods could rank the 5^{th} bicluster as the first based on the cumulative distributions of the similarity scores. This means that in terms of their fingerprints, the compounds in bicluster five share more of their chemical structures than the compounds in other biclusters. Both Tanimoto and Jaccard distance measures gave the same dendrogram when used with a hierarchical clustering algorithm to cluster the compounds, but this must not always be the case with the other similarity measures.

With respect to differential expressed genes, we were able to identify the top ten differential expressed genes using SAM and LIMMA, but both procedures did not result in the same list of top differentially expressed genes. Out of 7103 genes, just 417 were retained for downstream analysis after independent filtering was applied. According to the result of SAM, we had 202 up regulated genes and 176 down regulated genes. All these analysis were done at 5% significance level with BH correction for multiplicity.

There are several steps in our method that can be varied depending on the requirements. We used hierarchical clustering to cluster the compounds in the case of identifying differentially expressed genes, but different clustering methods could also be used. With respect to obtaining biclusters, the default models were used which could not be the best case in all situations. Irrespective of some minor limitations of our method, we were able to explore the function-structure relationship of compounds, which will lead to a better and more thorough understanding of drug actions *in vivo*.

.

Acknowledgement

I wish to extend my profound gratitude to members of the staff at Centre for Statistics, University of Hasselt who were more than generous with their expertise and precious time. A special thanks to my supervisor Prof.Dr. Ziv SKHEDY for his immense contribution from course work. Thank you Prof.Dr. Marc AERTS, chair Master of Epidemiology and Public Health Methodology, Prof.Dr. Geert MOLENBERGHS, chair Master of Biostatistics, Prof.Dr. Tomasz BURZYKOWSKI, chair Master of Statistical Bioinformatics and Mrs. Martine MACHIELS, Programs secretariat for not only accepting me to be a student of the statistic program but also transforming me to an able Bioinformatician.

I would like to acknowledge and thank my colleagues at CENSTAT for their collective effort from the start of the program till this faithful end. Special appreciation goes to Richard NJILEFAC, Kouam KAMANI and Susan NATIKAWEEWA for their contributions as my group members in one or two of the courses during this study period.

Finally I would like to thank Martin OTAVA who assisted me in developing the BFRank package. Your excitement and willingness to provide feedbacks made the completion of this research an enjoyable experience.

.

7 References

1. Yun L, Hang T, Siyuan Z, Jinfang W, Yixue L, Pei H, and Xuan L:**Association of feature gene epression with structural fingerprints of chemical compounds**. Journal of Bioinformatics and Computatioinal Biology, Vol 9, No. 4(2011) 503-519.

2. Sven B, Jan I, and Naama B: **Iterative signature algorithm for the analysis of large-scale gene expression data**. Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

3. Lin D, Shkedy Z, Burzykowski T, Ion R, Gohlmann H. W. H., Bondt A.De, Perera T, Geerts T, Van den Wyngaert I, and Bijnens L: **An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments with one Control in the Presence of Small-variance Genes**. Biometrical Journal 50 (2008) 5, 801823 DOI: 10.1002/bimj.200710467.

4. Richard B, Robert G, and Wolfgang H: **Independent filtering increases detection power for high-throughput experiments. European Bioinformatics Institute**. Cambridge CB10 1SD, United Kingdom.

5. Michael H, Christopher J, John N, and William L: **Applied Linear Statistical Models**. McGraw-hill International Edition, Fifth Edition, (2005).

6. Smyth G: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. Stat Appl Genet Mol Biol 2004, 3:Article3.

7. Carl M, Owen W, Anna Y and Robert N: **Comparison of small n statistical tests of differential expression applied to microarrays**. BMC Bioinformatics 2009, 10:45 doi:10.1186/1471-2105-10-45.

8. Scholtens D, von Heydebreck A: **Bioinformatics and computational biology solutions using R and Bioconductor**. Eds R Gentleman, VJ Carey, W Huber, RA Irizarry, and S Dudoit (Springer, New York,2005), pp 229248.

9. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. Journal of the Royal Statistical Society, (1995), Series B 57, 289300.

10. Sara C, Arlindo L: **Biclustering Algorithm for Biological Data Analysis: A Survay**, IEEE/ACM Transactions on Computatioinal Biology and Bioinformatics.

11. R Documentation: **Computation of Distance Matrices for Binary Data**, `http://pbil.univ-lyon1.fr/ADE-4/ade4-html/dist.binary.html`, accessed on 26/07/13.

12. Heather T, Trevor B, Wojtek K: **Improved biclustering of microarray data demonstrated through systematic performance tests**, Computational Statistics & Data Analysis, Department of Mathematical Sciences, University of Exeter, Laver Building, 48 (2005).

13. Sepp H: **FABIA: Factor Analysis for Bicluster Acquisition- Manual for the R package**, Institute of Bioinformatics, Johannes Kepler University Linz, Altenberger Str. 69, 4040 Linz, Austria.

14. Jurgen B: **Fingerprint Design and Molecular Complexity Effects**, Life science informatics, University of Bonn.

8 Appendix

BFRank user manual

The user manual of this package is presented in this section. This is an extra document added to this report so the numbering system will not be consistent with that of the main report.

Package 'BFRank'

September 11, 2013

Type Package

Title Ranking of biclusters based on fingerprints.

Version 1.0

Date 2013-08-26

Author Silenou Bernard, Martin Otava

Maintainer Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

Depends R (>= 2.0.9), Rcpp, fabia, biclust, isa2, ade4, colorspace,BcDiag, vegan, ALL, heatmap.plus, genefilter

Imports xcms

Description This package search for biclusters of genes using Plaid model,Fabia and ISA algorithms. It can also rank the biclusters based on on similarity of the compounds using different methods of similariy for binary data like Tanimoto coefficient described in Lin et al. (2011).

License GPL-3

R topics documented:

BFRank-package *Ranking of biclusters based on fingerprints.*

Description

The main functions ComInBicPlaid, ComInBicIsa and ComInBicFabia provides the Plaid model, ISA and FABIA model respectively in obtaining biclusters in two dimensional data. In addition, the package provides methods for comparing the similarity between compounds in a bicluster based on their fingerprints, visualisation and diagnostic of biclusters using line plots, and clustering of compounds base on their fingerprints using ten different specified distance measure of association (Tanimoto, Jacardin, etc).

Details

Package:	BFRank
Type:	Package
Version:	1.0
Date:	2013-08-26
License:	GPL-3

In using this package we start with any of the functions ComInBicPlaid, ComInBicFabia or ComInBicIsa that takes an expression matrix and its fingerprints matrix as parameters and returns a list where each element in the list corresponds to the fingerprints matrix of only compounds in the same bicluster. This can them be followed by the "Similarity" and "Simdiag" functions to obtain the similarity matrix of the compounds and finally ranking can be done using cumdc function.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Daniel C, Stephane D: Computation of Distance Matrices for Binary Data. R Documentation, http://pbil.univ-lyon1.fr/ADE-4/ade4-html/dist.binary.html, accessed on 20/08/13.

Yun L, Hang T, Siyuan Z, Jinfang W, Yixue L, Pei H, and Xuan L: Association of feature gene epression with structural fingerprints of chemical compounds. Journal of Bioinformatics and Computatioinal Biology, Vol 9, No. 4(2011) 503-519.

See Also

ComInBicPlaid, ComInBicFabia, Similarity, ClustCpd

Examples

```
## Not run:
set.seed = (1234)
r = ComInBicPlaid(expressionMatrixB,fingerprintsB)
FpLstCpds = r$FpLstCpdsP # list with each element a matrix of
```

```
#fingerprint of only compounds in the same bicluster.
bres = r$bresP
# Object of class biclust as returned by the biclust package.
head(FpLstCpds[[1]])


simtan = Similarity(mat=FpLstCpds,method=4)
# Similarity based on Tanimoto
#coefficient.

View(simtan[[1]])

diag = SimDiag(simtan)

linePlotBic(expressionMatrixB,bres)

cumdc(diag)


## End(Not run)
```

ClustCpd *Clustering*

Description

Takes a matrix of fingerprints of compounds and cluster the compounds based on their fingerprints using hierarchical clustering with ward linkage method.

Usage

```
ClustCpd(mydata, method, k)
```

Arguments

mydata	matrix with rows as fingerprints and columns as compounds.
method	method used for distance measure.
k	cut off limit for number of clusters required, can be any number in 2:n where n is the number of compounds.

Details

Ten different measures of distance can be used to cluster the compounds. These methods are :"Jaccard index"," Sockal & Michener","Sockal & Sneath", "Rogers & Tanimoto","Czekanowski","S9,Gower & Legendre"," Ochiai", "S13,Sockal & Sneath"," Phi of Pearson" and "S2,Gower & Legendre" and codded as 1:10. method = 1 means Jaccard index and same for the rest.

Value

A dendrogram showing how the compounds are clustered.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Daniel C, Stephane D: Computation of Distance Matrices for Binary Data. R Documentation, http://pbil.univ-lyon1.fr/ADE-4/ade4-html/dist.binary.html, accessed on 20/08/13.

`ComInBicFabia` *Compounds in bicluster using FABIA.*

Description

This function takes an expression set and its fingerprint data set and apply the fabia algorithm to obtain 5 biclusters using the default parameters. It further extract the fingerprint data for each bicluster obtained.

Usage

```
ComInBicFabia(EsetF, FpF)
```

Arguments

`EsetF`	The gene expression set.
`FpF`	The fingerprint data.

Value

A list of length two, the first object in the list is a list of size 5 containing the fingerprint data of the compounds belonging to the same bicluster. The second object is of type standardization containing the bicluster result as reported by the fabia package.

`FpLstCpdsF`: List of fingerprint data for compounds in bicluster.

`bresF`: Bicluster result of class standardization.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Sepp Hochreiter: FABIA: Factor Analysis for Bicluster Acquisition- Manual for the R package , Institute of Bioinformatics, Johannes Kepler University Linz,Altenberger Str. 69, 4040 Linz, Austria.

See Also

`ComInBicPlaid`, `ComInBicIsa`, `Similarity`

ComInBicIsa *Compounds in bicluster using ISA.*

Description

Takes an expression set and its fingerprint data set and extract biclusters using the Iterative Signature Algorithm (ISA).

Usage

```
ComInBicIsa(EsetI, FpI)
```

Arguments

EsetI	The expression set.
FpI	The fingerprint data.

Value

A list of length 2, the first object in the list is a list of size equal to the number of biclusters found. Each of the data sets in the list contain the fingerprint data of the compounds belonging to the same bicluster. The second object is of type biclust containing the bicluster result.

FpLstCpdsI: List of fingerprint data for compounds in bicluster.

bresI: Bicluster result.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Sven B, Jan I, and Naama B: Iterative signature algorithm for the analysis of large-scale gene expression data. Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

See Also

ComInBicPlaid, ComInBicFabia, Similarity

`ComInBicPlaid` *Compounds in bicluster using plaid model.*

Description

Takes an expression set and its fingerprint data set. Biclusters are extracted using the default model fit.model = ~m + a + b, and the corresponding fingerprint data for each bicluster is returned.

Usage

```
ComInBicPlaid(EsetP, FpP)
```

Arguments

`EsetP`	The expression set and.
`FpP`	The fingerprint data.

Value

A list of length 2, the first object in the list is a list of size equal to the number of biclusters found. Each of the data sets in the list contain the fingerprint data of the compounds belonging to the same bicluster. The second object is of type biclust containing the bicluster result.

`FpLstCpdsP`: List of fingerprint data for compounds in bicluster.

`bresP`: Bicluster result of class biclust.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Heather T, Trevor B, Wojtek K: Improved biclustering of microarray data demonstrated through systematic performance tests, Computational Statistics and Data Analysis, Department of Mathematical Sciences, University of Exeter, Laver Building, 48 (2005).

See Also

`ComInBicFabia`, `ComInBicIsa`, `Similarity`.

cumdc *Cumulative distribution function*

Description

Takes a list of similarity scores returns by the SimDiag function and produces a cumulative distribution plot for each bicluster so that ranking can be done by visual inspection.

Usage

```
cumdc(LstTot)
```

Arguments

LstTot	Object returned by the SimDiag function. This is a list and each element in the list is a numeric vector corresponding to the lower off diagonal of the similarity matrix.

Value

Cumulative distribution plot for all the biclusters on the same figure.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

See Also

Similarity, SimDiag.

Denplot *Density plot*

Description

Takes a list of similarity scores returns by the SimDiag function and produces a density plot for each bicluster. This plot shows how similar the compounds in a bicluster are based on their fingerprints.

Usage

```
Denplot(LstTot)
```

Arguments

LstTot	Object returned by the SimDiag function. This is a list and each element in the list is a numeric vector corresponding to the lower diagonal of the similarity matrix.

Value

The density plot of each similarity matrix. This plot reflects how similar the compounds in a bicluster are based on their fingerprints.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

See Also

SimDiag, Similarity.

---

expressionMatrixB *Gene Expression Data Example*

Description

This micro-array data contains 1000 genes and 96 compounds. Each array was got from cell line treated with a single compound. Due to confidentiality, this is only part of the real data set that has 7103 genes and 96 compounds.

Usage

```
data(expressionMatrixB)
```

Format

A double matrix with 1000 genes (row) on the following 96 compounds (columns).

`Cpd1` Array from compound 1

`Cpd2` Array from compound 2

`...` And so on

`Cpd96` Array from compound 96

Examples

```
data(expressionMatrixB)
dim(expressionMatrixB)
colnames(expressionMatrixB)
head(expressionMatrixB)
```

`filterCvPoverA` *Filtering based on CV and pOverA*

Description

Takes an expression matrix of compounds and threshold of to be used for filtering and returns only the genes that satisfy the filter.

Usage

```
filterCvPoverA(eset, P, A, a, b)
```

Arguments

`eset`	matrix with rows as genes and columns as compounds. This matrix should be of log transformed scale as that got from an expression set.
`P`	The proportion that need to exceed A for TRUE to be returned.
`A`	The value to be exceeded.
`a`	minimum required value for CV.
`b`	maximum required value for CV.

Details

Two conditions are applied to each gene; pOverA and coefficient of variation (CV= sd/mean)

Value

A list with two components:

`small.eset`: Selected expression set.

`propSelGenes`: Proportion of genes retained.

Author(s)

Silenou Bernard, Martin Otava <martin.otava@uhasselt.be>

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Richard B, Robert G, and Wolfgang H: Independent filtering increases detection power for high-throughput experiments. European Bioinformatics Institute. Cambridge CB10 1SD, United Kingdom.

fingerprintsB *Fingerprints Data Example*

Description

This fingerprints data contains 354 chemical structure characteristics and 96 compounds. It is a binary data with value 1 if the given compound have a particular characteristic and 0 otherwise.

Usage

```
data(fingerprintsB)
```

Format

A double matrix with 354 characteristics (row) on the following 96 compounds (columns).

`Cpd1` Characteristics for compound 1

`Cpd2` Characteristics for compound 2

`...` And so on

`Cpd96` Characteristics for compound 96

Examples

```
data(fingerprintsB)

dim(fingerprintsB)

View(head(fingerprintsB))
```

heat_map *Heat Map*

Description

Takes a list of similarity matrices and draws a heat map of each.

Usage

```
heat_map(simList)
```

Arguments

`simList`	List of similarity matrices returned by the similarity function.

Value

Heat map of each similarity matrix.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

See Also

`Similarity`.

---

linePlotBic *Line plot for Biclusters*

Description

Takes an expression matrix and bicluster results got from any of the "ComInBic" functions and provides profile plots for biclusterd and clusterd data parallely.

Usage

```
linePlotBic(dset, bres, mname)
```

Arguments

`dset`	Expression matrix
`bres`	Object returned by any of the "ComInBic" functions.
`mname`	Method name used to get biclusters, can be either 'biclust', 'isa2' or 'fabia'.

Details

This function uses the original code from the BcDiag package.

Value

A parallel line plot for compounds and genes in and out of biclusters.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

See Also

`ComInBicFabia`, `ComInBicIsa`, `ComInBicPlaid`.

SimDiag *Extract the lower off diagonal of the similarity matrix*

Description

Takes a list of datasets returned by the Similarity function. Since the similarity function is symmetrical, it extract just one section of the off diagonal values.

Usage

```
SimDiag(simMat)
```

Arguments

simMat	List object returned by the similarity function.

Details

Since the similarity matrix is symmetrical along the diagonal, there is no need to use all the values for downstream analysis, so we delete duplicates.

Value

A list of length n equal to the number of biclusters generated. Each element in the list is a vector corresponding to the lower off diagonal elements of the similarity matrix.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

See Also

ComInBicPlaid, ComInBicIsa, Similarity

Similarity *Similarity matrix*

Description

Takes a list of datasets returned by any of the ComInBic functions and returns a list where each element is a similarity matrix of the compounds in the given data set. Ten different methods of measuring similarity can be specified.

Usage

```
Similarity(mat, method)
```

Arguments

`method`	Method used for distance measure, take values in 1:10.
`mat`	Object returned by ComInBicIsa, ComInBicIsaFabia or ComInBicPlaid function.

Details

Ten different measures of similarity can be used to compare the compounds. These methods are :Jaccard index, Sockal and Michener, Sockal and Sneath, Rogers and Tanimoto, Czekanowski, S9,Gower and Legendre, Ochiai, S13,Sockal and Sneath, Phi of Pearson and S2, Gower and Legendre and codded as 1:10. method equal 1 means Jaccard index and same for the rest.

Value

A list of length equal to the number of biclusters. Each element in the list is a matrix of similarity scores.

Author(s)

Silenou Bernard, Martin Otava

Maintainers: Silenou Bernard <silenoubernard@yahoo.com>, Martin Otava <martin.otava@uhasselt.be>

References

Gower, J.C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3, 5-48.

Daniel C, Stephane D: Computation of Distance Matrices for Binary Data. R Documentation, http://pbil.univ-lyon1.fr/ADE-4/ade4-html/dist.binary.html, accessed on 20/08/13.

See Also

`ComInBicFabia`, `ComInBicIsa`.

Index

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
BFRank: An R package of fingerprints based ranking of gene expression biclusters in early drug development

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2013**

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,

Silenou, Bernard Chawo

Datum: **11/09/2013**