Evaluation of CC Biclustering ability to cluster differentially expressed genes with *biclust* R package

Daianna Gonzalez Padilla

23/2/2022

Keywords

Biclustering, biclust, CC algorithm, differential gene expression.

Abstract

This work accomplishes a CC biclustering analysis with real expression data from differentially expressed genes using **biclust** R package. It also compares the outcome with clusters obtained from hierarchical clustering and evaluates how well the biclusters adjust to the expected results based on the sample information. There were two biclusters found within the data from which at least one surely did not represent differentially expressed gene clusters across samples.

Introduction

Biclustering, also known as co-clustering, two-dimensional clustering and two-way clustering, is an unsupervised data mining technique that performs clustering on the rows and columns of a data matrix at the same time considering only a subset of relevant features when grouping objects into clusters, however, not only the objects are clustered but also their features [1]. The goal of this clustering algorithm is to find sub-matrices in the dataset (*e.g.* subsets of rows and columns) whose elements exhibit significant homogeneity or in other words, that are as similar as possible to each other and as different as possible to the rest [2,8]. Such submatrices consist of subsets of columns that in turn determine row assignment [3].

To understand how biclustering works, let's assume a data matrix M of dimensions $n \times m$ with a set of rows X = {x1 , . . . , xn } and a set of columns Y = {y1 , . . . , ym}. A bicluster B is defined as B ⊂ M, *i.e.* a submatrix of M of dimensions $k \times s$ where its rows I ⊂ X and its columns J ⊂ Y. All the elements a_{ij} of B with i ∈ I and j ∈ J are expected to constitute the smaller variance within the bicluster [4].

One of the biclustering methods (the first one) is Cheng and Church (CC) algorithm (Cheng & Church, 2000). It achieves biclustering by calculating the high similarity score called the Mean Residue Score (MRS): if this score for a certain set of rows and columns is above a predefined level, then the group is called a bicluster. The residue of the element a_{ij} in B is defined as following:

$$a_{ij} - a_{Ij} - a_{iJ} + a_{IJ}$$

where a_{ij} is the element with (i,j) coordinates in the matrix

$a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$

$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$

$a_{IJ} = \frac{1}{|I||J|} \sum_{\substack{i \in I \\ j \in J}} a_{ij}$

The Mean Residue Score of the bicluster B is:

$$H(I,J) = \frac{1}{|I||J|} \sum_{\substack{i \in I \\ j \in J}} (a_{ij} - a_{Ij} - a_{iJ} + a_{IJ})^2$$

These residues account for bicluster homogeneity, *i.e.* how different is each element of the bicluster from the rest of the values in its column and row and from the overall mean of the bicluster. The smaller these residues are, the smaller the MRS and more homogeneous the bicluster [3,4].

The main difference from clustering techniques lies in the fact that typical clustering methods define clusters by either rows or columns, finding groups relevant to one feature at the expense of the other. In contrast, biclustering techniques take into account both features or dimensions. These differences get clearer by visualizing their outcomes (See **Figure 1**). Clustering forms homogenous groups by either rows or columns, whilst biclustering creates homogeneous subgroups with high similarity from the entire data matrix defined by both the columns and rows.

Hence, biclustering overcomes the lost information of oversimplified clustering techniques and is therefore an useful bioinformatic tool for exploratory data analysis and data mining. The scientific aim of using this technique is to uncover biologically related elements under certain conditions. Particularly, it has become popular for the study of gene expression data sets since it is widely accepted to identify co-regulated or co-expressed genes under the same subset of experimental conditions [5,6]. In fact, the recent boom in biclustering finds its origin in the increased amount of genetic data [8]. In such cases, the interesting biclusters are those comprising genes that show similar up-regulation and down-regulation under a set of conditions [7]. Thus, through this type of analysis the involvement of genes and conditions in multiple pathways can be elucidated [4]. However, the applications of biclustering are not limited in genomics but in fields such as marketing in which consumers are grouped into market segments according to their preferences [8].

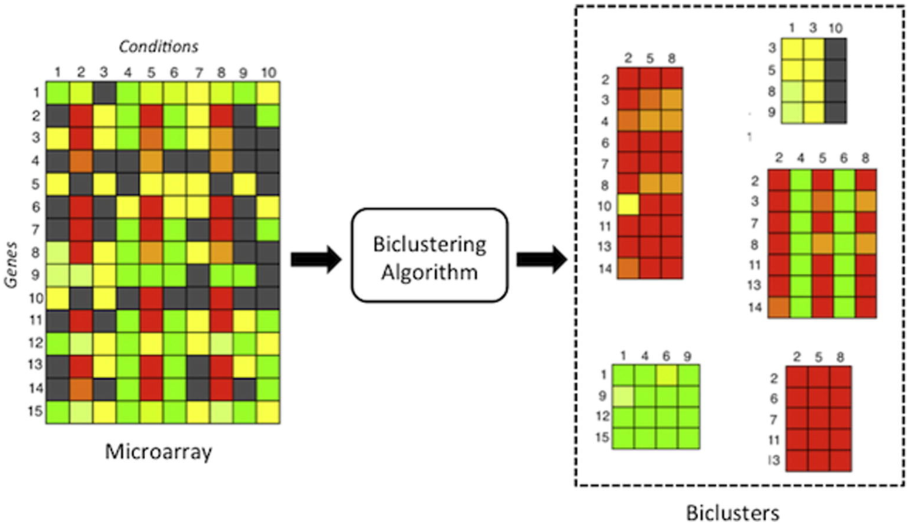

Figure 1: Clustering and Biclustering outcomes from the same data matrix. Source: Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015).

The aim of this work, however, is not to identify co-regulated or co-expressed genes but to evaluate CC biclustering method ability to cluster differentially expressed genes with samples where they show similar expression values. The motivation for doing this is to validate biclustering as an efficient method to get the particular set of genes that change their expression according to sample variables such as disease state in the particular dataset used.

There are available R tools allowing to implement biclustering such as *superbiclust, iBBiG, QUBIC, s4vd **and** BiBitR** each providing unique algorithms and implementations (Khamiakova, 2014; Gusenleitner and Culhane, 2019; Zhang et al., 2017; Sill and Kaiser, 2015; Ewoud, 2017) [9]. Another publicly package is **bioclust** by Kaiser and Leisch (2008). This is a commonly used package that obtains bioclusters in a data matrix using the algorithm specified in the method-argument and also allows to visualize the resultant biclusters. Currently, this package allows the implementation of 6 different methods applying different algorithms: BCCC, BCXmotifs, BCPlaid, BCSpectral, BCBimax and BCQuest [8].

Methods and Materials

In this work, the biclustering was performed using the version 2.0.3 of ***biclust*** R package and the method used was **BCCC** which performs CC biclustering based on the framework previously described.

On the other hand, the input data matrix consists of normalized expression data of 72 DE genes (as rows) in 10 samples (as columns) corresponding to isolated dermal endothelial cells from 4 diabetic patients and 6 control individuals. Original data is available on R package **recount3** as “SRP095512” project of type *data_source*, from *human*, with *gene* as data type and *gencode_v26* as selected annotation. Data processing and extraction of DE genes by disease state can be found at **RNAseq_2022_Finalproject** repository. Input data used in this work can be directly downloaded from **input_data**.

Results

When genes were grouped by biclustering (See code in **R_code** file), there were 2 biclusters found: one of 61 rows and 10 columns and another of 11 rows and 8 columns (See **Figure 2**), thus comprising all of the genes. In **Figure 2** it is possible to appreciate that each bicluster has a relatively homogeneous color distribution compared to the other which could be traduced as similar gene expression values in the samples that compose each bicluster.

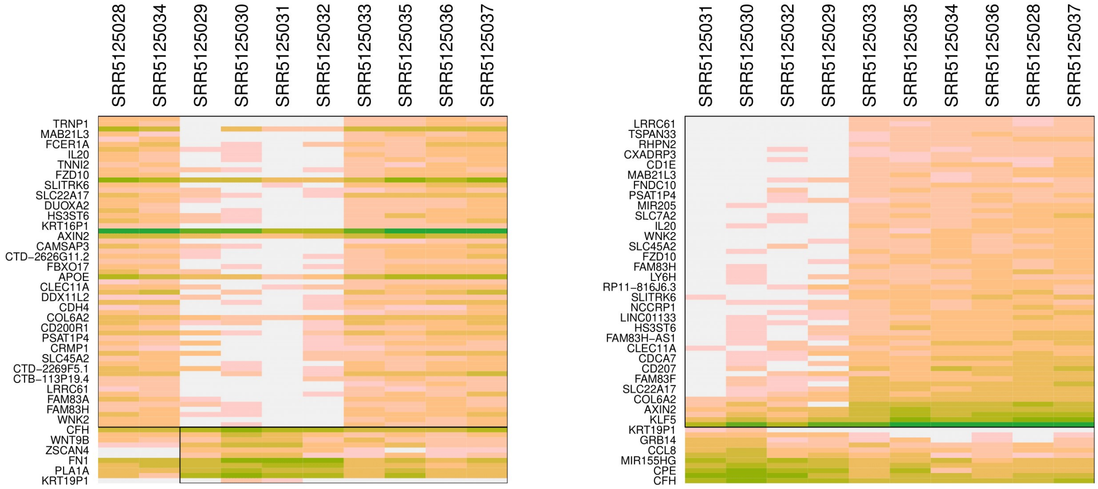

Figure 2: Biclusters of DE genes in 10 samples. At left there is a heatmap of the expression data of all genes in all samples in the original order. The two biclusters found are in black boxes. At right the rows and columns of the heatmap are ordered by their values. Note: the gene names that appear on the y-axis of heatmaps do not correspond to their rows since they encompass two rows each.

In **Figures 3** and **4** both biclusters are shown in the whole matrix and locally to make clearer which genes and samples are clustered in each one so their apparent relationships can come to light in the future.

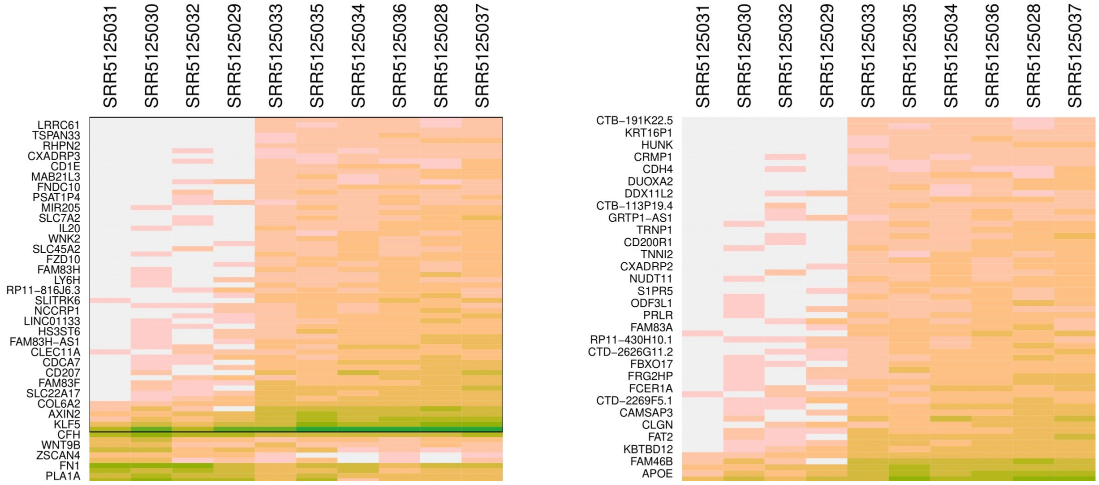

Figure 3: Bicluster 1. At left the bicluster 1 is framed in a black box in the complete heatmap. At right the local region of the bicluster is shown comprising 61 genes across the 10 samples.

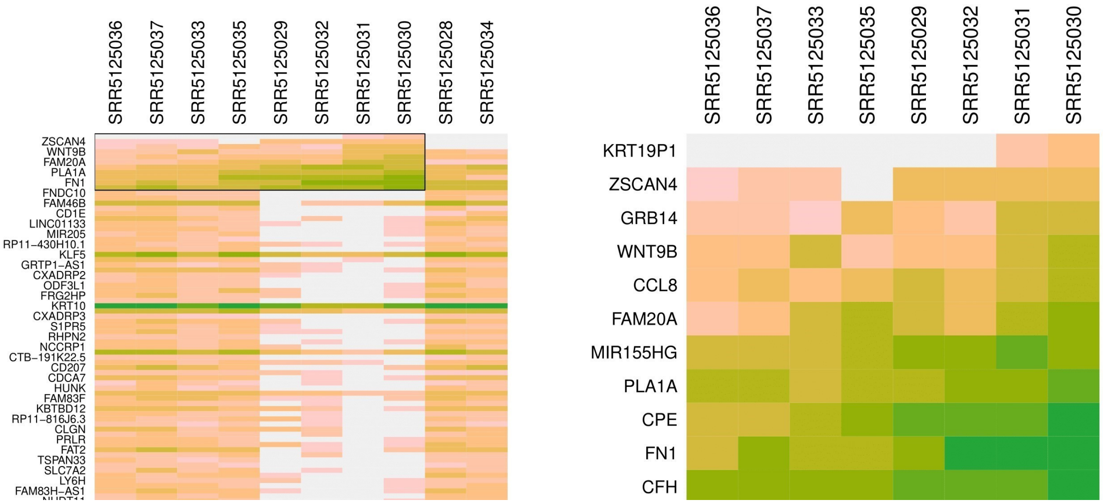

Figure 4: Bicluster 2. At left the bicluster 2 is framed in a black box in the complete heatmap. At right the local region of the bicluster is shown comprising 11 genes across the 8 samples.

Secondly, the expression levels of the genes in the set of samples for each bicluster were plotted and the results can be seen in **Figure 5** and **6**. These results show a subtle tendency in the gene expression behavior in all samples. It is expected to happen in a more noticeable way since it accounts for similar expression levels of each gene in all the samples or conditions that define a bicluster.

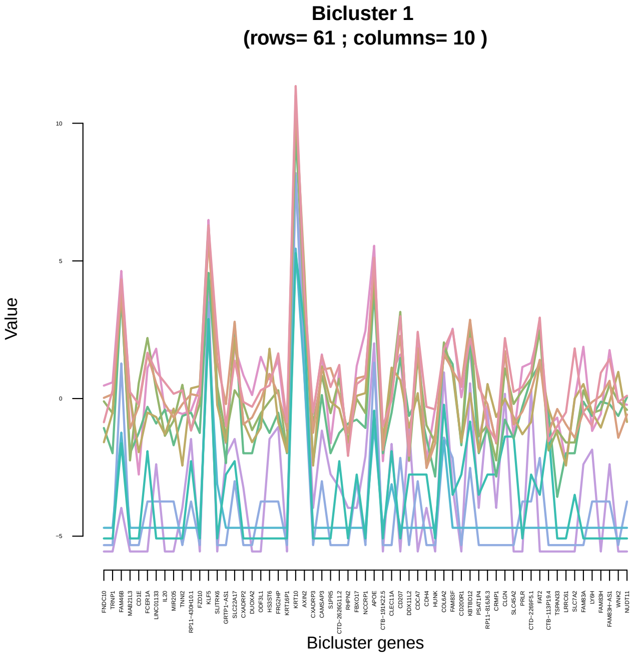

Figure 5: Expression levels of genes in bicluster 1. Each line represents a different condition of the bicluster 1 and describes the expression levels of the genes from the same bicluster.

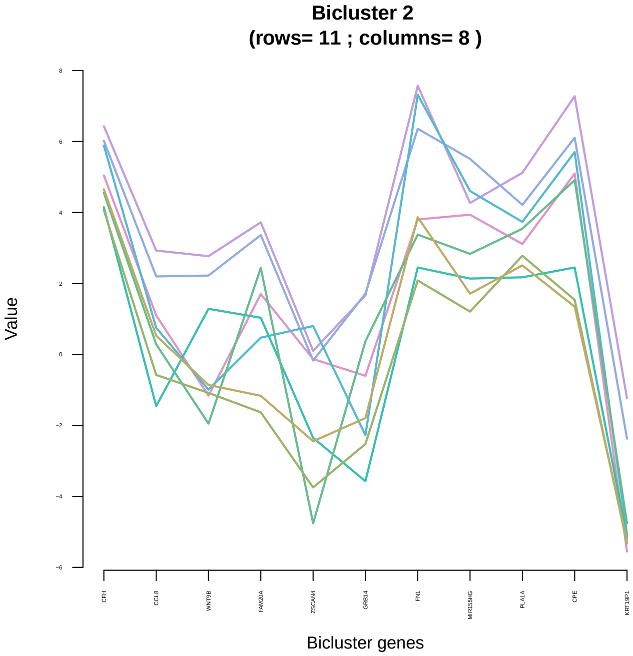

Figure 6: Expression levels of genes in bicluster 2. Each line represents a different condition of the bicluster 2 and describes the expression levels of the genes from the same bicluster.

Lastly, **Figure 7** shows a heatmap of the same DE genes expression levels grouped by the method *complete* of hierarchical clustering. In the figure both rows and columns are clustered but not in a simultaneous manner (as biclustering does). There it is possible to notice the similarity between the resultant clusters and the biclusters previously obtained: all 61 genes of bicluster 1 were grouped together and the 11 ones of bicluster 2 formed two groups (See **Figure 7**).

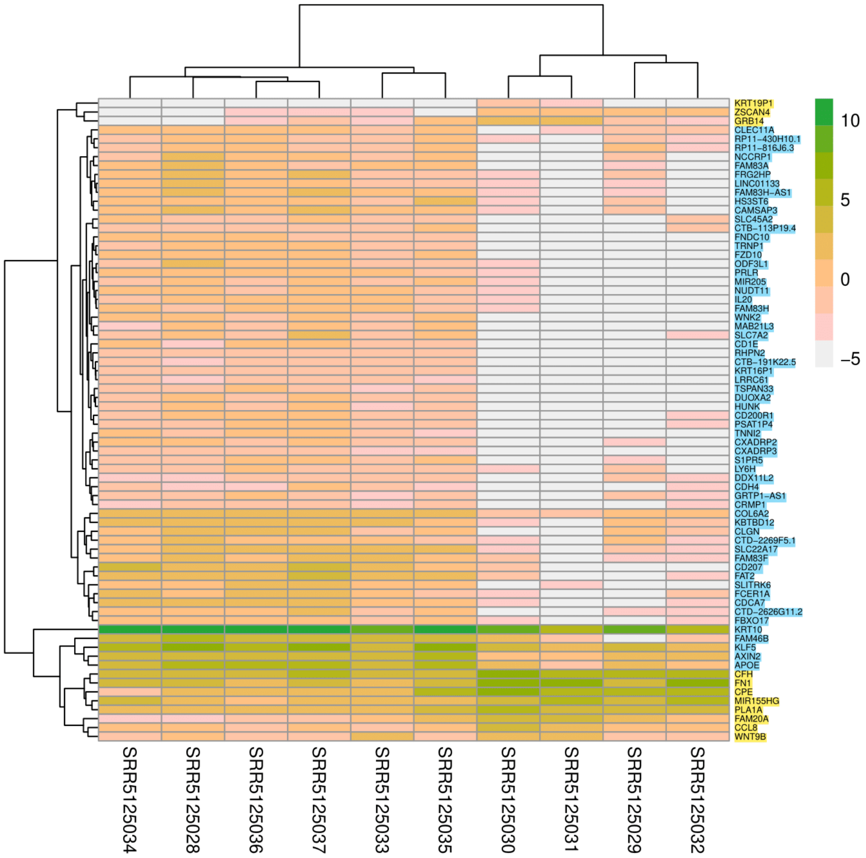

Figure 7: Hierarchical clustering of DE genes. This heatmap cluster genes by their expression values in the samples. Gene names highlighted in blue are those that belong to bicluster 1 and the ones in yellow to bicluster 2.

Discussion

Biclustering clearly achieved to obtain biclusters within the data and almost completely reproduce gene clusters previously obtained through hierarchical clustering methods [10], however, the similarities do not lie in the methods themselves but in the data. Since the data consists of differentially expressed genes, they are already expected to cluster into different groups according to their expression values across the samples, so it makes sense that *biclust* set all the genes into one of the two non-overlapping biclusters found. Another factor that could explain these similarities is having clustered rows and columns in the heatmap done by hierarchical clustering, taking the essential feature of biclustering.

Nonetheless, the fact of having a bicluster involving all 10 samples makes it non-informative for a DGE analysis since the main purpose of these type of studies is to find groups of genes that change their expression in certain samples with respect to the others so a gene with similar values in all samples can not account for differential expression, contradicting the initial and main feature of these presumed DE genes. The ideal scenario would have been to find two biclusters: one comprising samples from patients and one from controls, and both without shared genes. Perhaps, these contradictory results are due to the interest variables when doing DGE analysis: there are sample variables in the columns (not genes in the rows) from which a model is created to cluster genes differing their expression values across them. In simpler words, there are the columns that have a major importance when clustering. Also, having so many genes with similar expression patterns in the samples (even if genes have different expression values across them) accounts for it because CC biclustering interprets them as an homogeneous cluster, that was the case of bicluster 1.

Conclusion and future prospects

This work shows that defining the biological question and knowing the data features are two major issues in this study. This, in turn, determines the appropriate approaches or methods to use in order to get results. Here, CC biclustering did not prove to be a good method to extract biclusters from DE genes data, at least, with this particular set of genes and samples. However, one particular case does not account for the full demonstration, this same analysis must be applied to a broader set of genes not necessarily differentially expressed but normalized and comprissing well defined samples or conditions. Other biclustering methods must be tested as well. Finally, *biclust* R package is a good option to do this type of analysis providing a full range of biclustering methods options and visualization tools.

References

1. IGI Global (n.d.). *What is Biclustering.* Web site: What is Biclustering
2. Liew, A. W., Gan, X., Law, N. F., & Yan, H. (2015). *Bicluster Analysis for Coherent Pattern Discovery.* In M. Khosrow-Pour, D.B.A. (Ed.), Encyclopedia of Information Science and Technology, Third Edition (pp. 1665-1674). IGI Global.
3. Yuniarto, B., & Kurniawan, R. (2017). *Understanding structure of poverty dimensions in east java: Bicluster approach.* Signifikan: Jurnal Ilmu Ekonomi, 6(2), 289-300.
4. Zhou, R., Parida, L., Kapila, K., & Mudur, S. (2007). *PROTERAN: animated terrain evolution for visual analysis of patterns in protein folding trajectory.* Bioinformatics, 23(1), 99-106.
5. Pontes B, Giráldez R, Aguilar-Ruiz JS. Biclustering on expression data: A review. J Biomed Inform. 2015 Oct;57:163-80. doi: 10.1016/j.jbi.2015.06.028. Epub 2015 Jul 6. PMID: 26160444.
6. Yu Zhang, Juan Xie, Jinyu Yang, Anne Fennell, Chi Zhang, Qin Ma, *QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data*, Bioinformatics, Volume 33, Issue 3, 1 February 2017, Pages 450–452,
7. Yang, J., Wang H., Wang, W., and Yu, P. *Enhanced Biclustering on Expression Data* In Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering (BIBE), 321-327. 2003.
8. Kaiser, S., & Leisch, F. (2008). *A toolbox for bicluster analysis in R.*
9. Reisner, J., Pham, H., Olafsson, S., Vardeman, S., & Li, J. (2019). *biclustermd: An R Package for Biclustering with Missing Values.* R J., 11(2), 69.
10. González, D. (2022). *Differential expression analysis of dermal endothelial cells of healthy and type 2 diabetic patients.* Available at https://github.com/daianna21/RNAseq_2022_Finalproject/commit/a0ab1d71cc43d8a8898b5b17b31c048ae79dfca0