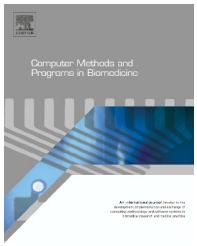
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 112 (2013) 599–606

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



pairheatmap: Comparing expression profiles of gene groups in heatmaps

Xiaoyong Sun[a,*], *Jun Li*[b,c]

[a] *Eugene Mcdermott Center For Human Growth and Development, The University of Texas Southwestern Medical Center, 6000 Harry Hines Boulevard, Dallas, TX 75390, USA*
[b] *Faculté de Pharmacie, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, Québec H3C 3J7, Canada*
[c] *Centre de Recherche Mathématiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, Québec H3C 3J7, Canada*

ARTICLE INFO

Article history:
Received 30 July 2012
Received in revised form
30 May 2013
Accepted 17 July 2013

Keywords:
Heatmap
Gene expression
Gene groups

ABSTRACT

This paper presents a new visualization software called pairheatmap, which is able to generate and compare two heatmaps so as to compare expression patterns of gene groups. It adds a conditioning variable such as time to the heatmap, and provides separate clustering for row groups in the first heatmap in order to visualize pattern changes between two heatmaps. pairheatmap is developed in R statistical environment. It provides: (1) the flexible framework for comparing two heatmaps; and (2) high-quality figures based on R package grid. The general architecture can be efficiently incorporated into bioinformatics pipeline. The package and user documentation are free to download at http://cran.r-project.org/web/packages/pairheatmap/index.html.

1. Introduction

It has been recognized that single-gene analysis has limitation for understanding the biological mechanism which involves hundreds of genes for even a simple biological function [1]. These genes do not work independently. Gene group-based analysis such as gene ontology (GO) analysis, on the other hand, has increasingly drawn attention for discovering the hidden links overlooked by the single-gene analysis. With emerging of high throughput technology, such as microarray and sequencing, researchers can simultaneously quantify thousands of genes and analyze these kinds of data within the context of GO. After Mootha et al. developed gene set enrichment analysis for gene expression data [2], many methods were developed to group genes as gene set based on biological knowledge, such as biological pathways, functions or phenotypes. In addition, gene prioritization has also been proposed. Many gene prioritization algorithms were developed for investigating heterogeneous data. By grouping genes based on functional similarity from various resources, these methods aimed to explore biological functions from different biological layers instead of investigating data from single sources. Grouping genes with specific research interests has become an important biological routine.

Heatmap, as a visualization tool for data matrix, is well-recognized in the biological field. Since Eisen et al. [3] introduced heatmap to visualize the gene groups based on biological functions, heatmap has been widely accepted as one of the main visualization tools for high-throughput data. By showing genes as rows, and conditions as columns, the gene expression values can be visualized with the density of

* *Corresponding author*. +1 214 648 5859.
E-mail addresses: sunx1@alumni.iastate.edu, johnsunx1@gmail.com (X. Sun).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.07.010

various colors. At the same time, hierarchical clustering can been applied to reorder the rows and/or columns so that similar genes or conditions with similar patterns can be grouped together. Dendrograms for rows/columns are drawn to show the hierarchy.

Clustering algorithm in heatmap has been one of the most important research topics for the last twenty years. The classical clustering algorithm in heatmap includes hierarchical clustering [4], k-means clustering [5], etc. All these methods investigated the expression pattern from global scale, and proved to be valuable in the biological research. However, the assumption for consistent pattern across conditions may not be the real case in the biology. Many researchers found many genes have co-regulated or co-expressed pattern in some conditions, but not in other conditions, suggesting that the sub-matrix or sub-pattern may be of great value to study this inconsistency across conditions. This discovery led to many algorithms developed on "biclustering" to search for sub-patterns inside a subset of genes and conditions.

In addition, some gene may be involved in several related biological groups while the classical algorithm assigned one gene to only one group. To allow gene clusters to overlap, Hastie proposed a gene shaving methods to search high variation across the samples and high correlation across the genes [6]. The method can be either "unsupervised" or "supervised". Lazzeroni developed an algorithm by combining clustering and ANOVA methods to address cluster overlapping [7]. Bergmann developed an iterative signature algorithm and introduced an "transcription module" (TM) notation to discuss the overlapping [8]. This module can be searched iteratively until certain criteria are met.

The purpose of this paper is to provide an R package to visualize gene expression data with three dimensions based on biclustering method. By comparing two heatmaps in one framework, we actually add the third dimension, i.e., a conditioning variable, so that researchers can achieve more comprehensive and dynamic view of the gene expression profile. In addition, this software allows separate group clustering for rows (genes) based on prior biological knowledge, such as gene ontology (GO). It advances our understanding in the patterns or relations among pre-determined gene groups when the third scale changes.

2. Methods

2.1. Matrix formulation

pairheatmap consists of two heatmaps represented by two data matrices. To cluster two data matrices simultaneously, we specify D1 be a $n \times p1$-dimensional data matrix, D2 a $n \times p2$-dimensional data matrix, g the number of the row groups. In gene expression analysis, the row of data matrix represents genes and the column shows conditions. a_{ij} is the expression value of gene i in condition j for the first data matrix; b_{ij} is the expression value of gene i in condition j for the second data matrix (Table 1). To simplify the formulation, we simply call the classical heatmap method as single heatmap approach, and the methods discussed here as two heatmap approach.

In this framework, we actually add an additional variable, i.e., a conditional variable in the column to split one heatmap to two heatmaps. Generally this third variable can be time or treatment in practice. Simultaneously, we can further divide genes into groups to display gene ontology or biological pathway information.

2.2. Clustering algorithm

Based on the definition above, we treat D1 as a "standalone matrix" for row clustering. After clustering the rows of D1 with the specified clustering algorithm, we apply the clustering results, i.e., the row order to the rows of D2. As for scaling, the two data matrices can be either scaled separately for comparison within matrix or scaled together for comparison across matrix.

Algorithm 1. Row clustering algorithm
Input: D1, D2
1. for $i = 1$ to g do
2. S_i = set of rows belonging to group g in D1.
3. $rowInd_i$ = the row order after clustering S_i in row.
4. $rowInd = (rowInd_1, rowInd_2, \ldots, rowInd_g)$
5. reorder the rows of D1, D2 based on rowInd
6. return the reordered D1, D2

On the other hand, the clustering algorithm for column is comparatively straightforward. In each heatmap, there is no group structure in the column. Since D1 may not have same number of columns as D2, the algorithm for clustering columns can take two approaches: (1) clustering D1 first, and then using the same column order for D2 and (2) clustering D1 and D2 independently.

Algorithm 2. Column clustering algorithm.
Input: D1, D2
1. If (ClusterColTogether)
2. $D1.colInd$ = column order after clustering column of D1
3. if (p1 > p2)
4. $D2.colInd = (D1.colInd_1, D1.colInd_1, \cdots, D1.colInd_{p2})$
5. else if (p1 < p2)
6. $D2.colInd = (D1.colInd_1, D1.colInd_1, \cdots, D1.colInd_{p1}, D2.colInd_{p1+1}, \cdots, D2.colInd_{p2})$
7. else
8. $D2.colInd = D1.colInd$
9. else
10. $D1.colInd$ = column order after clustering column of D1
11. $D2.colInd$ = column order after clustering column of D2
12. reorder columns of D1, D2 based on $D1.colInd$, $D2.colInd$ respectively
13. return the reordered D1, D2

2.3. Graphic layout

Based on R graphical package grid [9], pairheatmap takes advantage of different graphic objects, and transforms them