Volume 17, Issue 4, 2021

- ISSN: 1744-5485
- eISSN: 1744-5493

Editor in Chief

- **Pan**, Yi, Georgia State University, USA
  (editor.pan@hotmail.com)

Executive Editor

- **Wang**, Jianxin, Central South University, China

Editorial Board Members

- **Basu**, Mitra, United States Naval Academy, USA
- **Borodovsky**, Mark, Georgia Institute of Technology, USA
- **Bystroff**, Chris, Rensselaer Polytechnic Institute, USA
- **Chan**, Keith C.C., The Hong Kong Polytechnic University, Hong Kong SAR, China
- **Chen**, Jake Y., University of Alabama at Birmingham (UAB), USA
- **Chen**, Luonan, Shanghai Institutes for Biological Sciences, China
- **Claridge**, Ela, University of Birmingham, UK
- **Corne**, David, Heriot-Watt University, UK
- **Dehne**, Frank, Carleton University, Canada
- **Guerra**, Concettina, University of Padova, Italy
- **Guo**, Xuan, University of North Texas, USA
- **Harrison**, Robert, Georgia State University, USA
- **He**, Jieyue, Southeast University, China
- **He**, Matthew, Nova Southeastern University, USA
- **Hsu**, Wen-Lian, Academia Sinica, Taiwan, Province of China
- **Li**, Min, Central South University, China
- **Liang**, Jie, Univeristy of Illinois at Chicago, USA
- **Liu**, Jun S., Harvard University, USA
- **Luo**, Jingchu, Peking University, China
- **Mandoiu**, Ion, University of Connecticut, USA
- **Miyano**, Satoru, The University of Tokyo, Japan
- **Narasimhan**, Giri, Florida International University, USA
- **Nikravesh**, Masoud, University of California, USA
- **Niu**, Tianhua, Harvard Medical School, USA
- **Nussinov**, Ruth, Tel Aviv University, USA
- **Park**, Haesun, National Science Foundation, USA
- **Priami**, Corrado, Università di Trento, Italy
- **Rigoutsos**, Isidore, Thomas Jefferson University, USA
- **Singh**, Mona, Princeton University, USA
- **Srinivasan**, N., Indian Institute of Science, India
- **Szpankowski**, Wojciech, Purdue University, USA
- **Valencia**, Alfonso, Centro Nacional de Biotecnologia (C.N.B. - C.S.I.C.), Spain
- **Wang**, Jason T. L., New Jersey Institute of Technology, USA
- **Wolfson**, Haim J., Tel-Aviv University, Israel
- **Wu**, Weili, University of Texas at Dallas, USA
- **Xu**, Ying, University of Georgia, USA
- **Yang**, Jack Y., University of California at San Diego, USA

- **Yao**, Xin, The University of Birmingham, UK
- **Zelikovsky**, Alexander, Georgia State University, USA
- **Zhang**, Aidong, University at Buffalo, The State University of New York, USA
- **Zhang**, Louxin, National University of Singapore, Singapore
- **Zheng**, Hao, Hangzhou Nuowei Information Technology, China
- **Zimmermann**, Karl-Heinz, Technical University Hamburg-Harburg, Germany
- **Zomaya**, Albert, University of Sydney, Australia

DAFTAR ISI

(Source:https://www.inderscience.com/info/inarticletoc.php?jcode=ijbra&year=2021&vol=17&issue=4)

Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm

Titin Siswantining, Alhadi Bustamam*, Sofia Debi Puspa and Zuherman Rustam

Faculty of Mathematics and Natural Sciences,
Department of Mathematics,
Universitas Indonesia,
Pondok Cina, Depok, 16424, Indonesia
Email: titin@sci.ui.ac.id
Email: alhadi@sci.ui.ac.id
Email: sofiadebipuspa@sci.ui.ac.id
Email: zuherman@sci.ui.ac.id
*Corresponding author

Fahrezal Zubedi

Faculty of Mathematics and Natural Sciences,
Department of Mathematics,
Universitas Negeri Gorontalo,
Gorontalo, 96128, Indonesia
Email: fahrezal@ung.ac.id

Abstract: Similarity-based biclustering (SBB) algorithm consists of four main phases, transforming data, the construction of row (gene) and column (condition) similarity matrices, the clustering of each similarity matrix and the extraction of the bicluster. In this study, we modified the SBB algorithm at the stage of data transformation using min-max normalisation to identify significant biclusters in diabetic nephropathy and retinopathy microarray data after genes are selected using relative deviations and absolute deviations. Based on the comparison of the silhouette index validation experiments, SBB using partitioning around medoids (PAM) provided better clustering of genes and samples than K-means and agglomerative hierarchical clustering (AHC) (Ward's linkage). Furthermore, the proposed technique identified a meaningful non-overlapping bicluster on a real dataset. Using gene ontology (GO) enrichment analysis and the Bonferroni correction, we have identified biological evidence in each bicluster that is significant in terms of gene functions and biological processes.

Keywords: agglomerative hierarchical clustering; biclustering; diabetic nephropathy; diabetic retinopathy; gene expression; K-means; microarray data; PAM; partitioning around medoids; SBB; similarity-based biclustering.

Reference to this paper should be made as follows: Siswantining, T., Bustamam, A., Puspa, S.D., Rustam, Z. and Zubedi, F. (2021) 'Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm', *Int. J. Bioinformatics Research and Applications*, Vol. 17, No. 4, pp.343–362.

Biographical notes: Titin Siswantining received her Bachelor of Science degree in statistic from Sepuluh November Institute of Technology, Indonesia in 1984. The Master's degree in Applied Mathematics from Ecole Des Hautes Etudes En Sciences Sociales in 1990 and the PhD in Statistics from Bogor Agricultural Institute in 2013. Her research focuses on statistics and bioinformatics.

Alhadi Bustamam received his BSc (honour) degree in Computational Mathematics in 1996, the Master's degree in Computer Science from Universitas Indonesia in 2002, and the PhD in Bioinformatics from the University of Queensland, Australia in 2011. His research focuses on computational mathematics, computational biology, bioinformatics and computer sciences**.**

Sofia Debi Puspa received her Bachelor of Education degree in Mathematics from Universitas Negeri Jakarta in 2016, Master's degree in Mathematics from the Universitas Indonesia in 2018. Her research interests include bioinformatics.

Zuherman Rustam received his BSc in Mathematics in 1984, Master degree and PhD in Mathematics Science from Universitas Indonesia in 2006. His research focuses on bioinformatics and machine learning.

Fahrezal Zubedi received his Bachelor of Education degree in Mathematics from Universitas Negeri Gorontalo in 2016, Master's degree in Mathematics from the Universitas Indonesia in 2018. His research interests include bioinformatics.

This paper is a revised and expanded version of a paper entitled 'Implementation of co-similarity measure on microarray data of lymphoma by using K-means partition algorithm' presented at *3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (3rd ISCPMS 2017)*, 26–27 July, 2017.

1 Introduction

Global increases in the prevalence of diabetes mellitus (DM) are accompanied by the increased prevalence of diabetic microvascular disorders, such as diabetic nephropathy (DN) and diabetic retinopathy (DR). DN is a kidney abnormality that leads to decreased renal function in subjects with a history of DM. DM, particularly type II diabetes, is the sixth most common cause of death due to kidney failure in western countries. According to a microalbuminuria prevalence study (MAPS) in Asia, nearly 60% of patients with type 2 diabetic hypertension suffer from DN, with macroalbuminuria and microalbuminuria accounting for 18.8% and 39.8% of cases, respectively. DN is characterised by proteinuria (microalbuminuria), which precedes DN by months to years. Therefore, Foster (1994) has suggested that the early detection of microalbuminuria

allows interventions that inhibit further decline in renal function. DR currently affects more than 20 million diabetic people in the US The associated microvascular damage to the retina is characterised by the death of retinal neurons and decreased retinal vascularisation (Eldred and Katz, 2008). Although not lethal, this condition can decrease vision and lead to permanent blindness (Wilardjo, 2001).

Recent technological advances in molecular biology have greatly facilitated analyses of diverse biological phenomena. For example, microarray technologies have been used to profile changes in relative gene expression levels associated with DN and retinopathy. Microarray technology utilises gene sequences determined in the genome project to analyse specific organisms at any given time and under specific conditions. The microarrays return expression matrices for specific biological conditions, with genes in rows and expression levels in columns (Madeira and Oliveira, 2004).

Gene expression analyses can be performed in various ways, including by grouping data using clustering and co-clustering (biclustering) algorithms. Liu et al. (2014) have described clustering as a data classification process that partitions the data according to various criteria to find patterns. However, clustering is limited by the assumption that genes have the same expression levels under all measurable conditions. To address this limitation, bioinformatics analyses are conducted using biclustering or co-clustering algorithms. Prelić et al. (2006) simultaneously applied biclustering algorithms in two dimensions to identify subset pairs of genes and conditions of gene expression. Cheng and Church (2000) were the first to apply the metric block correlation method mean residual score (MRS), which measures the quality of the biclusters found. Subsequently, several studies have proposed algorithms based on MRS, such as that by Divina and Aguilar-Ruiz (2006). Several studies have shown the successful implementation of biclustering algorithms on Alzheimer's microarray data, such as Setyaningrum et a1. (2019), Wibawa et al. (2019) and Wutun et al. (2019).

One biclustering technique that can detect sets of genes with similar expression levels under certain conditions extracts biclusters according to biological functions and is known as similarity-based biclustering (SBB). Hussain and Ramazan (2016) generated biclusters that show gene expression levels in cellular processes. Their analyses define medically relevant groups of genes under certain conditions and can be used to inform prevention and treatment strategies.

Hussain and Ramazan (2016) implemented an SBB algorithm using agglomerative hierarchical clustering (AHC) to find biclusters. They also compared it with other bicluster algorithms such as ISA, SAMBA, Bimax, Cheng and Church algorithm and the other. The results of his research that AHC (Ward's linkage) produces the highest accuracy than other bicluster algorithms. However, AHC has a high complexity of $O(n^3)$, where n is the number of data points. Therefore, it is less suitable for application in large-scale gene expression datasets (Sasirekha and Baby, 2013). This paper explains that the weakness of the SBB algorithm lies in the clustering stage, namely the AHC method, and recommends using partition-based clustering methods. Partition-based clustering consists of several grouping methods, such as k-means clustering and partitioning around medoids (PAM). Popular clustering techniques, including K-means clustering and PAM, have been used to classify gene expression data.

Unlike hierarchical clustering algorithms, partition-based clustering methods begin by classifying the entire dataset into k partitions. After initial partitioning of data into k-clusters, heuristics are used to improve the grouping based on some objective functions.

The K-means clustering algorithm is designed to partition objects in the dataset into subsets so that all the points in particular subsets are closest to a given centre (Macqueen, 1996). The advantages of this approach include ease of implementation, comparatively less time to run the algorithm and adaptability. This algorithm has been commonly used, for example, by Wu and Kumar (2009). PAM is a widely used and powerful clustering technique (Park and Jun, 2009) and has the advantage of efficiency during application to large datasets. Bustamam et al. (2018) conducted clustering research using AHC and produced the best grouping accuracy. Therefore, we compared the PAM, K-means partitioning and AHC (Ward's linkage) methods to implement the SBB algorithm and identify significant biclusters in gene expression data.

We used a min–max normalisation method to transform the data. According to Chin et al. (2015), the main advantage of the min–max normalisation method is that it reflects a balance of comparative values before and after the normalisation process, thus eliminating biases in the data.

The remainder of this paper is organised as follows. The datasets and study method are described in Section 2. The results and the analysis are presented in Section 3. The purpose of this paper is to modify the clustering part of the SBB algorithm with k-means and PAM algorithm and compare their performance with the original one, i.e., SBB using AHC (Ward's Linkage) to obtain the biclusters from microarray gene expression data. Further, we extract important biological information from the biclusters which the Bonferroni correction provided by the database for annotation, visualisation and integrated discovery (DAVID) by Huang et al. (2009).

2 Datasets and methods

Diabetic gene expression data from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) can be accessed at https://www.ncbi.nlm.nih.gov/geo. The DN dataset was used with the GSE1009 ID, which contains gene expression profiles of kidney glomeruli from DN patients and comprises 12,626 genes from six samples. The DR dataset comprises 45,102 genes, and five samples and can be accessed using the code GSE12610 ID.

The following steps in the data analysis comprise the matrix of gene expression data A. We selected genes using absolute and relative deviation methods. We then used the min–max normalisation method to transform matrix A into matrix A_n and the chi similarity measure to generate a similarity matrix of gene, i.e., rows (SR) and a similarity matrix of condition, i.e., column (SC). This similarity matrix was then clustered into k and l clusters using PAM and K-means, and the best cluster was determined using the index silhouette method. The extract bicluster method was then applied to this cluster to determine the regulatory expression levels of each bicluster. The following flowchart (Figure 1) shows the data analysis steps.

2.1 Gene selection

Microarrays produce large-scale data that generally have experimentally generated noise due to, for example, transformation of the image, image segmentation, selection of parameters, different numbers of Cy3 and Cy5 on labelling mRNAs and inequality of initial RNA quantities. Such noise has a strong effect on the accuracy of grouping. In

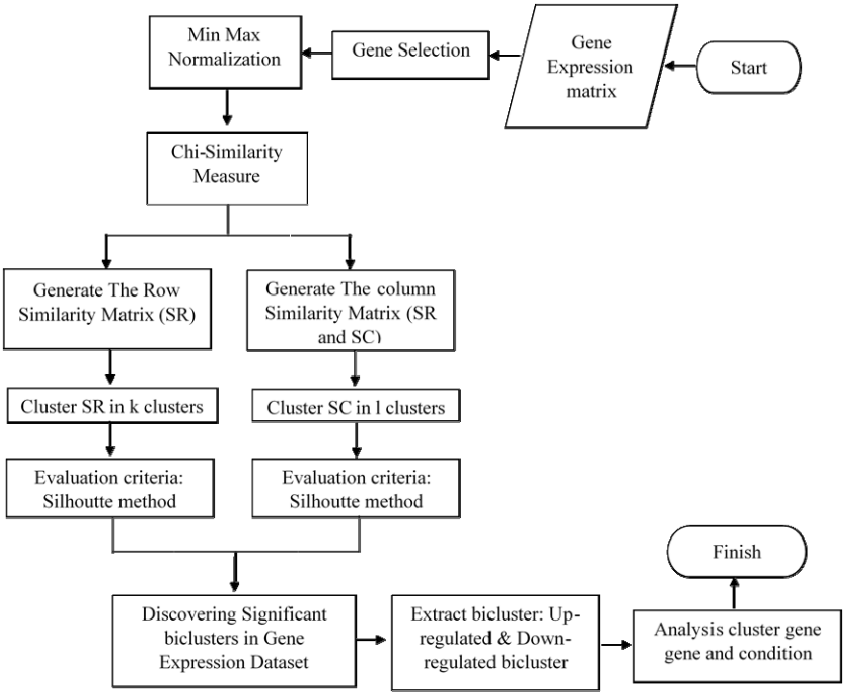
addition, greater numbers of features (genes) increase the time and cost of computing required to process the data. Therefore, genes are selected according to significant differential gene expression levels. We used two popular and simple selection techniques involving relative deviations and absolute deviations. In relative deviation techniques, selected genes have relative deviations that are greater than the threshold (δ), and genes with smaller deviations are considered noise, as described by Klebanov and Yakovlev (2007). The formula for relative deviations is as follows:

$$\delta \le \left| \frac{\max(a^i)}{\min(a^i)} \right| . \tag{1}$$

Golub and Van Loan (1999) have suggested that absolute deviation techniques considers the range of ranges of absolute values for all conditions. Genes that are selected using this technique have absolute deviations that are greater than the threshold (θ). Absolute deviation is calculated as follows:

$$\theta \le \left| \max(a^i) - \min(a^i) \right| . \tag{2}$$

Figure 1 Flowchart of the data analysis



2.2 Min–max normalisation

In this normalisation technique, data are transformed using minimum and maximum values. Min–max normalisation alters matrix entries to the interval [0,1], as indicated by Chin et al. (2015). The min–max normalisation formula reported by Patro and Sahu (2015) is as follows:

$$x'_{(a,b)} = \frac{x_{(a,b)} - \min(x_a)}{\max(x_a) - \min(x_a)} \times (new_{\max} - new_{\min}) + new_{\min}, \quad (3)$$

where $x'_{(a,b)}$ is the element value (a,b) in the data matrix after normalisation, $x_{(a,b)}$ is the element value (a,b) in the data matrix before normalisation, $\min(x_a)$ is the minimum value in the A^{th} row of the matrix before normalisation, $\max(x_a)$ is the maximum value in the A^{th} row of the matrix before normalisation, $new_{\max}$ is the limit of the new maximum value and $new_{\min}$ is the limit of the new minimum value.

2.3 X-Sim similarity measure

The χ-Sim similarity measure is a co-similarity-based approach that simultaneously uses two similarity matrices of similarity matrix between row (SR) and similarity matrix between column (SC). χ-Sim is calculated iteratively by updating these SR and SC matrices, and the resulting updated SR and SC matrices are interrelated so that each matrix construct has a similarity given by another co-similarity measure, as suggested by Hussain et al. (2010).

Let A be the data matrix representing a gene expression microarray with r rows (genes) and c columns (conditions). The intensity of gene activities between the Ith row and the Jth column is denoted by $a_{i:} = \left[a_{i1, \ldots,} a_{ic} \right]$, which is the row vector representing the gene i, and $a_{j:} = \left[a_{j1, \ldots,} a_{jc} \right]$, which is the column vector corresponding to condition j. In these computations, SR and SC represent square and symmetrical row and column similarity matrices of size r x r and c x c, respectively, with $\forall i, j = 1 \ldots r, sr_{ij} \in [0,1]$ and $\forall i, j = 1 \ldots c, sc_{ij} \in [0,1]$. In this equation, $F_s(\ldots)$ is a generic function that takes the elements a_{il} and a_{jn} of A and returns a measure of similarity $F_s(a_{il}, a_{jn})$ between them. The similarity (or distance) measure between two genes $a_{i:}$ and $a_{j:}$ is defined as a function and is denoted as $Sim(a_{i:}, a_{j:})$ to calculate the SR matrix (Hussain and Ramazan, 2016), as follows:

$$Sim\left(a_{i:}, a_{j:}\right) = F_s\left(a_{i1}, a_{ji}\right) + F_s\left(a_{i2}, a_{j2}\right) + \cdots + F_s\left(a_{ic}, a_{jc}\right). \quad (4)$$

Given a matrix SC with entries that provide a measure of similarity between columns (condition) of microarray data, we introduce a pseudo-norm λ that is analogous to the norm L_λ (Minkowski distance). Equation (4) can then be rewritten as follows without changing its meaning if $sc_{ll} = 1$ and $\lambda = 1$ (Hussain and Ramazan, 2016):

$$Sim\left(a_{i:}, a_{j:}\right) = F_s\left(a_{i1}, a_{ji}\right).sc_{11} + F_s\left(a_{i2}, a_{j2}\right).sc_{22} + \ldots + F_s\left(a_{ic}, a_{jc}\right).sc_{cc} \quad (5)$$

$$Sim\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\sum_{l=1}\left(F_s\left(a_{il}, a_{jl}\right)^{\lambda} \times sc_{ll}\right)}. \quad (6)$$

Equation (5) generalises (6) so that all possible pairs of features (genes) occurring under conditions $a_{i:}$ and $a_{:j}$ can be accounted for. The overall similarity between genes $a_{i:}$ and

$a_{:j}$ is defined in the following equation (7), in which the terms for $l = n$ are those in equation (6):

$$Sim^{\lambda}\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\sum_{l=1}^{c}\sum_{n=1}^{c}\left(F_s\left(a_{ll}, a_{jn}\right)^{\lambda} \times sc_{\ln}\right)}. \tag{7}$$

Assuming that the function $Sim(a_{il}, a_{jn})$ is defined as a product of the elements a_{il} and a_{jn} ($Sim(a_{il}, a_{jn}) = a_{il} x a_{jn}$, as with the cosine similarity measure), we can rewrite equation (7) as follows:

$$Sim^{\lambda}\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\left(a_{i:}\right)^{\lambda} \times sc \times \left(a_{j:}^{T}\right)^{\lambda}}, \tag{8}$$

where $\left(a_{i:}\right)^{\lambda} = \left[\left(a_{ij}\right)^{\lambda} ... \left(a_{ic}\right)^{\lambda}\right]$ and $a_{j:}^{T}$ denotes the transpose of the vector $a_{j:}$, as described by Hussain and Ramazan (2016).

2.4 The generic χ-Sim co-similarity measure

In general, expansion of all row and column pairs is achieved by matrix multiplication from SR and SC matrices. According to Hussain et al. (2010), the steps of this method are as follows:

a Similarity matrices of the gene or row (SR) and similarity matrix of condition or column (SC) are initiated with the identity matrix. We can write these matrices as $SR^{(\circ)}$ and $SC^{(\circ)}$, where the superscript is the iteration number.

b At each iteration *t*, we calculate the new similarity matrix between genes $SR^{(t)}$ using the similarity matrix between conditions $SC^{(t-1)}$, and we do the same thing for the columns' similarity matrix $SC^{(t)}$, as follows:

$$SR^{(t)} = A^{o\lambda} \times SC^{(t-1)} \times (A^{o\lambda})^{T}, \tag{9}$$

$$SC^{(t)} = (A^{o\lambda})^{T} \times SR^{(t-1)} \times A^{o\lambda}, \tag{10}$$

where $A^{o\lambda} = ((a_{ij})^{\lambda})_{i,j}$ is the element-wise exponent of *A*. Furthermore, each SR and SC matrix is normalised using the following pseudo-normalisation:

$$\forall i, j \in 1...r, sr_{ij}^{(t)} = \frac{\sqrt[\lambda]{sr_{ij}^{(t)}}}{2\sqrt[\lambda]{sr_{ii}^{(t)} x sr_{jj}^{(t)}}}, \tag{11}$$

$$\forall i, j \in 1...c, sc_{ij}^{(t)} = \frac{\sqrt[\lambda]{sc_{ij}^{(t)}}}{2\sqrt[\lambda]{sc_{ii}^{(t)} x sc_{jj}^{(t)}}}. \tag{12}$$

c $SR^{(t)}$ and $SC^{(t)}$ are updated iteratively ($t = 4$ is sufficient).

2.5 Partitioning around medoids

PAM uses the k-medoid method to identify clusters and chooses a medoid by calculating minimum total distances from others (Mondal and Choudhury, 2013). According to Cahyaningrum et al. (2017), a medoid is a representative of each group, and the number of medoids is the same as the number of clusters. The data distance formula with *n*-dimensional data is calculated using Euclidean distance, a distance matrix that adopts the Pythagoras principle. This is because the calculation pattern uses the rules of rank and the square root. Euclidean computations give relatively small distances because they use the square root rule, as follows:

$$d(x_a, x_b) = \sqrt{\sum_{k=1}^{n} (x_{ak} - x_{bk})^2} \ . \tag{13}$$

The purpose of PAM is to determine a representative object–a medoid–for each cluster that has a minimum total distance to other objects. PAM is performed in two stages, building and swapping. The building stage is a sequential object selection process that continues until the object is found. The swap stage is performed to improve group quality by exchanging selected objects for unselected objects (Mondal and Choudhury, 2013).

Kaufman and Rousseeuw (1990) have described the steps of the building stage as follows:

1 Suppose object *i* has not been selected.

2 Identify the unselected object *j* and calculate the difference between D_j (distance of object *j* from the previously selected object) and *j* (distance between objects *j* and *i*).

3 If the distance in step 2 is positive, then object *j* will contribute to the selection of object *i*. Then count

$$C_{ji} = \max(D_j - d(j,i), 0) \tag{14}$$

4 Calculate the total gain from the selection of object *i.*

$$\sum_j C_{ji} \tag{15}$$

5 Select an object that has not been selected *i.*

$$\underset{i}{\text{maximising}} \sum_i C_{ji} \tag{16}$$

6 Repeat steps 1 to 5 until *k* objects are found.

These steps can be used to calculate the effects of swapping between *i* and *h* objects on clustering results, as described by Kaufman and Rousseeuw (1990) in the following instructions:

1 Assume that object j was not selected and count the contribution of C_{jih} to the swap as follows:

   a If the distances between objects *j*, *i* and *h* are greater than between *j* and one of the representative objects, then $C_{jih} sil(x_i)$ 0.

b If the distance between objects *j* and *i* is further than the distance between *j* and any other selected representative objects $(d(j,i) = D_j$, then the following two conditions must be fulfilled:

1) The distance of object *j* is closer to object *h* than to the second nearest representative object, as follows:

$$d(j,h) < E_j, \tag{17}$$

where E_j .is the distance between object *j* and the second-nearest representative object. Under these conditions, the contribution of object *j* to the swap between objects *i* and *h* is calculated as follows:

$$C_{jih} = d(j,h) - (d(j,i)). \tag{18}$$

2) The distance of object *j* to *h* is at least equal to that between *j* and the second closest representative object, as follows:

$$d(j,h) < E_j. \tag{19}$$

Under these conditions, the contribution of the object to the swap is calculated as follows:

$$C_{jih} = E_j - D_j. \tag{20}$$

For condition 1), if object *j* is closer to object *i* than to object *h*, which contributes positively, then the swap with object *j* is not favourable.

c If the object distance *j* to *i* is greater than the object distance *j* to at least one of the other representative objects but closer to *h* than the distance of object *j* to any representative object, then the contribution of *j* to the swap is calculated as follows:

$$C_{jih} = d(j,h) - D_j. \tag{21}$$

2 Calculate the total result of a swap by adding a contribution.

3 To determine whether a swap has occurred, select the (*i*, *h*) pair in the following equation:

$$\underset{i,h}{\text{minimising}} T_{ih}. \tag{22}$$

4 If the minimum value of $T_{ih}sil(x_i)$ s negative, then the swap has occurred, and the swap stage returns to step 1. If the minimum value of $T_{ih}sil(x_i)$ s positive, then the swap has not occurred, and the stages are stopped.

2.6 K-means algorithm

The K-means algorithm is a partitioning method to partition existing data into one or more clusters and is performed using the following steps:

1 Determine the numbers of clusters (k) and centroids.

2 Determine the distance between each object and the centroids using the Euclidean distance formula, as follows:

$$d_{ik} = \sqrt{\sum_{i,k=1}^{n} (x_i - c_k)^2}\ , \tag{23}$$

where d_{ik} s the distance of object $sil(x_i)$ and centroid., n is dimensional data, x_i is a coordinate of objects *i* an $sil(x_i)$ c_k is the centroid coordinate.

3 Cluster the objects based on the minimum distance.

4 Determine new centroids using the following formula:

$$C_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}, \tag{24}$$

where $x_{ij} \in sil(x_i)$ s the.-cluster an $sil(x_i)$ p is the number of the member of the cluster. Iterate the above steps until no object moves from its assigned group.

2.7 Index silhouette

The silhouette method is used to determine the number of clusters and to measure the accuracy of grouping results (Rousseeuw, 1987). The silhouette method measures how similar an object is to other objects in its group relative to similarities between other groups. Silhouette values are calculated as follows:

$$sil(x_i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{25}$$

where $a(i)$ is the average dissimilarity (Euclidean distance) value of the ith data point to all members of the cluster containing data., and $b(i)$ is the smallest average dissimilarity of ith data to all members of the group not containing the data i. In this study, the silhouette index $sil(x_i)$ as used to determine the number of groups by choosing the highest value of $sil(x_i)$ as reported by Kaufman and Rousseeuw (1990).

2.8 Extract bicluster

The translation of mRNAs produces proteins and leads to the production of various other components. Therefore, genes are upregulated or downregulated depending on the cellular requirements of these components.

Clustering in SR and SC matrices using hard clustering algorithms produces non-overlapping bicluster or checkerboard structures. Elements of the checkerboard are average values of the set of genes in cluster under the conditions in cluster $\hat{y}_j$, as shown in Table 1. The most significant biclusters are the non-overlapping bicluster. The number of most significant biclusters is the number of obtained biclusters.

The bicluster extract algorithm of SBB extracts significant biclusters containing an upregulated gene cluster under one cluster of conditions only. Therefore, biclusters show greater variation under various conditions occurring in a cluster of genes. Mathematically, we can express this as a, in which the cluster gene is highly regulated under the cluster condition. For every cluster condition, *m* as low regulation, and the algorithm is written as $(\hat{x}_i, \hat{y}_{m \neq j} \ \forall m \in 1,\ldots,l)$.

Table 1 Identification of significant biclusters from gene expression data

		Condition clusters			
		$\hat{y}_1$.	$\hat{y}_2$	⋯.	*max – min*
Gene clusters	$\hat{x}_1$	$\overline{\hat{x}_1 \hat{y}_1}$ $\overline{\hat{x}_1 \hat{y}_1}$	$\overline{\hat{x}_1 \hat{y}_2}$	⋯.	$max\left(\overline{\hat{x}_l \hat{y}_i}\right) - min\left(\overline{\hat{x}_l \hat{y}_i}\right), \forall i \in 1...l$
	$\hat{x}_2$	$\overline{\hat{x}_2 \hat{y}_1}$	$\overline{\hat{x}_2 \hat{y}_2}$	⋯.	$max\left(\overline{\hat{x}_2 \hat{y}_i}\right) - min\left(\overline{\hat{x}_2 \hat{y}_i}\right), \forall i \in 1...l$
	⋯				
	$\hat{x}_k$	$\overline{\hat{x}_k \hat{y}_1}$	$\overline{\hat{x}_2 \hat{y}_2}$	⋯.	

3 Results and discussion

Gene expression matrices for DN and DR after selection were 6320×6 and 3689×5, respectively. The analyses were performed using R version 3.3.1 (an open-source programming tool) on a computer with an Intel® Core™ i5 CPU@ 2.40GHz processor and 4 GB of memory (RAM). Gene selection techniques are required to filter out informative genes by choosing those with values that are greater than each threshold and discarding those with lower relative and absolute deviations. The present relative and absolute deviation values of diabetic data are shown in Table 2.

The range of gene expression data values varied greatly for each row. This variation follows differences in sample conditions associated with patients, necessitating normalisation to produce datasets with uniform scales. Microarray data for DN and DR were subjected to min–max normalisation and are shown as a heat map in Figure 2.

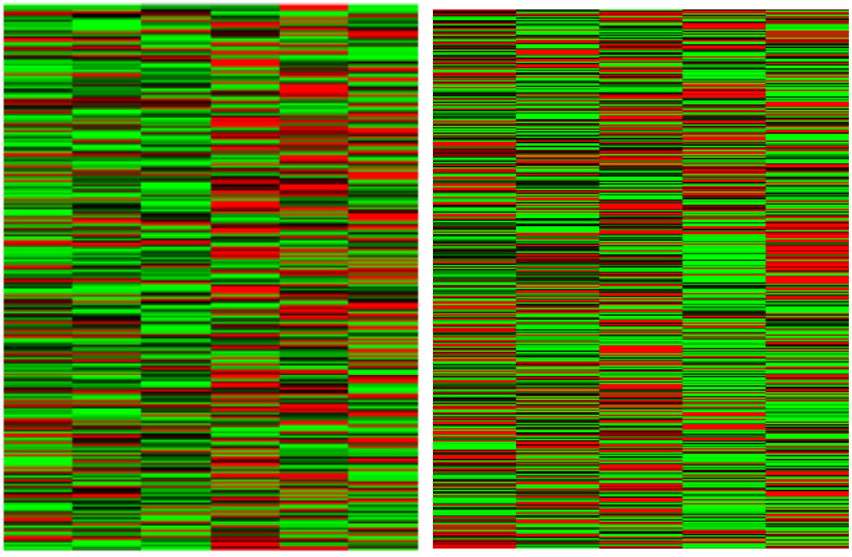
3.1 Comparison of clustering techniques

We ran the SBB algorithm for hard clustering and tested the K-means and PAM clustering techniques to obtain cluster genes and cluster conditions for SR and SC. After normalising the data, we constructed two matrices of the similarity between SR and SC using the χ-Sim co-similarity measure. Determinations of λ in pseudo-normalisations can be used to control the data distributions. The accuracy of grouping was affected by the silhouette index method. This method is needed to determine which clusters are optimal for grouping objects. Table 3 shows silhouette values for similarity matrices of SR and SC using the PAM and K-means algorithms.

Table 2 Description of gene expression datasets

Dataset	*DN*	*DR*
Number of original genes	12,626	45,102
Number of samples	6	5
Number of sample classes	2	2
Sample class names	Diabetic (3)	Diabetic (3)
	Normal (3)	Normal (2)
Relative deviation threshold	15	5
Absolute deviation threshold	4000	2000
No. of selected genes	6320	3689

Figure 2 Heat map normalisation of genes in: (a) DN and (b) DR (see online version for colours)



(a) (b)

Table 3 Silhouette values using PAM

		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.251	0.491	0.670	0.759	0.805	*0.830*	0.793	0.728	0.636
	SR	0.230	0.294	0.381	0.417	0.498	*0.538*	0.438	0.362	0.294
DR	SC	0.463	0.512	0.658	0.764	*0.785*	0.623	0.587	0.473	0.388
	SR	0.295	0.315	0.476	0.506	*0.548*	0.492	0.361	0.308	0.245

Based on Tables 3–5, the maximum threshold value of λ from both clustering algorithms for DN data in the similarity matrices of SR and SC was 0.6, whereas the maximum threshold of λ for DR was 0.5. Therefore, $\lambda = 0.6$ was used in χ-Sim co-similarity measures for DN data, and $\lambda = 0.5$ was used for DR data. For each value of $\lambda < 1$, where

$\lambda = 0.1, 0.2, ..., 0.9$, the number of clusters of the condition with the highest silhouette index value for diabetic data (DN and DR) was 2. Similarly, the number of gene clusters with the highest silhouette value was 2. Therefore, the number of optimal clusters of cluster genes and cluster conditions was 2 for the diabetic data. The silhouette index for the number of clusters for $k = 2, 3, 4, ..., 10$ is shown in Table 5; the number of clusters was selected based on the maximum value of the silhouette index (optimal).

Table 4 Silhouette values by using K-means

		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.232	0.472	0.592	0.684	0.743	*0.824*	0.781	0.674	0.601
	SR	0.226	0.285	0.376	0.372	0.785	*0.522*	0.425	0.351	0.288
DR	SC	0.418	0.485	0.646	0.748	*0.762*	0.601	0.577	0.452	0.364
	SR	0.288	0.291	0.456	0.482	*0.539*	0.487	0.329	0.256	0.237

Table 5 Silhouette values by using AHC (Ward linkage)

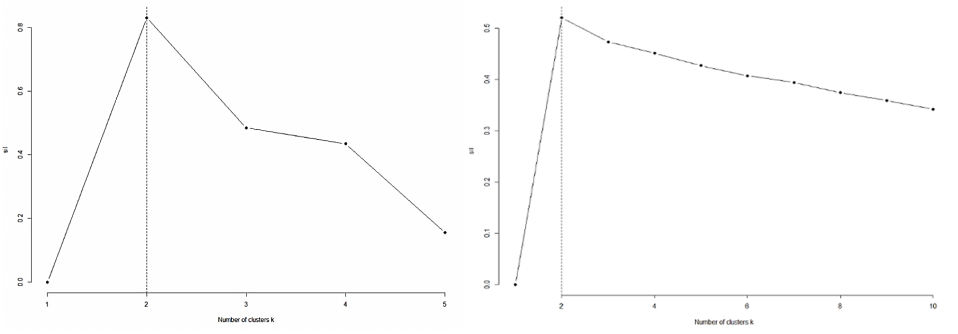
		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.228	0.463	0.558	0.641	0.726	0.754	0.731	0.653	0.587
	SR	0.218	0.263	0.358	0.362	0.736	0.520	0.418	0.337	0.252
DR	SC	0.385	0.468	0.623	0.708	0.648	0.585	0.551	0.447	0.324
	SR	0.277	0.284	0.447	0.453	0.528	0.475	0.318	0.236	0.208

Using PAM for clustering in SBB gave higher silhouette values than K-means and AHC in the range of $\lambda < 1$ for SR and SC matrices, indicating the greater clustering accuracy of the PAM method. Although the K-means algorithm has some advantages, it has a number of drawbacks. For example, K-means is sensitive to outliers and lacks a definite method for identifying optimal partition counts in the initial cluster determination, and the iterative procedure does not ensure movement toward a point of convergence, as described by Rui and Wunsch (2005). Because PAM gives better accuracy in SBB, we used this clustering algorithm for further analyses of genes and conditions. Plots of silhouette values for cluster conditions and cluster genes using PAM are shown in Figures 3 and 4. The horizontal axis represents numbers of clusters, and the vertical axis represents corresponding index silhouette values.

3.2 Analysis of gene clusters

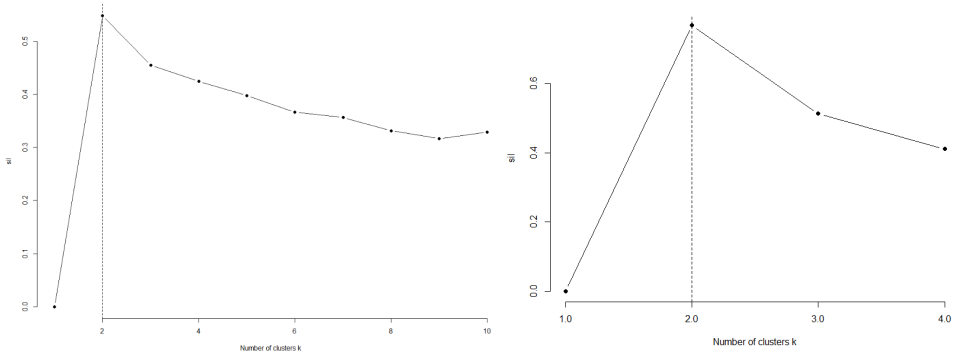
In this study, we generated a heat map to show high gene activities in the microarray dataset under certain conditions, where gene rows and column conditions were grouped simultaneously. Biclusters of similar gene expression values under certain conditions are evident in the heat map in Figure 5, which shows that the bicluster structure generated using the SBB algorithm is non-overlapping.

Figure 3 Silhouette index plots in: (a) SC and (b) SR matrices of DN



(a) (b)

Figure 4 Silhouette index plots in: (a) the SR matrix and (b) the SC matrix of DR



(a) (b)

Furthermore, the present biclusters were analysed according to expression levels using the extracted bicluster, which is the final stage of the SBB algorithm and indicates regulation that occurs in the bicluster–whether upregulation or downregulation–in quantitative terms. To support the results of our qualitative bicluster heat map (Figure 5), we conducted gene ontology (GO) enrichment analyses of cluster generation using the DAVID database at https://david.ncifcrf.gov/site.

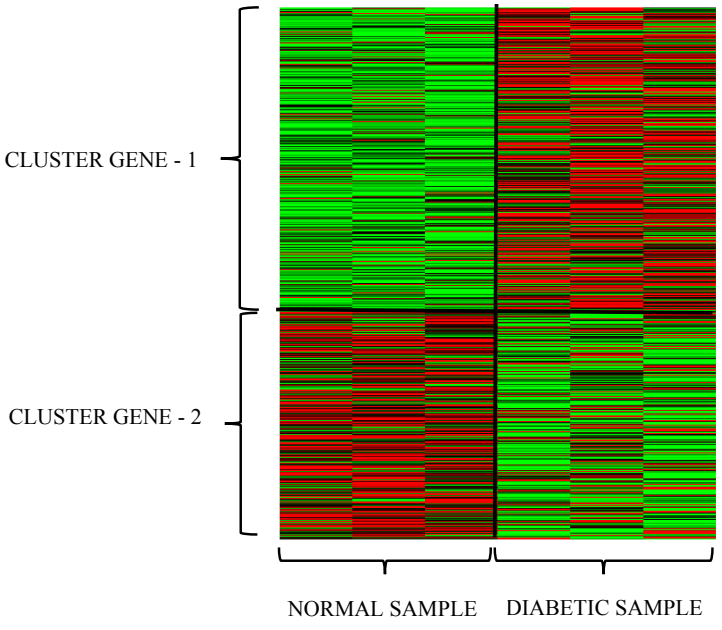
Based on clustering using the PAM algorithm, gene expression data for DN in row and column similarity matrices are grouped into two clusters. The clusters comprised 3615 genes (cluster-1) and 2705 genes (cluster-2) in SR and SC matrices, respectively, under three conditions.

Each cluster of genes corresponding to each sample becomes a bicluster. For example, in Figure 5, the cluster gene 1, which are also from the normal sample, is a submatrix from the microarray data is a bicluster. This is also applied for the cluster gene 1 and diabetic sample, cluster gene 2, and each normal sample and diabetics sample. Therefore, there are four biclusters in Figure 5 as the output.

In GO enrichments of the biological processes for each of the identified gene clusters (Table 6), cluster gen-1 under diabetic conditions had a high expression rate compared to under normal conditions. Genes in cluster-1 were identified as contributing to increases in blood glucose (hyperglycaemia) and glomerular hypertension, leading to kidney inflammation and major conditions of DN (such as high glucose and oxidative stress),

high expression of advanced glycation end-products (AGEs), angiotensin II, TGF-β, CTGF, protein kinase C, the receptor for advanced glycation end-products (RAGE) and NF-κB (Brennan et al., 2013). In contrast, under normal conditions, cluster gen-2 had higher gene expression levels than under diabetic conditions. This indicates that genes grouped using the SBB algorithm are enriched in terms of biological processes related to cellular energy metabolism, as may be expected in the diabetic dataset. Given the differences in expression levels of these genes, the diabetic sample network with expression levels that mathematically differ from those in normal tissue samples may indicate a useful focus for medical practitioners.

Figure 5 Heat map of the DN bicluster (see online version for colours)



In the DR dataset, differences in gene expression levels were evident between diabetic and normal conditions. In particular, genes in cluster-1 are upregulated in diabetic samples but downregulated in normal samples. The genes in cluster-2 are upregulated in normal samples but downregulated in diabetic samples. These gene clusters contained highly discriminative genes, many of which have been previously identified. Differences in gene expression levels between diabetic and normal conditions are shown in Figure 6. Genes of cluster-1, which were overexpressed under the conditions of DR, are known to affect blood glucose levels. Tables 7 and 8 show the biological functions of each gene cluster in diabetic nephropathy and retinopathy microarray respectively.

According to Servat et al. (2013), genetic variants in DR include those of the polyol pathway (hyperglycemia), AGE, hypoxia through vascular endothelial growth factor, RAGE, endothelial nitric oxide synthase, angiotensin I-converting enzyme and others. In contrast, genes in cluster-2 had higher expression levels in normal samples than in diabetic samples and included genes that control protein synthesis and immune activity.

Furthermore, these differences suggest the presence of novel sub-types of DR. The two bluslustering methods that we have developed, namely SBB PAM and SBB K-Means, work for gene expression data in the form of microarrays. The results of the biclusters that we obtained can be analysed further by getting the RNA sequencing (Wen et al., 2019) or Single-cell RNA (Wilson et al., 2019) data of each gene in each bicluster for accompanying cell-specific changes in gene expression.

Table 6 Number of clusters selected based on the maximum value of the silhouette index (optimal)

Number of clusters	*SBB–PAM*				*SBB–K-means*				*SBB-AHC (Ward)*			
	DN		*DR*		*DN*		*DR*		*DN*		*DR*	
	SR	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*
2	0.538	0.830	0.548	0.785	0.522	0.824	0.539	0.762	0.520	0.754	0.528	0.648
3	0.464	0.485	0.455	0.543	0.457	0.376	0.441	0.533	0.458	0.743	0.517	0.612
4	0.433	0.434	0.424	0.421	0.426	0.428	0.406	0.415	0.447	0.443	0.493	0.384
5	0.405	0.176	0.397	–	0.389	0.157	0.396	–	0.445	–	0.473	–
6	0.386	–	0.366	–	0.374	–	0.359	–	0.377	–	0.452	–
7	0.364	–	0.356	–	0.358	–	0.344	–	0.368	–	0.439	–
8	0.339	–	0.331	–	0.324	–	0.328	–	0.360	–	0.367	–
9	0.346	–	0.317	–	0.335	–	0.311	–	0.321	–	0.328	–
10	0.332	–	0.329	–	0.323	–	0.320	–	0.125	–	0.171	–

Figure 6 Heat map of the DR bicluster (see online version for colours)

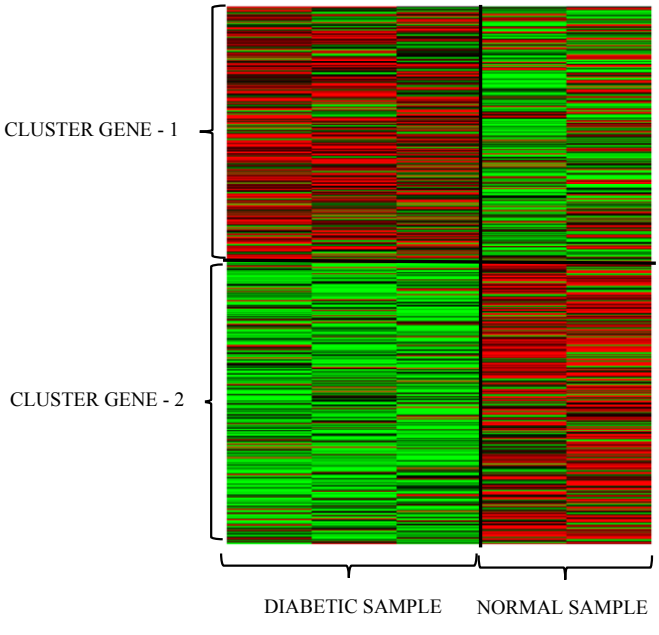


Table 7 Gene ontology enrichment in DN

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p -Value*
Cluster 1	Phosphoprotein	1.385	53.1	3.3×10^{-57}
	Disease mutation	530	20.3	5.2×10^{-39} .
	Kinase	189	7.5	1.6×10^{-25}
	Response to salt stress	19	0.8	5.2×10^{-3}
	Cellular response to glucose stimulus	58	2.7	6.3×10^{-5}
	TGF-beta signalling pathway	43	1.3	3.1×10^{-5}
	Protein Kinase C binding	27	1.2	3.6×10^{-4}
	Regulation of insulin secretion	34	1.9	1.9×10^{-5}
	Blood coagulation	18	2.7	2.1×10^{-5}
	Virus receptor activity	14	2.6	1.6×10^{-1}
Cluster 2	mRNA splicing	91	3.8	3.8×10^{-23}
	Extracellular exosome	484	20	1.5×10^{-14}
	Transcription regulation	382	15.8	5.4×10^{-14}
	Regulation of mRNA stability	33	1.4	2.4×10^{-6}
	GTPase activity	54	2.2	5.9×10^{-5}
	Transcription coactivator activity	69	2.9	2.4×10^{-9}

Table 8 Gene ontology enrichment in DR

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p-Value*
Cluster 1	Extracellular exosome	266	17.9	1.2×10^{-19}
	RNA-binding	73	4.9	2.9×10^{-10}
	Focal adhesion	55	3.7	3.6×10^{-9}
	Glycoprotein	270	18.2	2.3×10^{-6}
	Angiogenesis	27	1.8	1.9×10^{-3}
	Cellular glucose homeostatis	47	2.8	6.8×10^{-4}
	Glycolytic process	11	0.7	3.0×10^{-5}
	Response to hypoxia	46	2.8	3.1×10^{-5}
	Regulation of vascular endothelial growth factor	40	2.7	3.1×10^{-2}
	Stress response	14	1	1.5×10^{-4}
	Insulin like growth factor binding	6	0.4	9.8×10^{-3}

Table 8 Gene ontology enrichment in DR (continued)

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p-Value*
Cluster 2	Immunity	32	2.2	3.3×10^{-2}
	Cell body	26	1.8	3.9×10^{-10}
	Respiratory chain	17	1.2	9.5×10^{-8}
	ATPase activity	29	2.0	2.6×10^{-5}
	Biosynthesis of antibiotics	28	2.0	1.3×10^{-3}
	Regulation of cell growth	11	0.8	2.0×10^{-3}
	ATP metabolic process	9	0.6	2.1×10^{-3}

4 Conclusions

Several possible values of λ, where $\lambda < 1$, were determined to reach the maximum silhouette index value for accurate grouping of DN and DR microarray data. The results of these computations showed higher acquisition of silhouette indexes using PAM than using the K-means partition algorithm. Therefore, SBB–PAM gives better clustering accuracy than K-means.

Overall, SBB algorithms extracted important biological information from microarray gene expression data through biclustering, as demonstrated through regulation in each bicluster. The present heat map analyses show that differentiating genes have significantly different expression levels in different sample conditions. These observations may inform medical practitioners about genes that tend to affect disease. In addition, the detected genes corresponded with respective biological functions, and these were relevant to the conditions of DN and DR, as indicated by significant enrichment in GO analyses. Good clustering results can be used by medical experts to determine prevention and treatment strategies for patients with disorders characterised in terms of groups of genes that are affected by certain conditions, such DN and DR.

The limitation of the SBB algorithm is that it is only capable of producing non-overlapping biclusters. We intend to learn more about the SBB algorithm to find overlapping biclusters.

Acknowledgements

This research supported by research grant Universitas Indonesia 0677/UN.R3.1/HKP.05.00/2019.

References

Brennan, E., McEvoy. C., Sadlier, D., Godson, C. and Martin F. (2013) 'The genetics of diabetic nephropathy', *Genes*, ISSN 2073-4425.

Bustamam, A., Zubedi, F. and Siswantining, T. (2018) 'Implementation χ-sim co-similarity and agglomerative hierarchical to cluster gene expression data of lymphoma by gene and condition', *AIP Conference Proceedings,* Vol (2023) No. 1, pp.020221.

Cahyaningrum, R.D., Bustamam, A. and Siswantining, T. (2017) 'Implementation of spectral clustering with partitioning around medoids (PAM) algorithm on microarray data of carcinoma', *AIP. Conference Proceedings,* Vol. 1825, No. 1, pp.020007.

Cheng, Y. and Church, G.M. (2000) 'Biclustering of expression data', *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, Vol. 8, pp.93–103.

Chin, A.J., Mirzal, A. and Haron, H. (2015) 'Spectral clustering on gene expression profile to identify cancer types or subtypes', *Journal of Technology (Sciences and Engineering)*, Vol. 76, No. 1, pp.289–297.

Divina, F. and Aguilar-Ruiz, J.S. (2006) 'Biclustering of expression data with evolutionary computation', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 5, pp.590–602.

Eldred, G.E. and Katz, M.L. (1988) 'Fluorophores of the human retinal pigment epithelium: separation and spectral characterization', *Experimental Eye Research*, Vol. 47, No. 1, pp.71–86.

Foster, D.W. (1994) *Diabetes Mellitus in Harrison Prinsip-Prinsip Ilmu Penyakit Dalam*, 13th ed., Jakarta, EGC.

Golub, G. and Van Loan, C. (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protoc*., Vol. 4, No. 1, pp.44–57.

Hussain, S.F. and Ramazan, M. (2016) 'Biclustering of human cancer microarray data using co-similarity based co-clustering', *Expert System with Application*, Vol. 55, pp.520–531.

Hussain, S.F., Bisson, G. and Grimal, C. (2010) 'An improved co-similarity measure for document clustering', *Ninth International Conference on Machine Learning and Applications (ICMLA)*, pp.190–197.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Group in Data: An Introduction to Cluster Analysis*, Wiley.

Klebanov, L. and Yakovlev, A. (2007) 'How high is the level of technical noise in microarray data?', *Biology Direct*, Vol. 2, Vol. 2, No. 1, p.9.

Liu, B., Xin, Y., Cheung, R.C.C. and Yan, H. (2014) 'GPU-based biclustering for microarray data analysis in neurocomputing', *Neurocomputing*, Vol. 134, pp.239–246.

Madeira, S.C. and Oliveira, A.L. (2004) 'Biclustering algorithms for biological data analysis: a survey', *IEEE Transactions on Computational Biology and Bioinformatics*, Vol. 1, No. 1, pp.24–25.

Mondal, B. and Choudhury, J.P. (2013) 'A comparative study on k-means and PAM algorithm using physical characters of different varieties of mango in India', *International Journal of Computer Applications*, Vol. 78, No. 5, pp.0975–8887.

Park, H.S. and Jun, C.H. (2009) 'A simple and fast algorithm for K-medoids clustering', *Expert System with Application*, Vol. 36, No. 2, pp.3336–3341.

Patro, S. and Sahu, K.K. (2015) *Normalization: A Preprocessing Stage*, ArXiv Preprint ArXiv: 1503.06462.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig L., Thiele, L. and Zitzler E. (2006) 'A systematic comparison and evaluation of biclustering methods for gene expression data', *Bioinformatics*, Vol. 22, No. 9, pp.1122–1129.

Rousseeuw, P.J. (1987) 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis' *Journal of Computational and Applied Mathematics*, Vol. 20, pp.53–65.

Rui, X. and Wunsch, D.C. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, Vol. 16, No. 3.

Sasirekha, K. and Baby, P. (2013) 'Agglomerative Hierarchical Clustering Algorithm–A Review'*International Journal of Scientific and Research Publications*, Vol. 3, No. 3, ISSN 2250-3153.

Servat, O.S., Hernandez, C. and Simo, R. (2013) 'Genetics in diabetic retinopathy: current concepts and new insights', *Current Genomics*, Vol. 14, No. 5, pp.289–299.

Setyaningrum, N., Bustamam, A. and Siswantining, T. (2019) 'Finding correlated bicluster from gene expression data of alzheimer disease using FABIA biclustering method', *AIP Conference Proceedings,* Vol (2084) No. 1, p.020005.

Wen, L., Zhang, Z., Peng, R., Zhang, L., Liu, H., Peng, H. and Sun, Y. (2019) 'Whole transcriptome analysis of diabetic nephropathy in the db/db mouse model of type 2 diabetes', *J. Cell Biochem*., Vol. 120, pp.17520–17533.

Wibawa, N.A., Bustamam, A. and Siswantining, T. (2019) 'Differential gene co-expression network using bicMix', *AIP Conference Proceedings,* Vol (2084) No. 1, p.020006.

Wilson, P.C., Wu, H., Kirita, Y., Uchimura, K., Ledru, N., Rennke, H.G., Welling, P.A., Waikar, S.S. and Humphreys, B.D. (2019) 'The single-cell transcriptomic landscape of early human diabetic nephropathy', *Proceedings of the National Academy of Sciences*, Vol. 116, No. 39, pp.19619–19625.

Wu, X. and Kumar, V. (2009) *The Top Ten Algorithms in Data Mining*, University of Vermont, Chapman and Hall, USA.

Wutun, T.B., Bustamam, A. and Siswantining, T. (2019) 'Implementation of factor analysis for bicluster acquisition: sparseness projection (FABIAS) on microarray of Alzheimer's gene expression data', *AIP Conference Proceedings,* Vol. 2084, No. 1, p.020004.

Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm

by Sofia Debi Puspa Debi Puspa

Submission date: 10-Jun-2024 01:35PM (UTC+0700)
Submission ID: 2399409288
File name: IJBRA_17_4_Paper_5.pdf (489.66K)
Word count: 7410
Character count: 37521

Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm

Titin Siswantining, Alhadi Bustamam*,
Sofia Debi Puspa and Zuherman Rustam

Faculty of Mathematics and Natural Sciences,
Department of Mathematics,
Universitas Indonesia,
Pondok Cina, Depok, 16424, Indonesia
Email: titin@sci.ui.ac.id
Email: alhadi@sci.ui.ac.id
Email: sofiadebipuspa@sci.ui.ac.id
Email: zuherman@sci.ui.ac.id
*Corresponding author

Fahrezal Zubedi

Faculty of Mathematics and Natural Sciences,
Department of Mathematics,
Universitas Negeri Gorontalo,
Gorontalo, 96128, Indonesia
Email: fahrezal@ung.ac.id

Abstract: Similarity-based biclustering (SBB) algorithm consists of four main phases, transforming data, the construction of row (gene) and column (condition) similarity matrices, the clustering of each similarity matrix and the extraction of the bicluster. In this study, we modified the SBB algorithm at the stage of data transformation using min-max normalisation to identify significant biclusters in diabetic nephropathy and retinopathy microarray data after genes are selected using relative deviations and absolute deviations. Based on the comparison of the silhouette index validation experiments, SBB using partitioning around medoids (PAM) provided better clustering of genes and samples than K-means and agglomerative hierarchical clustering (AHC) (Ward's linkage). Furthermore, the proposed technique identified a meaningful non-overlapping bicluster on a real dataset. Using gene ontology (GO) enrichment analysis and the Bonferroni correction, we have identified biological evidence in each bicluster that is significant in terms of gene functions and biological processes.

Keywords: agglomerative hierarchical clustering; biclustering; diabetic nephropathy; diabetic retinopathy; gene expression; K-means; microarray data; PAM; partitioning around medoids; SBB; similarity-based biclustering.

Reference to this paper should be made as follows: Siswantining, T., Bustamam, A., Puspa, S.D., Rustam, Z. and Zubedi, F. (2021) 'Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm', *Int. J. Bioinformatics Research and Applications*, Vol. 17, No. 4, pp.343–362.

Biographical notes: Titin Siswantining received her Bachelor of Science degree in statistic from Sepuluh November Institute of Technology, Indonesia in 1984. The Master's degree in Applied Mathematics from Ecole Des Hautes Etudes En Sciences Sociales in 1990 and the PhD in Statistics from Bogor Agricultural Institute in 2013. Her research focuses on statistics and bioinformatics.

Alhadi Bustamam received his BSc (honour) degree in Computational Mathematics in 1996, the Master's degree in Computer Science from Universitas Indonesia in 2002, and the PhD in Bioinformatics from the University of Queensland, Australia in 2011. His research focuses on computational mathematics, computational biology, bioinformatics and computer sciences.

Sofia Debi Puspa received her Bachelor of Education degree in Mathematics from Universitas Negeri Jakarta in 2016, Master's degree in Mathematics from the Universitas Indonesia in 2018. Her research interests include bioinformatics.

Zuherman Rustam received his BSc in Mathematics in 1984, Master degree and PhD in Mathematics Science from Universitas Indonesia in 2006. His research focuses on bioinformatics and machine learning.

Fahrezal Zubedi received his Bachelor of Education degree in Mathematics from Universitas Negeri Gorontalo in 2016, Master's degree in Mathematics from the Universitas Indonesia in 2018. His research interests include bioinformatics.

This paper is a revised and expanded version of a paper entitled 'Implementation of co-similarity measure on microarray data of lymphoma by using K-means partition algorithm' presented at *3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (3rd ISCPMS 2017)*, 26–27 July, 2017.

1 Introduction

Global increases in the prevalence of diabetes mellitus (DM) are accompanied by the increased prevalence of diabetic microvascular disorders, such as diabetic nephropathy (DN) and diabetic retinopathy (DR). DN is a kidney abnormality that leads to decreased renal function in subjects with a history of DM. DM, particularly type II diabetes, is the sixth most common cause of death due to kidney failure in western countries. According to a microalbuminuria prevalence study (MAPS) in Asia, nearly 60% of patients with type 2 diabetic hypertension suffer from DN, with macroalbuminuria and microalbuminuria accounting for 18.8% and 39.8% of cases, respectively. DN is characterised by proteinuria (microalbuminuria), which precedes DN by months to years. Therefore, Foster (1994) has suggested that the early detection of microalbuminuria

allows interventions that inhibit further decline in renal function. DR currently affects more than 20 million diabetic people in the US The associated microvascular damage to the retina is characterised by the death of retinal neurons and decreased retinal vascularisation (Eldred and Katz, 2008). Although not lethal, this condition can decrease vision and lead to permanent blindness (Wilardjo, 2001).

Recent technological advances in molecular biology have greatly facilitated analyses of diverse biological phenomena. For example, microarray technologies have been used to profile changes in relative gene expression levels associated with DN and retinopathy. Microarray technology utilises gene sequences determined in the genome project to analyse specific organisms at any given time and under specific conditions. The microarrays return expression matrices for specific biological conditions, with genes in rows and expression levels in columns (Madeira and Oliveira, 2004).

Gene expression analyses can be performed in various ways, including by grouping data using clustering and co-clustering (biclustering) algorithms. Liu et al. (2014) have described clustering as a data classification process that partitions the data according to various criteria to find patterns. However, clustering is limited by the assumption that genes have the same expression levels under all measurable conditions. To address this limitation, bioinformatics analyses are conducted using biclustering or co-clustering algorithms. Prelić et al. (2006) simultaneously applied biclustering algorithms in two dimensions to identify subset pairs of genes and conditions of gene expression. Cheng and Church (2000) were the first to apply the metric block correlation method mean residual score (MRS), which measures the quality of the biclusters found. Subsequently, several studies have proposed algorithms based on MRS, such as that by Divina and Aguilar-Ruiz (2006). Several studies have shown the successful implementation of biclustering algorithms on Alzheimer's microarray data, such as Setyaningrum et a1. (2019), Wibawa et al. (2019) and Wutun et al. (2019).

One biclustering technique that can detect sets of genes with similar expression levels under certain conditions extracts biclusters according to biological functions and is known as similarity-based biclustering (SBB). Hussain and Ramazan (2016) generated biclusters that show gene expression levels in cellular processes. Their analyses define medically relevant groups of genes under certain conditions and can be used to inform prevention and treatment strategies.

Hussain and Ramazan (2016) implemented an SBB algorithm using agglomerative hierarchical clustering (AHC) to find biclusters. They also compared it with other bicluster algorithms such as ISA, SAMBA, Bimax, Cheng and Church algorithm and the other. The results of his research that AHC (Ward's linkage) produces the highest accuracy than other bicluster algorithms. However, AHC has a high complexity of $O(n^3)$, where n is the number of data points. Therefore, it is less suitable for application in large-scale gene expression datasets (Sasirekha and Baby, 2013). This paper explains that the weakness of the SBB algorithm lies in the clustering stage, namely the AHC method, and recommends using partition-based clustering methods. Partition-based clustering consists of several grouping methods, such as k-means clustering and partitioning around medoids (PAM). Popular clustering techniques, including K-means clustering and PAM, have been used to classify gene expression data.

Unlike hierarchical clustering algorithms, partition-based clustering methods begin by classifying the entire dataset into k partitions. After initial partitioning of data into k-clusters, heuristics are used to improve the grouping based on some objective functions.

The K-means clustering algorithm is designed to partition objects in the dataset into subsets so that all the points in particular subsets are closest to a given centre (Macqueen, 1996). The advantages of this approach include ease of implementation, comparatively less time to run the algorithm and adaptability. This algorithm has been commonly used, for example, by Wu and Kumar (2009). PAM is a widely used and powerful clustering technique (Park and Jun, 2009) and has the advantage of efficiency during application to large datasets. Bustamam et al. (2018) conducted clustering research using AHC and produced the best grouping accuracy. Therefore, we compared the PAM, K-means partitioning and AHC (Ward's linkage) methods to implement the SBB algorithm and identify significant biclusters in gene expression data.

We used a min–max normalisation method to transform the data. According to Chin et al. (2015), the main advantage of the min–max normalisation method is that it reflects a balance of comparative values before and after the normalisation process, thus eliminating biases in the data.

The remainder of this paper is organised as follows. The datasets and study method are described in Section 2. The results and the analysis are presented in Section 3. The purpose of this paper is to modify the clustering part of the SBB algorithm with k-means and PAM algorithm and compare their performance with the original one, i.e., SBB using AHC (Ward's Linkage) to obtain the biclusters from microarray gene expression data. Further, we extract important biological information from the biclusters which the Bonferroni correction provided by the database for annotation, visualisation and integrated discovery (DAVID) by Huang et al. (2009).

2 Datasets and methods

Diabetic gene expression data from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) can be accessed at https://www.ncbi.nlm.nih.gov/geo. The DN dataset was used with the GSE1009 ID, which contains gene expression profiles of kidney glomeruli from DN patients and comprises 12,626 genes from six samples. The DR dataset comprises 45,102 genes, and five samples and can be accessed using the code GSE12610 ID.

The following steps in the data analysis comprise the matrix of gene expression data A. We selected genes using absolute and relative deviation methods. We then used the min–max normalisation method to transform matrix A into matrix A_n and the chi similarity measure to generate a similarity matrix of gene, i.e., rows (SR) and a similarity matrix of condition, i.e., column (SC). This similarity matrix was then clustered into k and l clusters using PAM and K-means, and the best cluster was determined using the index silhouette method. The extract bicluster method was then applied to this cluster to determine the regulatory expression levels of each bicluster. The following flowchart (Figure 1) shows the data analysis steps.

2.1 Gene selection

Microarrays produce large-scale data that generally have experimentally generated noise due to, for example, transformation of the image, image segmentation, selection of parameters, different numbers of Cy3 and Cy5 on labelling mRNAs and inequality of initial RNA quantities. Such noise has a strong effect on the accuracy of grouping. In

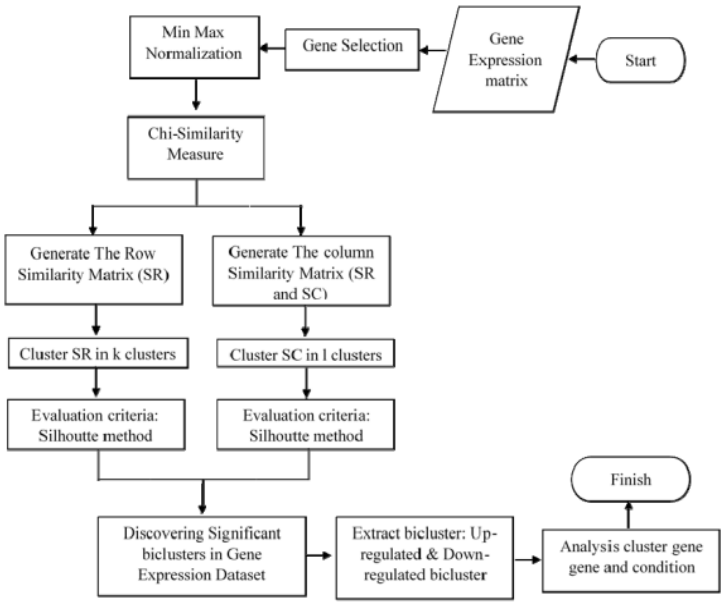
addition, greater numbers of features (genes) increase the time and cost of computing required to process the data. Therefore, genes are selected according to significant differential gene expression levels. We used two popular and simple selection techniques involving relative deviations and absolute deviations. In relative deviation techniques, selected genes have relative deviations that are greater than the threshold (δ), and genes with smaller deviations are considered noise, as described by Klebanov and Yakovlev (2007). The formula for relative deviations is as follows:

$$\delta \le \left| \frac{\max(a^i)}{\min(a^i)} \right|. \tag{1}$$

Golub and Van Loan (1999) have suggested that absolute deviation techniques considers the range of ranges of absolute values for all conditions. Genes that are selected using this technique have absolute deviations that are greater than the threshold (θ). Absolute deviation is calculated as follows:

$$\theta \le \left| \max(a^i) - \min(a^i) \right|. \tag{2}$$

Figure 1 Flowchart of the data analysis



2.2 Min–max normalisation

In this normalisation technique, data are transformed using minimum and maximum values. Min–max normalisation alters matrix entries to the interval [0,1], as indicated by Chin et al. (2015). The min–max normalisation formula reported by Patro and Sahu (2015) is as follows:

$$x'_{(a,b)} = \frac{x_{(a,b)} - \min(x_a)}{\max(x_a) - \min(x_a)} \times (new_{\max} - new_{\min}) + new_{\min}, \quad (3)$$

where $x'_{(a,b)}$ is the element value (a,b) in the data matrix after normalisation, $x_{(a,b)}$ is the element value (a,b) in the data matrix before normalisation, $\min(x_a)$ is the minimum value in the A^{th} row of the matrix before normalisation, $\max(x_a)$ is the maximum value in the A^{th} row of the matrix before normalisation, $new_{\max}$ is the limit of the new maximum value and $new_{\min}$ is the limit of the new minimum value.

2.3 X-Sim similarity measure

The χ-Sim similarity measure is a co-similarity-based approach that simultaneously uses two similarity matrices of similarity matrix between row (SR) and similarity matrix between column (SC). χ-Sim is calculated iteratively by updating these SR and SC matrices, and the resulting updated SR and SC matrices are interrelated so that each matrix construct has a similarity given by another co-similarity measure, as suggested by Hussain et al. (2010).

Let A be the data matrix representing a gene expression microarray with r rows (genes) and c columns (conditions). The intensity of gene activities between the Ith row and the Jth column is denoted by $a_{i:} = \left[a_{i1}, \ldots, a_{ic}\right]$, which is the row vector representing the gene i, and $a_{j:} = \left[a_{j1}, \ldots, a_{jc}\right]$, which is the column vector corresponding to condition j. In these computations, SR and SC represent square and symmetrical row and column similarity matrices of size r x r and c x c, respectively, with $\forall i, j = 1..r, sr_{ij} \in [0,1]$ and $\forall i, j = 1..c, sc_{ij} \in [0,1]$. In this equation, $F_s(...)$ is a generic function that takes the elements a_{il} and a_{jn} of A and returns a measure of similarity $F_s(a_{il}, a_{jn})$ between them. The similarity (or distance) measure between two genes $a_{i:}$ and $a_{j:}$ is defined as a function and is denoted as $Sim(a_{i:}, a_{j:})$ to calculate the SR matrix (Hussain and Ramazan, 2016), as follows:

$$Sim\left(a_{i:}, a_{j:}\right) = F_s\left(a_{i1}, a_{ji}\right) + F_s\left(a_{i2}, a_{j2}\right) + \cdots + F_s\left(a_{ic}, a_{jc}\right). \quad (4)$$

Given a matrix SC with entries that provide a measure of similarity between columns (condition) of microarray data, we introduce a pseudo-norm λ that is analogous to the norm L_λ (Minkowski distance). Equation (4) can then be rewritten as follows without changing its meaning if $sc_{ll} = 1$ and $\lambda = 1$ (Hussain and Ramazan, 2016):

$$Sim\left(a_{i:}, a_{j:}\right) = F_s\left(a_{i1}, a_{ji}\right).sc_{11} + F_s\left(a_{j2}, a_{j2}\right).sc_{22} + \ldots + F_s\left(a_{ic}, a_{jc}\right).sc_{cc} \quad (5)$$

$$Sim\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\sum_{l=1}\left(F_s\left(a_{il}, a_{jl}\right)^{\lambda} \times sc_{ll}\right)}. \quad (6)$$

Equation (5) generalises (6) so that all possible pairs of features (genes) occurring under conditions $a_{i:}$ and $a_{:j}$ can be accounted for. The overall similarity between genes $a_{i:}$ and

$a_{:j}$ is defined in the following equation (7), in which the terms for $l = n$ are those in equation (6):

$$Sim^{\lambda}\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\sum_{l=1}^{c}\sum_{n=1}^{c}\left(F_s\left(a_{ll}, a_{jn}\right)^{\lambda} \times sc_{\ln}\right)}. \tag{7}$$

Assuming that the function $Sim(a_{il}, a_{jn})$ is defined as a product of the elements a_{il} and a_{jn} ($Sim(a_{il}, a_{jn}) = a_{il} x a_{jn}$, as with the cosine similarity measure), we can rewrite equation (7) as follows:

$$Sim^{\lambda}\left(a_{i:}, a_{j:}\right) = \sqrt[\lambda]{\left(a_{i:}\right)^{\lambda} \times sc \times \left(a_{j:}^{T}\right)^{\lambda}}, \tag{8}$$

where $\left(a_{i:}\right)^{\lambda} = \left[\left(a_{ij}\right)^{\lambda} ... \left(a_{ic}\right)^{\lambda}\right]$ and $a_{j:}^{T}$ denotes the transpose of the vector $a_{j:}$, as described by Hussain and Ramazan (2016).

2.4 The generic χ-Sim co-similarity measure

In general, expansion of all row and column pairs is achieved by matrix multiplication from SR and SC matrices. According to Hussain et al. (2010), the steps of this method are as follows:

a Similarity matrices of the gene or row (SR) and similarity matrix of condition or column (SC) are initiated with the identity matrix. We can write these matrices as $SR^{(o)}$ and $SC^{(o)}$, where the superscript is the iteration number.

b At each iteration *t*, we calculate the new similarity matrix between genes $SR^{(t)}$ using the similarity matrix between conditions $SC^{(t-1)}$, and we do the same thing for the columns' similarity matrix $SC^{(t)}$, as follows:

$$SR^{(t)} = A^{o\lambda} \times SC^{(t-1)} \times (A^{o\lambda})^{T}, \tag{9}$$

$$SC^{(t)} = (A^{o\lambda})^{T} \times SR^{(t-1)} \times A^{o\lambda}, \tag{10}$$

where $A^{o\lambda} = ((a_{ij})^{\lambda})_{i,j}$ is the element-wise exponent of *A*. Furthermore, each SR and SC matrix is normalised using the following pseudo-normalisation:

$$\forall i, j \in 1..r, sr_{ij}^{(t)} = \frac{\sqrt[\lambda]{sr_{ij}^{(t)}}}{2\sqrt[\lambda]{sr_{ii}^{(t)} x sr_{jj}^{(t)}}}, \tag{11}$$

$$\forall i, j \in 1...c, sc_{ij}^{(t)} = \frac{\sqrt[\lambda]{sc_{ij}^{(t)}}}{2\sqrt[\lambda]{sc_{ii}^{(t)} x sc_{jj}^{(t)}}}. \tag{12}$$

c $SR^{(t)}$ and $SC^{(t)}$ are updated iteratively ($t = 4$ is sufficient).

2.5 Partitioning around medoids

PAM uses the k-medoid method to identify clusters and chooses a medoid by calculating minimum total distances from others (Mondal and Choudhury, 2013). According to Cahyaningrum et al. (2017), a medoid is a representative of each group, and the number of medoids is the same as the number of clusters. The data distance formula with *n*-dimensional data is calculated using Euclidean distance, a distance matrix that adopts the Pythagoras principle. This is because the calculation pattern uses the rules of rank and the square root. Euclidean computations give relatively small distances because they use the square root rule, as follows:

$$d(x_a, x_b) = \sqrt{\sum_{k=1}^{n} (x_{ak} - x_{bk})^2} \ . \tag{13}$$

The purpose of PAM is to determine a representative object–a medoid–for each cluster that has a minimum total distance to other objects. PAM is performed in two stages, building and swapping. The building stage is a sequential object selection process that continues until the object is found. The swap stage is performed to improve group quality by exchanging selected objects for unselected objects (Mondal and Choudhury, 2013).

Kaufman and Rousseeuw (1990) have described the steps of the building stage as follows:

1 Suppose object *i* has not been selected.

2 Identify the unselected object *j* and calculate the difference between D_j (distance of object *j* from the previously selected object) and *j* (distance between objects *j* and *i*).

3 If the distance in step 2 is positive, then object *j* will contribute to the selection of object *i*. Then count

$$C_{ji} = \max(D_j - d(j,i), 0) \tag{14}$$

4 Calculate the total gain from the selection of object *i*.

$$\sum_j C_{ji} \tag{15}$$

5 Select an object that has not been selected *i*.

$$\underset{i}{\text{maximising}} \sum_i C_{ji} \tag{16}$$

6 Repeat steps 1 to 5 until *k* objects are found.

These steps can be used to calculate the effects of swapping between *i* and *h* objects on clustering results, as described by Kaufman and Rousseeuw (1990) in the following instructions:

1 Assume that object j was not selected and count the contribution of C_{jih} to the swap as follows:

   a If the distances between objects *j*, *i* and *h* are greater than between *j* and one of the representative objects, then $C_{jih} sil(x_i)$ 0.

b If the distance between objects *j* and *i* is further than the distance between *j* and any other selected representative objects $(d(j,i) = D_j$, then the following two conditions must be fulfilled:

1) The distance of object *j* is closer to object *h* than to the second nearest representative object, as follows:

$$d(j,h) < E_j, \tag{17}$$

where E_j .is the distance between object *j* and the second-nearest representative object. Under these conditions, the contribution of object *j* to the swap between objects *i* and *h* is calculated as follows:

$$C_{jih} = d(j,h) - (d(j,i)). \tag{18}$$

2) The distance of object *j* to *h* is at least equal to that between *j* and the second closest representative object, as follows:

$$d(j,h) < E_j. \tag{19}$$

Under these conditions, the contribution of the object to the swap is calculated as follows:

$$C_{jih} = E_j - D_j. \tag{20}$$

For condition 1), if object *j* is closer to object *i* than to object *h*, which contributes positively, then the swap with object *j* is not favourable.

c If the object distance *j* to *i* is greater than the object distance *j* to at least one of the other representative objects but closer to *h* than the distance of object *j* to any representative object, then the contribution of *j* to the swap is calculated as follows:

$$C_{jih} = d(j,h) - D_j. \tag{21}$$

2 Calculate the total result of a swap by adding a contribution.

3 To determine whether a swap has occurred, select the (*i*, *h*) pair in the following equation:

$$\underset{i,h}{\text{minimising}} T_{ih}. \tag{22}$$

4 If the minimum value of $T_{ih}sil(x_i)$ s negative, then the swap has occurred, and the swap stage returns to step 1. If the minimum value of $T_{ih}sil(x_i)$ s positive, then the swap has not occurred, and the stages are stopped.

2.6 K-means algorithm

The K-means algorithm is a partitioning method to partition existing data into one or more clusters and is performed using the following steps:

1 Determine the numbers of clusters (k) and centroids.

2 Determine the distance between each object and the centroids using the Euclidean distance formula, as follows:

$$d_{ik} = \sqrt{\sum_{i,k=1}^{n} (x_i - c_k)^2}\,, \tag{23}$$

where d_{ik} s the distance of object $sil(x_i)$ and centroid., n is dimensional data, x_i is a coordinate of objects i an $sil(x_i)$ c_k is the centroid coordinate.

3 Cluster the objects based on the minimum distance.

4 Determine new centroids using the following formula:

$$C_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}\,, \tag{24}$$

where $x_{ij} \in sil(x_i)$ s the.-cluster an $sil(x_i)$ p is the number of the member of the cluster. Iterate the above steps until no object moves from its assigned group.

2.7 Index silhouette

The silhouette method is used to determine the number of clusters and to measure the accuracy of grouping results (Rousseeuw, 1987). The silhouette method measures how similar an object is to other objects in its group relative to similarities between other groups. Silhouette values are calculated as follows:

$$sil(x_i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}\,, \tag{25}$$

where $a(i)$ is the average dissimilarity (Euclidean distance) value of the ith data point to all members of the cluster containing data., and $b(i)$ is the smallest average dissimilarity of ith data to all members of the group not containing the data i. In this study, the silhouette index $sil(x_i)$ as used to determine the number of groups by choosing the highest value of $sil(x_i)$ as reported by Kaufman and Rousseeuw (1990).

2.8 Extract bicluster

The translation of mRNAs produces proteins and leads to the production of various other components. Therefore, genes are upregulated or downregulated depending on the cellular requirements of these components.

Clustering in SR and SC matrices using hard clustering algorithms produces non-overlapping bicluster or checkerboard structures. Elements of the checkerboard are average values of the set of genes in cluster under the conditions in cluster $\hat{y}_j$, as shown in Table 1. The most significant biclusters are the non-overlapping bicluster. The number of most significant biclusters is the number of obtained biclusters.

The bicluster extract algorithm of SBB extracts significant biclusters containing an upregulated gene cluster under one cluster of conditions only. Therefore, biclusters show greater variation under various conditions occurring in a cluster of genes. Mathematically, we can express this as a, in which the cluster gene is highly regulated under the cluster condition. For every cluster condition, m as low regulation, and the algorithm is written as $(\hat{x}_i, \hat{y}_{m \neq j} \ \forall m \in 1,\ldots,l)$.

Table 1 Identification of significant biclusters from gene expression data

		Condition clusters			
		$\hat{y}_1$.	$\hat{y}_2$	$\cdots$.	$max - min$
Gene clusters	$\hat{x}_1$	$\overline{\hat{x}_1 \hat{y}_1}$	$\overline{\hat{x}_1 \hat{y}_2}$	$\cdots$.	$max\left(\overline{\hat{x}_i \hat{y}_i}\right) - min\left(\overline{\hat{x}_i \hat{y}_i}\right), \forall i \in 1..l$
		$\overline{\hat{x}_1 \hat{y}_1}$			
	$\hat{x}_2$	$\overline{\hat{x}_2 \hat{y}_1}$	$\overline{\hat{x}_2 \hat{y}_2}$	$\cdots$.	$max\left(\overline{\hat{x}_2 \hat{y}_i}\right) - min\left(\overline{\hat{x}_2 \hat{y}_i}\right), \forall i \in 1..l$
	$\cdots$				
	$\hat{x}_k$	$\overline{\hat{x}_k \hat{y}_1}$	$\overline{\hat{x}_2 \hat{y}_2}$	$\cdots$.	

3 Results and discussion

Gene expression matrices for DN and DR after selection were 6320×6 and 3689×5, respectively. The analyses were performed using R version 3.3.1 (an open-source programming tool) on a computer with an Intel® Core™ i5 CPU@ 2.40GHz processor and 4 GB of memory (RAM). Gene selection techniques are required to filter out informative genes by choosing those with values that are greater than each threshold and discarding those with lower relative and absolute deviations. The present relative and absolute deviation values of diabetic data are shown in Table 2.

The range of gene expression data values varied greatly for each row. This variation follows differences in sample conditions associated with patients, necessitating normalisation to produce datasets with uniform scales. Microarray data for DN and DR were subjected to min–max normalisation and are shown as a heat map in Figure 2.

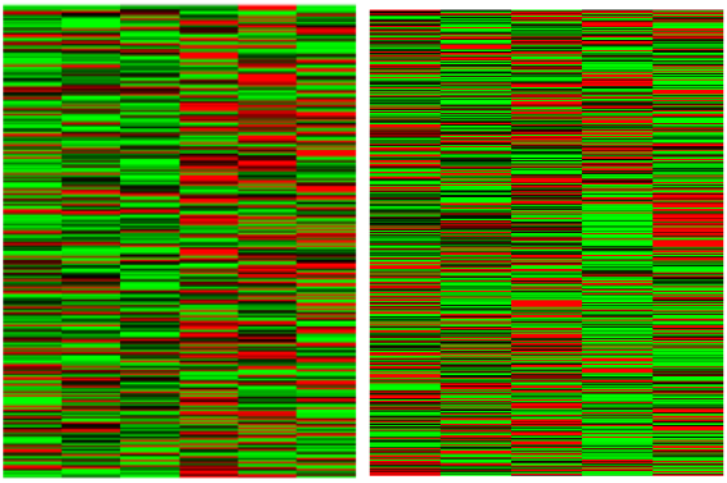
3.1 Comparison of clustering techniques

We ran the SBB algorithm for hard clustering and tested the K-means and PAM clustering techniques to obtain cluster genes and cluster conditions for SR and SC. After normalising the data, we constructed two matrices of the similarity between SR and SC using the χ-Sim co-similarity measure. Determinations of λ in pseudo-normalisations can be used to control the data distributions. The accuracy of grouping was affected by the silhouette index method. This method is needed to determine which clusters are optimal for grouping objects. Table 3 shows silhouette values for similarity matrices of SR and SC using the PAM and K-means algorithms.

Table 2 Description of gene expression datasets

Dataset	*DN*	*DR*
Number of original genes	12,626	45,102
Number of samples	6	5
Number of sample classes	2	2
Sample class names	Diabetic (3)	Diabetic (3)
	Normal (3)	Normal (2)
Relative deviation threshold	15	5
Absolute deviation threshold	4000	2000
No. of selected genes	6320	3689

Figure 2 Heat map normalisation of genes in: (a) DN and (b) DR (see online version for colours)



(a) (b)

Table 3 Silhouette values using PAM

		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.251	0.491	0.670	0.759	0.805	*0.830*	0.793	0.728	0.636
	SR	0.230	0.294	0.381	0.417	0.498	*0.538*	0.438	0.362	0.294
DR	SC	0.463	0.512	0.658	0.764	*0.785*	0.623	0.587	0.473	0.388
	SR	0.295	0.315	0.476	0.506	*0.548*	0.492	0.361	0.308	0.245

Based on Tables 3–5, the maximum threshold value of λ from both clustering algorithms for DN data in the similarity matrices of SR and SC was 0.6, whereas the maximum threshold of λ for DR was 0.5. Therefore, $\lambda = 0.6$ was used in χ-Sim co-similarity measures for DN data, and $\lambda = 0.5$ was used for DR data. For each value of $\lambda < 1$, where

$\lambda = 0.1, 0.2, ..., 0.9$, the number of clusters of the condition with the highest silhouette index value for diabetic data (DN and DR) was 2. Similarly, the number of gene clusters with the highest silhouette value was 2. Therefore, the number of optimal clusters of cluster genes and cluster conditions was 2 for the diabetic data. The silhouette index for the number of clusters for $k = 2, 3, 4, ..., 10$ is shown in Table 5; the number of clusters was selected based on the maximum value of the silhouette index (optimal).

Table 4 Silhouette values by using K-means

		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.232	0.472	0.592	0.684	0.743	*0.824*	0.781	0.674	0.601
	SR	0.226	0.285	0.376	0.372	0.785	*0.522*	0.425	0.351	0.288
DR	SC	0.418	0.485	0.646	0.748	*0.762*	0.601	0.577	0.452	0.364
	SR	0.288	0.291	0.456	0.482	*0.539*	0.487	0.329	0.256	0.237

Table 5 Silhouette values by using AHC (Ward linkage)

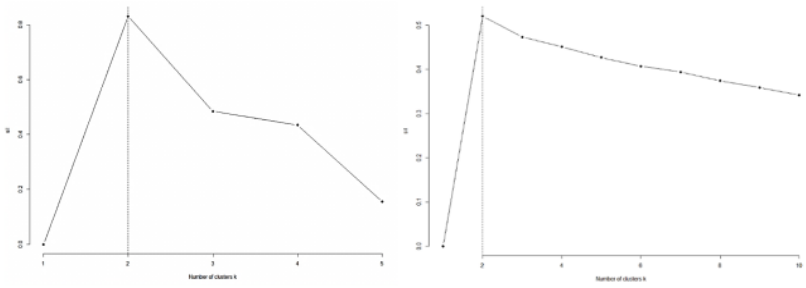
		Threshold λ								
		0.1	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
DN	SC	0.228	0.463	0.558	0.641	0.726	0.754	0.731	0.653	0.587
	SR	0.218	0.263	0.358	0.362	0.736	0.520	0.418	0.337	0.252
DR	SC	0.385	0.468	0.623	0.708	0.648	0.585	0.551	0.447	0.324
	SR	0.277	0.284	0.447	0.453	0.528	0.475	0.318	0.236	0.208

Using PAM for clustering in SBB gave higher silhouette values than K-means and AHC in the range of $\lambda < 1$ for SR and SC matrices, indicating the greater clustering accuracy of the PAM method. Although the K-means algorithm has some advantages, it has a number of drawbacks. For example, K-means is sensitive to outliers and lacks a definite method for identifying optimal partition counts in the initial cluster determination, and the iterative procedure does not ensure movement toward a point of convergence, as described by Rui and Wunsch (2005). Because PAM gives better accuracy in SBB, we used this clustering algorithm for further analyses of genes and conditions. Plots of silhouette values for cluster conditions and cluster genes using PAM are shown in Figures 3 and 4. The horizontal axis represents numbers of clusters, and the vertical axis represents corresponding index silhouette values.

3.2 Analysis of gene clusters

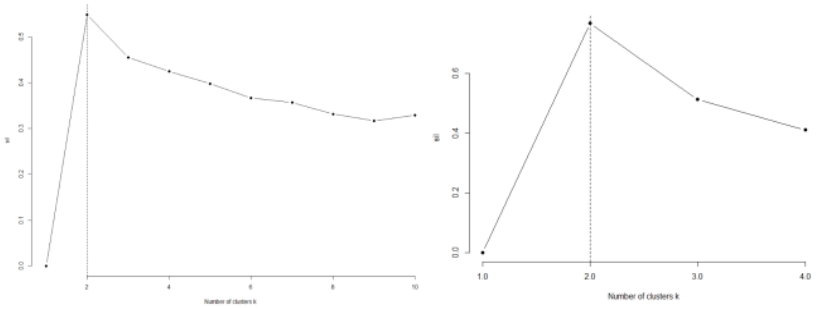
In this study, we generated a heat map to show high gene activities in the microarray dataset under certain conditions, where gene rows and column conditions were grouped simultaneously. Biclusters of similar gene expression values under certain conditions are evident in the heat map in Figure 5, which shows that the bicluster structure generated using the SBB algorithm is non-overlapping.

Figure 3 Silhouette index plots in: (a) SC and (b) SR matrices of DN



(a) (b)

Figure 4 Silhouette index plots in: (a) the SR matrix and (b) the SC matrix of DR



(a) (b)

Furthermore, the present biclusters were analysed according to expression levels using the extracted bicluster, which is the final stage of the SBB algorithm and indicates regulation that occurs in the bicluster–whether upregulation or downregulation–in quantitative terms. To support the results of our qualitative bicluster heat map (Figure 5), we conducted gene ontology (GO) enrichment analyses of cluster generation using the DAVID database at https://david.ncifcrf.gov/site.

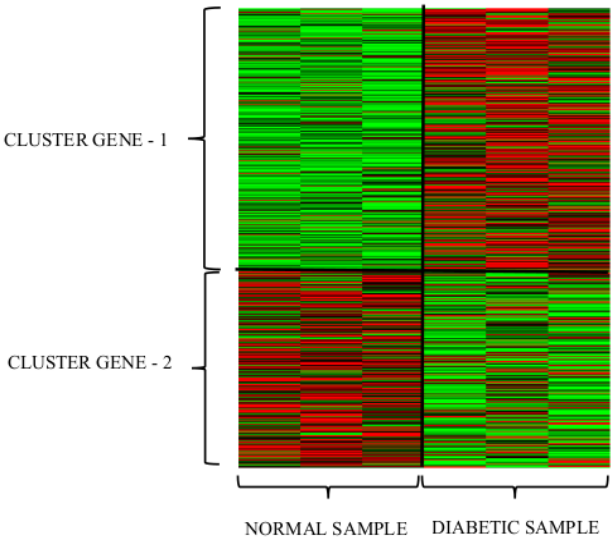
Based on clustering using the PAM algorithm, gene expression data for DN in row and column similarity matrices are grouped into two clusters. The clusters comprised 3615 genes (cluster-1) and 2705 genes (cluster-2) in SR and SC matrices, respectively, under three conditions.

Each cluster of genes corresponding to each sample becomes a bicluster. For example, in Figure 5, the cluster gene 1, which are also from the normal sample, is a submatrix from the microarray data is a bicluster. This is also applied for the cluster gene 1 and diabetic sample, cluster gene 2, and each normal sample and diabetics sample. Therefore, there are four biclusters in Figure 5 as the output.

In GO enrichments of the biological processes for each of the identified gene clusters (Table 6), cluster gen-1 under diabetic conditions had a high expression rate compared to under normal conditions. Genes in cluster-1 were identified as contributing to increases in blood glucose (hyperglycaemia) and glomerular hypertension, leading to kidney inflammation and major conditions of DN (such as high glucose and oxidative stress),

high expression of advanced glycation end-products (AGEs), angiotensin II, TGF-β, CTGF, protein kinase C, the receptor for advanced glycation end-products (RAGE) and NF-κB (Brennan et al., 2013). In contrast, under normal conditions, cluster gen-2 had higher gene expression levels than under diabetic conditions. This indicates that genes grouped using the SBB algorithm are enriched in terms of biological processes related to cellular energy metabolism, as may be expected in the diabetic dataset. Given the differences in expression levels of these genes, the diabetic sample network with expression levels that mathematically differ from those in normal tissue samples may indicate a useful focus for medical practitioners.

Figure 5 Heat map of the DN bicluster (see online version for colours)



In the DR dataset, differences in gene expression levels were evident between diabetic and normal conditions. In particular, genes in cluster-1 are upregulated in diabetic samples but downregulated in normal samples. The genes in cluster-2 are upregulated in normal samples but downregulated in diabetic samples. These gene clusters contained highly discriminative genes, many of which have been previously identified. Differences in gene expression levels between diabetic and normal conditions are shown in Figure 6. Genes of cluster-1, which were overexpressed under the conditions of DR, are known to affect blood glucose levels. Tables 7 and 8 show the biological functions of each gene cluster in diabetic nephropathy and retinopathy microarray respectively.

According to Servat et al. (2013), genetic variants in DR include those of the polyol pathway (hyperglycemia), AGE, hypoxia through vascular endothelial growth factor, RAGE, endothelial nitric oxide synthase, angiotensin I-converting enzyme and others. In contrast, genes in cluster-2 had higher expression levels in normal samples than in diabetic samples and included genes that control protein synthesis and immune activity.

Furthermore, these differences suggest the presence of novel sub-types of DR. The two bluslustering methods that we have developed, namely SBB PAM and SBB K-Means, work for gene expression data in the form of microarrays. The results of the biclusters that we obtained can be analysed further by getting the RNA sequencing (Wen et al., 2019) or Single-cell RNA (Wilson et al., 2019) data of each gene in each bicluster for accompanying cell-specific changes in gene expression.

Table 6 Number of clusters selected based on the maximum value of the silhouette index (optimal)

Number of clusters	*SBB–PAM*				*SBB–K-means*				*SBB-AHC (Ward)*			
	DN		*DR*		*DN*		*DR*		*DN*		*DR*	
	SR	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*	*SR*	*SC*
2	0.538	0.830	0.548	0.785	0.522	0.824	0.539	0.762	0.520	0.754	0.528	0.648
3	0.464	0.485	0.455	0.543	0.457	0.376	0.441	0.533	0.458	0.743	0.517	0.612
4	0.433	0.434	0.424	0.421	0.426	0.428	0.406	0.415	0.447	0.443	0.493	0.384
5	0.405	0.176	0.397	–	0.389	0.157	0.396	–	0.445	–	0.473	–
6	0.386	–	0.366	–	0.374	–	0.359	–	0.377	–	0.452	–
7	0.364	–	0.356	–	0.358	–	0.344	–	0.368	–	0.439	–
8	0.339	–	0.331	–	0.324	–	0.328	–	0.360	–	0.367	–
9	0.346	–	0.317	–	0.335	–	0.311	–	0.321	–	0.328	–
10	0.332	–	0.329	–	0.323	–	0.320	–	0.125	–	0.171	–

Figure 6 Heat map of the DR bicluster (see online version for colours)

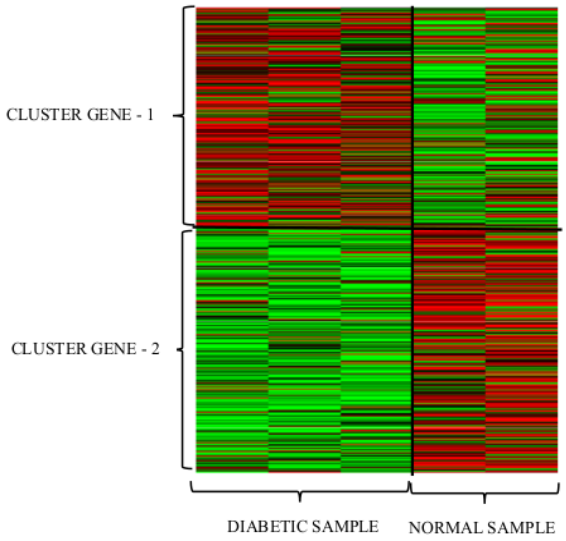


Table 7 Gene ontology enrichment in DN

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p -Value*
Cluster 1	Phosphoprotein	1.385	53.1	3.3×10^{-57}
	Disease mutation	530	20.3	5.2×10^{-39} .
	Kinase	189	7.5	1.6×10^{-25}
	Response to salt stress	19	0.8	5.2×10^{-3}
	Cellular response to glucose stimulus	58	2.7	6.3×10^{-5}
	TGF-beta signalling pathway	43	1.3	3.1×10^{-5}
	Protein Kinase C binding	27	1.2	3.6×10^{-4}
	Regulation of insulin secretion	34	1.9	1.9×10^{-5}
	Blood coagulation	18	2.7	2.1×10^{-5}
	Virus receptor activity	14	2.6	1.6×10^{-1}
Cluster 2	mRNA splicing	91	3.8	3.8×10^{-23}
	Extracellular exosome	484	20	1.5×10^{-14}
	Transcription regulation	382	15.8	5.4×10^{-14}
	Regulation of mRNA stability	33	1.4	2.4×10^{-6}
	GTPase activity	54	2.2	5.9×10^{-5}
	Transcription coactivator activity	69	2.9	2.4×10^{-9}

Table 8 Gene ontology enrichment in DR

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p-Value*
Cluster 1	Extracellular exosome	266	17.9	1.2×10^{-19}
	RNA-binding	73	4.9	2.9×10^{-10}
	Focal adhesion	55	3.7	3.6×10^{-9}
	Glycoprotein	270	18.2	2.3×10^{-6}
	Angiogenesis	27	1.8	1.9×10^{-3}
	Cellular glucose homeostatis	47	2.8	6.8×10^{-4}
	Glycolytic process	11	0.7	3.0×10^{-5}
	Response to hypoxia	46	2.8	3.1×10^{-5}
	Regulation of vascular endothelial growth factor	40	2.7	3.1×10^{-2}
	Stress response	14	1	1.5×10^{-4}
	Insulin like growth factor binding	6	0.4	9.8×10^{-3}

Table 8 Gene ontology enrichment in DR (continued)

Cluster	*Biological process*	*Number of gene*	*Percentage (%)*	*p-Value*
Cluster 2	Immunity	32	2.2	3.3×10^{-2}
	Cell body	26	1.8	3.9×10^{-10}
	Respiratory chain	17	1.2	9.5×10^{-8}
	ATPase activity	29	2.0	2.6×10^{-5}
	Biosynthesis of antibiotics	28	2.0	1.3×10^{-3}
	Regulation of cell growth	11	0.8	2.0×10^{-3}
	ATP metabolic process	9	0.6	2.1×10^{-3}

4 Conclusions

Several possible values of λ, where $\lambda < 1$, were determined to reach the maximum silhouette index value for accurate grouping of DN and DR microarray data. The results of these computations showed higher acquisition of silhouette indexes using PAM than using the K-means partition algorithm. Therefore, SBB–PAM gives better clustering accuracy than K-means.

Overall, SBB algorithms extracted important biological information from microarray gene expression data through biclustering, as demonstrated through regulation in each bicluster. The present heat map analyses show that differentiating genes have significantly different expression levels in different sample conditions. These observations may inform medical practitioners about genes that tend to affect disease. In addition, the detected genes corresponded with respective biological functions, and these were relevant to the conditions of DN and DR, as indicated by significant enrichment in GO analyses. Good clustering results can be used by medical experts to determine prevention and treatment strategies for patients with disorders characterised in terms of groups of genes that are affected by certain conditions, such DN and DR.

The limitation of the SBB algorithm is that it is only capable of producing non-overlapping biclusters. We intend to learn more about the SBB algorithm to find overlapping biclusters.

Acknowledgements

This research supported by research grant Universitas Indonesia 0677/UN.R3.1/HKP.05.00/2019.

References

Brennan, E., McEvoy. C., Sadlier, D., Godson, C. and Martin F. (2013) 'The genetics of diabetic nephropathy', *Genes*, ISSN 2073-4425.

Bustamam, A., Zubedi, F. and Siswantining, T. (2018) 'Implementation χ-sim co-similarity and agglomerative hierarchical to cluster gene expression data of lymphoma by gene and condition', *AIP Conference Proceedings,* Vol (2023) No. 1, pp.020221.

Cahyaningrum, R.D., Bustamam, A. and Siswantining, T. (2017) 'Implementation of spectral clustering with partitioning around medoids (PAM) algorithm on microarray data of carcinoma', *AIP. Conference Proceedings,* Vol. 1825, No. 1, pp.020007.

Cheng, Y. and Church, G.M. (2000) 'Biclustering of expression data', *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, Vol. 8, pp.93–103.

Chin, A.J., Mirzal, A. and Haron, H. (2015) 'Spectral clustering on gene expression profile to identify cancer types or subtypes', *Journal of Technology (Sciences and Engineering)*, Vol. 76, No. 1, pp.289–297.

Divina, F. and Aguilar-Ruiz, J.S. (2006) 'Biclustering of expression data with evolutionary computation', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 5, pp.590–602.

Eldred, G.E. and Katz, M.L. (1988) 'Fluorophores of the human retinal pigment epithelium: separation and spectral characterization', *Experimental Eye Research*, Vol. 47, No. 1, pp.71–86.

Foster, D.W. (1994) *Diabetes Mellitus in Harrison Prinsip-Prinsip Ilmu Penyakit Dalam*, 13th ed., Jakarta, EGC.

Golub, G. and Van Loan, C. (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protoc*., Vol. 4, No. 1, pp.44–57.

Hussain, S.F. and Ramazan, M. (2016) 'Biclustering of human cancer microarray data using co-similarity based co-clustering', *Expert System with Application*, Vol. 55, pp.520–531.

Hussain, S.F., Bisson, G. and Grimal, C. (2010) 'An improved co-similarity measure for document clustering', *Ninth International Conference on Machine Learning and Applications (ICMLA)*, pp.190–197.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Group in Data: An Introduction to Cluster Analysis*, Wiley.

Klebanov, L. and Yakovlev, A. (2007) 'How high is the level of technical noise in microarray data?', *Biology Direct*, Vol. 2, Vol. 2, No. 1, p.9.

Liu, B., Xin, Y., Cheung, R.C.C. and Yan, H. (2014) 'GPU-based biclustering for microarray data analysis in neurocomputing', *Neurocomputing*, Vol. 134, pp.239–246.

Madeira, S.C. and Oliveira, A.L. (2004) 'Biclustering algorithms for biological data analysis: a survey', *IEEE Transactions on Computational Biology and Bioinformatics*, Vol. 1, No. 1, pp.24–25.

Mondal, B. and Choudhury, J.P. (2013) 'A comparative study on k-means and PAM algorithm using physical characters of different varieties of mango in India', *International Journal of Computer Applications*, Vol. 78, No. 5, pp.0975–8887.

Park, H.S. and Jun, C.H. (2009) 'A simple and fast algorithm for K-medoids clustering', *Expert System with Application*, Vol. 36, No. 2, pp.3336–3341.

Patro, S. and Sahu, K.K. (2015) *Normalization: A Preprocessing Stage*, ArXiv Preprint ArXiv: 1503.06462.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig L., Thiele, L. and Zitzler E. (2006) 'A systematic comparison and evaluation of biclustering methods for gene expression data', *Bioinformatics*, Vol. 22, No. 9, pp.1122–1129.

Rousseeuw, P.J. (1987) 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis' *Journal of Computational and Applied Mathematics*, Vol. 20, pp.53–65.

Rui, X. and Wunsch, D.C. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, Vol. 16, No. 3.

Sasirekha, K. and Baby, P. (2013) 'Agglomerative Hierarchical Clustering Algorithm–A Review'*International Journal of Scientific and Research Publications*, Vol. 3, No. 3, ISSN 2250-3153.

Servat, O.S., Hernandez, C. and Simo, R. (2013) 'Genetics in diabetic retinopathy: current concepts and new insights', *Current Genomics*, Vol. 14, No. 5, pp.289–299.

Setyaningrum, N., Bustamam, A. and Siswantining, T. (2019) 'Finding correlated bicluster from gene expression data of alzheimer disease using FABIA biclustering method', *AIP Conference Proceedings,* Vol (2084) No. 1, p.020005.

Wen, L., Zhang, Z., Peng, R., Zhang, L., Liu, H., Peng, H. and Sun, Y. (2019) 'Whole transcriptome analysis of diabetic nephropathy in the db/db mouse model of type 2 diabetes', *J. Cell Biochem.*, Vol. 120, pp.17520–17533.

Wibawa, N.A., Bustamam, A. and Siswantining, T. (2019) 'Differential gene co-expression network using bicMix', *AIP Conference Proceedings,* Vol (2084) No. 1, p.020006.

Wilson, P.C., Wu, H., Kirita, Y., Uchimura, K., Ledru, N., Rennke, H.G., Welling, P.A., Waikar, S.S. and Humphreys, B.D. (2019) 'The single-cell transcriptomic landscape of early human diabetic nephropathy', *Proceedings of the National Academy of Sciences*, Vol. 116, No. 39, pp.19619–19625.

Wu, X. and Kumar, V. (2009) *The Top Ten Algorithms in Data Mining*, University of Vermont, Chapman and Hall, USA.

Wutun, T.B., Bustamam, A. and Siswantining, T. (2019) 'Implementation of factor analysis for bicluster acquisition: sparseness projection (FABIAS) on microarray of Alzheimer's gene expression data', *AIP Conference Proceedings,* Vol. 2084, No. 1, p.020004.

Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm

ORIGINALITY REPORT

10%	3%	9%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hussain, Syed Fawad, and Gilles Bisson. "Text Categorization Using Word Similarities Based on Higher Order Co-occurrences", Proceedings of the 2010 SIAM International Conference on Data Mining, 2010. Publication	1%
2	Anderlucci, Laura <1984>(Montanari, Angela). "Comparing Different Approaches for Clustering Categorical Data", Alma Mater Studiorum - Università di Bologna, 2012. Publication	1%
3	Aigul Kaskina. "Chapter 3 Citizen Privacy Profile Framework", Springer Science and Business Media LLC, 2022 Publication	1%
4	Submitted to Higher Education Commission Pakistan Student Paper	1%
5	Leonard Kaufman, Peter J. Rousseeuw. "Finding Groups in Data", Wiley, 1990 Publication	<1%

6	Logeswari, S., and K. Premalatha. "Ontology-based semantic smoothing model for biomedical document clustering", International Journal of Telemedicine and Clinical Practices, 2015. Publication	<1 %
7	www.science.gov Internet Source	<1 %
8	Gilles Bisson. "Chi-Sim: A New Similarity Measure for the Co-clustering Task", 2008 Seventh International Conference on Machine Learning and Applications, 12/2008 Publication	<1 %
9	Aparicio Carrasco, Roxana K. "Unsupervised classification of text documents", Proquest, 20111108 Publication	<1 %
10	Preeti Jain. "Combined Influence of Hall Current and Soret Effect on Chemically Reacting Magnetomicropolar Fluid Flow from Radiative Rotating Vertical Surface with Variable Suction in Slip-Flow Regime", International Scholarly Research Notices, 2014 Publication	<1 %
11	Sakinc, Eren. "Manufacturing Cost Prediction in the Presence of Categorical and Numeric Design Attributes", Auburn University, 2023 Publication	<1 %

12	Submitted to University of Birmingham Student Paper	<1 %
13	docplayer.net Internet Source	<1 %
14	zombiedoc.com Internet Source	<1 %
15	cobweb.cs.uga.edu Internet Source	<1 %
16	www.mdpi.com Internet Source	<1 %
17	www.coursehero.com Internet Source	<1 %
18	pure.tue.nl Internet Source	<1 %
19	research.birmingham.ac.uk Internet Source	<1 %
20	Bücker, Thies. "Costumer Clustering in the Insurance Sector by Means of Unsupervised Machine Learning : An Internship Report", Universidade NOVA de Lisboa (Portugal), 2024 Publication	<1 %
21	etheses.whiterose.ac.uk Internet Source	<1 %
22	spectrum.library.concordia.ca Internet Source	<1 %

23	link.springer.com Internet Source	<1%
24	discovery.ucl.ac.uk Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20