Concepts of relative sample outlier (RSO) and weighted sample similarity (WSS) for improving performance of clustering genes: co-function and co-regulation

Anindya Bhattacharya*

Department of Microbiology, Immunology and Biochemistry,
Center for Integrative and Translational Genomics,
University of Tennessee Health Science Center,
Memphis, TN 38163, USA
Email: abhatta3@uthsc.edu
*Corresponding author

Nirmalya Chowdhury

Department of Computer Science and Engineering,
Jadavpur University,
Kolkata 700032, West Bengal, India
Email: nirmalya_chowdhury@yahoo.com

Rajat K. De

Machine Intelligence Unit,
Indian Statistical Institute,
Kolkata 700108, West Bengal, India
Email: rajat@isical.ac.in

Abstract: Performance of clustering algorithms is largely dependent on selected similarity measure. Efficiency in handling outliers is a major contributor to the success of a similarity measure. Better the ability of similarity measure in measuring similarity between genes in the presence of outliers, better will be the performance of the clustering algorithm in forming biologically relevant groups of genes. In the present article, we discuss the problem of handling outliers with different existing similarity measures and introduce the concepts of Relative Sample Outlier (RSO). We formulate new similarity, called Weighted Sample Similarity (WSS), incorporated in Euclidean distance and Pearson correlation coefficient and then use them in various clustering and biclustering algorithms to group different gene expression profiles. Our results suggest that WSS improves performance, in terms of finding biologically relevant groups of genes, of all the considered clustering algorithms.

Keywords: similarity measure; *z*-score; *P*-value; functional enrichment; transcription factors.

Reference to this paper should be made as follows: Bhattacharya, A., Chowdhury, N. and De, R.K. (2015) 'Concepts of relative sample outlier (RSO) and weighted sample similarity (WSS) for improving performance of clustering genes: co-function and co-regulation', *Int. J. Data Mining and Bioinformatics*, Vol. 11, No. 3, pp.314–330.

Biographical notes: Anindya Bhattacharya is a Postdoctoral Fellow in the Department of Microbiology, Immunology and Biochemistry, Center for Integrative and Translational Genomics, University of Tennessee Health Science Center, Memphis, TN, USA. He obtained his PhD in Engineering from Jadavpur University, India in 2011. His research interest includes bioinformatics, data mining, pattern recognition and soft computing.

Nirmalya Chowdhury is working as an Associate Professor in the Department of Computer Science and Engineering, Jadavpur University, India. He did his PhD in Engineering from Jadavpur University, India in 1997. His fields of research include pattern recognition, soft computing, natural language processing and bioinformatics.

Rajat K. De is a Professor in the Indian Statistical Institute, Kolkata, India. He obtained his PhD degree from the Indian Statistical Institute, India, in 2000. He was a Distinguished Postdoctoral Fellow in the Whitaker Biomedical Engineering Institute, the Johns Hopkins University, USA, during 2002–2003. He has about 70 research articles published in international journals, conference proceedings and in edited books to his credit. His research interest includes bioinformatics, computational biology, systems biology, pattern recognition and soft computing.

This paper is a revised and expanded version of a paper entitled 'A methodology for handling a new kind of outliers present in gene expression patterns' presented at the '4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11)', Moscow, 26–30 June 2011.

1 Introduction

Any clustering algorithm involves measuring similarity between a pair of objects. Some standard similarity measures used in various clustering algorithms include Euclidean distance (Tou and Gonzalez, 1974; Gun et al., 2005), vector angle (Knudsen, 2001), Pearson's correlation coefficient (Tou and Gonzalez, 1974; Gun et al., 2005), rank correlation (Gun et al., 2005), Mahalanabis distance (Gun et al., 2005) and most recently developed Maximal Information Coefficient (MIC) score (Reshef et al., 2011). Choice of a similarity measure plays an important role in the performance of a clustering algorithm.

The aforesaid similarity measures are all right if the objects in a dataset are evenly distributed over the space. On the other hand, if some of objects due to noise or other factors, called outliers, are included in a dataset, these similarity measures may not lead to good performance of the clustering algorithms. They may be biased towards these outliers. There exist various methods for handling such outliers (Hawkins, 1980; Schiffman et al., 1981; Rousseeuw and Leory, 1987; Kaufman and Rousseeuw, 1990; Barnett and Lewis, 1994; Fawcett, and Provost, 1997; Han and Kamber, 2001; Shekhar and Chawla, 2002; Hu and Sung, 2003). They include, among others, statistical approach, distance-based approach, clustering-based approach, density-based local outlier detection

approach and deviation-based approach (Hawkins, 1980; Schiffman et al., 1981; Rousseeuw and Leory, 1987; Kaufman and Rousseeuw, 1990; Barnett and Lewis, 1994; Fawcett, and Provost, 1997; Han and Kamber, 2001; Shekhar and Chawla, 2002; Hu and Sung, 2003).

Statistical distribution-based methods assume a probability distribution (e.g., normal or Poisson distribution) for a given data set and then identify outliers with respect to the model (Hawkins, 1980 Rousseeuw and Leory, 1987; Barnett and Lewis, 1994) from the data set parameters (e.g., assumed data distribution), distribution parameters (e.g., mean and variance) and the expected number of outliers. Distance-based methods are usually based on local distance measures and are capable of handling large databases (Fawcett and Provost, 1997). In this approach, the distance of a point from its K-nearest points (or neighbours) is calculated. If the neighbouring points are relatively close to it, then the point is normal. Otherwise, the point is considered as an outlier. Another class of outlier detection methods is based on clustering techniques, where a cluster of small sizes can be considered as clustered outliers (Kaufman and Rousseeuw, 1990; Shekhar and Chawla, 2002; Hu and Sung, 2003). Density-based detection techniques for local outliers (Schiffman et al., 1981; Shekhar and Chawla, 2002) assign a degree to each object to be an outlier. This degree is called the Local Outlier Factor (LOF) of an object and depends on how isolated the object is with respect to the surrounding neighbourhood. In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighbourhood. Deviation-based outlier detection method (Han and Kamber, 2001) identifies outliers by examining the main characteristics of the objects in a group. Objects that 'deviate' from this description are considered as outliers. Hence, in this approach the term 'deviations' is typically used to refer to outliers.

In a gene expression data, there may be pairs of genes that have completely different expression values over a few samples under certain experimental condition(s), although they exhibit similar behaviour over the other samples. Depending on the algorithms, these outliers are either placed in single element clusters (hierarchical clustering) or allowed to be in a cluster that is more similar compared to others (partitioning clustering) or they may be completely discarded from grouping (density-based, grid-based and graph-based clustering). In all these cases outliers affect the outcome of a clustering result. Measurement errors or conditional changes during microarray experiments cause a sample differ in expression level compared to the other samples. This difference of gene expression value for an outlier sample may be ten times, hundred times or even thousand times more or less than the average expression value of the gene over all the samples. In this way expression values of the outlier samples may cause a gene to be an outlier.

If the expression value(s) of a single (both) gene(s) corresponding to a sample differ much from its (their) mean expression value(s) of the other samples, then the expression value(s) for this(ese) sample(s) gives rise to the notion of a different kind of outlier which is introduced in this article. That is, the sample is an outlier with respect to a gene pair. We call such an outlier as a Relative Sample Outlier (RSO) with respect to a pair of gene. Distance/similarity measures used by different clustering algorithms are unable to treat an outlier sample and a normal sample differently. All the samples contribute equally during the measurement of distance/similarity.

Pearson correlation coefficient measures the similarity in the pattern of changes in expression profiles of a pair of genes. Higher similarity in expression profiles of a gene pair induces a higher Pearson correlation coefficient value and vice versa. Pearson correlation coefficient dose not perform well in the presence of outliers (Heyer et al., 1999). If a gene pair have common peak or valley at a single sample, the correlation value

will be high, although the patterns at the remaining samples may be completely dissimilar. This observation motivated the development of a similarity measure, called Jackknife correlation coefficient (Heyer et al., 1999). It is defined as the minimum of Pearson correlation coefficient values obtained by considering all the samples and also considering one sample omitted (by taking one measurement for all the single sample omission).

Use of Jackknife correlation coefficient avoids the overwhelming effect of the single outlier sample over all the other samples. More general versions of Jackknife correlation coefficient that are robust to more than one outlier can similarly be derived. However, a generalised form of Jackknife correlation coefficient, which would involve the enumeration of different combinations of samples to be deleted, would be computationally costly and is rarely used. Moreover, Jackknife correlation coefficient cannot handle a situation with outlier, when one sample value of a pair of genes has large dissimilarity (outlier) than the others and remaining sample values are similar. In this case, instead of deleting the outlier sample value, Jackknife correlation coefficient will delete the sample value that has highest contribution to the similarity score.

Some of the other attempts for the detection of outlier samples in gene expression data are by Kadota et al. (2003), Ge et al. (2005) and Cheng and Wong (2001). Performances of these approaches depend largely to the maximum number of outlier samples and the number of samples in the dataset.

In order to improve performance of the various similarity measures (including Euclidean distance and Pearson correlation coefficient), with respect to better ability of handling outliers, we introduce the concept of Weighted Sample Similarity (WSS). Instead of using a sample value for the similarity measure, we multiply a weight value with the expression value of a sample and then use the resulting value. Weight values are determined in such a way that possible outliers are assigned smaller weight values (i.e. nearly equal to zero). Weight values for non outliers are large and are nearly equal to 'one'. With this new WSS, Euclidean distance or Pearson correlation coefficient involves low contribution of outlier samples and high contribution to non outlier samples. It is to be mentioned here the notion of WSS is introduced to take care of RSOs.

The concept of assigning weight value to each of the gene expression samples was first introduced by Bland and Altman (1995) who found that multiple observations from each subject produced a spurious increase in the sample size and a corresponding spurious significant relationship. They used the number of observations as weights and use weight values in the computation of Pearson correlation coefficient. Higher the value of weight, more important the corresponding sample is. While calculating weighted correlation coefficient between two genes, the samples with more importance should have more impact on correlation coefficient than the other samples.

WSS generalises the concept of assignment of a weight value to gene expression samples, for handling outliers. WSS uses the the Gaussian kernel function as a measure of closeness between gene expression values of a pair of genes under an experiment and their average gene expression values.

Any similarity measure that computes pair wise distance similarity, can use the WSS for better handling the outliers. For comparison, Euclidean distance with WSS (*WSS-D*) and without WSS (*D*), Pearson correlation coefficient with WSS (*WSS-Corr*) and without WSS (*Corr*), Spearman rank-order correlation coefficient (*Rank-Corr*) and Jackknife correlation coefficient (*Jack-Corr*) are used with clustering algorithms K-means (Jain and Dubes, 1988; Tavazoie et al., 1999; Han and Kamber, 2001), a constant factor approximation algorithm (MIND) by Bansal et al. (2004), DCCA (Bhattacharya and De, 2008) and ACCA (Bhattacharya and De, 2010), a biclustering algorithm BCCA

(Bhattacharya and De, 2009). All the instances of these algorithms are applied on different gene expression datasets and performances are assessed using different cluster validity indices.

2 Relative sample outlier (RSO)

Let us consider a set of experiments with microarray gene expression measurements of a particular tissue under the same treatment. That is, either different healthy individuals, different diseased individuals or different samples from the same individual are considered. In all the cases, the individual(s) were under the same treatment, if any. Due to error or other factors in microarray measurements, the expression profiles for a pair of genes may be similar over all the samples except for a few. For these few samples, expression values of the same pair of genes may either differ drastically or differ a lot from the other samples. Seven pairs of artificially generated gene expression profiles were used in Figure 1 to show some of these variations over samples. Figure 1 (a) depicts that the gene pair ($Gene_1$, $Gene_2$) have a large difference in expression values for one measurement although their expression patterns are similar over the remaining samples. The sample(s) for which the expression values differ drastically for the pair of genes, give rise to the notion of a different kind of outlier. That is, the sample is an outlier with respect to the gene pair. We call the outlier as RSO. It may be mentioned here that RSO is different from the notion of outliers already available in literature (Han and Kamber, 2001). In the later case, the gene as a whole needs to be treated as an outlier with respect to a group of genes, in contrary to the former one where a sample is considered as an outlier corresponding to a gene pair. This situation affects clustering if we consider similarity computation based on expression values only.

Figure 1 Various simulated expression patterns of seven artificial gene pairs

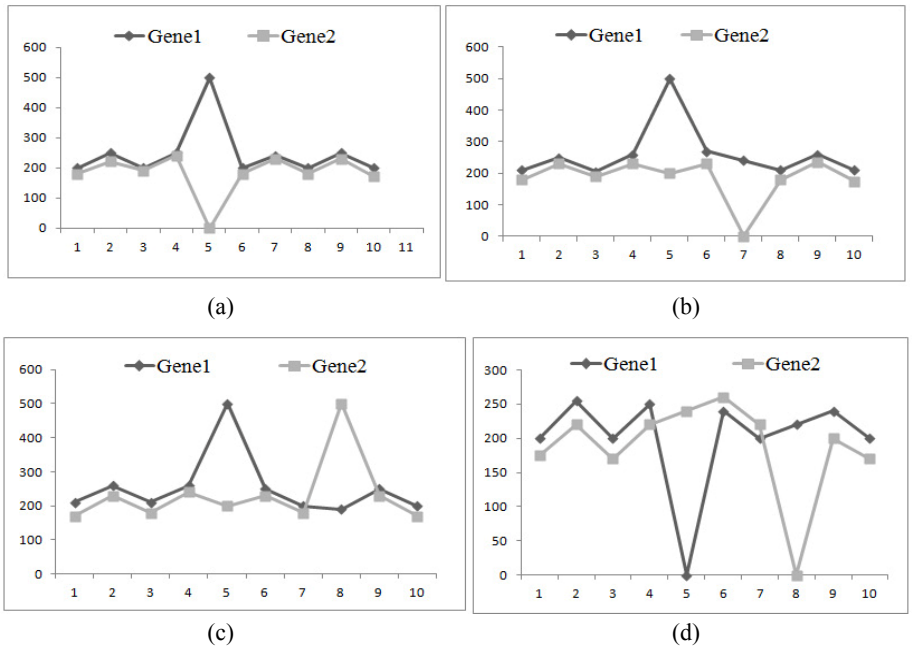
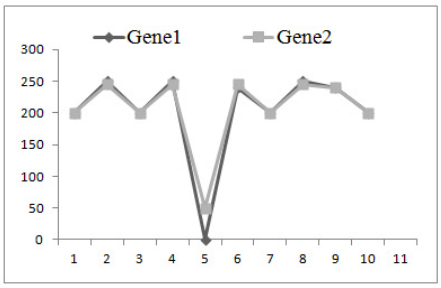
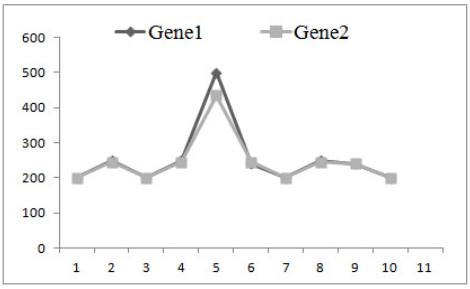


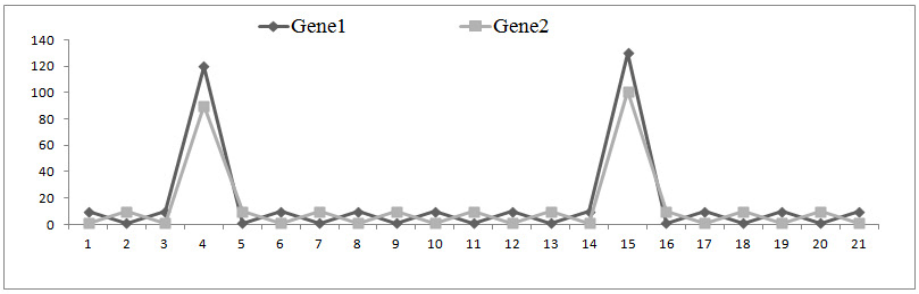
Figure 1 Various simulated expression patterns of seven artificial gene pairs (continued)



(e)



(f)



(g)

Similar situations arise for Figures 1 (b)–(d), where gene pairs have the similar variation in their expression profiles over all the samples except for two. Figures 1 (e) and (f) show that the expression values of the gene-pairs are similar over all the samples, but for one sample their expression values differ significantly from respective mean values computed over the other samples. In Figure 1 (g), the gene pair ($Gene_1$, $Gene_2$) has opposite pattern of variation in their expression values over all the samples except for two. For these two samples, they have the similar patterns of expression values. In order to tackle all these situations, one may find the sample(s) for which the said difference is much and then adopt some necessary arrangement to get rid of it. But this will enhance the complexity of a clustering algorithm that identifies the sample(s), if any, corresponding to every gene pair for which the above situation(s) arises. In order to reduce this complexity, we introduce WSS, to take care of the effect of such outliers, i.e. RSOs.

3 Formulation of WSS

Let us consider a set of n genes $X = \{g_1,g_2,\ldots,g_n\}$, for each of which m expression values are given. Let us also consider a set of m microarray experiments/samples (measurements) $Y = \{e_1,e_2,\ldots,e_m\}$. For each experiment, we have n expression values corresponding to n genes in X. That is, for each gene g_i, there is an m-dimensional vector $\mathbf{x}_i$, where x_{il} is the expression value of g_i in l-th experiment e_l.

Similarity between gene pair (g_i, g_j) may be computed using Euclidean distance $D(\mathbf{x}_i, \mathbf{x}_j)$ or Pearson correlation coefficient $Corr(\mathbf{x}_i, \mathbf{x}_j)$ and are defined, respectively, as:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2} \tag{1}$$

and

$$Corr(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{l=1}^{m} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{m} (x_{il} - \bar{x}_i)^2 \sum_{l=1}^{m} (x_{jl} - \bar{x}_j)^2}} \tag{2}$$

Here $\bar{x}_i$ and $\bar{x}_j$ are mean values over m expression values of i-th and j-th genes, respectively.

If l-th expression values of a co-expressed, co-regulated gene pair (g_i, g_j), corresponding to an experiment e_l, are such that the sample is a RSO with respect to gene pair (g_i, g_j), both equations (1) and (2) may be biased towards this outlier. That is, if we consider equation (2) for measuring similarity, the value should ideally be close to 1 for a pair of co-regulated genes. Due to this RSO, the correlation value will differ much from 1. In order to reduce this type of misleading contribution of RSO, we introduce the notion of weighting coefficient w_{ij} corresponding to l-th expression value and gene pair (g_i, g_j), for all l, i, j.

We determine the weight values to reduce the effect of such outliers by assigning lower weight values corresponding to the outlier samples of gene pair (g_i, g_j) and higher weight values to the other samples. In other words, higher the difference in l-th expression values of the genes in the pair (g_i, g_j) from their means, lower is the value of the weight w_{lij}.

Gaussian kernel function is a measure of closeness between data points with values between '0' and '1' where a value close to '1' represents closeness. In WSS for the weight, we used a Gaussian kernel function w_{lij} as:

$$w_{lij} = e^{(-\alpha \times D_{ijl}^2)} \tag{3}$$

where D_{ijl} is Euclidean distance for computing difference in l-th expression values of both the genes g_i and g_j from their means. We have:

$$D_{ijl} = \sqrt{(t_{il} - \bar{t}_i)^2 + (t_{jl} - \bar{t}_j)^2} \tag{4}$$

where t_{il} and t_{jl} are normalised expression values in $[0,1]$ of x_{il} and x_{jl}, respectively. Similarly, $\bar{t}_i$ and $\bar{t}_j$ are mean of normalised expression values, computed over all the samples of gene g_i and g_j, respectively. Here we have considered normalised expression values in equation (4), for keeping D_{ijl} bounded to a known value $\sqrt{2}$ ($\sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2}$). For an outlier sample, measured value of D_{ijl} should be high. Weight value w_{lij} for an outlier sample e_l corresponding to a pair of gene (g_i, g_j) should be low. Thus relationship between D_{ijl} and w_{lij} should be such that, increase in D_{ijl} should cause decrease in w_{lij} and vice versa.

In kernel function w, α is a adjustable parameter with a value always greater than zero. Here the value of α should be such that w_{lij} is nearly equal to zero for $D_{ijl}=\sqrt{2}$. On the other hand, w_{lij} should tend to one for a non-outlier sample. In fact this happens as D_{ijl} tends to zero for a non-outlier sample. To determine α, we used functional enrichment plot and it is described in the result section.

WSS function incorporates w_{lij}, for each l-th experiment, in equations (1) and (2). In these equations, x_{il} and x_{jl} are replaced by $x_{ijl}^{(w)}=w_{lij}\times x_{il}$ and $x_{jil}^{(w)}=w_{lij}\times x_{jl}$, respectively. Similarly, mean values $\bar{x}_i$ and $\bar{x}_j$, in equation (2), are replaced by $\bar{x}_{ijl}^{(w)}=\frac{1}{m}\sum_{l=1}^{m} w_{lij}\times x_{il}$ and $\bar{x}_{jil}^{(w)}=\frac{1}{m}\sum_{l=1}^{m} w_{lij}\times x_{jl}$, respectively. It is to be mentioned here that $\bar{x}_{ijl}^{(w)}\neq\bar{x}_{jil}^{(w)}$, although both of them involved the same w_{lij}s. It is further to be noted that the terms t_{il}, t_{jl}, $\bar{t}_i$ and $\bar{t}_j$ in equation (4) are computed using x_{il}, x_{jl} only. Thus we get WSS distance *WSS-D*($\mathbf{x}_i$, $\mathbf{x}_j$) and correlation coefficient *WSS-Corr*($\mathbf{x}_i$, $\mathbf{x}_j$) between a gene pair (g_i, g_j), based on Euclidian distance and Pearson correlation coefficient, respectively.

4 Results

The effectiveness of WSS along with comparative analysis with the aforesaid similarity measures is demonstrated with clustering algorithms K-means (Jain and Dubes, 1988; Tavazoie et al., 1999; Han and Kamber, 2001), MIND (Bansal et al., 2004), DCCA (Bhattacharya and De, 2008), ACCA (Bhattacharya and De, 2010) and a biclustering algorithm, BCC (Bhattacharya and De, 2009) using five gene expression datasets. These datasets deal with two yeasts (Yeast Cheng and Church Dataset (YCCD) (Tavazoie et al., 1999; Cheng and Church, 2000) and Spellman et al. Dataset (SPTD) (Spellman et al., 1998) and three mammals (GDS958 (Wills-Karp and Ewart, 2004), GDS2547 (Yu et al., 2004; Chandran et al., 2007) and GDS2938 (Wang et al., 2007)). The performance of all the algorithms is also demonstrated using several indices *viz.*, *z*-score (Gibbons and Roth, 2002) for homogeneity and *P*-value (on functional annotation and on transcription factors) (Gun et al., 2005) for cluster reliability. Null rows/columns and rows/columns with all zeros are deleted from the datasets before applying these clustering algorithms. For example, five such rows are deleted from original Yeast CC dataset (YCCD). Five datasets are described briefly in Table 1. Parameter settings used in different algorithms are presented in Table 2.

Table 1 A short description of the datasets used in clustering and biclustering algorithms with WSS

Name (organism)	*Number of genes*	*Number of samples*
Yeast Cheng and Church dataset (YCCD) (Yeast)	2879	17
Spellman et al. dataset (SPTD) (Yeast)	6178	77
GDS958 (Mouse)	22690	12
GDS2547 (Homosapiens)	12646	164
GDS2938 (Homosapiens)	22283	12

Table 2 Parameter settings used for different clustering methods

Name (organism)	*Algorithm*	*Parameters*
Yeast Cheng and Church dataset (YCCD) (Yeast)	K-means	$\alpha = 3$, $K = 5$
	MIND	$\alpha = 3$
	DCCA	$\alpha = 3$
	ACCA	$\alpha = 3$, $K = 5$
	BCCA	$\alpha = 3$, $r = 0.7$
Spellman et al. dataset (SPTD) (Yeast)	K-means	$\alpha = 4$, $K = 17$
	MIND	$\alpha = 4$
	DCCA	$\alpha = 4$
	ACCA	$\alpha = 4$, $K = 17$
	BCCA	$\alpha = 4$, $r = 0.7$
GDS958 (Mouse)	K-means	$\alpha = 3$, $K = 7$
	MIND	$\alpha = 3$
	DCCA	$\alpha = 3$
	ACCA	$\alpha = 3$, $K = 7$
	BCCA	$\alpha = 3$, $r = 0.7$
GDS2547 (Homosapiens)	K-means	$\alpha = 4$, $K = 23$
	MIND	$\alpha = 4$
	DCCA	$\alpha = 4$
	ACCA	$\alpha = 4$, $K = 23$
	BCCA	$\alpha = 4$, $r = 0.7$
GDS2938 (Homosapiens)	K-means	$\alpha = 4$, $K = 6$
	MIND	$\alpha = 3$
	DCCA	$\alpha = 3$
	ACCA	$\alpha = 3$, $K = 6$
	BCCA	$\alpha = 3$, $r = 0.7$

Notes: α is the adjustable parameter, K represents number of clusters and r represents Pearson correlation coefficient threshold.

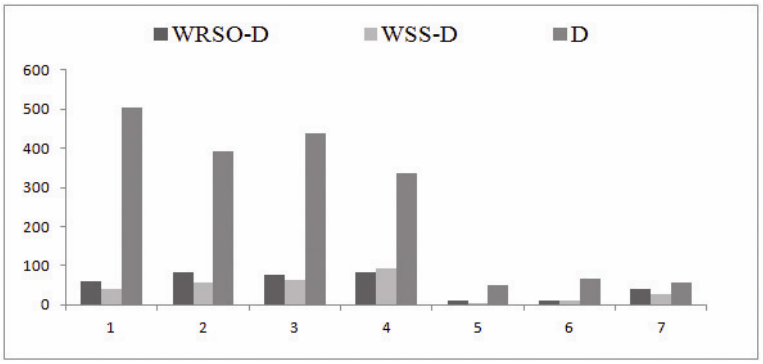
Before performing gene expression data analysis using clustering algorithms, we have compared different similarity measures for their ability to handle outliers with seven different pairs of gene expression data profile. This is described in the next subsection.

4.1 Comparison on the ability of handling RSO

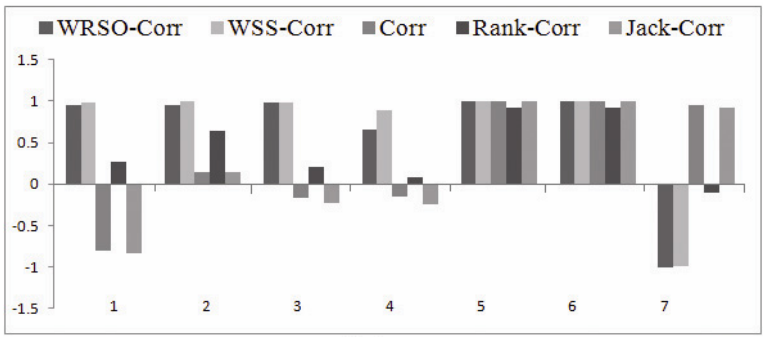
Standard similarity values for all seven gene pairs, in Figures 1 (a)–(g), are measured by Euclidean distance (*WRSO-D*) and Pearson correlation coefficient (*WRSO-Corr*) by ignoring all the samples that are RSOs corresponding a pair of genes. For comparison, Euclidean distance with WSS (*WSS-D*) and without WSS (*D*), Pearson correlation coefficient with WSS (*WSS-Corr*) and without WSS (*Corr*), Spearman rank-order

correlation coefficient (*Rank-Corr*) and Jackknife correlation coefficient (*Jack-Corr*) are used to measure similarity for all the seven gene pairs. Figure 2 (a) shows measured Euclidean distance values for seven gene pairs. It is evident from Figure 2 that Euclidean distance with WSS (*WSS-D*) is closer to the standard Euclidean distance (*WRSO-D*) than Euclidean distance without WSS (*D*). It is to be mentioned here that *WRSO-D* considers standard Euclidean distance with no RSO. That is, the gene pairs show the similar (opposite) expression pattern over all the samples considered for *WRSO-D*. On the other hand, Euclidean distance *D* consider all the samples irrespective of the fact whether the sample is a RSO or not corresponding to a gene pair. Thus the computation of *D* may be biased towards these outliers. Therefore, the results of *WSS-D* and *WRSO-D* may be considered as unbiased by these outliers. Similarly, *WSS-Corr* and *WRSO-Corr* should be unbiased and *Corr* may be biased towards the said outliers as expected. Moreover, it is observed from Figure 2 (b) that *Jack-Corr*is biased towards these outliers but *Rank-Corr* remains unbiased.

Figure 2 Values of similarity measures for seven gene pairs in Figure 1 (a) Euclidean distances; (b) correlation measures



(a)



(b)

4.2 Homogeneity comparison using z-score

For homogeneity comparison, we used *z*-score . *z*-score is calculated by investigating the relation between a clustering result and the functional annotation of the genes in the

cluster. To calculate *z*-score for two yeast datasets, Gibbons ClusterJudge tool has been used. Saccharomyces Genome Database (SGD) annotation of the yeast genes, along with the gene ontology developed by the Gene Ontology Consortium , has been used by ClusterJudge for calculation of *z*-scores. ClusterJudge only supports yeast datasets. For GDS958, annotation dataset GPL339 and for GDS2547 and GDS2938 annotation dataset GPL97 have been used. A higher value of *z* indicates that genes would be better clustered by function, indicating a more biologically relevant clustering result. Thus higher *z*-score indicates the resulting clusters being more homogenous. Figure 3 shows the results obtained by K-means algorithm with *WSS-D* (Euclidean distance with WSS) has a higher *z*-score compared to K-means algorithm with *D* (Euclidean distance without WSS) for all five datasets. Figure 4 shows that all the algorithms with *WSS-Corr* (Pearson correlation coefficient with WSS) result in higher *z*-scores compared to the algorithms with all the other correlation based measures for all the five datasets.

Figure 3 *z*-score for K-means algorithm with Euclidean distance

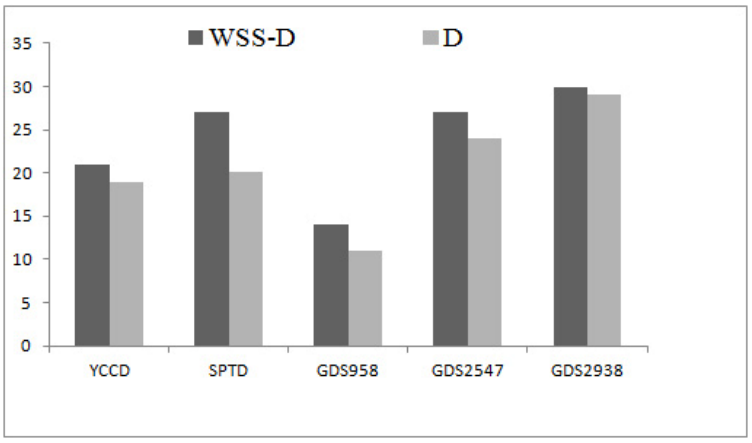
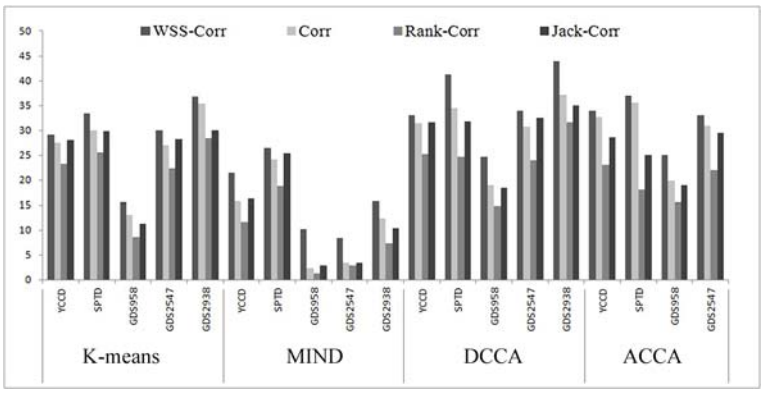


Figure 4 *z*-score for clustering algorithms with correlation coefficients



4.3 Functional enrichment in terms of P-value

For gene expression data analysis, *P*-value represents the probability of observing at least a given number of genes, in a cluster, from a specific GO functional category. A specific GO functional category is said to be 'enriched' if the corresponding *P*-value is less than a predefined threshold value. In the present article, only functional categories with *P*-value $<5.0 \times 10^{-7}$ are reported as 'enriched'. A low *P*-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. A clustering solution is considered to be more reliable if the number of functional categories obtained from a cluster is high. In order to compare the performance of different clustering algorithms for a microarray gene expression dataset, we used the average number of functionally enriched attributes found per cluster.

Figure 5 shows that the performance of K-means clustering algorithm with *WSS-D* (Euclidean distance with WSS) is much larger compared to K-means algorithm with *D* (Euclidean distance but without WSS) for all the five datasets. Similarly, Figure 6 provides the comparative analysis of the clustering algorithms with four similarity measures and shows that *WSS-Corr* (Pearson correlation coefficient with WSS) provides higher number of enriched attributes compared to algorithms with all the other correlation based measures for all the five datasets. Figure 7 shows that the biclustering algorithm BCCA with *WSS-Corr* outperforms BCCA with *Corr* in finding functionally enriched attributes per bicluster.

4.4 Significant transcription factors in terms of P-value

We considered the tool PRIMA available in EXPANDER for analysis of transcription factor binding sites corresponding to the resulting clusters and biclusters. Number of enriched transcription factors for each cluster/bicluster of two yeast datasets YCCD and SPTD was found based on *P*-values. Similar to the analysis on functional enrichment, here we compared performance of different similarity measures by average number of enriched transcription factors per cluster. Transcription factors with *P*-value $<1.0 \times 10^{-4}$ are considered for enrichment comparison. Higher the average number of enriched transcription factors per cluster, better is the chance of finding co-regulated groups of genes.

Figure 5 Average number of functionally enriched attributes per cluster for K-means algorithm with Euclidean distance

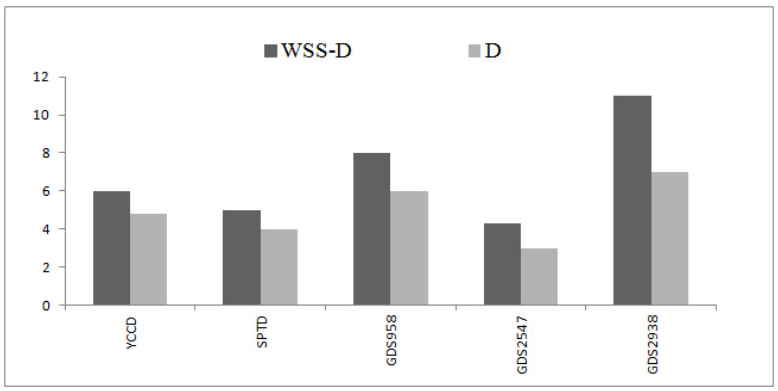


Figure 6 Average number of functionally enriched attributes per cluster for clustering algorithms with correlation coefficients

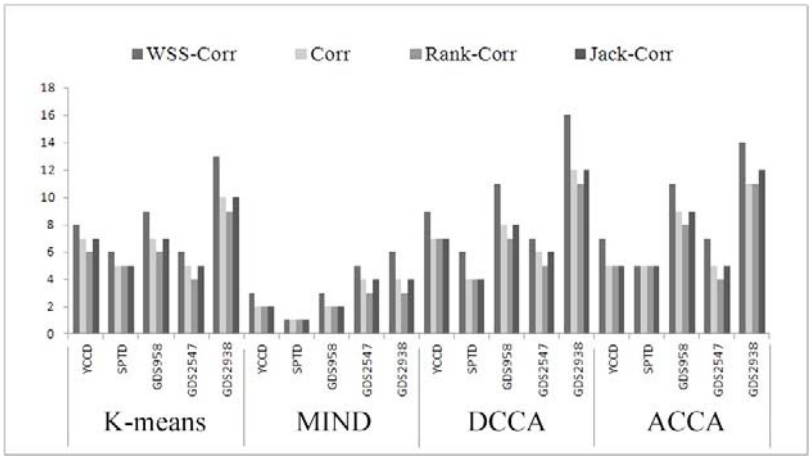
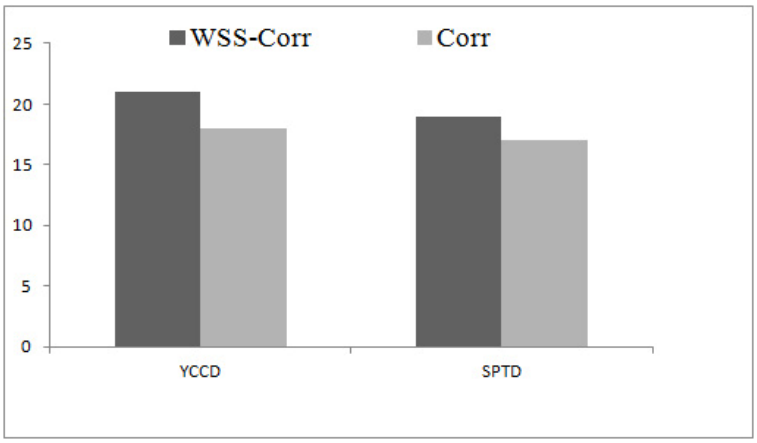


Figure 7 Average number of functionally enriched attributes per bicluster for BCCA with correlation coefficients



Figures 8–10 show the average number of enriched transcription factors per cluster or bicluster on two yeast datasets. The number of enriched transcription factors per cluster (Figure 8) corresponding to K-means algorithm with *WSS-D* (Euclidean distance with WSS) is higher compared to K-means algorithm with *D* (Euclidean distance without WSS) for both YCCD and SPTD datasets. Similarly, all the algorithms with *WSS-Corr* (Pearson correlation coefficient with WSS) (Figure 9) result in higher numbers of enriched transcription factors per cluster compared to the algorithms with all the other correlation based measures for two yeast datasets. The algorithm BCCA (Figure 10) with *WSS-Corr* outperforms BCCA with other correlation measures in obtaining functionally enriched transcription factors per bicluster.

Figure 8 Average number of enriched transcription factors per cluster for K-means algorithm with Euclidean distance

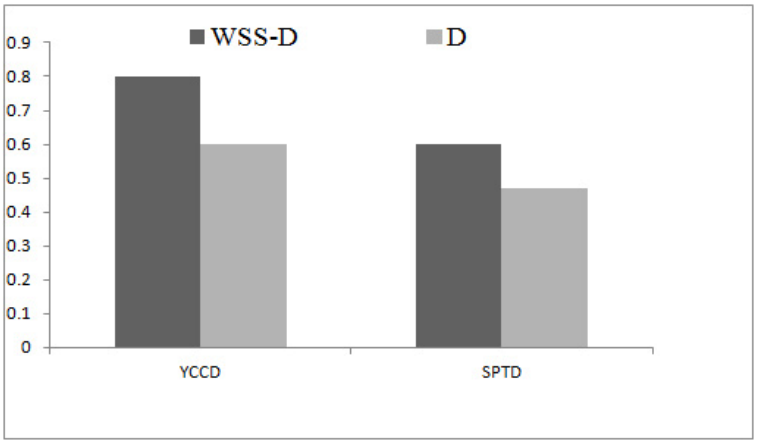
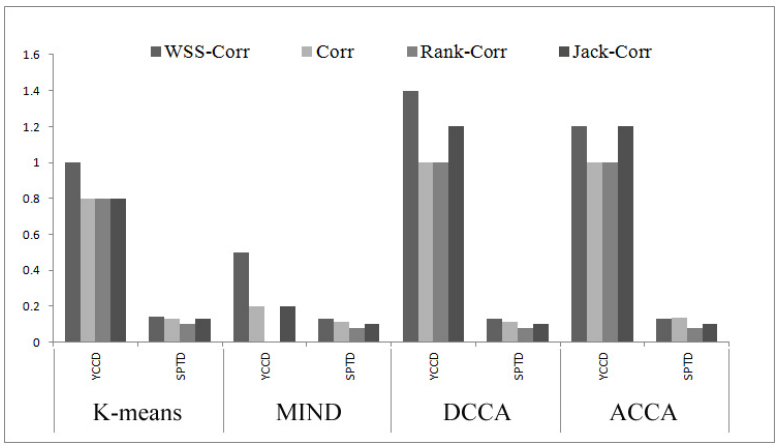


Figure 9 Average number of enriched transcription factors per cluster for clustering algorithms with correlation coefficients



4.5 Estimating the value of α

A suitable value of α for WSS has been determined from the plots of total number of functionally enriched GO functional category from a clustering output respect to α. Figure 11 shows variations of the total number of functionally enriched categories (Y axis) with respect to α (X axis) between 1 and 10 for clustering output of K-means algorithm with *WSS-D* and *WSS-Corr*, respectively, on YCCD dataset. From the plots, it is found that for α value 3, number of functionally enriched categories are highest. With a farther increase in α, the total number of functionally enriched categories are not improving. From this plot we have chosen α as 3. Table 2 shows selection of α for different algorithms with respect to different datasets. Table 2 also shows selection of other parameter values for execution of algorithms.

Figure 10 Average number of enriched transcription factors per bicluster for BCCA with correlation coefficients

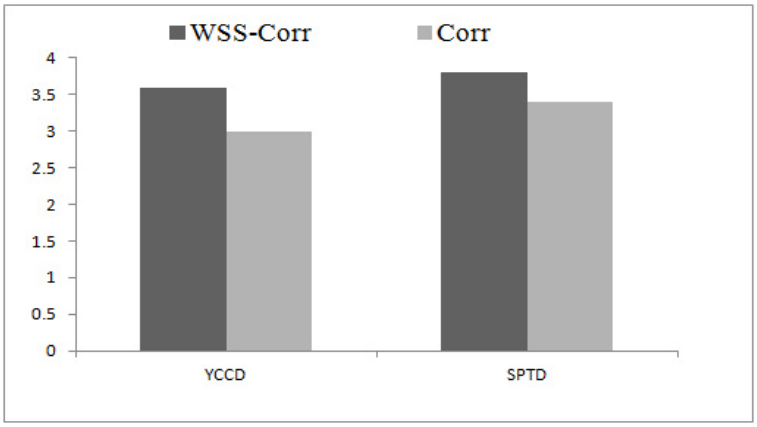
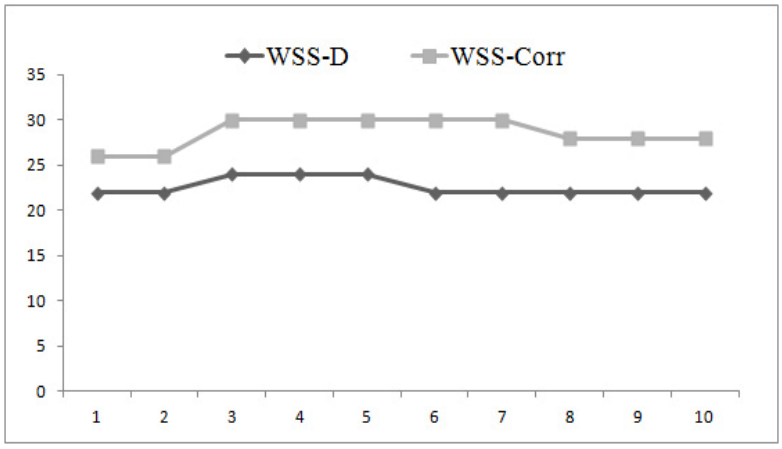


Figure 11 Variation of total number of functionally enriched GO categories from clustering output of K-means on YCCD dataset with the variation in α



5 Conclusions

Here we have introduced the concepts of RSO and WSS. RSOs are the samples with respect to a gene pair, for which sample values show large difference from the other samples corresponding to the gene pair. Incorporation of the notion of WSS helps in dealing with such outliers while measuring similarity between a pair of gene.

Analysis of the results on different clustering algorithms with the use of different similarity measures suggests that WSS used with a similarity measure improves the performance of clustering algorithms in obtaining more biologically significant clusters. This analysis also shows that incorporation of the notion of WSS enables to improve a

similarity measure including Euclidean distance, Pearson correlation coefficient in handling outliers and measure similarity more accurately between gene pairs. The main advantage of WSS is that it is able to deal with outliers without deleting them. This is an advantage because deletion of samples with outlier may cause loss of important information as the suspected outlier may not be an actual outlier. Another advantage of WSS is that it is compatible with different existing similarity measures. WSS can be used with any distance based similarity measure without changing formation of that similarity measure.

Determination of weight value for a sample pair in WSS using equation (3) depends on spatial distance of considered pair of samples from mean position. Equation (3) returns nearly equal to zero as the weight value for a pair of samples far away from their respective mean values and nearly equal to one for the weight for a sample pair close to their mean. With this weighting scheme, it is assumed that pair of values far away from mean values are outliers and to minimise the effects of these outliers they should be assigned a very low weight value. It may be possible in practice that such values are not actual outliers. Determination of weight value for a sample pair in WSS can be improved if we are able to incorporate information about experiments from which the expression values are generated along with the information associated with equation (3).

References

Bansal, N., Blum, A. and Chawla, S. (2004) 'Correlation clustering', *Machine Learning*, Vol. 56, pp.89–113.

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, Wiley, New York, NY, USA.

Bhattacharya, A. and De, R.K. (2008) 'Divisive correlation clustering algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles', *Bioinformatics*, Vol. 24, No. 11, pp.1359–1366.

Bhattacharya, A. and De, R.K. (2009) 'Bi-correlation clustering algorithm for determining a set of co-regulated genes', *Bioinformatics*, Vol. 25, No. 21, pp.2795–2801.

Bhattacharya, A. and De, R.K. (2010) 'Average correlation clustering algorithm (ACCA) for grouping of co-regulated genes with similar pattern of variation in their expression values', *Journal of Biomedical Informatics*, Vol. 43, No. 4, pp.560–568.

Bland, J.M. and Altman, D.G. (1995) Calculating correlation coefficients with repeated observations: part 2 - correlation between subjects', *British Medical Journal*, Vol. 310, No. 6980, p.633.

Chandran, U.R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M. and Monzon, F. (2007) 'Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process', *BMC Cancer*, Vol. 7, p.64.

Cheng, Y. and Church, G.M. (2000) 'Biclustering of expression data', *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 19–23 August, San Diego, CA, USA, pp.93–103.

Fawcett, T. and Provost, F. (1997) 'Adaptive fraud detection', *Data-mining and Knowledge Discovery*, Vol. 1, pp.291–316.

Ge, X.J., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M. and Aburatani, H. (2005) 'Interpreting expression profiles of cancers by genome-wide survey of breadth of expression', *Genomics*, Vol. 86, No. 2, pp.127–141.

Gibbons, F. and Roth, F. (2002) 'Judging the quality of gene expression-based clustering methods using gene annotation', *Genome Research*, Vol. 12, No. 10, pp.1574–1581.

Gun, A.M., Gupta, M.K. and Dasgupta, B. (2005) *Fundamentals of Statistics*, Vol. 2, The World Press Private Limited, Kolkata, India.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, USA.

Hawkins, D. (1980) *Identification of Outliers*, Chapman and Hall, London, UK.

Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999) 'Exploring expression data: Identification and analysis of coexpressed genes', *Genome Research*, Vol. 9, pp.1106–1115.

Hu, T. and Sung, S.Y. (2003) 'Detecting pattern-based outliers', *Pattern Recognition Letters*, Vol. 24, pp.3059–3068.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Prentice Hall, New Jersey, NJ, USA.

Kadota, K., Nishimura, S.I., Bono, H., Nakamura, S., Hayashizaki, Y., Okazaki, Y. and Takahashi, K. (2003) 'Detection of genes with tissue-specific expression patterns using akaikes information criterion', *Physiological Genomics*, Vol. 12, No. 3, pp.251–259.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, NY, USA.

Knudsen, S. (2001) *A Biologists Guide to Analysis of DNA Microarray Data*, Wiley, New York, NY, USA.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. and Sabeti, P.C. (2011) 'Detecting novel associations in large data sets', *Science*, Vol. 334, No. 6062, pp.1518–1524.

Rousseeuw, P. and Leory, A. (1987) *Robust Regression and Outlier Detection*, Wiley, New York, NY, USA.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981) *Introduction to Multidimensional Scaling: Theory, Methods and Applications*, Academic Press, New York, NY, USA.

Shekhar, S. and Chawla, S. (2002) *A Tour of Spatial Databases*, Prentice Hall, New Jersey, NJ, USA.

Spellman, P.T., Zhang, G.S.M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) 'Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization', *Molecular Biology of the Cell*, Vol. 9, No. 12, pp.3273–3297.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) 'Systematic determination of genetic network architecture', *Nature Genetics*, Vol. 22, No. 3, pp.281–285.

Tou, J.T. and Gonzalez, R.C. (1974) *Pattern Recognition Principles*, Addison-Wesley, Reading, England, UK.

Wang, S., Antwerp, M. V., Kuick, R. and Gauger, P. (2007) 'Microarray analysis of cytokine activation of apoptosis pathways in the thyroid', *Endocrinology*, Vol. 148, No. 10, pp.4844–4852.

Wills-Karp, M. and Ewart, S.L. (2004) Time to draw breath: asthma-susceptibility genes are identified', *Nature Reviews Genetics*, Vol. 5, No. 5, pp.376–387.

Yu, Y., Landsittel, D., Jing, L. and Nelson, J. (2004) 'Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy', *Journal of Clinical Oncology*, Vol. 22, pp.2790–2799.