TRIQ: A Comprehensive Evaluation Measure for Triclustering Algorithms

David Gutiérrez-Avilés and Cristina Rubio-Escudero

Department of Computer Science, University of Seville, Seville, Spain
davgutavi@gmail.com, crubioescudero@us.es

Abstract. Triclustering has shown to be a valuable tool for the analysis of microarray data since its appearance as an improvement of classical clustering and biclustering techniques. Triclustering relaxes the constraints for grouping and allows genes to be evaluated under a subset of experimental conditions and a subset of time points simultaneously. The authors previously presented a genetic algorithm, $TriGen$, that finds triclusters of gene expression dasta. They also defined three different fitness functions for $TriGen$: MSR_{3D}, LSL and MSL. In order to asses the results obtained by application of $TriGen$, a validity measure needs to be defined. Therefore, we present $TRIQ$, a validity measure which combines information from three different sources: (1) correlation among genes, conditions and times, (2) graphic validation of the patterns extracted and (3) functional annotations for the genes extracted.

Keywords: Triclustering · Validity measure · Genetic algorithms · Microarrays

1 Introduction

Data Mining has developed a vast amount of computational tools for the analysis of bioinformatics data and allows us to find new knowledge which is hidden for the human's eyesight. One of the most useful and studied approaches is the behavior pattern search in gene expression data from microarray experiments. These genes, that exhibit high correlation among their expression levels, could be involved in similar regulatory processes as relationship exists between correlation and functionality.

We focus on one behavior pattern searching technique, clustering, which analyzes the microarray dimensional space grouping genes and taking into account all experimental conditions. There are different approaches, having classic clustering techniques, which group genes based on all conditions [19], biclustering, which emerges as an evolution of clustering since it groups genes under some particular experimental conditions, and finally we have triclustering, that goes one step further by grouping genes under particular conditions and under particular time points [11], thus being capable of managing 3D data. Triclustering is therefore suitable for microarray experiment where time points are considered,

which has great interest since it allows for a deep analysis of biological processes where temporary development is important.

Both biclustering and triclustering attack NP-hard problems, and thus algorithms based on heuristics are well suited for them. In [11] we presented the TriGen algorithm, a triclustering-genetic algorithm based on an evolutionary heuristic which finds patterns of similarity for genes on a three dimensional space, thus taking into account the gene,conditions and time factors.

Definition of fitness functions for the genetic algorithm is an essential task. In biclustering, a classic measure is the Mean Squared Residue (MSR) [6]. We have defined a three dimensions adaptation of this measure, MSR_{3D} [9]. Furthermore, we have defined two other fitness functions improving the behavior of MSR_{3D}. Both are based on the similarity among the slopes of the angles. The first one, Least Squared Lines (LSL) [12], measures the quality of a tricluster based on the similarity among the slopes of the angles formed by the least squares lines from each of the profiles formed by the genes, conditions and times of the tricluster. The second one, Multi Slope Measure (MSL) [10], measures the quality of a tricluster based on the similarity among the angles of the slopes formed by each profile formed by the genes, conditions and times of the tricluster.

The results obtained by the three proposed fitness functions have been validated in three different ways. First, by analyzing the correlation among the genes, conditions and times in each tricluster using two different correlation measures: Pearson [4] and Spearman [13]. Second, by a graphic validation of the patterns extracted based on the graphic representation, and third we have provided functional annotations for the genes extracted from the Gene Ontology project (GO) [1]. These three validation measures have been chosen since they are the standard ones for this topic as can be seen in [14,15,17,22,23].

However, we consider that providing a single evaluation measure capable of combining the information from the three mentioned sources of validation will be a great improvement. Therefore, in this work we propose $TRIQ$, a validation measure which combines the three previously proposed validation mechanisms (correlation, graphic validation and functional annotation of the genes).

2 Related Works

This section is to provide a general overview of triclustering published in literature. We particularly focus on the validation methods applied to assess the quality of the triclusters obtained.

In 2005, Zhao and Zaki [23] introduced the triCluster algorithm to extract patterns in 3D gene expression data. They presented a measure to assess triclusters's quality based on the symmetry property. They validated their triclusters based on their graphical representation and Gene Ontology (GO) results.

g-triCluster, an extended and generalized version of Zhao and Zaki's proposal, was published one year later [15]. The authors claimed that the symmetry property is not suitable for all patterns present in biological data and propose the Spearman rank correlation [20] as a more appropriate tricluster evaluation measure. The also showed validation results based on GO.

An evolutionary computation proposal was made in [16]. The fitness function defined is a multi-objective measure which tries to optimize three conflicting objectives: clusters size, homogeneity and gene-dimension variance of the 3D cluster. The tricluster quality validation was based on GO.

LagMiner was introduced in [22] to find time-lagged 3D clusters, what allows to find regulatory relationships among genes. It is based on a novel 3D cluster model called S^2D^3 Cluster. They evaluated their triclusters on homogeneity, regulation, minimum gene number, sample subspace size and time periods length. Their validation was based on graphical representation and GO results.

Hu *et al.* presented an approach focusing on the concept of Low-Variance 3-Cluster [14], which obeys the constraint of a low-variance distribution of cell values. This proposal uses a different functional enrichment tool called CLEAN [8], which uses GO as one of their components.

The work in [17] was focused on finding Temporal Dependency Association Rules, which relate patterns of behavior among genes. The rules obtained are to represent regulated relations among genes. They also validated their triclusters based on their graphical representation and GO results.

Summarizing, we see that the standard for validation of triclusters quality is based on their graphical representation and GO functional annotations.

3 Methodology

This section describes a novel methodology to evaluate the performance of triclustering algorithms on gene expression data. The goal is to introduce a single measure that globally assesses the quality of the triclusters generated by any triclustering algorithm. The new measure is called TRIcluster Quality ($TRIQ$).

This index takes into account three key aspects to assess the quality of triclusters obtained from gene expression data:

1. The level of biological notoriety of the clustered genes.
2. The graphic quality of the patterns that forms the tricluster.
3. The level of correlation of the tricluster's values.

Biological notoriety and graphical quality are standards in literature for tricluster quality assessment as seen in Sect. 2. Correlation has been included since it provides statistical information on the relation of dependence among the components of the triclusters (gene, condition and time) [4,13].

These three aspects are reflected within the framework $TRIQ$ in its four terms of the general equation: Biological Quality ($BIOQ$) or biological quality of triclusters, Graphical Quality (GRQ) or graphical quality of triclusters, Pearson Quality (PEQ) or value for the Pearson correlation of triclusters and Spearman Quality (SPQ) or values for the Spearman correlation.

The influence of each term is reflected in Eq. 1, where $TRIQ$ is defined as the weighted sum of each of the four aforementioned terms. Therefore, four associated weights must be defined: the weight for ($BIOQ$), denoted as W_{bio};

the weight for (GRQ), denoted as W_{gr}; the weight for (PEQ), denoted as W_{pe}; and the weight for (SPQ), denoted as W_{sp}.

$$\begin{aligned} TRIQ(TRI) = & \frac{1}{W_{bio} + W_{gr} + W_{pe} + W_{sp}} * \\ & [W_{bio} * BIOQ(TRI) + W_{gr} * GRQ(TRI) + \\ & W_{pe} * PEQ(TRI) + W_{sp} * SPQ(TRI)] \end{aligned} \tag{1}$$

3.1 *BIOQ*

The Gene Ontology Project (GO) [1] is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The biological quality of a tricluster is calculated based on the GO analysis that identifies, for a set of genes in a tricluster, the terms listed in each of the three available ontologies: biological processes, cellular components and molecular functions.

The GO analysis used in the calculation of $BIOQ$ is done with the software Ontologizer [3]. This analysis, besides identifying the annotated terms, performs the statistical analysis for the over-representation of those terms, also providing their p-value. However, it is also important to take into account how deep in the ontology the terms are annotated, with the deeper terms being more specific than the superficial ones [18]. In order to represent that information in the $BIOQ$ measure, the biological quality of a tricluster TRI is defined in Eq. 2 as the normalization of the biological significance, SIG_{bio}, of the set of genes TRI_G:

$$BIOQ(TRI) = \frac{SIG_{bio}(TRI_G)}{S_{l_{|LV|}}} \tag{2}$$

where $S_{l_{|LV|}}$ denotes the value of maximum possible score for the total existing levels set LV, as established in Table 1.

The scoring system, in which the index is based, must be designed prior to the definition of SIG_{bio}. An interval for a given level $Inter_l$ is defined by a weight value w_l for the level, and by the lower and upper bounds (inf_l and sup_l, respectively), being an open-closed p-values interval (Eq. 3a). The set of existing LV consists of all levels with Inf_l smaller or equal to a minimum p-value, th. For each interval of each level $Inter_l$, the weight value w_l is the value of the previous level plus the level difference factor d (Eq. 3c.); Inf_l is defined as the division between the pitch factor interval s and the base interval b raised to the power of l factor (Eq. 3d), and sup_l is set to the division between s and b raised to the power of l minus 1 (Eq. 3e).

$$inter_l = \langle w_l,\ (inf_l, sup_l] \rangle \tag{3a}$$

$$LV = \forall\, l \in \mathbb{N} : inf_l \leq th \tag{3b}$$

$$w_l = [(l-1) * d] + 1 \tag{3c}$$

$$inf_l = \frac{s}{b^l} \tag{3d}$$

$$sup_l = \frac{s}{b^{(l-1)}} \tag{3e}$$

All biological significance intervals for the configuration detailed in Eq. 4 are shown in Table 1. For each row, weight (w_l) and range ($inter_l$) for each level (L) sorted in ascending order are shown. Each interval provides a set of p-values where their significance is directly related to the corresponding level, that is, a p-value is better the higher the level to which it belongs is (a p-value is better the closer to zero it is).

$$\begin{aligned} th &= 1.0 \times 10^{-40} \\ d &= 10.0 \\ b &= 10.0 \\ s &= 1.0 \\ LV &= \{1, \ldots, 41\} \end{aligned} \tag{4}$$

The biological significance for all genes in TRI_G is defined as the sum of the score for each level in the GO analysis (Eq. 5c), by taking into account all levels of l and predefined intervals $inter_l$.

S_l score for each level is defined as the multiplication of the concentration level (C_l) by the weight w_l and the level l, plus a function of maximum bonus dependent with the maximum level l_{max} found for all analyzed genes TRI_G (Eq. 5c). C_l is defined as the number of terms located on this level Te_l divided by the total of terms, Te, from the GO analysis results (Eq. 5c).

The bonus feature f_{bonus} is defined as the sum of the peak reached by TRI_G plus the bonus factor V_{bonus} (Eq. 5d).

$$SIG_{bio}(TRI_G) = \sum_{l \in LV} S_l \tag{5a}$$

$$S_l = [C_l * w_l * l] + f_{bonus}(l_{max}) \tag{5b}$$

$$C_l = \frac{Te_l}{Te} \tag{5c}$$

$$f_{bonus}(l_{max}) = l_{max} + V_{bonus} \tag{5d}$$

Table 1. Biological significance.

Level (l)	Weight (w_l)	Interval ($inter_l$)
41	401	(0.0E-00,1.0E-40]
40	391	(1.0E-40,1.0E-39]
39	381	(1.0E-39,1.0E-38]
38	371	(1.0E-38,1.0E-37]
37	361	(1.0E-37,1.0E-36]
36	351	(1.0E-36,1.0E-35]
35	341	(1.0E-35,1.0E-34]
34	331	(1.0E-34,1.0E-33]
33	321	(1.0E-33,1.0E-32]
32	311	(1.0E-32,1.0E-31]
31	301	(1.0E-31,1.0E-30]
30	291	(1.0E-30,1.0E-29]
29	281	(1.0E-29,1.0E-28]
28	271	(1.0E-28,1.0E-27]
27	261	(1.0E-27,1.0E-26]
26	251	(1.0E-26,1.0E-25]
25	241	(1.0E-25,1.0E-24]
24	231	(1.0E-24,1.0E-23]
23	221	(1.0E-23,1.0E-22]
22	211	(1.0E-22,1.0E-21]
21	201	(1.0E-21,1.0E-20]
20	191	(1.0E-20,1.0E-19]
19	181	(1.0E-19,1.0E-18]
18	171	(1.0E-18,1.0E-17]
17	161	(1.0E-17,1.0E-16]
16	151	(1.0E-16,1.0E-15]
15	141	(1.0E-15,1.0E-14]
14	131	(1.0E-14,1.0E-13]
13	121	(1.0E-13,1.0E-12]
12	111	(1.0E-12,1.0E-11]
11	101	(1.0E-11,1.0E-10]
10	91	(1.0E-10,1.0E-09]
9	81	(1.0E-09,1.0E-08]
8	71	(1.0E-08,1.0E-07]
7	61	(1.0E-07,1.0E-06]
6	51	(1.0E-06,1.0E-05]
5	41	(1.0E-05,1.0E-04]
4	31	(1.0E-04,1.0E-03]
3	21	(1.0E-03,1.0E-02]
2	11	(1.0E-02,1.0E-01]
1	1	(1.0E-01,1.0E-00]

3.2 *GRQ*

The graphic quality of a tricluster is a quantitative representation of a qualitative measure: how homogeneous the members of the tricluster are. This is widely used in literature for tricluster visual validation by means of graphically representing the triclusters on their three components: genes, conditions and time points [17,22,23].

We have quantified this information using the MSL measure [10], since it provides a numerical value of the similarity among the angles of the slopes formed by each profile for genes, conditions and times (see Fig. 1).

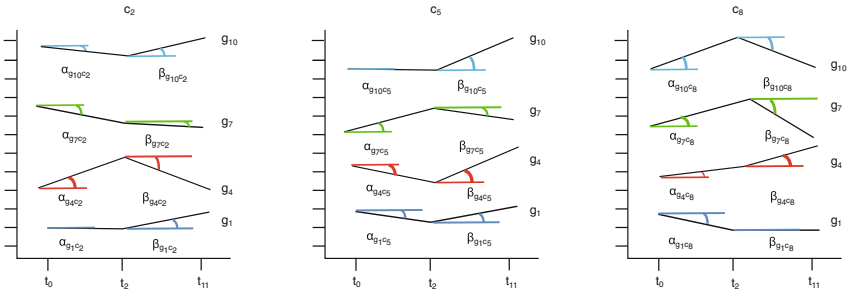

Fig. 1. Representation of how the MSL measure is calculated.

We define GRQ in Eq. 6 as one minus the normalization of MSL. Thus, a tricluster is graphically better the smaller the value of MSL is [10].

$$GRQ(TRI) = 1 - \frac{MSL(TRI)}{2\pi} \tag{6}$$

3.3 *PEQ* and *SPQ*

Pearson [4] and Spearman [13] correlations have been chosen since they are the standard correlations measures and they are widely used in literature [15].

Random variables for a tricluster TRI are defined to calculate PEQ and SPQ, based on its subset of genes (Eq. 7a), conditions (Eq. 7b) and time stamps (Eq. 7c). Thus, every tricluster will have a set of random variables $vars$ composed of the combination of each gene and each experimental condition (Eq. 7d). Each of these variables will have a expression level for each time stamp (Eq. 7e).

For example, for a tricluster consisting of four genes g_1, g_4, g_8, g_{10}, two conditions c_3, c_7 and three time points t_1, t_3, t_5, random variables for eight possible combinations will be considered, each having three values (one per time stamp): $V_{g_1c_3}$, $V_{g_1c_7}$, $V_{g_4c_3}$, $V_{g_4c_7}$, $V_{g_8c_3}$, V_{g_8,c_7}, $V_{g_{10}c_3}$ and $V_{g_1c_7}$.

$$TRI_G =< g_0, g_1, \ldots, g_{|G|} > \tag{7a}$$

$$TRI_C =< c_0, c_1, \ldots, c_{|C|} > \tag{7b}$$

$$TRI_T =< t_0, t_1, \ldots, t_{|T|} > \tag{7c}$$

$$\forall g_i \in TRI_G, c_j \in TRI_C \; vars = \{V_{g_0 c_0}, V_{g_1 c_1}, \ldots, V_{g_{|G|} c_{|C|}}\} \tag{7d}$$

$$V_{g_i c_j} =< el_{g_i c_j t_0}, el_{g_i c_j t_1}, \ldots, el_{g_i c_j t_{|T|}} > \forall g_i \in TRI_G, c_j \in TRI_C, t_k \in TRI_T \tag{7e}$$

Given the set of variables $vars$, PEQ is defined as the sum of the absolute value of the Pearson correlation coefficient for each combination of each pair of variables in the set $vars$ divided by the number of such combinations (Eq. 8).

$$PEQ(TRI) = \frac{\sum_{V_{g_i c_j}, V_{g_k c_l} \in vars} |PE(V_{g_i c_j}, V_{g_k c_l})|}{\left[\frac{(|G||C|)^2 - |G||C|}{2}\right]} \tag{8}$$

Similarly, SPQ is defined as the sum of the absolute value of the Spearman correlation coefficient for each combination of each pair of variables in the set $vars$ divided by the number of such combinations (Eq. 9).

$$SPQ(TRI) = \frac{\sum_{V_{g_i c_j}, V_{g_k c_l} \in vars} |SP(V_{g_i c_j}, V_{g_k c_l})|}{\left[\frac{(|G||C|)^2 - |G||C|}{2}\right]} \tag{9}$$

4 Results

In this section we show the results obtained by application of $TRIQ$ to the tricluster solutions obtained using the three fitness function MSR_{3D}, LSL and MSL embedded in the $TriGen$ algorithm. Three different datasets have been used: the yeast cell cycle (*Saccharomyces Cerevisiae*) [21], in particular the *elutriation* experiment, an experiment with mice (*Mus Musculus*) called $GDS4510$ [7] and data from an experiments with humans (*Homo Sapiens*) called $GDS4472$ [5]. The last two datasets have been retrieved from Gene Expression Omnibus [2], a repository of high throughput gene expression data. All experiments examine the behavior of genes under conditions at certain times.

We show, for each of the datasets, the average values obtained upon execution of TriGen with each of the fitness functions (MSR, LSL and MSL) for $TRIQ$ in the triclusters obtained, as well as the individual values obtained for each of the measures involved in $TRIQ$: $BIOQ$, GRQ, PEQ and SPQ.

The weights for calculation of $TRIQ$ (see Eq. 1) have been set to the values shown in Eq. 10. Triclusters with greater biological and graphical quality are

preferred, since Pearson and Spearman correlations are not typically critical, even if the information contained on them is greatly valuable.

$$\begin{aligned} W_{bio} &= 0.5 \\ W_{gr} &= 0.4 \\ W_{pe} &= 0.05 \\ W_{sp} &= 0.05 \end{aligned} \tag{10}$$

The value for the bonus parameter for the $BIOQ$ measure V_{bonus} is set to the value shown in Eq. 11:

$$\begin{aligned} V_{bonus} &= 0 \\ S_{l_{|LV|}} &= [C_{l_{|LV|}} * w_{l_{|LV|}} * l_{|LV|}] + [l_{|LV|} * V_{bonus}] \\ &= [1 * 401 * 41] + [41 + 0] = 16482 \end{aligned} \tag{11}$$

4.1 Yeast Cell Cycle Dataset

We have applied the TriGen algorithm to the yeast (*Saccharomyces Cerevisiae*) cell cycle problem [21]. The yeast cell cycle analysis project's goal is to identify all genes whose mRNA levels are regulated by the cell cycle. The resources used are public and available in http://genome-www.stanford.edu/cellcycle/. Here we can find information relative to gene expression values obtained from different experiments using microarrays. In particular, we have created a dataset $Delu_{3D}$ from the elutriation experiment with 7744 genes, 13 experimental conditions and 14 time points. Experimental conditions correspond to different statistical measures of the Cy3 and Cy5 channels while time points represent different moments of taking measures from 0 to 390 min.

Table 2. $TRIQ$ results for the Yeast Cell Cycle experiment.

Fitness function	$TRIQ$	$BIOQ$	GRQ	PEQ	SPQ
MSR_3D	0,2905	0,0001	0,5790	0,5901	0,5870
LSL	0,4530	0,0001	0,9181	0,8655	0,8479
MSL	0,4947	0,0001	0,9995	0,9400	0,9571

We see that MSL obtains the best values for $TRIQ$ in this experiment, as well as for GRQ, PEQ, and SPQ. LSL has the second better set of results, while the values for $BIOQ$ are very similar for the three fitness functions (Table 2).

4.2 Mouse GDS4510 Dataset

This dataset was obtained from GEO [2] with accession code GDS4510 and title *rd1 model of retinal degeneration: time course* [7]. In this experiment the degeneration of retinal cells in different individuals of home mouse (*Mus musculus*)

is analyzed over 4 days just after birth, specifically on days 2, 4, 6 and 8. Our input dataset $DGDS4510_{3D}$ is composed of 22690 genes, 8 experimental conditions (one for each individual involved in the biological experiment) and 4 time points.

Table 3. $TRIQ$ results for the Mouse GDS4510 experiment.

Fitness function	$TRIQ$	$BIOQ$	GRQ	PEQ	SPQ
MSR_3D	0,3601	0,0003	0,7412	0,6400	0,6289
LSL	0,4117	0,0006	0,8486	0,7094	0,7298
MSL	0,4693	0,0013	0,9767	0,7807	0,7780

We see that, again, MSL obtains the best values for $TRIQ$, as well as for all separate measures involved. LSL has the second position (Table 3).

4.3 Human GDS4472 Dataset

This dataset has been obtained from GEO [2] under code GDS4472 titled *Transcription factor oncogene OTX2 silencing effect on D425 medulloblastoma cell line: time course* [5]. In this experiment we analyze the effect of doxycycline on medulloblastoma cancerous cells at six times after induction: 0, 8, 16, 24, 48 and 96 h. Our input dataset $DGSD4472_{3D}$ is composed by 54675 genes, 4 conditions (one for each individual involved) and 6 time points (one per hour).

Table 4. $TRIQ$ results for the Human GDS4472 experiment.

Fitness function	$TRIQ$	$BIOQ$	GRQ	PEQ	SPQ
MSR_3D	0,3128	0,0001	0,6215	0,6334	0,6490
LSL	0,4111	0,0050	0,8326	0,7779	0,7344
MSL	0,4422	0,0048	0,8983	0,7905	0,8182

We see how, for this dasatet also, MSL obtains the best values for $TRIQ$, as well as for GRQ, PEQ, and SPQ. MSL and LSL have very similar values for $BIOQ$, both improving MSR. LSL has the second better set of results (Table 4).

In a previous publication [10] we stated that MSL was the fitness function capable of extracting best tricluster solutions in terms of the three validation measures considered: correlation, graphic validation and functional annotations. We see how $TRIQ$ has successfully represented the three validation measures yielding the same validation results as in [10].

5 Conclusions

In this work we have presented a tricluster validation measure, $TRIQ$ capable to combine information from three different sources: correlation among the genes, conditions and times, graphic validation of the patterns extracted and functional annotations for the genes extracted.

We have applied $TRIQ$ to the triclusters obtained with three fitness functions previously defined by the authors: MSR_{3D}, LSL and MSL. The datasets used are the yeast cell cycle (*Saccharomyces Cerevisiae*), in particular the *elutriation* experiment, an experiment with mice (*Mus Musculus*) called $GDS4510$ and data from an experiments with humans (*Homo Sapiens*) called $GDS4472$.

We have shown that $TRIQ$ has successfully represented the three validation measures yielding the same validation results as in [10] where each of the components of $TRIQ$ ($BIOQ$, GRQ, PEQ, and SPQ) where applied separately.

Acknowledgments. The authors thank financial support by the Spanish Ministry of Science and Technology, projects TIN2011-28956-C02-02 and TIN2014-55894-C2-1-R and Junta de Andalucía's project P12-TIC-7528.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1), 25–29 (2000)
2. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A.: NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. **41**, D991–D995 (2013)
3. Bauer, S., Grossmann, S., Vingron, M., Robinson, P.N.: Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics **24**(14), 1650–1651 (2008)
4. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise Reduction in Speech Processing, pp. 1–4 (2009)
5. Bunt, J., Hasselt, N.E., Zwijnenburg, D.A., Hamdi, M., Koster, J., Versteeg, R., Kool, M.: OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells. Int. J. Cancer **131**(2), E21–E32 (2012)
6. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology, pp. 93–103 (2000)
7. Dickison, V.M., Richmond, A.M., Abu-Irqeba, A., Martak, J.G., Hoge, S.C.E., Brooks, M.J., Othman, M.I., Khanna, R., Mears, A.J., Chowdhury, A.Y., Swaroop, A., Ogilvie, J.M.: A role for prenylated rab acceptor 1 in vertebrate photoreceptor development. BMC Neurosci. **13**(1), ID152 (2012)
8. Freudenberg, J.M., Joshi, V.K., Hu, Z., Medvedovic, M.: Clean: clustering enrichment analysis. BMC Bioinformatics **10**(1), 234 (2009)

9. Gutiérrez-Avilés, D., Rubio-Escudero, C.: Mining 3D patterns from gene expression temporal data: a new tricluster evaluation measure. Sci. World J. **1–16**, 2014 (2014)
10. Gutiérrez-Avilés, D., Rubio-Escudero, C.: MSL: a measure to evaluate three-dimensional patterns in gene expression data. Evol. Bioinform. **11**, 121–135 (2015)
11. Gutiérrez-Avilés, D., Rubio-Escudero, C., Martínez-Álvarez, F., Riquelme, J.C.: TriGen: a genetic algorithm to mine triclusters in temporal gene expression data. Neurocomputing **132**, 42–53 (2014)
12. Gutiérrez-Avilés, D., Rubio-escudero, C.: L.S.L: a new measure to evaluate triclusters. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 30–37 (2014)
13. Hauke, J., Kossowski, T.: Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae **30**(2), 87–93 (2011)
14. Hu, Z., Bhatnagar, R.: Algorithm for discovering low-variance 3-clusters from real-valued datasets. In: IEEE International Conference on Data Mining, pp. 236–245 (2010)
15. Jiang, H., Zhou, S., Guan, J., Zheng, Y.: gTRICLUSTER: a more general and effective 3D clustering algorithm for gene-sample-time microarray data. In: Li, J., Yang, Q., Tan, A.-H. (eds.) BioDM 2006. LNCS (LNBI), vol. 3916, pp. 48–59. Springer, Heidelberg (2006)
16. Liu, J., Li, Z., Hu, X., Chen, Y.: Multi-objective evolutionary algorithm for mining 3D clusters in gene-sample-time microarray data. In: 2008 IEEE International Conference on Granular Computing, No. 60573057, pp. 442–447. IEEE, August 2008
17. Liu, Y.-C., Lee, C.-H., Chen, W.-C., Shin, J.W., Hsu, H.-H., Tseng, V.S.: A novel method for mining temporally dependent association rules in three-dimensional microarray datasets. In: 2010 International Computer Symposium (ICS), pp. 759–764. IEEE (2010)
18. Romero-Zaliz, R.C., Rubio-Escudero, C., Cobb, J.P., Herrera, F., Cordón, O., Zwir, I.: A multiobjective evolutionary conceptual clustering methodology for gene annotation within structural databases: a case of study on the gene ontology database. IEEE Trans. Evol. Comput. **12**(6), 679–701 (2008)
19. Rubio-Escudero, C., Martínez-Álvarez, F., Romero-Zaliz, R.C., Zwir, I.: Classification of gene expression profiles: comparison of k-means and expectation maximization algorithms. In: Proceedings of IEEE International Conference on Hybrid Intelligent Systems, pp. 831–836 (2008)
20. Spearman, C.: Correlation calculated from faulty data. Br. J. Psychol. 1904–1920 **3**(3), 271–295 (1910)
21. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell **9**(12), 3273–3297 (1998)
22. Xu, X., Lu, Y., Tan, K.L., Tung, A.K.H.: Finding time-lagged 3D clusters. In: Proceedings of the IEEE International Conference on Data Engineering, pp. 445–456 (2009)
23. Zhao, L., Zaki, M.J.: triCluster: an effective algorithm for mining coherent clusters in 3D microarray data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 694–705 (2005)