COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 82 (2006) 44–50

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

Distribution-insensitive cluster analysis in SAS on real-time PCR gene expression data of steadily expressed genes

Aleš Tichopád[a,b], *Ladislav Pecen*[a], *Michael W. Pfaffl*[b,*]

[a] *IMFORM GmbH, International Clinical Research, Darmstadt, Germany*
[b] *Lehrstuhl für Physiologie, Fakultät Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising-Weihenstephan, Germany*

ARTICLE INFO

Article history:
Received 12 July 2005
Received in revised form 4 October 2005
Accepted 12 December 2005

Keywords:
Real-time PCR
Cluster analysis
Standardisation
Housekeeping genes
Expression pattern
Spearman coefficient

ABSTRACT

Cluster analysis is a tool often employed in the micro-array techniques but used less in the real-time PCR. Herein we present core SAS code that instead of the Euclidian distances takes correlation coefficient as a dissimilarity measure. The dissimilarity measure is made robust using a rank–order correlation coefficient rather than a parametric one. There is no need for an overall probability adjustment like in scoring methods based on repeated pair-wise comparisons. The rank–order correlation matrix gives a good base for the clustering procedure of gene expression data obtained by real-time RT-PCR as it disregards the different expression levels. Associated with each cluster is a linear combination of the variables in the cluster, which is the first principal component. Large set of variables can then be replaced by the set of cluster components with little loss of information. In this way, distinct clusters containing unregulated housekeeping genes along with other steadily expressed genes can be disclosed and utilized for standardization purposes. Simulated data in parallel with the data from a biological experiment were taken to validate the SAS macro. For both cases, good intuitive results were obtained.

1. Introduction

Quantitative real-time polymerase chain reaction (PCR) is a powerful method for reliable quantification of low-abundance mRNA in biological samples. The mRNA must be reverse transcribed into DNA as the enzyme used in PCR reaction is restricted to DNA only. Real-time PCR is based on enzymatic amplification of selected initial DNA sequence up to a detectable amount by repeated temperature changes. The entire development of the DNA amount in sample is closely monitored by fluorescence signal emission at the end of every cycle. Since the strength of the fluorescence signal emitted is proportional to the DNA amount produced, it facilitates visualisation of the reaction trajectory and produces so called amplification curve. Initial concentration of DNA in sample can be then estimated from the calculated number of cycles required to attain chosen signal strength or some strictly defined point on the amplification curve. This number called *crossing point* (CP) or *threshold cycle* (Ct) is the fundamental quantity obtained for each sample from the real-time PCR assay.

Since the CP reflects only an unknown amount of initial sequence, useful quantitative information cannot be obtained unless at least two samples are analysed. If at least two unknown samples are quantified, the difference between them can be obtained from the difference between their CPs and the amplification mode. In this way, effect of experimental treatment on up- or down-regulation of gene expression can be studied without the need for absolute quantities known.

* *Corresponding author*. Present address: Physiology – Weihenstephan, Center of Life and Food Science, Technische Universität München, Weihenstephan Berg 3, 85354 Freising-Weihenstephan, Germany. Tel.: + 49 8161 713511; fax: +49 8161 714204.
E-mail address: michael.pfaffl@wzw.tum.de (M.W. Pfaffl).

0169-2607/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2005.12.002

This is a principal of the so called relative quantification approach.

To make sure that the entire quantification assay proceeds in a similar way in all compared samples, amplification of another sequence assumed unregulated under applied study treatment is introduced into the assay. The search for such reference genes unregulated under treatment is therefore an essential task before any relative gene expression quantification is conducted. The reference gene is assumed to remain constant under applied treatment, and any shift in its amount observed between compared samples can be thus associated to assay disturbances. The same shift can be then expected to affect the results of the studied gene and can be subtracted from it. Computing method based on this principle was proposed [1] that incorporates a reference gene into calculation of expression ratio of studied gene in two compared samples.

There is a good reason and even bigger wish to believe, that so called housekeeping genes remain constantly expressed and provide thus a good references. Unfortunately, papers reporting about regulation of these genes are published too often to admit of this assumption (e.g., [2,3]). Physiological changes in untreated organisms can alone cause regulation of these genes [4]. Vandesompele et al. [5] proposed computing method based on the standard deviation that sorts candidates according to their best pair-wise score to other genes. This method, however, does not reflect the target genes and their relation to the reference genes. Repeated pair-wise analysis on more than two genes [6] is confronted with the need for an adjustment of the overall probability value. The Excel tool by Pfaffl et al. [6] also utilises the Pearson correlation coefficient as a similarity measure, assuming thus normal distribution along the data. This is however only seldom the case. Herein presented method is proposed with the aim to up-grade the tool presented by Pfaffl et al. [6].

2. Background

Some simple approach omitting the imaginary boundary between unregulated housekeeping genes and regulated genes is desired, that would group genes based on a robust distribution-insensitive dissimilarity measure. Spearman rank–order correlation coefficient is a non-parametric measure of association based on the rank of the data values. The formula is

$$\theta = \frac{\sum (R_i - R)(S_i - S)}{\sqrt{\sum (R_i - \overline{R})^2 \sum (S_i - \overline{S})^2}} \quad (1)$$

where R_i is the rank of the ith x value, S_i the rank of the ith y value, $\overline{R}$ the mean of the R_i values and $\overline{S}$ is the mean of the S_i values.

Clustering procedure based on the Spearman correlation coefficient prevents erroneous results due to non-normal distributed real-time PCR data [7]. In the here proposed method, associated with each cluster is a linear combination of the variables in the cluster, which is the first principal component. A large set of variables can often be replaced by the set of cluster components with little loss of information. A given number of cluster components does not generally explain as much variance as the same number of principal components on the full set of variables, but the cluster components are usually easier to interpret than the principal components. The first principal component is a weighted average of the variables that explains as much variance as possible.

The purpose of looking for *principal components* [8,9] is to derive a small number of linear combinations (principal components) of a set of variables that retain as much of the information in the original variables as possible. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering, and so on. Principal component analysis can also be viewed as an attempt to uncover approximate linear dependencies among variables.

The first j principal components provide a least-squares solution to the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (2)$$

where $\mathbf{Y}$ is an $n \times p$ matrix (with n columns and p rows) of the centered observed variables; $\mathbf{X}$ the $n \times j$ matrix of scores on the first j principal components; $\mathbf{B}$ the $j \times p$ matrix of eigenvectors; $\mathbf{E}$ the $n \times p$ matrix of residuals; and the trace ($\mathbf{E}'\mathbf{E}$), the sum of all the squared elements in $\mathbf{E}$, is to be minimized. In other words, the first j principal components are the best linear predictors of the original variables among all possible sets of j variables, although any nonsingular linear transformation of the first j principal components would provide equally good prediction.

3. Test data

3.1. *Simulated data set*

Data with five variables of $n = 11$ observations was simulated based on a real biological variable by introducing noise into four linearly associated variables. This data should facilitate a better intuitive understanding of the logic behind the dissimilarity measure applied for clustering. The first variable called RG contains 11 CP observations for ubiquitin with the arithmetic mean μ_{RG} and standard deviation σ^2_{RG}.

$$RG \sim U(\mu_{RG}, \sigma^2_{RG}) \quad (3)$$

Since $n = 11$ only, no strong departure from normal distribution was found. Subsequently, three semi-random simulated variables SRD_1, SRD_2 and SRD_3 were derived from RG as follows:

$$\begin{aligned} SRD_1 &= RG + \eta_1 & \eta_1 &\sim N(\mu_{\eta_1} = 0,\ \sigma^1_{\eta_1} = 1) \\ SRD_2 &= 1.3 * RG + \eta_2 & \eta_2 &\sim N(\mu_{\eta_2} = 0,\ \sigma^1_{\eta_2} = 0.1) \\ SRD_3 &= \eta_3 * RG & \eta_3 &\sim N(\mu_{\eta_3} = 1.3,\ \sigma^1_{\eta_3} = 0.03) \end{aligned} \quad (4)$$

where RG is the real gene's CP and η_1, η_2 denote random noise generated from the normal distribution with respective mean μ and standard deviation σ^2. These three variables were supposed to simulate different genes with the same expression pattern as the RG but more or less affected by sample specific disturbances. The η_1 and η_2 introduce some additive distur-

Table 1 – Descriptive statistics on the simulated data ($n = 11$)

Variable	Mean	Variance	S.D.
RG	20.44	0.38	0.61
SRD1	20.86	1.60	1.27
SRD2	26.57	0.64	0.80
SRD3	26.70	1.70	1.32
RD	20.71	1.71	0.08

bance into, otherwise perfect, correlations between SRD_1 and RG, and SRD_2 and RG. The SRD_1 and SRD_2 differ one from another not only by their multiplicative factor but also by the standard deviation of their added noise component $\sigma^2_{\eta_1}, \sigma^2_{\eta_2}$. As the noise is greater in SRD_1 then in SRD_2, which is given by its $\sigma^2_{\eta_1} > \sigma^2_{\eta_2}$, it is supposed, that RG should be more tightly correlated with SRD_2 in most of the simulation runs. As no multiplication of RD is introduced into SRD_1, it has the same central tendency as RG and from the Euclidian perspective is thus closer to RG than the SRD_2 with its multiplicative factor 1.3.

The SRD_3 is also derived from the RG, however the noise is introduced as a multiplicative rather than additive factor here. The noise has the same mean value 1.3 as the multiplicative factor in SRD_2, and small standard deviation 0.03. The aim of introducing SRD_3 is to simulate a gene regulation dependant on a baseline value, common phenomenon in biological regulation. In other words, expression change due to treatment is not only dependant on the treatment strength but also on the baseline expression value before the treatment was applied.

The fifth simulated variable RD has no linear association with RG as it is drown from the normal distribution with the same mean and standard deviation as the RG, therefore $RD \sim N(\mu_{RG}, \sigma^2_{RG})$. It is expected that in most simulations this variable should not be associated with the RG however, from Euclidian perspective, both variables are close. Descriptive statistics for the simulated data are shown in Table 1.

3.2. *Biological data set*

The same biological data as used in Pfaffl et al. [6] were used here also to show that similar results can be obtained with herein presented method, assuring however better adherence to a good statistical practice. Complete RNA from 31 bovine corpora lutea samples was extracted by method of Chomczynski [10] from small slices of deep frozen CL with peqGOLD according to the manufacturer's instruction.

The cDNA was reverse-transcribed from 1000 ng total RNA with 200 units of M-MLV Reverse Transcriptase (Promega Corp., Madison, USA) according to the manufacturer instructions.

The CP data related to expression levels [12] of studied factors were obtained on LightCycler (Roche, Basel, Switzerland) PCR instrument [11,12]. In the 31 cDNA samples, expression of four genes with assumed stable expression – housekeeping genes (HKG); ubiquitin (UBQ), glyceraldehyde-3-phosphate dehydrogenase (GAPD), β-actin and 18S ribosomal unit were quantified together with 10 studied target genes; IGF-1 (insulin-like growth factors type 1), IGF-2, IGFR-1 (insulin-like growth factor receptor type 1), IGFR-2, IGFBP-1 (insulin-like growth factor binding protein type 1) – IGF-6, whose expression is studied. In each biological sample, all 14 factors were quantified. The CP as used here is a fractional numbers of PCR cycles necessary to reach a strictly given point of the amplification curve geometry. Here, the maximum of the second derivative of the amplification curve [12] was used as computed by the LightCycler Software 3.5 (Roche Diagnostics). Descriptive statistics for this data are shown in Table 2.

Table 2 – Descriptive statistics on the biological data ($n = 31$)

Variable	Mean	Variance	S.D.
UBQ	20.86	1.05	1.03
GAPD	21.50	1.05	1.02
β-Actin	18.29	1.04	1.02
S18	12.97	3.66	1.91
IGF-1	29.31	1.00	1.00
IGF-2	23.14	1.17	1.08
IGF-1R	24.59	1.122	1.06
IGF-2R	37.89	0.70	0.84
BP-1	29.38	9.10	3.02
BP-2	30.53	1.042	1.02
BP-3	30.00	3.39	1.84
BP-4	31.13	2.18	1.48
BP-5	26.74	2.24	1.50
BP-6	30.36	2.12	1.46

4. System description

The SAS macro CLUSTER presented here (Table 3) is the minimal source that performs the fundamental computing. Data *genes* entered by the DATA step is a reduced example of the original biological dataset. The macro consists of following SAS/BASE and SAS/STAT procedures.

The CORR procedure is a statistical procedure that creates correlation matrix with the Spearman correlation coefficients. The correlation matrix is then saved as an output data set *cor*.

The VARCLUS procedure uses the recently created *cor* data set and omits observations with missing values from the analysis. The MAXCLUSTERS= option specifies the largest number of clusters desired. This can be determined in the macro invocation as &*clusno* parameter.

The VARCLUS procedure tries to maximize the sum across clusters of the variance of the original variables that is explained by the cluster components. The set of analysed genes is divided into non-overlapping clusters in so that each cluster can be interpreted as essentially unidimensional. For each cluster, PROC VARCLUS computes a component that is the first principal component and tries to maximize the sum across clusters of the variation accounted for by the cluster components. PROC VARCLUS is a type of oblique component analysis related to multiple group factor analysis [13]. By default, PROC VARCLUS begins with all genes in a single cluster. It then repeats the following steps:

(1) A cluster is chosen for splitting.
(2) The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvectors), and

Table 3 – The SAS macro CLUSTER

```
*Program source of the macro CLUSTER;
%macro CLUSTER (var=, clus=);
   PROC CORR outs=cor;
var &VAR;
  PROC VARCLUS data=cor outtree=tree
%if &clus. ne %str() %then
   %do; maxclusters=&clus; %end;
   ;
var &VAR;
   axis2 minor=none;
   axis1 label=('Proportion of Variation Explained')
    minor=none;
   PROC TREE horizontal vaxis=axis2 haxis=axis1
    lines=(width=2);
height _propor_;
   run;
%mend cluster;
*Entry of CP data into SAS by data step;
DATA genes;
   input UBQ GAPD Betaactin S18 IGF1;
   cards;
20.59 21.06 17.80 10.00 28.49
21.17 20.84 18.14 12.92 29.05
20.67 20.09 17.84  9.87 29.69
20.99 20.78 18.30 10.15 28.74
19.77 19.65 16.71 11.66 29.03
19.91 21.33 17.22 10.37 27.59
20.75 21.74 17.58 10.05 28.97
21.08 21.25 17.16 13.03 28.51
19.22 21.24 17.44 12.58 28.87
;
run;
*Call of the Macro;
%CLUSTER (var=UBQ GAPD Betaactin S18 IGF1, clus=);
```

Minimal SAS macro compiler source necessary for performing the described computation. The macro call bellow enables definition of analysed variables as well as optional definition of number of cluster produced.

assigning each gene to the rotated component with which it has the higher squared correlation.

(3) Genes are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components. The reassignment may be required to maintain a hierarchical structure.

If the MAXCLUSTERS is not defined, by default PROC VARCLUS stops splitting when each cluster has only a single eigenvalue greater than 1, thus satisfying the most popular criterion for determining the sufficiency of a single underlying factor dimension Kaiser [14]. The iterative reassignment of genes to clusters proceeds in two phases. The first is a nearest component sorting (NCS) phase, similar in principle to the nearest centroid sorting algorithms described by Anderberg [15]. In each iteration, the cluster components are computed, and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each gene is tested to see if assigning it to a different cluster increases the amount of variance explained. If a gene is reassigned during the search phase, the components of the two clusters involved are recomputed before the next gene is tested. The NCS phase is much faster than the search phase but is more likely to be trapped by a local optimum.

Table 4 – Cluster listing for six clusters computed on the biological data

Cluster	Variable	R-squared with		$1 - R^2$ ratio
		Own Cluster	Next Closest	
Cluster 1	UBQ	0.6362	0.3323	0.5448
	Betaactin	0.7239	0.3441	0.4210
	IGF-2	0.7547	0.5038	0.4944
	BP-4	0.8092	0.7010	0.6381
Cluster 2	BP-2	0.7111	0.0226	0.2956
	BP-6	0.7111	0.1439	0.3374
Cluster 3	IGF-2R	0.6429	0.0174	0.3635
	BP-5	0.6429	0.1200	0.4059
Cluster 4	IGF-1	1.0000	0.0921	0.0000
Cluster 5	BP-1	1.0000	0.0225	0.0000
Cluster 6	GAPD	0.6846	0.6351	0.8645
	S18	0.6217	0.4628	0.7043
	IGF-1R	0.7732	0.3552	0.3517
	BP-3	0.7960	0.4084	0.3449

To each cluster, the R^2-value of each variable with its own cluster and the R^2-value with its closest cluster are displayed. The R^2-value for a variable with the closest cluster should be low if the clusters are well separated. The last column displays the ratio of $(1 - R^2_{own})/(1 - R^2_{closest})$ for each variable. Small values of this ratio indicate good clustering.

The OUTTREE= option creates an output data set to contain information on the tree structure that can be used by the TREE procedure to print a tree diagram. PROC VARCLUS displays a cluster summary and a cluster listing (Table 4). The cluster summary gives the number of variables in each cluster and the variation explained by the cluster component. The proportion of variance explained is obtained by dividing the variance explained by the total variance of variables in the cluster. If the cluster contains two or more genes the second largest eigenvalue of the cluster is also printed. The cluster listing gives the genes in each cluster. Two squared correlations are calculated for each cluster. The column labeled "Own Cluster" gives the squared correlation of the variable with its own cluster component. This value should be higher than the squared correlation with any other cluster unless an iteration limit has been exceeded. The larger the squared correlation is, the better. The column labeled "Next Closest" contains the next highest squared correlation of the gene with a cluster component. This value is low if the clusters are well separated. The column headed "$1 - R^2$ ratio" gives the ratio of one minus the "Own Cluster" R^2 to one minus the "Next Closest" R^2. A small "$1 - R^2$ ratio" indicates a good clustering.

The TREE procedure produces a horizontally oriented tree diagram, using the data set created by the VARCLUS procedure. The AXIS statements create AXIS definitions that specify the characteristics of an axis. From left to right in the diagram, objects and clusters are progressively joined until a single, all-encompassing cluster is formed at the right (or root) of the diagram. Clusters exist at each level of the diagram, and

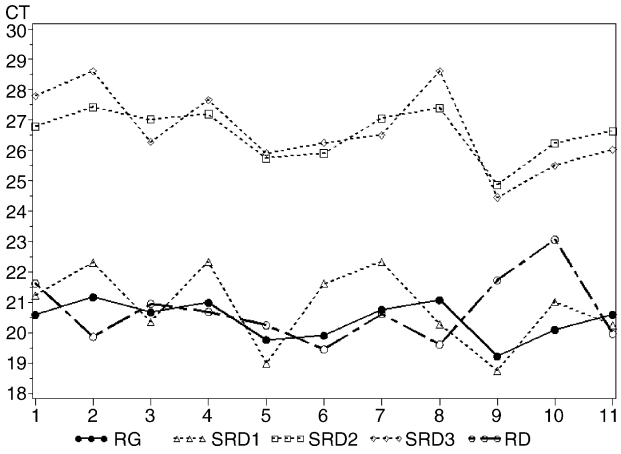

Fig. 1 – Plot of the simulated data vectors. Selected first simulation run.

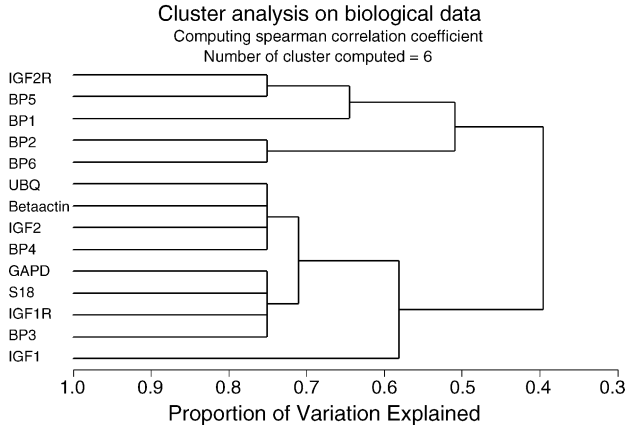

Fig. 3 – Cluster analysis with six clusters computed on the biological data.

every vertical line connects leaves and branches into progressively larger clusters (Figs. 2 and 3). The macro is terminated by the **%mend cluster**. Invocation of the macro consists of the **%cluster** sentence and the definitions of the three macro parameters for the names of genes analyzed (*var*) and number of clusters (*clus*). If definition of the *clus* is omitted and thus the **%if** condition in the program code fulfilled the default setting for number of clusters as described above will be activated.

5. Status report

5.1. *Results from the simulated data*

Twenty simulations were run producing more or less associated variables (Fig. 1). In most simulation the RG was closely associated with SRD_2 followed by SRD_3, whereas the RG was only weekly associated with SRD_1 (Fig. 2). As expected, in only one simulation runs there was a significant ($p < 0.05$) correlation between RG and RD, regardless of the same central tendency and spread of observations. Changing the multiplication factors of SRD_1 and SRD_2 as well as the μ_{σ_3} in SRD_3 had no effect, showing that the method can associate genes

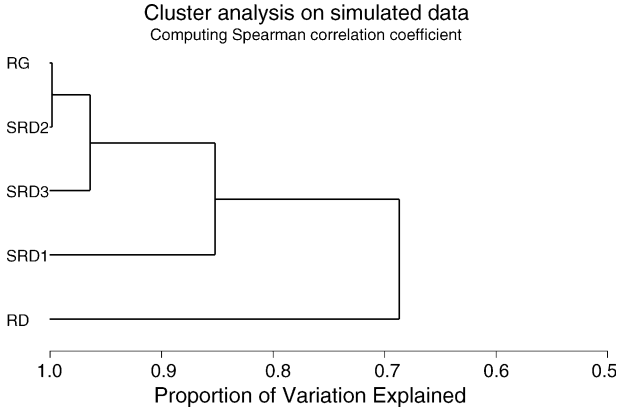

Fig. 2 – Cluster diagram on the simulated data. Similar diagrams were obtained in 14 of the 20 simulation runs.

with different expression levels but with the same patterns. The closest association was found between the RG and the SRD_2 as long as $\sigma^2_{\eta_1} > \sigma^2_{\eta_2}$. The SRD_3 was also tightly associated with the RG, regardless of the noisy multiplication factor η_3. This noise caused the pattern of SRD_3 to be less parallel to RG then the pattern of SRD_2, although $\sigma^2_{\eta_3} < \sigma^2_{\eta_2}$, and thus less associated with it. Such effects are however cushioned by performing the clustering on rank values and not on the parametric data. Therefore, the SRD_3 is still claimed strongly associated with RD here (Fig. 2).

In repeated simulation runs, slightly different results were obtained as the random values were newly generated, nevertheless, the hierarchy of the diagram remained unchanged as long as no extreme noise was added.

5.2. *Results from the biological data*

Taking a look at the diagram, some six discrete clusters, as listed in the first column of Table 4, come to the four (Fig. 3). By default, PROC VARCLUS stops when each cluster had only a single eigenvalue greater than 1, creating six clusters. Clusters 1 and 6 show a great explaining power and both always contain two housekeeping genes each. There are two ways how to deal with this result:

Cluster 1 considered as the best separated, will be taken with all four components for standardisation purposes.
Cluster 6 can also still be considered well separated and useful for the standardisation purposes. Both the cluster 1 and the cluster 2 contain some known ‘conservative’ housekeeping genes.

Further, some deeper insight into the regulation patterns of the target genes can be gained from Fig. 3. Surely, there is no association between the above proposed standards and IGF-2R, BP-5, BP-1, BP-2, BP-6 and IGF-1. These target genes can be well standardised using the genes from clusters 1 and 6.

Here, it is important to realise that the clustering procedure cannot give a clear answer as to what genes are absolutely suited for the standardisation and what not. This is due to fact that the border between still regulated and not regulated at all

genes need not necessarily respect the apriority definition of "conservative" housekeeping genes.

6. Lessons learned

Clustering approaches have been frequently adopted on micro-array data to disclose families of co-regulated genes [16–18]. Similar pattern of expression indicates either co-regulated genes ore genes those remain untouched by the experiment. This similarity is given by the stabile expression ratio between any two of the genes. Provided that sampling, extraction procedure, RT reaction, storage and the PCR were affected by erroneous factors, all genes achieve some common artificial shift. This shift then produces more visible pattern in genes that are not biologically regulated because any biological regulation would otherwise mask it.

The success of a cluster analysis depends on how well its underlying model describes the patterns of expression. Based on the above idea, the herein suggested method associates genes based on similar rank–order correlation patterns as given by the correlation matrix. Genes with different expression levels but correlating well due to steady expression ratio are clustered together. The Euclidean distances cannot be taken as a measure of dissimilarities here because the levels of expression can be different. Such method would associate genes with close CPs instead of genes with constant expression ration in various samples.

The real-time PCR yields so called crossing points or threshold cycles, and these are the fundamental quantitative units [12]. This data shows a skewed distribution and heterogeneous variance. The Gaussian distribution is only rarely given [7], therefore, the here proposed method clusters genes based on the non-parametric Spearman correlation coefficient, making the method distribution-insensitive. In addition, this procedure partially disregards non-linear effects of biological regulation among samples as well as sample specific disturbances as long as these are not too strong. The dissimilarity measure employed is less sensitive but more robust towards violation of proper data distribution. For this reason, more data are necessary than if a parametric measure of dissimilarity was used.

The clustering of the biological data was here limited to six clusters (Fig. 3). The decision on the cluster size is a trade-off between the strength of the associations within the cluster and the number of reference genes wanted. However, to see the entire association structure, the number of clusters equal to number of genes analysed is suggested. Alternatively, an algorithm deriving the cluster number from biological background was published [16].

The look at the associations between all genes, as facilitated by the cluster analysis, can exclude standardising with associated gene. If a distinct cluster contains predominantly known housekeeping genes, its genes can be applied for standardisation purposes in form of geometric mean as follows:

$$\text{Index} = \sqrt[n]{\text{CP}_1 \times \text{CP}_2 \times \text{CP}_3 \times \cdots \times \text{CP}_n} \quad (5)$$

where 1, 2, …, n are the genes [9]. Also genes, not *a priori* assumed to be unregulated, but tightly clustered with housekeeping genes can be included in the index. The standardisation model for relative quantification of change in expression was described by Pfaffl [1] and Excel based spreadsheet is also available Pfaffl et al. [19].

The presented SAS macro performs the simplest mostly default computing procedures. With some knowledge of SAS, it can be modified to perform with different settings or to produce more detailed output.

Acknowledgements

Authors thank their colleagues T.P. Neuvians and C. Prgomet of the chair for Physiology of the Technical University Munich for providing the data for analysis. Furthermore, authors thank Kathrin Kreyenberg for her critical review on the manuscript.

REFERENCES

[1] M.W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR, Nucleic Acids Res. 1 (2001) e45.
[2] O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, E. Heinen, Housekeeping genes as internal standards: use and limits, J. Biotechnol. 75 (1999) 291–295.
[3] T.D. Schmittgen, B.A. Zakrajsek, Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR, J. Biochem. Biophys. Methods 46 (2000) 69–81.
[4] H. Yamada, D. Chen, H.J. Monstein, R. Håkansen, Effect of fasting on the expression of gastrin, cholecystokinin, and somatostatin genes and of various housekeeping genes in the pancreas and upper digestive tract of rats, Biochem. Biophys. Res. Commun. 231 (1997) 835–838.
[5] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Accurate normalisation of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Gen. Biol. 3 (2002) 1–12.
[6] M.W. Pfaffl, A. Tichopád, Ch. Prgomet, T.P. Neuvians, Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations, Biotechnol. Lett. 26 (2004) 509–515.
[7] C. Urban, A. Schweinberger, M. Kundi, F. Dorner, T. Hammerle, Relationship between detection limit and bias of accuracy of quantification of RNA by RT-PCR, Mol. Cell. Probes 17 (2003) 171–174.
[8] C.R. Rao, The use and interpretation of principal component analysis in applied research, Sankhya A 26 (1964) 329–358.
[9] A.M. Kshirsagar, Multivariate Analysis, Marcel Dekker, Inc., New York, 1976.
[10] P.A. Chomczynski, Reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples, Biotechniques 15 (1993) 532–534.
[11] C.T. Wittwer, K.M. Ririe, R.V. Andrew, D.A. David, R.A. Gundry, U.J. Balis, The LightCycler: a microvolume multisample fluorimeter with rapid temperature control, Biotechniques 22 (1997) 176–181.
[12] R. Rasmussen, Quantification on the LightCycler instrument, in: S. Meuer, C. Wittwer, K. Nakagawara (Eds.), Rapid Cycle Real-Time PCR: Methods and Applications, Springer Press, Heidelberg, 2001, pp. 21–34.

[13] H.H. Harman, Modern Factor Analysis, third ed., University of Chicago Press, Chicago, 1976.
[14] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.
[15] M.R. Anderberg, Cluster Analysis for Applications, Academic Press, Inc., New York, 1973.
[16] D.R. Bickel, Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically, Bioinformatics 19 (2003) 818–824.
[17] V. Cherepinsky, J. Feng, M. Rejali, B. Mishra, Shrinkage-based similarity metric for cluster analysis of microarray data, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 9668–9673.
[18] S. Raychaudhuri, P.D. Sutphin, J.T. Chang, R.B. Altman, Basic microarray analysis: grouping and feature reduction, Trends Biotechnol. 19 (2001) 189–193.
[19] M.W. Pfaffl, G.W. Horgan, L. Dempfle, Relative Expression Software Tool (REST©) for group wise comparison and statistical analysis of relative expression results in real-time PCR, Nucleic Acids Res. 30 (2002) e36.