Search by cluster analysis for steadily expressed genes with application as normalization index in real-time RT-PCR

Ales Tichopad[1]& Michael W. Pfaffl[2]

[1]IMFORM GmbH. International Clinical Research, Birkenweg 14, D-94295 Darmstadt; [2]Physiology - Weihenstephan, Zentralinstitut für Ernährung- und Lebensmittel-forschung, Technische Universität München, 85354, Freising-Weihenstephan

BACKGROUND

Search for genes unregulated under treatment is an essential task before any relative gene-expression quantification can be conducted. Some simple approach ignoring the imaginary boundary between unregulated housekeeping genes and regulated genes is desired, that would group genes, based on a robust distribution-insensitive similarity measure.

MATERIAL & METHODS

Total RNA from 31 bovine Corpora Lutea was extracted. Data on expression levels of studied factors were obtained on LightCycler. In the 31 cDNA samples expression of four genes with assumed stable expression – housekeeping genes (HKG); Ubiquitin (UBQ), Glyceraldehyd-3-Phosphate Dehydrogenase (GAPD), ß-actin and 18S ribosomal unit was quantified together with ten 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, those expression is studied.

Similarity measure computation

Spearman rank-order correlation coefficient is a nonparametric measure of association based on the rank of the data values. The formula is

$$q = \frac{\Sigma(R_i - R)(S_i - S)}{\sqrt{\Sigma(R_i - \overline{R})^2 \Sigma(S_i - \overline{S})^2}}$$

where R_i is the rank of the *i-th* x value, S_i is the rank of the *i-th y* value, $\overline{R}$ is 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 the erroneous results due to non-normal distributed real-time PCR data.

Clustering procedure

Associated with each cluster is a linear combination of the genes in the cluster, which is the first principal component. A large set of genes can often be replaced by the set of cluster components with little loss of information. The first j principal components provide a least-squares solution to the model

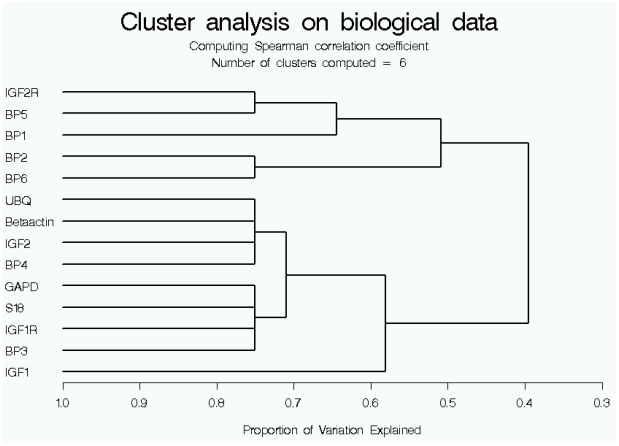
$$\mathbf{Y = XB + E}$$

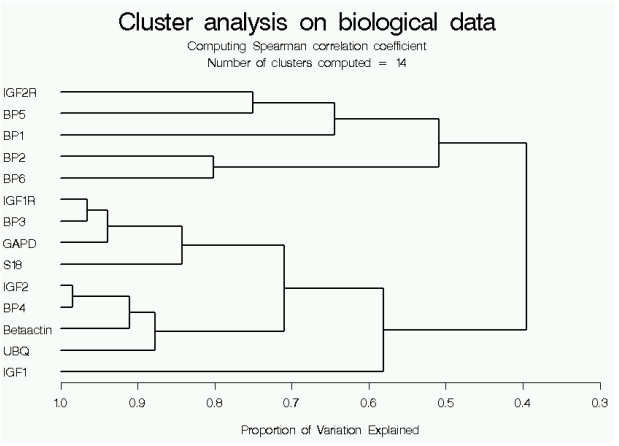
where **Y** is an $n \times p$ matrix of the centered observed variables; **X** is the $n \times j$ matrix of scores on the first j principal components; **B** is the $j \times p$ matrix of eigenvectors; **E** is an $n \times p$ matrix of residuals; and the trace(**E'E**), the sum of all the squared elements in **E**, is to be minimized.

			R-squared with	
		Own	Next	1-R²
Cluster	Variable	Cluster	Closest	Ratio
Cluster 1	UBQ	0.6362	0.3323	0.5448
	Betaactin	0.7239	0.3441	0.4210
	IGF2	0.7547	0.5038	0.4944
	BP4	0.8092	0.7010	0.6381
Cluster 2	BP2	0.7111	0.0226	0.2956
	BP6	0.7111	0.1439	0.3374
Cluster 3	IGF2R	0.6429	0.0174	0.3635
	BP5	0.6429	0.1200	0.4059
Cluster 4	IGF1	1.0000	0.0921	0.0000
Cluster 5	BP1	1.0000	0.0225	0.0000
Cluster 6	GAPD	0.6846	0.6351	0.8645
	S18	0.6217	0.4628	0.7043
	IGF1R	0.7732	0.3552	0.3517
	BP3	0.7960	0.4084	0.3449

TAB 1. Cluster listing. It displays the R^2 value of each variable with its own cluster and the R^2 value with its nearest cluster. The R2 value for a variable with the nearest cluster should be low if the clusters are well separated. The last column displays the ratio of $1\text{-}R_{own}^2/1\text{-}R_{nearest}^2$ for each variable. Small values of this ratio indicate good clustering.

International Clinical Research





RESULTS AND CONCLUSION

The cluster 1, considered the best separated, can be taken for normalisation purposes. Also the cluster 6 can still be considered well separated and useful for the normalisation purposes. Both the clusters contain some 'conservative' housekeeping genes. Alternatively, the clusters 1 and 6 can be joined. The encompassing cluster contains all the known housekeeping genes UBQ, Beta-actin, GAPD and 18S together with IGF-2, IGF-1R, BP-3 and BP-4. If a distinct cluster contains predominantly known housekeeping genes, its genes can be applied for normalization purposes in form of geometric mean as follows.

$$\text{Index} = \sqrt[n]{CP_1 \times CP_2 \times CP_3 \times \ldots\ldots \times CP_n}$$

where 1,2...n are the genes. Also genes, not *a priori* assumed to be unregulated, but those were tightly clustered with housekeeping genes can be included in the index.