Methods 50 (2010) 323–335

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Gene expression profiling – Clusters of possibilities

Anders Bergkvist [a], Vendula Rusnakova [b], Radek Sindelka [c], Jose Manuel Andrade Garda [d,1], Björn Sjögreen [a], Daniel Lindh [a], Amin Forootan [a], Mikael Kubista [b,e,*]

[a] MultiD Analyses AB, Gothenburg, Sweden
[b] Laboratory of Gene Expression, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[c] Whitehead Institute, Cambridge, UK
[d] Department of Analytical Chemistry, University of Corunna, Campus da Zapateira, 15071, Corunna, Spain
[e] TATAA Biocenter AB, Odinsgatan 28, SE-411 03 Gothenburg, Sweden

ARTICLE INFO

Article history:
Accepted 12 January 2010
Available online 15 January 2010

Keywords:
qPCR
Statistical analysis
Xenopus laevis
Hierarchical clustering
Heatmap
PCA
SOM
Hypotheses

ABSTRACT

Advances in qPCR technology allow studies of increasingly large systems comprising many genes and samples. The increasing data sizes allow expression profiling both in the gene and the samples dimension while also putting higher demands on sound statistical analysis and expertise to handle and interpret its results. We distinguish between exploratory and confirmatory statistical studies. In this paper we demonstrate several techniques available for exploratory studies on a system of *Xenopus laevis* development from egg to tadpole. Techniques include hierarchical clustering, heatmap, principal component analysis and self-organizing maps. We stress that even though exploratory studies are excellent for generating hypotheses, results have not been proven statistically significant until an independent confirmatory study has been performed. An exploratory study may certainly be valuable in its own right, and there are often not enough resources to report both an exploratory and a confirmatory study at the same time. However, exploratory and confirmatory studies are intimately connected and we would like to raise that awareness among qPCR practitioners. We suggest that scientific reports should always have a hypothesis focus. Reports are either hypothesis generating, from an exploratory study, or hypothesis validating, from a confirmatory study, or both. In either case, we suggest the generated or validated hypotheses be specifically stated.

1. Introduction

Modern scientific endeavours invariably involve reproducibility. To cite Werner von Siemens (1816–1892), 'good science needs good measurement'. As scientific questions continue to be resolved, the complexity of the remaining scientific questions increase. Complex scientific questions are likely to be obscured by unrelated processes that we as scientists often ascribe to random variables or noise. A distinction between the desired scientific knowledge and unrelated noise often requires statistical analysis. Although this is simple in principle, human intellect does not naturally seem to be optimally suited to parse such relationships and mistakes are often made because of subjective decisions. In this paper we will therefore attempt to describe some basic checkpoints that would facilitate correct handling of exploratory statistical analyses. We will also describe several alternative statistical methods relevant for pattern recognition by multivariate data analysis that are common for complex scientific questions.

A central checkpoint in statistical analysis is the hypothesis. We distinguish between an exploratory statistical study, if the aim of the study is to generate one or several hypothesis from a given data set, and confirmatory statistical study, if the aim of the study is to validate a given hypothesis or set of hypotheses by collection and analysing a new data set. We advocate that all statistical studies should provide a clear description of the hypothesis used, in case of a confirmatory study, or generated, in case of an exploratory study.

The statistical rigour required in a confirmatory study is very strict since we want to use the results for validation. Prior to collection and any testing of data, we need to define the hypothesis, what statistical test we are going to perform and what criteria we are going to use to judge whether we can or cannot reject the hypothesis, including whether the test will be one- or two-sided. It is important to make these definitions before data collection because a posterior selection of hypotheses, test and/or criteria may bias the conclusion and/or lead to an unintentional introduction of multiple testing which may compromise any statistical signifi-

[*] Corresponding author. Address: TATAA Biocenter AB, Odinsgatan 28, SE-411 03 Gothenburg, Sweden. Fax: +46 31 152890.
E-mail addresses: andrade@udc.es (J.M.A. Garda), mikael.kubista@tataa.com (M. Kubista).
[1] Fax: +34 981 167065.

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.01.009

cance found in this manner. In contrast to a confirmatory study, an exploratory study does not need to be defined prior to data collection, since the purpose of the exploratory study is to formulate a suitable hypothesis. Any approach that results in such a hypothesis may be a valid approach. However, it will not be clear whether the newly found hypothesis itself is statistically valid on the desired population and in the desired experimental setting until it has been validated in a confirmatory study.

Many mathematical and computational tools are available for exploratory studies. In this paper we will focus on visualization, clustering and partition methods for this purpose.

1.1. Embryonic development in *Xenopus laevis*

As an example to illustrate the power of visualization, clustering and partition methods for exploratory studies we will use measurements on transcripts coded by genes that are crucial for early embryonic development of African clawed frogs, *Xenopus laevis*.

RNA expression can be studied on two levels: as a spatial RNA distribution and a temporal RNA expression. The spatial expression profiles can be measured within different organs or different parts of the embryo [1], while temporal expression profiles are determined as function of development stage or time. Temporal expression profiling during early development of mammalian systems is limited by the small amounts of protein and RNA (few hundreds of pg per embryos) and time and money consumed preparing embryos. In contrast hundreds or thousands of embryos can be easily obtained by stimulating and in vitro fertilizing a single amphibian female. Further, amphibian eggs and embryos, such as those of *Xenopus laevis*, contain several μg of total RNA and proteins [2]. These features have made *Xenopus laevis* one of the most popular organisms for developmental studies. Two groups of mRNAs can be distinguished, based on temporal expression. One set, called maternal, is synthesized during oogenesis inside the mother's body. Proteins coded by maternal genes often have many roles in oogenesis and early developmental processes. Maternal mRNA molecules are translated into functional proteins in the oocyte during oogenesis and in the embryo after fertilization. At some point, the embryo reaches a stage when specialized gene products are needed and zygotic transcription is initiated. This point is called mid-blastula transition (MBT) and takes place at different developmental stages in different organisms. For example, mammalian zygotic transcription usually initiates after few cell divisions, while in *Xenopus* MBT occurs after 12 cell divisions in the gastrulation stage. The genes expressed after MBT are called zygotic genes.

Thirty-one genes that are crucial for early development of *Xenopus* were selected for high throughput qPCR expression profiling. Genes such as VegT, disheveled, p53, ubiquitin (ubc), Vg1, Xdazl, Xcad2, Oct-60, DEADSouth, alpha-tubulin, Stat3, U3-snoRNA, cytokeratin, Est1, Xmam1, An1, An2, 18S rRNA, mitochondrial cytochrome C (mt-cytC) and Wnt11 were previously found to be key components for early development and expressed in oocyte [3]. Similarly we selected genes, which were found to be important for developmental stages around and after MBT and therefore predicted to be expressed, such as siamois, chordin, HNF-3beta, Pax6, goosecoid, derriere, follistatin, cerberus and N-CAM. Thirteen developmental stages in biological triplicates from oocyte to tadpole were collected for high throughput qPCR analysis.

Primers and samples were loaded into a microfluidic chip, run and analyzed in the high-throughput BioMark qPCR platform [4]. Chips for dynamic qPCR analysis, which were used in our experiment, allowed us to run 48 cDNA samples times 48 primer pairs in parallel in a single run. Each run thus resulted in 2304 independent reactions with all cDNA sample:gene primer pair combinations. The analysis software GenEx developed by MultiD Analyses AB [5] was used for analysis and visualization of the data. Many other softwares are available to perform the studies presented herein; for instance, SPSS (SPSS Inc.), StatGraphics (StatPoint Technologies, Inc.) or open-code R and/or Matlab programs which can be downloaded freely, to mention but a few which are employed commonly.

1.2. Paper outline

The data analysis and visualization is performed using the analysis software GenEx, developed by MultiD Analyses AB [5].

Data measured with the BIOMARK platform are read and automatically annotated by the GenEx software. Analysis then starts with data preprocessing. This consists of normalization, imputation of missing data, removal of outliers, and scaling of data. Here normalization means scaling with endogenous reference or control samples to reduce systematic variations in the data. Scaling refers to rescaling needed to make the data analysis algorithms well conditioned. For example, mean centering of the data.

Some insight can be gained by visualizing the preprocessed data without further analysis. We will discuss visualization before proceeding to clustering. For clustering, we will consider four different methods to group the genes: three agglomerative hierarchical clustering methods, a divisive clustering approach, clustering by principal component analysis (PCA), and clustering by self organized maps (SOM).

2. Description of method

2.1. Sample collection and RNA isolation

Xenopus laevis females were stimulated by 500 U of hCG and kept overnight at 22 °C. Males were anesthetized in 0.1% tricaine solution for 20 min and testes were removed. A homogenized testes solution was used for in vitro fertilization by pouring to freshly squeezed oocytes. 0.1× Marc's Modified Ringers (MMR) medium was added after 5 min incubation at room temperature. After about 30 min after fertilization MMR medium was replaced by 2% cysteine solution to remove jelly coat. After short incubation, cysteine solution was removed and embryos were washed five times with 0.1× MMR solution. Thereafter the developing embryos were kept in 0.1 MMR at 25 °C.

Xenopus laevis embryos were staged according to [6]. Three sets of embryos (three embryos from the same female in each set) were collected from developmental stages 1, 2, 5, 6.5, 9, 10, 11, 13, 15, 21–22, 24–25, 38 and 44 and frozen at −70 °C. Total RNA from each sample was extracted using the RNeasy Mini kit (Qiagen) according to the manufacturer's instructions, including on column DNase treatment. Total RNA was eluted into 30 μl of elution buffer. Concentration of total RNA was measured with a Nanodrop instrument (Thermo Scientific). The RNA quality was analyzed on an Experion system (Bio-Rad).

2.2. Reverse transcription

cDNA was produced starting with 100 ng of total RNA, 1.5 μl of mixture 10 μM oligo-dT and 10 μM random hexamers (1:1) and water. The total volume was 6.7 μl. After incubation at 72 °C for 10 min, 100 U of MMLV reverse transcriptase (Promega), 12 U RNasin (Promega), 5 nmol dNTP and 2 μl buffer (5×) were added to a total volume of 10 μl, and incubated at 37 °C for another 70 min. The product was subsequently diluted to 100 μl and frozen.

2.3. Primer design and preamplification

Primers for the amplification of 31 selected genes were designed using Primer3 [7] and Beacon Designer (Premier Biosoft).

Primers' specificities and assay efficiencies were tested on control cDNA (mixture of cDNA from the three developmental stages). Criteria to accept a primer pair were: specific amplification of control cDNA with *Cq* lower than ~35, one peak in melting curve analysis, and no amplification of negative controls.

Preamplification PCR was run in 20 µl containing 2 µl of cDNA, 1 µl of all forward and reverse primers (500 µM each), 10 µl of Sigma A mastermix (kindly provided by Sigma, not yet a commercially available product) and water. CFX 96 cycler from Bio-Rad was used for preamplification with the cycling conditions: predenaturation at 95 °C for 2 min., followed by 14 cycles (95 °C 15 s., 59 °C 1 min. and 72 °C 1 min.). Product of the preamplification was diluted from 20 to 80 µl (4×) and stored at −20 °C. Preamplification efficiency was validated by comparing qPCR results of template that was and was not preamplified. Differences in *Cq* values between preamplified and not preamplified samples were similar for all genes [8], reflecting minimal bias and thus confirming the reliability of the preamplification.

2.4. High throughput qPCR performed on BioMark system

For qPCR analysis using the BioMark dynamic array (Fluidigm) a cDNA sample reaction mixture and a primer reaction mixture were prepared. The sample reaction mixture had a final volume of 5 µl and contained 1 µl of cDNA, 0.5 µl of SYBR Green Sample Loading reagent (Fluidigm), 2.77 µl Sigma A mastermix (Sigma, not provided yet), 0.165 µl of Chromophy, diluted 1:25 (TATAA), 0.025 µl of ROX (Invitrogen) and 0.1 µl of JumpStart DNA Taq polymerase (Sigma). The primer reaction mixture had a final volume of 5 µl and contained 2.5 µl of Assay Loading reagent (Fluidigm) and 2.5 µl mixture of reverse and forward primers corresponding to a final concentration of 10 µM. The chip was first primed with oil solution in the NanoFlex™ 4-IFC Controller (Fluidigm) to fill control valves. Bubbles were carefully removed from 5 µl of preamplified cDNA in sample reaction mixture and loaded into the sample wells, and 5 µl of the primer reaction mixtures was loaded into the assay wells of the dynamic array. The dynamic array was then placed on the NanoFlex™ 4-IFC Controller for automatic loading and mixing. After about 55 min the dynamic array was transferred to the BioMark qPCR platform (Fluidigm). The cycling program was 3 min at 95 °C for preactivation, followed by 30 cycles of denaturation at 95 °C for 15 s, annealing at 60 °C for 20 s, and elongation at 72 °C for 20 s. After completed qPCR melting curves were collected between 60 and 95 °C with 0.5 °C increments.

2.5. qPCR basic data analysis

An automatic exposure time with 72 °C calibration temperature was set up for measurement of fluorescence. Fluorescence signals were measured in the two channels: ROX and FAM-MGB. The raw FAM fluorescent data were normalized to the ROX signal. A linear baseline correction was used and the same threshold level was used for all assays. Quality threshold was set to 0.65.

2.6. Data preprocessing

qPCR data are frequently normalized by one of several options, including the expression of reference genes, number of cells, weight of tissue, DNA/RNA spike and total RNA concentration [9]. Expression of common *Xenopus laevis* reference genes in temporal developmental studies was found to be highly variable and unsuitable for normalization [10]. Endogenous reference genes are often powerful ways of reducing confounding variability introduced by technical handling. By not using endogenous reference genes, we emphasize careful technical handling. It may be more challenging to verify studied effects under these conditions, although it is certainly possible given that the studied biological variation is sufficiently larger than the confounding variation.

PCR products that gave unacceptable melting curves were classified as off-scale. Furthermore a cut-off of 28 cycles was used, and all *Cq*'s above 28 were treated as off-scale. The limit of 28 cycles was chosen due to our experience of this as a limit of reliable detection of true products for our BioMark instrument given the prior preamplification step. Measurements above 28 cycles were therefore judged to be extra sensitive to experimental handling errors. Off-scale *Cq* values were removed from the data set and subsequently treated as missing data. Patterns in remaining replicates indicated that missing data were due to experimental handling errors rather than concentration limiting errors. The information contained in the biological replicates was used to replacing missing data by the average of remaining biological replicates when available. If all biological replicates gave missing data, they were assigned the highest measured *Cq* for that particular gene +1. Since the highest measured *Cq* of a truly positive sample can be assumed to be the limit of detection (LOD) for that particular gene, assigning Cq(LOD)+1 to the off-scale samples corresponds to a concentration that is half of the LOD. Being below LOD is not equivalent to the sample being truly negative, because of sampling ambiguity. Hence, this is a rational correction of off-scale data for downstream processing of the results. The correction was further validated by reanalyzing the data assigning Cq(LOD)+2 to the off-scale measurement. This gave indistinguishable results evidencing that the correction does not affect the conclusions reached. How to deal with missing (outlying) data is a hot topic in statistics as any approach involves '*inventing the datum*' to some extent. The use of the average of the remaining values or, better, their mean assures that, at least, we do not disturb the major patterns on the dataset, which many times seems reasonable. Nevertheless, the reader should take into account that many other methods exist to cope with this difficult issue.

The relative expression among samples was calculated as [11]:

$$RQ = 2^{Cq_{\min}-Cq}$$

ΔCq values were calculated for each gene *Cq* by subtracting it from the lowest sample *Cq* and then converting the difference to linear scale as shown in the equation above. These relative quantities (RQ) indicate the level of expression, in each sample, of a particular gene relative to the sample in which the gene has highest expression. Hence, the RQ of the sample with highest expression for a particular gene is set to one and all other samples for that gene have RQ < 1.

Data must be normally distributed for analysis with parametric tests, such as the Student's *t*-test, linear regression, and ANOVA. If data transformed to RQ is normal distributed, further statistical analysis of data on this format may be preferred. However, gene expression data are usually not normal distributed when expressed as relative quantities, but usually become normal distributed by logarithmic transformation to fold differences (FD). Traditionally, log base 2 is used:

$$FD = \log_2(RQ)$$

Once the data has been transformed, it is, then, required to asses that outliers are not present into the series and, so, that normality is not clearly violated and that the parametric tests can be applied. The Grubbs' test [12–14] is generally recommended. For statistical analysis purposes fold differences are equivalent to the corresponding ΔCq values. However, for visualization purposes, the fold differences are preferred since they provide useful control over the relative pivot point as defined in the RQ transformation step.

2.7. Data pre-treatment

In general we analyze data with a hypothesis (or objective) in mind. This may have been determined *a priori* and our ambition is to validate it (confirmatory study) or our ambition is to browse through the data to propose a new suitable hypothesis (exploratory study). Regardless of whether the hypothesis is known *a priori* or not, we often know that some sort of comparison will be needed to study our data. To emphasize certain aspects of such a comparison it is therefore often appropriate to remove effects that may be present on the data, typically due to the different scales and ranges of variation of the genes. This requires scaling the data before further analysis. Here we describe two options: mean centering and autoscaling [11].

Genes are expressed at very different levels. Most genes have only a few transcripts per cell, while some few have tenths of thousands. In an analysis the few highly expressed genes will have much higher weight and may totally dominate the result. If this is not desired, the effect of the genes expression levels can be removed by subtracting the mean expression of every gene to the corresponding gene.

$$FD_{MC} = FD - \overline{FD}$$

Such transformed data are called mean centered. For mean centered data a certain deviation from normal (=mean) expression has the same weight independently of the expression level (or scale) of that gene.

Low expressed genes may show higher relative standard deviation (coefficient of variation) than high expressed genes, but hardly normal standard deviation. While a low abundant gene present in less than 10 copies can have standard deviations of perhaps 3–5 copies, a high expressed gene present in 100.000 copies will have a standard deviation of at least a few thousand copies. To remove also the effect of the magnitude of the change, data are further divided with the standard deviation:

$$FD_{AS} = \frac{FD - \overline{FD}}{SD} = \frac{FD_{MC}}{SD}$$

These new data are called autoscaled. An important drawback of autoscaled data is that subjects which vary randomly are given the same importance as subjects with systematic information because all variables are scaled to unit variance. Genes, whose expression is not sensitive to the studied parameters, should therefore be avoided, since they contribute only with noise to the analysis.

2.8. Direct visualization methods

As described in the introduction, our goal with the present report is to illustrate how an exploratory study can produce one or several testable hypotheses for future confirmatory statistical analyses. A hypothesis can be generated in many ways; however, it stands to reason that a good visualization improves a researcher's chances of proposing a good hypothesis. Here we distinguish between direct visualization methods that visualize the pre-treated or raw expression values of each (gene or sample) subject in the data set. Later we will see that visualizations of indirect descriptors, such as similarity measures or selected linear combinations of expression values, of the data set may be useful alternative ways to illustrate the data set.

One of the direct visualization tools is the *scatter plot*. In the scatter plot a subject (a sample or an mRNA target) is represented by a point in a coordinate system based on the expressions on each coordinate axis. However, due to human sensory limitations only two or three dimensions can conveniently be visualized in this way.

2D and 3D-line plots are ways to visualize one or several expression profiles. It is particularly useful to present trend studies where expression is measured as a function of time, drug load or other metric variable. However, when handling many and large expression profiles it may be difficult to distinguish features in such a plot. Once groups in the data set have been identified, typically by other means, the profiles in the trend plot can be coloured to emphasize different trend groups.

2.9. Hierarchical clustering methods

From a technical point of view, clustering can be performed either agglomeratively (i.e., by iteratively joining subjects (and small clusters) to form increasingly larger clusters, until all subjects have been accounted for in a comprehensive group), this is called *Hierarchical clustering*, or divisively (i.e., by iteratively breaking up clusters until only individual entities remain), this is called *Divisive clustering*. The latter option is not used frequently despite advanced algorithms exist, among them the '*k-means*' and the '*medoids*' ones. Both are partitioning methods in the sense that groups are found out from the initial overall set of data (which initially constitutes one group). The two are closely related and differ in how they minimize the distances, being the *medoids* method less sensitive to outliers. Medoids can be defined as those objects of a cluster, whose average (dis)similarity (mathematically, similarity and dissimilarity are closely related) to all the objects in the cluster is minimal; i.e., they are the most centrally located point in each cluster. The *medoids* or '*partitioning around medoids*' (PAM) algorithm clusters the objects (expression profiles, samples, etc.) into any of *k* clusters according to their (dis)similarity to each medoid. Its major difficulty is that *k* has to be decided in advance by some other means. PAM has been used successfully in the context of gene expression profiling [15,16]. Nevertheless, as agglomerative methods are employed much more frequently than the divisive ones, we will focus on them.

Clustering requires the selection of both a measure of similarity between samples (or genes) and a clustering algorithm.

2.10. Similarity measures

Clustering techniques are based on predefined criteria of similarity. In one of its most intuitive forms similarity is measured just as distances between points in the multidimensional space that is described by the expression vectors of assay samples or target genes. In fact, similarity has to be calculated as the inverse of the distance (the higher the distance, the lower the similarity). Note that it is equally possible to cluster samples (defined by the measured genes) and variables themselves (in this case the genes are grouped as a function of their expression throughout the different samples). To generalize the explanations, the term 'subject' will be used here to indistinctly refer to samples and variables.

There are many ways to mathematically define distance and, unfortunately, no '*golden rule*' can be given for a particular application. In general, the scientist has to try different options and select the most suited one (to what he/she is looking for). A note of caution is needed here as this does not mean that the scientist has to use the clustering methods to '*demonstrate*' its *preliminary* ideas; recall that clustering methods are intended to discover patterns among the datasets and, accordingly, something unexpected (or 'new' groups) should be studied carefully [17]. Here only some common distances will be discussed. The Euclidean distance is very intuitive as it generalizes the well-known Pythagoras' theorem.

$$E_{12} = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

E_{12} is the distance between samples 1 and 2 and x_{ji} represents the expression of each of the n genes ('i') measured on each sample ('j'). Small differences between each of the expression measurements x_{ji} result in a small distance E, which in turn can be interpreted as a measure of high similarity.

Another class of similarity measure is the correlation coefficient. In contrast to the distance measures correlation coefficients emphasize variational relationships between expression vectors. By variational relationship we mean that the differences between each of the expression measurements need not be small, but vary consistently throughout the sample vectors. For expression profiling purposes this is particularly useful to detect positive as well as negative correlations. The classical Pearson's correlation coefficient for two samples x and y, is [12,13]

$$r_{xy} = \frac{n\sum_{j=1}^{n} xy - \sum_{j=1}^{n} x \sum_{j=1}^{n} y}{\sqrt{\left[n\sum_{j=1}^{n} x^2 - \left(\sum_{j=1}^{n} x \right)^2 \right] \left[n\sum_{j=1}^{n} y^2 - \left(\sum_{j=1}^{n} y \right)^2 \right]}}$$

Here r_{xy} is the correlation between genes expression vectors of samples x and y where x_j and y_j indicate the expression values of the jth gene in the data set of n genes for the x and the y samples, respectively.

2.11. Clustering methods

As seen, measuring similarity between pairs of samples or variables (expression profiles) is relatively simple using a distance measure. Once samples (genes) with the highest similarity (lowest distance) were found, they are merged in a 'cluster' and the process is repeated. However, similarity measures between clusters need further definitions. To address this issue, a *clustering algorithm* has to be selected by the analyst. Four common options are reviewed here:

In the single linkage (or nearest neighbour) method groups are fused according to the distance between their nearest subjects, which are taken as representatives of their corresponding groups. The complete linkage (or furthest neighbour) method, behaves in the opposite way, as the distance between groups is now defined as the distance between their most remote subjects. The unweighted pairs or average linkage defines distance between groups as the average of the distances between all pairs of individuals in the two groups. It is sometimes also referred to as UPGMA (Unweighted Pair-Group Method using Arithmetic averages) [18,19]. It is a compromise between the single and complete linkages. Ward's method is more complex as it calculates the increase in the variance of the distances for the different possibilities of joining clusters. For each possibility, internal variance is computed as the sum of distances between each sample in the group and the group's centroid. The clustering that yields the lowest increment on the sum of the internal variances is then selected. Note that the Ward's method considers cluster analysis as an analysis of the variance problem, instead of using distance metrics. Therefore, a sample that was initially classified in a group might be removed from that group in a next step. Ward's method tends to produce compact clusters. In contrast, the single linkage tends to produce elongated groups, which sometimes are hard to interpret. The complete linkage tends to produce large numbers of groups.

2.12. Clustering visualization

The output of the clustering method is a figure which resembles a tree, which is called *dendrogram*, and it displays the distances among the individuals and the groups being formed. By careful selection of appropriate similarity measures (Table 1), and clustering algorithms (Table 2) the analyst may thus be able to highlight different aspects of the data set. Using the characteristics of different similarity measures and clustering algorithms (Tables 1 and 2), the significance of the groups has to be interpreted by the analyst and, so, explain the rationale of each group (i.e., what differentiates a group from the others). An advantage of the hierarchical clustering compared to the direct visualization methods is that a high dimensionality (a large number of genes and samples) of the data set is reduced to a convenient two-dimensional representation of subject similarities.

Hierarchical clustering can be performed either for the genes (comparing samples' expression profiles) or for the samples (comparing genes' expression profiles). The two classifications can be combined to produce a *heatmap* of the data set. The heatmap is a colour-coded two-dimensional mosaic that describes the whole expression matrix (samples vs. gene targets), each tile coloured with a different intensity according to the pre-processed data. In addition, the data set is reordered in each dimension of the mosaic according to the dendrograms calculated for samples and genes, respectively. The heatmap literally adds another dimension of information presented by the dendrogram, which may facilitate its interpretation.

2.13. Principal component analysis and its visualization

Principal components analysis, *PCA,* is a powerful approach to circumvent the dimensionality problem of scatter plots by projecting the high-dimensional data set onto two or three dimensions for easy visualization. Here, the original coordinate system of the data set (i.e., the measured expression profiles) is projected onto a new space with a lower number of new variables, so called principal components (PC's) or factors. Each PC is a linear combination of the subjects. Thus, when illustrating samples, each coordinate axis is a linear combination of genes' expression levels from the original measurements, and when illustrating genes each coordinate axis is a linear combination of sample expression levels from the original measurements.

By mathematical definition, the PC's are extracted in successive order of importance. This means that the first PC explains most of the information (variance) present in the data, the second less, and so forth. Therefore, we can use the first two or three PC coordinates (termed *scores*) not only to obtain a projection of the whole data set onto a conveniently small dimension, but also to obtain the projection that accounts for the most relevant variability in the data set. Variance from experimental design conditions is expected to be systematic, while confounding variance is expected to be random. Since the last PC's explain a very low amount of information, they can be considered to include noise or random information and can therefore be ignored. In this way, PCA can be a very efficient tool to separate systematic effects from noise.

2.14. Self-organizing maps

The *self organizing map, SOM*, or Kohonen artificial neural network, approach is an advanced iterative method based on comparisons, rather than an algorithm as used in hierarchical clustering and PCA. It is based on the use of artificial neural networks and their capability to 'learn', i.e., to change their internal values (called *weights*) to be as similar as possible to the measured samples. It can be used to visualize multivariate data sets in two dimensions (the

Table 1
Similarity measures.

Similarity measure	Effect
Euclidian distance	Favors direct one-to-one similarity. Results depend on scaling of expression profiles.
Pearson correlation	Emphasize variational relationships between expression vectors. Detects both positive and negative correlations.

Table 2
Clustering methods.

Clustering method	Effect
Single linkage	It has a chaining effect, useful for extended clusters with irregular shapes.
Complete linkage	Produces small compact clusters.
Ward's method	Produces very small and compact clusters.

graph obtained is called a *Kohonen map*). For the SOM approach, the analyst defines the size of a matrix of nodes, typically in two dimensions. The SOM algorithm then starts by associating the weights of each node with a random expression profile of the same dimensions (i.e., number of genes, if the object of study is gene expression profiles or number of samples, if the object of study is samples expression profiles) as the data set to be analyzed. Next, each expression profile in the real data set is sequentially assigned to the node that most resembles its own profile. The assigned expression profile is allowed to influence the expression profile in the node and its neighbours by a weighted factor that is large for the central node and progressively smaller for more distant neighbours. The procedure of assigning expression profiles and allowing them to influence the values of the weights of the related neuron is repeated iteratively until convergence is achieved. Finally, each node receives a code, according to the sample(s) that was(-were) assigned there and this is plotted in a 2D plot (the map), where the distribution of the samples can be visualized and, hopefully, the groups interpreted.

An advantageous property of the SOMs relative to, for example, PCA, is that the distances between the expression profiles of neighbouring nodes in the map are non-linear. This property may allow detection of otherwise obscured similarity patterns. Another advantage of the SOM is that the size of the node map may be adjusted to seamlessly transition from clustering (no prior assumptions of number of clustering classes) by using a large node map to partitioning (pre-defining a number of clustering classes) by using a small node map. For example by using a node map of dimensions 31 × 31 on a data set with 31 expression profiles, plenty of space is available for the process to distribute the expression profiles and the resulting clustering structure is thus free from any restraints. In this situation individual relationships between expression profiles can be studied under favourable conditions, but boundaries between groups may be difficult to discern. On the other hand, if a node map of dimensions 10 × 1 is used on the same data set the process is restrained to a maximum of 10 partitions. In this case, individual relationships between expression profiles are less apparent, but boundaries between groups are highlighted.

2.15. Analysis of visualization results

As when applying any chemometrics or bioinformatics technique, validation is a key stone. One has to be aware that by their own nature pattern recognition techniques (like clustering) yield results that, sometimes, may be surprising for the scientist. Even though hierarchical dendrograms and heatmaps provide powerful visualizations of data sets, they hold drawbacks in their own nature. This problem is well-known in clustering and can only be detected by close examination and interpretation of the final dendrogram and, probably, by repeating the clustering using different distances and clustering algorithms.

Many studies have been devoted to evaluating clustering quality. Some common terminology includes compactness, connectedness and spatial separation [17].

Compactness is a consequence of various clustering algorithms tendencies to keep intra-cluster variation small. Algorithms and parameters that push in this direction have a tendency to give results that are appropriate for spherical clusters, but may fail if the clusters have more complex shapes. Clustering algorithms that implement parameters to favour *connectedness* tend to give results that are appropriate for arbitrary shaped clusters, but lack robustness when there is little separation between clusters (typically, this happens with the single linkage method). *Spatial separation* is rarely used by itself as an algorithm objective, but is usually combined with the other objectives above.

In practical terms it is highly difficult to evaluate these parameters with an unknown dataset and, therefore, in this paper our ambition was to perform an exploratory study. We are therefore not interested in defining any objective measure of cluster quality, but we will use the terminology described above to subjectively describe some patterns that may help us propose hypotheses for future confirmatory studies. Many times we are mining in our (almost unknown) dataset and a multitude of different methods can be applied to draw conclusions. A careful and sound biological explanation of the results is mandatory before accepting any model. Recall always that, then, your model will be accepted under the explicit condition that the resulting conclusions are nothing more than hypotheses that will need to be analyzed in an independent confirmatory study before they can be considered truly validated.

3. Results

Many scientific questions remain on early development of *Xenopus laevis*. Typical questions may be: What is the function of each or gene? At what phase of development are they active? How are they regulated? How do they influence each other? With the increased availability of sequence data and gene-specific characterizations, large scale investigations into these matters are becoming feasible. However, with the increased amount of data we also face new challenges. Specifically: the challenge of finding new, testable, and biologically relevant hypotheses. Here we used the data set of 31 selected genes, measured at 13 stages throughout *Xenopos laevis* development, to illustrate how hypotheses may be generated from a multidimensional data set in an exploratory study.

Data organization is arguably one of the most time-consuming steps in large scale multivariate data analysis. The separate analysis algorithms can typically be performed quickly at a click of a button once input data has been arranged in appropriate format.

In our case, we have collected expression data of 31 genes in 13 developmental stages. Biological triplicates were collected for each stage. The triplicates were organized in the data set under the samples dimension since each replicate can be said to be of the same gene, but not of the same sample. The full data set thus forms a matrix of dimension 39 × 31. It is also useful to assign classes, when available, to the data set. GenEx allows for classification columns and rows to classify samples and genes. Analysis methods applied to the data read the data in rows, columns or the whole matrix. Analysis methods that are applied to either rows or columns thus analyze the data from either the gene or the sample perspective. By transposing the data matrix, analyses can be shifted from gene to sample perspective or vice versa. In the case of the Xenopus data we have clear classification of the samples as the mid-blastula transition is a well-known transition state during development. The 13 samples can therefore readily be classified as early, midblastular or late stages. Fig. 1 shows a 3D line-plot of sample expression profiles for each of the Xenopus developmental stages. To reduce cluttering and improve visibility the biological replicates have been averaged for Fig. 1. Off-sets may be present in the data sets for technical reasons, unrelated to the biological questions of interest, in the expression profiles of each sample. To further improve visibility and remove off-sets we therefore mean centered the data for Fig. 1. The colours of the 3D lines correspond to the three classes of developmental stages: early (red), midblastular (blue) and late (green) stages. Features that are characteristic for each group are readily visible and the distinction between the groups is clear although best appreciation for this is achieved on a computer screen where the 3D line plot can be rotated for improved 3D visualization.

Heatmaps are an alternative representation of expression profiles. The mosaic representation of the heatmap in Fig. 2 illustrates expression amplitude by colour intensities. It is often useful to recognize patterns in the heatmap mosaic and compare them to the expression profiles in a 3D line-plot such as shown in Fig. 1. The birds-eye-view of the heatmap allows this representation to be less sensitive to information overload and cluttering. In Fig. 2 the biological replicates have been retained and mean centered in the samples dimension. Comparisons should therefore primarily be performed in the samples dimension of the heatmap, i.e., treating it as a set of columns. For a 3D line-plot this large number of samples would have obscured any interpretation, but heatmap produces a very comprehensive overview. In addition, the rows and columns of the heatmap are organized through hierarchical clus-

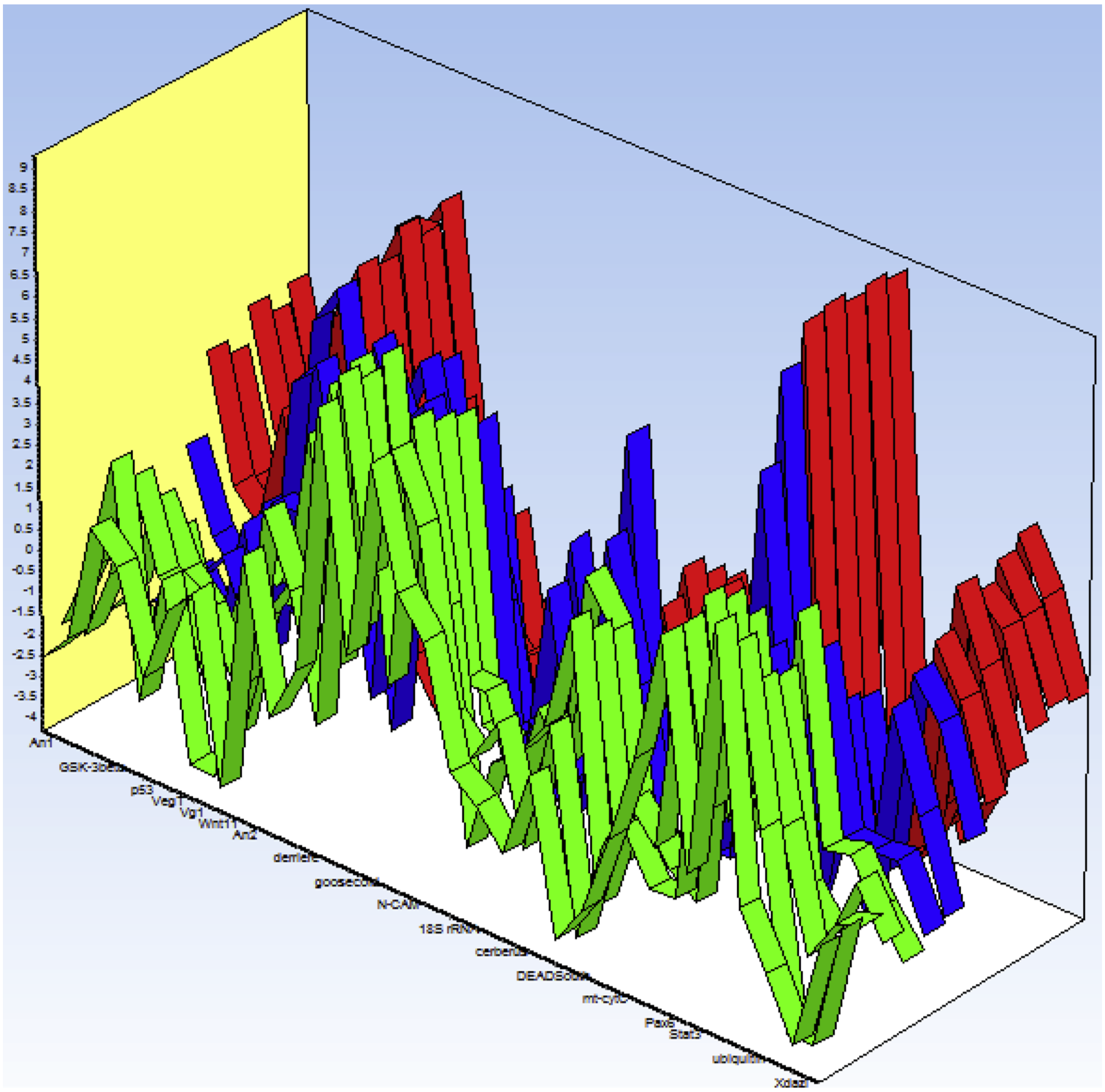

Fig. 1. A 3D line plot of mean centered sample expression profiles in *Xenopus laevis*. Red expression profiles are early, blue are intermediate and green are late developmental stages.

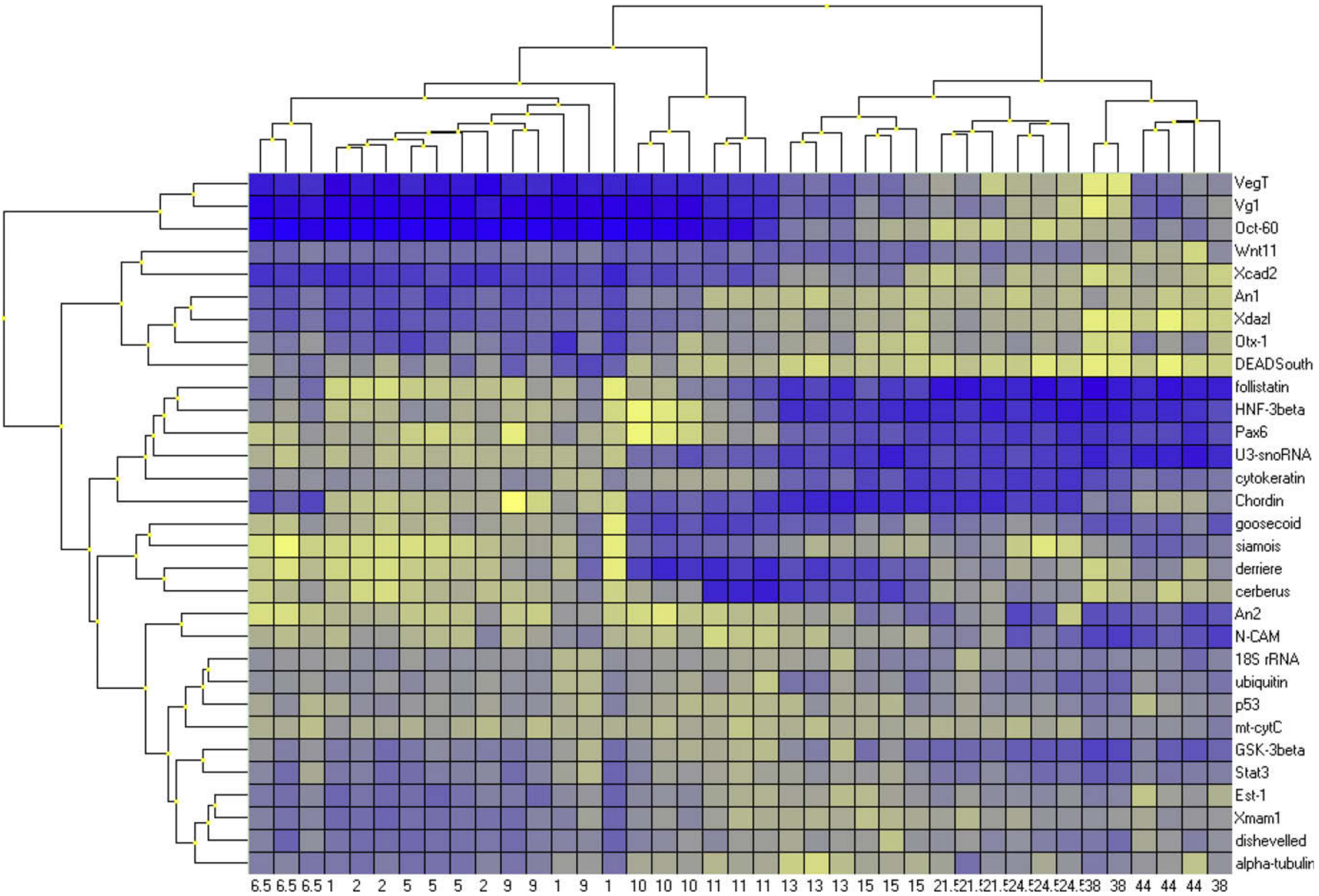

Fig. 2. Data with biological replicates, mean centered in the samples dimension. Light yellow corresponds to low expression, dark blue corresponds to high expression. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

tering. The dendrograms that indicate the relationships between the samples (columns) and genes (rows) thus gives additional information to the comprehensive overview. In our current data set we see that we have three distinct groups corresponding to the early (1, 2, 5, 6.5, and 9), midblastular (10 and 11) and late (13, 15, 21–22, 24–25, 38, and 44) developmental stages. We also see good reproducibility among the biological replicates, although some samples (notably sample 1) have variabilities large enough to make them overlap with several other neighboring developmental stages.

A heatmap is also very useful to characterize relationships between genes (Fig. 3). To emphasize the developmental characteristics of the gene expression profiles we average the biological replicates. This will give us a better estimate of the mean expression at each developmental stage and also reduce the information content in the heatmap for a more comprehensive overview. By autoscaling in the gene dimension we emphasize comparisons between gene expression profiles (rows in the heatmap), normalized so that the variability within each expression profile will be equal between different expression profiles. By inspecting the gene expression dendrogram on the left side of the heatmap in Fig. 3 we observe that, in the most simple, non-trivial clustering tier, they are organized into three distinct groups. The topmost genes (goosecoid, siamois, derriere, cerberus and chordin) have their highest expression levels primarily at the MBT (midblastular genes). The genes in the middle group (HNF-3beta to ubiquitin) have their highest expression levels primarily at the later developmental stages (zygotic genes). These group assignments are summarized in Table 3 and compared to assignments from Kubista et al. [20]. The bottom group of genes (An1–Wnt11) has high expression at the early developmental stages (maternal genes). However, there are also several interesting substructures in this data set. For instance, the maternal genes disheveled, Xmam1, Est-1 and alpha-tubulin seem to have an increased expression at the later stages of development compared to other maternal genes. Likewise, the midblastular genes goosecoid and siamois seem to have an increase expression at the developmental stages 38 and 44 compared to other MBT genes. The zygotic genes seem to divide into two characteristic groups one (genes HNF-3beta to U3-snoRNA) of which seems to have rather even expression throughout the later developmental stages (stages 13–44) and the other (genes GSK-3beta to ubiquitin) which seem to have a distinct increase in expression in the last few stages of development (stages 24, 25–44). These observations may inspire several interesting hypothesis to be validated in future confirmatory studies.

Now that we have used the heatmap and the associated hierarchical clustering to identify groups of genes, we can improve visibility in the 3D line-plot of the gene expression profiles by colouring each expression profile according to which group it has been assigned. Fig. 4 shows the autoscaled gene expression profiles of the Xenopus data set. Without the colour assignments it would be hard to discern patterns in the data set, but with the colour assignments the 3D line-plot literally gives another dimension of detailed information on the expression profiles. Now, we have a very good idea of what-is-going-on in our data set and, therefore, are ready to interpret the multivariate models that different multivariate techniques can offer us.

As explained above, hierarchical clustering is a way to visualize similarities in a multidimensional data set. An alternative is to produce a projection of the multidimensional data onto a two-dimen-

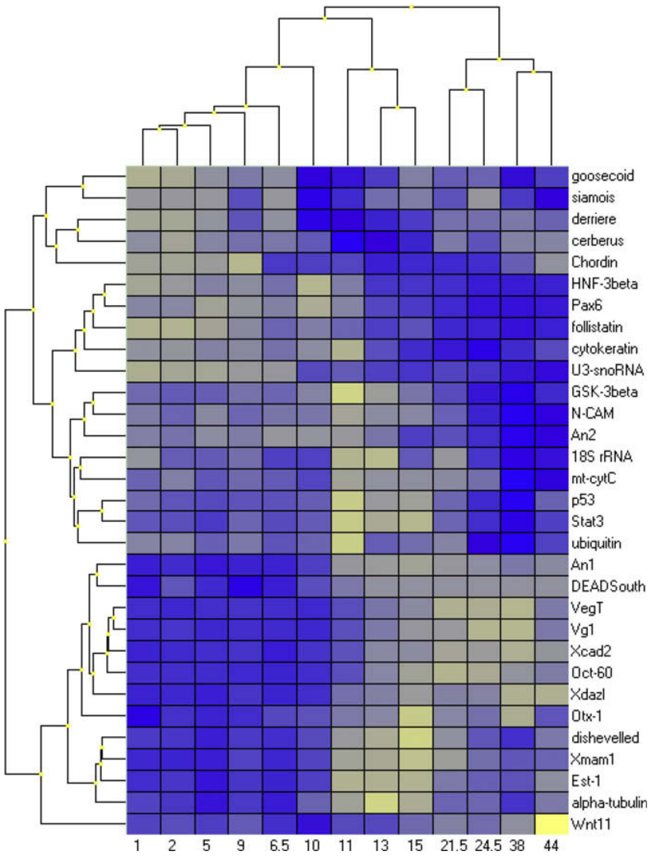

Fig. 3. Data autoscaled in the gene dimension. Samples 1–9 are from early stages in development, 10–13 are from the Mid-blastula transition, and 14–38 are from late stages in development. Light yellow corresponds to low expression, dark blue corresponds to high expression. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

sional scatterplot through PCA. The first two principal components of the mean centered samples expression profiles in the Xenopus data set is shown in Fig. 5. These two principal components account for 75.8% of the variation in the data, even though it is only a two-dimensional representation of a 31-dimensional data set (each sample vector contain expression measurements from 31 genes). As we saw in the heatmap dendrogram of the samples expression profiles (Fig. 2), the early, MBT, and late stages tend to be similar within each group than to members of other groups. An advantage of the PCA representation over the dendrogram is that the order of similarities is easier to discern. In Fig. 5 we see that there is a general trend from the early stages through the MBT stages to the late stages. However, there are some subpatterns that may be observed here too. For example, for the genes in this study it seems that stage 6.5 is more similar to later stages than stage 9. This we can interpret as a temporary regression in the development in stage 9 before development proceeds in stage 10 and beyond. Furthermore, stage 11 seems to constitute a significant jump in development compared to the other developmental stages in this study, as can be seen in the relatively large separation between this cluster of biological replicates and others. This fits with our expectations that the MBT is a dramatic transition during development. Finally, it seems that the Xenopus development goes through a local extreme around stage 38 before finally settling in (stage 44) to a state that is more similar to earlier developmental stages (stages 21–22 and 24–25).

The PCA can also be performed for the gene expression profiles. Fig. 6 shows that maternal, MBT and zygotic gene groups form distinct groups in the first two principal components of autoscaled gene expression profiles from the Xenopus data set. In this case, the first two PCs account for 80.5% of the total variation in the data. It is more than what we observed for the PCA in Fig. 5, although in this case, since we have averaged the biological replicates, the PCA is based on a 13-dimensional data set (of sample expression measurements). In either case it is a great focus on the systematic variation in the data set on only two easily visualized dimensions. In Fig. 6, the separation is clear between each group. Neither of the groups is perfectly connected. The maternal cluster may be interpreted to have three subgroups, and the MBT and zygotic groups may be interpreted to have two subgroups each. There seems to be a trend of sequential similarities of gene expression profiles from maternal to zygotic genes with the MBT genes being a separate group from this sequence of gene expression profiles. From a biological point-of-view it may make sense that different genes are activated and deactivated in sequence as they are required for each developmental stage. To further evaluate this as a tentative hypothesis we may take advantage of the inherent strengths and flexibilities of the SOM. Despite a two dimensional PC scores plot has been presented here, three dimensional plots can often be of much use as more information will be taken into account and, so, interesting features or patterns can appear.

Fig. 7 shows the SOM analogous to the PCA plot in Fig. 6. It is based on autoscaled gene expression profiles of the Xenopus data set. The default setting of the SOM in GenEx is a map of size equal to the number of expression profiles in each dimension. In our case we have 31 gene expression profile and thus a 31 × 31 matrix of SOM nodes. An obvious characteristic of the SOM is that the data points are distributed more evenly across the plot area. This is due to the fact that the plot area is non-linear and neighboring data points may be as similar or dissimilar as other data points further away in a different direction. This cause a greater separation among data points, and reduces the compactness of the clusters. With a large node matrix as the one in Fig. 7 it may be challenging to discover or confirm clustering tendencies. However, an interesting feature of the SOM is that the size of the node matrix can easily be reduced to funnel the data points into a specific number of available partitions. Keeping the notion of the tentative hypothesis of sequential activation of gene expression from the previous paragraph, we may suggest a specific organization of the SOM node matrix. In Fig. 8 we have three representations of 10 × 1 SOM matrices. By selecting the size of one side of the node matrix to

Table 3
Xenopus gene group assignments. Gene names in bold indicate genes assigned to different groups in Kubista et al. [20]. Chordin, derriere, goosecoid, and siamois were assigned to the late group in [20]. GSK-3beta and p53 were assigned to the early group in [20].

Developmental assignment from this study	Genes also analyzed in Kubista et al. 2006	Genes not analyzed in Kubista et al. 2006
Early	VegT, Vg1, dishevelled	Oct-60, Xcad2, Xdazl, An1, DEADSouth, Otx-1, Xmam1, Est-1, alpha-tubulin, Wnt11
MBT	Cerberus, **chordin**, **derriere**, **goosecoid**, **siamois**	
Late	Follistatin, HNF-3beta, N-CAM, **GSK-3beta**, **p53**	An2, Pax6, U3-snoRNA, 18S rRNA, mt-cytC, Stat3, ubiquitin, cytokeratin

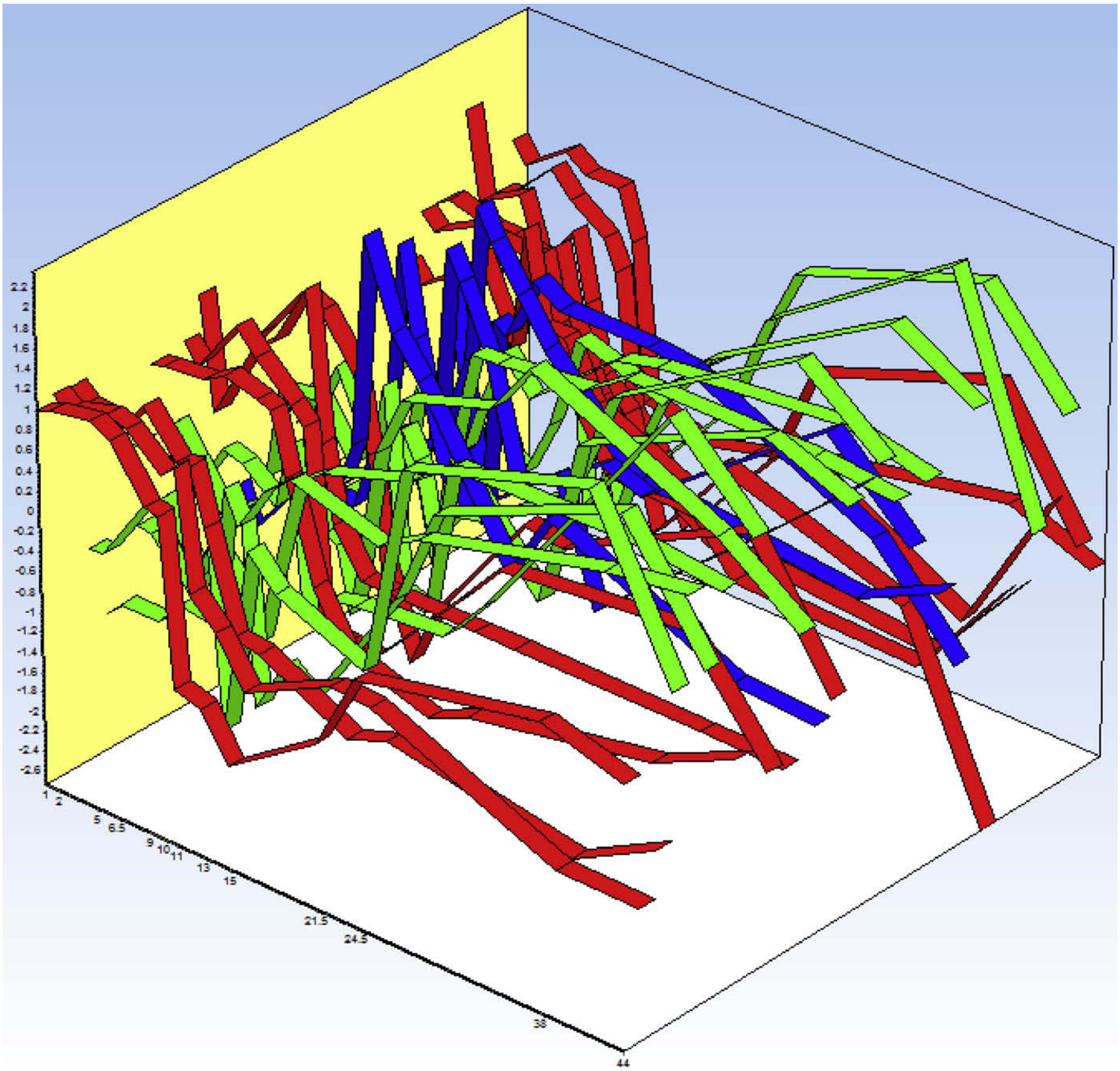

Fig. 4. A 3D line plot of autoscaled gene expression profiles in *Xenopus laevis*. Red expression profiles show maternal genes, green show zygotic genes and blue show midblastular or unknown genes.

one we naturally obtain a node structure conducive to investigate sequential patterns. Because there is an element of random starting points for each SOM generation, the SOMs are not identical. However, by comparing repeated SOM runs and identifying consistent patterns, the SOMs can be used to propose testable hypotheses. For example, here we find that expression profile of VegT seem to reach its higher levels before the expression profile of Otx-1 does. This hypothesis may be biologically sensible since VegT has previously been associated with mesoderm formation, and Otx-1 has been associated with brain development [2].

4. Concluding Remarks

From the observations we have made in the current study, we may propose several hypotheses for further study and potential validation in future confirmatory studies. Some examples may be:

- The average gene expression during the late stages (21–22, 24–25, 38 and 44) relative to the average gene expression during the early stages (1, 2, 5, 6.5, and 9) is higher for Otx-1 than for VegT.
- The expression of mt-cytC is more than 4 times larger during stage 38 than during stage 24-25.
- The average expression of cerberus is larger during the MBT stages (10 and 11) than during an average of other developmental stages (1, 2, 5, 6.5, 9, 21–22, 24–25, 38 and 44).

These examples illustrate how exploratory studies, and clustering and visualization techniques are powerful tools to generate hypotheses. It may be tempting to draw scientific conclusions already at this point of the analysis, in particular if the tendencies in the data seem to be very strong. However, without hypotheses defined before the study, we need to be mindful that we may, consciously or unconsciously, be testing a potentially unlimited

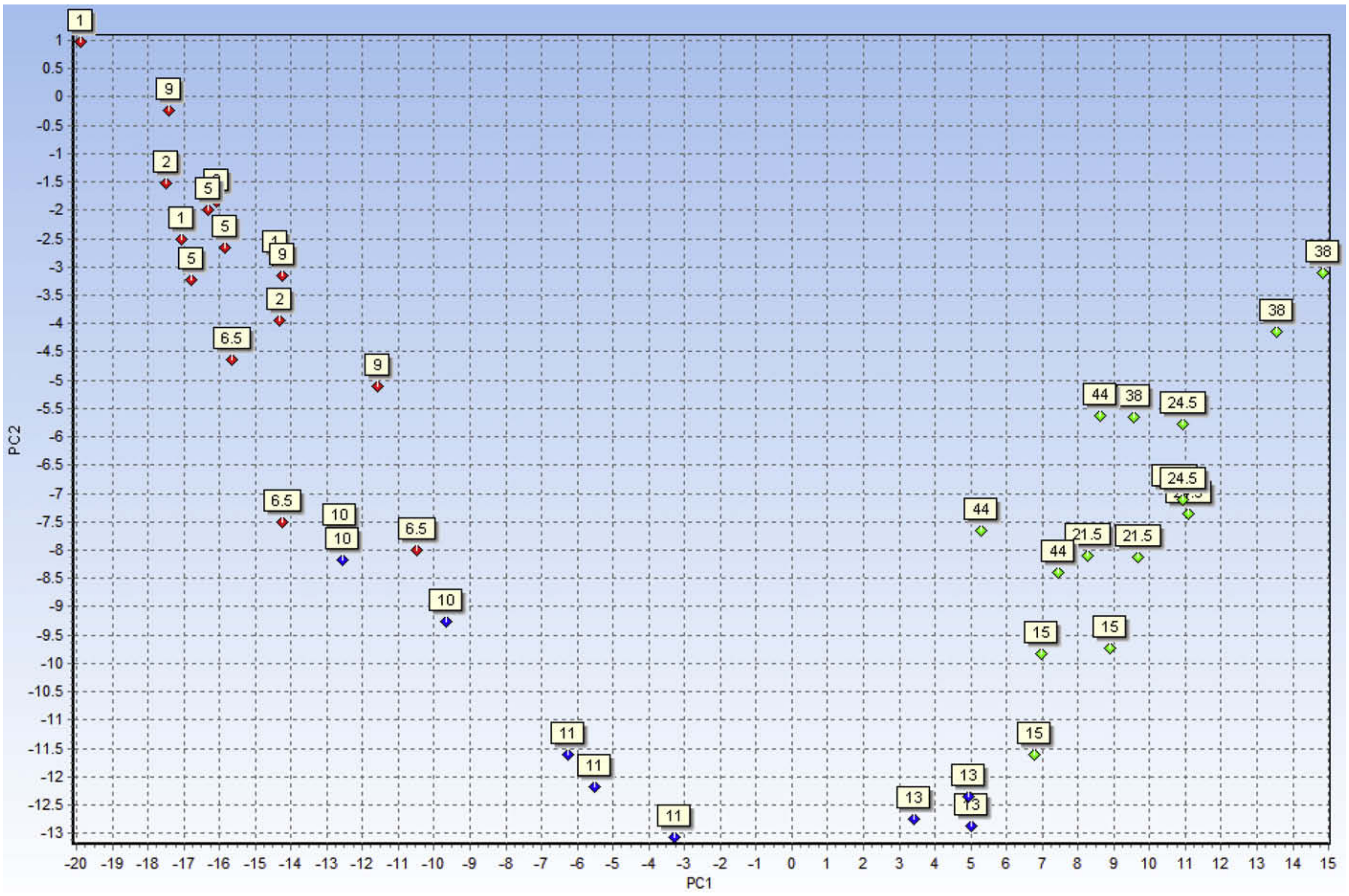

Fig. 5. PCA on Xenopus samples (developmental stages), mean centered, each developmental stage represented by triplicate measurements. Early stages in red, MBT transition stages in blue and late stages in green. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

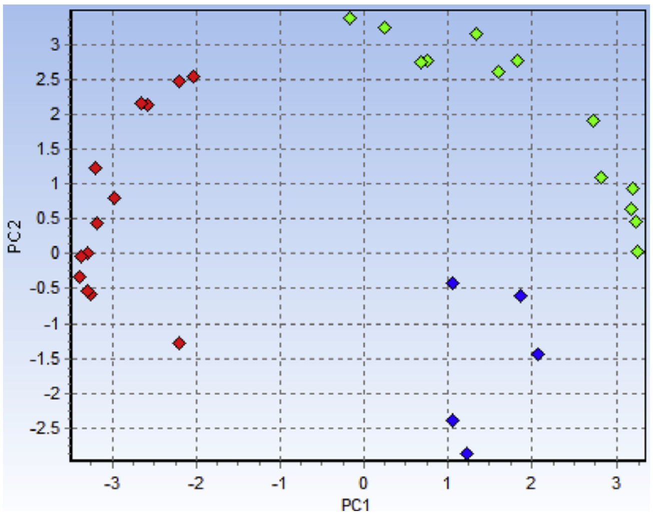

Fig. 6. PCA on Xenopus genes. Autoscaled in genes dimension, maternal genes in red, zygotic genes in green and other genes in blue.

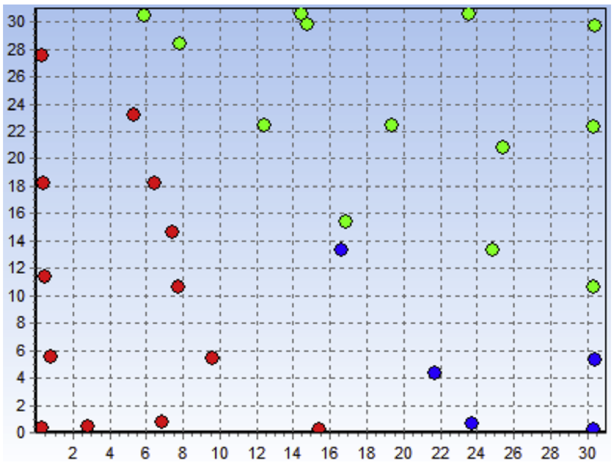

Fig. 7. SOM on Xenopus gene expression data with dimensions 31 × 31. Maternal genes in red, zygotic genes in green and other genes in blue.

number of hypotheses simultaneously. Testing many hypotheses simultaneously require corrections of the underlying statistics parameters in order to maintain overall significance level of the study [21]. Without these corrections there is a risk that observations of random events may be interpreted as biologically relevant processes. The generation of hypotheses is nevertheless an important part of the scientific process. We conclude that we have many useful tools for hypothesis generation and that statistical validation of new-found hypotheses is an integral part of the scientific process, even though it may be left to future studies.

Despite the power of the multivariate methods, we would like to stress that scientist must be aware that their sound knowledge and scientific skills are critical to validate/interpret/ascertain what they have found. The many clustering methods, distance metrics, and

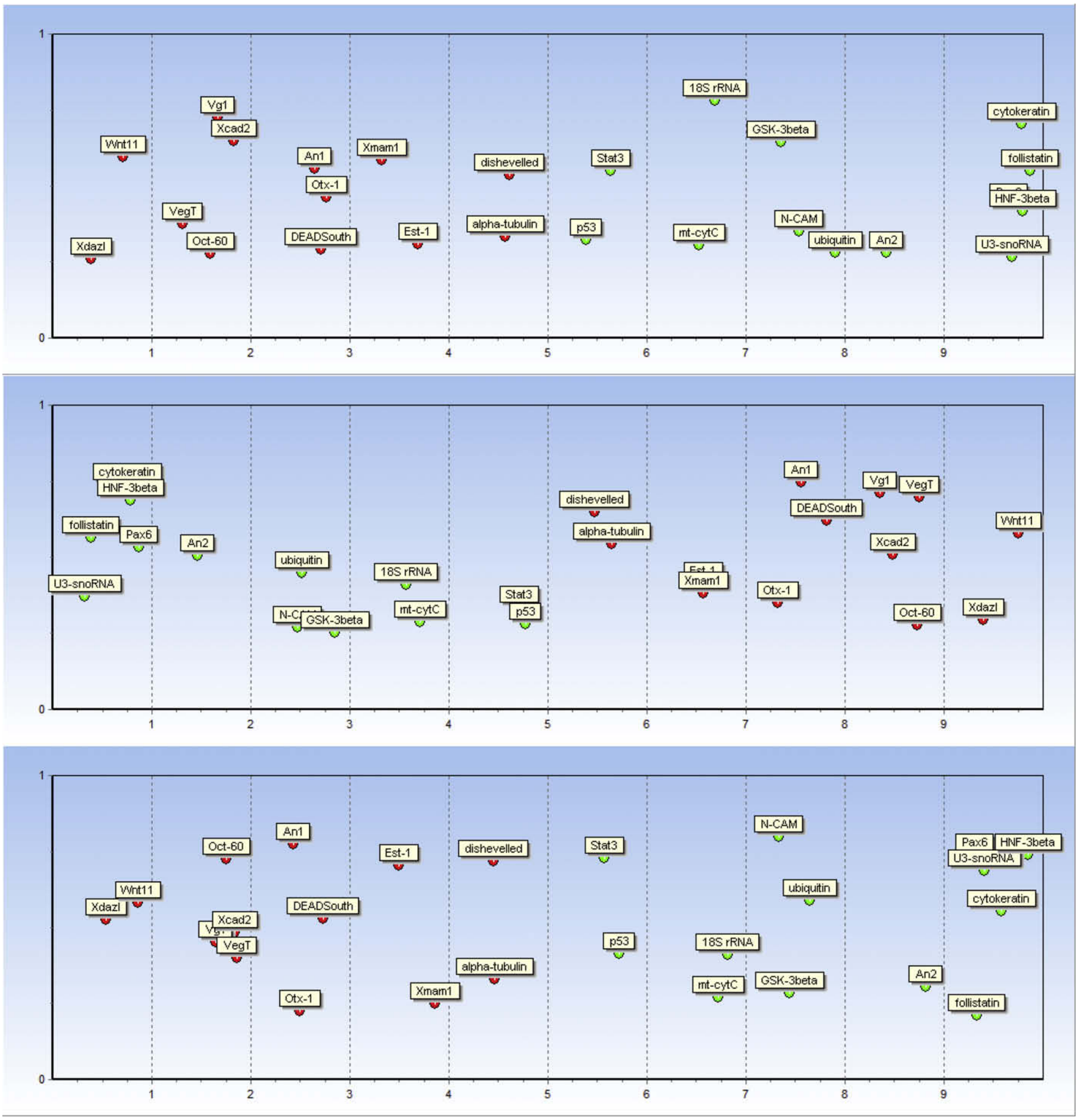

Fig. 8. Three SOMs on Xenopus genes with dimensions 10 × 1 show transition from maternal genes (red) to zygotic genes (green). Each SOM pattern is generated from a process that is influenced by random starting conditions and therefore the SOM patterns do not need to be identical. In the examples here we nevertheless see that there are consistent patterns in the data, such that the ordering of the gene expression profiles is preserved.

their combinations, make it too easy to bias the results to a subjective hypothesis on the data structure/patterns. Key point here is to be open to discover something unexpected but interesting.

Acknowledgments

Part of this work was supported by Grant agency of Academy of Science the Czech Republic (GAAV IAA500970904, IAA500520809) and research goal AV0Z50520701 Granted by Ministry of Youth, Education and Sports of the Czech Republic, together with support from Grant agency of the Czech Republic (GACR 301/09/1752).

References

[1] R. Sindelka, J. Jonák, R. Hands, S.A. Bustin, M. Kubista, Nucleic Acids Res. 36 (2) (2008) 387–392.
[2] J.B. Bowes, K.A. Snyder, E. Segerdell, R. Gibb, C. Jarabek, E. Noumen, N. Pollet, P.D. Vize, Nucleic Acids Res. 36 (2008) D761–D767.
[3] M.L. King, T.J. Messitt, K.L. Mowry, Biol. Cell 97 (1) (2005) 19–33.
[4] http://www.fluidigm.com

[5] http://www.multid.se
[6] P.D. Nieuwkoop, J. Faber, Normal Table of *Xenopus laevis*, Garland Publishing, Inc., New York & London, 1994.
[7] http://frodo.wi.mit.edu/primer3/
[8] R Sindelka, et al., to be published (2010).
[9] A. Bergkvist, A. Forootan, N. Zoric, L. Strömbom, R. Sjöback, M. Kubista, Genet. Eng. Biotechnol. News 28 (13) (2008) 26–28.
[10] R. Sindelka, Z. Ferjentsik, J. Jonák, Dev. Dyn. 235 (3) (2006) 754–758.
[11] M. Kubista, R. Sindelka, A. Tichopad, A. Bergkvist, D. Lindh, A. Forootan, G.I.T. Lab. J. 9–10 (2007) 33–35.
[12] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, fourth ed., Pearson, Harlow, 2000.
[13] L.L. Havilcek, R.D. Crain, Practical Statistics for the Physical Sciences, Edit ACS, Washington, US, 1988.
[14] V. Barnett, T. Lewis, Outliers in Statistical Data, third ed., Willey, Chichester, UK, 1994.
[15] M.J. van der Laan, K.S. Pollard, J.F. Bryan, A New Partitioning around Medoids Algorithm, The Berkeley Electronic Press, University of California, Berkeley, 2002. paper 105.
[16] M.J. van der Laan, J.F. Bryan, Biostatistics 2 (2001) 1–17.
[17] J. Handl, J. Knowles, D.B. Kell, Bioinformatics 21 (2005) 3201–3212.
[18] G.H. Lance, W.T. Williams, Comput. J. 9 (4) (1966) 373–380.
[19] J.H. Ward, J. Am. Stat. Assoc. 58 (1963) 236–244.
[20] M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, R. Sindelka, R. Sjöback, B. Sjögreen, L. Strömbom, A. Ståhlberg, N. Zoric, Mol. Aspects Med. 27 (2006) 95–125.
[21] H. Motulsky, Intuituve Biostatistics, Oxford University Press, New York, 1995. ISBN: 0-19-50-8607-4.