Data Analysis Software

5

Michael W. Pfaffl, Jo Vandesompele and Mikael Kubista

in
Real-Time PCR: Current Technology and Applications, Publisher: Caister Academic Press. Editor: Julie Logan, Kirstin Edwards and Nick Saunders ISBN: 978-1-904455-39-4 (2009)

Abstract
Quantitative real-time RT-PCR (qRT-PCR) is widely and increasingly used in any kind of mRNA quantification, because of its high sensitivity, good reproducibility and wide dynamic quantification range. While qRT-PCR has a tremendous potential for analytical and quantitative applications, a comprehensive understanding of its underlying principles is important. Beside the classical RT-PCR parameters, e.g. primer design, RNA quality, RT and polymerase performances, the fidelity of the quantification process is highly dependent on a valid data analysis. This review will cover all aspects of data acquisition (trueness, reproducibility, and robustness), potentials in data modification and will focus particularly on relative quantification methods. Furthermore, useful bioinformatic, biostatistical as well as multidimensional expression software tools will be presented.

Introduction

Real-time qPCR has a tremendous potential for analytical applications in quantitative DNA and RNA analysis. To tap the full potential, a comprehensive understanding of its underlying quantification principles is important. For most researchers, high reaction fidelity in the nucleic acid (NA) quantification procedure performed is key to a quick result and a biological answer. This is associated with highly standardized pre-analytical steps, like tissue sampling and storage, NA extraction and storage, NA quantity and quality control, and through an optimized RT and/or PCR performance in terms of specificity, sensitivity, reproducibility, and robustness (Bustin, 2004). However, we should not forget that post-qPCR data processing can influence or even change the final result. Events such as qPCR data generation, acquisition, evaluation, calculation and statistical analysis are essential to interpret the biological significance of an experiment.

Data analysis in general is the act of transforming and interpreting data with the aim of extracting useful information and drawing correct conclusions. Depending on the type of data and the biological question, qPCR data analysis might include curve fitting algorithms, data processing, selecting or discarding certain data subsets based on specific pre-set criteria, transformation of logarithm quantification cycle (Cq) values to relative quantities, normalization, rescaling, and a final statistical test of the derived qPCR data. The ultimate goal for qPCR and qRT-PCR is to get a meaningful biological answer at the initial DNA or RNA level, respectively. To understand qPCR data analysis in detail, we have to split the procedure in to various levels, which will be discussed below.

Levels of data analyses in qPCR

On the first level, qPCR data analysis takes place within one raw fluorescence acquisition point. Multiple fluorescence measurements are averaged to a final raw fluorescence reading value per cycle. How many data points are averaged and how the averaging algorithm works is often hidden in the quantification software. Some instrument manufactures, like Bio-Rad (Hercules, CA,

USA), promote the software option of inspecting single cycle fluorescence readings to optimize the polymerase elongation time length to prevent primer-dimers or unwanted and unspecific PCR by-products, or to evaluate reaction kinetics.

The next level of data evaluation is performed within one sample qPCR run, where multiple single fluorescence data points are analysed (mostly between 30 and 50), measured at each cycle and constituting the so-called amplification plot. Most software applications do curve smoothing to show a 'perfect' real-time qPCR history in the amplification plot. We have to realize that the 'real' fluorescence readings, generated in the reaction cup are noisy and rough, and not as 'smooth and beautiful' shaped as shown in the plot. To retrieve the maximal information out of the shape of the amplification plot, rather than a single 'quantification point', more advanced fitting procedures must be applied. In the recent literature various models have been described, either using single (Liu *et al.*, 2002a; Larionov *et al.*, 2005; Ma *et al.*, 2006; Goll *et al.*, 2006) or multiple algorithms fitted to the amplification plot (Tichopad *et al.*, 2003, Wilhelm *et al.*, 2003). More useful information can be generated from these mathematical model variables concerning the background, fluorescence increase and plateau phase (Pfaffl, 2004). Valuable conclusions on qPCR fidelity can be drawn from this data. From the heights of the background fluorescence we can observe if too much reporter dye (SYBR Green I or labelled probe) is present in the reaction setup, whether primer or probe mismatches occur in early cycles and how comparable the samples are. The fluorescence increase gives essential information about amplification fidelity and PCR efficiency. Plateau height informs about the total amount of generated PCR product (Liu *et al.*, 2002b; Tichopad *et al.*, 2004). Some early software applications (LightCycler Software 2001) used amplitude normalization to generate 'identical plateau levels' and unify plateau height (Larionov *et al.*, 2005). The amplitude normalization is based on the suggestion that in an ideal PCR procedure the outcome is determined by the initial available PCR resources. This assumption is valid for ideal PCR but in practice it may not be true. To improve any quantification procedure the amplitude normalization can be implemented for each individual factor analysed (Larionov *et al.*, 2005).

Further curve smoothing algorithms are applied in melting curve analysis, to test the qPCR product-specificity and to show whether amplimer-dimers, splice variants, or unspecific PCR products have been formed. New approaches to increase the sensitivity of melting curve analysis are the application of innovative saturation fluorescence dyes, e.g. using LC Green (Wittwer *et al.*, 2003) in the LightCycler (Roche Diagnostics, Mannheim, Germany) or Rotor-Gene 6000 (Corbett Life Science, Sydney, Australia) and the use of 'High Resolution Melt' (HRM) curve analysis (Corbett Life Science). These methods are designed to give better-specificity testing and reliable genotyping.

Background subtraction

A prerequisite for cycle threshold (C_t) or crossing point (CP) determination is background fluorescence subtraction. However, accurate measurement of the level of background fluorescence can be a challenge. While a stable and constant background is ideal for subtraction, often the background appears noisy, with a rising or decreasing fluorescence level. Real-time PCR reactions with significant background fluorescence variations occur, caused by drift-ups and drift-downs over the course of the reaction (Wilhelm *et al.*, 2003; Larionov *et al.*, 2005). Averaging over a drifting background will result in an overestimation of variance and thus increase the threshold level (Livak, 1997; Rasmussen, 2001). Key questions surround how the software deals with the background data. Are the data shown real raw fluorescence data or are they already manipulated, e.g. through an additional ROX adjustment or amplitude normalization? Has curve smoothing been applied to the fluorescence data? Which kind of fluorescence background correction and/or subtraction was applied on the hardware?

Most real time platforms show pre-adjusted fluorescence data and in consequence pre-adjusted CP values. After doing an automatic background correction the CP values are deter-

mined by various methods, at a idealized 'constant level' of fluorescence. These constant threshold methods assume that all samples have the same synthesized DNA concentration at the threshold fluorescence. In the recent literature it has been reported that the effect of ROX correction, applied to either an early or late raw fluorescence data, resulted in higher intra-assay variation (up to 35%) and in pre-biased results, which clearly shift the biological answer (Goll *et al.*, 2006).

Background subtraction is a common step in PCR data processing. Often it requires operators involvement to choose between several available options, e.g. subtraction of a minimal value throughout the run, subtraction of an average over a certain pre-platform defined cycle range, or the assumption of different kinds of 'background trends'. To avoid operator involvement we always subtract the minimal value observed in the run. This option has a clear interpretation and works well. It is important that the baseline subtraction is performed after smoothing. So the noise potentially affecting minimal values has already been reduced before baseline subtraction (Larionov *et al.*, 2005).

The real challenge lies in comparing various experimental biological samples. It is not always straightforward to define a constant background for all samples within one run, and in bigger studies between different real-time qPCR runs. These sample-to-sample and run-to-run differences in variance and absolute fluorescence values are leading to the development of a new user friendly CP acquisition modus. Several mathematical models are established to determine the amplification rate, using four parametric logistic or sigmoidal models (Tichopad *et al.*, 2003; Liu *et al.* 2002a,b). These mathematical fit models can also be consulted to determine the optimal CP (Tichopad *et al.*, 2003). Comparable algorithms are already implemented in the LightCycler Software (Roche Diagnostics 2001), and in the Rotor-Gene 3000 and 6000 software (Corbett Life Science). Both types of platform algorithms are more or less 'independent' of the background level, calculate on the basis of raw fluorescence and implement the background data in the CP determination modus (Tichopad *et al.*, 2003; Wilhelm *et al.*, 2003).

Determination of the quantification point

The most important event in qPCR data analysis is to decide at which point of the amplification curve to take the one and only 'quantification point'. This point is well known as cycle threshold (C_t), crossing point (CP) or take-off-point (TOP), depending on the platform used and the analysis software. The importance of the model and algorithm used to get the right 'quantification point' (CP value) is often underestimated (Pfaffl, 2004). All further real-time qPCR data processing events are based on the CP. CP value determination should be done on raw florescence reading data, in a highly standardized reproducible way and independent of any interfering side effects (e.g. ROX adjustment or amplitude normalization). In real-time PCR data analysis, various approaches are used to generate the 'quantification point'. First and widely distributed is the 'cycle threshold method' (C_t method), second is the 'second derivative maximum method' (SDM), and more recently the method of 'non-linear regression analysis' (NLR) and the 'CalQPlex' algorithm (Eppendorf, Germany) was introduced.

Cycle threshold method

For most researchers the C_t method is currently the gold standard. Some real-time cycler software packages offer curve-smoothing and normalization, but the basic C_t method algorithm remains unchanged (Goll *et al.*, 2006). Threshold fluorescence is calculated from the initial cycles, and in each reaction the C_t value is defined by the fractional cycle at which the fluorescence intensity equals the prior set threshold fluorescence. This method is based on an assumption of 'equal' PCR efficiency in all reactions, and accuracy may be compromised if this condition is not met.

The threshold level can be calculated by fitting the intersecting line upon the ten-times value of ground fluorescence standard deviation. This background acquisition mode can be easily automated and is therefore stable and robust (Livak *et al.*, 1997). In the 'fit point method' the user has to discard the uninformative background points, exclude the plateau values by entering

the number of log-linear points, and then fit a log-line to the linear portion of the amplification curve. These log-lines are extrapolated back to a common threshold line and the intersection of the two lines provides the C_t value. The strength of this method is that it is extremely robust. The weakness is that it is not easily automated and so requires a lot of user interaction, which can be arbitrary (Rasmussen, 2001; LightCycler Software, 2001).

Sample-to-sample changes in PCR efficiency, caused by RT and PCR inhibitors or enhancers, can question the C_t method and the derived results. As such, the shapes of fluorescence amplification curves differ due to the background level (noisy, constant or increasing), the take off point (early or late), the steepness (good or bad PCR efficiency), the change-over to the plateau phase (quick or steady), and in the appearance of the PCR plateau (constant, in- or decreasing trend) (Tichopad *et al.*, 2003, 2004). PCR amplification efficiency has the largest impact on amplification plot history. Therefore determination of the threshold level, the threshold cycle and in consequence the accuracy of the quantification results are strongly influenced by the amplification efficiency (Pfaffl, 2004).

Second derivative maximum method

Applying the *'second derivative maximum method'* the quantification point is automatically identified and measured at the maximum acceleration of fluorescence (Rasmussen, 2001; Tichopad *et al.*, 2003). The exact mathematical algorithm applied in the LightCycler software (Roche Diagnostics) is still unpublished, but is comparable to a logistic or polynomial fit (Tichopad *et al.*, 2003). Corbett Life Science cyclers use a *'comparative quantification'* method, where the TOP is calculated on the basis of a sigmoidal model. Both algorithms, the sigmoidal and polynomial curve models, work well with high significance ($P < 0.001$) and coefficient of correlation ($r > 0.99$), as determined in various qPCR experiments with different biological questions (Liu and Saint, 2002a,b; Tichopad *et al.*, 2003, 2004; Rutledge, 2004). Sigmoidal exponential curve fitting was shown to be the most precise method and increases the accuracy and precision of the CP measurements (Wilhelm *et al.*, 2003).

Non-linear regression analysis

Non-linear regression analysis (NLR) has been suggested as an alternative to the C_t method for absolute quantitation (Goll *et al.*, 2006). The advantages of NLR are that the individual sample efficiency is simulated by the model and that absolute quantitation is possible without a standard curve, releasing reaction wells for unknown samples. NLR can be fully automated and may be a powerful tool for analysis of fluorescence data from qPCR experiments. The unfavourable signal to noise ratio of the probe-based assays does not impair NLR analysis. The versatility of NLR depends on the precision needed but if it can be applied, this analysis method may save both time and resources in the laboratory. Further work is needed to improve the precision of the fluorescence copy number conversion factor in order to reduce the bias of NLR observed in this study. However, it is indeed possible to obtain absolute quantitation from real-time qPCR data without a standard curve. In an optimized assay, however, the C_t method remains the gold standard due to the inherent errors of the multiple estimates used in NLR (Goll *et al.*, 2006).

Validity of the crossing point

After the CP value has been determined by one of the above methods, the data point generated has to be validated. Several questions arise and these may be addressed using the assay standard, references and/or controls: is the generated CP data point valuable and reliable?; is the CP sufficiently different from my defined negative standard or non-template-control (NTC) value?; is the outcome comparable with the positive control CP?; is my data point within my defined quantification range, or is the data point out of range so that it must be discarded?.

Here a strict comparison and decision process between analysed CP values and given standards, references and controls is essential and assumed. Especially in so-called 'absolute' quantitative qPCR applications where the quantified NA amount will result in further decision making. A valid and reliable NA quantification is a prerequisite, e.g. in clinical diagnostics, quantitative microbiology and in analysis of genetically modified organisms (GMO) (Burns *et al.*, 2004; Paoletti and Mazzara, 2005). In an 'absolute'

quantification approach, the PCR signal is related to input copy number using a calibration curve (Bustin, 2000; Pfaffl and Hageleit, 2001; Fronhoffs *et al.*, 2002). The calibration curve itself can be derived from diluted PCR products, recombinant DNA or RNA, linearized plasmids, or spiked tissue samples. The reliability of such a an 'absolute' real-time RT-PCR assay depends on the condition of *'identical'* amplification and reverse transcription (if applicable) efficiencies for both the native mRNA target and the target RNA or DNA used in the calibration curve (Souaze *et al.*, 1996; Pfaffl, 2001). The efficiency evaluation remains one of the essential considerations in qPCR and efficiency correction is necessary in real-time PCR based gene quantification (Rasmussen, 2001; Tichopad *et al.*, 2003).

The synonym *'absolute'* quantification is misleading, because the quantification is in fact performed *'relative'* to the calibration curve. Clearly, all quantifications should be considered relative (Bustin, 2004). mRNA copy numbers should generally be corrected for differences in input and quality, using various standardization methods, such as mass of tissue, amount of total RNA or DNA (e.g. a verified internal standard), a defined amount of cells, or compared to a reference gene copy number (e.g. ribosomal RNA, or other commonly used reference genes). The current gold standard for gene expression analysis is the use of multiple carefully selected reference genes (Vandesompele *et al.*, 2002). The *'absolute'* quantification strategy using various calibration curves and applications has been summarized elsewhere in detail (Pfaffl and Hageleit; 2001; Donald *et al.*, 2005; Lai *et al.*, 2005).

Minimum performance requirements in genetically modified organism (GMO) analysis

The use of a validated method of analysis for 'absolute' quantification of GMOs is required for authorization of GM food or feed products within the European Union (EU). Various EU Commission Regulations (EC) 1829/2003, 641/2004, 882/2004, and a publication by the European Network of GMO Laboratories (ENGL) describe the validation procedure step-by-step. The assessment of method performance and its suitability for GMO testing consists of two independent steps. The purpose of the first phase is to verify that the method performance fulfils the requirements to enter the validation process. Acceptance criteria, like applicability, practicability,-specificity, dynamic range, accuracy, amplification efficiency, correlation of the calibration curve, reproducibility, limit of detection (LOD), limit of quantification (LOQ), and robustness are evaluated and strict cut off criteria are set. The accuracy of the qPCR method must be within ± 25%, the standard deviation should not exceed ± 25% over the whole range of quantification, and the real-time PCR amplification efficiency must be in the range of 90–110% (Paoletti and Mazzara, 2005). If all criteria fulfil the minimal criteria, an inter-laboratory collaborative trial is launched to validate the method in various countries. In the second phase evaluation of method performance characteristics must be done at the end of the validation study, by analysis of collaborative trial results. The purpose of the second phase is to verify that the method performance, as demonstrated by the results of the collaborative trial, confirms the characteristics indicated at the moment of submission in a multi-laboratory setting, and fulfils the requirements to be considered fit for regulatory purpose (Paoletti and Mazzara, 2005).

The GMO validation and strict exclusion criteria represent a good example and basis for further validation strategies in other areas of NA quantification. In clinical diagnostics, quantitative virology and microbiology such stringent and international accepted assay validation premises are urgently needed. They should be established in the near future and disseminated to analytical laboratories as 'quantitative PCR Good Laboratory Practice' (qPCR GLP).

Relative qPCR data analysis

Besides the 'absolute' gene quantification strategy, the relative expression strategy is commonly used by the academic research community. Relative or comparative gene expression analysis compares a gene-of-interest (GOI) in relation to a reference-gene (RG). Today various mathematical models are published and established to calculate the relative expression ratio (R), based on the comparison of the distinct cycle differences, called delta-CP model. In most applications the relative quantification is based on the *'delta-CP'* (ΔCP,

Wittwer *et al.*, 2001) or *'delta-delta-CP'* (ΔΔCP) values, described by Livak and Schmittgen (2001).

Screening the recent literature applying relative quantification and using the ΔΔCP method (equation 1), the majority of researchers assume an optimal doubling during each PCR cycle with an amplification rate of 100% (i.e. base of exponential amplification of 2) (Livak, 1997, Livak and Schmittgen, 2001).

$$R = 2^{-[\Delta CP\text{sample} - \Delta CP\text{control}]} \quad (1)$$

This is acceptable for a first approximation of the rough expression ratio (*R*) between the expressed genes. To obtain reliable relative expression data, more advanced and efficiency (E)-corrected models are published. The assessment of the sample-specific amplification efficiencies must be carried out for each gene and tissue analysed. Models are based on one analysed sample (equations 2 and 3) (Pfaffl, 2001; LightCycler Relative Quantification Software, 2001) or on a group of various samples (equation 4) (Pfaffl *et al.*, 2002).

$$R = \frac{(E_{target})^{\Delta CP_{target}(control-sample)}}{(E_{ref})^{\Delta CP_{ref}(control-sample)}} \quad (2)$$

$$R = \frac{(E_{ref})^{CP_{sample}}}{(E_{target})^{CP_{sample}}} \div \frac{(E_{ref})^{CP_{calibrator}}}{(E_{target})^{CP_{calibrator}}} \quad (3)$$

$$R = \frac{(E_{target})^{\Delta CP_{target}(MEAN\ control-MEAN\ sample)}}{(E_{ref})^{\Delta CP_{ref}(MEAN\ control-MEAN\ sample)}} \quad (4)$$

To be more stable and reliable when applying relative quantification strategy, new approaches using multiple reference genes were introduced (Vandesompele *et al.*, 2002). To base the relative quantification not only on one reference-gene, multi reference-gene normalization software was developed, e.g. Relative Expression Software Tool (REST-384; Pfaffl *et al.*, 2002), qBase (Hellemans *et al.*, 2007), and qPCR-DAMS (Jin *et al.*, 2006). These software applications apply efficiency corrected calculation models, based on multiple sample and on multiple reference-genes (RG index or normalization factor) consisting at least of three stably expressed reference-genes, as shown in equation 5 (Pfaffl *et al.*, 2004).

$$R = \frac{(E_{target})^{\Delta CP_{target}(MEAN\ control-MEAN\ sample)}}{(E_{RG\ index})^{\Delta CP_{RG\ index}(MEAN\ control-MEAN\ sample)}} \quad (5)$$

Some helpful software tools and Excel spreadsheets (Microsoft Corporation, USA) are available to correct for qPCR efficiency differences. The LightCycler Relative Expression Software (2001), Q-Gene (Muller *et al.* 2002), qBase (Hellemans *et al.*, 2006), SoFar (Wilhelm *et al.*, 2003), DART (Peirson *et al.*, 2003), qPCR-DAMS (Jin *et al.*, 2007) and various REST software applications (Pfaffl *et al.* 2002) allow the evaluation of amplification efficiency plots. In most of the applications a triplicate determination of real-time PCR efficiency for every sample is recommended. Therefore, PCR efficiency corrections should be included in the relative quantification procedure. Therefore a stand-alone software tool, named KEST 2008, was developed to implement the single-run PCR efficiency correction in relative quantification models (Pfaffl, Technical University Munich and Corbert Life Science; download: http://rest.gene-quantification.info/).

Complex qPCR data analysis

The next data analysis level is the comparison between different qPCR runs. The difference between organisms and/or treatment groups, e.g. treatment groups, sex, age, or along a time course, is the real biological question. Various non-treated control or reference treated groups must be implemented in the gene expression comparison. Further multiple inter-run calibrators have to be included in each run and on each plate to allow correction between different runs and repeats. The final goal is to make several runs fully comparable. At this point complex qPCR analysis software and databases are necessary, like qBase or qPCR-DAMS, to deal with large amounts of real-time data.

The next focus is on multiple target genes or multiple transcript analyses within one biological sample, and how to make them comparable.

Commonly this is done by using normalization strategies based on one or more reference-genes (Vandesompele *et al.*, 2002; Pfaffl *et al.*, 2004a). The quality of normalized quantitative expression data cannot be better than the quality of the normalizer itself. Any variation in the normalizer will obscure real changes and produce artefactual changes (Bustin *et al.*, 2002, 2005). These normalizing strategies are summarized and described in detail in published reviews (Huggett *et al.*, 2005, Wong and Medrano, 2005) or in Chapter 4.

Comparison between biological samples represent the next level of complexity. qPCR expression data within different organs, tissue, or types of cell culture can be analysed, using multi dimensional data analysis software applications, e.g. GenEx software. The results can be combined to give multi-tissue expression results or to provide a complex system biological overview. This represents a very high level of complexity and a lot of data conversion, corrections and normalization must be applied to make the real biological expression events visible and comparable.

The flowchart in real-time qPCR data analysis

The following flowchart of the key steps in the qPCR data analysis process, are necessary to give reliable and meaningful biological conclusions. Most of the steps are hidden in fixed software algorithms or models created and defined by the real-time PCR platform manufacturers. But some of the key steps are open and can be influenced by the researcher. The essential key steps are summarized below to indicate how they are influenced by given software applications.

(1) Scope of qPCR application

The researcher has to define the applicability, practicability and type of quantification (absolute or relative). One or more internal or external control samples, negative- and positive-controls, non-template controls (NTC), calibrators and/or reference samples, must be defined. At this stage the scientist has to define minimal and maximal detection limits, minimal and maximal quantification limits, quantification ranges and set unique characteristics for further data evaluation. These data can only be recruited from earlier experiments using identical chemistry, consumables and real-time PCR platforms. As an example the highly defined 'Minimum Performance Requirements for Analytical Methods of GMO Testing' can be mentioned as a gold standard for test validation (Paoletti and Mazzara, 2005).

(2) Data transfer and profiling

When the raw data is exported from the real-time cycler optical unit and imported to the calculation unit, the system software should (automatically) verify the completeness and integrity of transferred data sets, e.g. missing data or possible outliers. At this stage the auditor cannot perform any relevancy checks on the data sets. Missing or discarded data points must be reported according to the set of qPCR GLP regulations.

(3) Initial data conversion

The fluorescence data are modified and then presented for further calculations. These modifications are done automatically and software specific and cannot be influenced by the researcher. The following initial data conversions are possible: averaging from multiple measurements, corrections with other fluorescence channels, e.g. ROX correction, curve smoothing, amplitude normalization, deletion of outliers, and often an automatic fluorescence background adjustment.

(4) Data analysis

The data sets should be analysed according to the needs of the investigator and the given biological question, e.g. for further quality and integrity tests. Further data processing can be done to the 'raw fluorescence data':

- Data point evaluation, deletion of outliers, management of missing data, amplification plot curve smoothing, and amplitude normalization can be performed.
- A reliable and reproducible background subtraction should be performed, automatically or manually, to correct for differing background fluorescence levels. At this level a lot of different methods and algorithms can be used, mostly pre-set by the platform manufactures (C_t method, SDM, NLR).
- The threshold line should be drawn, automatically or manually, to determine the right

cycle threshold value (C_t, CP, or TOP).

- On some platforms mathematical models, e.g. sigmoidal or logistic models, help the researcher to find the 'right quantification point' and calculate the amplification efficiency in each analytical sample (sigmoidal fit, logistic fit, SDM, NLR, CalQPlex).

(5) Verification of CP data

At this stage the CP data must be verified according to the minimal requirements set for the experiment. These are as discussed in '(1) Scope of qPCR application'. At this point the melting curve analysis may be checked to verify the correct melting point, purity and integrity of amplified real-time PCR product.

(6) Advanced data analysis

Once the CP has been calculated more advanced data analysis can be performed. This depends upon the experimental setup or the biological question. Either an absolute or relative/comparative quantification strategy can be chosen. Data can be analysed using various macros, which will be discussed later in this chapter (qBase, REST, GenEx, qGene, SoFar, etc.). CP data can be further converted by normalization using one or more reference genes or internal, exogenous or artificial controls. Various software applications are distributed within the qPCR community, e.g. geNorm, BestKeeper, REST, Normfinder, Global Pattern Recognition, etc. Advantages and disadvantages of such applications are discussed in Chapter 4.

(7) Statistical data analysis

The statistical analysis of quantitative real-time PCR data is currently a topic with seven secrets! The group comparison has to be done with robust statistical methods, to get scientifically proven and reliable results. However, after several steps of normalization, correction, and adjustment using multiple references and calibrators, which statistical tool is reliable? In the literature only a few publications deal with the statistical analysis of expression data generated by real-time PCR and how they might be analysed in a reliable and statistically correct way (Pfaffl *et al.*, 2002; Muller *et al.*, 2002; Peirson *et al.*, 2003; Yuan *et al.*, 2006; Gilsbach *et al.*, 2006).

(8) Reporting and documentation

Research findings have to be summarized and reported. Results have to be documented and the type of report most suitable to describe the outcome of the applied experiment must be selected. Numbers of molecules, masses or molarities are reported in an absolute quantification, or the fold regulation in a relative quantification. Research findings have to be documented in the correct format for qualified scientific reports, original papers, spreadsheets, or flowcharts, to name a few. Clear and reliable documentation is essential to support the researchers experimental results, independent of whether they are positive or negative, significant or non-significant.

(9) Data visualization

A very important final part of data analysis is the clear data visualization and presentation in a scientific article. The reader should clearly see the experimental result, and the statistical significance without any additional confusing data. Graphical visualization, plot interpretation and the physiological conclusion should be straight forward. Error bars should always be provided.

qPCR data analysis software applications

At the time of writing qPCR analysis software has tended to be eclipsed by the production of new real-time PCR platforms and detection formats. Hardware and chemistry have developed much faster than detection and analysis software. However, we are now at the beginning of a new post qPCR data processing area. The challenge is the development of simple but reliable gene expression analysis and quantification software. The development of 'one-fits-all' software is the goal and this would appear to be the optimal solution. However, can we implement various detection chemistries bearing varying background levels and fluorescence acquisition modes in one analytical software? Optimized analysis models for each real-time platform and for each chemistry might be more effective.

In research and in clinical diagnostics, real-time qRT-PCR is the method of choice for expression profiling. However, accurate and straightforward mathematical and statistical analysis of qPCR data and management of

growing data sets have become the major hurdles to effective implementation. Nowadays up to 96-well applications are the standard in research, but in the near future high-throughput 384-well applications will generate huge amounts of qPCR data. qPCR data needs to be grouped, standardized, normalized, and documented by intelligent software applications (Hellemans *et al.*, 2007). Real-time qPCR data should be analysed according to automated statistical methods, e.g. kinetic outlier detection (KOD), to detect outliers and samples with dissimilar efficiencies (Bar *et al.*, 2003; Burns *et al.* 2005). Often the statistical data analysis is performed on the basis of classical parametric tests, such as analysis of variance (ANOVA) or *t*-tests. Parametric tests depend on assumptions, such as normality of C_t value distributions, the validity of which is often unknown to the researchers (Pfaffl *et al.*, 2002; Muller *et al.*, 2002). In relative gene expression analysis, the quantities are derived from ratios where variances can be high. Furthermore, normal distributions might not be expected as such. Logarithmic transformation is a prerequisite for gene expression analysis. It is unclear how a parametric test could be optimally constructed (Pfaffl *et al.*, 2002). New analysis formats, using new algorithms for data analysis and ideas for reliable statistical analysis of real-time qPCR data are urgently required.

To show the state-of-the art in qPCR analysis software the available applications are briefly described:

LightCycler Relative Quantification Software

The first commercially available software for relative quantification was the LightCycler Relative Quantification Software (Roche Diagnostics, 2001). It allows calculation and comparison of the relative quantification results for triplicates of a target versus a calibrator gene. Target genes are corrected via a reference-gene and calculated on the basis of the median value CP of the performed reaction triplets. In newer software versions more repeats can be compared. Real-time PCR efficiency correction is possible within the software and is calculated from the calibration curve slope. A given correction factor and a multiplication factor, which are provided in the product-specific applications by Roche Diagnostics (2001), have to be entered in the calculation process. Data output is done via a relative expression value, but the software application still lacks a reliable statistical analysis or confidence range of the analysed data (Relative Quantification Software 4.0, Roche Diagnostics, 2004).

REST

The Relative Expression Software Tool (REST) is Microsoft Excel-based and programmed in the Visual Basic Application (VBA), to compare several expressed genes. The basic version compares two treatment groups, with multiple data points in sample group versus control group, and calculates the relative expression ratio between them. The mathematical model used is published and based on the mean CP deviation between sample and control group of target genes, normalized by the mean crossing point deviation of one reference gene, as shown in equation 4. Furthermore, an efficiency correction can be performed, either based on the dilution method or an optimal efficiency of $E = 2.0$ is assumed. The big advantage of REST is the subsequent statistical test of the analysed CP values by a pairwise fixed reallocation randomization test (Pfaffl *et al.*, 2002). Permutation or randomization tests are a useful alternative to better known parametric tests for analysing experimental data (Manly, 1997; Horgan and Rouault, 2000). They have the advantage of making no distributional assumptions about the data, while remaining as powerful as conventional statistical tests. Randomization tests are based on what we know to be true: that treatments were randomly allocated. The randomization test repeatedly and randomly reallocates, at least 2000 times, the observed CP values to the two groups and notes the apparent effect on CP difference each time. The proportion of times random allocation produces a more significant result is reported. The REST software package makes full use of the advantages of a randomization test. In the applied two-sided pairwise fixed reallocation randomization test for each sample, the CP values for reference genes and target genes are jointly reallocated to control and sample groups (= pairwise fixed reallocation), and the expression ratios are calculated on the basis of the mean values. In practice, it is

impractical to examine all possible allocations of data to treatment groups, and a random sample is drawn. If at least 2000 or more randomizations are taken, a good estimate of *P*-value (standard error < 0.005 at $P = 0.05$) is obtained. Randomization tests with a pairwise reallocation are seen as the most appropriate approach for this type of application. New REST versions were developed and released in 2005, calculating a geometrically averaged RG index, according to earlier publications (Vandesompele *et al.*, 2002, Pfaffl *et al.*, 2004) analysing 15 target and reference genes (REST-384). Specialized REST-Multiple Condition Solver versions can compare six treatment group with one non-treated control (REST-MCS). During the last years two new REST tools, REST 2005 and REST 2008, were developed by Pfaffl and co-workers at the Technical University Munich and Corbert Life Science (Sydney, Australia). Beside the multiple reference gene normalization, in REST 2008 the single run-efficiency was implemented as a new tool. A further improved statistical testing method on the basis of a bootstrapping method is applied (http://rest.gene-quantification.info/).

Q-Gene

Q-Gene manages and expedites the planning, performance, and evaluation of quantitative real-time PCR experiments (Muller *et al.*, 2002). The software is able to perform a statistical test on the real-time C_t data. Efficiency correction according to the dilution method is also possible. The Q-Gene software application is a tool to cope with complex quantitative real-time PCR experiments at a high-throughput scale (96-well and 384-well format) and considerably expedites and rationalizes the experimental setup, data analysis, and data management while ensuring the highest reproducibility. The expression results are presented by graphical presentation. The Q-Gene Statistics Add-In is a collection of several VBA programs for the rapid and menu-guided performance of frequently used parametric and non-parametric statistical tests. To assess the level of significance between any two groups expression values, it is possible to perform a paired or an unpaired Student's test, a Mann–Whitney *U*-test, or Wilcoxon signed-rank test. In addition, the Pearson's correlation analysis can be applied between two matched groups of expression values. Furthermore, all statistical programs calculate the mean values of both groups analysed and their difference in per cent (Muller *et al.*, 2002).

qBase

The comprehensive software application qBase was recently developed as a generalized solution to accommodate virtually all relative quantification setups (Hellemans *et al.*, 2007). qBase is an Excel based tool for the management and automatic analysis of real-time quantitative PCR data (http://medgen.ugent.be/qbase/). It employs a proven delta-C_t quantification model with efficiency correction, multiple reference gene normalization and accurate error propagation along all calculations. The qBase browser allows data storage and annotation while keeping track of all real-time PCR runs by hierarchically organizing data into projects, experiments, and runs. It is compatible with the export files from many currently available PCR instruments and provides easy access to all data types (both raw and processed). The qBase analyser contains an easy plate editor, performs quality control, converts CP values into normalized and rescaled quantities with proper error propagation, and displays results in both tabulated and graphical forms. A big advantage of the program is that it does not limit the number of samples or genes and replicates, and allows data from multiple runs to be combined and processed together (Hellemans *et al.*, 2007). The possibility to use up to five reference genes allows reliable and robust normalization of gene expression levels, on the basis of the *geNorm* normalization procedure (Vandesompele *et al.*, 2002). qBase allows easy exchange of data between users, and exports tabulated data for further statistical analyses using dedicated software. qBase has been phased out and is now available as professional real-time PCR data analysis software qBase*Plus* from Biogazelle (http://www.biogazelle.com).

SoFar

The algorithms implemented in SoFar (distributed by Metralabs, Germany) allow fully automatic analysis of real-time PCR data obtained with a LightCycler (Roche Diagnostics)

instrument. The software yields results with considerably increased precision and accuracy of real-time quantification. This is achieved mainly by the correction of amplification-independent fluorescence signal trends and a robust fit of the exponential phase of the signal curves. The melting curve data are corrected for signal changes not due to the melting process and are smoothed by fitting cubic splines. Therefore, sensitivity, resolution, and accuracy of melting curve analyses are improved (Wilhelm *et al.*, 2003).

qCalculator

The qCalculator is a VBA-based program that enables the calculation of relative expression data derived from quantitative real-time PCR experiments (Gilsbach *et al.*, 2006). qCalculator enables a flexible calculation of 32 samples (in triplicate) for 20 experiments based on up to 10 reference genes. Each gene can act as target or reference gene without rearrangement of the data, as in REST-384 software. Calculation of the relative expression is done with efficiency correction (according to Pfaffl 2001), or without, on the basis of the delta-delta-C_t efficiency correction approach (Livak and Schmittgen, 2001). Samples for relative comparison can be freely selected and changed at any point of the analysis. Efficiency calculation and absolute quantification is done on the basis of standards curves. Extreme values like outliers can be excluded at any point of the qCalculator analysis procedure. Any changes in the parameters set leads to the immediate recalculation of data. qCalculator includes an automatic statistical comparison of different choices of reference genes. Data output is shown as arithmetic mean (mean ± s.e.m.) of *n* experiments. The two-tailed Student's t-test can be used to compare results under the precondition that their standard error of mean (s.e.m.) values do not differ significantly. The *F*-test is used to verify this requirement. GraphPad Prism 4.0 (GraphPad Software, San Diego, USA) is used to test the normal distribution of the results using the Kolmogorov–Smirnov test (KS test) and to calculate the associated *P* values according to Dallal and Wilkinson (1986). Results and error propagation is based on log2-values and are summarized and graphically displayed. The qCalculator manual and software can be downloaded from the homepage: www.pharmakologie.uni-bonn.de/frames/index_fr.htm

Dart-PCR

Dart-PCR (data analysis for real-time PCR) provides a simple means of analysing real-time PCR experiments from the raw fluorescence data (Peirson *et al.*, 2003). It allows an automatic calculation of amplification kinetics, as well as performing the subsequent calculations for relative quantification and calculation of assay variability. Amplification efficiencies are also tested to detect anomalous samples within groups (outlier detection) and differences between experimental groups (amplification equivalence). Data handling is simplified by automating all calculations in an Excel worksheet and this enables the rapid calculation of C_t values, amplification rate and resulting starting values, along with the associated error, from raw data. Differences in amplification efficiency are assessed using one-way analysis of variance (ANOVA), based upon the null hypotheses, that amplification rate is comparable within sample groups (outlier detection) and that amplification efficiency is comparable between sample groups (amplification equivalence) (Peirson *et al.*, 2003).

Gene Expression Macro

The Gene Expression Macro (Bio-Rad) is a simple tool for calculating relative expression values from real-time PCR data generated by the iCycler, iQ or MyiQ systems (Bio-Rad). The gene Expression Macro runs using Excel and contains specialized data analysis functions. The use of this macro can save valuable time by employing standard methods of relative gene expression analysis in predesigned, easy-to-use spreadsheets. The calculations in the spreadsheet are outlined in the methodology tab of the macro. They are derived from the algorithms outlined by geNorm (Vandesompele *et al.*, 2002). The calculations allow the user to analyse results using multiple reference genes. This often results in more consistent results and more confidence in the conclusions drawn from experiments. As demonstrated in the example calculations provided on the on Gene Expression Macro web page the algorithm does take PCR efficiencies of each assay into account. In general, all of the methods included

build upon on the delta-delta-C_t calculation (Livak and Schmittgen, 2001) and assume optimal doubling in each cycle and PCR efficiencies of two. Further an efficiency correction according to Pfaffl (2001) can be implemented in the calculation process. To exactly reproduce the results you would obtain using the 'Pfaffl model' you must define a control sample (calibrator sample in the publication). The Gene Expression Macro will express all values relative to this sample. The control sample is given a value of 1. Relative expression data output is done in a graph using mean expression values and standard deviations. No statistical tool for qPCR data verification is included in the Gene Expression Macro.

qPCR DAMS

Recently a new databank named qPCR-DAMS (Quantitative PCR Data Analysis and Management System) was developed (Jin *et al.*, 2006). It is a database management system implemented on MS Access 2003 (Microsoft Corporation) and VBA. The database tool works with integrated mathematical procedures and is designed to analyse, manage, and store relative and absolute quantitative real-time PCR data. The system consists of five independent blocks: three blocks (gene, plate, and experiment) for inputting, storing and describing raw data, and two blocks (view data and process data) for checking, evaluating, and processing data. Users are allowed to choose among four basic outputs: (1) ratio relative quantification, (2) absolute level, (3) normalized absolute expression, and (4) ratio absolute quantification. A further two advanced options are available within the software package: (5) multiple reference relative quantification and (6) multiple references absolute quantification. The coefficient of variation is monitored at each step during data processing and the accuracy is further improved by an easy data tracking and display system. Unfortunately, qPCR-DAMS lacks a statistical tool to verify the qPCR data. In summary, qPCR-DAMS is a handy and easy-to-use tool to host, manage, evaluate, process and store data from both relative and absolute quantitative real-time PCR. The qPCR-DAMS software and online manual are available free for academic use at http://www.cvm.okstate.edu/research/Facilities/LungBiologyLab/.

Expression profiling

In most studies the expression of a key marker gene in test samples is compared. A number of biological repeats are taken of each sample, and then robust statistical methods are used to infer if there is significant difference between the populations from which the samples where drawn. One may also compare a test sample with standards to estimate, for example, the viral load of a patient for clinical decisions. The approach, however, is only useful when there is a dominant marker gene whose expression signifies the biological condition or a pathogenic RNA. This is not the case for complex diseases, which are caused by a combination of factors and cannot be monitored through the expression of a single gene. Such diseases can only be studied in a meaningful way by expression profiling. In profiling experiments the expression of many genes, optionally all genes, are measured and the disease condition is identified from the pattern of the expressed genes (D'haeseleer, 2005). In some situations the detailed expression pattern can also have prognostic value. Traditionally expression profiles are measured using microarrays, by which the expression of all genes can be assessed in a single experiment. However, the quality of microarray expression data is usually not good enough for detailed classification and accurate prognosis. Real-time PCR is more sensitive, has a wider dynamic range and better reproducibility. Cost per test is also lower, and samples can be measured for the same money. The drawback is that it is not feasible to measure the expression of all the genes. However, this is rarely limiting, because only a fraction of genes have their expression significantly affected in most disease conditions. The expression of most genes show no or little disease dependence that can be discerned from noise and fluctuations caused by other factors. The preferred approach is to use microarrays to identify potentially informative genes, and then validate them by real-time PCR, before setting up a real-time PCR expression profiling study based on those that pass validation. Measuring the expression of only informative genes opens the possibility of data normalization by autoscaling (see below), which makes analysis very robust (Kubista *et al.*, 2006).

Data pre-processing

Real-time PCR raw data are expressed as C_t values. Raw C_t values are not readily comparable for many reasons and may not accurately reflect the initial number of target molecules in the sample. The C_t value depends on instrument factors (gain, filters, light source), detection chemistry (dye, probe), amplicon (longer amplicons bind more dye), and PCR efficiency, which may vary among assays as well as among samples. The C_t value may also be compromised by primer–dimer formation. Further, C_t values reflect the number of molecular copies of cDNAs in the samples, while we are more interested in their concentration. The raw C_t values require reliable normalization. In regular real-time PCR studies the expression of the reporter gene is normalized with the expression of properly selected reference(s) genes. Such normalization also compensates for non-gene-specific variations in extraction efficiency, reverse transcription yield and PCR inhibition. The proper way to pre-treat real-time PCR data for expression profiling is as follows:

1 Compensate for any variation between runs. If data were collected over several runs, C_t values should be related to a common reference sample included in all runs.
2 Correct any primer–dimer contributions. C_t values due to primer–dimer signals should be set to the same value that is higher than any C_t value reflecting product formation.
3 Correct for variations in PCR efficiency between assays (based on standard curve).
4 Normalize for variations in PCR efficiency between samples (normalize with spike).
5 Normalize to sample amount. The copy numbers should be normalized to the volume of biological material, e.g. serum volume or number of cells or alternatively, normalization to the amount of total RNA.
6 Optionally normalize to reference genes. If there is reason to believe there are sample-to-sample variations in extraction efficiency, reverse transcription or PCR inhibitors, these factors can be accounted for by normalization with a stably expressed gene or genes that are not affected under the conditions studied. Such genes may not always be available in an experimental design (Sindelka *et al.*, 2006). To help identifying suitable reference genes, panels of reference gene candidates have been made available for the most common species (www.tataa.com/referencepanels.htm and http://www.primerdesign.co.uk/geNorm.asp).
7 Calculate relative quantities.
8 Convert to log scale. Regulation of expression typically results in fold changes and data should be analysed in logarithmic scale. By convention base 2 is used.
9 Mean centre. Absolute gene expression levels are difficult to measure and often also to interpret. More robust and reliable are relative expression levels. In expression profiling expression levels are often expressed relative to the mean expression of each gene. This is called mean centring of data. Mean centred data has for every gene average expression of zero. Mean centred data are often used to classify samples.
10 Auto-scale. Often relative changes in expression levels are more informative than absolute changes. These are reflected by autoscaled data. Data are autoscaled by first mean centre them (see above) and then dividing with the standard deviation of the expression of each gene. Hence, auto-scaled gene expression data have mean expression of zero and a standard deviation of 1. Autoscaled data are used to classify genes.

Preprocessing is readily performed using software such as GenEx (www.multid.se) and in many cases the preprocessing scheme can be simplified. If there are no primer–dimer signals, the same PCR efficiency and assay sensitivity can be assumed, and if the samples were based on the same amount of material, steps 2–9 cancel and the measured C_t values can be auto-scaled directly. Another difference between expression profiling and comparisons based on the expression of a single gene is that in expression profiling biological repeats are not averaged. Treating them separately it is possible to compare the variation among different biological repeats to tell whether the samples are different or not.

The scatter plot

The minimum number of genes in an expression profiling experiment is two. For reciprocally regulated marker genes the profile based on just two genes can be very sensitive. A classical example is the test for lymphoma based on measuring the relative expression of the kappa and lambda variants of the immunoglobulin light chain. Each β-cell decides during maturation which of the light chain variants to express. In healthy humans 40% of the cells express the kappa variant while 60% express the lambda variant. Like all cancers, lymphoma is a clonal disease, resulting in a large number of identical ß cells originating from the same original cancer cell. All these cells will express either the kappa or the lambda variant, which will result in a kappa to lambda expression ratio that deviates from 60:40. A convenient way to visualize such data is in a scatter plot. If the samples are not normalized to the same sample amount, negative samples will fall on a straight line corresponding to the 60:40 expression ratio (Fig. 5.1). Below and above that line are positive samples with kappa and lambda clonality, respectively (Stahlberg *et al.*, 2003).

The expression of three genes can be visualized in three-dimensional scatter plots, where again similar samples will be close to each other. The idea of clustering samples in a space made up of expression levels can be generalized to any number of genes. Spaces of higher dimensionality than three cannot be visualized by conventional means, and a number of powerful clustering tools have been developed to analyse such spaces and to present the information. Some of the most common are described below.

Principal component analysis

The information in a multidimensional space can be collected in a space of lower dimension by means of so called principal components (PC). The PCs are defined as vectors that account for most of the variation in the data in the multidimensional space. The first PC is the best linear fit to the data in the multidimensional space; the second PC is orthogonal to the first and accounts for most of the variation not accounted for by the first PC; the third PC is orthogonal to the first two PCs and accounts for most of the remaining variation etc. Most of the information in the original multidimensional expression space can now be represented in a graph of lower dimensionality using the main PCs as axes.

Fig. 5.2 classifies the *Xenopus laevis* development stages based on the expression of 18 genes. Three clusters are seen. They can be identified

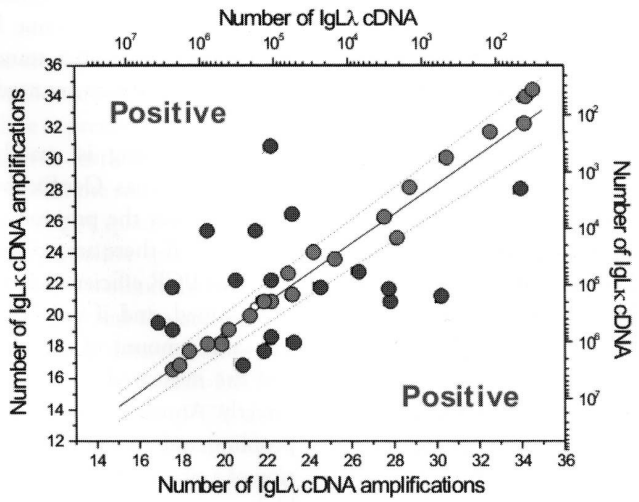

Figure 5.1 Classification of lymphoma samples in a scatter plot based on the expression of kappa and lambda variants of the immunoglobulin light chain. Each symbol represents a sample and its position in the plot is given by the CT values of the immunoglobulin κ and immunoglobulin λ reactions as indicated by the left and bottom axes. The right and top axes indicate the corresponding number of cDNA copies. Negative samples in the plot cluster around a diagonal line.

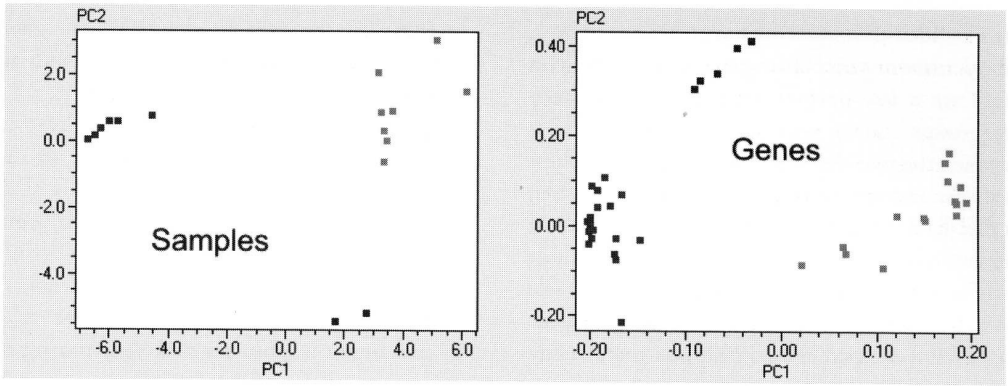

Figure 5.2 Classification of developmental stages (left) and of genes (right) in *Xenopus laevis* by Principal Component Analysis and visualized in a PC1 vs. PC2 plot. Early stages/genes are shown in red, midblastula stages/genes in blue and late stages/genes in green.

as the early 1–8.5 stages, stages 11 and 15 corresponding to the mid-blastula transition, and the late stages 17–44. Fig. 5.2 also shows corresponding classification of the genes, with three main clusters, corresponding to early, mid-blastula, and late genes. The PC1 vs. PC2 plots account for 74% of the information in the entire data set, as estimated from eigenvalues. Adding a third PC increases the amount of information to 82%. This can be viewed in a PC1 vs. PC2. vs. PC3 scatter plots, which indeed reveals some sub-clusters (Fig. 5.3).

Hierarchical clustering

An alternative to PC analysis is hierarchical agglomerate clustering. In hierarchical clustering all information in the data is accounted for, but the data are analysed sequentially, which means that all information is not considered at the same time. The procedure is:

1. In the multidimensional space find the two samples that are closest together and merge them into a cluster.
2. Find and merge the next two closest points, where a point is either an individual sample or a cluster of samples.
3. If more than one sample/cluster remains return to step 2.

The distance between two samples in the multidimensional space is typically calculated as

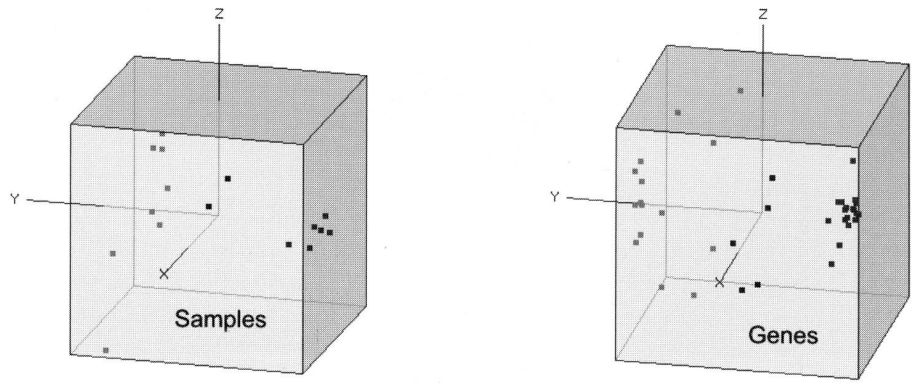

Figure 5.3 Same classification as in Fig. 5.3, visualized in a PC1 vs. PC2 vs. PC3 plot, which accounts for 82% of the variation in the experimental data.

the Euclidian (shortest) distance between them. To calculate distances between groups of samples there are a few options. The distance between the groups can be represented as the distance between the two closest samples, the two farthest samples, or some form of average distance, such as the distance between the centres of gravity of the two groups. Some more advanced methods such as the Ward's algorithm are also available (Ward, 1963). An alternative approach, proposed by Tichopad *et al.* (2006), is to calculate the correlation coefficient of the samples as a dissimilarity measure, and rank the correlation coefficients. Like auto-scaling this also eliminates the effect of expression levels.

Fig. 5.4 shows dendrograms of the *Xenopus laevis* developmental stages and genes, respectively. Both dendrograms show three main clusters, which correspond to the early stages/genes, mid-blastula stages/genes and late stages/genes. The members of the clusters are the same as those identified by the PCA. Also, the dendrograms reveal subclusters.

Self-organizing map

The idea behind the self-organizing map method (SOM; Kohonen, 2001) is to reflect variations in the expression profiles as a collection of cells, each with a representative expression profile, that are arranged to form a map with smooth changes in the profiles. When the expression profiles of the samples are located in the map, similar samples will be found close to each other. The SOM is generated as follows:

1. Initialize the cells in the SOM by giving each a random expression profile with the same genes as used in the experiment.
2. Choose by random one of the samples.
3. Compare the expression profile of the chosen sample with those in the cells of the SOM. The cell with the most similar profile is known as the best matching unit (BMU).
4. The neighbouring cells within a certain radius to BMU are now identified. The radius containing neighbours decreases during the training of the SOM. Initially, it is typically set to the 'radius' of the map, such that all cells are affected, and then it decreases with each time-step.
5. The expression profile of the BMU is adjusted to be similar to that of the selected sample. The expression profile of the neighbouring cells are also adjusted to be similar to the selected sample, but with a weight that decreases with the distance to BMU.
6. The procedure is repeated from step 2 a fixed number of iterations.

SOM trained with the expression profiles to classify samples and a SOM trained with samples to classify genes of *Xenopus laevis* are shown in Fig. 5.5.

Smoothing spline clustering

Gene expression over time is, biologically, a continuous process and as such it can be represented by a continuous function. Individual genes often share expression patterns. The shape of each

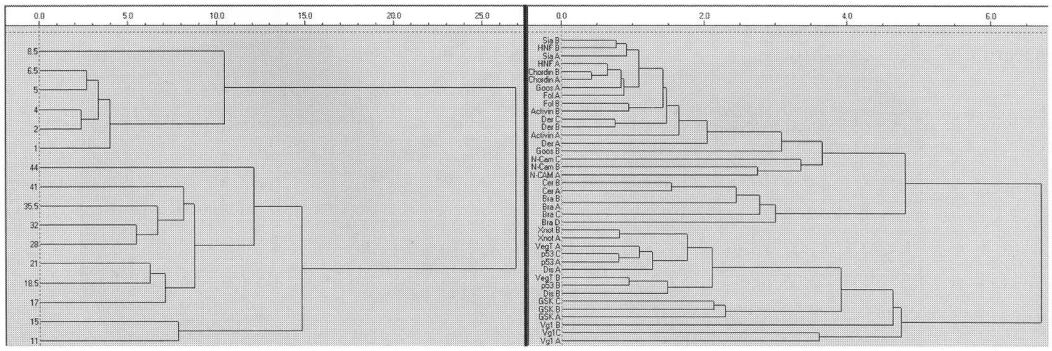

Figure 5.4 Hierarchical clustering of developmental stages (left) and of genes (right) in *Xenopus laevis*. Both dendrograms reveal three main clusters: early stages/genes, midblastula stages/genes, and late stages/genes. The unweighted pair model was used to calculate distances between groups.

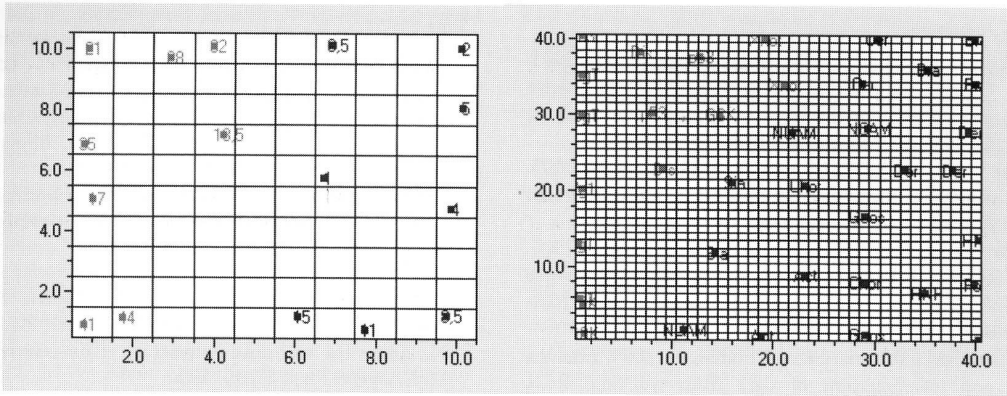

Figure 5.5 Example of a self-organized map (SOM) trained by stages (left) and genes (right) of *Xenopus laevis*. The three groups of stages/genes are seen to cluster.

function, the number of such functions, and the genes that share similar functional forms are typically unknown. Ping Ma *et al.* (2006) developed an approach to reveal related patterns of gene expression and their underlying functions. The method, smoothing spline clustering (SSC), models natural properties of gene expression over time, taking into account natural differences in gene expression within a cluster of similarly expressed genes. Furthermore, SSC provides a visual summary of each cluster's gene expression function and goodness-of-fit by way of a 'mean curve' construct and its associated confidence bands. The approach was used to model gene expression data over the life cycle of *Drosophila melanogaster* and *Caenorhabditis elegans* and revealed 17 and 16 unique patterns of gene expression respectively, in each species. The SSC software application and source code is freely available at the web page http://genemerge.bioteam.net/SSClust.html.

Data analysis in high-throughput screening

Malo *et al.* (2006) are developing statistical tools to detect quality hits in high throughput screens with a high degree of confidence. They examine statistical aspects of data pre-processing, focusing on concerns related to positional effects of wells within plates, choice of hit threshold and the importance of minimizing false-positive and false-negative rates. They argue that replicate measurements are needed to verify assumptions of current methods and they suggest data analysis strategies when the assumptions are not met.

Global pattern recognition

Akilesh *et al.*, (2003) developed global pattern recognition (GPR) to more reliably evaluate expression changes in real-time PCR data. GPR is inspired by triangulation techniques to determine positional information in cartography and astronomy. It goes through several iterations to compare the change of expression of a gene normalized to every other gene in a real-time PCR array. By comparing the expression of each gene to every other gene in the array, a global pattern is established, and significant changes are identified and ranked. GPR makes use of biological replicates to extract significant changes in gene expression, providing an alternative to relative normalization in real-time PCR experiments.

Conclusion

The successful application of qPCR and post-PCR data processing depends on a clear understanding of the problems. Facilitating data management and providing tools for automatic data analysis are the new goals in qPCR data processing and qPCR application software development. All of the calculation- and statistical-software applications described for this chapter are summarized and discussed at the following web page: http://bioinformatics.gene-quantification.info.

References

Akilesh, S., Shaffer, D.J. and Roopenian, D. 2003. Customized molecular phenotyping by quantitative gene expression and pattern recognition analysis. Genome Res. *13*, 1719–1727.

Bar, T., Stahlberg, A., Muszta, A. and Kubista, M. 2003. Kinetic outlier detection (KOD) in real-time PCR. Nucleic Acids Res. *31*, e105.

Burns, M.J., Nixon, G.J., Foy, C.A. and Harris, N. 2005. Standardisation of data from real-time quantitative PCR methods – evaluation of outliers and comparison of calibration curves. BMC Biotechnol. *7*, 31.

Burns, M.J., Valdivia, H. and Harris, N. 2004. Analysis and interpretation of data from real-time PCR trace detection methods using quantitation of GM soya as a model system. Anal. Bioanal. Chem. *378*, 1616–1623.

Bustin, S.A. 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. *25*, 169–193.

Bustin, S.A. 2002. Quantification of mRNA using real-time RT-PCR. Trends and problems. J. Mol. Endocrinol. *29*, 23–39.

Bustin, S.A. 2004. A–Z of Quantitative PCR, IUL Biotechnology Series, International University Line.

Bustin, S.A., Benes, V., Nolan, T. and Pfaffl, M.W. 2005. Quantitative real-time RT-PCR – a perspective. J. Mol. Endocrinol. 34(3), 597–601.

Dallal, G.E. and Wilkinson. L. (1986) An analytic approximation to the distribution of Lilliefors' test for normality. Am. Stat. *40*, 294–296.

D'Haeseleer P. 2005. How does gene expression clustering work? Nature Biotechnol. *23*, 1499 – 1501.

Donald, C.E., Qureshi, F., Burns, M.J., Holden, M.J., Blasic, J.R. Jr and Woolford, A.J. 2005. An interplatform repeatability study investigating real-time amplification of plasmid DNA. BMC Biotechnol. *5*, 15.

Fronhoffs, S., Totzke, G., Stier, S., Wernert, N., Rothe, M., Bruning, T., Koch, B., Sachinidis, A., Vetter, H. and Ko, Y. 2002. A method for the rapid construction of cRNA standard curves in quantitative real-time reverse transcription polymerase chain reaction. Mol. Cell Probes *16*, 99–110.

Gilsbach, R., Kouta, M., Bönisch, H. and Brüss, M. 2006. A comparison of in vitro and in vivo reference genes for internal standardization of quantitative real-time PCR data. Biotechniques *40*, 173

Goll, R., Olsen, T., Cui, G. and Florholmen, J.R. 2006. Evaluation of absolute quantitation by nonlinear regression in probe-based real-time PCR BMC Bioinformatics 7, 107

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. and Vandesompele, J. 2007. qBASE relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. 8:R19.

Horgan, G.W. and Rouault, J. 2000. Introduction to randomization tests, Biomathematics and Statistics Scotland.

Huggett, J., Dheda, K., Bustin, S.A., Zumla, A. 2005. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. *6*, 279–284.

Jin, N., He, K., Liu, L. 2006. qPCR-DAMS: a database tool to analyze, manage, and store both relative and absolute quantitative real-time PCR data. Physiol. Genom. *25*, 525–527.

Kohonen, T. (1995), Self-Organizing Maps, Springer Series in Information Sciences, Vol. *30*, Springer, Berlin, 1995. Third Extended Edition.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., Stahlberg, A. and Zoric, N. 2006. The real-time polymerase chain reaction. Mol. Aspects Med. *27*, 95–125.

Lai, K.K., Cook, L., Krantz, E.M., Corey, L. and Jerome, K.R. 2005. Calibration curves for real-time PCR. Clin. Chem. *51*, 1132–1136.

Larionov, A., Krause, A. and Miller, W. 2005. A standard curve based method for relative real time PCR data processing. BMC Bioinformatics 6, 62.

LightCycler Relative Quantification Software, Version 1.0 and 4.0, Roche Diagnostics, 2001 and 2004.

LightCycler Software®, Version 3.5; Roche Diagnostics, 2001.

Liu, W. and Saint, D.A. 2002a. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal. Biochem. *302*, 52–59.

Liu, W. and Saint, D.A. 2002b. Validation of a quantitative method for real time PCR kinetics. Biochem. Biophys. Res. Commun. *294*, 347–353.

Livak, K.J. (1997) ABI Prism 7700 Sequence detection System User Bulletin #2 Relative quantification of gene expression.

Livak, K.J. and Schmittgen, T.D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2^[-delta delta C(T)] Method. Methods *25*, 402–408.

Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J. and Nadon, R. 2006. Statistical practice in high-throughput screening data analysis Nature Biotechnol. *24*, 167–175.

Ma, P., Castillo-Davis, C.I., Zhong, W. and Liu, J.S. 2006. A data-driven clustering method for time course gene expression data. Nucleic Acids Res. *34*, 1261–1269.

Manly, B. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall.

Muller, P.Y., Janovjak, H., Miserez, A.R. and Dobbie, Z. 2002. Processing of gene expression data generated by quantitative real-time RT-PCR. Biotechniques *32*, 1372–1378.

Paoletti, C. and Mazzara, M. 2005. Definition of Minimum Performance Requirements for Analytical Methods of GMO Testing European Network of GMO Laboratories (ENGL) http://gmo-crl.jrc.it

Peirson, S.N., Butler, J.N. and Foster, R.G. 2003. Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. Nucleic Acids Res. *31*, e73.

Pfaffl M.W., Tichopad A., Prgomet C., Neuvians T. P. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample in-

tegrity: BestKeeper – Excel-based tool using pairwise correlations. Biotechnol. Lett. 26, 509–515.

Pfaffl, M.W. and Hageleit, M. 2001. Validities of mRNA quantification using recombinant RNA and recombinant DNA external calibration curves in real-time RT-PCR. Biotechnol. Lett. *23*, 275–282.

Pfaffl, M.W., Horgan, G.W. and Dempfle, L. 2002. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. *30*, e36.

Pfaffl, MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. *29*, e45.

Pfaffl, M.W. 2004. Quantification strategies in real-time PCR, pages 87–120 in, A–Z of Quantitative PCR, eds. Bustin SA., IUL Biotechnology Series, International University Line,

Ramakers, C., Ruijter, J.M., Deprez, R.H. and Moorman, A.F. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci, Lett. *339*, 62- 66.

Rasmussen, R. 2001. Quantification on the LightCycler. In: Meuer, S, Wittwer, C. and Nakagawara, K, eds. Rapid Cycle Real-time PCR, Methods and Applications. Springer Press, Heidelberg.

Regulation (EC) No. 1829/2003. of the European Parliament and Council of 22 September 2003 on genetically modified food and feed.

Regulation (EC) No. 641/2004. of 6 April 2004. on detailed rules for the implementation of Regulation (EC) No. 1829/2003. of the European Parliament and of the Council as regards the application for the authorization of new genetically modified food and feed, the notification of existing products and adventitious or technically unavoidable presence of genetically modified material which has benefited from a favourable risk evaluation.

Regulation (EC) No. 882/2004. of the European Parliament and Council of 29 April 2004 on official controls performed to ensure the verification of compliance with feed and food law, animal health and animal welfare rules.

REST©, REST-XL©, REST-384©, REST-MCS©, REST-RG©; Pfaffl, M.W., Horgan, G.W., Vainshtein, Y., Avery, P. 2002. and 2005, Institute of Physiology, Weihenstephan, Technical University of Munich, Germany; http://rest.gene-quantification.info/.

REST-2005© M.W. Pfaffl and M. Herman 2005, Technical University Munich and Corbett Life Science; http://rest.gene-quantification.info/.

Rutledge, R.G. 2004. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. Nucleic Acids Res. *32*, e178.

Sindelka, R., Ferjentsik, Z., Jonak, J. 2006. Developmental expression profiles of *Xenopus laevis* reference genes. Dev. Dyn. *5*, 754–8.

Souaze, F., Ntodou-Thome, A., Tran, C.Y., Rostene, W. and Forgez, P. 1996. Quantitative RT-PCR: limits and accuracy. Biotechniques *21*, 280–285.

Stahlberg, P. Aman, B. Ridell, P. Mostad and Kubista, M. 2003. Quantitative Real-Time PCR Method for Detection of B-Lymphocyte Monoclonality by Comparison of alpha and gamma Immunoglobulin Light Chain Expression. Clin. Chem. *49*, 51–59.

Tichopad, A., Didier, A., Pfaffl, M.W. 2004. Inhibition of real-time RT-PCR quantification due to tissue specific contaminants. Mol. Cell. Probes *18*, 45–50.

Tichopad, A., Dilger, M., Schwarz, G. and Pfaffl, M.W. 2003. Standardised determination of real-time PCR efficiency from a single reaction setup. Nucleic Acids Res. *31*, e122.

Tichopad, A., Pecen, L. and Pfaffl, M.W. 2006. Distribution-insensitive cluster analysis in SAS on real-time PCR gene expression data of steadily expressed genes. Comput. Methods Programs Biomed. *82*, 44–50

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3, 0034.1

Ward, J.H. (1963). Hierarchical Grouping to optimize an objective function. J. Am. Stat. Assoc. *58*, 236–244.

Wilhelm, J., Pingoud, A. and Hahn, M. 2003. *SoFAR* – Validation of an algorithm for automatic quantification of nucleic acid copy numbers by real-time polymerase chain reaction. Anal. Biochem. *317*, 218–225.

Wittwer, C.T., Reed, G.H., Gundry, C.N., Vandersteen, J.G. and Pryor, R,J. 2003. High-resolution genotyping by amplicon melting analysis using LC Green. Clin. Chem. *49*, 853–60.

Wittwer, C.T., Herrmann, M.G., Gundry, C.N. and Elenitoba-Johnson, K.S. 2001. Real-time multiplex PCR assays. Methods *25*, 430–42.

Wong, M.L., Medrano, J.F. 2005. Real-time PCR for mRNA quantitation. Biotechniques *39*, 75–85.

Yuan, J.S., Reed, A., Chen, F., Stewart, C.N. Jr. 2006. Statistical analysis of real-time PCR data. BMC Bioinformatics. 22(7), 85.