23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

Classifying the shape of qPCR expression data

P.A. Whigham [a], **J. Stanton** [b], **C. Mason** [c], **C.B. Rawle** [d] **and M. Macknight** [e]

[a]*Department of Information Science, University of Otago, Dunedin, New Zealand*
[b]*Department of Anatomy, University of Otago, New Zealand*
[c]*CMSoftware, Dunedin, New Zealand*
[d]*Department of Anatomy, University of Otago, New Zealand*
[e]*ADInstruments New Zealand, 77 Vogel St. Dunedin, New Zealand*
Email: peter.whigham@otago.ac.nz

Abstract: The application of real-time quantitative polymerase chain reaction (qPCR) methods for the identification and quantification of gene sequences requires automatic methods for the assessment of the resulting shape of the fluorescence readings. This requirement is especially important when we consider the use of rapid assessment hand-held qPCR devices, point of care operations and having to deal with a range of user expertise with qPCR interpretation. This paper describes a method for characterising different outputs from such devices by characterising the shape of the resulting fluorescence using a generalised sigmoid model based on the 5-parameter Richards function.

The shape of qPCR output should in theory follow a standard sigmoid shape, however with point of care devices with samples collected in the field there are a range of issues that may cause different responses. For example, biological samples can be contaminated, resulting in a reduced response for the PCR reaction, samples can have other chemicals that react with the amplification process, enzymes can be used that are not stable or have been destroyed due to heat or sunlight, and so on. Therefore a method is required to allow a range of output shapes to be classified, with an appropriate assessment for each type of output to then be presented to the user. In addition, qPCR devices have multiple wells for analysis with different behaviour. For example a well may be designed for a zero response, or have different DNA sequences, or use a specific concentration of DNA to confirm expected amplification to test that the device is working or for estimating DNA quantity. This means a flexible approach to assessing each well response is required.

This paper demonstrates the use of a simple one-dimensional line search algorithm for fitting the Richards equation, and the use of derived parameters from this model as input to a decision tree for shape analysis. The training data for shape analysis is created by adjusting the Richards parameters with associated noise. The advantage of designing different shapes by adjusting the parameters of the Richards equation is that unlimited training data can be sampled and automatically labelled for a specific shape. The use of appropriate noise levels ensures that the data is similar to real qPCR output. The resulting decision tree model demonstrates that a range of shape classes can be classified with high accuracy, and that the fitting of the parameters can be implemented without the use of complex matrix operations. The shape classification of the model to a range of test data of varying quality and fluorescence shapes is demonstrated on two real data sets created by different devices.

Keywords: *qPCR curve analysis, nonlinear optimization, decision tree, shape models*

1 INTRODUCTION

The application of fast quantitative polymerase chain reaction (qPCR) methods for the identification and quantification of gene sequences requires automatic methods for the assessment of the resulting shape of the fluorescence readings. For hand-held and field-based equipment this is critical given the possibly lack of skilled technicians for interpreting the output and the uncertainty of the result when using biological samples. This paper describes a method to classify the shape of qPCR output based on seven different types of representative curve. No assumption is made in terms of the scale of the data when classifying shape, and the resulting parameters from the model can be used to match control signals to determine quantification as well as assessing run quality.

Previous work to assess quality of qPCR runs have been developed, often based around the first and second derivative of the expression curve (Tichopad et al. (2010)). Since the shape of the expression is sigmoid a number of authors have fitted forms of generalised sigmoid functions to the data using a non-linear weighted estimate of the sigmoid parameters such as Levenger-Marquardt or Gauss-Newton (Liu and Saint (2002)). An extensive study on curve-fitting methods has been recently carried out Ruijter et al. (2013) which showed large variation in the methods but some consistencies depending on the type of data. Although much work has been done on estimating the expression inflection point there has been less emphasis on classifying overall shape. The work of Sisti et al. (2010) is a notable exception, where the emphasis was on the shape dynamics and the structure of the expression data. This work fitted the Richards function, an extension of the logistic growth model, to the expression data, with the aim to assess curves due to a range of sample contaminants. One benefit of the Richards function is that the first and second derivatives can be mathematically derived, allowing properties such as the inflection point and the tangent straight-line slope to be automatically calculated from the parameters of the fitted function.

This paper describes a method using normalised data fitted to a Richards equation to generate a range of expression curves for different classes of behaviour. The Richards parameters are fitted using a multi-parameter line search algorithm which does not require any matrix operations and is therefore safe under all forms of input. The example curves are used to train a decision tree for classifying these classes of behaviour. Classification accuracy is assessed using a cross-validation of the generated data and examples using independent data from two different types of PCR device. Section 2 describes the model, training data creation and shape classification, §3 examines the model properties, §4 shows the model shape classification applied to real data, and §5 concludes the paper.

2 METHODS

The Richards equation (Sisti et al. (2010); Spiess et al. (2008)) is defined as:

$$F(x) = F_b + \frac{F_{max}}{[1 + e(\frac{-1}{b}(x - c))]^d} \tag{1}$$

where x is the cycle number, F(x) is the fluorescence expression at cycle x, F_b is the background fluorescence, F_{max} is the maximal reaction fluorescence, with b,c and d as fitted coefficients of the model. Normally a non-linear optimisation method requiring matrix operations was used to fit the parameters (Sisti et al. (2010)) for Eqn. 1. However poor quality expression data can result in issues with convergence and reliability. An alternative approach is to use a simple single-parameter line search algorithm for each parameter. This requires multiple passes for each parameter being optimised, and multiple restarts for the initial parameter settings, but the approach can handle any quality of data. Since this approach is sensitive to initial parameter estimates, the model must be run many times with different random starting parameters, with the final selected parameters based on the lowest residual (mean-squared) error to the data. We have used a golden section search (Press et al. (2007)), although other methods, such as bisection or dichotomous search, would also be suitable.

The following additional parameters were derived from the solution of Richards equation: the y intercept of the inflection point (Y_f), the tangent straight line slope (m), the abscissa estimate for C_q and a measure of the asymmetry of the expression curve (*Asym*). Details for these calculations can be found in Sisti et al. (2010).

2.1 Training Data

The classification model was trained using examples generated from a set of models, rather than by using real data. This decision was made since it meant there was no limit to the number of examples that could be created,

Table 1. Parameter setting used to generate $F(x)$ shapes. $\mathcal{N}(0, \sigma^2)$ indicates sampling from a normal distribution with mean 0 and standard deviation σ^2. All shapes have $F_{max} = 40 + \mathcal{N}(0, 6)$ and $F_b = 1 + \mathcal{N}(0, 0.21)$.

Shape	b	c	d
good	$0.04 + \mathcal{N}(0, 0.005)$	$0.65 + \mathcal{N}(0, 0.04)$	$0.9 + \mathcal{N}(0, 0.01)$
saturated	$0.021 + \mathcal{N}(0, 0.005)$	$0.45 + \mathcal{N}(0, 0.07)$	$0.9 + \mathcal{N}(0, 0.01)$
incomplete	$0.13 + \mathcal{N}(0, 0.005)$	$0.797 + \mathcal{N}(0, 0.01)$	$0.4128 + \mathcal{N}(0, 0.01)$
late	$0.035 + \mathcal{N}(0, 0.005)$	$0.99 + \mathcal{N}(0, 0.01)$	$0.7 + \mathcal{N}(0, 0.01)$
poor chem	$0.035 + \mathcal{N}(0, 0.005)$	$0.99 + \mathcal{N}(0, 0.01)$	$0.2 + \mathcal{N}(0, 0.01)$

the classification of shape was automatic, and it allowed flexibility in terms of the types of example curves that could be identified. Examples of the training data for the defined $F(x)$ classes are shown in Fig 1. The training data for the classes *good, saturated, poor chemistry, late response* and *incomplete* were created using the Richards equation with selected parameters shown in Table 1. To allow variation for each class the parameters were sampled with some random variation, and by adding noise to the final generated expression data the curves simulated real data variation. Each example was created using normalised cycle numbers (between 0 and 1), but the final fluorescence values $F(x)$ were stored as absolute values. This meant that the training data could be directly compared with real data. The normalised cycle count meant that the resulting classification of shape was independent of cycle length. However, when producing the final parameter descriptions for each training example, both x (cycle number) and $F(x)$ (influorescence) were normalised. This meant that the classification model was independent of scale. The Richards equation could not create the double c_q pattern directly, and so was sampled by appending two examples randomly sampled from the good model. The first $10-14$ cycles used $F_{max} = 10+\mathcal{N}(0, 1)$ and the remaining cycles (to length 40) used $F_{max} = 30+\mathcal{N}(0, 6)$. The two patterns were linked, assuming the last value from the first model was X, by setting the second good model to have $F_b = X + \mathcal{N}(0, 0)$. The *error* class examples were created by randomly selecting from the five shape models with $F_{max} = 20 + \mathcal{N}(0, 1)$ and a model with a uniform random sampling of values in the range $5 - 15$ added to the resulting $F(x)$ sequence. The data was then reversed to form the final error sequence.

2.2 Shape Classification

Figure 1 shows an example of the target shape classes. The class of expressions that have no response are not considered for modelling with the Richards equation. Classifying expressions as no response uses the original expression data and satisfies Eqn. 2, where $threshold_1$ and $threshold_2$ are set based on the characteristic low signal for the device.

$$NoResponse(x) = (mean(x) < threshold_1) \wedge (var(x) < threshold_2) \tag{2}$$

Classifying the shape data independent of scale requires $F(x)$ to be normalised to a range between 0-1 for both axes. This means that the matching of shape is independent of the magnitude of expression and number of cycles. Low expression signals that have a shape similar to, say, a *good* shape, are categorised as *good*. Although this means the output may not indicate a good signal, this can be handled by examining the unscaled F_{max}, c_q and Y_f parameters after the shape analysis. One advantage of using normalised data is that the lower and upper bounds of the parameters for the Richards equation can be easily defined. Although the training data is built with the $F(x)$ unnormalised, the optimisation of the coefficients of Eqn. 1 use normalised x and $F(x)$ values. This allows a scale-free approach to interpret appropriate lower and upper bounds when doing the golden search optimisation. This simplifies and generalises the method for representing curve shapes. The device that will be used with this model is designed with control wells that establish the scale of the data - and hence will allow a low expression of a good shape to be categorised post identifying shape. Apart from the parameters defined or calculated based on Eqn. 1 the mean squared error(MSE) of the residuals for the fitted model is also used as a parameter for classification. The table of parameter values for each shape classification are created by using labelled class data with the parameters of the Richards equation (Eqn. 1) fitted using the multi-parameter golden search algorithm.

The resulting table of parameter values and class labels are used to train a decision tree using the rpart package (Therneau and Atkinson (2018)) in the $\mathbb{R}$ programming language (R Core Team (2017)). Classifying an

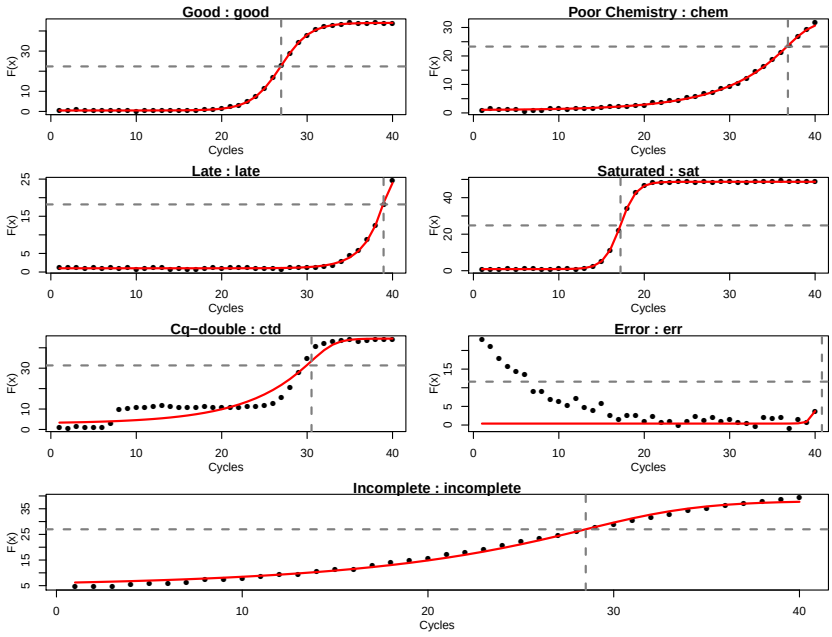

Figure 1. Example shape descriptions. Note the x and y-axes are shown with the original scale. Shapes with no response are not shown since they are handled separately. The vertical and horizontal grey lines show the calculated C_q and Y_f values respectively.

expression curve involves two stages: first, the data is assessed as having no response, and if this is not the case the normalised Richards equation parameters and derived measures are found and the expression data classified using the decision tree.

3 MODEL PROPERTIES

The Richards equation was used with different sampled parameter values (Table 1) to create 500 examples for each class shape. A range of different sized decision trees were created using the set of parameter values and tested for accuracy by randomly splitting the parameter table into 90% training and 10% testing, and repeating this 500 times. The resulting measure of overall accuracy (the sum of the diagonals of the confusion matrix divided by the total number of examples) is shown in Fig. 2. Note that the decision tree results in red just use the parameters F_{max}, m and Y_f, argued by Sisti et al. to be the only measurements required for a *fingerprint* model of shape. The black boxplots show the result using all of the parameters (although only a subset are finally used for the decision tree). A depth of seven creates a stable tree that has very close to perfect classification. When all of the data is used to build the final decision tree of depth seven (Fig. 3) the following parameters were used for classification: mse, b, c, F_{max} and m. Note that if a more extensive set of class examples is used, or a larger number of shape models sampled, the decision tree structure and selected parameters may change. It is worth noting that the process described here can easily be extended to other types of shapes - the only requirement is that the parameter ranges for Eqn. 1 can be determined to produce the shape.

The confusion matrix for 500 repeated training/test splits (90/10) using the decision tree for classification is shown in Table 2. This shows there is some overlap between the saturated and good classes, and some confusion between the late and poor chemistry classifications. Since some of the saturated examples are very similar to the good classes, and both would be assessed as runs with an appropriate response, this overlap is acceptable. The late and poor chemistry classes also have similar structure and therefore the confusion between these classes is acceptable.

4 RESULTS

The approach was tested using the Vermeulen A influorescence dataset (Ruijter et al. (2013)), which was a very clean dataset, and from the Freedom 4, a prototype hand-held qPCR device with a large amount of error

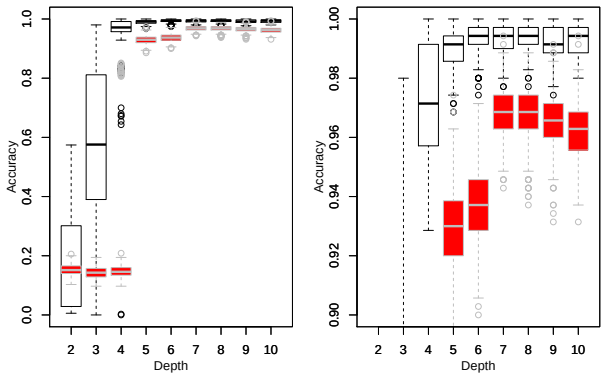

Figure 2. Overall accuracy based on 500 training (90% and test (10%) splits for a range of decision tree depths. The red boxplots show accuracy using the 3 parameter model as suggested by Sisti et al. (2010). The white boxplots are the model using all measured variables. A depth of seven using all explanatories is the simpliest model with good accuracy.

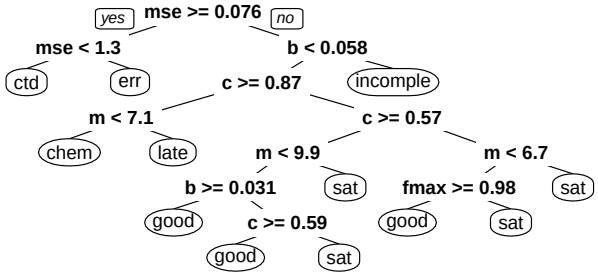

Figure 3. An example decision tree created using all of the generated data from the shape parameter set.

Table 2. Confusion Matrix for a decision tree of depth 7 with 90% Training, 10% Testing, repeated 500 times. Each class has 500 examples, resulting in $\approx$ 25000 test cases per class.

	Reference						
Prediction	chem	ctd	err	good	incomplete	late	sat
chem	25226	9	0	0	4	54	0
ctd	5	25055	0	0	8	0	0
err	0	0	24717	0	0	0	0
good	1	2	0	24661	3	0	325
incomplete	0	16	0	4	24946	0	0
late	51	10	0	2	0	24957	0
sat	0	0	0	321	2	0	24621

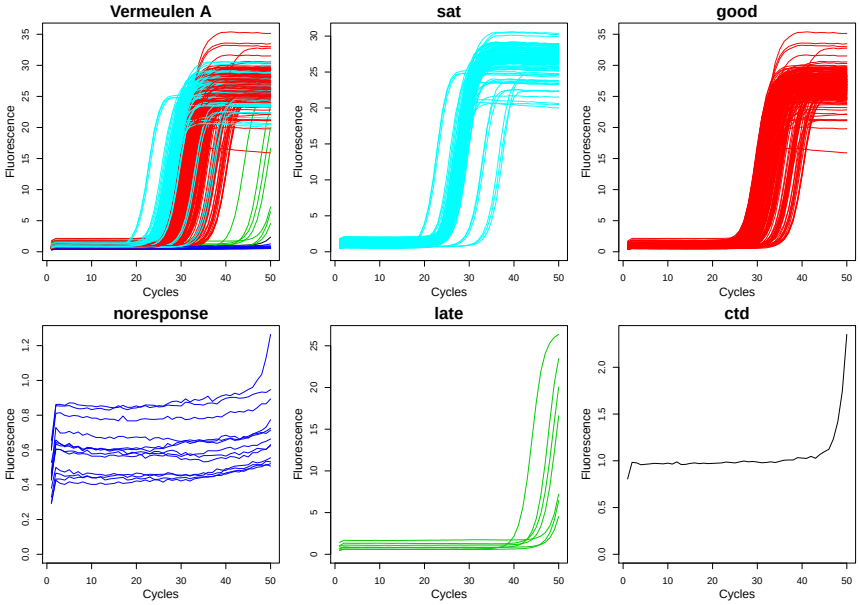

Figure 4. Classification of the Vermeulen A influoresence data. Note that the $F(x)$ scales are different for each class.

and noise. Vermeulen A had 384 examples of 50 cycle data. Several runs (wells) used water as a baseline and therefore had no response. The algorithm classified some examples as saturated, others as good, late, no response and ctd (ct double). Figure 4 shows the expression curves and associated (predicted) classes for this data. Since the data is real, and somewhat noisy, it shows that the current model can appropriately classify the shape of real data. It is also worth noting that the Vermeulen data is a different length from the data used for training (which had 40 cycles) and has a lower response (for the saturated and good classes) than most of the training data, demonstrating that the normalised approach to classification allows a generalisation to a range of output scales. Visual inspection indicates that no anomalous classifications were produced. Note also that the $ctdouble$ example is correctly classified given the small increase early in the response.

The 128 RunLog examples (Fig. 5) show a range of cycle number and expression response. There are, however, some runs classified as *good* that are very late in their sigmoid response. However these are not classified as *late* or *incomplete* since they do not have a similar structure to the training data. If this type of curve was necessary to be identified as a new class (for example, it could be called *weak* indicating a late expression but a possibly valid indication of the sequence) the model would have to be reconstructed using additional training data. This has been demonstrated successfully (not shown), indicating the flexibility of the method.

5 CONCLUSIONS

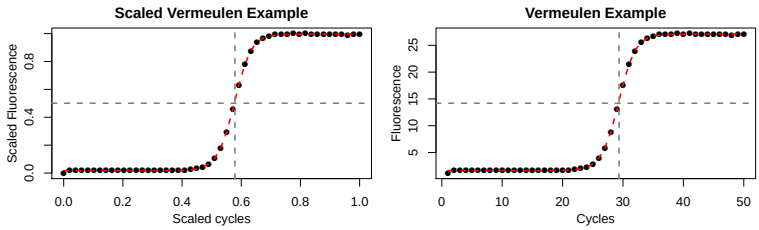

Figure 6. Example using Vermeulen data showing the scaled and unscaled calculation of C_q and Y_f. The points are the original data and the red line the result of fitting the Richards equation.

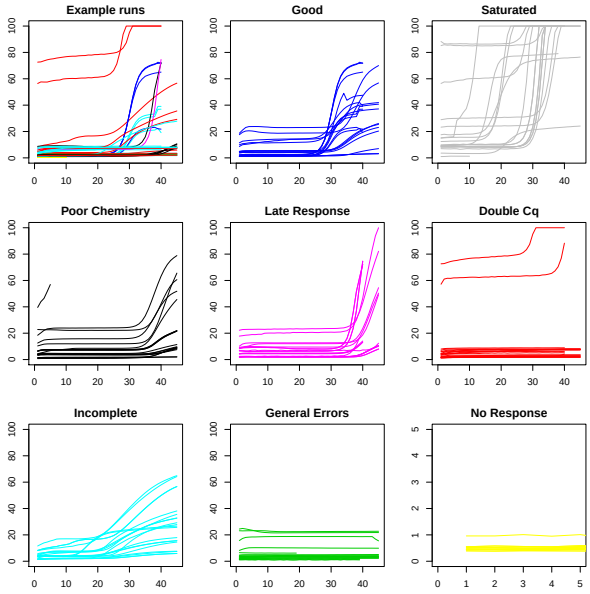

Figure 5. Freedom 4 data showing all runs and their classification.

The presented method shows that a normalised approach to shape classification is effective. Although further testing is required, the results are very encouraging. Tests have also been conducted to show that the resulting model can be re-scaled to the original data, and that the C_q and Y_f values are a good approximation to the inflection point of the expression data (Fig 6). There are some limitations to this approach, mainly in terms of variation in the shape towards the final cycles. For example, their are examples in Fig. 5 that have variation in the final cycles that could mean the run has not been successful. These types of issues need to be examined in relation to the behaviour of the individual device, but at present would not seem to be critical for the use of the presented approach.

ACKNOWLEDGEMENT

This work was funded through the Ministry of Business, Innovation and Education grant MBIE UOOX173.

REFERENCES

Liu, W. and D. Saint (2002). Validation of a quantitative method for real time pcr kinetics. *Biochem Biophys Res Commun 294*(2), 347–353.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ruijter, J. M., M. W. Pfaffl, S. Zhao, A. N. Spiess, G. Boggy, J. Blom, R. G. Rutledge, D. Sisti, A. Lievens, K. D. Preter, S. Derveaux, J. Hellemans, and J. Vandesompele (2013). Evaluation of qpcr curve analysis methods for reliable biomarker discovery: Bias, resolution, precision, and implications. *Methods 59*, 32–46.

Sisti, D., M. Guescini, M. Rocchi, P. Tibollo, M. D'Atri, and V. Stocchi (2010). Shape based kinetic outlier detection in real-time PCR. *BMC Bioinformatics 11*(186).

Spiess, A., C. Feig, and C. Ritz (2008). Highly accurate sigmoidal fitting of real-time pcr data by introducing a parameter for asymmetry. *BMC Bioinformatics 9*(221).

Therneau, T. and B. Atkinson (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Tichopad, A., T. Bar, L. Pecen, R. Kitchen, M. Kubista, and M. Pfaffl (2010). Quality control for quantitative pcr based on amplification compatibility test. *Methods 50*(4), 308–312.