Signal Identification of DNA Amplification Curves in Custom-PCR Platforms

1st Zhenzhe Han
Dept. of Electrical and Electronic Engineering
Imperial College London
London, UK
zhenzhe.han19@imperial.ac.uk

2nd Cavallo Francesca
Dept. of Electrical and Electronic Engineering
Imperial College London
London, UK
francesca.cavallo14@imperial.ac.uk

3rd Konstantin Nikolic
Dept. of Electrical and Electronic Engineering
Imperial College London
London, UK
k.nikolic@imperial.ac.uk

4th Khalid Mirza
Dept. of Biotechnology and Medical Engineering
National Institute of Technology
Rourkela, India
baigm@nitrkl.ac.in

5th Christofer Toumazou
Centre for Bio-Inspired Technology
Imperial College London
London, UK
c.toumazou@imperial.ac.uk

2021 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-7281-9201-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCAS51556.2021.9401777

***Abstract*—Custom-made, point-of-care PCR platforms are a necessary tool for rapid, point-of-care diagnostics in situations such as the current Covid-19 pandemic. However, a common issue faced by them is noisy fluorescence signals, which consist of a drifting baseline or noisy sigmoidal curve. This makes automated detection difficult and requires human verification. In this paper, we have tried to use nonlinear fitting for automated classification of PCR waveforms to identify whether amplification has taken place or not. We have presented several novel signal reconstruction techniques based on nonlinear fitting which will enable better pre-processing and automated differentiation of a valid or invalid PCR amplification curve. We have also tried to perform this classification at lower PCR cycles to reduce decision times in diagnostic tests.**

***Index Terms*—qPCR, noisy, automated detection, nonlinear curve fitting, classification, lower PCR cycles**

I. Introduction

Quantitative polymerase chain reaction (qPCR) is the gold standard for quantification of DNA sequences given its sensitivity, specificity and large dynamic range [1]. However, as shown in Fig. 1(a), it suffers from disturbances such as noise, drift, and non-specific amplification when performed on point-of-care devices. To obtain the real qPCR curves from an extremely large number of data samples, we need to find an algorithm which could identify the valid qPCR amplification signal automatically. Because the identification of qPCR does not have a very accurate numerical amplification template, i.e. different concentration and type of DNA has different threshold cycle (Ct), it makes automated detection using machine learning need extremely large training data which could be very difficult to be implemented. However, specific nonlinear model-fitting is a good choice for qPCR identification which can fit the qPCR curve automatically for any concentration and type of DNA if the curve has an amplification characteristic.

The specific nonlinear model-fitting detection could be classified as full-cycle detection and early-cycle detection. The full-cycle detection is based on model-fitting to the whole qPCR fluorescence data (full-cycle fitting) and therefore require waiting until the PCR reaction is complete. PCR reaction usually takes 45 minutes to 1 hour [2] [3], with each cycle about 1-1.5 minutes long. However, given that some time-limited applications require the time-to-result to be shortened, it is essential to reduce the number of cycles required, at which this classification can take place. For example, one of the current rapid Covid-19 tests deployed in the UK hospitals is CovidNudge – rapid point-of-care real time RT-PCR test, which requires no laboratory handling or sample pre-processing [4]. The platform comprises two components: a disposable cartridge (where a swab is inserted) and a processing unit ('Box'). It needs approximately 1.5h to complete the test, which is very good for a non-lab-based system, but still it means only 16 tests can be done over a 24 hour period with one box. If the PCR processing time can be halved then a single machine would double its throughput to 32 tests per day, which is in current circumstances a very significant improvement.

To solve the signal identification problem, we also propose a method called early-cycle detection to shorten the detec-

We thank UK EPSRC (EP/N002474/1), ERC Synergy Grant (no.319818) and ERC Proof of Concept grant (no.825796), for their financial support.

978-1-7281-9201-7/21/$31.00 ©2021 IEEE

tion time, by fitting a model only to the early part of the fluorescence signal (early-cycle fitting) and by evaluating the similarity between the reconstructed signal and real signal to identify a true sample or valid PCR amplification curve from noise. Depending on the initial concentration of DNA, the time-to-result could be shortened significantly as results would be available shortly after the signal enters the exponential phase. In this paper, we not only quoted the existing model such as Sigmoid, Log-Logistic and MAK2 to support our classification algorithm but also proposed some novel models such as expMAK2, dsMAk2 and SOE to supplement our model library.

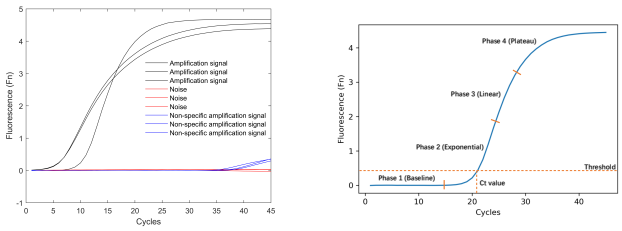
II. qPCR AMPLIFICATION CURVES

A. Mathematical Aspects

A qPCR amplification curve represents the fluorescence emitted by the target sequence, measured at each cycle of the PCR reaction. As shown in Fig. 1(b), a typical qPCR amplification curve has four phases [5]. Phase 1, approximately between 1 and 15 cycles, represents a baseline with a small variation of fluorescence. At this stage, the concentration of DNA is too low to be detected, which causes the background noise to cover the real signal. The background noise should be determined and removed by setting the range of baseline. Background noise can be removed automatically by the qPCR instrument by selecting the range of baseline. Processed qPCR curve is obtained by removing the background value. However, the result in phase 1 is still not clear, even if the background noise is removed. Because the noise is irregular which breaks the initial stage data irreversibly, so, a threshold line should be set as a starting point of correct qPCR quantification. Phase 2 is a region where the qPCR curve increases sharply as exponential. This phase represents the amplification characteristic, and the efficiency is large enough with the threshold line of detection equal to ten times the standard deviation of the baseline fluorescence signal [6]: this gives the Ct value, i.e. the cycle at which the threshold line crosses the fluorescence signal. Ct value will help for classification of the qPCR curve. Phase 3 is a region where the efficiency decreases, which is because of the limited amount of deoxyribonucleoside triphosphate (dNTPs) and enzymes. The qPCR curve reaches a plateau and stays almost unchanged in phase 4 [5]. At this stage, enzymes are gradually inactivated, dNTPs are exhausted, and the amount of product increase gradually decreases; until no product is produced, the number remains constant [5]. Four phases compose an s-shaped curve which is the characteristic of a valid amplification qPCR curve.

B. Ct of Amplification qPCR Curves

As shown in Fig. 2, different qPCR curves could be classified by Ct value, which means the Ct value is a unique characteristic of the qPCR curve. So, early-cycle detection could be used to classify qPCR curves by detecting the Ct value. Considering the real definition of Ct value is the cycle where fluorescence reaches the threshold value [7], early-cycle detecting cannot identify the Ct value extreme accurately.



(a) Signals in qPCR platform (b) Processed qPCR curve shape

Fig. 1: qPCR signals

However, setting several regions for Ct values should be helpful for classification such as three Ct values form an interval, which is shown as Fig. 2.

III. DNA AMPLIFICATION CURVES IDENTIFICATION

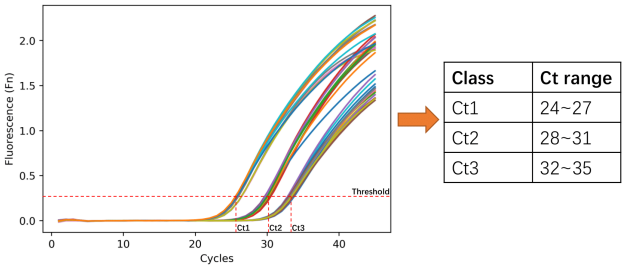
A. Full-Cycle Detection

Full-cycle fitting is implemented by fitting a model to the entire qPCR curve, including all characteristic phases. The choice of a suitable fitted model could be found in full-cycle fitting model. When performing full-cycle detection, a threshold R^2* should be set first (Fig. 3). The threshold R^2* should be set to around 0.99 theoretically. Secondly, a valid qPCR signal is fitted by the full-cycle fitting model, and if a high R^2 value is obtained, the amplification characteristic is reliably detected. If the R^2 is greater than the set threshold R^2*, the signal is classified as a valid qPCR amplification curve. If R^2 is smaller than R^2*, the signal is classified as noise.

B. Early-Cycle Detection

Early-cycle fitting is implemented by constructing a model which only fits phase 1 (baseline) and phase 2 (exponential). Early-cycle detection is dynamic and uses the R^2 value as a criterion for detection. The choice of a suitable fitted model could be found in early-cycle fitting model. Considering that phase 2 (exponential) is the main feature of the amplification curve, if phase 2 exists, the probability of the existence of true amplification in the qPCR curve is exceptionally high. The models for early-cycle fitting are therefore focused on the phase 2.

As shown in Fig. 3, a threshold value R^2* should be set first (around 0.99). At each cycle, the chosen model is fitted to the



Class	Ct range
Ct1	24~27
Ct2	28~31
Ct3	32~35

Fig. 2: qPCR curves with Ct value

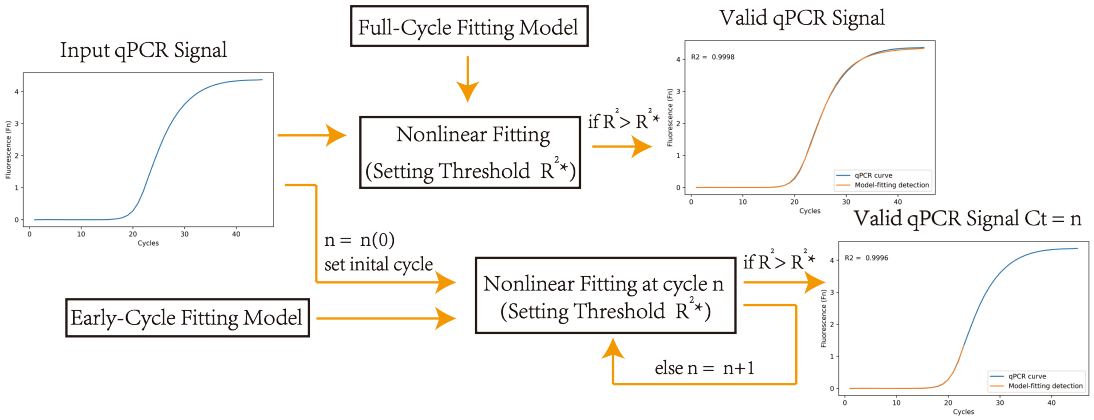


Fig. 3: Full-cycle detection and early-cycle detection

fluorescence curve obtained up until that cycle. If $R^2{*}{>}R^2$, the fluorescence signal is classified as valid amplification; otherwise, qPCR will keep running until the last cycle is reached. If the amplification characteristic is found before the end of the qPCR, the final fitting cycle n will be recorded as the Ct value. However, if all cycles have been completed, but R^2 is still lower than $R^2{*}$, the signal will be classified to noise.

IV. Models for Full-Cycle Fitting of qPCR Amplification Curves

A. Four and Five-Parameter Sigmoid Models

The description of four-Parameter sigmoid model is:

$$F_n = F_{\mathrm{b}} + \frac{F_{\max}}{\left(1 + e^{-(n-c)/b}\right)} \tag{1}$$

where F_n represents the fluorescence at cycle n; $F_{\max}$ represents the maximum fluorescence; F_b is the background fluorescence; and c is the fractional cycle of the inflection point and b is the slope of the curve [8] [9].

B. Log-Logistic Models

The expression of five-parameter log-logistic models is:

$$F_n = F_b + \frac{F_{\max} - F_b}{\left(1 + e^{q(\log(n) - \log(r))}\right)^s} \tag{2}$$

where F_n is fluorescence at cycle n; F_b is background fluorescence; $F_{\max}$ is the maximum fluorescence in the whole qPCR curve; q, r, s are used to adjust the shape of the fitting curve. Four-parameter log-logistic model is a specific case of the five-parameter log-logistic model when $s = 1$ and s are used to keep the asymmetry shape of the curve. The sigmoid model is same as log-logistic model only except the part of $(\log(n) - \log(r))$ [6].

C. Tanh Model

The tanh model can be represented as:

$$F_n = F_b + F_{\max}\left(\frac{1}{2} + \frac{\tanh(q(\log r - \log n))}{2}\right)^s \tag{3}$$

where F_n represents fluorescence at cycle n; F_b represents a constant background fluorescence; $F_{\max}$ represents the maximum fluorescence in the whole qPCR curve; q, r ,s are used to adjust the shape of the fitting curve. As the tanh model is derived from the sigmoid model, tanh model has the same fitting property as a sigmoid function.

V. Models for Early-Cycle Fitting of qPCR Amplification Curves

A. Two-Parameter Mass Action Kinetic Model of PCR

Two-Parameter Mass Action Kinetic Model (MAK2) is a mechanistic model obtained from reaction kinetics in the anneal/elongation steps of PCR [6]. The expression can be described as [6]:

$$D_n = D_{n-1} + k\ln\left(1 + \frac{D_{n-1}}{k}\right) \tag{4}$$

where n represents the cycle of double-stranded DNA (dsDNA) replication; D_n represents the fluorescence associated with dsDNA at n cycle; D_{n-1} represents the fluorescence associated with dsDNA at $n-1$ cycle; k is an adjustable parameter which represents the increasing rate of dsDNA in the process of PCR [6].

B. Modified MAK2 for Less Cycle Fitting

- Exponential MAK2 model (expMAK2) can be described as:

$$D_n = D_{n-1} + k\left[\ln\left(1 + \frac{D_{n-1}}{k}\right)\right]^{2k} \tag{5}$$

  where n represents the cycle of dsDNA replication; D_n represents the fluorescence associated with dsDNA at n cycle; D_{n-1} represents the fluorescence associated with dsDNA at $n-1$ cycle; k is adjustable value to adjust the amplification efficiency of fitting curve.
- The expression of double-strand MAK2 model (ds-MAK2) is:

$$D_n = D_{n-1} + k\ln\left(1 + \frac{kD_{n-1}}{2}\right) + q\ln\left(1 + \frac{qD_{n-1}}{2}\right) \tag{6}$$

compared to MAK2 and expMAK2, dsMAK2 introduces a new element to capture the reaction efficiency of another single-strand DNA. With the compensation of the new parameter q, amplification efficiency becomes more flexible, so dsMAK2 could fit the qPCR curve with high R^2 but uses fewer cycles. The slope of the curve is controlled by both k and q. So, dsMAK2 describes the reaction efficiency of two single-stranded DNA (ssDNA).

C. Second-Order Efficiency Model (SOE) for Less Cycle Fitting

The description of the second-order efficiency model (SOE) is:

$$D_n = D_{n-1} + \frac{(b+1) \times D_{n-1}}{D_{n-1} + a^2/b + a^2/D_{n-1}} \tag{7}$$

where n represents the cycle of dsDNA replication; D_n represents the fluorescence associated with dsDNA at n cycle; D_{n-1} represents the fluorescence associated with dsDNA at $n-1$ cycle; a and b are adjustable parameters which could adjust the amplification efficiency.

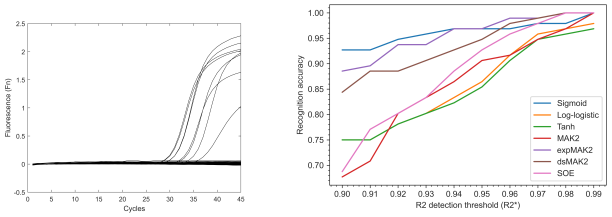
VI. Recognition Results

We used a real qPCR data-set for qPCR identification testing. As shown in Fig. 4(a), the data-set is composed of 96 qPCR curves, including both valid amplification signal and invalid noise signal.

All data in data-set were recognised by each model with different threshold value (R^2* = [0.90 to 0.99]). In our experiment, the recognition accuracy (AC) is defined as:

$$AC = P/N \tag{8}$$

where P represents the number of correct identifications and N represents the total number of test qPCR curves.



(a) Real qPCR data-set

(b) Recognition accuracy vs R^2* in each model

Fig. 4: Experiment data

TABLE I: Average Ct of each model

Model			
MAK2	***expMAK2***	***dsMAK2***	***SOE***
33.875	33.125	33.625	32.75

Fig. 4(b) shows the recognition accuracy of each model at different thresholds R^2*. R^2* can affect the recognition accuracy and the model can achieve a good recognition accuracy with a suitable R^2. The recognition accuracy increased as R^2* increased. All models reached the highest accuracy when R^2 = 0.99. When R^2* = 0.99, MAK2, expMAK2, dsMAK2, SOE and sigmoid give a recognition rate of 1.00. The accuracy of Log-logistic and Tanh cannot reach 1.00. In this data-set, we collected the average Ct value of each early-cycle detection model at R^2* = 0.99. Table I shows the average Ct of MAK2, expMAK2, dsMAK2 and SOE.

Here we also showed the classification results of each early-cycle fitting model. As shown in Fig. 5, the threshold R^2* = 0.99. The results have three parts: (1) Identification, which represents the model recognition results of amplification curve and noise. (2) Ct, which shows the minimum cycles required for model-fitting to reach R^2*. (3) Classification, which is the results of the qPCR amplification curve classification by Ct value. The range of each class is 5 cycles.

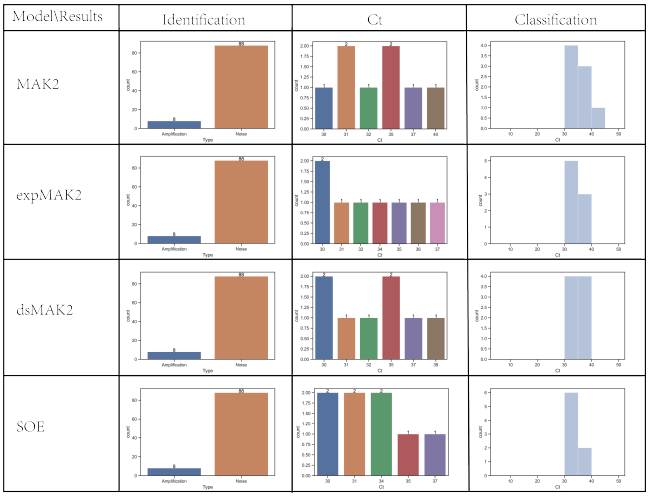


Fig. 5: Identification, Ct and classification result of 4 models

VII. Conclusion

Since the full cycle detection only needs to perform one fitting, full-cycle detection is easy to implement and has fast detection speed, but it can only be performed after the qPCR reaction. Early-cycle detection has high accuracy and can help shorten the time required to obtain qPCR results, but requires fitting a new model at each amplification cycle. R^2* should be set in a suitable range to achieve high recognition accuracy, usually, we set R^2* = 0.99 to reach the highest recognition accuracy.

Considering early-cycle detection is always run with thermal cycler simultaneously, it is necessary to be fast and accurate. So, the reduction of Ct value would be an improvement of time-limited amplification detection. By using Ct value obtained from early-cycle detection, amplification signal could be classified further. The normal range of Ct value is between 15 and 35, and some signal with very high or very low Ct value could be found from a large number of amplification signals. So, early-cycle detection could help us find the non-specific amplification with abnormal Ct value for further processing.

REFERENCES

[1] M. Wong and J. Medrano, "Real-Time PCR for mRNA quantification," *BioTechniques*, vol. 39, no. 1, pp. 75–85, 2005. [Online]. Available: www.BioTechniques.com

[2] R. Saiki, D. Gelfand, S. Stoffel, S. Scharf, R. Higuchi, G. Horn, K. Mullis, and H. Erlich, "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase," *Science*, vol. 239, no. 4839, pp. 487–491, jan 1988. [Online]. Available: https://science.sciencemag.org/content/239/4839/487 https://science.sciencemag.org/content/239/4839/487.abstract

[3] S. A. Bustin, "How to speed up the polymerase chain reaction," *Biomolecular Detection and Quantification*, vol. 12, pp. 10–14, jun 2017.

[4] M. M. Gibani, C. Toumazou, M. Sohbati, R. Sahoo, M. Karvela, T.-K. Hon, S. De Mateo, A. Burdett, K. Y. F. Leung, J. Barnett, A. Orbeladze, S. Luan, S. Pournias, J. Sun, B. Flower, J. Bedzo-Nutakor, M. Amran, R. Quinlan, K. Skolimowska, C. Herrera, A. Rowan, A. Badhan, R. Klaber, G. Davies, D. Muir, P. Randell, D. Crook, G. P. Taylor, W. Barclay, N. Mughal, L. S. P. Moore, K. Jeffery, and G. S. Cooke, "Assessing a novel, lab-free, point-of-care test for sars-cov-2 (covidnudge): a diagnostic accuracy study," *The Lancet Microbe*, vol. 1, no. 7, pp. e300 – e307, 2020.

[5] J. Brunstein, "Interpretation of qPCR curve shapes," *MLO: medical laboratory observer*, vol. 47, no. 6, jun 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26281509/

[6] G. J. Boggy and P. J. Woolf, "A Mechanistic Model of PCR for Accurate Quantification of Quantitative PCR Data," *PLoS ONE*, vol. 5, no. 8, p. e12355, aug 2010. [Online]. Available: www.plosone.org https://dx.plos.org/10.1371/journal.pone.0012355

[7] T. D. Schmittgen and K. J. Livak, "Analyzing real-time PCR data by the comparative CT method," *Nature Protocols*, 2008.

[8] J. Rodriguez-Manzano, A. Moniri, K. Malpartida-Cardenas, J. Dronavalli, F. Davies, A. Holmes, and P. Georgiou, "Simultaneous Single-Channel Multiplexing and Quantification of Carbapenem-Resistant Genes Using Multidimensional Standard Curves," *Analytical Chemistry*, 2019.

[9] R. G. Rutledge, "Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications." *Nucleic acids research*, 2004.