Filomat 30:4 (2016), 1103–1110
DOI 10.2298/FIL1604103G

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Robust CT-Prediction Algorithm for RT-PCR

Melih Gunay[a], Rajarajeswari Balasubramaniyan[b]

[a]*Akdeniz University*
[b]*MrPIBiotech LLC*

Abstract. Introduction of fluorescence-based Real-Time PCR (RT-PCR) also called qPCR is increasingly used to detect multiple pathogens simultaneously and rapidly by gene expression analysis of PCR amplification data.

Real-time PCR data are analyzed after setting an arbitrary threshold that must intersect the signal curve in its exponential phase. The point at which the curve crosses the threshold is called CT (Cycle Threshold). This simple and arbitrary value however is not an elegant definition of CT value sometimes leads to conclusions that are either false positive or negative. Therefore, the purpose of this paper is to present a stable and consistent alternative approach for the definition and determination of CT value that leads to near zero false positives and negatives.

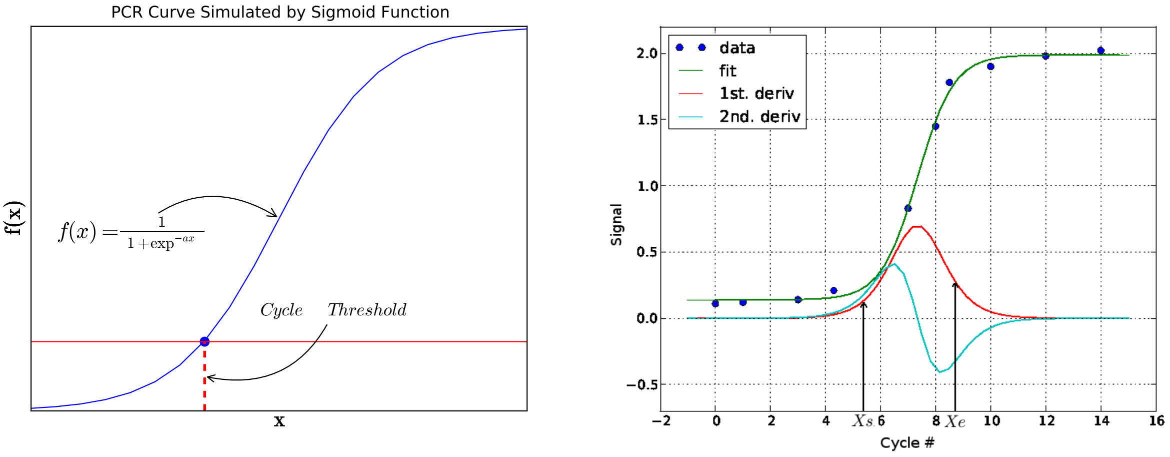
1. Introduction:

The advent of real-time PCR has enabled rapid and reproducible high throughput RTPCR quantification, with an unparalleled dynamic range and extremely high sensitivity. Real-time PCR is fast becoming the method of choice for the quantification of gene expression. In the past decade, polymerase chain reaction (PCR) assay has become a standard method for the detection of a wide range of pathogens and biomarkers in diagnostics. Improved nucleic acid amplification and detection technologies have facilitated the identification of multiple pathogens for a limited number of genes rapidly when only a small number of cells are available([2]). Employing these technologies, the simultaneous measurement of gene expression in many different samples can be invaluable tool during outbreaks and routine surveillance that had previously proved labor intensive and/or impossible to detect using traditional culture or immunofluorescent techniques ([3–5]).

Among the technologies that are developed, Real-Time PCR (RT-PCR), also called quantitative PCR or qPCR measures the amount of PCR product produced at each thermo cycle step of the reaction or in ”real time. The amount of PCR product is determined by the quantity of the fluorescent signal, provides a simple and elegant method for determining the amount of a target sequence or gene that is present in a sample. However, its simplicity can sometimes lead to problems of overlooking some of the critical factors that make it work [6].

In ideal conditions, PCR product grows exponentially. Because of impurities and it takes several cycles for enough product to be readily detectable, early in the PCR amplification process the plot of fluorescence

2010 *Mathematics Subject Classification*. Numerical Methods; Computer Science, Mathematical Programming
Keywords. (Prediction, CT-value, Algorithm, RT-PCR)
Received: 27 August 2015; Accepted: 23 November 2015
Communicated by Gradimir Milovanović and Yilmaz Simsek
Email addresses: mgunay@akdeniz.edu.tr (Melih Gunay), mrpibiotech@gmail.com (Rajarajeswari Balasubramaniyan)



(a) Logistic Function with CT mark

(b) Fitting of Sigmoid Curve onto the representative data

Figure 1: Fitting of data to the logistic function

vs. .cycle number however starts out flat. Next, as conditions become optimal and the amount of DNA copy theoretically doubles at every cycle, the emitted fluorescent signal increases proportional with the amount of PCR product exponentially. At later cycles, the reaction substrates become depleted, PCR product no longer doubles, and the curve begins to flatten. Consequently, if cycle number vs. fluorescence signal were to be plotted, it will exhibit a sigmoidal appearance as shown in Figure 1a and is given as a logistic function ([7]) in Equation 1.

$$f(x) = \frac{1}{1 + e^{-ax}} \tag{1}$$

The point on the curve in which the amount of fluorescence begins to increase rapidly, usually a few standard deviations above the baseline, is termed the threshold cycle value (CT value). This arbitrarily set cycle threshold value must intersect the signal curve in its exponential phase as demonstrated in Figure 1a.

Biologically, the higher the starting copy number of the nucleic acid target is, the sooner a significant increase in fluorescence is detected, and the lower the CT value [6, 8]. In other words, the lower the CT value, the higher the amount of target gene in the sample is. If the presence of target gene in the sample is **none**, then one should not expect the PCR curve to rise significantly, but rather flat with some random noise around the baseline. The same random signal may sometimes make subtraction of the baseline value in real-time RT-PCR data analysis difficult ([3, 9]).

Real-time PCR results differ from conventional diagnostic assays (e.g., enzyme-linked immune absorbent assay) in that negative specimens do not yield CT values as the fluorescent signal stays below the specified threshold [5]. Consequently, the distributions of CT values are generally non-normal, heteroscedastic, and truncated ([10]). In addition, the quantitative CT value is negatively associated with the (log) concentration of nucleic acids detected (i.e., high CT value reflects a low target concentration and vice versa) ([11]). Generally, when a CT value is obtained, the specimen tested is deemed to be positive. However, there is an tendency among laboratory operators to consider as negative (i.e., false-positive) when a CT value is above an arbitrary cutoff value ([4]). A sample with a CT value greater than the subjective cutoff might, therefore, be classified and reported as a negative result without additional information to distinguish it from results that did not produce a CT value ([12], ([4]).

Many models have been created and available in the literature for predicting CT cut off values for RT-PCR experiments concentrating on mRNA quantification; Real-time PCR Miner ([13]), TAQMan threshold method (MyIQ software) ([14]), the first derivative maximum (FDM) method ([15]), the second derivative maximum (SDM) method ([16]),[17]. Latest version of Real-time PCR Miner from Tellinghuisen et al., ([18]) have consolidated many of these methods and compared their precision and accuracy.

Although several methods are available for predicting CT values, they mostly test the abundance of mRNA. Method from ([12]) addresses the complexity of CT determination in diagnostic applications where the probability of classifying samples as false positive and false negative is higher. Therefore, the purpose of this research is to introduce another objective CT prediction algorithm that is robust in diagnostic applications.

2. Theoretical Model

The algorithm 1 introduced in this study tries to locate the point on the curve at which the amount of fluorescence signal begins to increase rapidly regardless of the baseline. The model fits modified sigmoid function (See Equation 2) to the signal and estimates the parameters of the function. The rate at which the response signal begins to change significantly with respect to its baseline is computed by simply taking the derivative of the function. For a typical signal, the curve will look as shown in Figure 1b, significant changes occur at cycles around the dotted line that intersects the CT threshold line. Furthermore, for a typical PCR curve, the rate of change should increase during what we call the **acceleration phase** and reach a maximum value at some cycle beyond the threshold cycle. Once the maximum rate of change is observed, the rate of change begins to decrease during what we call the **deceleration phase** as demonstrated in the logistic function curve (Figure 1a). The cycle where rate of change moves from acceleration phase to deceleration phase may be called the turning point. At turning point the rate of change (velocity) is the maximum. The second derivative of the signal (acceleration) crosses the zero line as shown in Figure 1b.

$$f(x) = \frac{a}{1 + e^{-k(x-x_0)}} + c \tag{2}$$

Equation 2 is the modified sigmoid function for CT prediction. The algorithm 1 develop first tries to estimate the parameters: a, k, x_0 and constant c. Once these parameters are identified, both the first and second derivative of the modified sigmoid function is determined as follows ([1]):

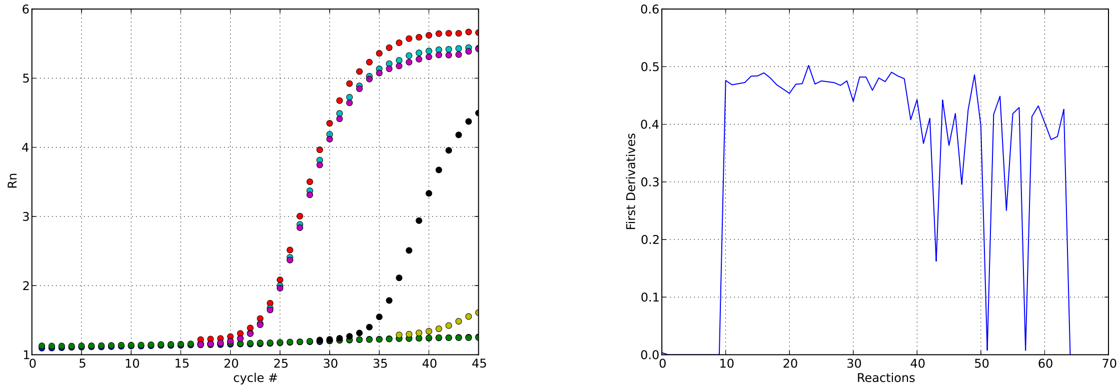
$$\frac{df(x)}{dx} = \frac{ae^{-k(x-x_0)}}{(1 + e^{-k(x-x_0)})^2} \tag{3}$$

If we let $z = e^{-k(x-x_0)}$, Equation 3 may be rewritten as follows:
The algorithm 1 uses the Levenburg-Marquardt algorithm through leastsq.

$$y' = df(x)/dx = \frac{az}{(1+z)^2} \tag{4}$$

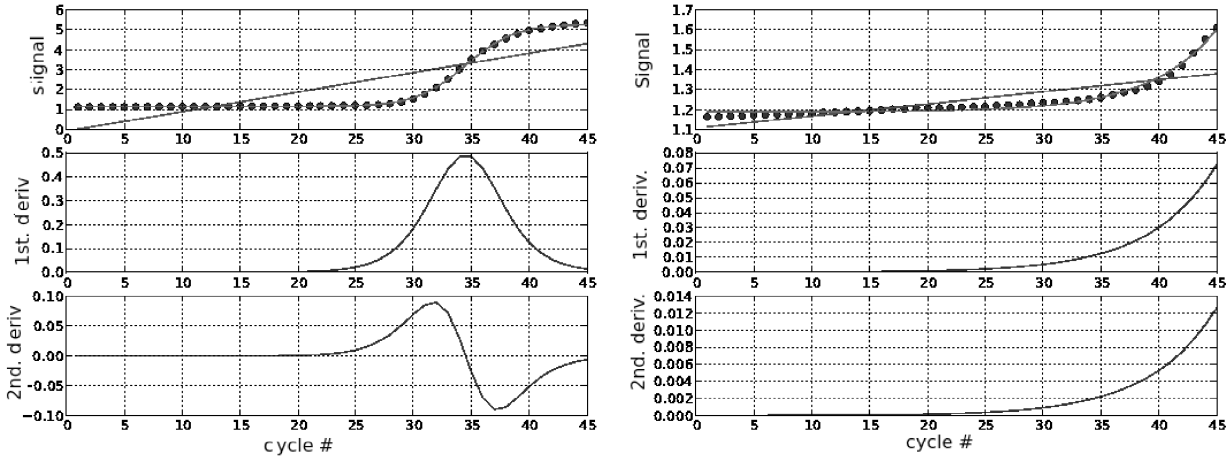
Using the Quotient Rule for derivation of fraction, one may obtain the second derivative of the sigmoid function as follows.

$$\begin{aligned} y'' = \frac{d^2 f(x)}{dx^2} &= -[\frac{azk^2}{(1+z)^2} - \frac{2ak^2z^2(1+z)}{(1+z)^4}] && (5) \\ &= -\frac{azk^2}{(1+z)^2}[1 + \frac{2z}{1+z}] && (6) \end{aligned}$$



(a) Cases to consider for the development of the algorithm (b) Maximum value of the first derivative for samples

Figure 2: Various cases for algorithm development



(a) Full sigmoid curve (b) Partial sigmoid curve

Figure 3: Full and Partial Sigmoid Curves

3. CT Prediction Algorithm

The determination of the CT is done in two steps; a) the determination of negative samples, b) the prediction of CT value for positive samples.

In the algorithm 1, for a full sigmoid curve (with positive ct) the points X_s and X_e are marked on Figure 1b. X_s is the first point on the left side of the peak at which $y'(x) \geq max(y'(x)) \times 0.20$. X_e is the first point on the left side of the peak at which $y'(x) \geq max(y'(x)) \times 0.5$, respectively. Although, the constants 0.2 and 0.5 seem arbitary, these values are selected using a large set of The samples. Nevertheless, the algorithm is robust to minor changes in these values.

4. Results and Discussions

1536 samples are analyzed and used to develop a robust algorithm for CT prediction in diagnostic applications. Samples are selected such that they include both negative and positive results at different

Algorithm 1 CT Prediction Algorithm

```
procedure isSigmoid(x, y(x), y′(x), y″(x))
    $slope, intercept \leftarrow fitLine(x, y)$   [$y = slope \quad x + intercept$]
    if $slope$ <= 0.003 then return -1
    if $max(y'(x)$ <= 0.05 then return -1
    $i \leftarrow X_s$
    while $i \leq X_e$ do
        if y″($x_i$) == max(y″(x)) then
            $peakFound \leftarrow True$
        if peakFound and y″($x_i$) == 0 then return 1
    return 0

procedure predictCT
    $y(x) \leftarrow X_0, k, a, c \leftarrow$ scipy.optimize.leastsq$(x, y)$
    $y'(x) \leftarrow sigmoidFirstDerivative(x, X_0, k, a, c)$
    $y''(x) \leftarrow sigmoidSecondDerivative(x, X_0, k, a, c)$
    $result = isSigmoid(y(x), y''(x), y''(x))$
    if  not result == 1 then return result
                                    ▷ In case it is not determined then return undermined
    $i \leftarrow X_s$
    while $i \leq X_e$ do
        if y″($x_i$) == max(y″(x)) then
            $peakFound \leftarrow True$
        if peakFound and y″($x_i$) == 0 then return $x_s$
    $i \leftarrow X_s$
    while $i \leq X_e$ do
        if y″($x_i$) ¡ 0  then return -1          ▷ If not full or partial sigmoid curve then negative
    return $X_s$
```

concentrations such as shown in Figure 2a. These samples are grouped together so that in the first group with 192 samples no traces of target exists. In the following groups, the traces of target is decreased gradually which in turn resulted in an increase in the CT value. Therefore, according to the simulation setup, the algorithm should produce no CT value for the first group of samples and a decreasing CT for the remained.

Figure 2b show the maximum value of the first derivative calculated using Equation 3 for all samples. In the first set of samples, where no target exists; the maximum values of the first derivative is nearly zero. However, for the rest of the samples, the maximum value of the derivatives are consistently around 0.5 except where the traces of target is minute in the last group of samples.

Consequently, it was decided that using the first derivative results to establish a threshold value above 0.1 could safely identify the existence of target. Again, analyzing the data and setting a threshold value below 0.05 for the first derivative leads to the selection of negative samples. However, for the 25 cases among the last set of samples a negative CT value is unexpectedly predicted. When each of these cases are studied, the curves they produce are quite identical to the ones with no target. Therefore, the robustness of the first derivative threshold values is sufficient for the classification of positive and negative samples.

In addition to the first derivative threshold, a quick check based on the slope of the 1st degree polynomial fitted to the raw data is incorporated into the algorithm (See line 4 of the CT prediction Algorithm). Both lines 4 and 4 determine whether the curve produced by the raw data is sigmoid or not. For a number of cases, the algorithm on the other hand could not definitely classify them as sigmoid. These curves are determined to be mostly partially sigmoid. Therefore, additional logic is added into the algorithm that checks if the data could be represented by either the full or partial sigmoid curve (See algorithm lines 23-25).

Once the data is determined to be represented by either the partial or full sigmoid curve, the CT value is calculated by counting down the cycles starting from the peak of first derivate till 20% of the maximum peak of first derivate value is reached (See lines -). This is also the same point marked as X_s as shown in Figure 1b.

We validated our algorithm on a benchmark dataset that was used to test conventional RT-PCR data analysis approaches ([2]. The results suggested that in all the cases our algorithm predict the CT values correctly as per the experimental validation. However, in this dataset all samples were positive with varying CT values.

For diagnostic purposes, we evaluated the proposed algorithm using the data generated at Akdeniz University. For this test, 384 samples with replicates were evaluated. CT values were calculated using PCR Miner software online (http://miner.ewindup.info/) and our algorithm and the Results were summarized in Table 1.

Out of 384 PCR reactions tested, 17 of them were falsely predicted with a positive CT value by PCRminer. In most cases a CT value is assigned to both replicates falsely confirming the existence of pathogen in the given sample. In a typical epidemiological study this error could critically cause a misdiagnosis.

In these cases low availability of the mRNA, one could see a partial sigmoid curve with nearly flat noisy introductory region appended with a half sigmoid curve without the plateau region. In these cases of low availability mRNA, where PCR reaction is delayed (typically after cycle #35), PCRminer have not predicted a positive CT value. See Figure 3b.

	True Positive	True Negative	False Positive	False Negative
Expected	301	83	0	0
PCRminer	318	66	17	2
Proposed Algorithm	300	83	0	1

Table 1: Comparision of Proposed Algorithm with Gold Standard PCRminer

Figure 4 demonstrate CT prediction and expected values for a select set of cases. As given in Table 1. Our proposed algorithm predicts all true Positive without an error. It predicted all true negatives correctly except 1. The in correctly predicted false positive case is characterized by a partial sigmoid curve. To download and evaluate the software visit: http://baum.akdeniz.edu.tr/yayinlar.

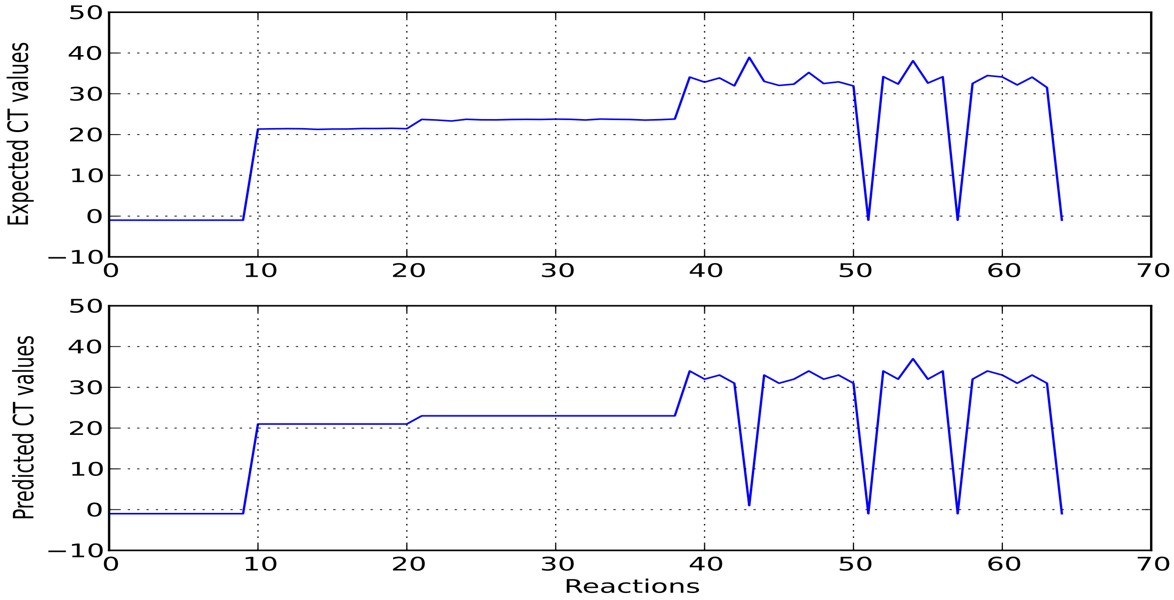


Figure 4: CT prediction using proposed algorithm vs Expected CT values

5. Conclusion

Several analytical and epidemiologic approaches exist to justify and select a cutoff based on evidence for real-time amplification experiments. Mostly, the justification made by the laboratory person may lead to the same cutoffs and that could lead to erroneous prediction of pathogens in an epidemiological set up. Although, epidemiologic cutoffs are population dependent, and their validity is directly associated with the targeted population. Cost-based determination of an optimal cutoff is likely to depend strongly on prevalence; hence, reliable prior prevalence information is needed.

Our algorithm could automate the process of selecting the best cutoff and reducing the error made by person selecting the CT temperature in an typical epidemiological setup.

References

[1] J. A. Goguen, L-fuzzy sets, Journal of Mathematical Analysis and Applications 18 (1967) 145–174.

[2] S. Peirsonö, Jason N. Butler, and Russell G. Foster Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis Nucleic Acids Research, 2003, Vol. 31, No. 14 e73

[3] Glenys R. Chidlow and Gerry B. Harnett and Geoffrey R. Shellam and David W. Smith, An Economical Tandem Multiplex Real-Time PCR Technique for the Detection of a Comprehensive Range of Respiratory Pathogens, 1:42-56, 2009.

[4] M. Kodani and et.al, Application of TaqMan Low-Density Arrays for Simultaneous Detection of Multiple Respiratory Pathogens, Journal of Clinical Microbiology, (49), 6:2175-2182, 2011.

[5] S.R. Kim and C. S. Ki and N. Y. Lee and et.al, Rapid detection and identification of 12 respiratory viruses using a dual priming oligonucleotide system-based multiplex PCR assay, Journal of Virology Methods, 156:111-116, 2009.

[6] C.A. Heid, J. Stevens, K.J. Livak, P.M. Williams, Real time quantitative PCR, Genome Research, (6), 10:986-994, 1996.

[7] W. Rudin, Real and Complex Analysis, (3rd edition), McGraw-Hill, New York, 1986.

[8] J. Vandesompele and et.al, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biology, (3), 7, 2002.

[9] T. Schmittgenö, and Kenneth J Livak, Analyzing real-time PCR data by the comparative CT method Nature Protocols 3, - 1101 - 1108,2008

[10] Burns Mö, Valdivia H, Modelling the limit of detection in real-time quantitative PCR. Eur Food Res Technol 226: 15131524.

[11] Wong MLö,, Medrano JF, Real-time PCR for mRNA quantitation. Biotechniques 39:7585.

[12] C. G. B.Caraguelö, Henrik Stryhn, Nellie Gagne, Ian R. Dohoo, K. Larry Hammell Selection of a cutoff value for real-time polymerase chain reaction results to fit a diagnostic purpose: analytical and epidemiologic approaches J Vet Diagn Invest 23:215, 2011

[13] Zhao Sø", Fernald.R.D Comprehensive algorithm for quantitative real-time polymerase chain reaction. J. Comput. Biol. 2005 Oct;12(8):1045-62

[14] Holland, ø" P.M., Abramson, R.D., Watson, R., and Gelfand, D.H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5f 3f exonuclease activity of Thermus aquaticus DNA polymerase. Proc. Natl. Acad. Sci. USA 88, 72767280

[15] Tichopadø", A., Dilger, M., Schwarz, G., and Pfaffl, M.W. 2003. Standardized determination of real-time PCR efficiency from a single reaction set-up. Nucl. Acids Res. 31, e122.

[16] Tichopad,ø" A., Didier, A., and Pfaffl, M.W. 2004. Inhibition of real-time RT-PCR quantification due to tissue-specific contaminants. Mol. Cell. Probes 18, 4550.

[17] Wittwer,ø" C.T., Gutekunst, M., and Lohmann, S. 2001. Method for quantification of an analyte. US 6,303,305 B1. United States.

[18] Tellinghuisen. Jø", Andrej-Nikolai Spiess. Comparing real-time quantitative polymerase chain reaction analysis methods for precision, linearity, and accuracy of estimating amplification efficiency Analytical Biochemistry 2014 Mar 15;449:76-82.