White paper [PRELIMINARY- March 12, 2021]

Artificial Intelligence for PCR Medical Machines

An innovative AI/ML platform deploys Real time PCR to provide reliable sequence detection and number of copies of DNA(CT) estimation.

Overview

Fluorescence-based Real-Time PCR (RT-PCR) is increasingly being used to detect multiple pathogens simultaneously and rapidly by gene expression analysis of PCR amplification data.
PCR data is often analyzed by estimating the cycle number at which the PCR product crosses a threshold of detection, called Threshold Cycle (CT) for **positive** samples. On the other hand, when such cross of threshold does not occur, the sample is identified as **negative**.

The amount of PCR product which 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 very simplicity can sometimes lead to conclusions that are either **false positive or negative**.

Bluewind and AB ANALITICA present here a stable and consistent alternative approach that is based on **ML-machine learning** for the definition and determination of CT values, dramatically decreasing overall rate of **false detections**.

Background

The PCR amplification process is performed by Thermocycler, which is an instrument used to amplify DNA and RNA samples by the polymerase chain reaction. The thermocycler raises and lowers the temperature of the samples in a holding block in discrete, pre-programmed steps, allowing for denaturation and reannealing of samples with various reagents. In ideal conditions, PCR product grows exponentially. Because of presence of impurities and since it takes several cycles for enough product to be readily detectable, the plot of fluorescence vs cycle number of the PCR amplification process 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 exhibits a sigmoidal appearance as shown in Figure 1 and given by logistic function.

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 (Ct value). This somewhat arbitrarily set cycle threshold value must intersect the signal curve in its exponential phase as demonstrated in the following Figure 1.

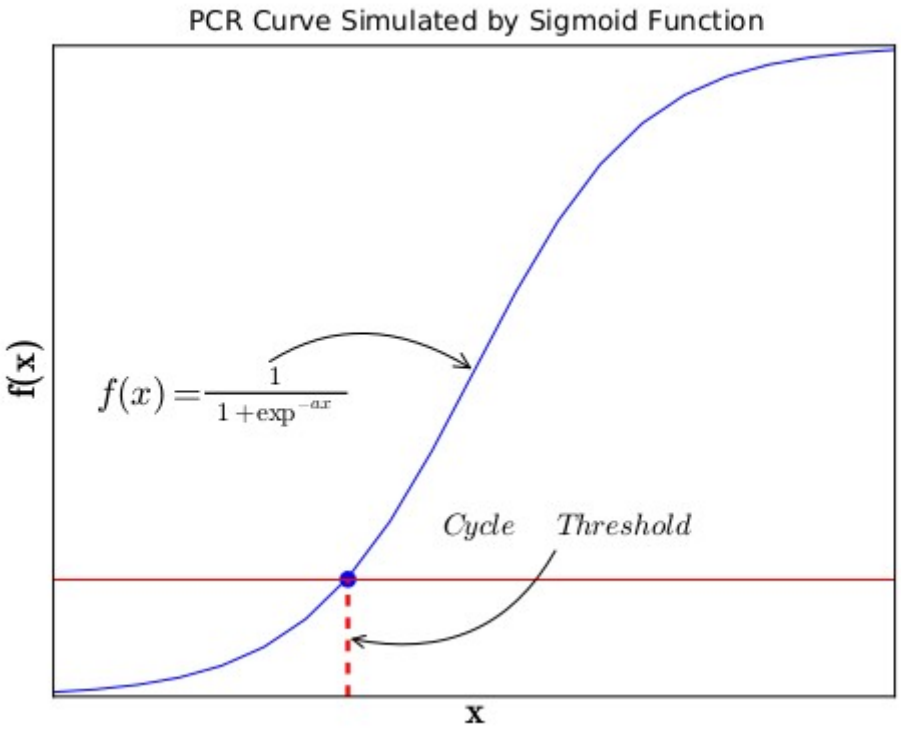


Figure 1

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. 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.

New standard amplification curve normalization method

The input data is its raw fluorescence read by one of the real thermocyclers. The preliminary operation is the normalization of the amplification curves, specifically they must be brought to a common value, so as to be all superimposed.

The algorithm works as follows, for each amplification curve:

1. Averaging curves to decrease the noise amplitude: for each cycle c, the software computes the average value of the raw curve values from cycle c-4 to c+4. Then, by concatenating the obtained values, there will be a new averaged curve
2. Linear regression: for each cycle c, the software computes the second order gradient and its standard deviation of the averaged raw curve values from cycle c to c+5. Then, by searching the cycles where the associated standard deviation is below a predefined threshold, it is possible to perform a linear regression from the lowest and highest cycle that fulfill the condition.
3. From the interpolation of these points a straight line is obtained, y = mx + q;
4. The values of this line are calculated for each cycle;
5. Finally, the points of the curve are subtracted from those of the straight line.

Through these steps, all the curves will have the value 0 RFU (relative florescence units) as baseline.

In this Figure 2, there is a representation of raw fluorescence.

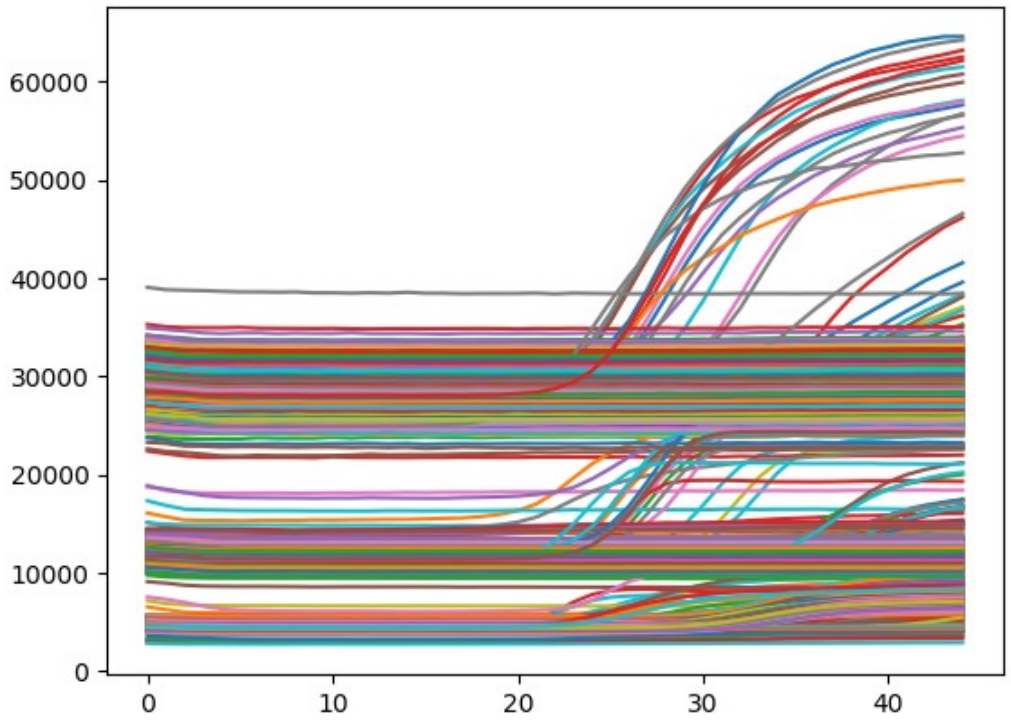


Figure 2

In this Figure 3, there is a representation of raw fluorescence after normalization.

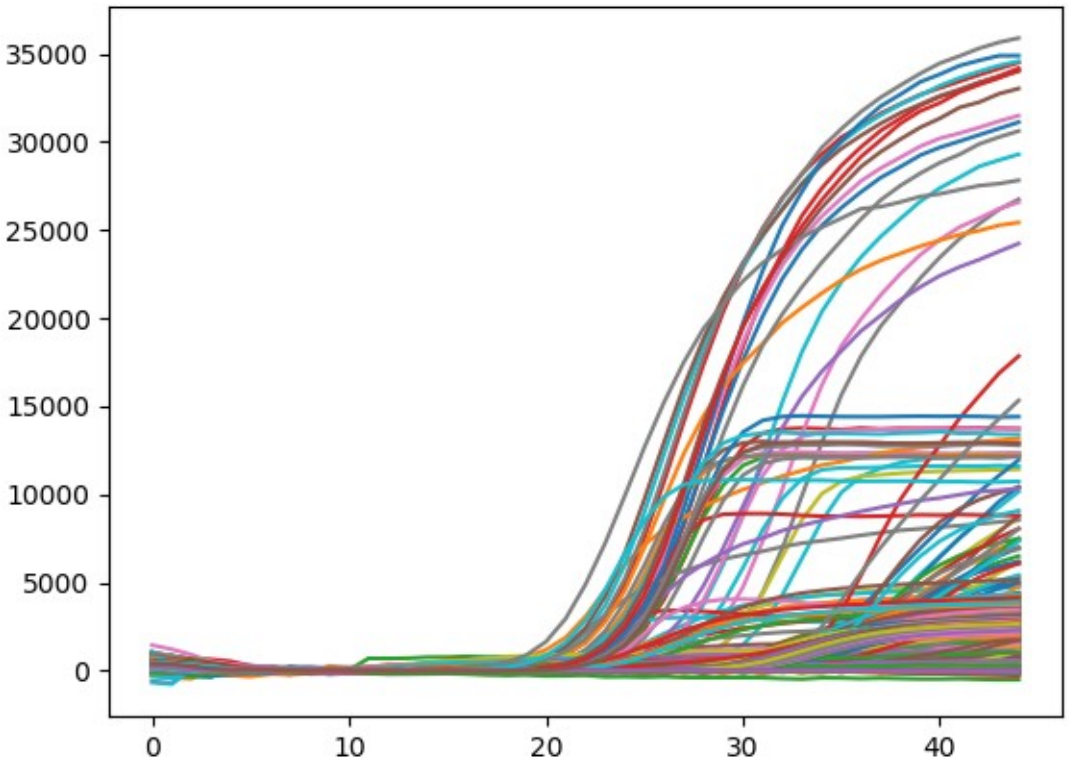


Figure 3

Qualitative PCR based on ML
A qualitative PCR is used to detect the presence or absence of a sequence. For some applications, qualitative nucleic acid detection is sufficient (example: Covid-19). Today with the spread of Covid-19, the qualitative nucleic acid detection must be performed with high accuracy and in real time.

From prior analysis, it has been easy to find that detecting the presence or the absence of a sequence is a classification problem. Since different target sequences have different effects on PCR product, some simple machine learning methods can separate these known activities (presence or absence) based on fluorescence data.

Machine Learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Today, a vast proportion of ML-based applications are supported by PC platforms, cloud computing or powerful smartphones.

The proposed solution involves a machine learning approach and demonstrates how data-driven approaches meet the challenge of reaching high accuracy while being light and computationally efficient.

The proposed application's task is to detect the presence or the absence of a target sequence within a sample.

The application is designed to output the sample state (negative or positive) from fluorescence data streams coming from a real thermocycler.

The actual computation is performed by a deep learning pipeline composed of two main items:

- a denoising algorithm used to remove noise from the data stream
- a convolutional neural network aimed to classify the input data

The experiments show that by using the proposed mechanism, the accuracy is around 99%. Thus, the proposed method performed a qualitative PCR in real time and with high accuracy.

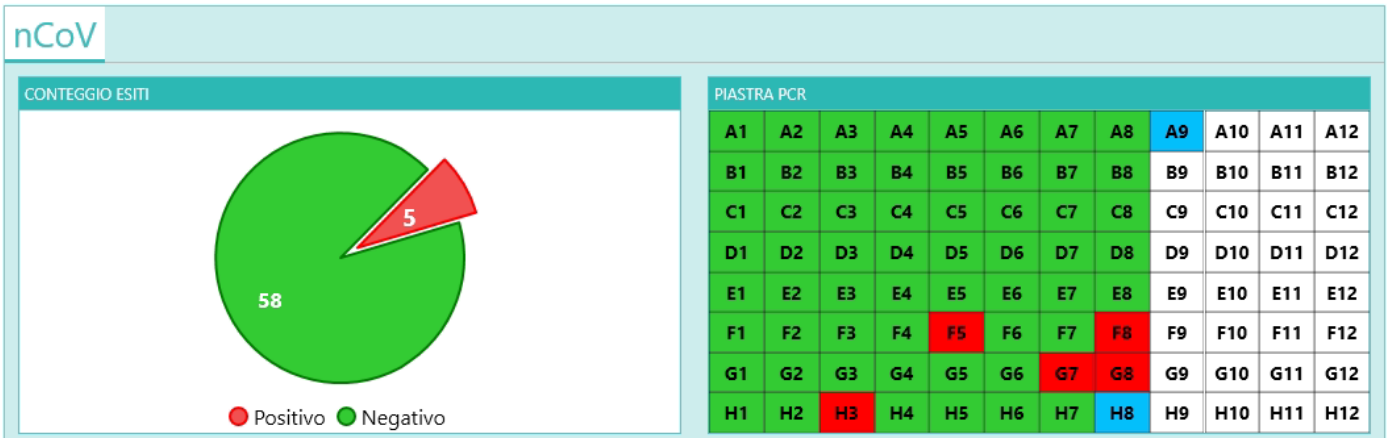


Figure 4

Quantitative PCR

Real-time PCR results can either be qualitative (presence or absence of a sequence) or quantitative (number of copies of DNA). Real-time PCR that is quantitative is also known as qPCR.

The CT of a reaction is determined mainly by the amount of template present at the start of the amplification reaction. If a large amount of template is present at the start of the reaction, relatively few amplification cycles will be required to accumulate enough product to give a fluorescent signal above background.

Thus, the reaction will have a low, or early, CT. In contrast, if a small amount of template is present at the start of the reaction, more amplification cycles will be required for the fluorescent signal to rise above background. Thus, the reaction will have a high, or late, CT. This relationship forms the basis for the quantitative aspect of real-time PCR.

In this study, the proposed approach performs quantitative PCR after doing a qualitative analysis. If a sample was classified positive, it is possible to proceed with the quantitative analysis, otherwise no. This approach is useful to save computational time.

For the quantitative analysis, two methods are proposed:

Analytical Method

The Ct values are obtained from the intersection of the amplification curves with the straight threshold, which has a different value depending on the thermocycler and the target gene, such that it intersects the curve at the beginning of its exponential phase.

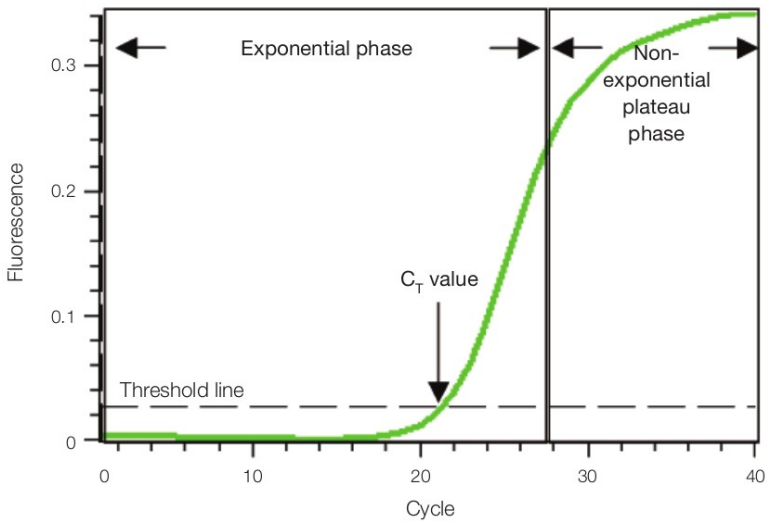
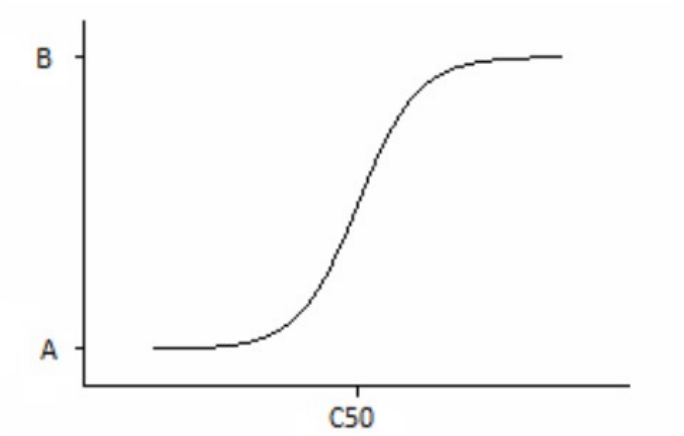


Figure 5

Since the values of the amplification curves are discrete, model fitting is required to find the points of intersection with the threshold. The model used is Boltzmann (4-parameter Sigmoid).



$$y = A + \frac{B - A}{1 + e^{\left(\frac{C50 - x}{Slope}\right)}} \qquad x = C_{50} - Slope \log\left(\frac{B - y}{y - A}\right)$$

Figure 6

ML based method

In order to compute the CT value, the proposed process exploits a deep learning method. This method is based on a *convolutional neural network* (*CNN*, or ConvNet which is a class of deep *neural networks*) regression model based on data collected from different thermocyclers, so it is possible to build a robust ANN regression model.

Several different experimentation of this method show very promising results.

Nevertheless, this method is currently under improvement in order to reach a smaller error rate.

Pozzetto	Nome	Tipo	IC	Esito	Ct [FAM - RdRp]	Ct [HEX - IC RNA]	Ct [Cy5 - E]
F1		Campione	Idoneo	Negativo	0	29,04	0
F2		Campione	Idoneo	Negativo	0	29,19	0
F3		Campione	Idoneo	Negativo	0	28,94	0
F4		Campione	Idoneo	Negativo	0	28,96	0
F5		Campione	Idoneo	Positivo	29,82	28,69	30,79
F6		Campione	Idoneo	Negativo	0	28,94	0
F7		Campione	Idoneo	Negativo	0	28,83	0
F8		Campione	Idoneo	Positivo	39,44	28,71	39,48
G1		Campione	Idoneo	Negativo	0	28,81	0
G2		Campione	Idoneo	Negativo	0	29,2	0

Figure 7

Conclusions

The proposed platform uses AI- Artificial Intelligence to analyze the raw data of PCR (Polymerase Chain Reaction) from multiple commercial PCR devices, resulting in a more reliable measurement, accurate results and almost negligible calculation time.

The platform is used in molecular laboratories to dramatically improve diagnostic quality and decrease the overall computing time of results, effectively allowing laboratories to automate their initial technical validation with software that can analyze curves in exactly the same way as expert laboratory scientists.

About Bluewind

Bluewind, an independent engineering company, provides world-class products, engineering and software solutions in the domains of electronics, safety critical applications, and connected devices.
As a qualified researcher for Artificial Intelligence technologies, Bluewind is actively involved in designing next generation products in the Automotive, Industrial and Medical industries.

About AB ANALITICA

AB ANALITICA is specialized in R&D and production of advanced molecular diagnostic solutions based on Real-Time PCR, Reverse line blot, and NGS technologies.

We produces a vast range of CE-IVD diagnostic tests for respiratory infections, transplantation, HIV DNA, sexually transmitted infections, HPV screening and genotyping, oncohematology and genetics.

Our experience in biotechnology combined with strong skills in engineering and IT, has allowed the development of a branded fully automated Real Time PCR platform the GENEQUALITY™ X120 which offer an innovative molecular diagnostic solution for the labs.

Recently we have implemented another important turning point for Real Time PCR quantitative tests: our V-Quant™ System allows an immediate Nucleic acid Quantification for every sample without the classical standards use.

Everyday we innovate, imagine and grow with our Customers.

AB ANALITICA Srl. Via Svizzera 16, 35127 Padova (Italy), www.abanalitica.com

Bluewind Srl

Via della Borsa, 16/A - 31033
Castelfranco Veneto (TV) - Italy
+39 0423 723431 - info@bluewind.it

www.bluewind.it