ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Detection and quantification in real-time polymerase chain reaction

Abou Keita[1,2], Romain Hérault[1,2], Colas Calbrix[1,3] and Stéphane Canu[1,2]

1 - Normandie Univ, France

2 - INSARouen, LITIS, F-76801 Saint Etienne du Rouvray, France
{abou.keita,romain.herault,stephane.canu}@insa-rouen.fr

3 - Université de Rouen, PRIMACEN, F-76821 Mont Saint-Aignan, France
colas.calbrix@univ-rouen.fr

Abstract. The estimation of the concentration of an infectious agent in the environment is a key step to trigger an alert when there is a biological threat. This concentration can be obtained trough a quantitative polymerase chain reaction (qPCR). Nevertheless, standard real-time procedure do not address detection delay which is a main concern in alert triggering. Therefore, we propose a method based on Lasso regression and CUSUM change detection to accurately estimate the concentration while minimizing the detection delay. The trade-off between accuracy and delay can be managed through a parameter. We compare our results with those found by a standard method (threshold method) and promising results are obtained.

Keywords: Real-time PCR, quantitative PCR (qPCR), Change detection, Detection delay, CUSUM, Lasso, Biological threat.

1 Context

A sample of interest is taken from environment in order to be tested by *Polymerase Chain Reaction* (PCR). At each reaction cycle, genetic information supports are doubled [1, 2]. When a fluorophore matching specific genetic information from an infectious agent reaches its target, it emits light. Thus at each cycle of PCR, fluorescent light signal increases as genetic information increases (fig. 1(a)). This signal has three steps: baseline, exponential and plateau. The break time between baseline and exponential steps is log-linear to the initial concentration of targeted agent [3, 1]. Thus if we know accurately this break time (or specific cycle), we can compute the initial concentration of the agent in the sample. Moreover, in the context of alert triggering, the time between this break and the actual detection, that is detection delay, should be short.

Among methods described for specific cycle (change) detection in Biology works, we can cite: a) The **threshold method** in which the specific cycle corresponds to the last time the fluorescence curve intersects a threshold [2, 1].

Authors would like to thanks the Agence Nationale de la Rercherche (ANR) for its support through the Genetic EquipemeNt for biothrEat environmental Analysis and SurveillancE (GENEASE) project and our partners Bertin technologies, CEA LETI and CEA SBTN.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

b) The **second derivative method** in which the specific cycle is defined by the maximum of the fluorescence curve second derivative [1]. c) The **sigmoid curve fitting method** in which the specific cycle is linked to a parameter of a curve model [3]. None of them study the alert delay. Methods described for change (specific cycle) detection in Signal Processing can be split into two families: a) **Off-line methods**. For example, derivation method studied in [4] or the sigmoid curve fitting method (see above). Here, the full knowledge of the signal is assumed before the decision is taken. b) **On-line methods**. For example, the Shewhart rule, the moving average rule, the CUSUM method [4, 5]. Here, observations arrive continuously, alert can be triggered as soon as possible. Under this latter setting, a trade off between precision on change detection and alert delay can be made accordingly to the application context.

Our proposal: To address detection delay meanwhile taking into account accuracy in quantitative real-time PCR, we use an on-line CUSUM method for fluorescence change detection. A kernelized-Lasso regression is done as a preprocessing step to overcome signal drift and to get rid of outlier samples.

Results will be compared with the threshold method included in the apparatus from which our data are recorded.

2 CUSUM

2.1 Introduction

To detect a potential break point t_0 from a time series in a sequential manner Page [6] introduced the CUSUM method that optimizes the detection delay. It is a statistical test between two hypothesis: no break point ($\mathcal{H}_0$) or break point ($\mathcal{H}_1$) occurs. Under ($\mathcal{H}_0$) observed data is assumed to be i.i.d. from some distribution P_{θ_0} where θ_0 is a parameter while under ($\mathcal{H}_1$) observed data underlying distribution parameter change from θ_0 to θ_1 at break time r that is

$$(\mathcal{H}_0):\ X = \{X_t\}_{t=1,\dots,s} \ \square\ P_{\theta_0} \quad \text{vs} \qquad (\mathcal{H}_1):\ \begin{array}{l} \exists\ r \text{ such that} \\ \{X_t\}_{t=1,\dots,r-1} \ \square\ P_{\theta_0} \\ \{X_t\}_{t=r,\dots,s} \ \square\ P_{\theta_1} \end{array}$$

where s is the length of the signal. Thus, under null and alternative hypothesis, data distributions are

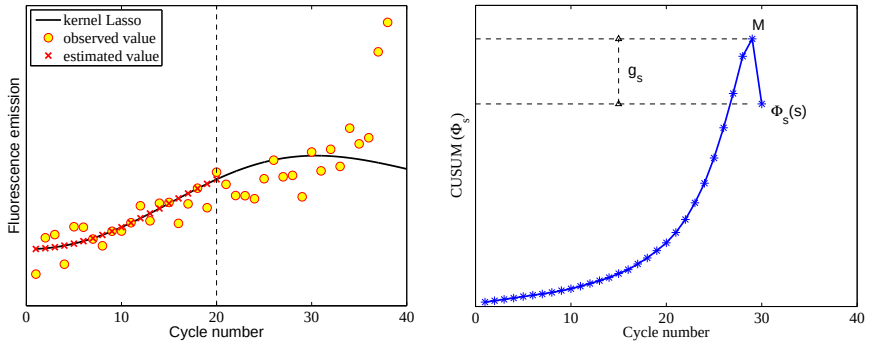
$$L(X|\mathcal{H}_0) = \prod_{t=1}^{s} P_{\theta_0}(X_t), \quad L(X|\mathcal{H}_1) = \prod_{t=1}^{r-1} P_{\theta_0}(X_t) \prod_{t=r}^{s} P_{\theta_1}(X_t)\ , \tag{1}$$

leading to the following log-ratio [4, 5]:

$$S(X) = \log \frac{L(X|\mathcal{H}_1)}{L(X|\mathcal{H}_0)} = \log \frac{\prod_{t=1}^{r-1} P_{\theta_0}(X_t) \prod_{t=r}^{s} P_{\theta_1}(X_t)}{\prod_{t=1}^{s} P_{\theta_0}(X_t)} = \sum_{t=r}^{s} \log \frac{P_{\theta_1}(X_t)}{P_{\theta_0}(X_t)}\ . \tag{2}$$

$S(x)$ as a function of unknown break time r is denoted by $\Phi_s(r)$, the CUSUM indicator. The distributions P_{θ_0} and P_{θ_1} are assumed to follow Gaussian distri-

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.



(a) Kernelized-lasso regression (b) CUSUM indicator Φ_s

Fig. 1: Preprocessing and CUSUM steps

butions respectively parametrized by (μ_0, σ_0) and (μ_1, σ_1). In our application context, providing by experts that breaks do not occurs until $s_{min} = 20$, $(\hat{\mu}_0, \hat{\sigma}_0)$ are estimated on $\{x_t\}_{t=1,\ldots,s_{min}}$, $\hat{\sigma}_1$ is set to $\hat{\sigma}_0$, $\hat{\mu}_1$ is estimated on $\{x_t\}_{t=r,\ldots,s}$.

As r is unknown, detection is based on M, the *log* maximum of the generalized likelihood ratio,

$$M = \max_{1 \leq r \leq s} \Phi_s(r) \ . \tag{3}$$

As we see from the figure 1(b), the indicator increases until it reachs M and then decreases. We perform the following statistic test $g_s = M - \Phi_s(s)$. If g_s is below a chosen threshold h, the signal is in $(\mathcal{H}_0)$ state; otherwise, we have switched to $(\mathcal{H}_1)$ state. Here, an alert is triggered (i.e. the alarm time t_a equals s) and the break point, t_0, is given by the position of M, that is, $t_0 = \arg \max_{1 \leq r \leq t_a} \Phi_s(r)$.

2.2 Application to fluorescence signals obtained by PCR

CUSUM can not be applied directly to a fluorescence signal. In our application context, the fluorescence is not i.i.d. and outliers are encountered at the beginning of the signal, disturbing the estimation of (μ_0,σ_0). That is why a kernelized-lasso regression [7] is estimated on the twenty first observations by solving the following linear program

$$\underset{\boldsymbol{\alpha}}{argmin} \ ||\mathbf{y} - \boldsymbol{\alpha} K||_1 + \lambda ||\boldsymbol{\alpha}||_1 \ , \tag{4}$$

where $\boldsymbol{\alpha}$ is coefficient vector of size s_{min}, λ the regularization parameter, K the Gaussian kernel matrix and $\mathbf{y}$ the s_{min} first samples where $s_{min} = 20$. Figure 1(a) illustrates on an example the way kernel lasso regression is used.

Moreover, the DNA duplication reaction is continuous but fluorescence is not: fluorophore need light excitation in order to, at there turn, emit light. This excitation process takes one minute (a cycle) which is far from the precision needed to quantify the concentration. We need a better time precision than the

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

sampling rate! This is overcome by using a spline approximation of the CUSUM indicator, then, the break time (or specific cycle) t_0 is redefined to the spline maximum.

Algorithm 1 CUSUM algorithm for real-time qPCR

Pre-processing :
Wait for s_{min} samples
Compute kernel regression on s_{min} samples then (μ_0,σ_0) on residuals
Main algorithm:
$s \leftarrow s_{min} + 1$, $decision \leftarrow 0$
while $decision == 0$ **do**
  Compute Φ_s
  $[M, position] \leftarrow \max(\Phi_s)$
  $g_s \leftarrow M - \Phi_s(s)$
  if $g_s >= h$ **then**
    $decision \leftarrow 1$, $t_a \leftarrow s$, $t_0 \leftarrow position$
  else
    $s \leftarrow s + 1$ (Wait for a new sample)
  end if
end while
Post-processing:
$S\Phi_s \leftarrow spline(\Phi_s(t_0 - 2 : t_0 + 1))$
$t_0 \leftarrow arg\max S\Phi_s$

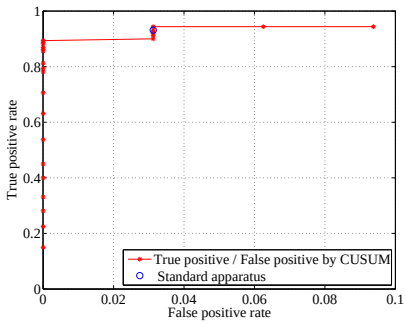
3 Experimentation

3.1 Set-up

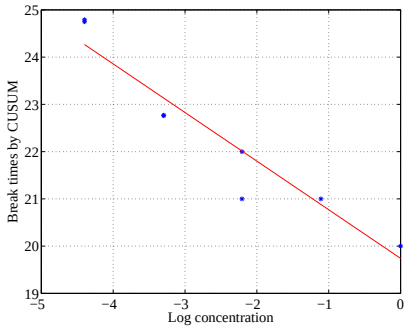
Fluorescence signals have been recorded on a standard apparatus, 7500 Fast Real-Time PCR System of Applied Biosystems, which itself gives the specific cycle by a simple threshold method at the end of the record. It can process multiple samples in plate consisting in 8 lines by 12 columns of wells. Materials of different kinds and at different concentrations are analysed in the wells: Each line containing only one combination of tested material and fluorophore; Columns (2 by 2) containing a specific concentration. The two last columns are negative. The record of two plates have been processed, that is 192 signals (160 positive and 32 negative).

As the threshold method and the CUSUM method do not look for the same event in the signal, results can not be compared directly. Nevertheless, line by line (material wisely), they should be linear together. Eventually, line by line, CUSUM results and threshold results should both be log-linear to the concentration. We will first consider results in term of detection (true positive versus false positive) and, in a second step, look for the accuracy of the detection (log-linearity to concentration) and detection delay only on true positive examples. Accuracy is evaluated on the mean of residuals of log-regression performed line by line. A trade-off between these two indicators can be made trough the h parameter in the CUSUM method.

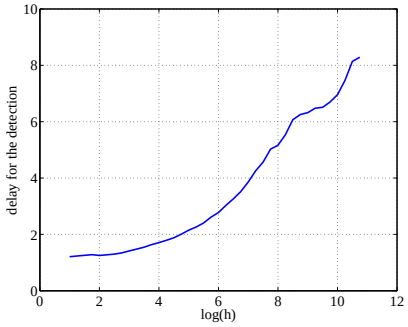
ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.



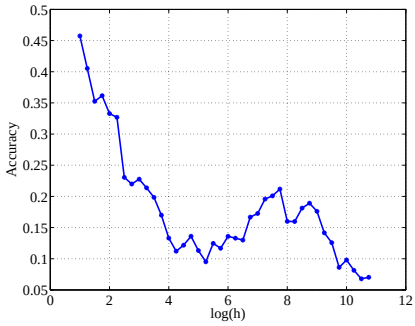
(a) Curve operational



(b) Break times by the CUSUM with respect to the log concentration on online for $h = exp(1.5)$.



(c) Delay detection in cycles with respect to h with CUSUM method (compared to 13.88 cycles for apparatus)



(d) Accuracy with respect to h with CUSUM method (compared to 0.77 for apparatus)

Fig. 2: False alarm rate compared to the rate of correct detection, delay, accuracy

3.2 Results

Standard apparatus has a fixed point operation that gives 93.13% of true positive rate with 3.12% of false alarm. On the true positive examples, the mean residual of the log-regression between break time and log-concentration is 0.77 and the detection delay 13.88 cycles.

The best operation point is obtained by CUSUM method at $h = 1.5$. The proposed CUSUM method gives 94.37% of true positive rate with 3.12% of false alarm. At this operation point, on the true positive examples, the mean residual of the log-regression between break time and log-concentration is 0.35 and the detection delay 1.26 cycles. The CUSUM method is on all indicators better than the standard apparatus. Especially, it reduces by 10 the detection delay. An example of log-linearity on a specific plate line is shown on figure 2(b).

By changing h, we can draw an operational curve (fig. 2(a)) in term of true positive rate and false alarm rate. Moreover h influences the detection delay (fig.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2(c)) and accuracy (fig. 2(d)). From those two last curves we can induced two possible strategies: A) Taking h between $exp(4)$ and $exp(6)$ that leads to good accuracy with a delay inferior to 2 cycles; B) Or taking h superior to $exp(10)$ that leads to an even better accuracy at the price of a higher delay.

For pratical use, we can recomend to use the operation point closer to 100% true positive and 0% false positive. Error estimation can be estimed through cross validation.

4 Conclusion

The determination of a characteristic cycle of the fluorescence from qPCR can be done in continuous time and without prior knowledge of the total signal through a CUSUM method. Care must be taken to estimate statistical models at the beginning of the signal: a kernelized lasso regression is performed to avoid outliers. We obtain similar results in term of true positive and false alarm rates as standard apparatus but with better performance in term of concentration estimation accuracy and detection delay. Moreover, the method is easily tunable to application context (delay vs accuracy) trough a trade-off parameter.

References

[1] J.D. Durtschi, J. Stevenson, W. Hymas, and K.V. Voelkerding. Evaluation of quantification methods for real-time PCR minor groove binding hybridization probe assays. *Analytical biochemistry*, 361(1):55–64, 2007.

[2] M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, B. Sjögreen L. Strömbom A. Stahlberg Sindelka, R. Sjöback, and N. Zoric. The real-time polymerase chain reaction. *Molecular aspects of medicine*, 27(2-3):95–125, 2006.

[3] R.G Rutledge. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research*, 32(22):e178, 2004.

[4] M. Basseville and I.V. Nikiforov. *Detection of abrupt changes: theory and application*, volume 10. Prentice-Hall, 1993.

[5] G. Verdier, N. Hilgert, and J.P. Vila. Optimality of cusum rule approximations in change-point detection problems: application to nonlinear state–space systems. *Information Theory, IEEE Transactions on*, 54(11):5102–5112, 2008.

[6] E.S Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[7] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.