K-MEANS CLUSTERING FOR DNA COMPUTING READOUT METHOD IMPLEMENTED ON LIGHTCYCLER SYSTEM

[1]Muhammad Faiz Mohamed Saaid, [1]Zuwairie Ibrahim, [1]Marzuki Khalid, and [2]Nor Haniza Sarmin

[1]Center for Artificial Intelligence and Robotic, Department of Mechatronic and Robotic, Faculty of Electrical Engineering, Universiti Teknologi Malaysia

Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia

ABSTRACT

In the previous paper, a readout approach for the Hamiltonian Path Problem (HPP) in DNA computing based on the real-time polymerase chain reaction (PCR) was proposed. Based on this approach, real-time amplification was performed with the TaqMan probes and the TaqMan detection mechanism was exploited for the design and development of the readout approach. The readout approach consists of two steps: real-time amplification *in vitro* using TaqMan-based real-time PCR, followed by information processing *in silico* to assess the results of real-time amplification, which in turn, enables extraction of the Hamiltonian path. However, the previous method used manual classification of two different output reactions of real-time PCR. In this paper, K-means clustering algorithm is used to identify automatically two different reactions in real-time PCR. It is shown that K-means clustering technique can be implemented for clustering output results of DNA computing readout method based on LightCycler System.

1. INTRODUCTION

Since the discovery of the polymerase chain reaction (PCR) (Mullis, 1986), numerous applications have been explored, primarily in the life sciences and medicine, and importantly, in DNA computing as well. The subsequent innovation of real-time PCR has rapidly gained popularity and plays a crucial role in molecular medicine and clinical diagnostics (Overbergh, 2003). All real-time amplification instruments require a fluorescence reporter molecule for detection and quantitation, whose signal increase is proportional to the amount of amplified product. Although a number of reporter molecules currently exist, it has been found that the mechanism of the TaqMan hydrolysis probe is very suitable for the design and development of a readout method for DNA computing, and is thus selected for the current study.

A TaqMan DNA probe is a modified, nonextendable dual-labeled oligonucleotides. The 5' and 3' ends of the oligonucleotide are terminated with an attached reporter, such as FAM, and quencher fluorophores dyes, such as TAMRA, respectively, as shown in Figure 1 (Walker, 2003). Upon laser excitation at 488 nm, the FAM fluorophore, in isolation emits fluorescence at 518 nm. Given proximity of the TAMRA quencher, however, based on the principle of fluorescence resonance energy transfer (FRET), the excitation energy is not emitted by the FAM fluorophore, but rather is transferred to TAMRA via the dipole-dipole interaction between FAM and TAMRA. As TAMRA emits this absorbed energy at significantly wavelengths (580 nm), the resulting fluorescence is not observable in Channel 1 of real-time PCR instruments (Lakowicz, 1999).

The combination of dual-labeled TaqMan DNA probes with forward and reverse primers is a must for a successful real-time PCR. As PCR is a repeated cycle of three steps (denaturation, annealing, and polymerization), a TaqMan DNA probe will anneal to a site within the DNA template in between the forward and reverse primers during the annealing step, if a subsequence of the DNA template is complementary to the sequence of the DNA probe. During polymerization, *Thermus aquaticus (Taq)* DNA polymerase will extend the primers in a 5' to 3' direction. At the same time, the *Taq* polymerase also acts as a "scissor" to degrade the probe via cleavage, thus separating the reporter from the quencher, as shown in Figure 2 (Heid, 1996), where R and Q denote the reporter dye and quencher dye, respectively. This separation subsequently allows the reporter to emit its fluorescence (Holland, 1991). This process occurs in every PCR cycle and does not interfere with the exponential accumulation of PCR product. As a result of PCR, the amount of DNA template increases exponentially, which is accompanied by a proportionate increase in the overall fluorescence intensity emitted by the reporter group of the excised TaqMan probes. Hence, the intensity of the measured fluorescence at the end of each PCR polymerization is correlated to the total amount of PCR product, which can then be detected, using a real-time PCR instrument for visualization.

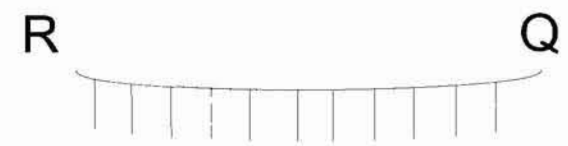

Fig. 1 Illustration of the structure of a TaqMan DNA probe. Here, R and Q denote the reporter and quencher fluorophores, respectively.

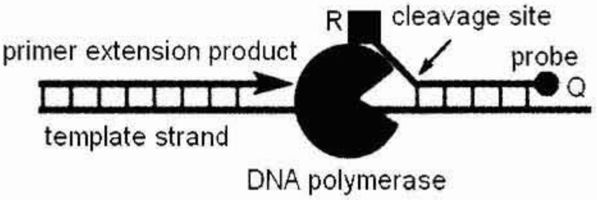

Fig. 2 Degradation of a TaqMan probe, via cleavage by DNA polymerase.

Previously, we proposed a readout method tailored specifically to the HPP in DNA computing, which employs a hybrid *in vitro-in silico* approach (Ibrahim, 2006). In the *in vitro* phase, $O(|V|^2)$ TaqMan-based real-time PCR reactions are performed in parallel, to investigate the ordering of pairs of nodes in the Hamiltonian path of a $|V|$-node instance graph, in terms of relative distance from the DNA sequence encoding the known start node. The resulting relative orderings are then processed *in silico*, which efficiently returns the complete Hamiltonian path. The proposed approach is experimentally validated optical method specifically designed for the quick readout of HPP instances, in DNA computing. Previously, graduated PCR, which was originally demonstrated by Adleman (Adleman, 1994), was employed to perform such operations. While a DNA chip based methodology, which makes use of biochip hybridization for the same purpose has been proposed (Rose, 1997, Wood, 1998, Wood, 1999), this method is more costly, and has yet to be experimentally implemented.

As shown in Figure 3, the output of the real-time PCR consist of two kinds of reactions, namely "YES" reaction and "NO" reaction. However, the amplification for the "NO" reactions appear as an amplification-like signal which make the interpretation of amplification response more difficult to be done. In this study, we utilize K-means algorithm to cluster the output results of real-time PCR, followed by additional algorithms to classify "YES" and "NO" reactions of the real-time PCR.

2. READOUT APPROACH OF DNA COMPUTATION BASED ON REAL-TIME PCR

2.1 Notation and Basic Principle

First of all, $v_{1(a)}v_{2(b)}v_{3(c)}v_{4(d)}$ denotes a double-stranded DNA (dsDNA) which contains the base-pairs subsequences, v_1, v_2, v_3, and v_4, respectively. Here, the subscripts in parenthesis (a, b, c, and d) indicate the length of each respective base-pair subsequence. For instance, $v_{1(a)}$ indicates that the length of the double-stranded subsequence, v_1 is 20 base-pairs (bp). When convenient, a dsDNA may also be represented without indicating segment lengths (*e.g.*, $v_1v_2v_3v_4$).

A reaction denoted by TaqMan(v_0,v_k,v_l) indicates that real-time PCR is performed using forward primer v_0, reverse primer v_l , and TaqMan probe v_k . Based on the proposed approach, there are two possible reaction conditions regarding the relative locations of the TaqMan probe and reverse primer. In particular, the first condition occurs when the TaqMan probe specifically hybridizes to the template, between the forward and reverse primers, while the second occurs when the reverse primer hybridizes between the forward primer and the TaqMan probe. As shown in Figure 3, these two conditions would result in different amplification patterns during real-time PCR, given the same DNA template (*i.e.*, assuming that they occurred separately, in two different PCR reactions). The higher fluorescent output of the first condition is a typical amplification plot for real-time PCR. In contrast, the low fluorescent output of the second condition reflects the cleavage of a few of the TaqMan probes via DNA polymerase due to the 'unfavourable' hybridization position of the reverse primer. Thus, TaqMan(v_0,v_k,v_l) = YES if an amplification plot similar to the first condition is observed, while TaqMan(v_0,v_k,v_l) = NO if an amplification plot similar to the second condition is observed.

2.2 The *in vitro* Phase

Let the output of an *in vitro* computation of an HPP instance of the input graph be represented by a 120-bp dsDNA $v_{0(20)}v_{2(20)}v_{4(20)}v_{1(20)}v_{3(20)}v_{5(20)}$, where the Hamiltonian path $V_0 \rightarrow V_2 \rightarrow V_4 \rightarrow V_1 \rightarrow V_3 \rightarrow V_5$, begins at node V_0, ends at node V_5, and contains intermediate nodes V_2, V_4, V_1, and V_3, respectively. Note that in practice, only the identities of the starting and ending nodes, and the presence of all intermediate nodes will be known in advance to characterize a solving path. The specific order of the intermediate nodes within such a path is unknown.

The first part of the approach, which is performed *in vitro*, consists of $[(|V|-2)^2-(|V|-2)]/2$ real-time PCR reactions, each denoted by TaqMan(v_0,v_k,v_l) for all k and l, such that $0 < k < |V|-2$, $1 < l < |V|-1$, and $k < l$. For this example instance, so that the DNA template is dsDNA $v_0v_2v_4v_1v_3v_5$, these 6 reactions are as follows:

(1) TaqMan(v_0,v_1,v_2) = NO
(2) TaqMan(v_0,v_1,v_3) = YES
(3) TaqMan(v_0,v_1,v_4) = NO
(4) TaqMan(v_0,v_2,v_3) = YES
(5) TaqMan(v_0,v_2,v_4) = YES
(6) TaqMan(v_0,v_3,v_4) = NO

Note that the overall process consists of a set of parallel real-time PCR reactions, and thus requires $O(1)$ laboratory steps for in vitro amplification. The accompanying SPACE complexity, in terms of the required number of tubes is $O(|V|^2)$.

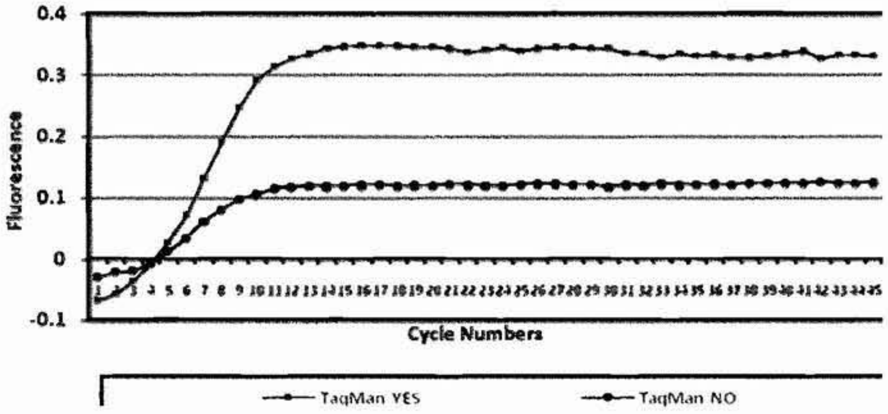

Fig. 3 An example of reaction plots corresponding to TaqMan(v_0,v_k,v_l) = YES (first condition) and TaqMan(v_0,v_k,v_l) = NO (second condition).

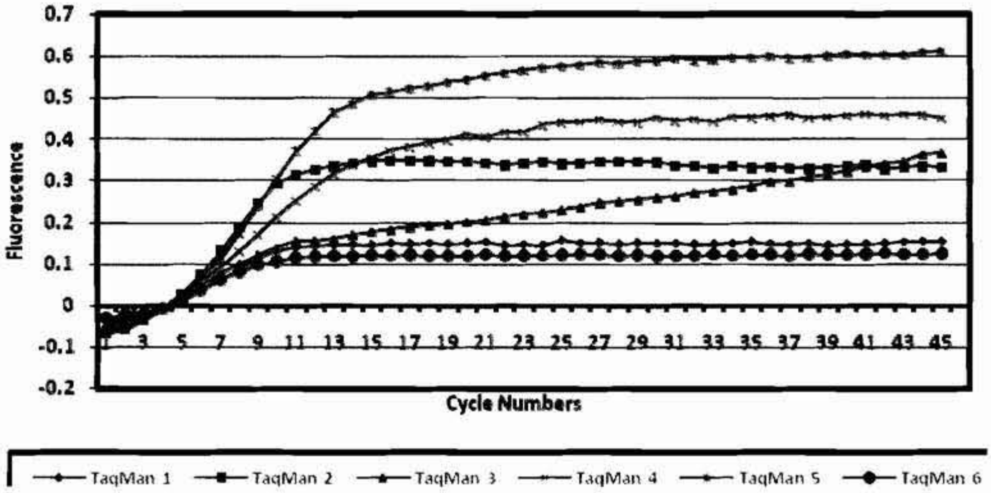

Fig. 4 Output of real-time PCR and grouping of output signals into two regions: amplification region (YES), and non-amplification region (NO).

Clearly, only one forward primer is required for all real-time PCR reactions, while the number of reverse primers and TaqMan probes required with respect to the size of input graph are each $|V|$-3.

The real-time PCR reaction involves primers (Proligo, Japan), TaqMan probes (Proligo, Japan), and LightCycler TaqMan Master (Roche Applied Science, Germany). Six separate real-time PCR reactions, including a negative control were performed, in order to implement the first stage of the HPP readout. The amplification consists of 45 cycles of denaturation, annealing, and extension, performed at 95°C, 48°C, and 72°C, respectively. The resulting real-time PCR amplification plots are illustrated in Figure 4.

After all real-time PCR reactions are completed, the *in vitro* output is subjected to an algorithm for *in silico* information processing, producing the satisfying Hamiltonian path of the HPP instance in $O(n^2)$ TIME (here, n denotes vertex number).

The next step is to use all the information from 6 TaqMan reactions to allocate the each nodes of the Hamiltonian path. This can be done by applying the *in silico* algorithm as the following:

```
Input: N[0…|V|-1]=2 // N[0, ?, ?, ?, ?, 5]
A[1…|V|-2]=|V| // A[1, 1, 1, 1]
    for k=1 to |V|-3
        for l=k+1 to |V|-2
            if TaqMan($v_0$,$v_k$,$v_l$) = YES
                A[l] = A[l]+1
            else A[k] = A[k]+1
            endif
        endfor
        N[ A[k] ] = k
    endfor
N[ A[ |V|-2] ]=|V|-2
```

In this algorithm, an array (N[0…$|V|$-1) that stores all of the nodes of the Hamiltonian path is defined. In addition, an array of aggregation values (A[1..$|V|$-2]) used to locate the Hamiltonian path in each array of nodes is also defined. Based on the modified algorithm, the input array N is first initialized to N={0,?,?,?,?,5}, since the start and the end of the path are known, in advance. Next, the aggregation array A is initialized to A={1,1,1,1}. During the loop operations of the algorithm, the value of array A is increased in each iteration step. The aggregation array A[*i*] is used to indexing the nodes array for each value of *k*. After the loop operation |*V*|-2 is assigned to the N[A[|*V*|-2]. The output of the *in silico* algorithm can be viewed by calling back all the nodes array N[0] to N[|*V*|-1]. The outcome of the current instance of the *in silico* algorithm is N={0,2,4,1,3,5}. Note that this algorithm can be executed if all of the information regarding the TaqMan reactions has already been determined. This only can be done if clustering is applied to investigate the "YES" and "NO" reactions.

3. K-MEANS ALGORITHM

K-means clustering (MacQueen, 1967) is a very popular clustering technique which is used in numerous applications. This algorithm aims at minimizing an objective function,

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij} \| x_j - v_i \|^2 \qquad (1)$$

where is the number of clusters and is the total number of data points, and $\| x_j - v_i \|$ is the Euclidean distance between and . $U = [u_{ij}]$ is a binary partition matrix, in which elements . have values of 0 and 1 only, to indicate the degree of membership of each data point in the data set in cluster . Each data point in the data set is required to belong to exactly one cluster. Let $X = \{x_1, x_2, \cdots, x_N\}$ be a collection of data. By minimizing the objective function (1), is classified into homogeneous clusters, where the values in $V = \{v_1, v_2, \cdots\}$ are the cluster centers. The cluster center can be calculated as:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij} x_j}{\sum_{j=1}^{N} u_{ij}}, \quad i = 1, \cdots, C \qquad (2)$$

where $\sum_{j=1}^{N} u_{ij}$ is defined as the number of data points belonging to cluster . The partition matrix, *U* is updated using the following condition:

$$u_{ij} = 1 \text{ if } i = \arg\min_i \| x_j - v_i \|, \text{ else } u_{ij} = 0 \qquad (3)$$

In order to cluster TaqMan reaction, results into, "YES" and "NO" reactions, each reaction plot is represented as a vector, $x_j = \{x_{1,j}, x_{2,j}, \cdots, x_{n,j}\}$, where $x_{i,j}$ denotes the fluorescence intensity measured after amplification cycle *I* in TaqMan reaction *j*. The reactions are then clustered into two groups, with centers at $v_1 = \{v_{1,1}, v_{1,2}, \cdots, v_{1,n}\}$ and $v_2 = \{v_{2,1}, v_{2,2}, \cdots, v_{2,n}\}$. These two centers can be viewed as plot similar to the TaqMan reaction "YES" and "NO". Based on Figure 5, it may be noticed that the center located in the amplification region always has greater value than the other center, located in the non-amplification region. We call the two centers as the "YES" and "NO" centers, and use this information to classify the TaqMan reactions into "YES" and "NO" reactions, by comparing the partition matrix . Let us say that represents the "YES" center, and represent the "NO" center (note that does not always represent the "YES" center, when the clustering algorithm is run). We can say that $v_1 > v_2$. Consider example values, and , which are equal to 1 and 0, respectively. The "YES" and "NO" reactions can be determined by the following rule:

```
if(( $v_1 >$ $v_2$ and  >  ) or ( $v_2 >$ $v_1$
    and  >  ))
        ="YES"
else    ="NO"
```

Based on the proposed rule, we can classify as a "NO" reaction since > and $v_1 < v_2$. Applying this rule, we can classify the "YES" and "NO" reactions for each set of TaqMan reactions. The whole classification process can be described by the following steps:

Step 1: Initialize the membership matrix with random values (0 and 1, only)
Step 2: Calculate cluster center using equation (2)
Step 3: Update the partition matrix using condition (3)
Step 4: Calculate cost function using equation (1)
Step 5: If $|J(t-1) - J(t)| < \varepsilon$ then stop; otherwise, go to step 2.
Step 6: Determine "YES" and "NO" centers (either $v_1 >$ v_2 or $v_2 >$ v_1)
Step 7: Classify each TaqMan reaction, using the rule stated above

4. RESULT AND DISCUSSION

Figure 5 shows the two centers for "YES" and "NO" reactions, which will be used to classify the TaqMan reactions. Table 1 shows the value of partition matrix that represents the cluster. Based on the results from Figure 5 and Table 1, the work successfully clustered the two different TaqMan reactions. The result proves that the K-means clustering algorithm can be implemented to automatically classify the TaqMan reactions.

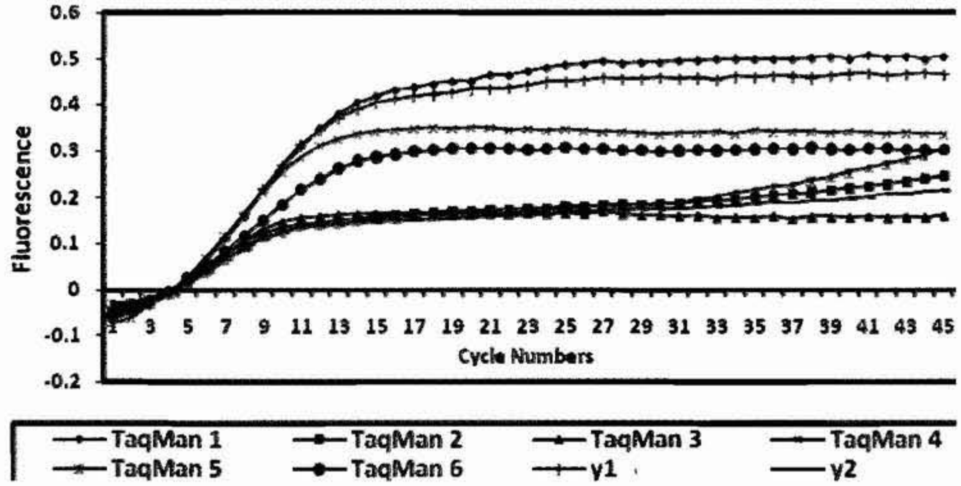

Fig. 5 Output of real-time PCR with "YES" and "NO" centers. In this case y_1 is the "YES" center and y_2 is the "NO" center, which show that $y_{1(45)}>y_{2(45)}$.

Table 1. Partition value for each TaqMan reaction

TaqMan	[illegible]	[illegible]	Reaction ($y_{1(45)}> y_{2(45)}$)
1 TaqMan(v_0,v_1,v_2)	0	1	"NO"
2 TaqMan(v_0,v_1,v_3)	1	0	"YES"
3 TaqMan(v_0,v_1,v_4)	0	1	"NO'
4 TaqMan(v_0,v_2,v_3)	1	0	"YES"
5 TaqMan(v_0,v_2,v_4)	1	0	"YES"
6 TaqMan(v_0,v_3,v_4)	0	1	"NO"

CONCLUSION

In the DNA computing readout approach based on the real-time PCR, the output of real-time PCR must be correctly clustered for automatically implementation of *in silico* information processing algorithm. By applying the K-means algorithm on the output of real-time PCR, two different TaqMan reactions, "YES" and "NO", can be clearly distinguished.

ACKNOWLEDGEMENT

This research is supported financially by the Ministry of Science, Technology, and Innovation (MOSTI), Malaysia under eScienceFund Research Funding (Vot 79033) and by the Ministry of Higher Education (MOHE) under Fundamental Research Grant Scheme (Vot 78225). Muhammad Faiz Mohamed Saaid is indebted to Universiti Teknologi Malaysia for granting him a financial support and opportunity to do this research.

REFERENCE

Adleman, L.M., Molecular Computation of Solutions to Combinatorial Problems, *Science*, vol. 266, pp. 1021-1024, 1994.

Heid, C.A., Real-time quantitative PCR, *Genome Research*, vol. 6, pp. 986-994, 1996.

Holland, P.M., Detection of specific polymerase chain reaction product by utilizing the 5'→3' exonuclease activity of termus aquaticus DNA polymerase, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, pp. 7276-7280, 1991.

Ibrahim, Z., Rose, J.A., Tsuboi, Y., Ono, O., and Khalid, M., A New Readout Approach in DNA Computing Based on Real-Time PCR with TaqMan Probes, *Lecture Notes in Computer Science (LNCS), Springer-Verlag, C. Mao and T. Yokomori (Eds.)*, vol. 4287, pp. 350-359, 2006.

Lakowicz, J.R. *Principles of fluorescence spectroscopy*, 2nd Ed., Kluwer Academic/Plenum Publishers, New York, 1999.

MacQueen, J. B., Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1, pp. 281-297, 1967.

Mullis, K., Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction, *Cold Spring Harbor Symposium on Quantitative Biology*, vol. 51, pp. 263-273, 1986.

Overbergh, L., The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression, *Journal of Biomolecular Techniques*, vol. 14, pp. 557-559, 2003.

Rose, J.A., The Effect of Uniform Melting Temperatures on the Efficiency of DNA Computing, *DIMACS Workshop on DNA Based Computers III*, pp. 35-42, 1997.

Walker, N.J., A technique whose time has come, *Science,* vol. 296, pp. 557-559, 2002.

Wood, D.H., A DNA Computing Algorithm for Directed Hamiltonian Paths, *Proceedings of the Third Annual Conference on Genetic Programming,* pp. 731-734, 1998.

Wood, D.H., Universal Biochip Readout of Directed Hamiltonian Path Problems, *Lecture Notes in Computer Science*, vol. 2568, pp. 168-181, 1999.