SICE Annual Conference 2008
August 20-22, 2008, The University Electro-Communications, Japan

Fuzzy C-Means Clustering for DNA Computing Readout Method Implemented on LightCycler System

Muhammad Faiz Mohamed Saaid[1], Zuwairie Ibrahim[1], Marzuki Khalid[1],
Nor Haniza Sarmin[2], and John A. Rose[3]

[1]Center for Artificial Intelligence and Robotics, Faculty of Electrical Engineering
Universiti Teknologi Malaysia, Malaysia
(Tel: +607-5535304; E-mail: ninjagaban@yahoo.com, zuwairiee@fke.utm.my, marzuki@utmkl.utm.my)
[2]Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, Malaysia
(Tel: +607-5534266; E-mail: nhs@mel.fs.utm.my)
[3]Institute of Information Communication Technology, Ritsumeikan Asia Pacific University,
1-1 Jumonjibaru, Beppu-shi, 874-8577 Oita, Japan
(E-mail: jarose@apu.ac.jp)

Abstract: In the previous work, 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 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, the Fuzzy C-Means (FCM) clustering algorithm is used to automatically identify two different reactions in real-time PCR. We show that the FCM clustering technique can be implemented for clustering the output results of the DNA computing readout method based on the LightCycler System.

Keywords: DNA computing, Fuzzy C-Means, real-time PCR, TaqMan probes.

1. INTRODUCTION

Since the discovery of the polymerase chain reaction (PCR) [1], 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 [2]. 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 oligonucleotide. The 5' and 3' ends of the oligonucleotide are terminated with attached reporter, such as FAM, and quencher fluorophores dyes, such as TAMRA, respectively, as shown in Figure 1 [3]. 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 a significantly longer wavelength (580 nm), the resulting fluorescence is not observable in Channel 1 of real-time PCR instruments [4].

The combination of dual-labeled TaqMan DNA probes with forward and reverse primers is a must for 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 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 [5], where R and Q denote the reporter dye and quencher dye, respectively. This separation subsequently allows the reporter to emit its fluorescence [6]. 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 with the total amount of PCR product, which can then be detected, using a real-time PCR instrument for visualization.

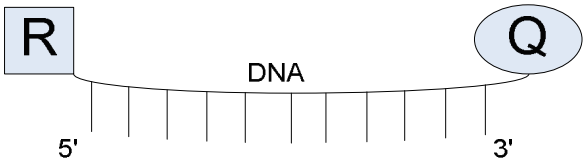

Fig. 1. Illustration of the structure of a TaqMan DNA probe. Here, R and Q denote the reporter and quencher fluorophores, respectively.

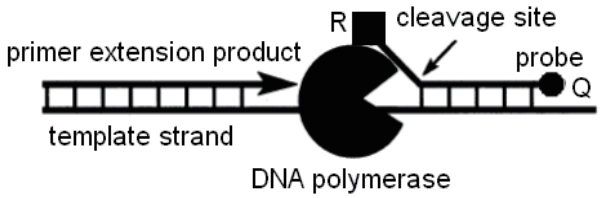

Fig. 2. Degradation of a TaqMan probe, via cleavage by DNA polymerase.

Previously, we proposed a readout method tailored specifically to the HPP in DNA computing, which employs a hybrid *in vitro-in silico* approach [7]. 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 an 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 [8], was employed to perform such operations. While a DNA chip-based methodology, which makes use of biochip hybridization for the same purpose has also been proposed [9-11], this method is more costly, and has yet to be experimentally implemented.

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

2. READOUT APPROACH OF DNA COMPUTATION BASED ON REAL-TIME POLYMERASE CHAIN REACTION

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 the 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•V_2•V_4•V_1•V_3•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 $[(|V|\text{-}2)^2\text{-}(|V|\text{-}2)]/2$ real-time PCR reactions, each denoted by TaqMan(v_0,v_k,v_l) for all k and l, such that $0 < k < |V|\text{-}2$, $1 < l < |V|\text{-}1$, and $k < l$. For this example instance, so that the DNA template is the 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

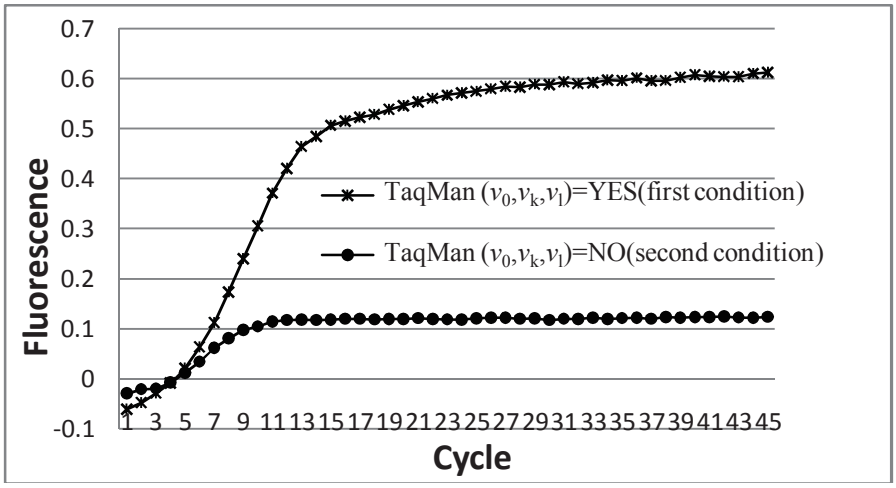

Fig. 3. An example of reaction plots corresponding to TaqMan(v_0,v_k,v_l) = YES (first condition) and TaqMan(v_0,v_k,v_l) = NO (second condition).

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 $O(|V|^2)$. 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 95ºC, 48ºC, and 72ºC, respectively. The resulting real-time PCR amplification plots are illustrated in Figure 4.

2.3 The *in silico* information processing

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 the six TaqMan reactions to allocate each node of the Hamiltonian path. This can be done by applying the *in silico* algorithm, as follows:

```
Input: A[0…|V|-1]=2                    // A[2, 2, 2, 2, 2, 2]
       A[0]=1, A[|V|-1]=|V|            // A[1, 2, 2, 2, 2, 6]
       for k=1 to |V|-3
          for l=2 to |V|-2
             while l>k
                if TaqMan(v0,vk,vl) = YES
                   A[l] = A[l]+1
                   else A[k] = A[k]+1
                endif
             endwhile
          endfor
       endfor
```

It is assumed that a Hamiltonian path is stored *in silico*, in an array (e.g., A[0…|V|-1]), for storage, information retrieval, and processing, such that A[i]∈A returns the exact location of a node, V_i∈V, in the Hamiltonian path. Based on the pseudo-code, and the example instance, the input array A is first initialized to A = {1, 2, 2, 2, 2, 6}. During the loop operations of the pseudo-code, the elements A[0]∈A and A[|V|-1]∈A, are not involved, as these two elements may conveniently be initialized to the correct values, as the distinguished starting and ending nodes of the Hamiltonian path are known in advance. The loop operations are thus strictly necessary only for the remaining elements, A[1,2,3…, |V|-2]∈A. Again, for the example instance, the output of the *in silico* information processing is A = {1, 4, 2, 5, 3, 6}, which represents the Hamiltonian path, $V_0 \rightarrow V_2 \rightarrow V_4 \rightarrow V_1 \rightarrow V_3 \rightarrow V_5$. For instance, in this case, it is indicated that V_3 is the fifth node in the Hamiltonian path, since A[3] = 5, and so on.

3. FUZZY C-MEANS ALGORITHM

FCM is a data clustering technique based on optimization of the objective function [12]:

$$J(U,V) = \sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^m \left\|x_j - v_i\right\|^2 \qquad (1)$$

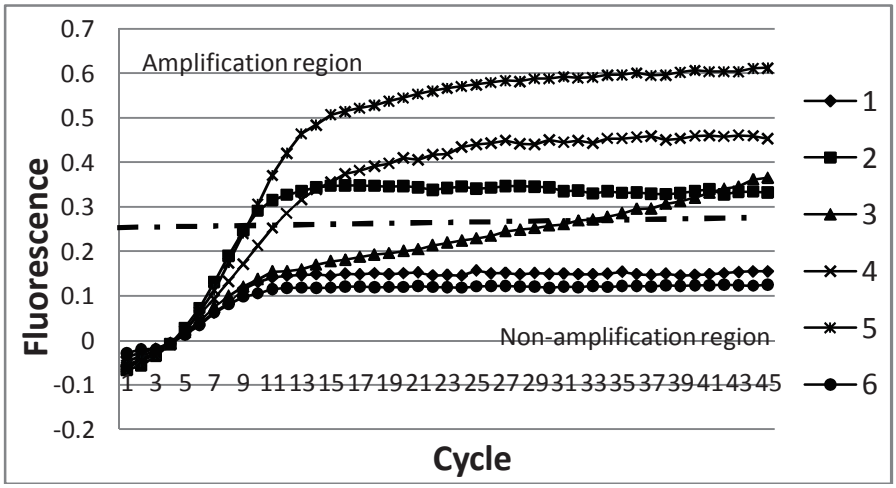

Fig. 4. Output of real-time PCR and grouping of output signals into two regions: amplification region (YES), and non-amplification region (NO).

Every data point in the data set is required to belong to a cluster at a particular membership degree. The purpose of FCM is to group the data points into different specific clusters. Let $X = \{x_1, x_2, \cdots, x_N\}$ be a collection of data. By minimizing the objective function (1), X is classified into C homogeneous clusters, where μ_{ij} is the membership degree of data x_j to fuzzy cluster setv_i. $V = \{v_1, v_2, \cdots v_N\}$are the cluster centers. $U = (\mu_{ij})_{N*C}$ is a fuzzy partition matrix, in which μ_{ij} indicates the membership degree of each data point in the data set to cluster j. The value of U should satisfy the following conditions:

$$\mu_{ij} \in [0,1] \quad \forall i = 1, \cdots, C, \ \forall j = 1, \cdots, N \tag{2}$$

$$\sum_{i=1}^{C} \mu_{ij} = 1, \ \forall j = 1, \cdots, N \tag{3}$$

In addition, $\|x_j - v_i\|$ is the Euclidean distance between x_j and v_i. The parameter m is called the fuzziness value index, which controls the fuzziness value of membership of each datum. The cluster center can be calculated by using the following equation:

$$v_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m x_j}{\sum_{j=1}^{N} (\mu_{ij})^m}, \quad \forall i = 1, \cdots, C \tag{4}$$

Clustering can then be achieved by iteratively minimizing the aggregate distance between each data point in the data set and cluster centers, until no further minimization is possible. The fuzzy partition matrix $U$is updated using the following equation:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}} \tag{5}$$

In order to cluster the results of the TaqMan reaction, namely ‘‘YES’’ and ‘‘NO’’ reactions, each graph of the reactions are represented as vector $\boldsymbol{x_j} = \{x_{j(1)}, x_{j(2)}, \cdots, x_{j(45)}\}$. The reactions are clustered into two groups, having their centers at $\boldsymbol{v_1} = \{v_{1(1)}, v_{1(2)}, \cdots, v_{1(45)}\}$ and $\boldsymbol{v_2} = \{v_{2(1)}, v_{2(2)}, \cdots, v_{2(45)}\}$. These two centers can be viewed as graphs that are similar to the TaqMan reaction ‘‘YES’’ and ‘‘NO’’. Based on Figure 5, it is noticed that the center that is located in the amplification region always has greater value than the other center in the non-amplification region. We call the two centers as the ‘‘YES’’ and ‘‘NO’’ centers, where the ‘‘YES’’ center is greater than ‘‘NO’’ center. We use this information to classify the TaqMan ‘‘YES’’ and ‘‘NO’’ reactions by comparing the fuzzy partition matrix U. Let us say that $\boldsymbol{v_2}$ represents the ‘‘YES’’ center and $\boldsymbol{v_1}$ represent the ‘‘NO’’ center (Note that $\boldsymbol{v_2}$ does not always represent the ‘‘YES’’ center when the FCM algorithm is run). Then, we can say that $\boldsymbol{v_2} > \boldsymbol{v_1}$. Let’s take an example U_{11} and U_{21}, which equals to 0.6 and 0.4. The ‘‘YES’’ and ‘‘NO’’ reaction can be determined by following this rule:

if($\boldsymbol{v_1} > \boldsymbol{v_2}$ and $U_{1j} > U_{2j}$) or ($\boldsymbol{v_2} > \boldsymbol{v_1}$ and $U_{2j} > U_{1j}$)
$\boldsymbol{x_j}$ =’’YES’’
else $\boldsymbol{x_j}$ =’’NO’’

Based on the proposed rule, we can classify x_j as ‘‘NO’’ reaction since $U_{1j} > U_{2j}$ and $v_1 < v_2$. Applying this rule, we can classify each of the TaqMan reactions as ‘‘YES’’ and ‘‘NO’’ reactions. The entire classification process can be described in the following steps:

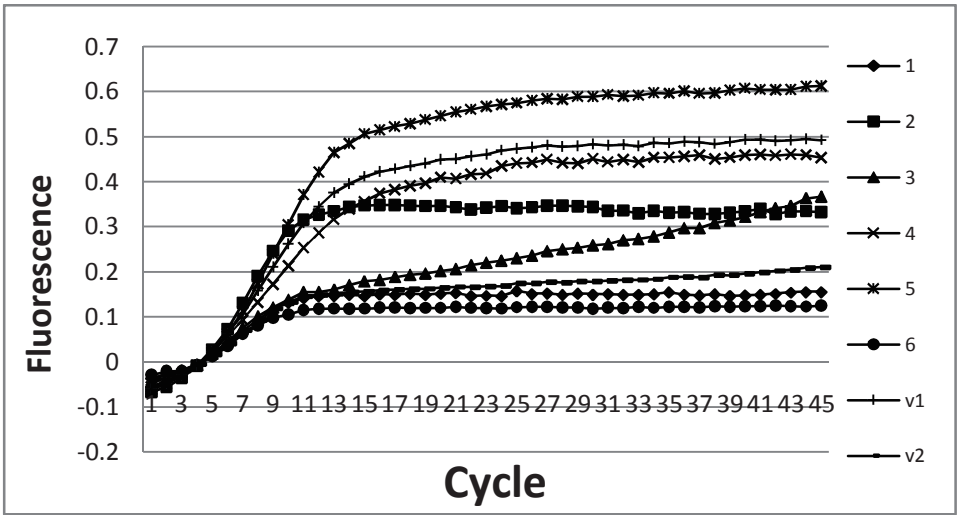

Fig. 5. Output of real-time PCR, with the "YES" and "NO" centers. In this case, v1 is the "YES" center and v2 is the "NO" center, which show that v1>v2.

Table 1: Fuzzy Partition Value for each TaqMan reactions.

TaqMan	U_{i1}	U_{i2}	Reaction ($\mathbf{v}_1 > \mathbf{v}_2$)
TaqMan(v_0,v_1,v_2)	0.009751	0.99025	"NO"
TaqMan(v_0,v_1,v_3)	0.63009	0.36991	"YES"
TaqMan(v_0,v_1,v_4)	0.14071	0.85929	"NO'
TaqMan(v_0,v_2,v_3)	0.97123	0.028771	"YES"
TaqMan(v_0,v_2,v_4)	0.9331	0.0669	"YES"
TaqMan(v_0,v_3,v_4)	0.027844	0.97216	"NO"

Step 1: Initialize the membership matrix, U with random values, subject to conditions (2) and (3).
Step 2: Calculate the cluster center V using equation (4).
Step 3: Update Fuzzy partition matrix using equation (5).
Step 4: If $\|U(t+1) - U(t)\| < \varepsilon$ then stop, otherwise go to step 2.
Step 5: Determine the ''YES'' and ''NO'' centers (either $\boldsymbol{v}_1 > \boldsymbol{v}_2$or $\boldsymbol{v}_2 > \boldsymbol{v}_1$).
Step 6: Classify each of the TaqMan reactions using the rule, stated above.

4. RESULTS 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 fuzzy partition matrix that represents the membership degree for each cluster. Based on the results from Figure 5 and Table 1, we successfully clustered the two different TaqMan reactions. The result proved that the FCM clustering algorithm can be implemented to automatically classify the TaqMan reactions.

5. 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 FCM 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 & 79034). Muhammad Faiz Mohamed Saaid is indebted to MOSTI and Universiti Teknologi Malaysia for granting him a financial support and opportunity to do this research

REFERENCES

[1] K. Mullis, "Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction", *Cold Spring Harbor Symposium on Quantitative Biology 51,* 1986, pp 263-273.
[2] L. Overbergh, "The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression", *Journal of Biomolecular Techniques 14,*2003, pp. 557-559.
[3] N.J. Walker, "A technique whose time has come." *Science 296*, 2002, pp 557-559.

[4] J.R. Lakowicz, *Principles of fluorescence spectroscopy*, 2nd Ed., Kluwer Academic/Plenum Publishers, New York, 1999.
[5] C.A. Heid, "Real-time quantitative PCR." *Genome Research 6*, 1996, pp 986-994.
[6] P.M. Holland, "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 88*, 1991, pp 7276-7280.
[7] Zuwairie Ibrahim, John A. Rose, Yusei Tsuboi, Osamu Ono, and Marzuki Khalid, "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*, 2006, pp. 350-359.
[8] L.M. Adleman, "Molecular Computation of Solutions to Combinatorial Problems", Science, Vol. 266 (1994) 1021-1024.
[9] J.A. Rose, "The Effect of Uniform Melting Temperatures on the Efficiency of DNA Computing", DIMACS Workshop on DNA Based Computers III (1997) 35-42.
[10] D.H. Wood, "A DNA Computing Algorithm for Directed Hamiltonian Paths", *Proceedings of the Third Annual Conference on Genetic Programming* (1998) 731-734.
[11] D.H. Wood, "Universal Biochip Readout of Directed Hamiltonian Path Problems, Lecture Notes in Computer Science, Vol. 2568 (1999) 168-181.
[12] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms,* Plenum Press, New York. (1981).