BioSystems 100 (2010) 1–7

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

In vitro molecular pattern classification via DNA-based weighted-sum operation

Hee-Woong Lim[a,1], Seung Hwan Lee[c,1], Kyung-Ae Yang[b], Ji Youn Lee[d], Suk-In Yoo[b], Tai Hyun Park[c], Byoung-Tak Zhang[b,*]

[a] Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-750, Republic of Korea
[b] School of Computer Science and Engineering and the Center for Biointelligence Technology (CBIT), Seoul National University, Seoul 151-744, Republic of Korea
[c] School of Chemical and Biological Engineering, Seoul National University, Seoul 151-744, Republic of Korea
[d] Department of Biomedical Engineering, University of California, Davis, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 4 June 2009
Received in revised form
30 November 2009
Accepted 2 December 2009

Keywords:
Molecular pattern classification
DNA computing
DNA-based weighted-sum operation
Competitive hybridization

ABSTRACT

Recent progress in molecular computation suggests the possibility of pattern classification *in vitro*. Weighted sum is a primitive operation required by many pattern classification problems. Here we present a DNA-based molecular computation method for implementing the weighted-sum operation and its use for molecular pattern classification in a test tube. The weights of the classifier are encoded as the mixing ratios of the differentially labeled probe DNA molecules, which are competitively hybridized with the input-encoding target molecules to compute the decision boundary of classification. The computation result is detected by fluorescence signals. We experimentally verify the underlying weight encoding scheme and demonstrate successful discrimination of two-group labels of synthetic DNA mixture patterns. The method can be used for direct computation on biomolecular data in a liquid state.
© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The identification of biomarkers and their analysis for disease diagnosis have recently emerged as important issues. Many of those problems involve pattern classification (Bishop, 2006), where the target patterns are represented as mixtures of biomarkers such as messenger RNAs (mRNA), proteins, or microRNAs (miRNA), and their expression levels provide informative features for the classification (Khan et al., 2001; Ramaswamy et al., 2001; Lu et al., 2005). Whereas the conventional analysis methods for these biomolecular patterns require quantitative detection *in vitro* prior to the analysis *in silico*, recent progress in molecular information processing technology suggests that direct computation on such biochemical information *in vitro* is realizable (Paun et al., 1998); examples include gene expression analysis and control, extraction of molecular features, and algebraic operations *in vitro* (Oliver, 1997; Mills et al., 1999; Sakakibara and Suyama, 2000; Mills, 2002; Benenson et al., 2004; Lim et al., 2004).

Many DNA computing approaches have been developed previously for solving computational problems or for implementing Boolean logic circuits *in vitro* (Henkel et al., 2007; Bakar et al., 2008; Wang et al., 2008; Zoraida et al., 2009). However, more advanced analysis of biochemical information *in vitro* requires efficient arithmetic operations capable of handling molecular quantitative information. Although a few theoretically sound models of DNA-based algebraic operations have been proposed (Oliver, 1997; Mills et al., 1999, 2001), they have not yet been implemented experimentally. Here, instead, we focus on a simple but significant primitive molecular algebraic operation and propose a molecular pattern classification model that is implemented in DNA computing. Operating *in vitro*, it directly takes biological inputs such as DNA or RNA molecules and computes the group indices that are displayed optically as fluorescence signals. As a primitive operation, we define a DNA-based weighted-sum computation and formulate simple competitive hybridization reactions between the input molecules and differentially labeled probe mixtures into the weighted sum of inputs. We present the verification results of our weight encoding scheme and demonstrate experimentally the successful classification of *in vitro* patterns of synthetic DNA mixtures. This paper is an extended written version of the oral presentation given at the 13th international meeting on DNA computing (Lim et al., 2007).

The paper is organized as follows. In Section 2, we describe the basic architecture and theoretical background of the molecular pattern classifier. Section 3 gives a DNA-based implementation of the molecular pattern classifier *in vitro* including the DNA-based weighted-sum operation. The experimental results are presented in Section 4. Section 5 summarizes the result. More details on mate-

* Corresponding author.
E-mail addresses: hwlim@ewha.ac.kr (H.-W. Lim), skulsh78@snu.ac.kr (S.H. Lee), kayang@bi.snu.ac.kr (K.-A. Yang), jiyounlee@gmail.com (J.Y. Lee), siyoo@ailab.snu.ac.kr (S.-I. Yoo), thpark@snu.ac.kr (T.H. Park), btzhang@bi.snu.ac.kr (B.-T. Zhang).
[1] These authors contributed equally to this work.

0303-2647/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2009.12.001

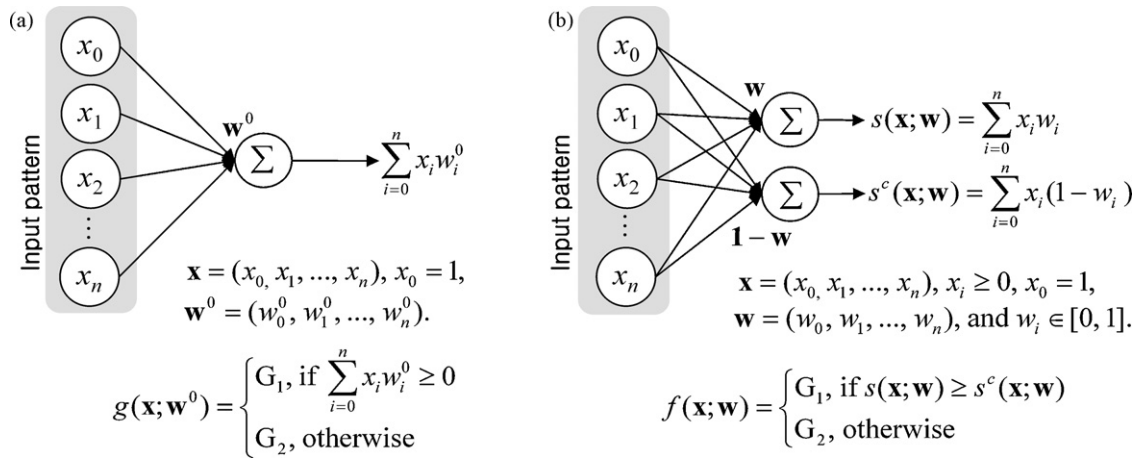

Fig. 1. Comparison of pattern classification models. (a) Conventional linear classification model using both positive and negative weights. It takes an input pattern, **x**, computes the weighted sum of the input elements, and compares the result with a threshold value to determine the class. It should be noted that the weights, w_i^0, may be either positive or negative. (b) Molecular pattern classification model, proposed in this paper, using only positive weights for its DNA-based realization *in vitro*. It computes two complementary weighted sums, $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$, as classification scores for the classes using only positive weights unlike the conventional model. See text for more explanations.

rials and methods for DNA computing experiments are provided in Appendix B.

2. A Molecular Computation Model of Pattern Classification

A conventional linear discriminant function takes real-valued vectors as input patterns and computes the weighted sum of the input elements to produce a classification score, which is then compared with a threshold for decision-making (Fig. 1a). Despite its simplicity, this is the common architecture of many pattern classification models and the combination with appropriate preprocessing techniques enables it to classify complex patterns (Bishop, 2006). Therefore, implementing this linear model using DNA molecules could provide the basis for more sophisticated molecular computing applications.

Nucleic acids are attractive materials as information carriers and the analogy between numerical information and molecular quantities makes real number representation plausible. Especially, the primary targets of our molecular pattern classifier are the mixtures of biomarkers such as DNA or RNA molecules. Therefore, it is reasonable to assume that our target patterns are nonnegative real-valued vectors, $\mathbf{x} = (x_0, x_1, \ldots, x_n)$, where n is the number of biomarkers under consideration, x_i is the quantity of each marker, and $x_0 = 1$ for a bias term, i.e. a constant term in a linear classification model. Each biomarker shows a different expression level depending on the group, and these differences are reflected in the weight factors, w_i^0, which can be either positive or negative. The inputs are weighted and summed into a classification score, which is used for decision-making as follows:

$$g(\mathbf{x};\mathbf{w}^0) = \begin{cases} G_1 & \text{if } \sum_{i=0}^{n} x_i w_i^0 \geq 0, \\ G_2 & \text{otherwise,} \end{cases} \tag{1}$$

where $x_i \geq 0$, $\mathbf{w}^0 = (w_0^0, w_1^0, \ldots, w_n^0)$, w_0^0 is a bias term, and G_1 and G_2 are labels for the two groups, Group 1 and Group 2, respectively.

However, the representation of the negative weights by DNA molecules is not feasible because molecular quantities are always nonnegative. Therefore, we transform the conventional linear classification model into an alternative one, $f(\mathbf{x};\mathbf{w})$, that is equivalent to $g(\mathbf{x};\mathbf{w}^0)$. To do this, we define two complementary weighted-sum functions:

$$s(\mathbf{x};\mathbf{w}) = \sum_{i=0}^{n} x_i w_i \quad \text{and} \quad s^c(\mathbf{x};\mathbf{w}) = \sum_{i=0}^{n} x_i (1 - w_i) \tag{2}$$

for computing the scores for the classes G_1 and G_2, respectively. Then, the molecular classifier uses these two values to decide the class (Fig. 1b):

$$f(\mathbf{x};\mathbf{w}) = \begin{cases} G_1 & \text{if } s(\mathbf{x};\mathbf{w}) \geq s^c(\mathbf{x};\mathbf{w}), \\ G_2 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_n)$, w_0 is a bias term, and $w_i \in [0, 1]$. The two complementary scores, $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$, are computed *in vitro* via a DNA-based weighted-sum operation simultaneously and the concrete molecular realization of the operation will be described in the next section.

This transformed classification model can be shown to be equivalent to the conventional one (Fig. 1a) by selecting appropriate weight values. To show this, note that

$$\begin{aligned} s(\mathbf{x};\mathbf{w}) \geq s^c(\mathbf{x};\mathbf{w}) &\Rightarrow s(\mathbf{x};\mathbf{w}) - s^c(\mathbf{x};\mathbf{w}) \geq 0 \\ &\Rightarrow \textstyle\sum_{i=0}^{n} x_i w_i - \sum_{i=0}^{n} x_i (1 - w_i) \geq 0 \\ &\Rightarrow \textstyle\sum_{i=0}^{n} x_i (2w_i - 1) \geq 0. \end{aligned} \tag{4}$$

Thus, we have an alternative form of (3) as

$$f(\mathbf{x};\mathbf{w}) = \begin{cases} G_1 & \text{if } \sum_{i=0}^{n} x_i (2w_i - 1) \geq 0, \\ G_2 & \text{otherwise.} \end{cases} \tag{5}$$

If we let

$$w_i = \frac{(1 + \lambda w_i^0)}{2}, \tag{6}$$

where w_i^0 is the weight for the conventional classifier and λ is a positive scaling constant to keep w_i in the range [0, 1], then (5) becomes:

$$f(\mathbf{x};\mathbf{w}) = \begin{cases} G_1 & \text{if } \sum_{i=0}^{n} \lambda x_i w_i^0 \geq 0, \\ G_2 & \text{otherwise.} \end{cases} \tag{7}$$

which is the same as $g(\mathbf{x};\mathbf{w}^0)$ in (1) for any positive constant λ. Consequently, for any given linear pattern classifier, we can determine a positive weight vector to build an equivalent molecular pattern classifier using a weight transformation rule (6).

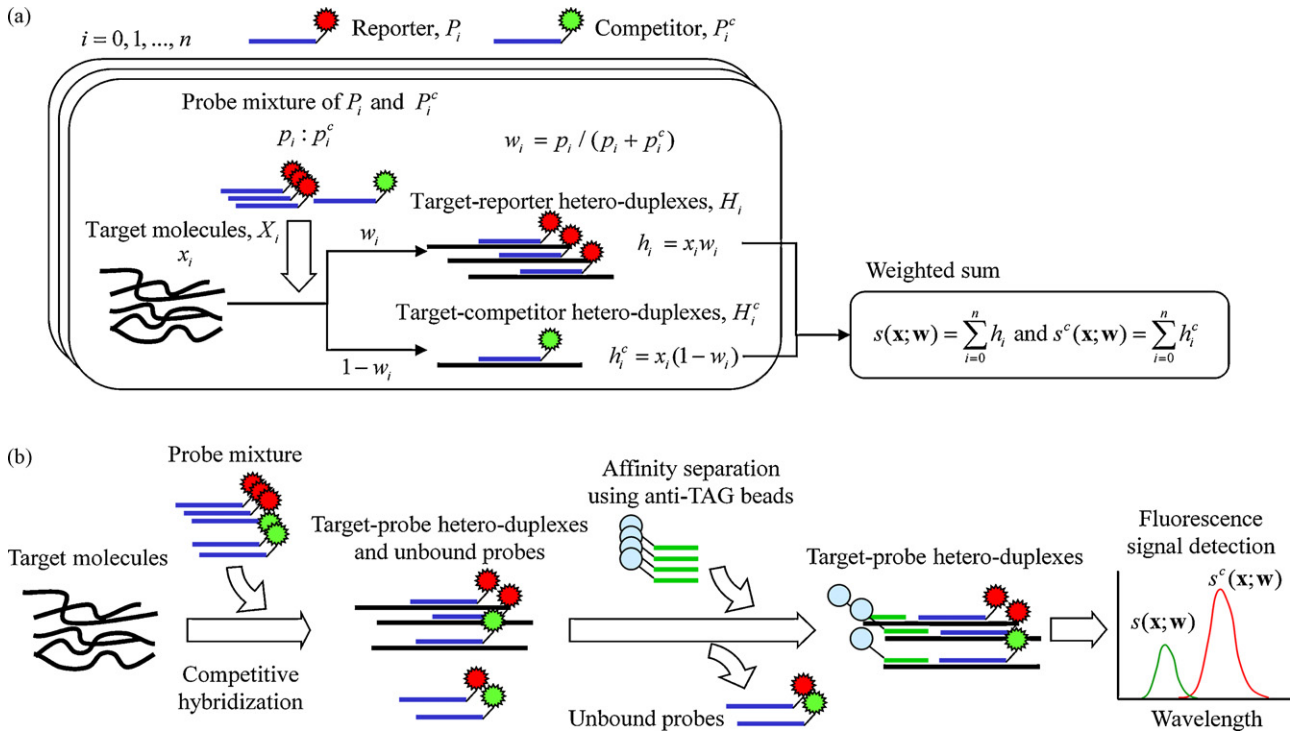

Fig. 2. DNA-based weighted-sum computation. (a) The weighted-sum operation is performed via competitive hybridization reaction between the input DNA molecules, X_i, and a differentially labeled probe mixture of P_i and P_i^c, which competitively bind to X_i to form target-probe hetero-duplexes, H_i or H_i^c. The resulting amount of H_i and H_i^c depends on the initial input levels, x_i, and the mixing ratios between P_i and P_i^c. Therefore, the total amount of H_i and H_i^c corresponds to $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$, respectively, where w_i is represented as the ratio of P_i in the probe mixture. $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$ are complementary to each other in the sense that their sum is always equal to the sum of the inputs. (b) Overall process for the computation and the result detection. The weighted-sum computation is performed via competitive hybridization reaction between inputs and a probe mixture. The computation results, $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$, are detected by fluorescence signal intensities after selecting target-probe hetero-duplexes via affinity separation.

3. Realization in DNA Computing

The massively parallel interaction among DNA molecules based on their complementarity has inspired a novel architecture for cognitive information processing realizable *in vitro* (Zhang, 2008). Likewise, appropriate mapping from reactants and products of a chemical reaction to inputs and outputs could be the key to a new molecular operation. Here, we develop a DNA-based method for weighted-sum computation by forming hybridization reactions between input DNA molecules and differentially labeled probe molecules (Fig. 2).

The weight encoding scheme and the DNA-based weighted-sum computation are motivated by a simple quantitative assay using fluorophore-labeled probe DNAs (Wetmur, 1991; Kessler, 1994). In the assay, the resulting fluorescence intensity is proportional to the amount of target molecules. It is also possible to perform this assay for multiple targets simultaneously in a test tube using a probe mixture. If the probes are labeled with the same fluorophore for multiple targets, the resulting fluorescence signal can be viewed as the sum of inputs, where each input is represented by the amount of each target. We will refer to these probes as reporter probes.

Now, the problem is how to assign an arbitrary weight to each input. Weight values are assigned to the inputs by using competitor probes, which have the same nucleotide sequence as the reporter but have a different fluorophore (Fig. 2a). During the hybridization, the reporter and competitor probes, P_i and P_i^c, compete to bind to a corresponding target, X_i, and to form target-probe hetero-duplexes, H_i and H_i^c, respectively. Assuming symmetric thermodynamics of the two competitive hybridization reactions due to the same probe sequence and complete target hybridization under the excess probe condition, it is reasonable to expect that the final amount of target-probe hetero-duplexes, H_i and H_i^c, depend on the mixing ratio of the reporter and competitor molecules (see Appendix A):

$$\begin{aligned} h_i &= x_i w_i, \\ h_i^c &= x_i(1 - w_i), \end{aligned} \tag{8}$$

where w_i denotes the portion of the reporter in the probe mixture for X_i, i.e.:

$$w_i = \frac{p_i}{p_i + p_i^c}. \tag{9}$$

p_i and p_i^c denote the initial amount of the reporter and competitor probes, respectively. This competitive hybridization reaction occurs for all the target species ($i = 0, \ldots, n$) in a test tube. Therefore, the total amount of H_i corresponds to the weighted sum of inputs, $s(\mathbf{x};\mathbf{w})$, and the mixing ratios of P_i and P_i^c encode the weights as follows:

$$s(\mathbf{x};\mathbf{w}) = \sum_{i=0}^{n} h_i, \tag{10}$$

where $\mathbf{x} = (x_0, x_1, \ldots, x_n)$, $\mathbf{w} = (w_0, w_1, \ldots, w_n)$, $x_i \geq 0$, and $w_i \in [0, 1]$ (Fig. 2a). Likewise, H_i^c forms a complementary score, $s^c(\mathbf{x};\mathbf{w})$, as follows:

$$s^c(x;w) = \sum_{i=0}^{n} h_i^c, \tag{11}$$

which is in a trade-off relation with (10) in the sense that the sum of the two scores is always equal to the sum of the inputs. It should be noted that inputs and weights are both nonnegative.

The computation results are detected by fluorescence signals in the emission wavelengths of the reporter and competitor probes. Typically, an excess amount of probes is used in the hybridization reaction; there remains a surplus of unbound probes, which must be eliminated before the signal detection. Only hetero-duplexes are

Table 1
Target and probe sequences used for *in vitro* pattern classification.

Symbol	Sequence (5′ → 3′)	Length (mer)	T_m (°C)
X_1	ACCTGATGACTCTAAGCCT AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	49	60.3
X_2	AAGTGTTACAAGTGTGAGGA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	50	60.2
X_3	AAGATGAAGACAATGTTCTCAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	52	59.5
X_4	TCTTACAGAAGCAGAGATTGA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	51	59.9
P_1/P_1^c	Cy5/Cy3-AGGCTTAGAGTCATCAGGT	19	52.3
P_2/P_2^c	Cy5/Cy3-TCCTCACACTTGTAACACTT	20	51.6
P_3/P_3^c	Cy5/Cy3-TTGAGAACATTGTCTTCATCTT	22	50.7
P_4/P_4^c	Cy5/Cy3-TCAATCTCTGCTTCTGTAAGA	21	51.3

Oligonucleotide sequences used in the experiments. X_i is the target strands, where the first part of X_i is the probe-binding region and the last 30 mer of adenine is reserved for the affinity bead separation. P_i is a reporter probe and P_i^c is a competitor probe; they are attached with Cy5 and Cy3, respectively. The T_m was calculated using the nearest neighbor hybridization model (SantaLucia and Hicks, 2004) under the conditions, Na^+ concentration = 0.05 M and oligonucleotide DNA concentration = 0.25 μM.

extracted from the hybridization product through affinity bead separation using paramagnetic beads (Lambert and Williamson, 1993) attached with nucleotide strands that are capable of binding to the single-stranded region of the hetero-duplexes. The binding regions can be easily selected as subsequences of the target strands except the probe-binding regions. For example, poly-adenine tails would be appropriate for the separation when the targets are mRNAs.

4. Experimental Results

As an illustrative example and a verification experiment, we demonstrate binary classification of *in vitro* patterns made of synthetic oligonucleotide DNA mixtures.

4.1. Model System and Target Patterns

As our primary goal is to develop an *in vitro* molecular pattern classifier that takes wet biological data directly, it is assumed that the parameter, i.e. the weight vector for the decision boundary is already known. Accordingly, we prepare two groups, Group 1 and Group 2, of *in vitro* patterns from a known decision boundary and show that they are correctly classified by the molecular pattern classifier built using the known weights. The model system is motivated by gene expression-based disease diagnosis. The difference is that the target patterns are made of synthetic oligonucleotide DNAs rather than gene transcripts.

In defining the decision boundary of the two groups in the model system, we set the bias $w_0^0 = 0$ for convenience, and consequently, can ignore the bias and the corresponding constant input element x_0. This is reasonable setting and does not restrict the generality of the experiments, because the existence of the bias term does not affect the general computation scheme. If the bias is not zero, for example, the only things we need to do are to add constant amount of extra input molecules as x_0 to each example and to add another pair of probe mixture for the bias according to (6). Therefore, from now on we will omit w_0, w_0^0, and x_0 in describing the model system and the experimental results. First, a target separating hyperplane with a normal vector $\mathbf{w}^0 = (0.9, 0.3, -0.6, -0.3)$ was determined in a four-dimensional vector space, arbitrarily. Five points were selected from each side of the hyperplane in close proximity as representative patterns of each group. Then, four oligonucleotide sequences were designed, synthesized, and mixed according to the selected points to prepare target *in vitro* patterns (see Tables 1 and 2). The nonnegative weight vector for the *in vitro* pattern classifier was calculated as $\mathbf{w} = (0.95, 0.65, 0.2, 0.35)$ from the above $\mathbf{w}^0$ using (6), and a weight encoding probe mixture was prepared (see Table 1).

4.2. Verification of Weight Encoding Scheme

The DNA-based weighted-sum operation assumes that the thermodynamics in the competitive hybridization reactions of differentially labeled probe pairs is symmetric. We test this assumption. A series of competitive hybridization reaction was performed using single target DNA and probe mixtures with various mixing ratios, and the resulting fluorescence signals were detected (see Appendix B). As shown in Fig. 3, high input levels led to high fluorescence intensities. The intensities increased linearly with the ratio of the corresponding probe in the mixture, which is consistent with our assumption.

4.3. In vitro Pattern Classification

Two groups of patterns were prepared and classified using the *in vitro* molecular pattern classification method experimentally. Fig. 4a shows the plot of the score for Group 1, $s(\mathbf{x}; \mathbf{w})$, versus the score for Group 2, $s^c(\mathbf{x}; \mathbf{w})$, for each pattern. Here the diagonal solid line denotes the decision boundary for the classification. The two groups of patterns do not overlap, which indicates successful classification. A large difference between $s(\mathbf{x}; \mathbf{w})$ and $s^c(\mathbf{x}; \mathbf{w})$ for a pattern means that the pattern is located away from the separating hyperplane in the input vector space, and vice versa. Although two patterns in Group 1 are adjacent to the decision boundary because the patterns were chosen close to the separating hyperplane as mentioned above, the classification results are still correct. It is expected that other patterns located further away from the target

Table 2
Composition of the target patterns used for *in vitro* classification.

Group 1					Group 2				
Number	x_1	x_2	x_3	x_4	Number	x_1	x_2	x_3	x_4
1	350	150	350	50	6	150	250	350	250
2	250	150	250	50	7	50	250	350	150
3	250	150	250	350	8	50	150	350	50
4	250	50	150	350	9	50	350	250	250
5	150	50	50	350	10	50	250	250	150

Each row denotes the composition of the *in vitro* pattern used in the classification task. Each column, x_i, denotes the amount (fmol) of the corresponding target strands, X_i, in the patterns.

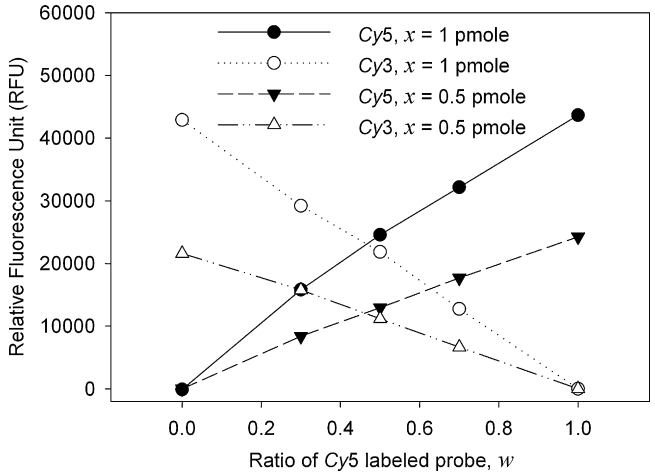

Fig. 3. Weight encoding verification. The resulting relative fluorescence unit (RFU) representing the multiplication results of single-input and weight pairs increased linearly with the input level, x, and the mixing ratio of the corresponding probe. The black solid symbols denote the signal from Cy5, i.e. xw, and the white symbols denote the signal from Cy3, i.e. $x(1-w)$.

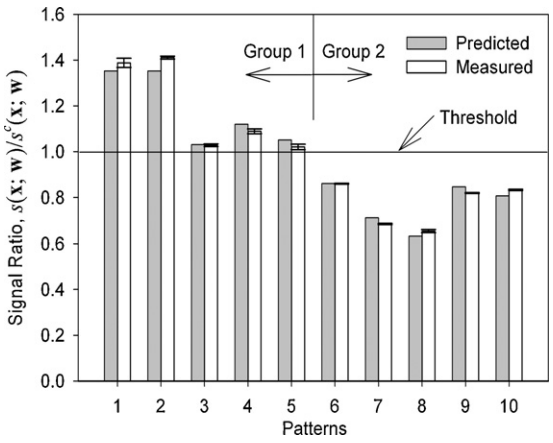

Fig. 5. Average score ratio, $s(\mathbf{x};\mathbf{w})/s^c(\mathbf{x};\mathbf{w})$, for each pattern $\mathbf{x}$. The final decision-making is done by comparing two fluorescence intensities. Hence, if the ratio is larger than 1, i.e. $s(\mathbf{x};\mathbf{w}) > s^c(\mathbf{x};\mathbf{w})$, the pattern is classified as Group 1, else as Group 2. Patterns 1–5 and 6–10 belong to Group 1 and Group 2, respectively.

hyperplane but not used in these experiments would be classified correctly as well.

To confirm the classification results, we compared the detected scores with the predicted scores. Fig. 4b shows the correlation with the predicted values and a fitting line. Pearson correlation coefficient was 0.96, and this high correlation supports the soundness of the classification results as well as the reliability of the weighted-sum operation in terms of algebraic computation.

We repeated the classification experiment to confirm the reproduction of the experimental results. As shown in Fig. 5, the average ratio, $s(\mathbf{x};\mathbf{w})/s^c(\mathbf{x};\mathbf{w})$, for each pattern was very close to that of the predicted one. It is the same for the patterns with relatively large errors of outputs. This means that the errors during the experimental process affect both $s(\mathbf{x};\mathbf{w})$ and $s^c(\mathbf{x};\mathbf{w})$ by scaling them with a similar magnitude, which supports the error tolerance of the proposed method as an *in vitro* molecular pattern classifier. As for the computation or classification scale, our experiments were fmole scale, and further research would be needed to investigate and improve the performance in smaller scale. However, if enough amounts of samples are available, performing the classification task multiple times could reduce the stochastic errors. Under clinical situation, this can be viewed as performing multiple diagnoses for a disease.

Although it was already known that the patterns are linearly separable, it is important to note that the patterns are wet DNA molecules and that the computation to obtain the classification scores is performed *in vitro*. Given a set of target patterns and their separating hyperplane, it is possible to classify *in vitro* patterns using the prepared weight encoding probe mixture without explicitly knowing the pattern compositions. This concept can be generalized to the problem of gene expression analysis or molecular diagnosis (Golub et al., 1999). A mixture of marker gene transcripts can be viewed as an *in vitro* pattern, and weight values mean their significance for the disease diagnosis. Once the marker genes and weight values are known via typical methods such as microarray technology (Duggan et al., 1999; Lipshutz et al., 1999), then we can prepare a probe mixture in advance and diagnose the disease. In this framework, the diagnosis can be simplified since probe molecules are already labeled from manufacturing, and the labeling step, which was needed in microarray-based methods, is not needed for mRNA substrate extracted from a biopsy.

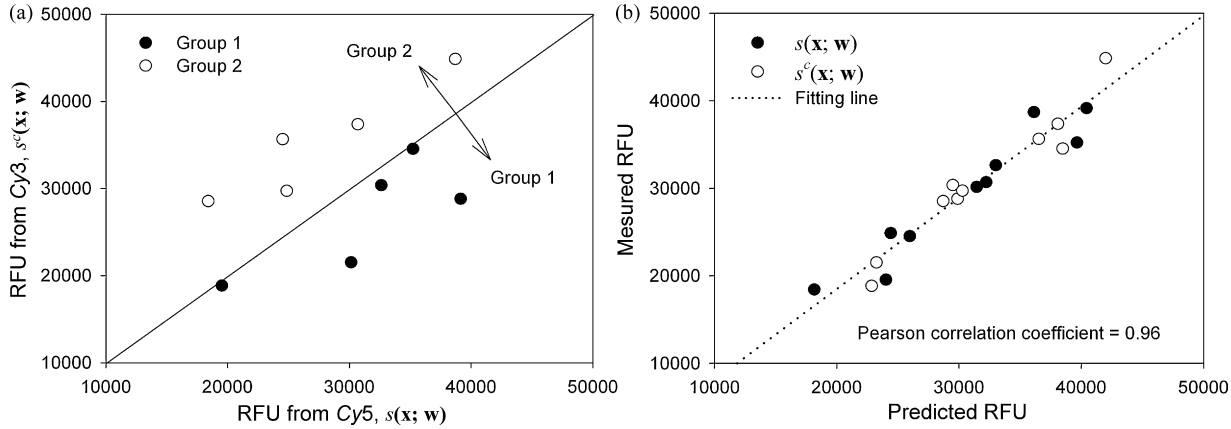

Fig. 4. Molecular pattern classification *in vitro*. (a) Plot of the score for Group 1, $s(\mathbf{x};\mathbf{w})$, vs. the score for Group 2, $s^c(\mathbf{x};\mathbf{w})$, for the patterns showing the classification results. The diagonal solid line denotes the decision boundary between the two groups. The two groups of patterns are separated without overlap, which means successful classification. (b) Correlation between the measured scores and the predicted values. The predicted relative fluorescence unit (RFU) was obtained by numerically calculating the weighted sum of inputs for each pattern. The high correlation supports the accuracy of the DNA-based weighted-sum operation and the soundness of the classification results.

5. Conclusion

We proposed a DNA-based molecular pattern classification method. It takes biological inputs such as DNA or RNA mixtures directly and the computation is performed *in vitro*. DNA-based weighted sum was defined as a primitive operation, and its molecular algorithm was developed based on the competitive hybridization reaction between input molecules and a differentially labeled probe mixture. We verified and confirmed our weight encoding scheme experimentally. This demonstrates a successful classification of molecular patterns made of synthetic oligonucleotide DNA mixture and further confirms the accuracy of the DNA-based weighted-sum operation and the classifier. It should be noted that the molecular algorithm requires no enzymatic reaction and takes only three experimental steps: hybridization, separation, and detection. The proposed method can be used in combination with other molecular learning algorithms that automatically generate the weight encoding probe mixture from training samples via DNA computing *in vitro* (Chen et al., 2005; Zhang and Jang, 2005; Zhang, 2008).

Acknowledgements

This research was supported in part by the Ministry of Commerce, Industry, and Energy through the Molecular Evolutionary Computing (MEC) project, the Ministry of Education, Science, and Technology under the BK21-IT Program and Pioneer Project, and the Ministry of Science and Technology through the National Research Laboratory (NRL) Program. Yang was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-532-C00031). The ICT at Seoul National University provided research facilities for this study.

Appendix A. Mass Action Equation of the Competitive Hybridization Reaction

In the hybridization reactions among input molecules and differentially labeled probe pairs, the final amounts of hetero-duplexes are determined by the initial input concentration and the mixing ratio of the probe pairs as (8). This equation can be easily derived from the competitive hybridization reactions:

$$\begin{aligned} X_i + P_i &\overset{K}{\longleftrightarrow} H_i, \\ X_i + P_i^c &\overset{K}{\longleftrightarrow} H_i^c, \end{aligned} \tag{12}$$

and their binding constant K given by the law of mass action:

$$K = \frac{[H_i]}{[X_i][P_i]} = \frac{[H_i^c]}{[X_i][P_i^c]}, \tag{13}$$

where the bracketed symbols denote the concentrations of the corresponding molecules under equilibrium state. Eliminating $[X_i]$ from (13) gives us

$$\frac{[H_i]}{[H_i^c]} = \frac{[P_i]}{[P_i^c]}. \tag{14}$$

And it is a reasonable assumption that the sum of the final concentrations of H_i and H_i^c is approximately equal to the initial concentration of X_i under sufficiently low reaction temperature as follows:

$$x_i \cong h_i + h_i^c, \tag{15}$$

where x_i denote the initial concentration of X_i, and h_i and h_i^c denote the final concentrations of H_i and H_i^c. From (14) and (15), we can obtain the final concentrations of H_i and H_i^c as:

$$\begin{aligned} h_i &= x_i w_i, \\ h_i^c &= x_i(1 - w_i), \end{aligned} \tag{16}$$

where w_i denotes the initial ratio of P_i in the probe mixture for X_i as (9).

Appendix B. Experimental Details: Materials and Methods

B.1. Target and Probe Oligonucleotide Sequences Design

First, four oligonucleotide DNA sequences, X_i ($i = 1, 2, 3, 4$) were designed in the sense orientation of the transcripts from the genes, fumarylacetoacetate (FAH), zyxin (ZYX), c-myb (MYB), and proteasome subunit alpha type 6 (PSMA6), respectively (Table 1). In designing, thermodynamic properties were considered to avoid cross homology between the sequences and undesired secondary structures such as hairpins or unspecific dimers. In addition, the melting temperature (T_m) and length were considered so that the maximum differences were less than 2 °C and 3 mer, respectively. Then a sequence of 30 repeats of adenine was concatenated to the 3′ ends of the designed sequences for affinity bead separation to select hetero-duplexes.

The sequences of the reporter and competitor probes, P_i and P_i^c, were determined to be reverse-complementary to X_i except the poly-adenine tail, and labeled with Cy5 and Cy3, respectively, at the 5′ ends. Table 1 shows the final set of designed sequences. These 12 sequences were purchased from Bioneer (Daejeon, Korea), and each sequence pellet was brought to a stock concentration 100 μM in TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH 8.0) and stored at −20 °C.

B.2. Weighted Probe Mixture Hybridization

In the verification of the weight encoding method, the hybridization reaction was performed in 50 μl of buffer containing 10 mM Tris–HCl (pH 7.5), 100 mM NaCl, and 1 mM EDTA, using the target sequence X_2 and the corresponding probes P_2 and P_2^c. The final target oligonucleotide amount was either 0.5 or 1 pmol per reaction; the total probe amount was set to be 10 pmol, constantly, but their composition (p_2, p_2^c) was varied as (0, 10), (3, 7), (5, 5), (7, 3), and (10, 0) pmol in each reaction. The reaction mixture was incubated at 95 °C for 3 min and the temperature was steadily lowered to 42 °C by 2 °C/min and incubated at 42 °C for 1 h using a thermal cycler (iCycler, Bio-Rad, Hercules, CA, USA).

In the pattern classification *in vitro*, the hybridization condition was the same with that of the weight encoding verification. The target *in vitro* patterns were prepared by mixing X_i ($i = 1, 2, 3, 4$) as shown in Table 2. In preparing the weight encoding probe mixture of P_i and P_i^c ($i = 1, 2, 3, 4$), the total amount of the probes for each i was kept to be 2500 fmol, constantly, and their compositions, (p_i, p_i^c), were (2375, 125), (1625, 875), (500, 2000), and (875, 1625) fmol, respectively, according to the determined mixing ratios.

B.3. Affinity Bead Separation and Fluorescence Signal Detection

To separate the target-probe hetero-duplexes, we performed post-hybridization at 42 °C for 1 h after adding 50 μl of oligo $(dT)_{25}$ magnetic beads (Dynal Biotech ASA, Oslo, Norway). Following the manufacturer's instructions (Dynal Biotech ASA), the hybridized samples were washed with 100 μl of washing buffer (10 mM Tris–HCl (pH 7.5), 0.15 M NaCl, 1 mM EDTA) four times using a magnet separation stand (Promega, Madison, WI, USA) to select out hetero-duplexes. Finally, the hetero-duplexes captured

on beads were eluted by heating at 75 °C for 10 min with 100 μl of elution buffer (10 mM Tris–HCl, pH 7.5), and the same volume of the eluted solution was collected and transferred into a 384-well black plate (Greiner, Frickenhausen, Germany). The fluorescence signal intensities were measured using a computer-controlled fluorescence plate reader (GENios Pro, Tecan, Mannedorf, Switzerland) at the wavelength 590 nm and 670 nm for Cy3 and Cy5, respectively.

References

Bakar, R.B.A., Watada, J., Pedrycz, W., 2008. DNA approach to solve clustering problem based on a mutual order. Biosystems 91 (1), 1–12.

Benenson, Y., Gil, B., Ben-Dor, U., Adar, R., Shapiro, E., 2004. An autonomous molecular computer for logical control of gene expression. Nature 429 (6990), 423–429.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer Science + Business Media.

Chen, J., Deaton, R., Wang, Y.Z., 2005. A DNA-based memory with *in vitro* learning and associative recall. Nat. Comput. 4 (2), 83–101.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M., 1999. Expression profiling using cDNA microarrays. Nat. Genet. 21 (Suppl. 1), 10–14.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531–537.

Henkel, C.V., Bäck, T., Kok, J.N., Rozenberg, G., Spaink, H.P., 2007. DNA computing of solutions to knapsack problems. Biosystems 88 (1–2), 156–162.

Kessler, C., 1994. Nonradioactive analysis of biomolecules. J. Biotechnol. 35 (2–3), 165–189.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7 (6), 673–679.

Lambert, K.N., Williamson, V.M., 1993. cDNA library construction from small amounts of RNA using paramagnetic beads and PCR. Nucleic Acids Res. 21 (3), 775–776.

Lim, H.-W., Lee, J.Y., Kim, S.N., Yoo, S.-I., Park, T.H., Zhang, B.-T., 2004. DNA-based perceptron and application to binary classification of gene expression. In: Ferretti, C., et al. (Eds.), Preliminary Proceedings of The 10th International Meeting on DNA Computing 2004. pp. 439.

Lim, H.-W., Lee, S.H., Yang, K.A., Yoo, S.-I., Park, T.H., Zhang, B.-T., 2007. *In vitro* molecular pattern classification using DNA-based weighted sum computation. In: Garzon, M., Yan, H. (Eds.), Preliminary Proceedings of The 13th International Meeting on DNA Computing 2007, p. 90.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat. Genet. 21 (Suppl. 1), 20–24.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebet, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., Golub, T.R., 2005. MicroRNA expression profiles classify human cancers. Nature 435 (7043), 834–838.

Mills Jr., A.P., 2002. Gene expression profiling diagnosis through DNA molecular computation. Trends Biotechnol. 20 (4), 137–140.

Mills Jr., A.P., Turberfield, M., Turberfield, A.J., Yurke, B., Platzman, P.M., 2001. Experimental aspects of DNA neural network computation. Soft Comput. 5 (1), 10–18.

Mills Jr., A.P., Yurke, B., Platzman, P.M., 1999. Article for analog vector algebra computation. Biosystems 52 (1–3), 175–180.

Oliver, J.S., 1997. Matrix multiplication with DNA. J. Mol. Evol. 45 (2), 161–167.

Paun, G., Rozenberg, G., Salomaa, A., 1998. DNA Computing: New Computing Paradigms. Springer-Verlag, Berlin/Heidelberg.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. P. Natl. Acad. Sci. USA 98 (26), 15149–15154.

Sakakibara, Y., Suyama, A., 2000. Intelligent DNA chips: logical operation of gene expression profiles on DNA computers. Genome Inform. 11, 33–42.

SantaLucia Jr., J., Hicks, D., 2004. The thermodynamics of DNA structural motifs. Annu. Rev. Biophys. Biomol. Struct. 33, 415–440.

Wang, X., Bao, Z., Hu, J., Wang, S., Zhan, A., 2008. Solving the SAT problem using a DNA computing algorithm based on ligase chain reaction. Biosystems 91 (1), 117–125.

Wetmur, J.G., 1991. DNA probes: applications of the principles of nucleic acid hybridization. Crit. Rev. Biochem. Mol. Biol. 26 (3–4), 227–259.

Zhang, B.-T., 2008. Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory. IEEE Comput. Intell. Mag. 3 (3), 49–63.

Zhang, B.-T., Jang, H.-Y., 2005. A bayesian algorithm for *in vitro* molecular evolution of pattern classifiers. Lecture Notes Comput. Sci. (DNA10) 3384, 458–467.

Zoraida, B.S.E., Arock, M., Ronald, B.S.M., Ponalagusamy, R., 2009. A novel generalized design methodology and realization of Boolean operations using DNA. Biosystems 97 (3), 146–153.