Non-linear molecular pattern classification using molecular beacons with multiple targets

In-Hee Lee a,1,2, Seung Hwan Lee b,1, Tai Hyun Park b, Byoung-Tak Zhang a,3,*

a School of Computer Science and Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea
b School of Chemical and Biological Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea

1. Introduction

DNA computing was originally highlighted as an alternative approach toward difficult computational problems such as the Hamiltonian path problem (Adleman, 1994; Martinez-Perez et al., 2005), the satisfiability problem (Lipton, 1995; Sakamoto et al., 2000; Braich et al., 2002; Rozenberg and Spaink, 2003), and the maximal clique problem (Ouyang et al., 1997). In recent years, some researchers have shifted their focus on the potential of DNA computing in biological data analysis, especially for medical or diagnostic purposes (Benenson et al., 2004; Bayer and Smolke, 2005; Shapiro and Benenson, 2006). Since both information and algorithm are represented by molecules and implemented by in vitro experiments, DNA computing can naturally handle biological data without converting them into digital data. An automated mechanism for molecular diagnosis or drug discovery based on Boolean logic has been demonstrated in Benenson et al. (2004). Bayer and Smolke (2005) suggested an anti-sense technology to develop an anti-switch that can be programmed to regulate a genetic network in response to cellular states. Shapiro and Benenson (2006) introduced a DNA computing algorithm to release a piece of DNA, designed to act as a drug by interfering with pathogenes, depending on the expression of particular genes.

Another DNA computing approach to detecting a set of microRNAs that shall be expressed when the cell is in an abnormal state is demonstrated in Lee et al. (2008a).

Most of the existing DNA computing methods for biological data analysis were based on binary or Boolean data representation. However, for more advanced analysis of biological data, it requires arithmetic operators capable of handling quantitative information. A few theoretical models towards algebraic operations have been proposed (Oiler, 1997; Mills et al., 1999, 2001), but without experimental verification. Recently, a DNA computing model that can perform quantitative analysis using weighted sum operations via molecular beacons has been proposed and experimentally verified (Lee et al., 2008; Lim et al., 2010). In spite of practical advantages such as simplicity of experimental procedure, its power of data analysis is limited to linear classification. More complex problems, however, require non-linear classification capability. Here we develop a model for non-linear in vitro classification by designing molecular beacons that can hybridize with multiple targets. The fluorescence signal from the molecular beacon changes in a non-linear way according to the pattern of its targets. Non-linear pattern classification can be achieved by combining the signals from multiple molecular beacons as the way in the weighted sum operation.

The rest of the paper is organized as follows. Section 2 gives the theoretical background. Section 3 describes the design of the proposed molecular pattern classification in detail. In Section 4, we present experimental results on a mathematical problem and a real-life biological data analysis. Section 5 discusses points to be considered to apply the work in more general context. Section 6 draws conclusions.
2. Theoretical background for non-linear pattern classification

2.1. Separability of patterns and Cover’s theorem

Let us consider a general pattern classification problem between two classes: a set \( \mathcal{X} \) of \( N \) patterns \( x_1, x_2, \ldots, x_N \), each of which belongs to one of two classes \( C_1 \) and \( C_2 \). A binary partition (dichotomy) \( \{C_1, C_2\} \) of patterns \( \mathcal{X} \) is said to be separable with respect to the family of surfaces if a surface exists in the family that separates the points in the class \( C_1 \) from those in the class \( C_2 \) (Haykin, 1999). Those surfaces can be either linear or non-linear, but the problems that are separable by linear surfaces are considered relatively easy to solve.

Suppose that the pattern \( x \) is a vector in an \( m_0 \)-dimensional input space. We can define a mapping function, \( \varphi(x) = [\varphi_1(x), \ldots, \varphi_{m_1}(x)] \), that maps points in \( m_0 \)-dimensional input space into corresponding points in \( m_1 \)-dimensional space. Then the given dichotomy is said to be \( \varphi \)-separable if the mapped patterns \( \varphi(x) \) are linearly separable (Cover, 1965): there exists a vector \( w = (w_1, w_2, \ldots, w_{m_1}) \) such that \( s(x; w) \geq 0 \) for \( x \in C_1 \) and \( s(x; w) < 0 \) for \( x \in C_2 \), where \( s(x; w) = \sum_{i=1}^{m_1} w_i \varphi_i(x) \). Cover’s theorem (Cover, 1965) states that a complex pattern classification problem can be more likely to be \( \varphi \)-separable for non-linear mapping function \( \varphi \) as we take the higher value for \( m_1 \), and that, in some cases, the use of non-linear mapping function may be sufficient for \( \varphi \)-separability without having to increase \( m_1 \). Fig. 1(a) shows an example of \( \varphi \)-separable problem which was not linearly separable in original \((x_1, x_2)\) space becomes linearly separable by non-linear mapping into \((\varphi_1, \varphi_2)\) space.

2.2. Radial-basis function network

Radial-basis function (RBF) networks are a class of artificial neural network which implements the core principle of Cover’s theorem. An RBF network typically consists of three layers: an input layer, a hidden or middle layer and an output layer (Fig. 1(b)). The input layer consists of individual elements, \( x_i \), of input pattern \( x \). Each unit in the hidden layer is associated with a radial-basis function \( \varphi(||x - e_i||) \) with its own parameter \( e_i \). The output units define a weighted sum of middle-layer units. Thus, for the RBF network with \( n \) units in the middle layer, the output of the network can be represented as a function \( f(x) = \sum_{i=1}^{n} w_i \varphi(||x - e_i||) \), where \( w_i \) denotes the weight for the \( i \)-th unit in the middle layer.

RBF networks carry out non-linear mapping of input pattern into \( m \)-dimensional space of RBF units, \( \varphi(||x - e_i||) \), in the middle layer. The most common choice of an RBF is a real-valued unimodal function whose value depends only on the distance between input \( x \) and pre-defined ‘center’ vector \( e \) and reaches maximum when the distance is zero. For example, Gaussian radial-basis functions are defined as

\[
\varphi(||x - e||) = \exp(-\beta||x - e||^2),
\]

for some \( \beta > 0 \). The weight vector \( w \) that connects the middle layer with output layer defines a linear separating surface in \( \varphi \)-space. Thus a trained RBF network embeds a non-linear mapping that can \( \varphi \)-separate the given input patterns. RBF networks with an enough number of units in the middle layer can be trained to solve any complex pattern classification problem.

To develop a DNA computing model capable of non-linear classification based on the same principle as RBF networks, we need to be able to carry out two operations. One is the non-linear mapping similar to the radial-basis function and the other is the weighted sum. One solution to the latter is to use the quantity of molecules to represent the weights as demonstrated in Lim et al. (2010) (used linear DNA probes) and Lee et al. (2008) (used conventional molecular beacons). In the following, we focus on the molecular beacon design that can also do the former.
3. Molecular beacon design for non-linear classification

3.1. Molecular beacon design with multiple targets

Molecular beacons are single-stranded DNA probes useful for solution-based nucleic acid detection (Tyagi and Kramer, 1996). Typical molecular beacons, with a fluorophore attached at the 5’ end and a quencher at the 3’ end, form a hairpin-like, stem-loop structure by itself. In this closed state, the fluorophore is located within a short distance of the quencher and the energy absorbed by the fluorophore is not emitted as fluorescence but transferred to the quencher and released as heat. As a result the system is unable to fluoresce strongly on its own. When a target nucleic acid is introduced, the formation of the rigid helical structure between the loop of the molecular beacon and the target causes the dissociation of the hairpin stem and the separation of the fluorophore from the quencher. Since the distantly located quencher is no longer able to absorb the energy from the excited fluorophore, it emits strong fluorescence. Since its first introduction in 1996, the applications of molecular beacons have been explored in many research fields such as gene expression, biosensor development and clinical diagnosis (Kim et al., 2008; Wang et al., 2009). But in most works, molecular beacons usually interact with a single target. We design a molecular beacon that can hybridize with multiple targets so that it can respond to complex input patterns consisting of target molecules.

Fig. 2 shows the working principle of the proposed beacon design. Here, the loop of molecular beacon is a concatenation of complementary sequences for its targets. If both of the targets exist in the solution, the loop and the targets will form helical structure. Consequently the fluorophore and the quencher will be separated and strong fluorescence will be emitted (Fig. 2(a)). If only one of the targets exists in the solution, only half of the loop will form a helical structure. Therefore, the fluorophore and the quencher will not be separated as much as when both targets exist (Fig. 2(b)). This half-opened molecular beacon can still emit weak fluorescence. If none of the targets exists, the beacon will stay closed and emit no fluorescence (Fig. 2(c)).

3.2. Use of a molecular beacon as a radial-basis function unit

The signals from the beacons as designed above will reach their maximum when the amount of the target molecules reaches its maximum (assuming the amount of beacon is sufficient to react with the target molecules). But, for molecular beacons to be used as an RBF unit, \( \psi(||x - c||) \), it is more natural to assume that the amount of target molecules is proportional to \( ||x - c|| \). Thus the signals from the beacons need to be inverted so that the RBF unit has the maximum value when \( ||x - c|| = 0 \), i.e., there is little target molecules. This can be achieved by making beacons to start from an open state instead of a closed state. That is, the beacons are already hybridized with complements of the targets from the beginning. But these complements of the targets still have hangovers that can hybridize with the targets. If none of the targets exists, the beacons will stay open and emit a weak signal. If any one of the targets exists, it will first hybridize to the hangover part of its complement and eventually take the complement off from the beacon. Therefore, the beacons will be partially open and emit a weak signal. If both targets exist, the beacons will return to the closed state. Therefore, the signal intensity from molecular beacons will be inverted from that of Fig. 2.

Now we could implement a type of radial-basis function that takes an input pattern \( x \) of target molecules and a center \( c \) of RBF. Suppose an input pattern \( x = (x_1, x_2) \) and a center for RBF \( c = (c_1, c_2) \). \( x_1 \) and \( x_2 \) are complementary to \( c_1 \) and \( c_2 \), respectively. The values of \( c_1, c_2, x_1 \) and \( x_2 \) are represented as the amounts of corresponding molecules in the solution. \( x \) and \( c \) are hybridized to each other and then mixed with four kinds of molecular beacons as shown in Fig. 3. If the amounts of \( c \) and \( x \) are similar, then almost all of \( c_1 \) and \( c_2 \) will be hybridized to \( x_1 \) and \( x_2 \), respectively, and little will be left to hybridize with their complements attached to the beacons.

![Fig. 2. Molecular beacons with multiple targets. (a) When both targets are present, the beacon is fully opened emitting the strongest signal. (b) When only one of the targets is present, the beacon is half-opened and emits a weaker signal than in (a). (c) None of the targets is present. The beacon is fully closed and emits no signal.](image-url)
Therefore, most of the four molecular beacons will stay open and the summed signal will be the strongest. However, if any components of $\mathbf{x}$ and $\mathbf{c}$ are not similar, at least one of four molecules ($x_1, x_2, c_1$, and $c_2$) will remain to hybridize with the complements on the beacons, closing the beacons. Thus the total signal from the four beacons will achieve the maximum level only when $\mathbf{x}$ and $\mathbf{c}$ is similar to each other and decrease as the difference $|\mathbf{x} - \mathbf{c}|$ increase, which fits the characteristics of radial-basis function.

With those four molecular beacons in Fig. 3 as a single RBF unit, various non-linear functions can be generated by weighted summation of multiple RBF units. The weighted sum of multiple RBF units can be easily computed in the same way as proposed in Lim et al. (2010), i.e., using larger amounts of molecular beacons for RBF units with larger weights and vice versa.

### 3.3. Experimental test of molecular beacons as an RBF unit

We first verified the basic workings of individual molecular beacons for an RBF unit. Fig. 4(a) shows the responses of a single molecular beacon started from open state. Since it has started from open state, its signals should be the inverse of those shown in Fig. 2. The results show that the beacons emit the weakest signal when both targets exist; the strongest signal when none of targets exists. When hybridized with only one target, the beacons emit significantly reduced signal than when hybridized with both targets.

Next, we tested the functionality of a set of such beacons as an RBF unit (Fig. 3). For this purpose, we measured output signals from an RBF unit for various input patterns $\mathbf{x}=(x_1, x_2)$ and center $\mathbf{c}=(c_1, c_2)$. The results in Fig. 4(b) show that the actual behavior (right) of the RBF unit is concordant with the expected behavior (left): the actual output is strongest when the expected output is highest and weakest when the expected output is lowest. And, for all cases when the unit is expected to output intermediate signal, the actual output signals were between the strongest and the weakest. However, it should be also noted that the actual output signal fluctuates among cases when the expected output levels are the same. It could be attributed to the sensitivity of molecular beacon to the base composition of input molecules, which has been utilized to detect single-nucleotide polymorphism (Mhlanga and Malmberg, 2001). Other factors for the variance in signal could include GC content or melting temperature of input molecules and beacons.

Fig. 4(c) plots the signals as a function of Euclidean distance between $\mathbf{x}$ and $\mathbf{c}$, $|\mathbf{x} - \mathbf{c}|$. The signal degrades as the distance increases as expected for a radial-basis function. From these test results, we can conclude that the beacon design in Fig. 3 can implement an RBF unit for molecular pattern classification.

### 4. Application to mathematical problem and complex bioanalysis

#### 4.1. Mathematical problem: XOR

The RBF network built from molecular beacons is first applied to solve a classical non-linear classification problem: exclusive OR (XOR). It is originally defined for binary input patterns, (0, 0), (0, 1), (1, 1), and (1, 0), where each pattern is either in class 0 or class 1. The first and third patterns are in class 0, as shown by $0 \oplus 0 = 0$ and $1 \oplus 1 = 0$ where $\oplus$ denotes the Boolean operator, exclusive OR. And the other patterns are in class 1, as shown by $0 \oplus 1 = 1$ and $1 \oplus 0 = 1$. Since patterns in class 0 and 1 are located in opposite diagonal lines of unit square, they are not linearly separable.

We can redefine XOR problem in real values by shifting and scaling input patterns without changing class memberships (Fig. 5(a)). It is still not linearly separable, but it is $\psi$-separable. For example, the input patterns in Fig. 5(a) can be linearly separated by a set of curves,

$$\psi(|\mathbf{x} - \mathbf{c}_1|) + \psi(|\mathbf{x} - \mathbf{c}_2|) - C = 0,$$

where $\mathbf{c}_1 = (10, 50)$ and $\mathbf{c}_2 = (600, 600)$. Fig. 5(a) also shows one of the possible separating curves.

As shown above, the XOR problem can be solved in vitro by an RBF network with two RBF units. We set $c_1 = 10$ (pmole) and $c_2 = 50$ (pmole) for unit 1 ($\psi(|\mathbf{x} - \mathbf{c}_1|)$), and $c_1 = c_2 = 600$ (pmole) for unit 2 ($\psi(|\mathbf{x} - \mathbf{c}_2|)$).

Fig. 5(b) shows the input patterns with their classes and output signals from RBF unit as well as sum of output signals. Input patterns are shown in pmole and output signals are in
Fig. 4. (a) Average signal from individual molecular beacon when started from the open state. As expected, beacons generally show strongest signals when none of the targets exists. (b) Test results for an RBF unit. For given $x$ and $c$, the expected output (left) and the actual output (right) show similar trends. (c) Plot of the total signal in (b) as a function of distance between $x$ and $c$, $||x-c||$.

Fig. 5. (a) Input patterns for XOR problem and one of possible separating lines. (b) Input patterns and the corresponding signals of individual RBF units in the middle layer and the final output of the RBF network (sum of signals from RBF units in the middle layer). Input patterns are in pmole and the output signals from molecular beacons are in arbitrary unit (AU). (c) Plot of outputs from the RBF units. Outputs from different classes are separable by a linear line. (d) Distribution of final output signals from the RBF network for different classes.
Fig. 6. (a) Input expression profiles of microRNAs and the corresponding output signals from an RBF network. Inputs are in pmole and output signals are in arbitrary unit (AU). (b) Plot of input expression profiles. (c) Discrimination of different classes by output signals.

4.2. Application to bioanalysis with microRNAs

As an application to biological data analysis, the RBF network is also used to classify the expression profiles of microRNAs. The expression profile data are obtained from Lu et al. (2005). The original data consist of 89 samples containing expression levels of 151 microRNAs that are known to be related to cancers. From this data, we chose 2 microRNAs (hsa-miR-210 and hsa-let-7c) for 4 samples with prostate cancer and 4 normal samples (Fig. 6(a)). As shown in Fig. 6(b), these 8 samples are not linearly separable. One of the possible solutions for this dataset can be formulated as

$$\varphi(||x - c||) - C = 0,$$

where \(x = (\text{hsa-miR-210, hsa-let-7c})\) and \(c = (577.5, 140)\). Similar to the XOR problem, the above function can be implemented by an RBF network with one RBF unit. We set \(c_1 = 577.5\) (pmole) and \(c_2 = 140\) (pmole) for the RBF unit. We used DNAs with the same sequence as microRNAs to simplify the procedure, but we expect the same results with RNAs.

Fig. 6(a) shows input patterns and their output signals from the RBF network. Input patterns are shown in pmole and output signals are in AU. Fig. 6(c) shows that the outputs of the RBF network can discriminate the input patterns of different classes (147386 (AU, normal) and 92188.25 (AU, cancer); p-value = 0.0067).

5. Discussion

So far we successfully showed the application of trained RBF network to solve non-linear problems. To be utilized in more general context, the “training” of RBF network should be considered. The training process can be divided into two levels. One is the computational level: to obtain numeric solutions for the parameters of RBF network such as \(w_i\) and \(c_i\) in \(f(x) = \sum_{i=1}^{n} w_i \varphi(||x - c_i||)\). The other is the biological level: to represent numeric values with appropriate molecule quantities. The first level is relatively easy compared to the second one since RBF network is one of the most studied computational model for non-linear classification problem and several programs developed to that purpose are available.

But the second level of training requires intensive trial and error experiments. We considered the concentration of salt, the amount of beacon and target DNA, the sensitivity of fluorescence reader, the specificity of beacon sequence and the cross-homology of beacon sequence to obtain best results. In our experiments, the problems were simple enough that we did not require intensive trials. But it would become more complicated and demanding for complex problems.

Another point to be discussed is the robustness of the designed RBF network. From Fig. 4(b), it could be noted that the actual output signal fluctuates among cases when the expected output levels are the same. Similarly, variances were observed in outputs from individual RBF units (Figs. 5(b) and 6(a)). It could be due to the inherent sensitivity of molecular beacon to the nucleotide composition in the input molecules. Also the differences in melting temperature of input molecules and molecular beacons could have resulted in different reaction rate and output signal level. Lastly, it could be attributed to the variance in signal when only one input is given. Since only half of beacon is hybridized to the input, the other half can move relatively freely. It could be bent towards the other end reducing signal level close to the minimum. Or it could be stretched away from the other end increasing signal level close to the maximum. The effect of each of these factors on our RBF network design is not clear and requires further investigation.

6. Conclusion

We proposed a DNA computing method for non-linear pattern classification in vitro. We designed molecular beacons to hybridize with multiple targets to construct RBF (radial basis function) units. The signals from RBF units are combined to make an RBF network, the final non-linear pattern classifier. We experimentally verified that the molecular signals from the constructed classifier were able to discriminate input patterns. The proposed method was successfully applied to XOR function and microRNA expression profile classification, showing its capability of solving non-linear problems. Compared to other DNA computing methods for (biological) pattern classification, this is a unique advantage for real-world application. Also it should be noted that the proposed method does not require an enzymatic reaction, which makes its implementation easier than other methods.

In this paper, we have focused on implementing an already-trained classifier using molecular beacons. Integrating training process into our system would be one of important next steps. There exist several studies about training molecular classifier (termed as hypernetworks) based on evolutionary process (Zhang
and Jang, 2005; Zhang, 2008). Since their representation of classifier has similarity to the RBF network in our system, we expect the learning process in these studies could be the most promising approach towards expanding our system.

Acknowledgements

This research was supported in part by the Ministry of Knowledge Economy (MKE) through the Molecular Evolutionary Computing (MEC) project, the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science & Technology (MEST) (No. 0421-20110032), and the BK21-IT Program. The ICT at Seoul National University provided research facilities for this study.

Appendix A. Materials and methods

A.1. Molecular beacon and target sequence preparation

Four molecular beacon sequences and four target sequences were designed as in Table A.1. 5’-End and 3’-end of beacon was labeled with Cy3 dye and BHQ-2 quencher, respectively. The bold-faced bases indicate the stem sequence. The loop part of beacon was complementary to two pair of targets. The target sequences (C1, C2, x1, and x2) were purchased from Bioneer (Deaen, Korea) and molecular beacon were purchased from IDT (Integrated DNA Technologies, Inc., Coralville, IA, USA). Each sequence was brought to a stock concentration of 100 µM in distilled water and stored at −20 °C.

A.2. Verification of a molecular RBF unit

As indicated in Fig. 3, each molecular beacon in an RBF unit should start from open state. For this purpose, we prepared each beacon and its targets as follows. The amount of beacon was set to be 20 pmole. Each oligonucleotide target was set to be 100 pmole, resulting in 200 pmole for a total amount of target. All the hybridization reactions for each molecular beacon were performed in 50 µl reaction buffer containing 3.5 mM MgCl2, 400 mM KCl and 10 mM Tris–HCl (pH 8.0). The reaction mixture was incubated at 95 °C for 3 min and the temperature was steadily lowered to 10 °C by 0.5 °C/min using a thermal cycler (iCycler, Bio-Rad, Hercules, CA, USA).

Next, various combinations in terms of concentration of targets were prepared. The amounts of targets were varied as shown in Fig. 4(b). Then x1 and x2 were hybridized with C1 and C2, respectively. The hybridization condition was the same as the above. After hybridization, each hybridized solution was mixed with four kinds of open state molecular beacons as shown in Fig. 3. Fluorescence signal intensities were measured using a computer-controlled fluorescence plate reader (GENios Pro, Tecan, Mannedorf, Switzerland)

Table A.1

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1, x2</td>
<td>Cy3-CCCGACACCGCGTGCTAGGTACTATGCCG-BHQ2</td>
</tr>
<tr>
<td>C1, C2</td>
<td>Cy3-CCCGACACCGCGTGCTAGGTACTATGCCG-BHQ2</td>
</tr>
<tr>
<td>x1, x2</td>
<td>Cy3-CCCGACACCGCGTGCTAGGTACTATGCCG-BHQ2</td>
</tr>
<tr>
<td>C1, C2</td>
<td>Cy3-CCCGACACCGCGTGCTAGGTACTATGCCG-BHQ2</td>
</tr>
<tr>
<td>C1</td>
<td>CTGGTCGTGACGCGGCTGGA</td>
</tr>
<tr>
<td>x1</td>
<td>TACGGGCGTTCACAGCCACAG</td>
</tr>
<tr>
<td>C2</td>
<td>TGAGTGGATGCTGTATCAGGTT</td>
</tr>
<tr>
<td>C2</td>
<td>AACCACAACACTACTCCCA</td>
</tr>
</tbody>
</table>

at the wavelength of 590 nm. Finally, the signal from four molecular beacons was added.

A.3. Experiment for XOR problem

Two RBF units were prepared to solve the XOR problem. For unit 1, C1 and C2 were set to 10 pmole and 50 pmole, respectively. For unit 2, C1 = C2 = 600 pmole. Input patterns, x1 and x2, were prepared as shown in Fig. 5(b). Each input pattern was mixed with C1 and C2 of unit 1 and unit 2, respectively. The hybridization was performed using the same method as described in Section A.2 then the reacted solutions were mixed with the molecular beacons for the two RBF units respectively, each consisting of four open state molecular beacons as shown in Fig. 3. Finally, the intensities of the fluorescence signals were measured as described in Section A.2 and combined.

A.4. Experiment for microRNA

Four normal samples and four cancer samples were prepared as shown in Fig. 6(a). Only one RBF unit was prepared with C1 = 577.5 pmole and C2 = 140 pmole. Each input pattern was *X* (hsa-miR-210, hsa-let-7c) was hybridized with the molecular RBF unit. The hybridization condition was the same as described in Sections A.2 and A.3. The hybridized solutions were mixed with four kinds of open state molecular beacons. The fluorescence signals were measured as in Section A.2 and their intensities were summed together.

References


