In vitro molecular machine learning algorithm via symmetric internal loops of DNA

Ji-Hoon Lee a, 1, Seung Hwan Lee b, 1, Christina Baek c, Hyosun Chun d, Je-hwan Ryu c, Jin-Woo Kim e,f, Russell Deaton g, Byoung-Tak Zhang a,c,d,h,*

a Graduate Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
b School of Chemical and Biological Engineering, Seoul National University, Seoul, Republic of Korea
c Graduate Program in Brain Science, Seoul National University, Seoul, Republic of Korea
d School of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
e Biological and Agricultural Engineering, University of Arkansas, Fayetteville, AR, USA
f BioNano Technology Laboratory, Institute for Nanoscience and Engineering, University of Arkansas, Fayetteville, AR, USA
g Electrical and Computer Engineering, University of Memphis, Memphis, TN, USA
h Graduate Program in Cognitive Science, Seoul National University, Seoul, Republic of Korea

A R T I C L E   I N F O
Article history:
Received 1 February 2017
Received in revised form 12 April 2017
Accepted 24 April 2017
Available online 29 April 2017

Keywords:
Biomolecular computation
Hypernetwork
Machine learning
Classification

A B S T R A C T
Programmable biomolecules, such as DNA strands, deoxyribozymes, and restriction enzymes, have been used to solve computational problems, construct large-scale logic circuits, and program simple molecular games. Although studies have shown the potential of molecular computing, the capability of computational learning with DNA molecules, i.e., molecular machine learning, has yet to be experimentally verified. Here, we present a novel molecular learning in vitro model in which symmetric internal loops of double-stranded DNA are exploited to measure the differences between training instances, thus enabling the molecules to learn from small errors. The model was evaluated on a data set of twenty dialogue sentences obtained from the television shows Friends and Prison Break. The wet DNA-computing experiments confirmed that the molecular learning machine was able to generalize the dialogue patterns of each show and successfully identify the show from which the sentences originated. The molecular machine learning model described here opens the way for solving machine learning problems in computer science and biology using in vitro molecular computing with the data encoded in DNA molecules.

© 2017 Published by Elsevier Ireland Ltd.

1. Introduction

Machine learning is increasingly important for the development of intelligent systems in digital computers. Despite substantial advances in the field, developing machines with human-level intelligence remains a major challenge because of the fundamental differences between the two entities. In the human brain, learning is a highly complex process comprising many biological functional modules such as cells, biomolecules, and endogenous chemicals. Therefore, to bridge the gap between humans and machines, machine learning should ideally be implemented at the molecular level.

However, previous studies have reported findings regarding molecular computing (Adleman, 1994; Amir et al., 2014; Benenson et al., 2001; Brown et al., 2014; Chen et al., 2013; Mao et al., 2000; Pei et al., 2010; Qian et al., 2011; Seelig et al., 2006; Stojanovic and Stefanovic, 2003; Wang et al., 2014; Winfree et al., 1998; Yurke et al., 2000; Zhang and Kim, 2006; Zhang and Seelig, 2011), pattern classification, and associative recall (Chen et al., 2005; de Murieta and Rodríguez-Patón, 2012; Lakin et al., 2012; Lee et al., 2011a; Lee et al., 2011b; Lim et al., 2010; Lim et al., 2002; Pei et al., 2010; Qian et al., 2011; Zhang, 2008; Zhang and Jang, 2004; Zhang and Kim, 2006). Although these models have demonstrated the potential to reduce the gap between human and machines through implementation of molecules to the machine, there are some limitations, particularly the absence of learning and generalization. Learning and generalization from general training data encoded by molecules should be necessary for in vitro analysis at the molecular level because they enhance the accuracy of machines.

http://dx.doi.org/10.1016/j.biosystems.2017.04.005
0303-2647/© 2017 Published by Elsevier Ireland Ltd.
We have proposed the hypernetworks model (Lee et al., 2011a; Zhang, 2008; Zhang and Kim, 2006) that has a molecular evolutionary architecture based on a probabilistic graphical model for cognitive learning and memory (Zhang, 2008). The hypernetworks are based on weighted hyperedges formed by several vertices as features. It has been used for several tasks that require for human-level intelligence (Zhang, 2008). Because this model was inspired by the complex molecular networks in biological systems, machine learning can be implemented at the molecular level. In our previous study, the hypernetwork models were implemented in silico (Kim et al., 2006; Zhang and Jang, 2005), and with in vitro experimental design for learning and generalization. However, the implementation of the hypernetwork model to the wet-lab experimental setting has not yet been demonstrated.

In this study, a molecular machine was used to address learning problems involving the recognition and classification of natural language sentences, which were constructed from dialogue taken from the television shows Friends and Prison Break. In the molecular machine, learning was achieved using a type of DNA secondary structure, i.e., symmetric internal loops of double stranded DNA (dsDNA) (Cekan and Sigurdsson, 2012; Peritz et al., 1991; Zacharias and Hagerman, 1996; Zeng and Zocchi, 2006). These molecules were first trained using the full set of television show sentences with given classifications. Subsequently, learning performance was measured by calculating the percentage of correctly classified test sentences, which also consisted of sentences taken from the television shows’ dialogue, which were unknown to the molecular machine. Therefore, the molecular learning machine was designed to recognize natural language and correctly classify it according to a given question. Our results showed that the molecular learning machine was able to generalize unknown data. Indeed, a 100% success rate was observed during our 20-sentence experiments.

2. Model: DNA hypernetwork for classification problems

A hypernetwork $H$ is defined as triple $H = \{X, E, W\}$, where $X = \{x_1, x_2, \ldots, x_n\}$, $E = \{E_1, E_2, \ldots, E_j\}$, and $W = \{w_1, w_2, \ldots, w_j\}$ are the sets of vertices, hyperedges, and weights, respectively. A vertex is the value of an attribute or a feature. A hyperedge is the combination of vertices, and can contain more than two vertices. A weight is the probability of hyperedges.

The hypernetwork can be expressed as an ensemble of DNA molecules as described in our previous study (Lee et al., 2011a). The molecular ensemble is maintained in a microtube, and can be changed when new training examples were added. The trained ensemble is stored in microwells to complete the training process. Subsequently, when a new example was encountered, similar or identical training instances were retrieved from the hypernetworks and used for classification problems. Specifically, we used a set of training data from target distribution $A$ to train hypernetwork $A$ ($H_A$). Similarly, we used the training data from distribution $B$ to create hypernetwork $B$ ($H_B$). These hypernetworks have hyperedges that resemble their target distributions, which allows us to use the hypernetworks for classification problems. Suppose we have $H_A$ and $H_B$ trained by A1–A5 and B1–B5, respectively, and we have A6–A10 and B6–B10, we can then hybridize A6–A10 to $H_A$ and $H_B$ and perform electrophoresis of the two samples in two different lanes. After electrophoresis, we observed different band patterns between the two lanes. $H_A$ and A6–A10 duplexes moved faster and migrated further down the gel than $H_B$ and A6–A10 duplexes. However, when B6–B10 sequences were used, we obtained opposite results. This process was the classification test. If an A11 sequence was not used for any training processes, and has more similar characteristics to $H_A$ than to $H_B$, then the classification test shows that similarity of A11 is closer to $H_A$ than to $H_B$. A parallel DNA hybridization process was used to calculate the distances between queries and training examples.

The classification algorithm is as follows:

- Give a query input $x_q$ to be classified:
  - Hybridize $x_q$ into Hypernetwork A ($H_A$) and Hypernetwork B ($H_B$).
  - For each distance $d_i$ (Hamming distances between instances, $i = 1, \ldots, n$):
    1. Measure the intensity $I_A$ of the hybridized molecules with $H_A$ and $x_q$ within distance $d_i$.
    2. Measure the intensity $I_B$ of the hybridized molecules with $H_B$ and $x_q$ within distance $d_i$.

Return $y^* = \text{argmax}_{y \in \{A, B\}} I_y$

End

In this classification algorithm, the given query $x_q$ hybridizes with $H_A$ and $H_B$, respectively, in each tube. The distance refers to the similarity of sequence $x_q$ with $H_A$ and $H_B$. A smaller distance means they are more similar. The distance $d_i$ of the perfect match case, one-variable mismatch, two-variable mismatch, and three-variable mismatch are 0, 1, 2, and 3, respectively. Intensity $I_x$ can

![Diagram](image-url)
be relatively measured by gel band analysis. If $l_A > l_B$ in the same distance, we classify $q_A$ to class A, and if $l_A < l_B$, we classify $q_A$ to class B. If they do not have the same distance matches, we classify $q_A$ to the class that has the lower band spot in polyacrylamide gel electrophoresis (PAGE).

3. Realization in DNA computing

3.1. Basic architecture based on symmetric internal loops

In this study, the DNA hypernetwork was applied to classify unknown test sentences based on whether the style of corpus was similar to the sentences taken from the television shows’ dialogue. Sentences were selected from the television shows Friends and Prison Break as the training dataset. Each sentence was encoded by DNA molecules. A DNA hypernetwork consists of DNA hyperedges that represent hypergraphs containing several vertices as features or attributes. In the present study, a word is a variable, whereas a hyperedge is an example that is part of a full sentence collected from the script dialogue. We used an order-three hyperedge

![Diagram](image)

**Fig. 2.** (A) A typical hyperedge structure. The black and blue arrows indicate single stranded DNAs (ssDNAs) in the 5' to 3' direction. The black regions have fixed nucleotide sequences and include the primer and inner tag, whereas the blue variable regions contain different sequences for encoding specific words. (B) DNA sequences for the hyperedge of the phrase “I love babies.” (C) Predicted structures of matches and mismatches (hybridization structures predicted by Mfold (Zuker, 2003)). From top to bottom: i) a perfect match; ii) a 1-variable mismatch in the v3 region; iii) a 2-variable mismatch in the v1 and v2 regions; iv) a 3-variable mismatch in the v1, v2, and v3 regions; and v) an example of the small partial mismatches that can occur in variable regions. (D) The degree of mismatching was measured by polyacrylamide gel electrophoresis (PAGE). The internal loops produce resistance, which gradually decreases the running speed of PAGE (PM; perfect match, 1MM; one-variable mismatch, 2MM; two-variable mismatch, 3MM; three-variable mismatch). All of the sequences were designed using EGNAS (Kick et al., 2012). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The training and test datasets.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td><strong>Friends (A)</strong></td>
</tr>
<tr>
<td>Training (Hypernetwork)</td>
<td>She's the only one (A1)</td>
</tr>
<tr>
<td></td>
<td>Put Joey on the phone (A2)</td>
</tr>
<tr>
<td></td>
<td>I went to the bathroom (A3)</td>
</tr>
<tr>
<td></td>
<td>I've been intimate with a man (A4)</td>
</tr>
<tr>
<td></td>
<td>The sex of the baby (A5)</td>
</tr>
<tr>
<td></td>
<td>She's the only woman (A6)</td>
</tr>
<tr>
<td>Test (Unknown)</td>
<td>His mom's on the phone (A7)</td>
</tr>
<tr>
<td></td>
<td>I'm going to the bathroom (A8)</td>
</tr>
<tr>
<td></td>
<td>Date with a man (A9)</td>
</tr>
<tr>
<td></td>
<td>Take care of the baby (A10)</td>
</tr>
</tbody>
</table>
with three variable regions (i.e., a trigram sentence). The number of example molecules in a hypernetwork provides the weight of the hyperedge. In short, a hypernetwork is a collection of hyperedges in a class. Fig. 1 shows a schematic diagram for the DNA hypernetwork model for learning and classification based on the symmetric internal loops. The DNA hyperedges encoding selected sentences from Friends were stored in microtubes that represent the gold hypernetwork ($H_A$). The DNA hyperedges encoding selected sentences from Prison Break were stored in microtubes that represent the red hypernetwork ($H_B$). The hypernetwork can be used to produce answers to questions regarding unknown test sentences that were also selected from the two television shows (in this example, “It’s the only way”). The DNA molecules encoding the question sentences were hybridized to all the DNA molecules in the $H_A$ and $H_B$, respectively.

To classify the question sentences, we used mismatched DNA sequences during hybridization. The DNA hybridization technique, typically employed in DNA computing, limits the design of DNA sequences to complementary base pairs (bp), where only perfectly matched sequences are relevant. An alternative approach involves using mismatched DNA sequences during hybridization, which introduces the use of symmetric internal loops into DNA computing. The mismatches can be employed to determine the distances between given instances, which is essential for recognizing similar patterns and classification of the molecules in a noisy environment. Symmetric internal loops were formed according to the similarities and differences in the nucleic acid sequences. Differences in the size of the internal loops were detected by electrophoresis. DNA that was perfectly hybridized moved faster than DNA with internal loops due to the added resistance of the latter; the larger the internal loops, the slower the DNA moved. In Fig. 1, lane 1 represents the hybridization between a question and $H_A$. Lane 2 represents the hybridization between question and $H_B$. DNA travelled further in lane 2 than in lane 1, indicating a greater degree of matching in lane 2. This pattern suggests that the test sentences probably originated from Prison Break. When the question sentences are classified, the molecular learning machine generalizes the unknown sentences (in this example, “It’s the only way”) and answers the question (“It seems to come from Prison Break”).

### 3.2. Design for wet-lab experiments

Fig. 2A shows the structure of the DNA hyperedges for this study. Each hyperedge has three variables, two inner tag regions and a primer region. The arrows indicate single stranded DNAs (ssDNAs) in the 5' to 3' direction. The blue arrows contain different sequences for encoding specific words (V1–V3). The black arrows contain two inner tag regions and a primer region. The sequences of the inner tag regions and the primers were the same in all hyperedges. The inner tag regions and a primer region were used to anchor the top strand to the bottom strand, thereby preventing the matching sequence regions from shifting. Through these regions, a clear boundary is drawn between the variables. They reduce errors that could otherwise have occurred during the molecular learning and pattern classification process within or near the variables. The top strand starts from the forward primer region, and the bottom strand starts from the backward primer region. If required, the primer regions can be used during the amplification process: the forward primer can attach to the F primer region on the bottom strand for the elongation of the top strand, and the backward primer can attach to the B primer region on the top strand for the elongation of the bottom strand. Specifically, Fig. 2B shows a DNA strand for a typical hyperedge structure of the phrase “I love babies.” Each variable region was encoded by nucleotides (nts) that represented each word. The inner tag regions were inserted between the variable regions with 8 nts. Additionally, primer regions with 20 nts were located on either side of the DNA strands.

Fig. 2C shows the predicted structures of matches as well as the structures of mismatches that formed the symmetric internal loops between the top and bottom strands. If the top and bottom strand form perfect double stranded DNA through the hybridization process, this indicates identical content (Fig. 2C-i). Except in this case, mismatched hybridizations between the top and bottom strands produced symmetric internal loops of DNA. Consider the following example: the molecular machine observes the sentence “I love babies” from Friends and then classifies the unknown input sentence “I love coffee,” which does not have a class label. The symmetric internal loops occurred between the variables “babies” and “coffee,” whereas both the side variable sequences “I” and “love” are matched perfectly (Fig. 2C-ii). If the sentence “She cuddles babies” exists in Prison Break, the internal loops are produced from “I love babies” between the variables “She”–“I” and “cuddles”–“love” (Fig. 2C-iii). Consequently, the input sentence can be classified by measuring partial mismatches using electrophoresis. When the input sentence is “We got married,” the internal loops are produced between the variables “We”–“I,” “got”–“love,” and “married”–“babies” (Fig. 2C-iv). The small partial mismatch sequences may contain a 4- or 6-bp mismatch in one or two variables, respectively (Fig. 2C-v). The DNA sequences used in this study were designed to avoid the formation of small partial mismatch structures, which in turn ensured the maximum possible distance between results and, thus, enabled a clear distinction between different variables (Kick et al., 2012) (Table S1 and Fig. S1).

The degree of mismatch was measured by PAGE. The internal loops produce resistance, which gradually decreases the running speed of PAGE for dsDNAs of the same size. Fig. 2D shows the results of PAGE for the examples from Fig. 2C. Perfect match (PM; lane 1), one-variable mismatch (1MM; lane 2), two-variable mismatch (2MM; lane 3), and three-variable mismatch (3MM; lane 4) were successfully discriminated by PAGE.

These can be categorized into four cases according to the number of symmetric internal loops created. In a hyperedge, the three variables consisted of 24 nucleotides; thus, 24 possible mismatches existed in the hybridization. Because the data features were encoded in the variable regions, 1–8 bp mismatches indicated 2-variable matches, 9–16 bp mismatches indicated 1-variable matches, and 17–24 bp mismatches indicated zero-variable matches; this was true regardless of whether or not the mismatches occurred in a contiguous variable region. For example, if a top strand hyperedge encoded by the variables vA, vB, and vC consecutively (i.e., “I love babies”) was hybridized with a bottom strand hyperedge encoded by vA, vB, and vD consecutively (i.e., “I love coffee”), then vA and vB (“I love”) would match perfectly, but vC and vD (“babies” and “coffee”) would not be hybridized. In this example, a 2-variable match would exist regardless of the order on the hyperedge. We used four types of internal loops for calculating the hamming distances of each match. As an example of these distance matches, the matches between the top strand “I love babies” and the bottom strands “I love babies,” “I love coffee,” “you love babies,” “she cuddles babies,” and “We got married” were zero, 1, 1, 2, and 3, respectively.

### 4. Experimental results

#### 4.1. Verification of learning algorithm

In this study, the hypernetwork is a model for training, learning, and classification. To explain this, 20 sentences, A1–A10 and B1–B10, were selected from the two television shows Friends and Prison Break, respectively (Table 1). A1–A5 were used for training
Fig. 3. Hypernetwork trainings were performed using the training data A1–A5 and B1–B5. (A) A1 and B1 were trained from empty tubes (HA1 and HB1), and A2 and B2 were tested using HAa and HBb. A2 and B2 were not correctly classified using untrained HAa and HBb. (B) After training [HAb and HBb], A2 and B2 were correctly classified. (C) A3 and B3 were not correctly classified using the untrained HAc and HAd. (D) After training [HAc and HAd], A3 and B3 were correctly classified. (E) A4 and B4 were not correctly classified using the untrained HAE and HAF. (F) After training [HAE and HAF], A4 and B4 were correctly classified. (G) A5 and B5 were not correctly classified using the untrained HAG and HAH. (H) After training [HAG and HAH], A5 and B5 were correctly classified. (I) The predicted structure of DNA for (G) and (H). Before training, the hybridization structure between B2-4 and A5-3 had fewer mismatches than the hybridization structure between A1-1 and A5-3. Therefore, the gel image shows that the hybridizations between HAg with A5 had fewer mismatched structures than HAa with A5 and the A5 test sequence was incorrectly classified. After training, A5 was correctly classified because HAg had perfect matches and was identifiable in the gel. All sequences were designed using EGNAS (Kick et al., 2012).
HA, and A6–A10 were used for classification as the unknown test data. B1–B5 were used for training HA, and B6–B10 were used for classification as the unknown test data. Each of the 20 sentences consisted of common words and each had at least one perfectly matched order-three hyperedge of its own class. However, these sentences did not exist as full sentences in the training data. The number of sentences used in this experiment was much smaller than the original corpus; however, the subset we used was sufficient to explain the molecular learning process and demonstrate the general properties of the learning algorithm. Each order-three hyperedge was encoded by DNA molecules (Table S1). For example, the sentence A1 (“she’s the only one”) consisted of two order-three hyperedges (“she’s the only” and “the only one”). Each hyperedge was encoded by the DNA sequences of A1–1 and A1–2.

Top strand DNA for A1–1 and A1–2 were prepared in microtube A and Top strand DNA for B1–1 and B1–2 were prepared in microtube B (Table S1). Accordingly, A1 was sets of training data from the target distribution A to train HA, B1 were sets of training data from the target distribution B to train HB. In other words, HA and HB were trained using A1 and B1 from empty tubes and became HA1 and HB1. The bottom strand DNA for A2 was then hybridized with the top strand of DNA for HA1 and HB1 sentences. This means that A2 were tested using HA1 and HB1. The hybridized samples were analyzed by 15% PAGE. The control sample was loaded onto the gel. We used a control sample containing a PM, 1MM, 2MM, and 3MM. The lowest band was used as a control for the classification of each sentence in each lane and provided the cross-section at which the band intensities were measured.

**Fig. 3A** shows the results of PAGE analysis. In the gel image, lane 1 represents the hybridized sample of HA1 + A2 (top strand: A1, bottom strand: A2). Lane 2 represents the hybridized HB1 + A2 sample (top strand: B1, bottom strand: A2). Lane 3 represents the control sample. The DNA band of lane 2 moved further down than band of lane 1, indicating a greater degree of matching in lane 2. However, this could be interpreted as an inaccurate result. Although we tested using A2, which came from HA, the result showed that A2 is more matched with HB. Therefore, A2 was not correctly classified using untrained HA1 and HB1. The bottom strand DNA for B2 was also tested with HA1 and HB1. In this case, B2 was correctly classified but the size of the band was incorrect. In addition to gel images, the band intensity around PMs (60 bp) was graphed. The band intensity was weaker than the control sample, because the size of the band
was wrong. These results mean that the classification of H_{A3} and H_{B1} was not correct.

H_{A} and H_{B} were trained through the addition of A2 and B2 sentences into H_{A1} and H_{B1}, respectively, and changed to H_{A2} and H_{B2}. The top strand of A2 and B2 were added into microtubes A and B, respectively. Then, the H_{A2} and H_{B2} were tested using sentence A2. The bottom strand of A2 were hybridized with the top strand of A1 and A2 in microtube A, and the bottom strand of A2 was hybridized with the top strand of B1 and B2 in microtube B. Fig. 3B shows the gel image and profile of band intensities. Lane 1 represents the hybridized sample of H_{A2} + A2 (top strand: A1 + A2, bottom strand: A2). In addition, lane 2 represents the hybridized sample of H_{B2} + A2 (top strand: B1 + B2, bottom strand: A2). Lane 3 represents the control sample. The band in lane 1 moved further down and the intensity was brighter than the control sample around PMs. This indicates that A2 was correctly classified. In addition, the B2 sentence was also tested with H_{A2} and H_{B2}. The gel image showed that the band for lane 2 for H_{B2} + B2 (top strand: B1 + B2, bottom strand: B2) moved further than lane 1 H_{A2} + B2 (top strand: A1 + A2, bottom strand: B2) and the intensity was brighter than the control sample around PMs. This indicates that B2 was correctly classified. These results demonstrate that the classification for A2 was correct after H_{A1} and H_{B1} were trained by A2 and learning was achieved. Identical results are shown for A3-A5 and B3-B5 in Fig. 3C-3H. Because H_{A} and H_{B} were trained through the addition of sentences, the classification results were correct compared to the results from untrained H_{A} and H_{B}.

Fig. 3I provides a detail illustration of the possible DNA hybridization structures of a sentence in the final training steps, when the A5 sentence was tested using H_{A4}, H_{B4}, H_{A5}, and H_{B5} in Fig. 3G and H. The A5 test sequence ("The sex of the baby," encoded by the three hyperedges A5-1, A5-2, and A5-3) was hybridized with H_{A4} and H_{B4}, which was not trained by A5. The A5 test sequence was incorrectly classified because the hybridizations in H_{B4} with A5 had fewer mismatches than H_{A4} with A5. After the top strand of A5 and B5 were added to H_{A4} and H_{A5}, respectively, as a training process, A5 was correctly classified by H_{A5} and H_{B5}. Because H_{A5} perfectly matched with A5, the band moved further down than the H_{B5} and A5 bands. These results indicate that our DNA molecule-based hypernetwork model can be trained.

4.2. Accuracy of classification

Based on these results, the classification accuracy was monitored as training of the hypernetwork was processed. The number of sentences increased in each hypernetwork (H_{A} and H_{B}) sequentially. H_{N1}, H_{N2}, H_{N3}, H_{N4}, and H_{N5} were prepared, which means H_{A1} and H_{B1}, H_{A2} and H_{B2}, H_{A3} and H_{B3}, H_{A4} and H_{B4}, and H_{A5} and H_{B5}, respectively. In other words, H_{N1} was trained by A1 and B1, H_{N2} was trained by A1–A2 and B1–B2, H_{N3} was trained by A1–A3 and B1–B3, H_{N4} was trained by A1–A4 and B1–B4, and H_{N5} was trained by A1–A5 and B1–B5. The classification test for each sentence (A1, A2, A3, A4, and A5) was performed in each training step. Fig. 4 shows the verification of training steps by classification of the Friends and Prison Break training examples (A and B, respectively). Fig. 4A shows the accuracy for the classification of A1–A5 using H_{N1–N5} based on the PAGE analysis (Fig. S2). Regardless of the low accuracy in the initial phases, the accuracy of the hypernetwork gradually increased to 100% when the H_{A5} and H_{B5} were used. The classification test for each sentence (B1–B5) was also performed in each training step (H_{N1–N5}). The classification accuracy increased and achieved 100% by the H_{N5} in Fig. 4B. After training, learning was carried out using two experimental processes: hybridization and electrophoresis. Generalization of the training examples was postponed until the training (top strand hyperedges) and test (unknown, bottom strand hyperedges) instances were hybridized. The classification experiment was performed in each training step using unknown test data (A6–A10) and the accuracy was analyzed (Fig. 4C). In addition, the same classification test was repeated using unknown test data (B6–B10) with H_{N1–N5} (Fig. 4D). As the training was processed and the number of sentences increased in each training step, the accuracy was increased and achieved 100% accuracy finally. We verified the robustness of trained information in a gradually increasing set of training data (Fig. S3). These results indicate that training of the hypernetwork provides high accuracy for classification.

4.3. In vitro classification

Fig. 5 shows the gel image for the classification test of sentences (A6–A10) with H_{N5}. The hybridized samples were loaded onto the gel in numerical order: for Fig. 5A (H_{A5} + A6, H_{B5} + A6,
$H_{A5} + A7, H_{B5} + A7, \ldots, H_{A5} + A10, H_{B5} + A10$ and for Fig. 5B ($H_{A5} + B6, H_{B5} + B6, H_{A5} + B7, H_{B5} + B7, \ldots, H_{A5} + B10, H_{B5} + B10$). Below the gel images, the horizontal band intensities around the PMs (60 bp) are graphed for ease of inspection. In Fig. 5A, the bands hybridized with $H_{A5}$ and A6–A10 moved further down the gel than the bands hybridized with $H_{B5}$ and A6–A10. Similarly, the highest intensities were observed in the lanes between $H_{A5}$ and A6–A10 (near PM). Thus, the results demonstrate that the correct classification was achieved. In Fig. 5B, the classification experiments of the test sentences B6–B10 followed the bands hybridized with $H_{A5}$ and B6–B10 moved further down the gel than the bands hybridized with $H_{A5}$ and B6–B10.

5. Discussion

Each sentence consisted of 2–4 hyperedges that were chosen in a sequential order from the full sentence. In wet-lab experiments, the training data contained at least one hyperedge that was identical to the unknown test sentence. Therefore, the molecular learner could classify the unknown test sentences with high accuracy. We designed the experiment in this way for the following reasons. First, the original set of sentences from each show contained a minimum of 10,000 sentences. If test sentences had been randomly selected from a dataset this large it would have been problematic for our molecular experiment because the probability of perfect or near perfect matching hybridizations would have been extremely low. Therefore, specific sentences with a common word between the two classes were deliberately selected. Second, our sentence selection criteria ensured that the classification could be sufficiently visualized via the size of the bands on the gel electrophoresis because the correctly classified sentence would always contain a 3-variable perfect match. If this match were absent, the similar sized bands that could arise from both classes would require further analysis (e.g., comparison of band intensity) in order to complete the classification.

We acknowledge that the small training dataset and the existence of a guaranteed 3-variable perfect match in the hybridization experiments, which ensured that the system did not have to produce an answer based on comparisons of fewer variable matches, each represent a limitation of our study. Nevertheless, the molecular learning that was conducted using our small training dataset is entirely valid because we demonstrate generalization of unknown sentences by existing hypernetworks and the classifications were 100% successful.

Designing the sequences of each sentence requires careful analysis of the possible DNA hybridization structures because the size of the internal loops that form at a mismatch variable can cause substantial variation in the distances travelled down the electrophoresis gel, independent of the number of internal loops formed. Therefore, analyzing sentence classification from the size of the bands alone can produce misleading results. To avoid these possible errors, we designed the DNA sequences of the mismatch regions to form maximally large internal loops (Fig. S1).

Although recent electrophoresis technology can identify a difference as low as one base pair, small partial mismatches are difficult to distinguish by electrophoresis owing to the 3D conformations of mismatched DNA. In our research, we focused on the distinction between perfect matched variables and partial mismatched variables to prove DNA-based learning process and applied them for in vitro classification. However, several challenges persist in solving this issue. An accurate detection method would be developed, such as distinction of DNA duplex based on the 3D structure or number of mismatched sequences/locations in DNA. In addition, DNA encoding method should be improved to make the English word distance equal to the DNA word distance at a character level. These would be useful to improve DNA-based learning algorithms and enable their application for more correct and sophisticated in vitro classification.

Additionally, it may be possible to employ fluorescent markers or next generation sequencing techniques to train and classify large amounts of data in future research. Accordingly, the full set of 10,000+ sentences from each television show could potentially be learned (Classification simulation of sentences from Friends or Prison Break with varying training set sizes, shown in Fig. S4 and S5). For example, microchip-synthesized oligonucleotides could be used to scale up the experiment because the training DNA sequences could be mixed together without separating each DNA sequence from the microchip. Longer variable regions would also allow encoding of more words (e.g., a 15-mer variable region could encode 415 = 1,073,741,824 words). Even if cross-hybridized sequences were removed, this region is still large enough to encode every word currently in existence.

6. Conclusion

Our major contribution is the development of a molecular machine learning algorithm and its realization in DNA computing, exploiting the symmetric internal loops, which we evaluated using a series of in vitro learning and classification experiments. We verified each molecular learning step and performed classification experiments using the test data taken from the dialogue sentences from real-life television shows. The molecular learning machine was able to generalize from example sentences and perform sentence classifications with 100% success. By exploiting the generality of machine learning and DNA computing, our molecular machine learning model in vitro can be used to solve molecular computing and other machine learning problems in computer science and molecular data analysis in biology if the data can be encoded in DNA molecules.

Acknowledgements

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project number SRFC-IT1401-12.

Appendix A. Experimental Details: Materials and Methods

The training experiments were performed in microtubes. Each DNA sequence was stored in nuclease-free sterile water and 1 pmol was used in each PAGE experiment. In the hybridization step, we confirmed the difference in running speed of the dsDNA hyperedges by their degree of mismatch (Fig. 2). We hybridized all the sequences in 1X SSC buffer for 1 min at 95 °C and gradually decreased the temperature 1 °C every minute until we reached 4 °C. A 15% polyacrylamide gel was used for PAGE and the DNA was run for about 3 h at 120 V. The gels were stained with SYBR® Safe DNA gel stain (1:10,000 dilution in 0.5X TBE) for 20 min, and then viewed using the Bio-Rad Gel Doc 2000.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biosystems.2017.04.005.

References


