Random Ensemble Hypernetwork for Pattern Recognition with Enzymatic Weight Update

Christina Baek[1], Dong-Hyun Kwak[1], Byoung-Tak Zhang[2]

[1]Interdisciplinary Program in Neuroscience, Seoul National University,

[2]Department of Computer Science & Engineering, Seoul National University

패턴인식을 위한 엔자이메틱 웨이트 업테이트 기반의 랜덤 앙상블 하이퍼네트워크

백다솜 [1], 곽동현 [1], 장병탁 [2]

[1]서울대학교 뇌과학 협동과정, [2]서울대학교 컴퓨터공학부

dsbaek@bi.snu.ac.kr

Abstract

Pattern recognition is a major division of machine learning which focuses on learning the patterns and regularities in data. It differs to that of pattern matching where only exact matches are found. However in the field of DNA computing, molecular pattern recognition has not been well established due to the lack of control of molecules in liquid state, instability and inaccuracy to solve such problems. Also the cost of designing the mass amount of DNA to represent data is a realistic issue. Here, we propose the random ensemble Hypernetwork as a model for pattern recognition *in* vitro with handwritten digit data encoded to DNA. For the manipulation of DNA *in vitro,* this molecular programming model is proposed to build a massively-parallel classification device, with the use of enzymatic weight update. Furthermore, a novel method of encoding vast amounts of data to DNA is introduced, also allowing the production of random hyperedges, a key difference to previous studies or computational implementations which only focused on fixed hyperedges.

1. Introduction

DNA (deoxyribonucleic acid) encodes the genetic information of cellular organisms. It consists of 2 strands, each with a chain of bases or nucleotides attached to a sugar-phosphate "backbone". With these features, DNA is now being used as the primary database as molecules in the field of DNA computing. The double-stranded DNA allows complementary base pairing which offers specificity in molecular recognition and self-assembly properties. Furthermore, 1μm of DNA contains about 1026 reactions which mean massively parallel reactions can take place in a minute volume of DNA sample, thus, a very large storage medium. These characteristics of DNA have given rise to new fields such as DNA origami and DNA nanotechnology [1]. Extensions of DNA sequencing technology with the ability to sequence specific DNA oligomers have even lead to programmable and autonomous computing machines [2, 3] for problem solving [4] and *in vivo* applications, taking advantage of DNA as a biologically compatible material [5].

However, the implementation of learning with DNA is still a goal pursued by many researchers. This is where the cross between machine learning and DNA computing research is focused. In this study, we propose a model for massively-parallel pattern recognition using enzymatic weight update in random ensemble Hypernetworks. The Hypernetwork is a learning algorithm based on evolutionary learning mechanisms. It is a graphical model with nodes and connections between nodes called hyperedges. The connections between these nodes are strengthened or weakened through the process of weight update or error correction during learning [6]. This model is fit for implementation to DNA computing for many reasons. As first inspired by the idea of *in vitro* evolution, it is a clear framework for molecular programming to be designed for artificial evolution in a test tube. The probability of hyperedges, or weights are represented by the concentration of DNA species in the tube. In addition to this, the idea of weight update is implemented here with specific enzymes. All these are processed in a massively-parallel manner.

As our dataset, we use the handwritten digit dataset, MNIST to encode to DNA using the random ensemble process to address the need for vast amounts of DNA to represent digit data, a key difference to previous studies which required the sequencing of two DNA oligomers for every digit pixel. Furthermore, a method of producing random order hyperedges are used to further reduce the cost of DNA sequences required and a method of producing all possible combinations of pixels in DNA through the advantages of

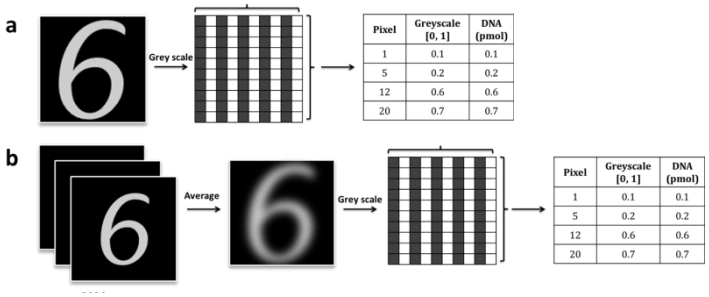

Figure 1: DNA encoding of pixel variables. a. Single image encoding b. Batch image encoding

DNA ligation techniques.

Two-class digit classification problem for digits '6' and '7' is solved by the molecular learning system. Here, each class is labelled with a different **florescent protein** to allow visualization of classes, a significant difference to previous studies as labelling allows learning to occur for two classes in one tube. The specifically designed DNA molecules are trained via the random ensemble Hypernetwork to build trained molecular Hypernetworks [6, 7].

The applications of implementing machine learning algorithms with molecules are still DNA computing technologies to develop towards in the future. However, with the achievement of learning in small scale experiments with classical machine learning problems is a critical starting point for future computational molecular devices, implicated in a diverse range of fields such as medical diagnostics and assembly of nanodevices.

2. Method and Results

MNIST handwritten digit data is dimensionally reduced to 10 × 10 images which are used as the input data. From these images, 6 sets of 100 images are averaged into 6 separate images for batch learning. 5 sets are used as the training data and one is used as the validation data. From each batch image, 25 pixels are randomly selected in a non-replacement manner, for each ensemble. Four ensembles are produced for each batch image. These 25 unique pixels are encoded into DNA by allocation to each unique DNA sequences consisting of 15 base pairs. Once each DNA oligomer is assigned to a pixel in the batch image, it is the grey scale value (between 0 and 1) for each pixel that determines the amount of DNA to be added. Each DNA oligomer is added at an amount corresponding to that pixel value. The initial Hypernetwork is made from equal amounts of each 25 DNA oligomers to ensure it is a random sample to be trained with training data. 8].

It is important to note here that DNA sequences designed for this purpose requires a high degree of uniqueness to avoid unwanted inter or intramolecular binding. Furthermore, the CG content must be controlled to consistent value for controlled hybridization and separation of DNA sequences during PCR re3ctions. An exhaustive DNA sequence design algorithm, EGNAS was used to design DNA sequences in a controlled manner [8].

Following the addition of relative amounts of DNA oligomers representing each pixel, these pixels, or variables are joined together to produce random order hyperedges (Figure 2.). In order to do this, first, the single-stranded DNA sequences are annealed by controlled temperature conditions in the PCR machine (Figure 2. a). There are three types of DNA sequences, the variable, forward and backward sequences. Each variable sequence consists of a sticky end sequence of 15 base pairs on the 3' end of the DNA sequence. The forward primer sequence contains of the forward primer, class label sequence and equivalent tag sequence all of 15 base pairs each from the 3' end to 5' end. The reverse primers contain the equivalent tag sequence and reverse primer at the 3' end. At end of the DNA sequence, Biotin is added for later use in strand separation. Furthermore, the training data is made with forward variables with Cy3 or Cy5 tag as class 6 and 7 labels respectively and validation data is made with forward primer without the fluorescent tag.

Free-order hyperedges, containing 2 to 8 variables (number of pixels) are designed to be used during the learning process and as the initial random Hypernetwork (Figure 2. c). This is critical in producing enough variable combinations to represent the possible hyperedges that can be produced. This addresses the issue of high cost and experimental time that would otherwise be required to produce various lengths of hyperedges with DNA. In order to produce a sample of random order hyperedges, the forward and reverse primers in their double-stranded form following annealing with complementary base pairs are added to a tube containing variable sequences. Enzymatic treatment with ligase and PCR conditions are controlled so that the optimal length of hyperedge produced is 4 –order (Figure 2. b). It is worth noting that it is not

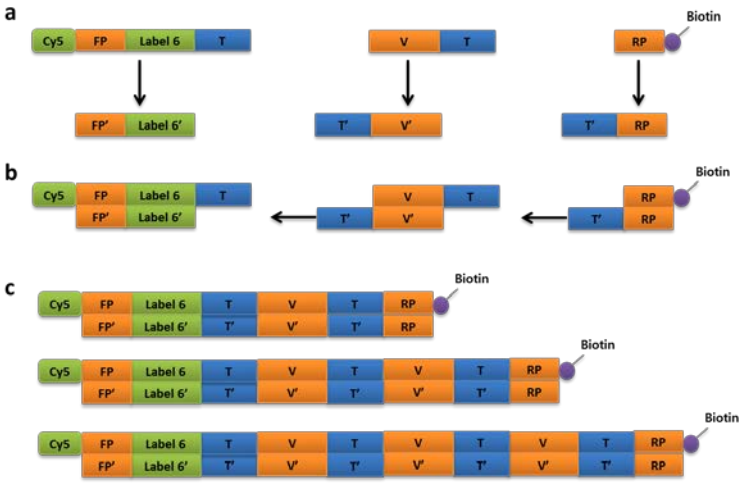

Figure 2: DNA sequence design for free-order hyperedge production a. Annealing of single-stranded DNA b. Ligation of sticky ended DNA c. Production of free-order hyperedges. (Cy3 = fluorescent tag, FP = Forward Primer, RP = Reverse Primer, T = Tag and V= Variable sequence

only the variables that are learned through the random ensemble Hypernetwork, but also the order of hyperedges as well.

After completing the encoding of image data to DNA, the molecular learning algorithm is implemented using an experimental protocol designed to replicate, *in vitro,* the random ensemble Hypernetwork using enzymatic weight update (Figure 3.). First, the preprocessing of data occurs with the hybridization of initial Hypernetwork with training data for digits 6 and 7. Next, T7 endonuclease is added to the sample for preprocessing and feature selection. This enzyme is able to cleave mismatched DNA sequences which form internal loops in the DNA structure, allowing selection of only the best matched DNA sequences. After enzyme use, DNA is purified using common DNA purification protocols and separated by biotin bead and streptavidin methods resulting in single-stranded DNA in the preprocessed Hypernetwork 0 (HN0) selected from the added training data. This is then amplified using PCR for the following procedure (Figure 3. a). The molecular learning step is implemented similarly with the hybridization of preprocessed Hypernetwork and training data for digits 6 and 7. This time the S1 nuclease is used for decreasing weight. The best matched DNA sequences are cleaved from the sample leaving only that which consist of internal loops. This discrepancy is represented by this pool of DNA sequences, as those sequences or hyperedges important for learning the digit '6' is divergent to those learning '7' with the use of labelled DNA sequences. Consequently, these sequences are amplified for increasing weight. The final sample from the first iteration of learning results in the Hypernetwork 1 (HN1). Repetition of these learning iterations will demonstrate learning of MNIST digits '6' and '7' for correct classification by measuring the ratio of fluorescence for Cy3 (Label for 6) and Cy5 (Label for 7).

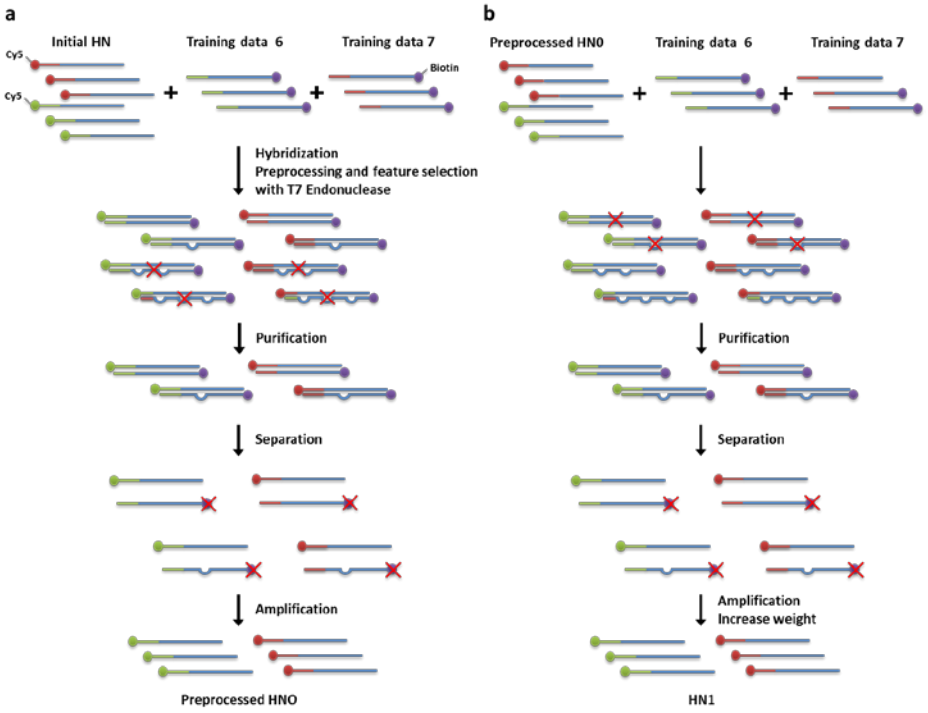

Figure 3: Hypernetwork molecular learning algorithm with enzymatic weight update. a. Preprocessing of MNIST image data b. Molecular learning with weight update protocol

3. Conclusion

Here, we propose a novel experimental design for implementing pattern recognition *in vitro* working side-by-side the random ensemble Hypernetwork algorithm. This learning model has been modified for suitability to realistic experimental conditions and costs. Although full iterations of learning must be experimentally carried out, the key significance of this paper is to provide a framework, with sound theoretical and experimental plausibility for the implementation of molecular learning. For future works, we will carry out the protocol and repeat 5 iterations, consisting of 4 ensembles each for the demonstration of this design in wet laboratory conditions.

Acknowledgment

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

References

[1] Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998)

[2] Benenson, Y., Paz-Elizur, T., Adar, R., Keinin, E. & Shapiro, E. Programmable and autonomous computing machine made of biomolecules, *Nature* **414**, 430-434 (2001)

[3] Stojanovic, M. N. & Stefanovic, D. A deoxyribozyme-based molecular automaton. *Nature Nanotech*. **21**, 1069–1074 (2003)

[4] Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994)

[5] Perrault, S. D., Shih, W. M, Virus-inspired membrane encapsulation of DNA nanostructures to achieve in vivo stability. *ACS Publications. **8**, 5132-5140 (2014)*

[6] Zhang, B.-T. Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE Comput. Intell. Mag*. **3**, 49-63 (2008)

[7] Lee, J.-H., Lee, B., Kim, J., Deaton, R. & Zhang, B.-T. A molecular evolutionary algorithm for learning hypernetworks on simulated DNA computers. *IEEE Cong. Evol. Comp. (CEC)*, 2735–2742 (2011)

[8] Kick, A., Bonsch, M. & Mertig, M., EGNAS: an exhaustive DNA sequence design algorithm. *BMC Bioinformatics.***13**, 138-155 (2012)