Massively-Parallel Pattern Recognition through Evolutionary Molecular Hypernetwork in DNA Computing

Christina Baek\textsuperscript{1}, Je-Hwan Ryu\textsuperscript{1}, Ji-Hoon Lee\textsuperscript{2}, Byoung-Tak Zhang\textsuperscript{3}

\textsuperscript{1}Interdisciplinary Program in Neuroscience, Seoul National University, \textsuperscript{2}Interdisciplinary Program in Bioinformatics, Seoul National University, \textsuperscript{3}Department of Computer Science & Engineering, Seoul National University

dsbaek@bi.snu.ac.kr

Abstract

Handwritten digit classification is a classical problem in field of machine learning or artificial intelligence. Here, we propose a DNA computing method to explore the intersection between biology and computation. We implement of the Molecular Evolutionary Hypernetwork \textit{in vitro} using DNA sequences as data. The pixels of the images are encoded into DNA sequences which are designed to create 3-order hyperedges to proceed as the training or test data to a random DNA library of 3-order hyperedges, ie. The Hypernetwork. This machine learning algorithm consists of the hybridization of training or test data to the Hypernetwork (random DNA library in the initial stage), selection and amplification of best matched sequences, isolation, classification, feedback and update of the Hypernetwork for the next learning phase. However, for this algorithm to be demonstrated experimentally, the creation of a 3-order hyperedge random library is crucial for the success of this learning process using a realistic sized dataset. This paper demonstrates the making of a random single-stranded DNA library through the use of ligation and isolation techniques, and validates the experimental steps required to implement the evolutionary molecular Hypernetwork in DNA Computing.

1. Introduction

DNA computing is a fast-developing interdisciplinary field. 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, the use of DNA offers many great advantages. Massively parallel search of all possibilities is possible, as 1μm of DNA contains about 1026 reactions. Furthermore, the molecular self-assembly property has given rise to fields such as DNA origami and DNA nanotechnology \cite{1}, specific molecular recognition and huge memory capacity in a minimal spaces are being exploited to create programmable and autonomous computing machines \cite{2, 3}, the ability to perform several operations has been used to solve various problems \cite{4} and as this biological material is naturally fit for \textit{in vitro} or \textit{in vivo} processing, applications in cells and animals models have also been demonstrated \cite{5}.

In this study, we propose a method of massively-parallel pattern recognition through molecular learning with the use of DNA operations. As evident with the described characteristics of using this biomolecule, this method allows us to manipulate DNA for solving the handwritten digit classification problem. Two digits, 6 and 7 will be recognized and classified into their correct classes. The pixels of these digits are encoded as specific DNA sequences, each strand representing a variable. The specifically designed DNA molecules are trained via an \textit{in vitro} evolutionary process to build molecular Hypernetworks \cite{6, 7}.

The conceptual process of solving the handwritten digit classification problem using DNA is illustrated in Figure 1. Dimensionally reduced MNIST handwritten digit data is used as input data. This is encoded into molecular DNA sequences. The encoded pixels in DNA are named as the nodes of the Hypernetwork. These pixels or variables are joined in threes, forming the 3-order hyperedge. This will be represented by a single-stranded DNA sequence.

3-order hyperedges, containing 3 variables (3 pixels) are designed to be used during the learning process and as the initial random Hypernetwork. However, to sequence all possibilities of random hyperedges would be impractical and costly. To address this problem, a ligation method is used to create the 3-order hyperedges randomly from individual variable components. Thus, a large collection of hyperedges are made which can be used for the learning process.

These experimental steps will be demonstrated through the
creation of a 3-order hyperedge random library which acts as the initial Hypernetwork, through the use of ligation and isolation techniques and through confirming that the library is successfully made.

2. Method and Results

The process of the molecular evolutionary Hypernetwork in vitro using DNA molecules require that each component is produced and manipulated in a controlled and orderly manner. For this reason, the preparation steps to create each component are a large and significant part of this study.

First, a technique is needed to design and produce the random, test and training data in the form of a 3-order hyperedge. Depending on the size of the digits, the number of pixels and thus the number of DNA sequences needed escalates exponentially. Ordering each DNA sequence individually would be impossible, and the experiments to follow, unrealistic. Thus, it is crucial that DNA sequences can be manipulated to create 3-order hyperedges experimentally as it is impossible to order all combinations of sequences corresponding to the data from the digits. For the purpose of this study, the techniques required to create and confirm such DNA sequences experimentally will be focused on.

Ligation of DNA sequences with specifically designed tags, ie sticky ends (red and orange in Figure 2) will be implemented to create the random, test and training hyperedges. It is necessary to take caution in designing the DNA sequences so that minimal matches occur with other regions of the DNA sequences. This was done using an exhaustive DNA sequence design algorithm, EGNAS [8].

First we strategized a protocol to produce these 3-order hyperedges start with the formation of double-stranded variable DNA (ordered DNA is single-stranded). Each double-stranded variable DNA, consists of the forward primer (FP), backward primer (BP) and relevant Tag sequence (T) depending on weather it is the first, second or third variable in the hyperedge (Figure 2a). Next, each variable can be joined together through the ligation of specific tags (ie Left, middle or right variable as shown in Figure 2d). Finally the double-stranded hyperedge can be stripped of its complementary strand to create single-stranded hyperedges (Figure 2c) using Streptavidin and Biotin beads attached to one end of the right variable sequence to produce a random library of Hypernetworks (Figure 3d).

The evolutionary Hypernetwork is constructed through a series of experimental steps. Training and testing of the classification process is performed during this process. The hybridization of training and random library in the initial step is required to create the first Hypernetwork, called Hypernetwork A or B (HNA or HNB) depending on which class of digits it correlates to. Next, selection and amplification of the best matched sequences is targeted through the Polymerase Chain Reaction (PCR) technique. These amplified double-stranded DNA is isolated using Biotin and Streptavidin and enter the classification test. Test data is added to the isolated single-stranded DNA. Through Gel electrophoresis the DNA produced the good results and the poor results are differentiated. The feedback to the random library proceeds, where the hyperedges which produced better results are added in greater amounts (double the amount) to the random library to create the HNA or HNB and the hyperedges which produced poor results are added in less amounts (half the amount). Further dilution during the full experimental process will ensure that the hyperedges which provided poor results are diluted out.

Figure 1. Conceptual process of handwritten classification through molecular machine learning.
Finally, the first cycle concludes and the next starts with the addition to new training data to the created HNA and HNB. This cycle is repeated and it is expected that each Hypernetwork will become specialized to its class and thus contain features, specific to each digit type, allowing more accurate digit classification as the molecular learning proceeds.

A series of preliminary experiments were undertaken to validate the experimental steps involved in demonstrating the evolutionary molecular Hypernetwork. First, the formation of a random double-stranded library was critical in verifying the success of the full experimental scheme. The experimental steps and results are as follows:

A. Hybridization of upper and bottom strands of variable units
B. Ligation of these variables in a random fashion, all in one tube to create a double-stranded DNA random library
C. Purification of the sample from ligase
D. Separation of the double-stranded to a single-stranded random library using Streptavidin and Biotin
E. Centrifugal filtering of DNA for concentration

Verification of library formation with the use of complementary strands (C9)

Each step listed above was carried out using wet DNA in a test-tube, and at each step, the DNA concentration was measured (Figure 3 and b). There is significant loss of DNA content however, a sufficient concentration was recovered from the centrifugal filtering step which allowed hybridization of the 9 complementary strands possible from the combination of the 3 variables used. Bands at the 70bp marker are present for all 9 types of sequences which confirm that all possible sequences were successfully constructed and retrieved during the experimental process.

The results shown in Figure 3 present DNA concentrations at various stages of the learning process, and the final verification of the success in making a random double-stranded library.

3. Conclusion

We propose the use of DNA and its manipulation through many experimental techniques to devise an in vitro representation of the evolutionary molecular Hypernetwork to perform handwritten digit classification. Through the development of ligation techniques to produce specific Hypernetworks freely, the overall process became plausible and more realistic to implement experimentally.

Acknowledgment

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

References