DNAzyme을 이용한 멀티모달 숫자패턴 학습

천효선 장병탁
서울대학교 컴퓨터공학부
{hschun, btzhang}@bi.snu.ac.kr

Multimodal Digit Pattern Learning using Deoxyribozymes

Hyo-Sun Chun Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

본 연구에서는 정보를 인코딩한 DNA 가닥을 이용하여 멀티모달 숫자 패턴을 학습하는 방법을 제시한다. DNA 분자는 많은 양의 정보를 안정적으로 저장할 수 있다는 장점으로 주목을 받아왔으며, 연상 메모리의 특징을 가진다. 본 연구에서 제안하는 학습 시스템은 확률적 연상 메모리 모델인 하이퍼네트워크 모델을 기반으로 구현되며, 하이퍼네트워크 구조를 DNA 가닥 상에서 표현하기 위해서 효소 기능을 하는 DNAzyme을 활용한다. 본 논문을 통해 DNAzyme 기반 논리 게이트는 감독학습을 위한 하이퍼네트워크 구조로 표현될 수 있음을 보이고, DNA 컴퓨팅 기법을 이용한 학습 과정의 구현을 설명한다. 또한 MNIST 필기체 데이터셋과 TIDIGITS 음성 데이터셋을 활용하여 멀티모달 숫자 패턴의 학습 과정을 컴퓨터 상에서 시뮬레이션하고, 이를 통해 분자 컴퓨팅적 구현의 실현 가능성을 탐색해본다.

1. 서 론

DNA(deoxyribonucleic acid)는 차세대 저장매체로 주목을 받아왔다[1]. DNA는 아데닌(adenine), 시토신(cytosine), 구아닌(guanine), 티민(thymine)으로 이루어진 염기서열에 유전정보를 저장하는데, 실리콘 기반의 컴퓨터에 비해 훨씬 더 많은 양의 정보를 안정적으로 저장할 수 있고[2], 연상 기억장치(content-addressable memory or associative memory)의 구축에 사용될 수 있다[3].

한편, DNA의 자가조립(self-assembly)적 성질과 화학반응에 내재된 초병렬적 연산능력은 DNA 컴퓨팅 분야의 탄생을 이끌었고[4], 생체 내에서 동작할 수 있는 나노머신 구현에 DNA가 활발히 활용되고 있다[5-8]. 뿐만 아니라 DNA 컴퓨팅이 인공신경망(artificial neural network), 유전 알고리즘(genetic algorithm), 베이지안 추론(Bayesian reasoning), 메시지 전달 추론(message passing inference), 선형 회귀(linear regression) 등 인공지능 분야에도 응용될 수 있음이 이론적으로 증명되어 왔으며[9-13], 다양한 패턴인식 문제에 적용되어 *in vitro* 환경에서 구현되었다[14-16].

본 논문에서는 분자 하이퍼네트워크 모델[17]을 기반으로 DNA 분자를 이용하여 멀티모달 숫자 패턴을 학습하는 방법을 제시한다. 2장에서는 하이퍼네트워크에 대해 알아보고, 3장에서는 하이퍼네트워크를 DNA 분자 상에서 표현하기 위해 DNAzyme(deoxyribozyme) AND 게이트[18] 구조를 사용하는 것을 보인다. 4장에서 MNIST 이미지[19]와 TIDIGITS 음성 코퍼스[20]를 사용한 컴퓨터 시뮬레이션을 통해 분자 컴퓨팅 학습 알고리즘의 성능을 검증한다.

2. 하이퍼네트워크 모델

사람의 뇌는 10^{11}개의 뉴런과 10^{14}~10^{15}개의 시냅스 연결을 가지고 있는 것으로 추정되는데[21], DNA 분자는 1 uM 농도의 용액 1 ml에 약 6×10^{14} 개의 분자가 존재하기 때문에 뇌에서 다루어지는 연산 규모에 비견될 수 있다[3,17]. 이러한 특징에 의거하여, DNA 분자를 이용한 대규모 연상 메모리 시스템 구현[22]과 메모리 기반 학습[23] 방법이 제시되어 왔다.

하이퍼네트워크(hypernetwork) 모델은 기존의 DNA 기반 분자 연상 메모리 모델에서 더 나아가, 뇌에서 이루어지는 희소 집단 코딩(sparse population coding)을 모델링하였다[24]. 하이퍼네트워크는 정점(vertex), 하이퍼에지(hyperedge), 가중치(weight)로 구성된다. 정점은 데이터의 세부적 특징(feature)을 직접적으로 표현하고, 하이퍼에지는 두 개 이상의 정점들을 연결하여 그들 간의 관계를 나타내며, 그 연관 강도는 하이퍼에지의 가중치를 통해 표현된다. 하이퍼네트워크는 멀티모달 데이터들을 연결하는 하이퍼에지를 정의함으로써 멀티모달 데이터 간의 관계를 표현할 수 있다[25].

3. DNA 하이퍼네트워크 구현 방법

3.1 DNAzyme 기반 AND 게이트

DNAzyme은 효소의 기능을 가진 DNA 가닥으로, RNA-cleaving 또는 RNA-ligating 기능을 하는 다양한 종류의 DNAzyme이 발명되었다[26]. 본 연구에서는 하이퍼네트워크 구조를 표현하기 위해 RNA-cleaving DNAzyme을 이용한다.

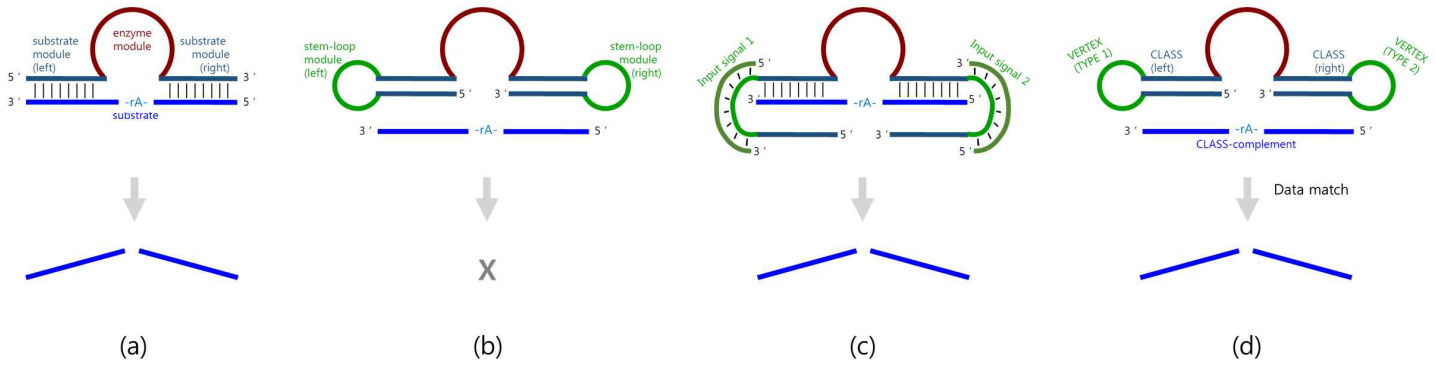

그림 1 **그림 1** (a) DNAzyme (b) DNAzyme 기반 AND 게이트 – 닫힌 상태
(c) DNAzyme 기반 AND 게이트 – 열린 상태 (d) DNAzyme 기반 AND 게이트를 활용한 하이퍼에지 구현 방법

일반적인 형태의 DNAzyme은 효소모듈(enzyme module)과 기질모듈(substrate module)로 구성된다(그림 1(a)). 효소모듈은 DNAzyme이 효소 기능을 수행하기 위해 필수적인 부분으로 염기서열이 보존되어야 한다. 기질모듈은 기질(substrate)과 결합하는 부분으로, 기질에 상보적인 가닥으로 구성되기 때문에 기질에 따라 변형될 수 있는 부분이다.

DNAzyme 기반 AND 게이트[18]는 효소모듈과 기질모듈에 더해서 헤어핀모듈(stem-loop module)을 가진다(그림 (b)). 헤어핀모듈은 평상시에는 닫힌 구조를 유지하기 때문에 기질모듈이 기질과 결합하지 못하지만, 헤어핀모듈에 상보적인 서열을 가진 분자가 존재할 때에는 열린 구조로 변환되어 기질이 결합할 수 있다 (그림 (c)). 이러한 구조의 DNAzyme은 양쪽의 헤어핀모듈에 해당하는 두 개의 신호가 모두 존재할 때에만 기질에 온전히 결합하여 효소의 기능을 수행하게 되므로 AND 게이트의 역할을 한다.

3.2 DNA 하이퍼네트워크 학습

DNAzyme 기반 AND 게이트는 감독학습을 위한 하이퍼네트워크 구조를 표현하는 데 사용될 수 있다(그림 (d)). 정점에 대한 정보를 헤어핀모듈에 인코딩하면, 정점을 표현하는 DNA 가닥이 존재할 때에만 헤어핀 구조가 열리게 된다. 따라서 클래스에 대한 정보를 기질모듈에 인코딩한다면, 클래스를 표현하는 DNA 가닥의 변형 유무를 통해서 정점에 대한 각 클래스의 확률을 예측할 수 있다. 이 때, DNAzyme 분자는 하이퍼에지가 되고, 그 농도가 하이퍼에지의 가중치가 된다.

하이퍼네트워크 감독학습은 예측(prediction), 선택(selection), 증폭(amplification)의 반복적인 실행에 의해 이루어진다[27-28]. 초기에 랜덤한 정보가 인코딩된 하이퍼에지 풀(pool)로부터 클래스를 올바로 예측한 하이퍼에지를 선택하여 그 가중치를 증폭시키는 방식으로 학습이 진행된다. DNAzyme을 활용한 구조 상에서 클래스의 예측은 형광소광법(fluorescence quenching), 선택과정은 바이오틴-스트렙타비딘(biotin-streptavidin)을 이용한 분리법, 증폭과정은 중합효소 연쇄 반응(polymerase chain reaction)을 통해서 실현될 수 있다.

4. 멀티모달 숫자 패턴 학습 및 결과

본 연구에서는 이미지와 음성으로 이루어진 데이터 셋을 구성하여, 컴퓨터 시뮬레이션을 통해 분자 컴퓨팅 알고리즘의 실현성을 알아보았다. MNIST 필기체 이미지 데이터셋[19]와 TIDIGITS 숫자 발화 음성 데이터셋[20]에서 6과 7을 나타내는 데이터를 학습시킨 후, 분류(classification)를 제대로 수행하는지를 확인해보았다.

멀티모달 패턴의 학습을 위해 DNAzyme 상에서 두 헤어핀 모듈 중 한 쪽은 이미지 정보를, 다른 쪽은 음성 정보를 인코딩하였다. 이미지 데이터는 그레이스케일의 이미지를 모노크롬 이미지로 변환시킨 후, 픽셀의 값과 위치를 DNA 염기서열로 인코딩 하였다. 음성 데이터의 경우 MFCC (Mel-frequency Cepstral Coefficients)[29] 벡터를 추출하여 k-means 클러스터링 통해 얻은 20개 클러스터에 대한 정보를 DNA 염기서열로 변환하였다. 각 MFCC 벡터는 유클리디안 거리(Euclidean distance)가 가장 가까운 클러스터의 염기서열로 맵핑하여 사용하였다.

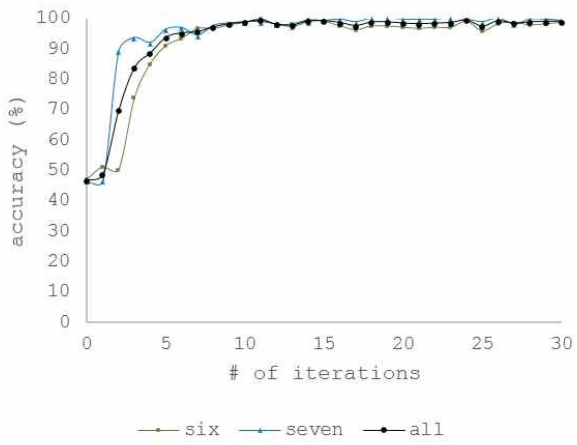

그림 2 멀티모달 숫자패턴 감독학습 시뮬레이션 결과

학습 과정에서 동일한 클래스에서 추출한 이미지 데이터와 음성 데이터를 이용해 예측, 선택, 증폭 과정을 실행하였고, 학습에 쓰이지 않은 독립적인 200개 테스트 데이터에 대해서 성능을 검증하였다. 그 결과 학습이 진행됨에 따라 성능이 점진적으로 증가하였고, 약 10번의 학습 사이클을 거친 후 정확도가 약 98%로 수렴하였다

(그림 2).

5. 결 론

본 논문에서는 DNAzyme을 이용하여 멀티모달 숫자 패턴을 학습하는 방법을 제시하였다. 학습을 위한 메모리 모델로 하이퍼네트워크 구조를 사용하였고, DNAzyme 기반의 AND 게이트 구조를 활용하여 분자 상에서 멀티모달 패턴을 학습하는 방법을 보였다. 또한 멀티모달 숫자 데이터를 이용한 컴퓨터 시뮬레이션을 통해 분자 학습의 실현가능성을 검증하였다.

DNAzyme은 바이오엔지니어링 분야에서 활발하게 사용되고 있어서[30], 본 논문에서 제시한 DNAzyme 기반 학습을 *in vitro* 환경에서 구현한다면 개인화된 질병 진단과 치료에 응용될 수 있을 것으로 기대된다.

감사의 글

이 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임(SRFCIT1401-12).

참고문헌

[1] G. M. Church et al., “Next-generation digital information storage in DNA,” *Science 337 (6102)*, 1628-1628, 2012.

[2] N. Goldman et al., “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA,” *Nature 494 (7435)*, 77-80, 2013.

[3] E. B. Baum, “Building an associative memory vastly larger than the brain,” *Science 268 (5210)*, 583-585, 1995.

[4] L. M. Adleman, “Molecular computation of solutions to combinatorial problems,” *Science 266 (5187)*, 1021-1024, 1994.

[5] B. Yurke, et al., "A DNA-fuelled molecular machine made of DNA," *Nature 406 (6796)*, 605-608, 2000.

[6] E. Mokany et al., “MNAzymes, a versatile new class of nucleic acid enzymes that can function as biosensors and molecular switches,” *Journal of the American Chemical Society 132 (3)*, 1051-1059, 2009.

[7] S. F. J. Wickham et al., “A DNA-based molecular motor that can navigate a network of tracks,” *Nature Nanotechnology 7 (3)*, 169-173, 2012.

[8] S. M. Douglas et al., “A logic-gated nanorobot for targeted transport of molecular payloads,” *Science 335 (6070)*, 831-834, 2012.

[9] A. P. Mills et al., “Experimental aspects ofDNA neural network computation,” *Soft Computing 5 (1)*, 10-18, 2001.

[10] P. Wasiewicz and J. J. Mulawka, “Molecular genetic programming,” *Soft Computing 5 (2)*, 106-113, 2001.

[11] I. S. de Murieta and A. R. Paton, “Probabilistic reasoning with a Bayesian DNA device based on strand displacement,” *Lecture Notes in Computer Science 7433 (DNA18),* 110-122, 2012.

[12] N. Napp and R. P. Adams, “Message passing inference with chemical reaction networks,” *Proceedings of Advances in Neural Information Processing Systems (NIPS 2013),* 2013.

[13] M. R. Lakin and D. Stefanovic, “Supervised learning in an adaptive DNA strand displacement circuit,” *Lecture Notes in Computer Science 9211 (DNA21)*, 154-167, 2015.

[14] H.-W. Lim et al., “In vitro molecular patternclassification via DNA-based weighted- sum operation,” *BioSystems 100 (1)*, 1-7, 2010.

[15] L. Qian et al., “Neural network computation with DNA strand displacement cascades,” *Nature 475 (7356)*, 368-372, 2011.

[16] J.-H. Lee et al., “A DNA assembly model ofsentence generation,” *BioSystems 106 (1)*, 51-56, 2011

[17] B.-T. Zhang, “Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory,” *Computational Intelligence Magazine 3 (3)*, 49-63, 2008.

[18] M. N. Stojanovic et al., “Deoxyribozyme-based logic gates,” *Journal of the American Chemical Society 124 (14),* 3555-3561, 2002.

[19] Y. LeCun et al., “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE 86 (11)*, 2278-2324, 1998.

[20] R. G. Leonard and G. Doddington, TIDIGITS LDC93S10. *Linguistic Data Consortium*, 1993.

[21] S. Herculano-Houzel, “The human brain in numbers: a linearly scaled-up primate brain,” *Frontiers in Human Neuroscience 3*, 1-11, 2009.

[22] J. H. Rief et al., “Experimental construction of very large scale DNA databases with associative search capability,” *Lecture Notes in Computer Sci- ence 2340 (DNA7)*, 231-247, 2001.

[23] J. S. Lee et al., “A DNA-based pattern classi- fier with in vitro learning and associative recall for genomic characterization and biosensing without explicit sequence knowledge,” *Journal of Biological Engineering 8 (1)*, 1-12, 2014.

[24] B.-T. Zhang et al., “Sparse population code models of word learning in concept drift,” *Proceedings of Annual Meeting of the Cognitive Sciene Society (CogSci 2012)*. 1221-1226, 2012.

[25] J.-W. Ha et al., “Text-to-image retrieval based on incremental association via multimodal hypernetworks,” *Proceedins of IEEE International Conference on Systems, Man and Cybernetics (IEEE SMC 2012),* 3245-3250, 2012.

[26] R. R. Breaker, “DNA enzymes,” *Nature Biotechnology 15 (5)*, 427-431, 1997.

[27] B.-T. Zhang, and H.-Y. Jang, “A Bayesian algorithm for In vitro molecular evolution of pattern classifiers,” *Lecture Notes in Computer Science 3384 (DNA10)*, 458-467, 2004.

[28] B.-T. Zhang, and J.-K. Kim, “DNA Hypernetworks for Information Storage and Retrieval,” *Lecture Notes in Computer Science 4287 (DNA12)*, 298-307, 2006.

[29] D. Jurafsky and J. H. Martin, “Speech and language processing,” *Pearson Education*, 2000.

[30] Y. Lu and L. Juewen, “Functional DNA nanotechnology: emerging applications of DNAzymes and aptamers,” *Current opinion in Biotechnology 17 (6)*, 580-588, 2006.