# DNA 기반 드라마 문장 분자패턴인식 시뮬레이션

이지훈 O1 천효선 2 장병탁 1234
서울대학교 생물정보학 협동과정[1]
서울대학교 컴퓨터공학과[2]
서울대학교 인지과학 협동과정[3]
서울대학교 뇌과학 협동과정 [4]
{jhlee, hschun, btzhang}@bi.snu.ac.kr

# Simulation for Molecular Pattern Classification of Drama Sentences based on DNA Molecules

Ji-Hoon Lee O1 Hyo-Sun Chun[2] Byoung-Tak Zhang[1234]
[1]Graduate Program in Bioinformatics, [2]Computer Science and Engineering
[3]Cognitive Science Program, and [4]Brain Science Program, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea

**요 약**

분자컴퓨팅은 DNA, RNA와 같은 생체분자 물질을 사용하여 정보처리를 수행하는 것을 말한다. 1994년 Adleman의 DNA 분자를 사용한 TSP 문제 풀이는 분자물질이 현재의 컴퓨터와 같이 사용될 수 있다는 능력을 보여주었다. 최근 분자컴퓨팅 학계에서는 분자 나노 구조물 혹은 DNA Tile과 같은 분자 구조를 만드는 연구가 주류를 이루고 있으며, 정보처리적 관점 연구에서는 분자 논리소자, 분자 유한 상태 기계와 같은 기초적인 연산 기능이나, 기초 컴퓨팅 모델연구가 많이 진행되고 있다. 그러나 이런 연구들은 주로 분자적 특성을 살려 기능적 도약을 추구하기 보다는 기존 컴퓨터의 연산 구조를 비슷하게 모사하는 수준에서 진행되고 있는 실정이다. 본 연구에서는 DNA의 분자적인 특성을 살려 논리회로와 같은 기초 연산 소자를 사용하지 않고도 기계학습과 같은 고차원적인 정보처리를 분자컴퓨터가 성공적으로 수행할 수 있음을 컴퓨터시뮬레이션 연구를 통해 보여준다. 드라마 코퍼스를 학습 데이터로 사용한 분자학습 시뮬레이션 결과, 기초 논리소자를 사용하지 않고도 간단한 분자학습 알고리즘으로 문장과 같은 언어데이터가 학습되며 문장 패턴인식과 같은 상위 레벨의 정보처리를 수행할 수 있음을 보여주었다. 본 시뮬레이션 연구를 통해 논문에서 제시된 분자학습알고리즘이 실제 DNA 분자물질을 사용한 분자생물학 실험을 통해 성공적으로 수행될 경우 나올 수 있는 결과를 미리 예측할 수 있었으며, 실험과정에서 생길 수 있는 문제점들을 미리 파악하고 대비할 수 있게 되었다.

## 1. 서론

1994년 Adleman의 분자 실험을 통해 시작된 분자컴퓨팅 연구는 현재들어 DNA 나노구조, DNA Tile, 분자 논리소자, 나노 로봇, 분자 메모리 등 다양한 응용 분야를 창출하고 있다 [1] [2] [3] [4] [5] [6]. 분자컴퓨터는 기존 실리콘 컴퓨터와는 다르게 DNA, RNA, 제한효소 와 같은 생체분자물질을 사용하여 컴퓨팅을 수행한다. 초기에 DNA 컴퓨팅 연구자들이 외판원 문제(TSP) 같은 NP-hard나 충족가능성 문제(SAT) 같은 NP-complete 문제를 많이 다루었다면 현재는 보다 응용과 기초 분자 논리 소자와 같은 것을 만드는데 더욱 노력을 하고 있다 [7] [8] [9].

최근에는 기계학습과 같은 보다 고차원적인 컴퓨팅을 분자 수준에서 수행하기 위하여 관련 연구들이 진행되고 있다. 하지만 아직 분자가 학습되는 과정을 생체분자를 사용하여 연구된 경우는 발표되지 않았고, 분자 논리소자를 사용하여 분류기를 만드는 연구는 몇몇 연구자들을 통해 진행되었다 [10] [11]. 분자하이퍼네트워크 학습 알고리즘은 DNA와 같은 생체분자를 사용하여 분자수준에서 기계학습을 하기에 용이한 학습 알고리즘이다 [12] [13]. 하이퍼네트워크 그래프 구조는 언어와 같은 인지적인 문제에 적용하기에 적합하여 분자 문장 생성, 분자학습 알고리즘 연구, 분자 애너그램 풀이 등에 적용 되었다 [14] [15] [16].

본 논문에서는 기존의 DNA 분자기계학습 알고리즘의 [15] 분자실험 단계를 간소화 하여 학습과정의 분자적 적용 가능성을 높인 학습 알고리즘을 제안 하고 드라마 코퍼스를 학습 데이터로 사용하여 시뮬레이션한 결과를 소개한다.

## 2. 분자학습 알고리즘

본 연구에서는 기존에 제안된 분자학습알고리즘 [15] 과는 다르게 랜덤 DNA 시퀀스를 사용하지 않았고 PCR 과정을 생략하여 분자생물학 실험 과정을 간소화 하였다. 제안된 분자학습 알고리즘은 다음과 같다.

〈훈련 알고리즘〉
모든 입력 데이터 (x, f(x))를 클래스에 따라 마이크로 튜브 A, B에 넣는다. 이것이 하이퍼네트워크 A ($H_A$), B ($H_B$)가 된다.
〈분류 알고리즘〉
분류 테스트 샘플 $x_q$ 이 주어지면,
1. $x_q$를 하이퍼네트워크 A와 하이퍼네트워크 B에 결합 (hybridization) 한다.
2. 결합된 dsDNA 분자들을 전기영동을 통해 결합된 구조에 따라 분류한다.
3. $x_q$ 가 결합된 하이퍼네트워크 A, B에서 보다 Perfect 결합이 많은 것을 조사하여 가장 빈도가 높은 클래스로 분류한다.

## 3. 실험결과

드라마 자막 코퍼스 (Friends, Prison Break) 에서 학습 문장을 10개에서 10000개 까지 증가시키며 학습 곡선을 그려 보았다(그림 1). 테스트 데이터는 학습에 참여하지 않은 100개의 문장을 학습 단계별 테스트에 사용하였다. 학습에 참여하는 문장의 크기가 증가할 수록 분별력이 좋아지는 결과를 확인할 수 있었다.

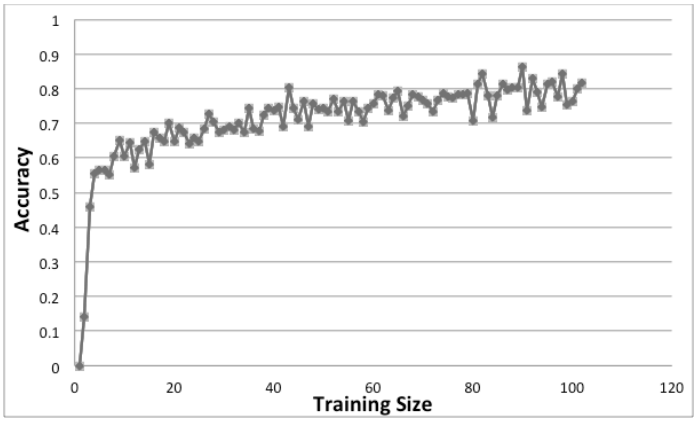


**그림 1 학습곡선**

## 4. 결론

본 시뮬레이션 연구를 통해 실제 DNA 분자를 사용해 분자기계학습을 수행하였을 때의 결과를 분자 실험 없이 예측할 수 있었다.
이를 통해 본 연구에서 사용된 간소화된 분자학습 알고리즘을 기존에 연구된 하이퍼네트워크 DNA 분자구조의 학습에 사용할 수 있음을 확인 하였다 [15]. 학습에 참여하지 않은 테스트 데이터 분류에 대한 성능 증가는 제안된 학습 알고리즘이 일반화 능력을 가지고 있음을 보여준다.

## 참고문헌

[1] Adleman, L. Molecular computation of solutions to combinatorial problems. *Science*, 266, 1021–4 (1994).
[2] Lipton, R. J. DNA solution of hard computational problems. *Science*, 268, 542–5 (1995).
[3] Braich, R. S., Chelyapov, N., Johnson, C., Rothemund, P. W. & Adleman, L. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, 296, 499–502 (2002).
[4] Mao, C., LaBean, T. H., Relf, J. H. & Seeman, N. C. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature*, 407, 493–6 (2000).
[5] Amir, Y., Ben-Ishay, E., Levner, D., Ittah, S. Abu-Horowitz, A. & Bachelet, I. Universal computing by DNA origami robots in a living animal. *Nature nanotechnology*, 9, 353–357 (2014).
[6] Chen, J., Deaton, R. & Wang, Y.-Z. A DNA-based memory with in vitro learning and associative recall. *Natural Computing* 4, 83101 (2005).
[7] Stojanovic, M. N., Mitchell, T. E. & Stefanovic, D. Deoxyribozyme-based logic gates. *Journal of the American Chemical Society* 124, 3555–61 (2002).
[8] Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. *Science* 314, 1585–8 (2006).
[9] Pei, R., Matamoros, E., Liu, M., Stefanovic, D. & Stojanovic, M. N. Training a molecular automaton to play a game. *Nature nanotechnology* 5, 773–7 (2010).
[10] Lim, H.-W., Lee, S. H., Yang, K., Lee, J.-Y., Yoo, S. Park, T.-H. & Zhang, B.-T. In vitro molecular pattern classification via DNA-based weighted-sum operation. *Biosystems* 100, 17 (2010).
[11] Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* 475, 368–72 (2011).
[12] Kim, J.-K. & Zhang, B.-T. Evolving hypernetworks for pattern classification. *IEEE Congress on Evolutionary Computation (CEC 2007)*, 1856–1862 (2007).
[13] Zhang, B.-T. Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory. *IEEE Computational Intelligence Magazine* 3, 4963 (2008).
[14] Lee, J.-H., Lee, S. H., Chung, W.-H., Lee, E. S., Park, T.-H., Deaton, R. & Zhang, B.-T. A DNA assembly model of sentence generation. *BioSystems* 106, 51–6 (2011).
[15] Lee, J.-H., Lee, B., Kim, J., Deaton, R. & Zhang, B.-T. A molecular evolutionary algorithm for learning hypernetworks on simulated DNA computers. *IEEE Congress on Evolutionary Computation (CEC 2011)*, 2735–2742 (2011).
[16] Lee, J.-H., Lee, E. S., Ryu, J.-H., Chun, H.-S. & Zhang, B.-T. Molecular computational simulation of cognitive processes for anagram solving, *International Conference on DNA Computing and Molecular Programming (DNA 19)*, 40 (2013).

## 감사의글

이 논문은 미공군연구소의 지원(FA2386-12-1-4087)과 한국연구재단의 지원(NRF-2013M3B5A2035921)을 받아 수행된 연구이다.