DNA 연산을 이용한 기억 인출 시뮬레이션

김준식 [0,1,2] 이은석 [2,3] 노영균 [2,3] 장병탁 [2,3]
[1]서울대 물리학과
[2]서울대 컴퓨터 공학과 바이오 지능 연구실
[3]서울대 인지 과학 협동과정
{jskim, eslee, yknoh, btzhang}@bi.snu.ac.kr

Memory retrieval with a DNA computing

Joon Shik Kim, Eun Seok Lee, Yung-Kyun Noh, Byoung-Tak Zhang
School of Physics, Seoul National University
Biointelligence laboratory, School of Computer Science and Engineering, Seoul National University
Interdisciplinary Program in Cognitive Science, Seoul National University

요약

본 연구는 특정 사물을 계속 접하면서 그 사물에 대한 기억 강도가 의식적 노력 없이도 점점 강화되는 암묵적 기억 인출과정associative memory retrieval의 DNA 연산 가능성을 논한다. 예를 들어 한 표적 단어에 대한 노출이 이를 관찰하는 시스템에게 그 단어의 기억 강도를 강화시키는 반면, 그와 유사한 다른 단어는 천천히 감소되고 나머지 가장 다른 단어는 일찍 잊혀지는 현상을 생각할 수 있다. 이들 단어들과 알파벳 철자들을 DNA 염기서열로 표현하고 simulated annealing을 통하여 결합 결과를 얻는다. Ridge regression 형태의 supervised 학습을 통하여 한 가지 표적 단어가 많이 생성되도록 DNA 조각들의 개수 분포를 변화시켜 진행한다. 실험 예로 'tic' 'tac' 'toe' 세 가지 단어를 그 아이템으로 정하여 계속 자극받는 표적 단어의 갯수가 증가함을 DNA annealing 시뮬레이션을 통하여 확인할 수 있다. 또한 'tac' 과 't' 와 'c'를 공유하는 'tic' 의 감소 정도가 't'만을 공유하는 'toe' 보다 느림을 확인할 수 있다. 위의 실험들을 통해 연합기억associative memory의 암묵적 인출과정을 분자 층위에서 표현할 수 있음을 확인할 수 있다.

1. 서론

DNA를 소자로 하여 전산학적으로 의미있는 문제를 푼 이후로 [1] DNA 연산은 꾸준히 발전되어왔다. 인공지능의 기본이 되는 연산 과정을 분자 층위에서 구현하려는 시도 [3,4] 들도 있어왔으며 이는 DNA 분자의 많은 갯수와 그의 초병렬성을 이용하는 방법을 모색하는 연구들이다. 또한 DNA와 RNA를 사용하여 홉필드망 같은 뉴럴 네트워크를 구연하려는 시도 [5] 또한 생물계의 집합적 정보처리 현상을 이용하여 인간에게 유용한 계산을 하려는 노력으로 볼 수 있다.

기억의 인출은 크게 두 가지 방식을 통해 이루어진다. 기억인출을 위한 신호가 주어졌을 때, 인간은 먼저 장기기억을 탐색하여 그 안에 목표 아이템이 있을 경우 이를 그대로 인출하는 방식을 사용한다. 이를 intact memory retrieval, perfect-match recall이라 부를 수 있다. 또 다른 방식은 위의 방식으로 인출에 어려움을 겪을 때 자동적으로 이루어지는 것으로, 부분적인 정보partial information만을 이용하여 신호와 비슷한 아이템을 연합하여 형성하는 방식이다. 이를 associative memory retrieval, best-match recall이라 부를 수 있다.

Associative memory retrieval은 그 심리학적 고유성으로 인해 인지과학, 심리학, 인공지능 영역에서 주목을 받아왔다. 이 현상의 특성을 보면 크게 다음과 같다. 첫째, 아주 작은 부분적 정보만으로 전체 아이템을 재인, 회상해낼 수 있다. 둘째, 패턴 자체가 기억에 대한 주소로 사용될 수 있다. 세째, 자극이 주어졌을 때 암묵적으로 (implicitly) 의식적인 노력 없이 그 자극의 부분 정보들로부터 그 자극과는 전혀 다른 (혹은 달라 보이는) 아이템들이 자동적으로 회상된다. 이러한 특성들은 인간의 지능이 관여하는 거의 모든 영역 (언어, 음악, 시각 패턴 인식 등)에서 발견된다.

그러면 분자 연산의 초병렬성을 이용하여 암묵적인 연합기억 인출을 연구할 수는 없을까? 이에 대한 답으로 알파벳의 순서를 다시 정렬하여 의미있는 단어를 찾는 Anagram이라는 놀이를 DNA 연산을 통해 구현하려는 시도 [6] 가 있었다. 역시 인간의 인지 기능 중 하나인 기억을 DNA의 생화학적 반응을 이용하여 실험으로 구현한 시도 [7]도 있었다. 이는 주소 address 중심이 아닌 내용content 중심의 기억 인출을 시도했다는 데 그 중요성이 있다. 본 논문에서는 심리학 실험을 바탕으로 한 기억의 공고화 과정을 DNA 연산으로 시뮬레이션 하여 본다. DNA에 magnetic bead를 붙여서 사용하면

특정 sequence를 가진 DNA 조각을 끌어낼 수 있고 이를 이용하여 기억을 시뮬레이션한다. 온도를 낮출 때 DNA 조각들이 결합하는 시뮬레이터는 [10] simulated annealing을 이용하여 만들었다.

표적 단어 또는 그와 비슷한 단어 또는 아주 다른 단어 등을 미리 짧은 시간 보여주고 다시 그 세 단어중 하나를 보여주어 그 전에 제시된 단어가 맞는지를 묻는 심리 실험 [8]을 상정할 수 있다. 처음 단어를 보여주는 시간이 길어질수록 표적 단어를 잘 맞추게 되며 유사한 단어를 보여주었을 때가 아주 다른 형태의 단어를 제시했을 때보다 더 잘 맞추게 된다. [9]에서는 이 실험을 probabilistic information transmission이라는 hidden Markov model로 설명하고 있다. 이는 확률 모델이며 간단한 조건 확률 강화 규칙으로 표현된다. 우리는 확률 모델이 아닌 많은 수 분자의 초병렬성을 이용하여 보다 무의식적인 인지 기능을 설명하려 한다. 분자로는 단어와 알파벳을 표현하는 DNA 조각들을 사용하고 wet 실험이 아닌 Monte Carlo simulation [10] 을 통하여 실험한다. 'tic' 'tac' 'toe'의 세 단어에 대해 실험해보며 이들을 표상하는 DNA 조각의 개수 분포 변화 규칙으로는 Ridge regression 과 형태가 유사한 supervised learning 을 이용한다.

2. 실험 방법

표 1 과 같이 세 단어와 알파벳을 인코딩 한다.

단어와 알파벳		DNA 표현
0	tic	AAAGGG
1	tac	AATCGG
2	toe	AAACTG
3	t	TT
4	i	CT
5	a	GA
6	o	GT
7	c	CC
8	e	TG

표 1. 세 단어와 알파벳의 DNA 인코딩.

Annealing은 95 도에서 10 도까지 천천히 내리는 방법을 따랐다. 초기 DNA 갯수는 't'를 위해서는 120 개를, 'c'를 위해서는 80 개를, 나머지들을 위해서는 각각 40 개를 준비했다. 구체적인 알고리즘은 아래와 같다.

a. 표적 단어를 정한다. 그리고 표적 단어에 해당하는 표적 분포인 벡터 A를 정한다. 예를 들자면 표적 단어 'tic' 의 표적 분포 A 벡터는 표 1 을 참조하여 $(120,0,0,120,120,0,0,120,0)^T$ 로 나타낼 수 있다. 이것은 'tic' template인 첫번째 원소와 't', 'i' 'c' 를 나타내는 염기서열들만 값을 가지도록 나타낸 것이다. 갯수가 120 인 이유는 초기 DNA 조각의 총 갯수가 480 이라서 이를 사등분할 필요가 있었기 때문이다.

b. Simulated annealing을 통해 DNA 조각들을 결합시키고 그 관계를 살핀다. (i,j) 원소가 표 1 기준으로 i번째 조각과 j번째 조각이 결합된 덩어리의 개수를 나타내는 9X9 행렬을 T 를 구한다.

DNA 조각의 갯수를 변경할 보정 벡터 C를 다음과 같이 Ridge regression을 닮은공식으로 구한다.

$$C=[T+I]^{-1}A. \qquad (1)$$

위에서 I는 identity matrix를 나타내며 역행렬이 singular하는 것을 피하는 역할을 한다. 보정벡터 C 는 DNA 조각들의 similarity (이 실험의 경우 hybridization 되는 정도)를 고려하여 구한 값이다.

c. 과정 b에서 구한 보정 벡터 C를 DNA 갯수 분포 벡터에 더하고 DNA 조각들의 총 수가 이전과 같게 되도록 비례적으로 갯수를 조정한다. 보정값을 더했을때 갯수값이 음수가 되면 이를 0 으로 놓는다.

d. 위의 a, b, c 를 순서대로 여섯 번 반복하여 실험하고 세 가지 단어가 각각 표적 단어일 때의 combined character 들의 개수의 평균을 본다.

3. 실험 결과

단어와 결합하는 알파벳 문자들의 갯수를 평균하여 그래프로 그려본다. 예를 들어 'tic' 의 대표값은 template 'tic'과 결합된 't' 'i' 'c' 갯수의 합을 으로 나눈 값이다. 이제 우리는 'tic' 'tac' 'toe'각각이 표적 단어인 경우의 알파벳 갯수 평균값의 변화를 iteration 순서에 따라 살펴본다.

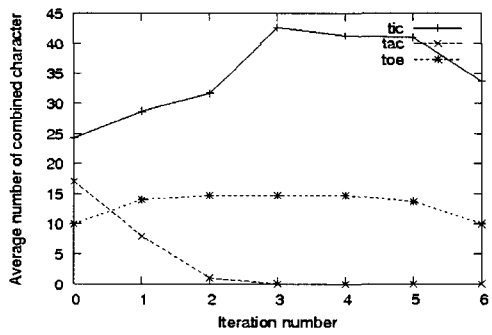

그림 1. 표적 단어가 'tic' 일 때의 단어들과 결합하는 alphabet 문자 DNA 조각들의 갯수 평균.

먼저 그림 1 은 'tic' 이 표적 단어일 경우의 결과이다. iteration 횟에 따라 'tic'과 결합된 't' 'i' 'c' 의 갯수 평균은 증가하고 나머지 단어들의 값들은 감소함을 알 수 있다.

아래의 그림 2 에서는 'tac'의 값이 증가함을 볼 수 있고 'toe'보다는 'tic'의 값이 더 서서히 감소함을 알 수 있다. 이는 'tic' 이 'toe'보다 'tac' 과 더 닮아서 그렇다고 볼 수 있다. 즉 'tic' 은 'tac' 과 't' 'c' 두 개의 철자를 공유하며 'toe'는 't' 한 개만 공유한다.

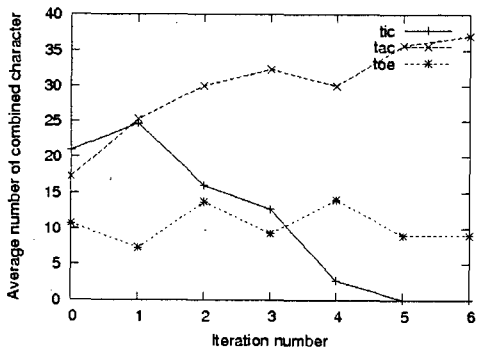

그림 2. 'tac'이 표적 단어일 때.

그림 3 은 'toe' 가 표적 단어일 경우이며 앞의 경우들과 같이 'toe'와 결합하는 다른 철자의 수는 증가하고 나머지 단어들에 대해서는 감소한다.

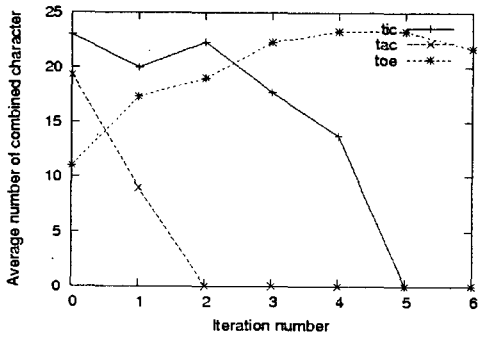

그림 3. 'toe' 가 표적 단어일때의 결과.

4. 결론

DNA 결합 행렬을 이용한 새로운 갯수 분포 보정 방법으로 표적 단어로의 supervised learning을 구현했다. 이는 먼저 보여준 단어를 기억해내는 실험에서 표적단어 기억이 공고해지는 현상을 시뮬레이션한다. 유사 단어 간 기억 친밀성도 확인할 수 있었다. DNA 조각에 magnetic bead를 달아 특정 DNA 조각을 끌어낼 수 있으며 이는 흡사 낚시에서 미끼로 물고기를 낚는 원리와 비슷하다. 이런 방식의 기억인출 방법으로 암묵적인 과정을 통해 연합기억이 확실해지는 과정을 DNA 염기서열과 simulated annealing 방법을 이용하여 시뮬레이션하였다.

감사의 글

손정욱, 강윤정 그리고 태강수 선생님의 친절한 논의에 감사를 드린다. 이 연구는산업자원부의 MEC 과제와 과학기술부의 NRL 그리고 교육부의 BK-21 들의 지원을 받았다.

참고문헌

1. L. M. Adleman, Molecular computation of solutions to combinatorial problems, Science 266, 1021-1024, 1994.
2. Y. Benenson et al., An autonomous molecular computer for logical control of gene expression, Nature, 423-429, 2004.
3. B,-T, Zhang and H,-Y, Jang, A Bayesian algorithm for in vitro molecular evolution of pattern classifiers, *Preliminary Proceedings of the Tenth International Meeting on DNA Computing (DNA10)*, pp.294-303, 2004.
4. B.-T. Zhang and H.-Y. Jang, Molecular programming: evolving genetic programs in a test tube, *The Genetic and Evolutionary Computation Conference (GECCO 2005)*, vol. 2, pp. 1761-1768, 2005.
5. J. Kim et al., Neural network computation by *in vitro* Transcriptional Circuits, Advances in Neural Information Processing Systems (NIPS) 17, 681-688, 2004.
6. 이은석, DNAgram: Anagram 문제 해결에 관한 분자 컴퓨팅 시뮬레이션 연구, 서울대학교 석사학위 논문, 2003. 2.
7. J. Chen, R. Deaton, Y. Wang, A DNA-based memory with *in vitro* learning and association recall, Natural computing 4, 83-101, 2005.
8. R. Ratcliff and G. Mckoon, A counter model for implicit priming in perceptual word identification, Psychological Review 104, 319-343, 1997.
9. M. C. Mozer et al. Mechanisms of long-term repetition priming and skill refinement: A probabilistic pathway model, In *Proceedings of the Twenty Fifth Annual Conference of the Cognitive Science Society, 2003.*
10. 김준식, 김종찬, 노영균, 이동윤, 장병탁, DNA 컴퓨팅 연산 과정의 통계 물리적 예측, 한국컴퓨터 종합학술대회(Korea Computer Congress), 253-255, 2005. 7.