분자컴퓨터를 통한 애너그램 시뮬레이션

이지훈 O1 천효선 2 이은석 3 류제환 4 장병탁 1234

서울대학교 생물정보학 협동과정[1]

서울대학교 컴퓨터공학과[2]

서울대학교 인지과학 협동과정[3]

서울대학교 뇌과학 협동과정 [4]

{jhlee, hschun, eslee, jhryu, btzhang}@bi.snu.ac.kr

Simulation of Anagram Solving by Molecular Computer

Ji-Hoon Lee O1 Hyo-Sun Chun[2] Eun Seok Lee[3] Je-Hwan Ryu[4] 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

이 논문은 분자컴퓨터를 통한 애너그램 시뮬레이션에 관한 것이다. 사람의 인지과정을 컴퓨터가 모사하기 위해서는 알고리즘적 접근도 중요하지만 컴퓨팅 자체를 DNA와 같은 생체물질을 사용하여 본질적으로 인지과정이 일어나는 것과 유사한 환경을 만들어 모사하는 것도 중요하다. 애너그램 문제는 무작위로 섞인 문자가 주어졌을 때 가능한 단어를 찾는 문제로 사람의 인지현상을 연구하는데 많이 사용되어왔다. 애너그램의 답을 찾는 과정에서 숙련자는 초보자 보다 제약조건의 빈도와 같은 보다 구조적인 단어 정보를 사용한다는 것이 밝혀져 있고, 이것은 애너그램을 풀이하는 과정을 분자로 모사하는데 핵심적인 역할을 한다. 본 연구에서는 애너그램 풀이 과정을 확률적인 제약 만족 문제로 보았고, 생체분자물질인 DNA를 사용해 정보를 인코딩 하고 애너그램 풀이과정을 모사할 수 있는 분자애너그램 알고리즘을 디자인 하였다. 컴퓨터 시뮬레이션 실험 결과, 이 분자알고리즘이 인지적 애너그램 풀이과정을 모사함을 보였고 실제 분자실험과정에서 나타날 수 있는 문제점들을 미리 예측할 수 있게 되었다.

1. 서 론

데이비드 럼멜하트는 인간의 인지현상을 간단한 유닛들로 이루어진 복잡한 네트워크를 통해 설명할 수 있다고 하였다[1]. 이 네트워크는 결국 뇌를 모사한 것인데, 뇌는 뉴런과 시냅스라는 간단한 유닛들의 연결로 만들어져있다. 대략적으로 1011 개의 뉴런이 서로 연결되어 1014 개의 연결을 만들고 있다[2]. 각 시냅스에는 신경전달물질, 단백질 수용기, 이온물질 등 수천가지의 분자물질들이 각자의 역할을 하고 있다. 연결주의자(Connectionist) 입장에서 뇌의 기능을 보았을 때 우리는 사람의 인지활동이 이러한 작은 분자물질들의 상호작용을 기본으로 하여 이루어지고 있다는 것을 쉽게 유추해 볼 수 있다[1][3]. 최근에 뇌의 활동을 시뮬레이션 하고자 하는 시도가 많이 이루어지고 있는데 핸리 마크램의 경우 슈퍼컴퓨터를 사용하여 수많은 뉴런을 모델링 하여 인공 두뇌를 만들고자 하고있다[4]. 이러한 시도들에는 중요한 한계점이 있는데 우선 인지현상을 구현하는 컴퓨터가 생물학적 두뇌와 본질적으로 차이가 있다는 것이다. 만약 생물학적 두뇌를 모사하기 위해서 기존 컴퓨터 대신 생체물질을 사용한 컴퓨팅 기술을 사용한다면 기존 컴퓨터가 잘 하지 못하는 부분을 극복할 수 있을 것이다.

DNA를 사용한 최초 분자컴퓨터가 만들어진 후[5], 최근에는 분자컴퓨터에서 영향 받은 인지기계학습에 대한 연구가 시작되었다[1][6][7]. DNA분자의 자기조립 현상과 초병렬적 연산 특징은 뉴런들의 상호연결을 통한 컴퓨팅 현상과 유사한 부분이 많다[2]. 본 연구에서는 분자컴퓨터를 통한 사람의 인지현상 모사를 위해 애너그램을 선택하였다. 애너그램은 철자를 맞추는 문제로 예를 들면 철자가 뒤섞인 ‘ogod’ 라는 글자를 보고 알파벳의 순서를 바꾸어 존재하는 단어인 ‘good’을 맞추는 것이다. 애너그램은 풀이과정에 대해 충분한 연구 결과가 존재하고 있으며 사람의 고도의 인지현상을 잘 보여준다[8]. 또한 한국에서 이은석 및 다른 연구자들을 통해 DNA 컴퓨팅 개념을 사용해 애너그램을 풀고자 하는 시도가 있었다[9][10][11]. 본 연구에서는 이전 연구에서 다루지 못했던 보다 구체적인 분자알고리즘디자인과 애너그램을 해결하는 과정에서 나타나는 인지현상을 모사하는 시뮬레이션 결과를 제시한다.

2. 분자 실험 디자인

애너그램 문제는 다음 그림1과 같이 제약만족 네트워크로 표현할 수 있다. 애너그램문제는 문제를 많이 풀어본 숙련자의 경우 양질의 제약조건을 알고 있고, 초보자의 경우 알고 있는 유용한 제약조건이 많지 않기 때문에[8] 제약만족네트워크에서 알파벳과 제약조건들 사이의 가중치값 w_i이 다르게 구성 된다. 이와 같이 제약조건들 사이와 어휘집과 연결된 가중치값 w_i도 다르게 구성되어 있을 것이다. 분자실험에서 이 가중치 값은 각 제약조건과 어휘를 표현하는 DNA 분자의 절대적 양으로 표현이 된다.

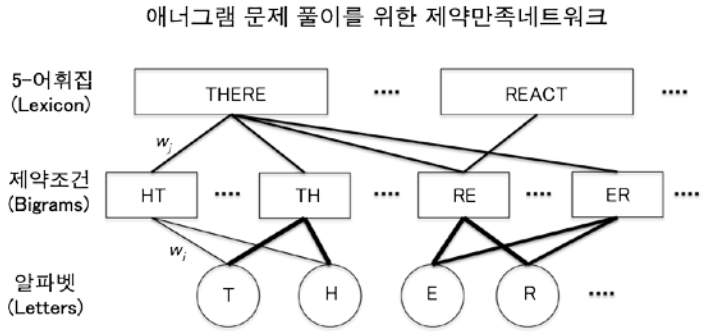

그림1. 3단계로 구성된 애너그램 제약만족네트워크. 네트워크의 최하층에서부터 영문 알파벳, 제약조건, n-어휘집(Lexicon, n개 문자로 구성된 단어 어휘집)으로 구성되어 있다. 각 노드를 연결하는 에지의 가중치는 굵기로 구분할 수 있다.

분자애너그램 제약만족네트워크는 다음과 같은 분자알고리즘에 의해서 구현이 된다. 우선 분자실험에 사용된 모든 알파벳, 제약조건, 어휘집은 DNA간의 결합에너지, GC 빈도, cross-hybridization 과 같은 실험시 발생할 수 있는 문제를 최소화 할 수 있도록한 DNA 서열 디자인 방법을 사용하여 DNA 서열로 인코딩하였다[7]. 실험에 사용되는 알파벳 서열은 제약조건서열과 또한 다음 단계에서 어휘집과 결합이 이루어져야 하기 때문에 dsDNA의 한 가닥인 (+)서열로 정의를 하였고 제약조건과 어휘집은 (+)서열의 상보가닥인 (-)서열로 정의를 하였다. Hybridization 단계에서는 주어진 알파벳들과 제약만족네트워크상에 존재하는 제약조건과의 결합이 이루어진다. 제약만족네트워크는 애너그램 숙련자와 초보자의 네트워크가 다르게 구성된다. 숙련자의 경우 이 단계에서 알파벳들이 제약조건과 더욱 많은 결합이 이루어지고, 초보자의 경우 결합이 거의 이루어지지 않게 된다. Hybridization 후 연결된 알파벳 사이에는 실제 DNA 서열이 화학적으로 연결되어 있지 않은 상태이다. 따라서 Ligase 효소를 사용하여 알파벳 사이에 생긴 공간을 메워준다. 이렇게 알파벳이 연결된 DNA 서열은 제약조건과 상보적인(Complementary) 서열을 갖게 된다(이하 cBigrams)

분자애너그램 알고리즘(*In vitro*)

1. Hybridization : 알파벳 (+), Bigrams (-)
2. Ligation
3. cBigrams 추출 (+)
4. cBigrams 증폭 (Polymerase Chain Reaction)
5. Hybridization : 추출된 cBigrams (+) + n-어휘집 (-)
6. 전기영동(Electrophoresis)
7. 어휘집의 단어와 완전하게 결합된 dsDNA 추출
8. 정답확인 (Sequencing)
9. END

그림 2 분자애너그램 알고리즘

이 모든 과정은 한 개의 마이크로 튜브상에서 이루어 지는데 Ligation후의 튜브 내부에는 생성된 cBigrams 외에도 결합되지 않은 알파벳, 제약조건, 같은 시퀀스들이 여전히 존재하게 된다. 이 중 순수하게 cBigrams만 추출하기 위해 마이크로 마그네틱 비드(Micro magnetic beads)를 사용하여 cBigrams를 추출한다. 추출과정을 거친 DNA서열은 충분하게 어휘집과 결합시키기 위해서 PCR(Polymerase Chain Reaction)을 사용하여 증폭을 한다. 두번째 Hybridization은 추출된 cBigrams와 n-어휘집과의 결합이다. 어휘집은 주어진 문제 알파벳 수 n값에 따라 다르게 선택한다. cBigrams들은 어휘집과 공백이 없게 결합이 되면 완벽한 단어를 형성하게 되고 이것은 전기영동을 통해 구분할 수 있다. 전기영동 후 n 개의 알파벳으로 만들어진 단어를 전기영동 젤에서 추출한다. 추출된 DNA서열은 시퀀싱을 통해 서열을 분석하게 되고 어떤 단어가 생성되었는지 정답을 확인한다.

실제 애너그램 숙련자와 초보자를 통한 인지실험결과에 의하면 제약만족네트워크는 제약조건을 얼마나 학습하여 알고있는가에 따라 애너그램 풀이에 걸리는 시간이 차이가 난다[12]. 숙련자의 경우 주어진 알파벳을 보는 순간 제약조건과의 병렬적 검색과정을 통해 빠르게 답을 찾아낼 수 있게 된다(Pop-out solution). 그에 반해 초보자의 경우 주어진 알파벳들과 연결된 제약조건들 사이의 연결이 약해 알파벳을 순차적인 가설검증 과정(Serial hypothesis-testing)을 거치게되어 답을 찾는데 보다 많은 시간이 소요된다.

3. 실험 결과

컴퓨터 시뮬레이션 결과는 다음 그림3과 같다. 시뮬레이션은 초보자와 숙련자의 경우로 나누어서 수행되었다.

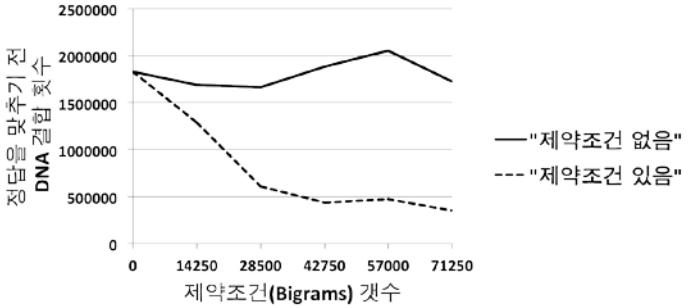

그림 3 분자애너그램 시뮬레이션 결과

애너그램 초보자의 경우 학습된 제약조건이 없다고 가정하여 제약조건을 넣지 않았고, 숙련자의 경우 제약조건을 증가시켜 가며 실험을 수행하였다. 그림 3의 세로축은 시뮬레이션 과정 중에 DNA 서열들이 정답을 맞추기 전 다른 DNA 서열들과 결합한 총 횟수인데 결합 횟수가 적어질 수록 정답을 빨리 맞추는 것을 뜻한다. 가로축은 제약조건들 갯수인데 숙련자의 경우 0에서부터 계속 증가시키며 실험하였으며 초보자의 경우 제약조건은 없는 상태에서 숙련자의 제약조건 증가분 만큼 알파벳 서열을 늘려주었다. 실험 결과 제약조건이 증가 할 수록 정답을 처음으로 맞추는 속도가 빠르게 증가하는 것을 확인할 수 있었으며, 제약조건의 개수가 약40000개를 넘어갈 경우 정답을 맞추는 속도가 4배 가까이 증가하는 것을 확인할 수 있었다. 이것은 숙련자의 애너그램 풀이과정 중 나타나는 pop out 현상을 분자애너그램이 잘 모사하는 것이라 볼 수 있다. 또한 이것을 통해 실제 분자실험시 유용한 정보인 DNA 결합 횟수와 정답이 Pop out 되는데 필요한 제약조건의 갯수를 어느정도 예측할 수 있었다.

4. 결론

본 논문에서는 인간의 인지현상을 모사하기 위해 DNA를 사용한 분자애너그램 알고리즘을 디자인하였다. 여기서 애너그램 풀이과정은 확률적인 제약만족 문제로 보았고 분자실험 과정은 컴퓨터를 사용해 시뮬레이션 실험으로 수행하였다. 본 연구에서 제시한 분자애너그램은 실제 DNA실험 과정을 고려하여 디자인 되었으며, 컴퓨터 시뮬레이션 실험을 통해 분자애너그램이 실제 인지적인 애너그램풀이과정을 잘 모사할 수 있는지를 확인 하였다. 시뮬레이션은 분자 하나하나를 비율적으로 모사하여 수행되었으며 실제 분자실험과정에서 생길 수 있는 필요한 조건 및 문제점 등을 미리 예측할 수 있었다.

감사의글

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

참고문헌

[1] Rumelhart, D. E., McClelland, J. L., Parallel distributed processing: explorations in the microstructure of cognition, Volume 1: Foundations, MIT press, 1986.

[2] Zhang, B. T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *Computational Intelligence Magazine*, 3, 3, 49-63, 2008.

[3] Feldman, Jerome A., Dana H. Ballard, Connectionist models and their properties, *Cognitive science*, 6, 3, 205-254, 1982.

[4] Markram, H., The blue brain project, *Nature Reviews Neuroscience*, *7,* 2, 153-160, 2006.

[5] Adleman, L. M., Molecular computation of solutions to combinatorial problems, *Science*, 266, 5187, 1021-1024, 1994.

[6] Zhang, B. T., Random hypergraph models of learning and memory in biomolecular networks: shorter-term adaptability vs. longer-term persistency, *In Foundations of Computational Intelligence*, 344-349, 2007.

[7] 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*(1), 51-56, 2011.

[8] Novick, L. R., Sherman, S. J., The effects of superficial and structural information on online problem solving for good versus poor anagram solvers, *The Quarterly Journal of Experimental Psychology*, *61*(7), 1098-1120, 2008.

[9] 이은석, 윤지은, 장병탁, DNAgram: Anagram문제해결에 관한 분자 컴퓨팅 연구, *한국인지과학회*, 2003.

[10] 강윤정, 이은석, 태강수, 장병탁, 확률 라이브러리 모델(PLM)에 의한 애너그램 문제 해결, *한국인지과학회*, 130-134, 2005.

[11] 김수동, 이은석, 장병탁, Plasmid-DNAgram: 녹색형광단백질 발현 Plasmid DNA 기반 분자컴퓨팅에 의한 언어 퍼즐 문제 해결, *한글 및 한국어 정보처리 학술대회*, 293-299, 2003.

[12] Novick, L. R., Sherman, S. J., On the nature of insight solutions: Evidence from skill differences in anagram solution. *The Quarterly Journal of Experimental Psychology: Section A*, 56(2), 351-382, 2003.