하이퍼망 모델을 이용한 분자자기조립기반 언어처리:

DNA 컴퓨팅 실험 설계

이지훈 [O1] 장하영 [2] 정원형 [4] 이승환 [3] 박태현 [3] 박성배 [4] 장병탁 [1,2]

서울대학교 생물정보학 협동과정 [1]
서울대학교 컴퓨터공학부 [2]
서울대학교 화학생물공학부 [3]
경북대학교 컴퓨터공학과 [4]

jhlee@bi.snu.ac.kr, hyjang@bi.snu.ac.kr, whchung@sejong.knu.ac.kr, skulsh78@snu.ac.kr ,
thpark@snu.ac.kr, sbpark@sejong.knu.ac.kr, btzhang@bi.snu.ac.kr

Molecular Self-assembly-Based Language Generation Using the Hypernetwork Model: Design of DNA Computing Experiments

Ji-Hoon Lee[O1] Ha-Young Jang[2] Won-Hyong Chung[4] Seung Hwan Lee[3] Tai Hyun Park[3] Seong-Bae Park[4]
Byoung-Tak Zhang[1,2]

Graduate Program in Bioinformatics, Seoul National University[1]
Department of Computer Science & Engineering, Seoul National University[2]
School of Chemical and Biological Engineering, Seoul National University[3]
Department of Computer Engineering, Kyungpook National University[4]

사람들의 언어 활동을 컴퓨터로 모사하는 방법 중 한 가지가 문장 생성을 해 보는 것이다. 사람들은 문장을 자유 자제로 생성하고 조합하는 능력이 있다. 이것은 언어 학습의 결과로 인한 것인데, 현상은 알지만 실제 언어 학습과 문장 생성의 프로세스에 관해서는 정확히 알기가 어렵다. 이러한 문제를 언어 하이퍼망 모델을 만들어 유아의 언어 학습 과정에서 문장 생성을 바탕으로 이해하려고 하는 시도가 있었다[1]. 언어 하이퍼망은 분자들의 초상호작용 반응에 기반한 분자정보처리 모델인 하이퍼망(Hypernetwork) 모델을 언어에 적용한 언어 학습 및 문장 생성 모델이다[3]. 이 연구에서 유아용 비디오 자막 문장을 학습 데이터로 하여 생성되는 문장들의 구문론적 또는 의미론적으로 정합적인 문장 비율을 분석 하였는데, 이것은 언어 하이퍼망 모델이 언어 학습과 문장 생성 연구에 적합함을 보인 것이다. 본 연구에서는 이러한 언어 하이퍼망을 바탕으로 DNA 하이퍼망 언어모델을 디자인 하여 직접 DNA를 사용한 언어 학습과 문장 생성이 가능함을 보이고자 한다.

본 연구에서는 기존의 컴퓨터를 사용하지 않고 DNA를 문장 학습과 생성에 사용하는데, 이것은 DNA의 특성인 초병렬분산 정보저장 및 처리능력이 문장 학습 및 생성을 효률적으로 하기에 적절하기 때문이다. 언어는 복합성(Compositionality)을 가지고 있고, 사실상 무한에 가까운 문장을 생성하는 생산성(Productivity)을 가지고 있다[2]. 이러한 특성은 DNA 하이퍼망 언어모델을 통해 적절하게 모사될 수 있다.

DNA 하이퍼망 언어모델은 크게 3가지 단계로 나누어 기술할 수 있다. 첫째, 학습할 문장들의 집합에서 각 문장의 하이퍼에지를 추출한 후 각 하이퍼에지를 DNA 서열로 인코딩 한다. 둘째, 인코딩된 DNA 하이퍼에지 서열들을 한 마이크로튜브에 모아 학습시키고, In vitro 상에서 자기결합을 통해 문장을 생성한다. 셋째, 생성된 DNA 문장들에서 키워드를 포함하는 문장들을 추출한 후 서열분석을 통한 디코딩을 수행한다. 그림 1에서처럼 문장(I'll talk to my friend and my sister)의 각 단어들은 고유한 DNA 서열로 치환된다. 인코딩된 DNA 서열들은 그림 2에서와 같이 양쪽으로 연결되어 DNA 문장을 생성하게 된다.

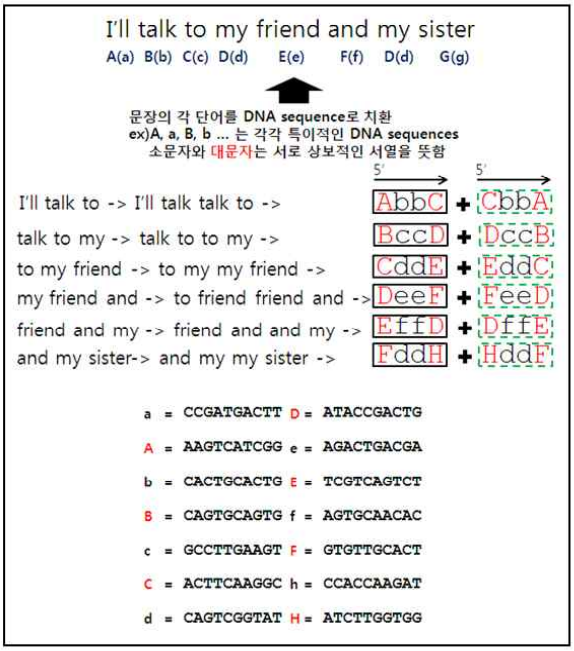

그림 1 DNA sequence 인코딩 과정

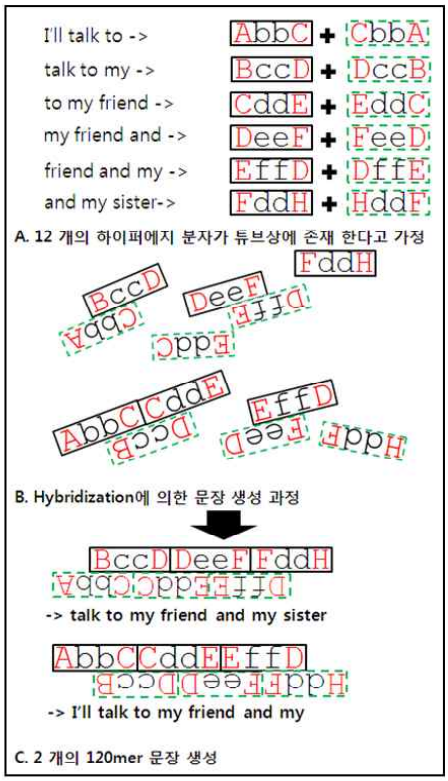

그림 2 문장 생성 과정

본 연구에서는 컴퓨터과학에서의 인공지능 학습 이론과, 인지과학의 언어현상, 분자생물학의 실험 도구를 융합한 연구를 통한 새로운 언어정보처리 방법을 제시하고 이를 실제 DNA 실험으로 구현하기 위한 설계를 제시하였다. DNA를 언어 학습에 사용하는 것은 DNA의 분자메모리적 특성을 자연스럽게 활용한 것이며, DNA 분자의 조합적인 문장 생성은 DNA의 초병렬적 자기결합특성을 효과적으로 이용한 것이다. 이처럼 DNA를 적절하게 언어학습과 확률적인 문장 생성을 가능하게 한 점이 DNA 하이퍼망 언어모델의 장점이라고 할 수 있다. 현 시점에서 대규모 시퀀스를 합성해야 하는 실험에서는 비용 문제와 정확한 실험적 재현을 위해서는 아직 해결해야 할 기술들이 남아 있다. 그러나 최근 시퀀스 합성 및 분석 기술이 급격히 발전하고 있으며 현재 이 논문에서 제시한 인코딩 및 디코딩 방법을 사용하여 DNA 실험으로 구현하는 연구가 진행 중이다.

감사의 글

본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업 [2009-F-051-01, 차세대 맞춤형 서비스를 위한 기계학습 기반 멀티모달 복합 정보 추출 및 추천기술 개발] 및 한국연구재단 기초연구사업에 의해서 지원되었음.

참 고 문 헌

[1] Lee, J.-H., Lee, E.-S. and Zhang, B.-T., A hypernetwork memory-Based model of sentence learning and generation in children: How a child learns to produce language from a video corpus, *Proc. of the Korea Computer Congress 2009*, vol.36, no.1(A), pp.128-129, 2009. (in Korean)
[2] Fodor, J.A., and Pylyshyn, Z.W., Connectionism and Cognitive Architecture: A Critical Analysis, Cognition, vol.28, pp.3-71, 1988.
[3] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, vol.3, no.3, pp.49-63, 2008.
[4] Zhang, B.-T. and Park, C.-H., Self-assembling hypernetworks for cognitive learning of linguistic memory, *Proceedings of International Conference on Computer, Electrical, and Systems Science, and Engineering (CESSE2008)(WASET)*, vol. 27, pp.134-138, 2008.
[5] SantaLucia, Jr., J., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci*. vol. 95, pp.1460-1465, 1998.
[6] Markham, N. R. and Zuker, M., UNAFold: software for nucleic acid folding and hybridization. In Keith, J. M., editor, Bioinformatics, vol. 2, Structure, Functions and Applications, number 453 in Methods in Molecular Biology, pp.3-31, Humana Press, Totowa, NJ. 2008.
[7] Shendure, J. and Ji, H., Next-generation DNA sequencing, *Nature Biotechnology*, vol.26, no.10, pp.1135-1145, 2008.