의미적·철자적 거리를 고려한 단어의 DNA 부호화

류제환 O1 이지훈 2 장병탁 1,2,3
[1]서울대 뇌과학 협동과정
[2]서울대 생물정보학 협동과정
[3]서울대 컴퓨터공학부
{jhryu, jhlee, btzhang}@bi.snu.ac.kr

Integrated DNA Encoding of Semantic and Orthographic Distances between words

Je-Hwan Ryu O1 Ji-Hoon Lee 2 Byoung-Tak Zhang 3
[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Interdisciplinary Program in Bioinformatics, Seoul National University
[3]Department of Computer Science & Engineering, Seoul National University

요 약

DNA는 그 분자적인 특성 때문에 계산 매질로써 인간 수준의 지능적 계산을 수행하는 데 있어서 여러 가지 장점이 있다. 이미 DNA를 사용한 문장 생성 실험을 통해 이러한 특성을 활용하려는 연구도 수행된 바 있다. 하지만 기존 연구에서는 단어를 DNA 서열로 치환하는 과정이 임의적이었기 때문에 DNA 서열로 표현된 단어는 다른 단어들과의 관련성을 잃어버리게 된다는 문제점이 있다. 본 연구에서는 단어간 의미 및 철자적 거리를 고려하여 언어를 DNA 서열로 부호화하는 방법을 제시하고 인코딩 결과를 소개한다.

1. 서론

컴퓨터를 이용하여 인간 수준의 지능을 성취하려는 노력은 컴퓨터 과학 분야에서 꾸준하게 계속되고 있다. 한편 DNA Computing은 DNA의 분자적인 성질을 활용하여 계산을 수행하는 학문으로, 계산 과정이 분자의 고유한 성질에 영향을 받아 자발적인(spontaneous) 방향으로 진행된다. 이러한 DNA Computing의 특징이 인간 수준의 지능적 계산을 성취하는 데에 도움이 될 것이라는 견해가 있으며[1][2], DNA를 사용한 문장 생성 실험을 통해 이러한 특징을 활용하려는 연구도 수행되었다[3][4][5]. 하지만 이 연구에서는 단어를 DNA 서열로 부호화 할 때에 DNA 서열들간의 부적절한 결합만 고려했기 때문에 DNA 서열로 표현된 단어는 다른 단어들과 의미 및 철자적 거리를 잃어버리게 된다는 문제점이 있다.

이런 약점을 극복하고자, 본 논문에서는 DNA 서열간 거리가 단어간 거리를 반영하는 방법을 제안한다. 이에 본 논문에서는 단어간 관계가 이미 정해져 있는 말뭉치(corpus)에서 관계가 긴밀한 여러 군의 단어 집합들을 추출하고, 추출된 단어들에 적절한 변이 과정(mutation)을 통해 조절된 DNA 서열들을 할당한다. 그리고 DNA 서열간 결합 에너지[6]를 비교하여 단어간 거리가 적절하게 반영되었음을 결과로 제시한다.

2. 단어의 DNA 부호화

2.1. Wordnet

워드넷(Wordnet)은 영어의 의미적 어휘목록이다[7]. 워드넷의 기본적인 단위는 동의어 집단(Synset)으로, 한 동의어 집단은 간략한 정의와 이 집단에 속하는 몇 개의 어휘들로 구성된다. 한 동의어 집단은 의미적으로 연관관계가 있는 다른 동의어 집단을 가질 수 있으며, 이러한 연관관계로는 상위어(Hypernym), 하위어(Hyponym), 부분어(Meronym), 반의어(Antonym) 관계가 있다. 특히 상위어와 하위어 관계는 워드넷의 기본 구조를 형성하는 핵심적인 관계로, 이러한 상위-하위 관계 때문에 워드넷은 계층 구조를 형성하게 된다[8].

2.2. 어휘간 의미적 · 철자적 거리

워드넷에서 각 동의어 집단들은 여러 연관관계를

통해 이어지게 되므로, 적절한 거리 계산법을 정의하면 동의어 집단간 의미적 거리를 계산할 수 있다. 이러한 의미적 거리 계산에는 워드넷 계층구조에서 단어간 경로 길이나 출현 빈도 등이 사용된다. 본 연구에서는 문제의 복잡도를 낮추기 위해 단어의 상위어 · 하위어 관계만 거리 정의에 포함시킨다. 한편, 어휘간 철자적 거리는 레벤스타인(Levenshtein) 거리를 통해 쉽게 계산할 수 있다. 계산된 의미적 거리와 철자적 거리에 각각 가중치를 곱하여 통합적 거리를 형성하게 된다.

2.3. 변이를 통한 DNA 부호화

두 어휘간 거리를 DNA 서열 부호화에 나타내는 것은 DNA 서열간 일치도를 조절함으로 가능하다. 그러나 모든 동의어 집단에 대해서 서열 일치도를 조정하는 것은 동의어 집단의 수가 많아질수록 복잡해지는 문제이므로, 이번 연구에서는 한 어휘를 기준으로 그 하위에 있는 단어들만 부호화 대상에 포함시키도록 한다.

서열 일치도 조정 문제는 변이 문제로도 표현이 가능하다. 즉, 어떤 동의어 집단과 그 집단을 부호화하는 DNA 서열이 있으면, 서열을 변이시켜 다른 관계 집단에 대한 DNA 서열을 생성할 수 있다. 변이의 정도나 횟수는 어휘간 의미 및 철자적 거리에 비례하게 결정한다. DNA 부호화 과정은 다음과 같다.

1. 어휘 계층 구조에서 가장 상위에 위치한 어휘에 초기 DNA 서열을 할당
2. 서열이 할당된 어휘의 하위어 집단을 추출
3. 통합적인 거리를 계산
   A. 워드넷을 이용한 의미적 거리 계산
   B. 레벤스타인 거리를 이용한 철자적 거리의 계산
   C. 가중치를 반영한 거리의 통합
4. 부모 서열을 재귀적으로 변이시켜 하위어의 DNA 서열을 생성
   A. 중복이 없는 변이
   B. 중복이 있는 변이

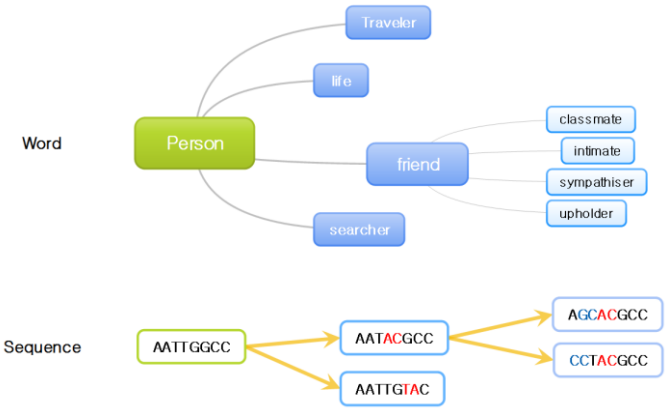

그림 1. 단어 위계구조와 서열 변이

변이 과정에서는 전 단계에서 변이된 위치의 재변이를 허용하느냐 아니냐의 두 가지 방법이 있다. 재변이를 허용하지 않는다면 계층 구조 내의 모든 어휘는 일관된 거리를 갖지만, 필연적으로 변이 단계에 제한을 받게 된다.

본 연구는 DNA를 사용한 문장 생성 문제에 대해 개선점을 제시하는 연구이므로, 성능 평가의 척도로 언어의 의미적 거리가 실제 DNA 서열간 결합 확률에 반영되었는지를 측정하는 것이 적절하다. 화학적인 결합 확률은 결합 에너지에 의존적이므로, 직접적인 성능 평가에 DNA 서열간 결합 에너지를 사용한다. 변이시킨 자리수가 같으면 문자적 거리는 동일하게 증가하지만, 화학적 결합의 경우 핵산의 종류나 결합 분포에 따라 에너지가 조금씩 차이가 난다는 특징이 있다. 따라서 변이 서열이 여러 개가 있다면, 같은 수준의 변이 서열이라도 본래 서열과 결합하는 비율은 조금씩 차이가 나게 된다.

3. 실험 결과

제안한 방법의 타당성을 평가하기 위해 각 어휘당 추출하는 하위 집단수는 10개, 변이가 일어날 때 바꾸는 서열은 핵산 8개 분량으로 조절하여 컴퓨터로 시뮬레이션 하였다. 어휘간 거리 계산은 의미적 거리만 반영되도록 가중치를 조정하였으며, 위계도에서 한 단계의 수준차가 있으면 그대로 한 변이 단계를 진행했다. 변이되었던 서열도 재변이가 가능하도록 하였다. 총 서열길이는 45-mer 였으며, 이 중 머리 태그가 8-mer, 꼬리 태그가 7-mer로 실제 변이가 일어나는 서열 길이는 30-mer로 고정되었다. 가장 상위에 위치한 어휘로부터 얼마나 다른 서열이 생성되는지를 그림 2를 통해 볼 수 있다.

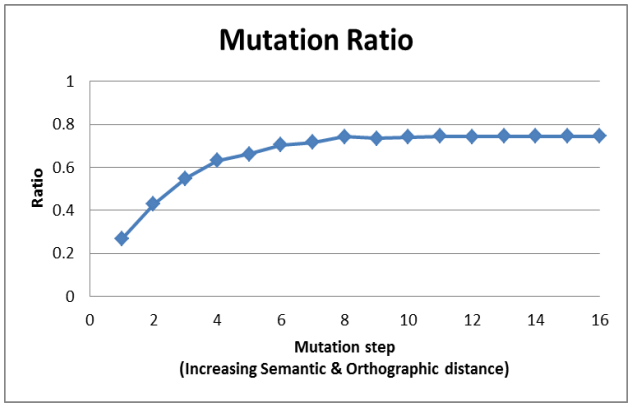

그림 2. 변이 단계에 따른 서열 변이 비율

총 16단계를 변이시켰으며, 핵산이 일치하지 않는 부분의 비율이 약 0.75로 수렴하였다. 실제로 중복이 가능한 변이를 거듭하면 단계가 증가할수록 변이된 서열은 원래의 서열과 관계가 점차 사라져 임의의 서열에 근접한다. 임의 서열이 원래 서열과 한 부분이 일치할 확률은 25%이므로 변이비율이 0.75에 근접하게 된다. 이러한 거리 정보의 반영은 화학적 결합 에너지로도 확인할 수 있다.

표 1. 변이 수준에 따른 결합 에너지

		변이 수준		
		0	1	2
표본(Kcal/mole)	1	-81.9	-19.59	-19.89
	2		-20.85	-20.85
	3		-20.85	-19.89
	4		-17.95	-17.95
	5		-22.19	-17.95
	6		-25.74	-17.95
	7		-25.74	-20.85
	8		-19.5	-19.89
	9		-20.85	-17.95
	10		-20.85	-17.95
평균 자유 에너지 (Kcal/mole)		-81.90	-21.41	-19.11
반응 속도		1.86E+27	9.96E+00	1.00E+00

표 1을 보면 변이 단계가 높아질수록 결합 에너지가 낮아지는 것을 볼 수 있다. 반응속도 상수는 결합 에너지에 지수적으로 비례하므로 단계 1과 2의 반응속도는 약 10배정도의 차이를 보인다. 따라서 변이를 통해 생성한 DNA 서열들에 어휘간 거리 정보가 반영되어 있음을 확인할 수 있다.

4. 결론

본 논문에서는 의미적 · 철자적 거리 정보를 반영하는 어휘의 DNA 부호화 문제를 해결하는 방안으로 DNA 변이를 제안하였다. 시뮬레이션 결과 변이로 인해 생성된 DNA 서열들은 어휘간 거리를 반영하고 있음을 서열간 일치도와 결합 에너지를 통해 확인할 수 있었다. 이를 통해 언어 데이터를 DNA를 통해서 계산할 때 보다 정보 손실이 없는 계산이 가능하리라 기대한다.

감사의 글

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

참고문헌

[1] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *Computational Intelligence Magazine*, IEEE3.3, p. 49-63, 2008.

[2] 이은석, 윤지은, 장병탁, DNAgram: Anagram 문제 해결에 관한 분자 컴퓨팅 시뮬레이션 연구, *한국인지과학회 춘계학술대회 논문집*, p. 258-262, 2003.

[3] Lee, J.-H., Lee, S.H., Chung, W.-H, Lee, E.S., Park, T.H., R. Deaton, and Zhang, B.-T., A DNA assembly model of sentence generation, *BioSystems*, 106, p. 51-56, 2011.

[4] Lee, J.-H., Kim, J.-W., R. Deaton, Lee, S. H., Park, T. H., and Zhang, B.-T., Molecular machine learning in vitro, *International Conference on DNA Computing and Molecular Programming (DNA 18)*, p.55, 2012.

[5] 이지훈, 이은석, 장병탁, 하이퍼망 메모리 기반 유아 언어학습 및 생성 모델, *한국컴퓨터종합학술대회*, 제36권 1(A), p. 128-129, 2009.

[6] N. Markham and M. Zuker., UNAFold, *Bioinformatics*, p. 3-31, 2008.

[7] G. Miller, WordNet: a lexical database for English, *Communications of the ACM* 38.11, p. 39-41, 1995.

[8] Budanitsky, Alexander, and G. Hirst., Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Workshop on WordNet and Other Lexical Resources*. 2, 2001.