분자 컴퓨팅을 위한 효율적인 DNA 서열 생성 시스템

김동민[O] 신수용 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
{dmkim, syshin, btzhang}@bi.scai.ac.kr

Effective Sequence Generation for Molecular Computing

Dong-Min Kim[O] Soo-Yong Shin Byoung-Tak Zhang
School of Computer Science and Engineering,
Seoul National University

요 약

최근 DNA 분자의 병렬성을 이용한 DNA 컴퓨팅 기법들이 활발히 개발되고 있다. 그러나, DNA 컴퓨팅은 실제 생체 분자인 DNA를 사용하기 때문에 생체분자의 화학적 성질에 의한 오류의 가능성을 항상 내포하고 있다. 이러한 문제를 극복하고자 오류의 가능성을 최소화시키는 방법들이 연구되고 있고, 특히 DNA 서열을 만들 때 오류의 가능성을 최소화시키는 방법들이 많이 연구되고 있다. 본 논문에서는 현재 개발하고 있는 시스템인 NACST를 간단히 소개한 후, DNA 컴퓨팅에 사용할 DNA 서열을 생성하기 위해서 유전자 알고리즘을 사용하는 방법을 제안하며, 유전자 알고리즘을 이용하여 DNA 서열을 효율적으로 생성하기 위한 적합도 함수들에 대해서 구체적으로 살펴보았다.

1. 서 론

Adleman[1]이 1994년 DNA 분자를 이용하여 Hamiltonian Path Problem을 해결한 이후, DNA 분자의 병렬성을 이용한 다양한 DNA 컴퓨팅 기법들이 개발되고 있다[2]. DNA 컴퓨팅은 생체분자인 DNA를 계산 및 저장의 매체로 사용하고, 생물학 실험실에서 사용되는 여러 가지 실험 방법들을 연산자로 이용하는 계산 모델이다. 기존의 컴퓨터와 비교할 때의 특징을 몇 가지 살펴보면, 첫째, 기존의 컴퓨터가 2진수를 사용하는데 반하여, DNA 컴퓨팅은 4진수 체계를 사용한다. DNA의 4종류인 A(Adenine), C(Cytosine), G(Guanine), T(Thymine)를 사용하여 자료를 2진수 체계가 아닌 4진수로 표현한다. 둘째, 막대한 병렬성을 가지고 있다. 대략 핵산 용액 1 mole당 $6x10^{23}$개의 엄청난 양의 DNA가 존재하고 이들 각각 DNA가 연산의 대상으로 사용이 된다. 따라서 기존의 컴퓨터에 비해서 막대한 양의 자료를 저장하는 도구로 사용하거나, 엄청난 병렬성을 가진 컴퓨터를 개발할 수 있을 것으로 기대된다. 지금 현재 DNA 컴퓨팅은 계산학적으로 어려운 문제들인 NP-complete 문제들을 해결하거나[1, 2], 명제 증명[2], 대용량 associative memory 개발[2] 등 여러 분야에 적용되고 있다.

그러나 DNA 컴퓨팅은 DNA 분자를 사용하기 때문에 몇 가지 문제점을 가지고 있다. 특히 연산자로 사용되는 실험 방법들이 정확한 결과를 보여주지 않는 다는 것이 가장 큰 문제점이다. 연산자로 사용되는 실험 방법들이 DNA의 화학적 성질을 이용하여 생화학적인 반응을 통해서 결과를 도출해 내는데, 생화학 반응의 특징상 항상 오류의 가능성을 가지고 있다. 일반적으로는 화학 반응들이 아주 높은 안정성을 가지고 있다고 알려져 있지만, 반응 중 생기는 아주 작은 오류일지라도 항상 동일한 결과를 배출해야 하는 컴퓨터로써는 큰 문제가 될 가능성을 내포하고 있는 것이다. 이러한 단점을 극복하기 위해서 DNA 컴퓨팅 시에 오류를 최소화하는 방법들이 다각도로 연구되고 있는데, 특히 DNA 서열을 효율적으로 생성하여 오류의 가능성을 사전에 최소화하는 방법이 부각되고 있다[3]. DNA 서열을 디자인할 때 DNA 컴퓨팅 연산자인 실험실 실험에서 발생할 수 있는 오류의 가능성들을 미리 제거하면, 연산 과정에서 오류가 최소화되기 때문에 보다 정확한 결과를 얻을 수 있을 것이다. 이러한 목적에 따라 여러 연구가 진행되고 있었고, 본 연구팀도 이미 NACST라는 DNA 컴퓨팅 시뮬레이터의 서열 생성 부분 개발에서 이 문제를 고려하였다.[3]. 본 논문에서는 NACST의 DNA 서열 생성 부분을 좀더 정확한 모델로 개선하고자 하였다. 2장에서는 분자 컴퓨터 시뮬레이터인 NACST에 대해서 간단히 설명을 하고, 3장에서는 개선된 DNA 서열 생성 방법에 대해서 기술하였다. 그리고 마지막으로 4장에서 요약 및 결론을 하고자 한다.

2. 분자 컴퓨터 시뮬레이터

NACST[3](Nucleic Acid Computing Simulation Toolbox)은 DNA 서열 생성, 실험 과정 등 실제 DNA 컴퓨팅 과정에서 필요한 모든 사항을 컴퓨터상에서 시뮬레이션 할 수 있도록 설계된 프로그램이다. 전체적인 구조는 크게 서열 생성 부분과 실험 과정 시뮬레이션의 2부분으로 나뉘어 지며, 부가적으로 기타 데이터 처리, 결과 확인을 위한 그래프 작성 부분, 사용자의 편의를 위한 입력 스크립트 등의

부가적인 부분으로 구성되어 있다.

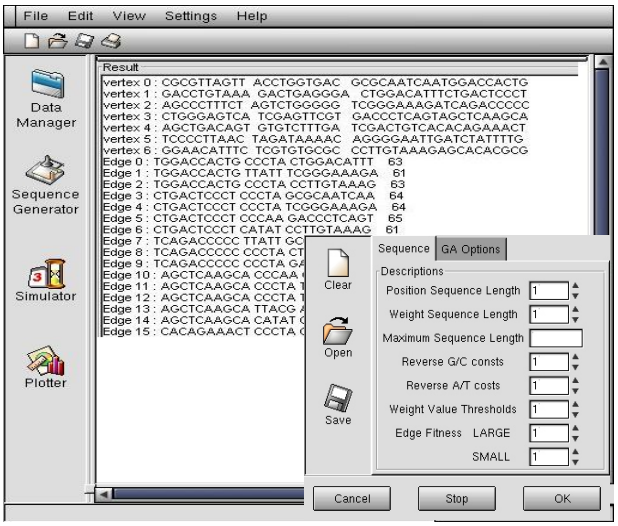

그림 1. NACST 서열 생성기 화면

먼저 DNA 서열 생성 부분의 특징을 살펴보면 다음과 같다. 첫째 효율적인 DNA 서열을 생성하기 위해서 인공지능 기법 중에서 유전자 알고리즘을 기반으로 하였다. 따라서 최근 개발되고 있는 DNA 서열 시스템들이 그래프를 이용한 단순한 생성 방식[2]이나, 비교적 단순한 적합화 과정[2]를 갖는 것에 비하여, 다양한 적합도 함수를 선택하여 가장 최적화된 DNA 서열을 생성할 수 있다. 둘째로 기존의 시스템들이 고유한 서열을 생성하는 것에만 초점을 둔 것에 비해서 임의의 서열들 간을 연결하는 DNA 서열을 생성하여 보다 복잡한 구조의 DNA 구조를 생성하는 과정도 지원할 수 있도록 하였다. 셋째로 hairpin 구조와 같은 2차적인 구조도 사용자가 생성할 수 있는 기능을 지원한다.

실험 과정 시뮬레이션 부분은 DNA 컴퓨팅으로 풀고자 하는 그래프 문제를 파일 형태로 입력받으며, 앞서 생성된 서열을 이용하여 실험 과정을 시뮬레이션을 할 수 있도록 설계되어 있다. 각 시뮬레이션은 실제 실험을 기준으로 모듈화되어 다양한 조합이 가능하며, 각 모듈은 통계적인 방법을 기준으로 작성되어 있다.

표 1. DNA 적합도 함수

함수명	비고
Similarity	측정 대상 서열이 다른 서열과 비슷한 정도를 측정
H-measure	다른 서열들과의 상보성 (Complementary)을 고려
GC-contents	서열에서 G, C 염기가 차지하는 비율
Continuity	특정 염기가 연속되어 나타나는 정도
Self-complementary	서열이 자체로 상보 결합하여 2차 구조를 형성할 가능성
Melting Temperature	상보 결합된 서열의 50%가 분리되는 온도
PCR primer	서열이 PCR primer로서 기능할 가능성

다음 장에서는 DNA서열 생성 부분에서 사용하는 최적화 기준들에 대해서 설명하고자 한다.

3. DNA 서열 최적화 방법

DNA 컴퓨팅의 성공 여부는 문제 해결에 사용된 DNA 서열의 종류와 성질에 크게 의존하므로, DNA의 화학적 특성을 수치화 하여 실험에서의 적합성을 고려하는 일이 중요하다. 표 1에서 NACST에서 구현한 DNA 적합도 함수들을 정리하였다.

3.1 Similarity

Similarity는 한 서열이 다른 서열과 다른 고유한 정도를 측정하기 위한 방법이다. 기본적으로 임의의 두 개의 서열에서 동일한 위치에 같은 종류의 염기가 있는 정도를 측정한다. 부가적으로 보다 정확한 고유성을 보장하기 위해서 한 서열을 한 염기만큼 이동시키며 다른 서열과의 고유성도 비교하도록 하였다. 그림 2에서 좌측의 그림의 적합도는 3이고, 우측의 경우도 적합도는 3이나 한 염기만큼 이동한 경우의 결과를 보여주고 있다.

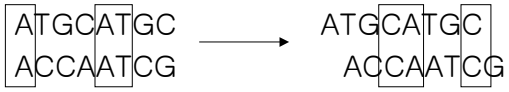

그림 2. Similarity. 염기 서열은 5‘ → 3’ 방향

3.2 H-measure

Similarity는 방향이 동일한 서열간의 고유성만을 점수화하는 기준인 반면, H-measure는 서로 염기 서열 방향이 상보적이어서(5‘→3’과 3‘→5’) DNA duplex를 형성하는 경우에 내가 원하는 염기 서열들 간에 결합이 생성되는 지의 여부를 판단하는 기준이다. 각 서열이 서로 상보 서열이라는 가정 하에서 각 염기가 서로 상보 결합을 형성하는 횟수를 계산하였다. 역시 Similarity와 마찬가지로 한 염기씩 이동을 하면서 계산하는 과정도 포함되어 있다. 이 H-measure는 실험과정에서 발생하는 mismatch hybridization과 shifted hybridization을 사전에 예방하는 효과를 가져다준다. 그림 3에 그 과정이 설명되어 있고, 좌측의 그림에서 적합도는 3이고, 우측의 결합은 그림 2처럼 한 염기가 이동한 경우를 보여주는데 적합도는 1이다.

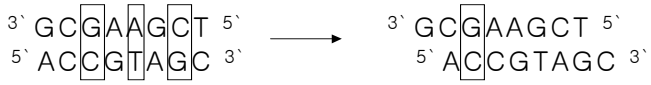

그림 3. H-measure

3.3 GC-contents

DNA의 상보 결합은 A-T, G-C 염기간의 결합으로 이루어진다. 이중 G-C 결합은 A-T 결합에 비해 결합강도가 강하므로, 전체 서열의 결합 정도는 서열 안에 존재하는 G, C 염기의 비율에 크게 의존하게 된다. 결합의 강도는 구조의 안정성으로 이어지며, 이는 DNA 컴퓨팅의 안정성을 결정하는 요인이 된다. GC-contents 수치는 식 (1)을 통해 구한다.

$$(GC-contents) = \frac{number\ of\ G,C}{number\ of\ all} \qquad (1)$$

3.4 Continuity

같은 종류의 염기가 연속적으로 나타나는 서열은 실제 실험에서 의도하지 않은 결합을 형성하는 경우가 많이 발생하며, 그 반응을 제어하는 것이 어렵다. 실제 실험에서 생길 수 있는 상황을 고려하여 NACST에서는 2개의 동일 염기 연속은 허용하고, 3-5개의 연속은 점수화하며 약간의 페널티를 주었고, 한 염기가 6개 이상 연속되는 서열은 생성하지 않도록 하였다.

3.5 Self-complementary

Self-complementary의 대표적인 현상은 hairpin이라고도 불리는 현상으로 한 염기 서열이 휘어져서 자신의 염기들이 상보 결합을 형성하여 2차 구조를 만들어 내는 경우를 일컫는다. DNA 컴퓨팅에서 다른 서열과 결합하도록 생성된 서열들이 예상치 못한 hairpin을 일으킴으로서 실험의 안정성을 떨어뜨리는 요인이 될 수 있다. 또는 그 반대로 특정 염기 서열이 hairpin 구조를 생성하여 컴퓨팅 과정에 사용될 수도 있다. 따라서 NACST에서는 hairpin 구조를 형성하는 것을 수치화하도록 고려하였다. 그림 4에서 보는 바와 같이, 한 서열을 결합하는 양 끝단, 휘어지는 부분의 세 부분으로 나누고 양 끝단의 부분 서열(sub sequence)들을 비교하는데, 염기 서열이 휘기 위해서 최소한 5개 이상의 염기 서열이 필요하고, 상보 결합을 형성하여 안정화되기 위해서 6개 이상의 염기 서열이 결합을 형성하여야 한다. 따라서 이 기준을 만족하는 경우가 발생하는지의 여부를 판단하였다. 역시 한 염기 서열에서 발생할 수 있는 모든 가능성을 고려하였다.

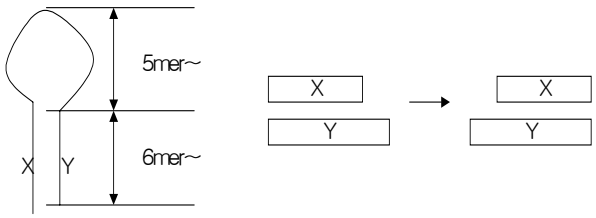

그림 4. Self-complementary

3.6 Melting Temperature

실제 실험에서는 각 서열들을 상보결합-분리하는 과정이 반복적으로 이용되므로, 생성된 서열 집합이 어느 정도의 Melting Temperature를 갖는가를 예측하는 것이 의미가 있다. NACST에서는 2가지 식을 사용하였다. 식 (2)는 nearest-neighbor model[4]로써 염기 서열의 길이가 대략 50 미만일 경우에 사용하였다. 식 (3)은 염기 서열의 길이가 50 이상일 경우에 사용하는 식이다.

$$T_m = \frac{\Delta H^\circ}{\Delta S^\circ + R\ln(|C_T|/4)} \quad (2)$$

(R: Boltzmann constant, $|C_T|$: 전체 서열 농도

$\Delta H^0, \Delta S^0$ 계산식은 [4] 참조)

$$Tm = 81.5 + 41 \times GC\,Ratio - 500 \div Length \quad (3)$$

3.7 PCR primer

PCR은 서열의 개수를 증폭시키는 실험이며, PCR primer란 PCR에 사용되는 특정한 서열을 지칭한다. PCR primer로 사용될 서열은 GC-contents, Continuity, Melting Temperature등에서 다른 서열들과는 약간 다른 적합도를 적용할 필요가 있으며, NACST는 이를 고려하도록 설계되어 있다.

생성된 DNA 서열 집합의 전체 적합도는 이제까지 살펴본 각 적합도 함수들의 가중합(weighted sum)인 식 (4)로 결정되며, 이는 곧 유전자 알고리즘의 적합도로 이용되어 DNA 서열을 최적화하는데 사용된다.

$$F = \sum_{i=1}^{n} w_i f_i \quad (4)$$

4. 결론 및 향후 과제

DNA 컴퓨팅에서는 DNA의 생화학적인 특성에 의해 오류의 가능성이 항상 존재하며, 본 논문에서는 이를 최소화하기 위한 방법으로서 서열 생성 단계에서의 적합화 기준을 제시하였다. 이렇게 수치화된 DNA 서열 집합의 적합도는 유전자 알고리즘을 통해 최적화되며, 그 결과 서열은 실제 DNA 컴퓨팅의 서열로서 이용되기에도 충분할 것으로 생각된다. 앞으로는 최적화 알고리즘을 개선하여 DNA 서열 생성에 특화된 유전자 알고리즘을 개발하고, 실험 과정 시뮬레이션 부분의 모델 개선 및 구현을 통해 보다 실제적인 DNA 컴퓨팅 시뮬레이터를 구현하고자 한다. 또한 부가적으로 실험 데이터 처리, 결과 해석 부분 등을 확장하여, NACST 안에서 DNA 컴퓨팅 문제 디자인부터 결과 리포팅까지 할 수 있게 되기를 기대하고 있다.

감사의 글

본 연구는 산업자원부 차세대신기술사업과 BK21 프로그램에 의하여 일부 지원되었음. 이 연구를 위해 연구 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터신기술공동연구소에 감사드립니다.

참고문헌

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

[2] N. Jonoska & N.C. Seedman (Eds.), *Preliminary Proceedings of 7th International Meeting on DNA Based Computers*, University of South Florida, Tampa, FL, June10-13, 2001.

[3] B.-T. Zhang S.-Y. Shin, "Molecular Algorithms for Efficient and Reliable DNA Computing", in *Proc. Genetic Programming 1998*, pp. 735-742, Morgan Kaufmann, 1998.

[4] J. SantaLucia, Jr., "A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbor thermodynamics", *Proc. Natil. Acad. Sci. USA*, Vol.95, pp.1460-1465, 1998.