열역학적 데이터를 이용한 26도시 Traveling-Salesman Problem 시뮬레이션

장하영O 신수용 장병탁
서울대학교 공과대학 컴퓨터공학부 바이오지능 연구실
{hyjangO, syshin, btzhang}@bi.snu.ac.kr

DNA Computing Simulation Using Thermodynamic Data For TSP With 26 Nodes.

Ha-Young Jang0 Soo-Yong Shin Byoung-Tak Zhang
BioIntelligence Lab. School of Computer Science and Engineering, Seoul National University

요 약

DNA 컴퓨팅에 대한 연구가 진행되어 감에 따라 기존의 튜링 머신과 동등한 계산 능력을 가진 다양한 계산 모델이 제안되고 있으며, 이와 함께 DNA의 병렬성을 이용하여 NP 문제들을 풀고자 하는 시도가 계속되고 있다. 그러나 전통적인 폰 노이만 기계에서의 알고리즘이 해집합에 대한 순차적 탐색을 하는 것과는 달리 가능한 모든 해를 미리 생성해 놓고 그 중에서 해를 찾아내는 기존의 DNA 컴퓨팅 알고리즘으로는 NP 문제의 크기가 증가함에 따라 초기 해의 생성조차도 불가능하게 된다. 이에 대한 해법의 하나로 진화적인 방법론을 생각할 수 있지만, 이 경우에는 진화 연산을 위한 추가적인 연산자의 고안과 이의 적용에 따른 어려움이 생긴다. 따라서 본 논문에서는 DNA 컴퓨팅에서 가능한 초기 해를 모두 생성할 수 있는지를 열역학적인 데이터에 근거한 시뮬레이션을 통하여 검증하였다. 이러한 과정을 통해서 값비싼 실제 실험의 성공 여부나 실험 디자인의 정당성 등을 미리 예측할 수 있을 뿐만 아니라, DNA 컴퓨팅이 보다 큰 크기의 NP 문제를 해결할 수 있는 가능성을 제공할 수 있다.

1. 서 론

DNA 컴퓨팅은 생체 분자인 DNA를 계산 및 저장의 매체로 사용하고, 생물학 실험실에서 사용되는 여러 가지 실험 방법들을 연산자로 이용하는 계산 모델이다. 그러나 현재의 분자 생물학 기술로는 문제의 크기가 커서 현재의 컴퓨터로는 해를 구하기 힘든 문제를 풀기는 사실상 불가능하다. 왜냐하면 지금까지 개발된 DNA 컴퓨팅 기법은 모든 가능한 해들(혹은 해가 될 수 없는 것까지도)을 DNA를 사용하여 생성하고 이 중에서 문제의 요구 조건에 맞는 답을 찾아가는 이른 바 "exhaustive 탐색"을 사용하기 때문이다.

예를 들어 해가 Boolean 값들의 조합으로 이루어진 문제의 경우, 문제의 크기가 40일 때 $2^{40} \fallingdotseq 10^{12}$ 의 가능한 해가 존재하지만 일반적으로 생물학 실험에서 사용하는 수 *pico mole*의 농도로는 이 정도 크기의 문제조차 그 해를 정확하게 전부 표현하기 어렵다. 또한 이 정도 크기의 문제는 개인용 컴퓨터로도 생물학 실험을 사용하는 DNA 컴퓨팅 보다 훨씬 더 정확하게 빨리 풀 수 있다. 실제로 현재까지 크기가 20인 3-SAT 문제가 DNA 컴퓨팅으로 푼 가장 크기가 큰 NP-complete 문제이다 [1].

이와 같은 난점을 극복할 수 있는 한 가지 방법은 해의 전체 탐색 과정 또는 초기해의 생성을 위해 진화적인 개념을 도입하는 것이다 [2, 3]. 이러한 진화적인 방법론들은 분명 앞에서 언급한 DNA 컴퓨팅의 문제점을 개선하는데 있어서 큰 도움을 주고 있다. 그러나 이러한 진화적 방법론의 도입만으로 초기해의 생성을 보장할 수 있다고 생각하는 것은 여전히 어려움이 있다. 따라서 본 논문에서는 26도시 traveling salesman problem (TSP)의 해를 실험적으로 검증하기에 앞서서, 실제 실험에서 초기 해의 생성을 보장해 줄 수 있도록 1-base mismatch와 dangling ends데이터를 포함한 열역학적인 시뮬레이션[4]을 통해서 실제 초기 해의 생성을 위한 실험 과정을 검증하고자 한다. 이러한 과정을 통해서 비싼 비용이 드는 실제 실험을 수행하기 전에 실험의 성공 가능성을 예측할 수 있을 뿐 아니라 보다 큰 NP 문제의 해결에도 도움이 될 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2장과 3장에서는 문제의 정의와 우리가 선택한 실제 문제에 대해서 설명하고 초기해 생성의 진화적 개념을 살펴보겠다. 그리고 4장에서는 초기 해의 생성과 유효해의 선택 과정에 대한 시뮬레이션을 이용한 검증에 대해 논의하겠다. 이후 5장에서 결론을 내리겠다.

2. The traveling-salesman problem

Traveling-salesman problem (TSP) 은 제한된 수의 도시들 간의 여행을 하는데, 모든 도시를 방문하면서 가장 여행 비용이 싸게 들도록 하고 출발지점으로 되돌아 오는 문제이다. 좀 더 명확한 정의는 다음과 같다 [5].

TSP = {(G, c, k): G = (V, E) is a complete graph,
c is a function from $V \times V \rightarrow Z$, $k \in Z$, and
G has a traveling-salesman tour with cost at most k}

우리가 DNA 컴퓨팅을 통해서 풀고자 하는 문제는 그림 1에서 보여지는 바와 같이 26개의 노드와 46개의 에지를 가지고 있는 문제이다. exhaustive serach를 통해서 가능한 모든 경로를 탐색해 본 결과 총 81개의 경로와 minimum cost 58을 가지는 경로를 찾을 수 있었다. 이와 같이 일반적으로 논의되는 complete graph가 아닌 에지의 수가 상당히 적은 문제이기 때문에 DNA 컴퓨팅으로 초기 해의 생성이 가능할 수도 있지만, 초기 해의 생성시에 해로서의 조건을 만족시키지 못하는 많은 부수적인 해들이 같이 생성된다는 점을 감안할 때에 단순히 각각의 node와 edge들을 나타내는 DNA 조각들을 한데 섞어줌으로 해서 초기 해를 만드는 것은 거의 불가능하다고 생각할 수 있다.

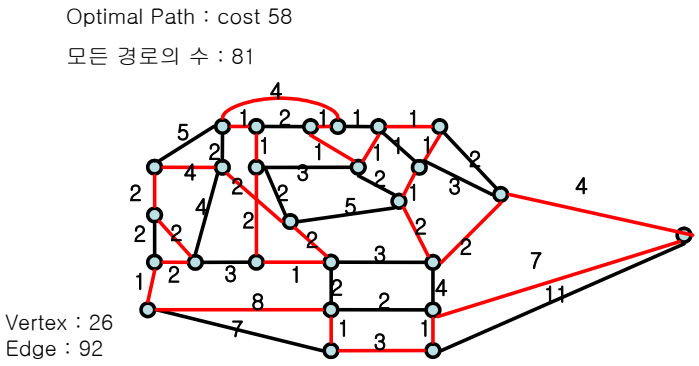

그림 1 The traveling salesman problem with 26 nodes.

3. 초기해의 진화적 생성

진화적인 방법론을 사용하고 있지만 해의 전체 탐색 과정에서 진화적인 방법을 이용하는 것이 아니라 초기해의 생성을 위해 진화적인 개념을 도입한 방법으로 DNA shuffling이 있다. DNA Shuffling은 재조합하고자 하는 DNA를 여러 가지 수단에 의해 절단(cleavage) 혹은 분쇄(breaking)하고 여기서 얻어지는 이중가닥의 상보적인 DNA조각들을 섞은 다음 PCR을 하게되면 이종의 상보적인 DNA 조각들이부분접합(partial annealing)과 신장(extension)을 하게 됨으로써 재조합 DNA가 만들어지는 기법이다. 즉, 그림 1에서 보여지는 것과 같은 분자진화의 개념을 도입하여 보다 효율적으로 초기해를 생성하고자 하는 것이다.

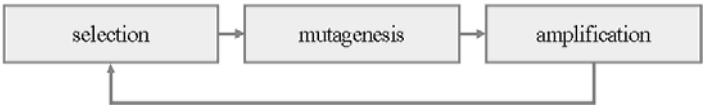

그림 2 분자진화.

그러나 이러한 실험 과정에 있어서 가장 큰 문제점은 그러한 과정을 얼마나 반복해야 충분한지를 알 수가 없다는 점이다. 극단적으로 이야기한다면, sequence의 길이가 모든 도시를 포함하기에 충분하도록 늘이기 위해서 ligation을 반복하다 보면 결과적으로 우리가 원하는 길이의 sequence보다 긴sequence들만이 남을 수도 있다는 것이다.

따라서 실제 실험에 들어가기 전에 실험의 가이드라인으로써 시뮬레이션을 실행하는 것은 성공적인 실험을 위해서 매우 큰 도움이 될 수 있다. 이를 위해서 NACST/Sim [4]를 이용하여 실험에 사용될 DNA sequence의 primary structure간의 self-homology와 cross-homology을 검증하고, 실제 hybridization의 시뮬레이션을 통해 반응 시간이 진행되어 나감에 따라서 전체 해집합에 존재하는 DNA sequence의 길이 변화를 관찰함으로써, 실제 실험자들이 실험을 하는데 있어서 가이드 라인을 제시하고자 한다.

4. 실험 결과

simulation은 NACST/seq [6]를 이용하여 만들어낸 총 232개의 candidate sequence set 중에서 NACST/Sim [4]을 이용하여 self-homology와 cross-homology 또는 잘못된 결합 가능성이 있는 sequence를 검사하여 선별된 sequence set에 대해서 수행하였다.

Simulation은 하나의 sequence set에 대해서 세가지 방법으로 수행되었다. 먼저 각각의 sequence들이 3차원 상에서의 위치 정보를 가지게 하여 수행을 하였고, 다음에는 위와 동일한 실험을 2차원 위치정보를 이용하여 수행하였다. 마지막으로 각각의 sequence들이 자신의 위치 정보를 가지는 것이 아니라 tube 내에 uniform하게 분포하여 서로 random하게 결합이 이루어진다는 가정 하에 실험하였다. 2차원과 3차원 위치 정보의 경우에는 sequence의 분포를 유지하기 위하여 일정한 시간 간격마다 perturbation이 이루어지게 하였다.

하나의 sequence set은 각각 26mer로 이루어진 26개의 ssDNA로 구성이 되어 있는데, simulation에서는 이 26개의 ssDNA를 모두 합쳐 750만개 정도의 크기를 가진 sequence pool을 구성하여 이용하였고, sequence pool 내부에는 26개의sequence가 uniform하게 분포하고 있다고 가정하였다.

그림 3에서 볼 수 있는 것처럼 3종류의 simulation에서 모두 sequence의 최대 길이는 유사한 것으로 나타나고 있다. 더욱 복잡한 위치 정보를 유지한 simulation의 경우에 좀 더 빠른 시간 내에 길이의 변화가 일어난 것처럼 보이고 있으나, 사실 이러한 문제는 위치 정보를 담고 있는 simulation의 경우에는 한번의 iteration에서 보다 많은 수의 hybridization이 이루어 지도록 구현이 되어 있기 때문에 일어나는 현상으로 생각된다. 즉, 개개의 sequence가 위치 정보를 가지고 있지 않는 경우와 실제 위치 정보를 가지고 있는 경우에 큰 차이가 없음을 알 수가 있다.

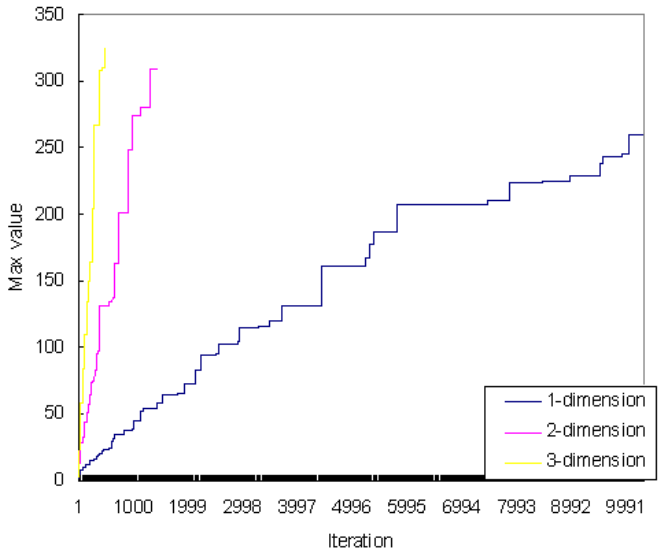

그림 3 최대 길이의 변화.

그림 4에서는 hybridization이 일정 정도 지난 후에 각각의 sequence 길이에 따라 존재하는 sequence의 개수를 보여주고 있다. 여기에서도 앞에서 보았던 최대 길이의 변화 그래프에서와 마찬가지로 위치정보를 가지고 있는 경우와 그렇지 않은 경우가 큰 차이가 없음이 나타나고 있다. 또 하나 주목할 수 있는 사실은 길이에 따른 sequence의 분포가 예상했던 것보다 훨씬 급격히 감소하는 로그함수의 모습을 하고 있다는 사실이다. 실제 실험 과정에서 sequence의 길이 분포가 완만한 로그 함수 또는 정규분포의 형태를 따를 것이라는 예상과는 달리 상대적으로 길이가 짧은 sequence가 많이 존재한다는 결과가 나왔다. 이러한 사실은 iteration이 진행되어서 hybridization/ligation이 더 많이 일어남에 따라 더욱 극명하게 보여졌다. 즉, hybridization/ligation이 완료된 후의 결과에서는 이보다 더욱 급격히 감소하는 로그 함수의 형태가 보여진다.

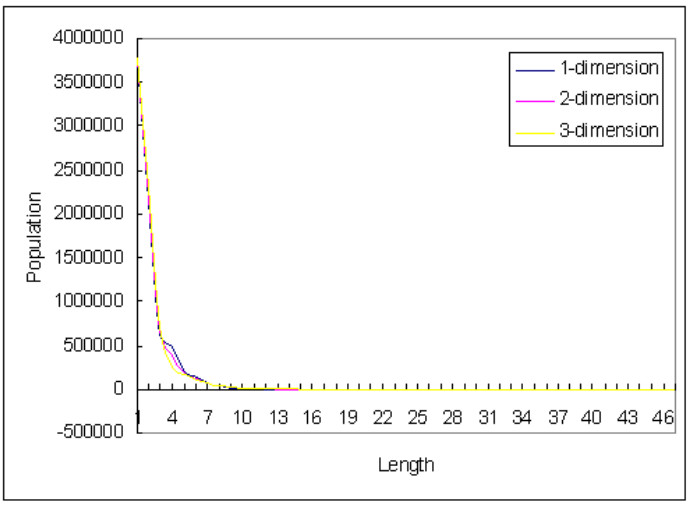

그림 4 길이별 sequence의 분포(초기 개수의 50%).

simulation의 수행에서 위치 정보와 함께 중요한 역할을 하는 것이 바로 전체 DNA sequence의 개수다. DNA sequence의 개수가 변화함에 따라서 두개의 sequence가 만날 수 있는 확률이나 전체 반응 속도 등이 변화할 수 밖에 없기 때문에 전체 sequence의 개수가 어떻게 변화하고 있는지 아는 것이 중요하고, 이를 위해서 각각의 simulation에서 전체 sequence 개수의 변화를 관찰하였다. 그림 5를 보면 시간의 증가에 따라 전체 sequence의 개수가 log scale로 감소하고 있는 것을 관찰할 수 있다. 예상했던 것 처럼 반응이 계속해서 일어남에 따라서 반응에 참가할 수 있는 전체 DNA sequence의 개수가 줄어들고 이와 동시에 서로 결합할 수 있는 DNA의 개수도 줄어들기 때문에 시간이 흐름에 따라서 점점 반응 속도가 느려지고, 일정 시간 이후에는 충분한 시간이 경과하더라도 거의 반응이 일어나지 않아서 전체 DNA의 개수가 변화하지 않을 것이라는 것을 예상할 수 있다.

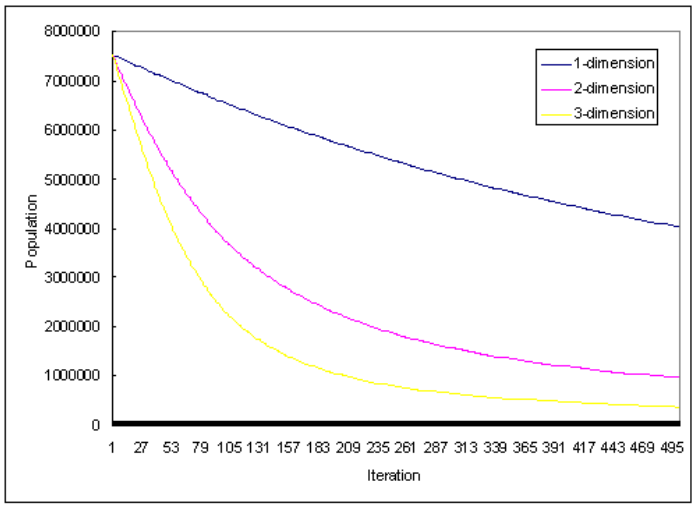

그림 5 전체 sequence 개수의 변화.

그래프 상에서 위치 정보를 가지고 시행한 simulation과 그렇지 않은 simulation의 전체 개수가 변화하는 속도가 다른 것은 그림 3에서 보여지고 있는 것과 마찬가지로 실제 구현상의 특징으로 인해서 위치 정보를 담고 있는 simulation의 경우에는 한번의 iteration에 더 많은 결합이 일어나기 때문인 것으로 생각된다.

5. 결론

실험 결과 총 750만개 정도의 population을 가지고 simulation을 했을 경우에, 우리가 원하는 길이의 sequence가 평균적으로 60-70개 정도 발생하였다. 일반적인 생물학 실험에서 한번의 실험에서 사용되는 DNA sequence가 수십억개 가량이라는 점을 감안할 때, 이 결과에서 생성되는 feasible solution의 개수는 수천개 정도가 될 것이라 예상할 수 있다. 앞에서 이야기 한 대로 우리가 선택한 문제의 가능한 경로의 수가 81개밖에 되지 않기 때문에 이러한 경우 모든 solution을 생성해낼 수도 있다고 생각할 수 있지만, 실제로 생성되는 solution들은 TSP 상에 존재하는 cycle로 인해 81개의feasible한 solution만이 생성되는 것이 아니라 길이는 만족시키지만 잘못된 경로를 담고 있는 많은 solution들을 생성해 낼 수 있다는 점을 생각하면 실험의 성공을 장담할 수는 없는 것이 사실이다.

따라서 보다 성공적인 실험의 수행을 위해서는 앞에서 논의한 DNA shuffling등의 방법을 이용한 분자진화 개념의 적용을 통한 초기 해집합의 생성이 필요하리라 생각되고, 이러한 DNA shuffling 또한 실제 실험 전에 그 성공여부를 simulation 하는 것이 가능할 것이다.

감사의 글: 본 연구는 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정 연구실 과제에 의해 지원되었음. 이 연구를 위해 연구 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터 연구소에 감사드립니다.

5. 참고문헌

[1] R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L. M. Adleman, Solution of a 20-variable 3-SAT problem on a DNA computer, *Science*, vol. 296, pp. 499-502, April 2002.

[2] D. H. Wood, J. Chen, E. Antipov, B. Lemieux, and W. Cedeno, In vitro selection for a OneMax DNA evolutionary computation, *DNA Based Computers V*, pp. 23-37, 2000.

[3] D. H. Wood, J. Chen, E. Antipov, B. Lemieux, and W. Cedeno, A design for DNA computation of the OneMax problem, *Soft Computing*, vol. 5, no. 1, pp. 19-24, 2001.

[4] 장하영, 신수용, 장병탁, 열역학적 데이터에 기반한 DNA/DNA 연쇄 결합 반응 시뮬레이션, *한국정보과학회 가을 학술발표 논문집 (II)*, 제30권 2호, pp. 772-774, 2003.

[5] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, McGraw-Hill, New York, NY, 1990.

[6] D. Kim, S.-Y. Shin, I.-H. Lee, and B.-T. Zhang, NACST/Seq: A Sequence Design System with Multiobjective Optimization, *Lecture Notes in Computer Science*, vol. 2568, pp. 242-251, 2003.