유전알고리즘을 이용한 범용 올리고뉴클레오타이드 태그 디자인

임희웅[O,1] 유석인[1] 장병탁[2]
서울대학교 인공지능연구실[1]
서울대학교 바이오지능연구실[2]
{hwlim[O], siyoo, btzhang}@bi.snu.ac.kr

Universal Oligonucleotide Tag Design using Genetic Algorithm

Hee-Woong Lim[O] Suk-In Yoo[1] Byound-Tak Zhang[2]
Artificial Intelligence Lab. Seoul National University
Biointelligence Lab. Seoul National University

요 약

올리고뉴크레오타이드 서열의 디자인은 일반 분자 생물학 뿐만 아니라 DNA 컴퓨팅 분야에서도 중요한 문제이다. DNA나 RNA와 같은 생체 물질간의 화학반응을 이용하여 계산을 수행하는데 사용되는 염기 서열의 품질은 계산의 정확도에 큰 영향을 미치기 때문에, 문제의 특성에 따른 요구 조건에 맞는 염기 서열을 디자인 하기위한 방법에 대해 여러 가지 연구가 있어왔다. 기존의 DNA 컴퓨팅을 위한 염기서열 디자인은 주어진 녹는점의 범위에서 단순히 서로 독립적인 염기서열들의 집합을 디자인 하거나, 분자생물학 실험에 사용되는 올리고 프로브나 프라이머 셋을 디자인 하는 것을 중심으로 이루어졌다. 반면, 본 논문에서는 세포에서 추출된 DNA/RNA 분자가 섞여있는 환경에서 어느 DNA/RNA 분자와도 혼성화 반응을 하지 않는 범용 올리고뉴클레오타이드 태그를 디자인 하는 간단한 유전 알고리즘을 제시하며, 이를 이용해서 디자인된 염기서열 결과를 제시한다.

1. 서 론

바이오 분자 컴퓨팅은 생체에 존재하는 물질을 직접 사용해서 정보를 처리하는 방법으로서, 생체 물질간의 화학반응을 정보 처리의 관점에서 바라보고 이를 이용해서 연산을 수행하고자 하는 연구이다. 현재까지 DNA, RNA, Protein등의 여러 가지 물질들과 분자생물학 실험 방법이 응용되어 많은 연구가 있어왔고 최근에는 실제 임상 환경에 적용 하려는 시도가 이루어지고 있다 [1-4]. 특히 생체 정보의 핵심이라고 할 수 있는 DNA를 이용한 계산에 관한 연구가 많이 이루어지고 있다.

DNA를 구성하는 각 단위 염기들은 A, T, G, C 네 가지 종류가 있는데 이들은 두개씩 상보적으로 쌍을 이루어 결합하는 성질이 있다. 이들이 사슬구조로 연결되어 DNA 단일 가닥을 이루고, 이 단일 가닥 역시 상보성에 따라 다른 DNA 단일 가닥과 결합하여 DNA 이중 가닥을 형성한다. 이때 두 가닥간의 결합력은 둘 사이에 상보적인 염기들의 비율에 크게 좌우되며, 이러한 상보성의 조절을 이용해서 DNA 가닥들을 디자인하여 DNA 컴퓨팅에 사용한다. 결과적으로 DNA 컴퓨팅의 신뢰성은 컴퓨팅을 위해 디자인된 DNA 염기 서열의 품질에 좌우된다. 뿐만 아니라, PCR 반응을 위한 프라이머 염기 서열의 디자인이나, 유전자 발현 패턴 분석을 위한 마이크로어레이 칩에 사용되는 프로브 시퀀스 등, 일반 분자생물학이나 생물정보학 분야에서도 DNA 염기 서열의 디자인은 중요한 문제이다[5].

DNA 컴퓨팅을 위한 기존의 염기서열 디자인은 선택적인 결합을 위한 서로 독립적인 DNA 가닥들의 집합을 생성하거나, 원하는 이차구조를 가지는 DNA 단일 가닥을 생성하는 것을 중심으로 이루어졌다[6-8]. 그러나 이러한 대부분의 연구들은 연산에 참여하는 DNA 가닥들만 존재하는 환경을 가정한 상태에서 이루어진 것이기 때문에 실제 임상환경과 관련된 응용에 바로 사용하기에는 무리가 있다. 예를 들어 실제 세포에서 추출된 DNA나 RNA를 입력으로 사용하는 정보처리에서는 계산에 참여하지 않는 일종의 배경에 불과한 분자들이 함께 존재한다. 본 논문에서는 이러한 조건에서 어느 DNA 혹은 RNA 분자와도 결합하지 않는, 다시 말해서 태그로 사용할 수 있는 올리고뉴클레오타이드 서열을 디자인 하는 방법을 제시한다. 본 방법은 간단한 유전 알고리즘을 기반으로 하고 있으며, 이를 이용해서 올리고뉴크레오타이드 태그를 효과적으로 디자인 할 수 있음을 보이고, 디자인된 태그의 염기 서열을 함께 제시한다.

본 논문은 다음과 같이 구성되어있다. 다음 장에서는 범용 올리고뉴크레오타이드 태그의 정의와 문제를 정의하고, 3장에서는 이를 디자인 하기위한 유전 알고리즘을 제시한다. 그리고 4장에서 실험 결과를 제시하고, 마지막으로 5장에서 결론과 함께 앞으로의 보완 사항과 추후에 이루어질 추가적인 연구에 대해서 언급한다.

2. 범용 올리고뉴클레오타이드 태그

'범용 올리고뉴클레오타이드 태그'를 정의하기 전에, 일반적인 올리고 뉴클레오타이드 프로브의 디자인에 대해서 살펴보자. 프로브는 특정 DNA/RNA 단일 가닥과 특이적으로 결합하는 DNA/RNA가닥을 뜻한다. 예를 들어 유전자 발현 패턴이나 유전자 질병 진단에 광범위하게 이용되는 마이크로어레이에 올라가있는 DNA 가닥들은 각각 특정 유전자 (DNA/RNA)에 특이적으로 결합하는 성질이 있다. 이러한 프로브를 디자인 할 때는 목표로 하는 유전자에 대한 특이적인 결합하도록 해야할 뿐만, 아니라 목표가 아닌 유전자 혹은 뉴클레오타이드 분자에 결합하지 않도록 해야 한다. 이때 전자를 '특이성'이라고 하고 후자를 '교차 상동' (cross homology)이라고 한다.

이러한 프로브의 디자인은, 목표 유전자에 대해 상보적인 서열에서 부분서열을 골라내되, 다른 비 목적 유전자에 대해서 특이성을 갖지 않는 부분서열을 골라내는 과정으로 생각할 수 있다. 따라서 원하는 프로브의 길이와 녹는점이 결정되어있다면, 목표 유전자에 대해 상보적인 서열로부터 주어진 녹는점과 길이 조건을 만족하는 가능한 모든 부분 서열 가운데서, 비 목표 유전자에 대해 결합력이 가장 약한 부분 서열을 골라냄으로써 간단히 프로브를 디자인 할 수 있다[5].

한편 프로브와 반대되는 개념으로서 '태그'를 생각해 볼 수 있다. 다시 말해서 목적 유전자가 없이 어떠한 유전자와도 결합하지 않는 염기서열로서 올리고머 길이에 해당하는 태그를 '올리고뉴클레오타이드 태그'라고 정의하자. 이와 같은 태그 가닥은 기존의 프로브와 함께 사용되어 여러 프로브들을 필요에 그룹 짓는 용도로 사용될 수 있다. 각각의 프로브들은 해당 목적 유전자에 특이성을 가져야하므로 각각 다른 서열을 가질 수밖에 없다. 그런데 만일 이러한 프로브들에 태그 가닥을 연결해서 제작을 한다면 같은 태그를 가진 서로 다른 프로브들을 모아서 처리하는 것이 가능해진다. 예를 들어 임상 세포에서 RNA를 추출하여 특정 유전자들의 발현 강도의 합과 같은 유전자 발현 정보를 이용한 연산과 같은 실제 임상 환경에 보다 가까운 DNA 컴퓨팅을 수행하는데 사용 될 수 있다. (그림 1)

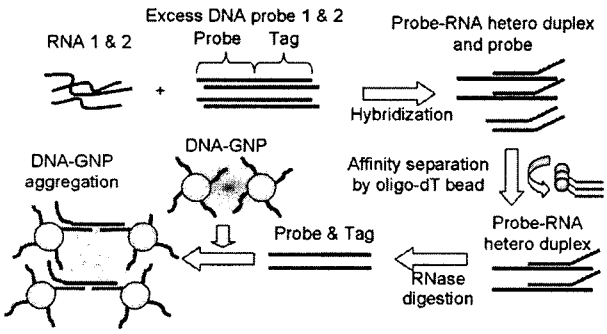

그림 1. 범용 올리고뉴클레오타이드 태그의 이용 예: 두 유전자에 대해 각각 프로브를 제작하고 각각의 프로브에 동일한 태그를 붙인다. 이러한 공통된 태그는 DNA Gold-Nanoparticle (DNA-GNP) [9]들을 뭉치도록 하는 일종의 브릿지 역할을 하게 된다. 결과적으로 두 유전자의 발현 양에 따라 태그 가닥의 양이 결정되고, DNA-GNP의 뭉침의 정도에 따라서 발색반응의 강도가 달라진다.

기존의 프로브의 디자인은 목적 유전자의 염기서열로부터 가능한 모든 부분 서열들을 검사한 후 선택함으로써 간단하게 수행될 수 있다. 그러나 태그의 경우 검사를 수행할 후보가 전혀 존재하지 않기 때문에 후보 서열을 랜덤하게 생성하면서 그 품질, 즉 반응 용액 상에 존재하는 다른 염기서열들과의 교차 상동을 검사하여 디자인해야 한다. 또한 염기 서열의 열역학적인 특성상, 교차 상동이 비교적 적은 염기 서열로부터 더 나은 염기 서열을 생성할 수 있는 가능성이 높다는 점에 착안하여, 유전 알고리즘을 이용해서 태그 서열을 디자인하는 것을 자연스럽게 생각해볼 수 있다.

3. 알고리즘

기본적으로 본 알고리즘은 단일 적합도를 가지는 간단한 유전 알고리즘을 채용하였으며 한번에 한 개의 자식만을 생성하는 steady-state GA를 사용하였다. Pseudo 코드는 다음과 같다.

```
N ← 최대 반복 횟수
P ← Tm, 길이 조건을 만족하는 랜덤한 초기 해집단
for i=1~N
        s1, s2 ← Selection(P) where s1≠s2
        offspring ← Crossover(s1, s2)
        while (offspring이 Tm 조건을 만족 못함)
                offspring ← Mutation(offspring)
        If Fitness(offspring) > Fitness(Worst(P))
        then P = (P - Worst(P)) U {offsping}
```

염기 서열이 직접적으로 각각의 해를 표현하도록 하였고, 태그 디자인을 위해 주어지는 녹는점 (T_m)과 길이 조건을 제한 조건(constraint)으로 사용하여 해집단 내의 모든 원소들은 항상 이 조건을 만족하도록 하였다. 선택 방법은 순위 기반 방법(Rank based selection)을 사용했고, 이렇게 선택된 두개의 염기 서열을 균등 교차 (Uniform crossover) 시켜 새로운 서열 (offspring)을 생성해냈다. 이 새로운 서열에 대해 주어진 녹는점(T_m) 조건을 만족 할 때까지 점 변이 (Point mutation)을 적용했다. 그 후, 이 새롭게 생성된 염기 서열이 현재 해집단에 존재하는 최악의 염기서열보다 더 나을 경우 그것을 대체하도록 하는 elitism을 사용했다.

염기서열의 녹는점은 올리고머 길이의 뉴클레오타이드 서열에 잘 맞는 Neareast-neighbor 모델을 이용해서 계산했다 [10]. 그리고 염기서열의 적합도는 교차 상동을 피해야하는 염기 서열과의 가장 안정한 결합에 대한

free energy[1]로 하였으며 OligoArray [5]에서 사용한 OligoArrayAux를 이용해서 계산하였다.

4. 실험 결과

먼저 교차 상동을 피해야 하는 유전자가 한개만 있는 상황에서 태그의 디자인을 수행했다. 또한 새로운 서열을 만들어내는데 사용할 부모 서열의 선택 방법의 효용을 보기 위해 무작위 선택을 수행했을 경우와 순위 기반 선택을 비교하였다. 표 1에 제시된 바와 같이 두 경우 0에 가까운 free energy를 갖는 서열을 생성함을 알 수 있지만, 순위 기반 선택을 이용한 경우 더 안정적인 성능을 보임을 알 수 있다. 그리고 품질에 관계없이 균등 선택을 통해서 교차연산을 수행했음에도 불구하고 좋은 품질의 결과를 내는 것은 교체 과정에서 elitism을 사용했기 때문으로 생각할 수 있다.

Free energy	평균	표준편차
균등 선택	-3.9836	0.1871
순위 기반 선택	-3.6586	0.1012

표 1. 균등 선택 vs. 순위 기반 선택: 각각 총 시행횟수 14번, 해집단 크기 500, 세대수 20000

두 번째로, 교차 상동을 피해야하는 염기서열이 두개가 존재할 경우에 대해서 태그의 디자인을 수행했다. 다음 표 2에 수행 결과의 예가 제시되어 있는데 한 번의 수행에서 생성된 해집단에서 상위 3개에 해당하는 서열을 보여준다.

염기 서열	T_m (℃)	free energy
1 ACTTACGTCACCCACACTCG	57.2	-5.53 -5.51
2 GCGAGAATTTCGCACACACT	57.0	-5.30 -5.73
3 CGGAACTGATCACCCACACT	57.2	-5.59 -5.77

표 2. 교차 상동 고려 대상이 2개일 때. 해집단 크기 500, 세대수 20000

5. 결론 및 추가 연구

지금까지 유전 알고리즘을 이용한 올리고뉴클레오타이드 태그의 디자인 방법을 기술 했으며, 태그를 성공적으로 디자인 할 수 있음을 실험 결과를 통해서 보였다., 추후에는 각 태그의 이차구조와 여러 태그를 동시에 디자인 할 경우 태그들 간의 교차 상동까지도 고려할 예정이다. 또한 사용자 편이성을 위한 유저 인터페이스의 추가도 이루어져야 할 것이다.

[1] free energy (kcal/mol)는 화학구조물의 안정상태를 나타내는 지표로서 그 구조물이 만들어지는 반응에서의 에너지 변화를 나타내며, 그 값이 작을수록 두 염기서열 간의 결합도가 높음(화학적으로 안정함)을 나타낸다.

Acknowledgement

본 연구는 교육인적자언부의 BK21-IT 프로그램, 산업자원부의 Molecular Evolutionary Computing (MEC) 프로젝트로부터 지원받았으며, 서울대학교의 컴퓨터연구소가 연구에 필요한 기자재를 제공하였습니다.

참고 문헌

1. L. M. Adleman, Molecularcomputation of solutions to combinatorial problems, *Science,* vol. 266, pp. 1201-1204, 1994
2. C. C. Maley, DNA computation: theory, practice, and prospects, *Evol. Comput.*, vol. 6, no. 3, p. 201-229, 1998
3. M. H. Garzon and R. J. Deaton, Biomolecular computation in the US, *New Generation Comput.* vol. 20, no. 3, pp. 217-236, 2002
4. Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, An autonomous molecular computer for logical control of gene expression. *Nature* vol. 429, pp. 423-429, 2004
5. J.-M. Rouillard, M. Zuker, and E. Gulari, OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach, *Nucleic Acids Research*, vol. 31, no. 12, pp. 3057-3062, 2003
6. S.-Y. Shin, I.-H. Lee, D.-M. Kim, and B.-T. Zhang, Multiobjective Evolutionary optimization of DNA sequences for reliable DNA computing, *IEEE Trans. of Evolutionary Computations,* vol 9, no. 2, 2005
7. M. Arita, A. Nishikawa, M. Hagiya, K. Komiya, H. Gouzu, and K. Sakamoto, Improving sequence design for DNA computing, *Proc. Genetic Evol. Comput. Conf. (GECCO)* pp. 875-882, 2000
8. R. M. Dirks, M. Lin, E. Winfree, and N. A. Pierce, Paradigms for computational nucleic acid design, *Nucleic Acids Research*, vol. 32, no. 4, pp. 1392-1403, 2004
9. J. J. Storhoff, A. D. Lucas, V. Garimella, and U. R. Muller, Homogeneous detection of unamplified enomic DNA sequences based on colormetric scatter of gold nanoparticle probes, *Nat. Biotech.* vol. 22, no. 7, pp. 883-887, 2004
10. J. Santalucia Jr. and D. Hicks, The thermodynamics of DNA structural motifs, *Annu. Rev. Biophys. Biomol. Struct.* vol. 33, pp 415-440, 2004