# 효과적인 프로브 설계를 위한 핵산 이차구조 예측

장하영[0] 장병탁
서울대학교 공과대학 컴퓨터공학부 바이오지능 연구실
{hyjang[0], btzhang}@bi.snu.ac.kr

# DNA secondary structure prediction for effective probe design

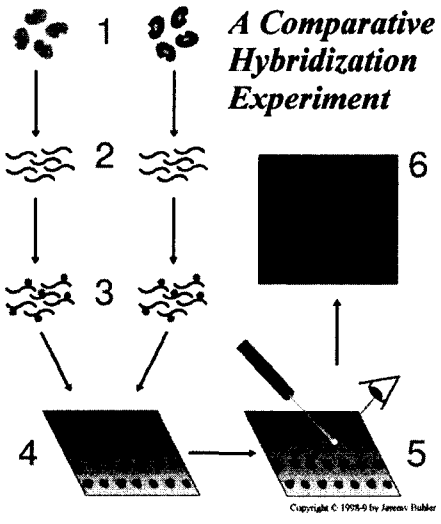
Ha-Young Jang[0] Byoung-Tak Zhang
BioIntelligence Lab. School of Computer Science and Engineering, Seoul National University

**요 약**

DNA 칩에서 사용되는 프로브를 가장 효과적으로 설계하기 위해서는 상보결합을 위한 1차구조뿐만이 아니라 열역학적인 움직임과 함께 2차구조가 고려되어야만 한다. 그러나 핵산의 기능에 큰 영향을 미치는 2차구조에 대한 연구는 일찍부터 진행되어 왔지만, 상대적으로 DNA에 대한 연구는 크게 미흡한 것이 현실이다. 이에 우리는 유전자 알고리즘을 이용한 핵산의 이차구조 예측을 통해서 보다 효과적인 프로브의 설계를 위한 방법을 고안했다.

## 1. 서 론

DNA는 아데닌과 티민, 구아닌과 시토신 간의 강하고 선택적인 결합으로 인해 이중나선 구조를 형성하는 고유한 특징을 가지고 있다. DNA 칩은 바로 이러한 DNA의 성질을 활용한 것이다. 일반적인 과정은 매우 간단하다. '프로브'라 불리우는 대개는 24머의 길이를 가지는 특정한 염기 서열의 DNA 가닥이 칩 위의 특정 부분에 고정시킨다. 그 후 검색하고자 하는 DNA 시료를 칩 위에 뿌려주면, 상보적인 서열을 가진 DNA만이 프로브와 함께 상보 결합을 이루게 되는 것이다.[1] 이 후 그림 1에서 보여지는 것처럼 그 결합 여부를 형광을 이용하여 확인하게 된다.




<그림 1> DNA 칩

일반적으로 사용되는 프로브들의 길이는 경제적, 실험적인 이유로 인해서 검출하고자 하는 DNA의 길이에 비해서 매우 짧기 때문에 목표가 되는 DNA의 모든 상보적인 서열을 담고 있을 수는 없다. 그렇기 때문에 한정된 물리적 공간 내에 보다 많은 DNA를 검출할 수 있는 프로브들을 효율적으로 담기 위해서는 이 프로브에 대한 효과적인 설계가 매우 중요하게 된다. DNA 칩상에서 일어나는 상보 결합은 기본적으로 목표가 되는 DNA의 1차 구조에 기반해서 일어나게 되지만, 이에 부가해서 DNA가 가질 수 있는 열역학적인 특성과 1차 구조에 기반해 발생하는 2차 구조에 의한 영향도 매우 큰 영향을 미치게 된다.

그러나 많은 프로브 설계 방법들은 1차 구조만을 고려하고 있을 뿐, 아직까지도 확실히 밝혀지지 않고 있는 DNA 2차 구조까지를 고려하고 있지는 못하다. 이에 우리는 보다 효율적인 프로브 설계를 위해서 유전자 알고리즘에 기반한 DNA의 2차 구조 예측 방법을 제안하고자 한다.

## 2. Probe 설계

DNA 칩에서 사용되는 프로브의 개수는 칩의 유전자 검색 성능뿐만 아니라 DNA 칩 개발의 난이도와 경제적인 측면에까지 직접적인 영향을 미치기 때문에, 보다 적은 개수의 프로브를 사용하여 DNA 칩을 설계하는 것은 매우 중요한 문제이다.

프로브의 설계를 위해서는 가장 먼저 목표로 하는 많은 유전자들 각각에 존재하는 유일한 서열을 찾아 각각의 프로브들이 특정한 유전자하고만 결합할 수 있도록 프로브의 염기서열을 결정해야만 한다. 이와 동시에 이중 나선 구조를 형성하고 있는 DNA가 두가닥으로 분리되는 온도인 melting temperature를 모든 probe 간에 일정하게 만들어줘야 하는데, melting temperature는 서열내의 아데닌, 티민, 구아닌과 시토신의 조성에 의해 결정되기 때문에 이를 고려해서 프로브의 서열을 결정해야 한다.[2] 이처럼 프로브의 설계는 동시에 여러 가지 요

소를 고려해주어야만 하고, 또한 이러한 각각의 요소들은 서로 영향을 주고 받기 때문에 계산상으로 많은 어려움을 가지고 있는 것이 사실이다.

### 3. GA 기반 이차구조 예측

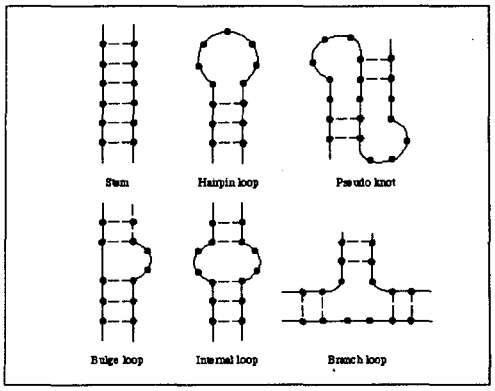
위에서 언급한 프로브 설계를 위한 요소들 외에도 부가적으로 고려할 수 있는 것이 프로브가 가지는 2차구조의 안정성이다.[2] 생체 내에서 핵산의 기능은 그 핵산이 가지는 이차구조와 직접적으로 연관되어 있기 때문에, 핵산의 이차구조 예측을 위한 연구는 RNA의 이차구조를 중심으로 1960년대부터 활발하게 진행되어 왔다. 현재 RNA의 이차구조를 예측하는 방법은 크게 두가지로 분류될 수 있는데, 비슷한 계통의 RNA는 비슷한 형태의 이차구조를 가지게 될 것이라는 가정 하에 발생학적인 분석을 하는 방법과 열역학적인 파라메터들에 기반한 자유 에너지를 최소화 하는 방향으로 이차구조를 형성하는 성질에 기반한 자유 에너지 방법이 있다.[3]

그러나 이와 반대로 DNA의 경우에는 아직도 그다지 많은 연구가 이루어지지 못하고 있는 것이 현실이기 때문에 기존의 연구 방법을 그대로 사용하는 것에는 다소 무리가 있다. 비록 DNA와 RNA의 기본적인 구성은 티민과 우라실의 차이밖에 존재하지 않고 대부분의 특성은 거의 동일하다고 생각할 수도 있지만, 기존의 RNA 구조 연구에 의해 제안된 방법론에 있어서도 어떠한 에너지 모델을 사용하느냐에 따라 서로 상이한 RNA의 이차 구조가 예측되는 상황에서 RNA의 에너지 모델로 제시된 내용들을 그대로 DNA에 적용시키는 것은 그다지 올바른 판단은 아닐 것이다. 또한 DNA의 경우 발생학적인 분류 또한 아직까지는 그다지 많은 연구가 이루어지지 못하고 있기 때문에 이에 기반한 방법 또한 사용하기가 여의치 않다.

본 논문에서는 문제의 특성을 최대한 고려하여 이에 대한 해결방법으로 상보결합을 최대로 하는 2차 구조의 선택을 목표로 하는 유전자 알고리즘을 선택했다. 이 방법은 기본적으로는 자유 에너지 최소화 방법과 같은 맥락에서 이해될 수 있다. 자유 에너지를 최소화 시키는데 있어서 가장 많은 부분을 역할을 하는 것이 바로 상보 결합을 통한 2중 나선 구조의 형성이고, 현재 사용되고 있는 대부분의 자유 에너지 모델들 또한 이를 기반으로하여 실험적인 과정을 통해서 얻어진 각종 열역학적인 파라메터들을 이용하여 만들어진 것이기 때문이다.[4]

물론 이러한 단순화 시킨 모델로 DNA의 정확한 구조를 예측하는 것은 사실상 불가능하다. 그러나 기술한 바와 마찬가지로 DNA의 이차구조가 프로브의 효율성에 있어서 중요한 역할을 하고 있음은 결코 간과할 수 없는 사실이지만, 이것은 어디까지나 부가적인 요구사항일 뿐이며 이러한 방법만으로도 최악의 경우 자기 스스로 상보적인 결합을 일으키게 되는 문제점등은 해결이 가능하다는 점에서 우리가 선택한 해결 방안은 문제의 특성과 가장 잘 부합한다고 볼 수 있다.

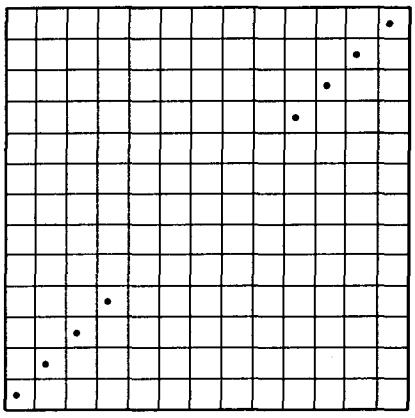
그림 2에서 보여지고 있는 바와 같이 기본적인 핵산의 2차구조들은 방향성이 없는 그래프로 생각할 수 있다. 따라서 핵산의 2차 구조는 모든 노드들이 다른 노드(상보적인 노드)와 최대 1개의 연결만을 가질 수 있는 그래프로 모델링 될 수 있다. 따라서 그림 2에서 보여지는 Hairpin loop과 같은 형태의 이차구조는 그림 3와 같은 표현형으로 나타나게 된다. 즉, 각각의 노드는 행과 열의 인덱스에서 위치하면서 자신이 상보적으로 결합하는 노드와의 관계를 희소행렬(sparse matrix)로 나타나게 되는 것이다. 이런 경우에 필요한 정보는 삼각행렬만으로도 충분히 표현이 가능하지만, 교차연산이나 돌연변이 등의 GA 연산의 효율적인 수행을 위해서는 삼각행렬이 아닌 정방행렬을 사용하는 것이 더욱 바람직하다.




<그림 2> 핵산의 이차구조

이러한 표현형을 사용하게 될 경우 그래프가 가지는 제약 조건은 그림 3에서도 보여지는 것처럼 다음과 같은 형태로 나타나게 된다.

1. 대각 행렬을 기준으로 한 대칭형.
2. 동일한 행이나 열에는 하나 이하의 연결만이 존재.



<그림 3> Hairpin Loop의 표현형

위와 같은 제약조건들을 가지고 생성된 초기해 들은 각각의 연결들에 점수를 매겨서 이를 최대화 하는 방향으로 진화하게 된다. 이러한 경우에 목적함수는 다음과

같이 간단히 정의가 가능하다.

F = 연결의 개수

문제의 목적이 가장 효율적인 프로브를 찾아내는 것임에도 불구하고, 연결이 가장 많은(즉, 자체적으로 상보 결합이 일어나 프로브로서의 성능이 떨어지는) 방향으로 진화를 시키는 것은 평가 함수 자체가 정확한 구조를 평가하는데는 한계가 있기 때문에 1차 구조의 특성을 보완하는 2차 구조로서의 역할을 최대한 활용해서 주어진 서열에서 가능한 가장 나쁜 구조를 찾아내기 위해서이다. 다시 말하면 프로브 설계에서 그 효과를 결정하는 대부분의 요소는 1차구조가 될 수 밖에 없지만, 여기에 부가적으로 2차구조에 대한 정보를 동시에 사용함으로써 좀 더 효과적인 프로브를 설계하고자 하는 것이다.

물론 이러한 베이스 페어를 최대화 시키는 연결을 찾아내는 방법은 다이나믹 프로그래밍 기법을 사용하여 $O(n^3)$에 해결될 수 있는 간단한 문제이다. 그러나 우리가 제시한 방법의 장점은 기존의 방법에서 가지고 있던 많은 제약점 들을 해결할 수 있다는 것이다. 예를 들면, 그림 1에서 보여지고 있는 pseudo knot의 경우에 대부분의 에너지 모델에서는 이를 고려하지 못하고 있고[5] 확률적 문맥자유 문법을 이용해 RNA의 이차구조를 모델링하는 방법에서도 이의 해결을 위한 별도의 문법 규칙을 필요로 하는 어려움이 존재하는 반면에[3] 위와 같은 방법을 사용할 경우에는 핵산이 가질 수 있는 모든 구조를 동시에 고려할 수 있다는 장점이 있다.

## 4. 결론 및 차후 연구방향

한정된 물리적 공간에서 보다 많은 일을 수행하기 위한 DNA 칩의 제작을 위해서는 주어진 작업에 가장 효과적인 프로브를 선택하는 것이 무엇보다 중요하고, 이를 위해서는 프로브의 1차 구조 뿐만이 아니라 2차 구조까지도 고려해 주는 것이 필요하다. 기존의 프로브 설계를 위해 고안된 기법들은 단순히 1차 구조만을 사용하고 있었던 것에 반해 2차 구조까지도 동시에 고려해 줄 수 있다면 좀 더 효과적인 프로브의 설계가 가능할 것이다.

그러나 프로브의 가장자리 부분보다는 가운데 부분에 결합할 가능성이 더욱 많다는 Tobler등[6]의 연구 결과에서도 알 수 있듯이, 단순히 수용액 속을 부유하고 있는 DNA 가닥과 칩에 고정되어 있는 프로브의 경우에는 열역학적인 매개변수들에 의한 움직임이 다를 수 밖에 없기 때문에 단순히 염기서열 내에서 상보적인 결합이 일어나는 경우만을 고려하는 것이 아니라 이러한 문제에 대한 추가적인 고려가 필요할 것이다. 또한 본 연구에서는 모든 결합에 대해서 동일한 가중치를 가정하였지만, 이보다는 각각의 결합이 가질 수 있는 확률적인 가중치에 대한 좀 더 면밀한 고려를 통해서 각각의 연결에 대한 가중치의 변화를 정하는 것이 바람직 할 것이다. 이를 위해서는 RNA에서 사용되고 있는 것과 같은 에너지 모델의 구축이 반드시 필요하다.

이러한 에너지 모델의 구축을 위해서는 RNA에서 사용되었던 기법들을 상당 부분 차용할 수 도 있지만, 아직까지도 일부 2차구조에 대해서는 RNA에서도 그다지 효율적인 에너지 모델이 존재하지 못하고 있다는 사실등으로 미루어 볼 때, DNA의 경우에 적용될 수 있는 에너지 모델을 구축하는 것에는 많은 어려움이 있을 것이라 예상되지만, 좀 더 정확한 2차구조의 예측을 위해서는 반드시 해결되어야만 할 문제일 것이다.

감사의 글: 본 연구는 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정 연구실 과제에 의해 지원되었음.

## 5. 참고문헌

[1] Snustad, D.P. and Simmons, M.J., *Principles of Genetics*, 2nd Edition, Wiley, New York, 1999.

[2] Kurata, K. and Suyama, A., Probe Design for DNA Chips, *Genome Informatics*, 10:225-226, 1999.

[3] Sakakibara, Y., M. Brown, R. Underwood, I. S. Mian, and D. Haussler. Stochastic context-free grammars for modeling RNA. Technical Report UCSE-CRL-93-16, University of California at Santa Cruz, Computer Science, UCSantaCruz, CA95064, 1993.

[4] Durbin, R., Eddy, S., and Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 1st Edition, Cambridge University Press, 1997.

[5] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591-621, 1984.

[6] J. Tobler, M. Molla, E. Nuwaysir, R. Green & J. Shavlik. Evaluating Machine Learning Approaches for Aiding Probe Selection for Gene-Expression Arrays. *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology*, Vol. 18 Suppls. 1, Pages S164-S171, 2002.