

유전알고리즘을 이용한 MicroRNA 이중나선 구조의 Dinucleotide Step 파라미터 추정

남진우^{0,1,2} 장병탁^{1,2,3}

서울대학교 대학원 생물정보학 협동과정¹
서울대학교 바이오정보기술 연구센터(CBIT)²
서울대학교 컴퓨터 공학부 바이오지능 연구실³
{jwnam, btzhang}@bi.snu.ac.kr

Estimation of Dinucleotide Step Parameters of Double Helix Structure on MicroRNA Using Genetic Algorithms

Jin-Wu Nam^{0,1,2} Byoung-Tak Zhang^{1,2,3}
Graduate Program in Bioinformatics¹

Center for Bioinformation Technology (CBIT)²
Biointelligence Laboratory, School of Computer Science and Engineering³
.Seoul National University, Seoul 151-742, Korea

요 약

MicroRNA는 약 22 nucleotide 길이로, 세포질에서 유전자의 전사 후 조절 기능을 맡는 small RNA의 한 종류이다. MiRNA는 긴 전사체인 pri-miRNA에서 Drosha에 의해 절단 되어 핵 밖으로 나가 최종 Dicer에 의해 성숙 된다. 하지만, 아직까지 이 효소들이 pri-miRNA를 잘라내는 3차 구조상의 메커니즘을 이해하지 못하고 있다. 본 연구에서는 완숙한 miRNA 이중나선이 약 2 회전을 이루게 된다는 정보를 바탕으로, Drosha가 붙는 miRNA stem구조의 dinucleotide step 파라미터를 유전 알고리즘을 이용하여 추정한다. 추정된 파라미터로부터 실제 miRNA들의 3차구조를 예측할 수 있으며, 3차 구조상의 Drosha의 절단 메커니즘을 이해하는데 도움을 줄 것이다.

1. 서 론

MicroRNA (miRNA)는 18~23 nucleotide (nt) 범위의 길이를 갖는 small non-coding (nc) RNA의 한 종류이다 [1]. MiRNA는 RNA polymerase II에 의해서 전사되며, 그 전사체를 primary miRNA (pri-miRNA)라 지칭한다 [2]. Pri-miRNA는 핵 내의 RNase III 형태의 효소인 Drosha에 의해서 non-looped 말단이 절단되어 pre-miRNA로 만들어 진다 [2,3]. 이 pre-miRNA는 세포질로 나와 또 다른 RNase III 타입 효소인 Dicer에 의해 looped 말단이 잘리게 된다 [3]. 최근 Drosha에 의해서 절단되는 위치가 loop 경계로부터 약 2-turn helix 떨어져 있고, non-looped의 단일 가닥 서열 경계에서 약 1-turn helix 위치라는 사실을 보고 했다 [4]. 하지만 아직까지 3차 구조상의 절단 메커니즘은 정확히 알려지지 않고 있다.

RNA 이중나선의 3차 구조는 RNA에 작용하는, 단백질이 인식하기 위한 구조적인 모티프가 존재한다고 예상할 수 있다. 하지만 RNA 이중나선 구조는 DNA 이중나선 구조에 비해 많이 알려지지 않고 있다. 다양한 형태의

DNA에서 결정화 방법을 이용한 실험 결과에서, 이중나선 구조와 그 파라미터들이 보고되었다 [5, 6]. 이 결과는 DNA 이중나선 구조에 따라 특이하게 작용하는 전사 조절인자들의 결합 메커니즘을 이해하는데 도움을 주었다. 이와 유사하게 이중나선 구조를 갖고 있는 pri-miRNA의 stem의 3차 구조를 밝히는 것은, Drosha 절단 메커니즘을 이해하는데 도움을 줄 것이다.

본 연구에서는 유전알고리즘을 이용하여 miRNA의 이중나선에 존재하는 dinucleotide step의 파라미터를 추정하고 이를 바탕으로 miRNA의 이중나선구조를 예측한다.

2. RNA 3차 구조

2.1 RNA dinucleotide steps

RNA의 dinucleotide step은 염기인 A, G (purine), U, G (pyrimidine)의 염기 step에 의해 결정된다. RNA에는 5'→3' 방향으로 10개의 dinucleotide step이 존재하고, 각 step은 반대쪽 가닥의 base step과 상보 결합을 하게 된다. 이 dinucleotide step은 염기의 기하학상의 경향 또는, 영향력에 따라 아래와 같이 크게 세 가지 클래스로 나눌 수 있다.

- (1) Pyrimidine (pyr) / Purine (Pur) steps
UA, CA(UG), CG
- (2) Pyr /Pyr (Pur /Pur) steps
AA(UU), AG(CU), GA(UC), GG(CC)
- (3) Pur /Pyr steps
AU, GU(AC), GC(GC)

2.2 Dinucleotide step 파라미터

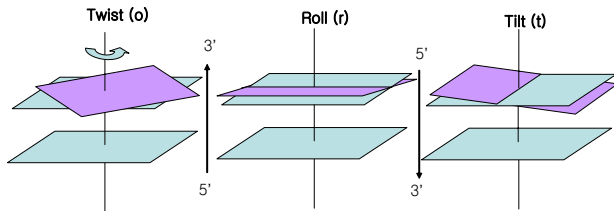


그림 1. 이중나선 구조의 각도 파라미터

RNA 이중나선 구조는 간단하게 본다면 세 가지 각도에 의해서 3차원 구조가 결정된다. 첫 번째, 'Twist' (o)는 나선의 각도를 의미하며 이에 의해 왼 나선, 오른 나선 구조가 결정 된다 (그림 1). 두 번째, 'Roll' (r)은 이중나선 구조를 앞뒤 (주 나선 홈 - 보조 나선 홈)로 구부러지는 회전각으로, 각 나선 홈의 dimer step을 열리게 한다. 세 번째 'Tilt' (t)는 좌우로 구부러지게 하는 파라미터로 사용 된다 (그림 1). 위 세 가지 파라미터에 의해 RNA 이중나선의 각 염기가 3차원상 위치하는 곳이 결정 된다. 이 각도 파라미터들은 기본적으로 RNA dimer step의 값으로 정의 되며, 이를 dinucleotide step 파라미터라 정의한다. 본 연구에서는 'Twist' 만을 추정하여 miRNA 이중나선에 대해 고찰한다.

3. 유전 알고리즘 구현

유전 알고리즘은 목적함수의 탐색공간에서 최적의 해를 찾기 위해, 생명체 진화 과정의 모델을 적용하여 병렬 탐색하는 학습 전략을 갖는다 [7]. 특히, 비선형 시스템의 실수 공간에서의 최적화 문제는 유전 알고리즘에 적합한 문제라고 할 수 있다.

3.1 염색체 설계

각 염색체는 총 15개의 실수 값의 1차 서열로 구성되며, 10개의 RNA dinucleotide step과 indel을 포함한 step 5개로 구성 된다 (그림 3). 각 파라미터는 각도 범위를 제한하고 있는데, 10개의 dimer step은 $\phi/9 \sim \phi/4$ 의 범위로 제한하고, indel의 step은 $\phi/100 \sim \phi/4$ 로 제한한다.

3.2 알고리즘

본 연구에서는 단순 유전알고리즘을 사용하였으며 알고리즘은 다음과 같다.

개체군의 초기화는 각 유전자들이 갖는 값의 범위 내에서 무작위 적으로 생성된다. selection은 elitism을 사용하며 (10%), 자연선택설 개념으로 두 개의 부모 중 좋은 적합도의 염색체를 각 유전 연산자에 적용하여 다음 세대에 전달한다. 교차는 균일하게 실행되며,

```

procedure GA()
  initialize(random generation);
  evaluate();
  while not (terminal condition satisfied) do
    selection(elitism);
    crossover(1%);
    mutation(15%);
    evaluate();
  end while
end procedure
    
```

돌연변이는 유전자의 실수 값의 정규분포에 따라 변화를 준다. 각 유전 연산자의 적용 확률은 1%와 15%이다. 진화의 종료 조건은 정해진 세대수이다.

3.3 적합도 함수

우선 적합도를 위해 우리는 세 가지 관측 값을 정의 한다. MDOP (Mean of Distance between pairs)는 총 N개의 mature miRNA data에 대해 mature miRNA stem 구조의 양쪽 strand (x, \bar{x})에 염색체 c의 dinucleotide step 파라미터를 적용하여 계산한 전체 회전각의 차이의 평균으로 계산된다. 이는 양쪽 가닥의 최종 회전 각도가 같다는 가정에서 출발한다. MOD (Mean of Distance)는 Drosha가 2회 회전하는 나선구조를 인지한다는 가정에서, mature miRNA의 전체 회전 각도가 4phi에서 떨어져 있는 차이에 대한 평균값으로 계산된다. SDD (Standard Deviation)는 각 차이에 대한 표준편차 값을 제공한다. 위 세 개의 식을 이용하여 우리는 적합도 (Fitness)를 위와 같이 계산한다.

$$MDOP = \frac{\sum_{i=1}^N (x_i^c - \bar{x}_i^c)}{N}$$

$$MOD = \frac{\sum_{i=1}^N (T - x_i^c)}{N}, (T = 4\pi \|\bar{x}\|)$$

$$SDD = \sqrt{\frac{\sum_{i=1}^N (MOD - (T - x_i^c))^2}{N}}$$

$$Fitness = \frac{1}{SDD^2} + \frac{1}{MDOP}$$

4. 실험 데이터 및 결과

4.1 실험 데이터 및 설정

본 실험에서는 northern blot을 통해 정확한 Drosha 절단 위치가 결정된 38개의 인간 pri-miRNA 서열을 사용한다 (<http://www.sanger.ac.uk/Software/Rfam>). 이 서열은 mfold 패키지 (<http://www.bioinfo.rpi.edu/applications/mfold>)를 이용하여 2차 구조를 예측한 후, mature miRNA 부분만을 선택하여 염기 쌍 형태로 작성하였다.

유전 알고리즘의 파라미터 설정으로 세대수는 200, 개체수는 2000으로 설정한다. 또한, 총 4번의 반복실험을 하여 파라미터와 적합도에 대해 평균과 표준편차로 결과를 표시한다.

4.2 적합도 결과

적합도 함수는 큰 값으로 최적화되며 4번의 실험결과에서 706 근처의 적합도 값으로 수렴했다 (그림 2). 대부분 100세대 이전에 대부분의 최적화가 이루어 졌음을

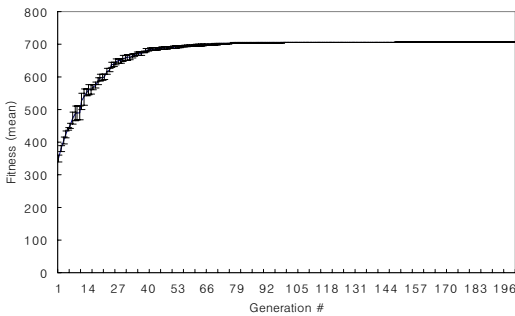


그림 2. 세대별 최적함도의 평균값과 표준편차 그래프

알 수 있다. 각 세대별 오차막대는 표준편차를 기준으로 설정되었다. 초기 최적화에는 SDD에 의해서 급격히 학습되며, 100세대 이후부터 MDOF에 의해서 최적화 되는 경향을 보여 주었다.

4.3 예측된 파라미터

(a)	MIRNA	Pur A	Pur G	Pyr T(U)	Pyr C	B
Pur A		31.16				
Pur G		29.38	38.99			
Pyr U		35.83	22.01	33.64		
Pyr C		40.83	35.62	37.7	39.2	
B		34.61	17.98	14.89	33.14	22.03

(b)	DNA+p	Pur A	Pur G	Pyr T(U)	Pyr C	B
Pur A		37.8				
Pur G		37.3	36.1			
Pyr T		35.1	31.9	29.3		
Pyr C		36.3	32.9	31.5	33.6	
B						

그림 3. (a) 추정된 Twist 파라미터; (b) DNA Twist 파라미터

그림 3a는 유전알고리즘을 통해 추정된 dinucleotide step의 회전 각도를 보여주고 있다. 추정된 회전 각도는 pri-miRNA에 Drosha 등의 단백질이 붙어 있는 상태의 파라미터임을 주의해야 한다. pri-miRNA가 고유하게 갖고 있는 나선 구조는 Drosha 등의 단백질과의 상호작용을 통해 구조상 변형을 일으킴으로써, 정확한 위치의 절단을 유도하게 될 것이다. 그림 3b는 실험을 통해 규명된 DNA dinucleotide 파라미터에 대한 값을 보여주고 있다. 이 테이블의 값은 상호 대칭이다.

4.4 miRNA 3차 구조

마지막으로 최적화된 dimer step의 회전 각도를 이용해 학습데이터 38개에 대해 나선 구조를 그려 보았다 (그림 4). 그림 4는 sin(x) 함수를 이용하여 파동으로 나선 회전을 나타냈었다. 결과에서 보듯이 나선이 2 회전을 하고 있으며, 많은 데이터의 끝나는 지점이 유사함을 알 수 있다. 하지만 몇몇 데이터의 경우 2회전을 넘거나 모

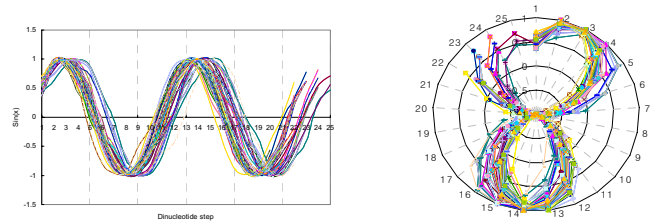


그림 4. 추정된 파라미터를 이용해 예측한 miRNA 이중나선 구조, 왼쪽은 sin 함수를 이용한 그래프이고, 오른쪽은 sin 함수 값을 방사형으로 그린 그래프

자라는 경우가 있는데, 이것은 적합도 함수의 문제이거나, 데이터 자체의 노이즈라고 생각 할 수 있다. 실제로 miRNA cloning에서 1~2 nucleotide 추가가 종종 일어나기도 한다. 이는 유전알고리즘을 이용한 dinucleotide step 파라미터의 추정 값이 어느 정도 신뢰 할 수 있음을 의미한다.

5. 결론

유전알고리즘을 통해 추정된 dinucleotide step의 Twist 값들은 아직 정확한 3차 구조 정보를 준다고 할 수 없을 것이다. 앞에서 언급한 Roll과 Tilt에 대해서 같은 방법으로 추정을 함으로써, 현실적인 pri-miRNA 이중나선 구조를 완성할 수 있을 것이다. 이렇게 완성된 이중나선의 3차 구조는 아직 까지 완벽히 이해하지 못한 miRNA의 성숙 과정과 유전자 조절 과정의 이해에 대한 새로운 관점을 제시 할 수 있을 것이다.

감사의 글

이 논문은 교육부 BK21 사업, 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음.

참고문헌

- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116:281-297, 2004.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425:415-419, 2003.
- Kim VN. Small RNAs: Classification, Biogenesis, and Function. *Mol Cells*, 19:1-15, 2005.
- Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *Embo J*, 24:138-148, 2005.
- Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M. and Zhurkin, B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *PNAS*, 95:11163-11168, 1998.
- Strahs, D. and Schlick, T. A-Tract Bending: Insight into Experimental Structures by Computational Models. *J. Mol. Biol.*, 301:643-663, 2000.
- Goldberg, D.E (1989). Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley Longman.