유전자 프로그래밍을 이용한 RNA 구조 문법 학습

남진우[O,1,2] 정제균[1,2] 장병탁[1,2,3]
서울대학교 대학원 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터(CBIT)[2]
서울대학교 컴퓨터 공학부 바이오지능 연구실[3]
{jwnam[O], jgjoung, btzhang}@bi.snu.ac.kr

Learning of RNA Structural Grammar using Genetic Programming

Jin-Wu Nam[O,1,2] Je-Gun Joung[1,2] Byoung-Tak Zhang[1,2,3]
Interdisciplinary Program in Bioinformatics[1]
Center for Bioinformation Technology(CBIT), Seoul National University[2]
Biointelligence Laboratory, School of Computer Science and Engineering[3]
.Seoul National University, Seoul 151-742, Korea

요 약

RNA는 세포내에서 유전자 발현에 직, 간접적으로 중요한 역할을 하며, RNA 구조는 세포 내에서의 기능과 깊은 연관이 있기 때문에 RNA 구조를 예측하는 것은 중요한 의미를 갖는다. 본 논문에서는 진화연산의 한가지인 유전자 프로그래밍(genetic programming) 방법을 사용하여 염기서열 정보를 참고하는 RNA 구조 문법의 학습 방법을 보여 준다. 이 RNA 구조를 의미하는 문법을 트리(tree)형태의 함수로 코드화(encoding) 한 후 이것을 유전자 프로그래밍 방법으로 진화시킨다. 진화를 통해 최적의 적합도를 갖는 트리의 문법을 테스트 데이터를 통해 평가한 결과 0.893의 특이도(speicificity)와 0.752의 민감도(sensitivity)를 보였다.

1. 서 론

세포 내에서 RNA 구조는 염기 서열과 2, 3차 모티프의 조합에 의해서 결정되는데, 특히 mRNA의 발현과 안정성에 중요한 역할을 하는 ncRNA의 2, 3차 RNA 구조간의 상호 작용은 그 기능과 밀접한 관련을 가지고 있다[1][2]. 이러한 RNA 2, 3차 구조의 중요한 역할 때문에 RNA의 중요 모티프는 유전적으로 그 구조가 보존되어 왔다. RNA 구조의 예측을 위해 여러 가지 방법론이 연구되었는데, 동적 알고리즘을 이용한 연구[3]와 유전 알고리즘을 이용한 연구[4]가 대표적이다.

한편, 유사한 구조와 서열을 가진 RNA를 검색하기 위한 알고리즘도 연구되어 왔다[5]. 유사 RNA 검색 알고리즘은 단순한 서열검색 알고리즘과는 달리 유사 2, 3차 구조를 갖는 RNA를 검색해야 하므로 RNA 구조를 문법화한 후 그것으로 유사 RNA를 검색하는 방법이 사용되고 있다. 그러나 문법에 기반을 둔 유사 RNA 검색은 엄격한 문법을 찾지 못한다면 잘못된 유사 구조 결과가 많이 생성되는 큰 한계를 갖고 있으며[6], 특히 엄격한 문법을 찾는 것은 상당히 어려운 작업이 될 것이다.

유전자 프로그래밍은 염기 서열의 2차 구조뿐 만 아니라 단백질 구조 탐색에도 사용되고 있으며 구조 예측과 분석 연구에 적합한 방법이다[7]. 이 논문에서는 문법을 기반으로 한 유사 RNA 검색 방법의 단점을 극복하기 위해 유전자 프로그래밍(genetic programming)을 통해 최적의 문법을 찾아내는 알고리즘에 관하여 논하고자 한다.

2. RNA 구조의 문법과 트리구조 코드화

2.1 RNA 구조의 문법

RNA 2차구조 중 가장 기본적인 형태가 머리핀(hairpin) 구조이다. 이 머리핀 구조[그림1(b)]는 염기쌍을 이루는 부분과 루프를 이루는 간단한 구조로 되어 있다. 이 구조를 문법화 하기 위해 5′쪽의 이중나선 구조로 되어 있는 염기쌍 부분 h5, 3’쪽을 h3 로 정하고 루프를 이루는 부분은 단일 가닥(single st

(a)

```
descr
h5(minlen=8, maxlen=16, mispair=1)
        ss ( len=7 )
h3
```

(b)

그림 1. (a) 머리핀 구조 문법 (b) 머리핀 구조

rand)으로 되어 있어 ss라 정했다. 또한 각각의 h5, h3, ss에 해당하는 염기의 길이와 비염기쌍(mispair)의 수를 정할 수 있게 문법을 정의했다[그림1(a)]. 몇몇의 RNA 검색 알고리즘은 이 문법을 이용하여 유사 RNA를 검색할 수 있으며, 우리는 그 중에 RNAmotif 프로그램[5]을 이용하여 검색을 하였다.

2.1 트리 구조 코드화(encoding)

RNA 구조를 위해 표시된 문법을 곧바로 트리로 표현하는데 몇 가지 어려운 점이 있다. 우선 염기쌍은 5′, 3’방향에서 동시에 존재하는 것이기 때문에 문법에서 h5와 h3의 수는 동일하게 나타나야 하며, 항상 h5는 h3 보다 선행되어 나와야 한다.

함수	문법	소속변수
f1	h5 (f1 or f2) h3	minlen/maxlen, len, mispair
f2	ss	minlen/maxlen, len
root	descr	

표 1. 함수를 이용한 문법의 코드화와 소속변수

또한 문법이 완전히 종료하기 전까지 h5개수는 h3개수보다 많거나 같아야 하며 ss는 동시에 나와서는 안 된다. 마지막으

로 각 구조의 길이를 나타내는 minlen/maxlen len mispair는 독자적으로 존재하지 못한다. 이것은 모두 문법에 맞지 않거나 생물학적으로 존재할 수 없기 때문이다. 이 트리의 구조적인 제약을 극복하기 위해 우리는 함수 f1과 f2를 도입하여 이것을 문법으로 정의하였다[표1]. 다음 이 함수를 이용하여 유전 프로그래밍의 변이(variation)에도 안정된 구조를 갖는 트리를 만들었다[그림2]. 이 트리를 이용한 유전자 프로그래밍은 말단의 입력 데이터를 받지 않으면서 구조에 기반을 둔 진화만을 계산하게 된다.

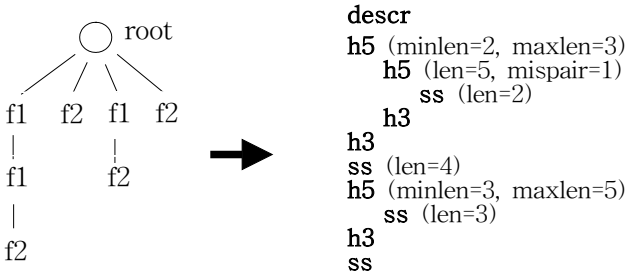


그림 2. 함수 트리의 문법 변환

3. 유전자 프로그래밍(genetic programming) 알고리즘

<유전자 프로그래밍 알고리즘>

1. 개체군을 초기화한다
2. 개체군에 속한 개체를 평가하고, 개체의 적합도를 반환한다
3. 새 개체군이 모두 생성될 때까지 다음을 반복한다
   3.1 최적 적합도 개체를 그대로 새 개체군에 첨가한다
   3.2 선택 알고리즘을 사용해 개체들을 선택한다
   3.3 선택된 개체는 변이 과정을 거친다
   3.4 새 개체군에 변이 과정의 결과를 첨가한다
4. 종료 조건에 도달하면 종료, 아니면 새 개체군을 기존 개체군과 바꾸어 주고, 2-4번 과정을 반복한다

초기화(initialization) 과정은 무작위로 300개의 트리를 생성하였으며, 머리핀 구조를 만들기 위해서는 비염기쌍이 있어야 하므로 f1 함수가 말단에 놓이지 않도록 제한사항을 두었으며, f2 함수가 연속되지 않도록 제한사항을 두었다. 선택과정(selection)에서는 모든 트리의 적합도(fitness) 값을 정렬하여 상위 50%를 부모로 선택(selection) 하였다. 여기서 적합도는

$$\text{Fitness} = \frac{(AH+1)}{(BH+1)} \times \left(\frac{AS}{BS}\right) - \text{Complexity} \qquad (1)$$

로 정의된다. (1)식에서 AH 와 BH는 각각 음성데이터와 양성데이터에 대한 RNAmotif의 검색 점수를 의미하며 +1은 분모가 0이 되는 것을 방지하기 위한 보정 값이다. AS/BS는 훈련과정 점수를 일반화시키기 위한 음성데이터와 양성데이터의 비율 값이다. 또 진화되는 RNA 구조의 복잡도를 안정하게 하기 위해 Complexity라는 복잡도 수치를 추가했다.

다음 변이 과정중 돌연변이(mutation)는 포아송(poisson) 분포에 따라 각 함수의 소속 변수 값을 변화시키며 0.2의 확률로 진행된다. 교차 연산(crossover)은 모든 개체에서 일어난다. 진화의 종료 조건은 30세대로 정했다.

4. 실험 데이터 및 결과

4.1 실험 데이터

tRNA는 구조가 정확히 알려져 있으며 많은 종에서 발견되었기 때문에 실험의 데이터로 좋은 조건을 가지고 있다. 본 연구에서는 초파리의 tRNA 염기서열을 훈련 집단과 테스트 집단으로 나누어 초파리 tRNA의 구조를 예측하였다. tRNA는 3개의 루프 구조를 갖고 있고 3′ 끝이 긴 형태를 가지고 있다. 대부분의 tRNA의 구조와 염기서열은 그림4(a)와 상당히 유사하지만, 종에 따라, 코돈(codon)의 종류에 따라 조금씩 다른 차이를 보이고 있다. 염기서열의 변이가 있는 경우 염기쌍의 변화로 인해 구조적인 변이가 생기게 되며, 염기서열의 길이 변화는 단일 가닥의 변화를 일으켜 전체적인 tRNA 구조의 변이를 가지고 오기 때문에 다양한 tRNA 구조가 존재하는 것을 설명해 준다.

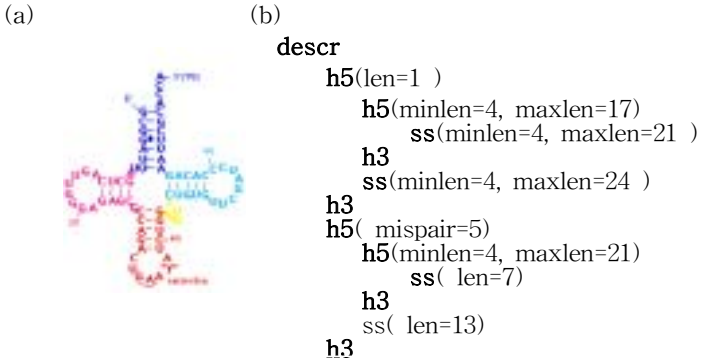


그림 3. (a) tRNA 일반적인 구조 (b) 30번째 세대의 최적 적합도 0.031224인 tRNA 문법

4.2 실험 결과 – 최적 적합도 문법

RNA 염기서열의 변이는 문법구조를 이용한 RNA 검색과 예측에 있어서 어려움을 주고 있다. 그러나 유전자 프로그래밍을 통해 진화시킨 그림3(b)의 문법은 우리가 찾고자 하는 tRNA의 통합된 구조의 형태를 가지고 있어 다른 tRNA의 예측에 높은 민감도와 특이도를 나타낸다. 30번째 세대에서 나타난 이 문법은 민감도가 0.79 특이도가 0.975로 나타나 학습이 잘 이루어 졌고, 테스트 과정에서도 좋은 결과가 나왔다[그림3(b)].

4.3 적합도와 트리 복잡도의 변화

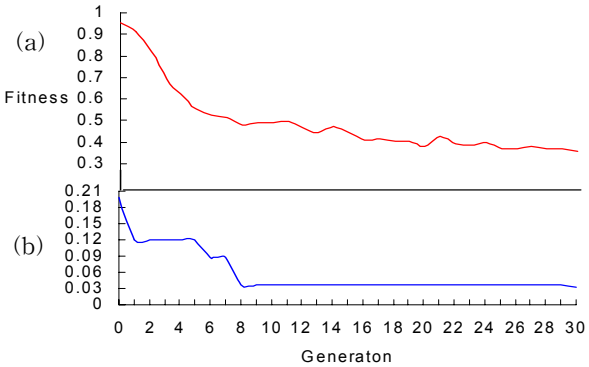


그림 4. 세대별 최적 적합도(b)와 평균 적합도(a)

세대별 최적 적합도의 변화는 그림4(b)를 통해 보여 주고 있다. 최적 적합도는 초기 세대에서 많은 학습이 이루어지며 10세대 이후부터는 거의 변화가 없게 된다. 또한 급격한 변화 전후에는 우성인 적합도의 평형 상태가 상당히 유지되고 있음을 보여 주고 있어 생물계에서의 진화 현상과 유사한 모습을 보여 주고 있음을 관찰할 수 있다. 평균 적합도의 변화는 초기 빠른 학습 성장을 보이며 30세대까지도 완만한 성장을 계속해서 보이고 있다[그림4(a)]. 이는 트리 구조의 문법이 학습에 좋은 결과를 보여 주고 있음을 의미한다.

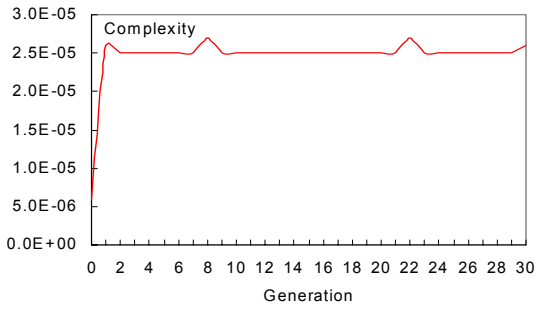


그림 5. 세대별 최적화 개체의 복잡도

적합도의 측정에서 복잡도의 개념을 도입한 결과 초기 세대를 제외한 거의 모든 세대에서 일정한 정도의 복잡도를 갖게 되었으며[그림5] 트리의 노드 개수도 6개에서 9개 사이에서 최적화 되었다. 이렇게 RNA 구조를 나타내는 함수의 트리는 적합도의 기준에 따라 안정된 형태로 학습하고 있음을 보여주고 있다.

4.4 최적 적합도 문법의 특이도, 민감도, 상관 계수

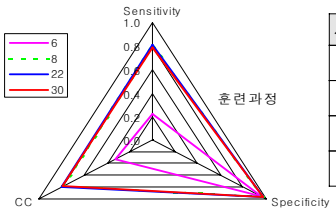
양성인 tRNA 염기서열 데이터 100개와 음성인 염기서열 데이터 100개에서 훈련된 최적 적합도의 문법은 평가를 위해 tRNA 염기서열 290개의 양성 데이터와 음성인 염기서열 290개로 테스트를 실시한 후, 특이도와 민감도를 분석하였다. 또한 특이도와 민감도를 이용해 구한 상관 계수[(2)식]를 적합도와 비교해 보았다. 음성데이터는 mRNA, rRNA, smallRNA, IRE, 루프 구조의 염기서열과 유전자간 염기서열의 집합으로 구성하였다.

	Actual +	Actual -
Predicted +	TP	FP
Predicted -	FN	TN

표 2. 오분류행렬(confusion matrix)

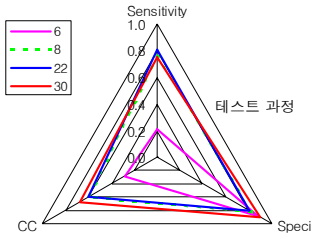
$$CC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \quad (2)$$

6, 8, 22, 30세대에서 훈련 과정과 테스트 과정에서 적합도는 모두 줄어 드는 반면, 상관 계수는 두 과정에서 모두 증가한다. 이는 최적화된 문법을 이용한 RNA 구조 예측은 적합도를 최소화, 상관계수를 최대화하는 방향으로 해야 함을 의미한다. 특히 30번째 세대에서 나타난 최적 적합도 문법은 훈련 과정에서 8, 22세대와의 차이가 거의 없었으나 테스트 과정에서는 30세대에서 더 높은 상관 계수를 보여 주고 있다.



세대	민감도	특이도	CC
6	0.220	0.957	0.329
8	0.790	0.975	0.784
22	0.810	0.976	0.802
30	0.790	0.975	0.784

그림 6. 훈련 과정의 민감도, 특이도, 상관 계수의 관계



세대	민감도	특이도	CC
6	0.207	0.882	0.279
8	0.752	0.826	0.596
22	0.803	0.798	0.600
30	0.752	0.893	0.671

그림 7. 테스트 과정의 민감도, 특이도, 상관 계수의 관계

훈련 과정과 테스트 과정에서 후세대의 문법이 훨씬 좋은 성능을 보여 주고 있다[그림6]. 특히 훈련 과정에서의 복잡도 도입으로, 테스트 과정에서 세대가 지날수록 RNA 구조 예측에 좋은 성능을 보이고 있다[그림7].

.

5. 결론

유전자 프로그래밍을 RNA 구조 문법 학습 방법에 적용함으로써, 특히 기존에 알고 있던 염기서열 정보를 이용하여 밝혀지지 않은 RNA들의 예측에 큰 성능을 발휘할 수 있을 것이라 생각되며 테스트 평가에서 민감도 결과가 80% 가까이 나왔다는 것이 이를 뒷받침 할 수 있는 것이다. 특히 특이도에도 높은 결과를 보임으로서 잘못된 예측을 많이 줄일 수 있는 성능을 보이고 있다. 또한 유전자 프로그래밍을 사용한 RNA 구조 문법 학습 방법은 여러 개의 RNA 염기서열을 함축하여 학습하기 때문에 tRNA 외에 다른 종류의 RNA나 DNA에 도 충분히 적용할 수 있을 것이다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업과 IMT-2000 과제에 의하여 지원되었음.

참고문헌

[1] Victor R. Ambros. microRNAs: Tiny regulators with great potential. *Cell*, Vol. 107, pp. 823-826, 2001

[2] Alexander Huttenhofer, Jurgen Brosius and Jean-Pierre Bachellerie. RNomics: Identification and function of small, non-messenger RNAs. *Current Opinion in Chemical Biology*, Vol. 6 pp. 835-843, 2002

[3] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction, *Mathematical Methods for DNA Sequences. CRC Press, Inc.* pp. 159-184, 1989

[4] Shapiro, B.A. and Navetta, J. A. massively parallel genetic algorithm for RNA secondary structure prediction , *The Journal of Supercomputing.* Vol. 8, pp. 195-207, 1994

[5] Thomas J. Macke, David J. Ecker, Robin R. Gutell, Daniel Gautheret, David A. Case and Rangarajan Sampath, RNAMotif : an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, Vol. 29 No. 22, pp. 4724-4735, 2001

[6] Gary B. Fogel, V. William et al. , Discovery of RNA structural elements using evolutionary computation, *Nucleic Acids Research*, Vol. 30, No 23, pp. 5310-5317, 2002

[7] John R. Koza and Andre, David. Automatic discovery of protein motifs using genetic programming. *Proceedings of AAAI-95 Fall Symposium Series -Genetic Programming. Menlo Park, CA: AAAI Press.*