Computational Method for Searching Human miRNA Precursors

인간 miRNA 전구체 탐색을 위한 계산학적 방법

Jin-Wu Nam[1,2], Je-Gun Joung[1,2], Wha-Jin Lee[1,2], Byoung-Tak Zhang[1,2,3*]

[1] **Graduate Program in Bioinformatics**

[2] **Center for Bioinformation Technology**

[3] **Biointelligence Laboratory, School of Computer Science and Engineering**
Seoul National University, Seoul 151-742, Korea

***To whom correspondence should be addressed. E-mail: btzang@bi.snu.ac.kr**

Abstract

본 논문은 진화 알고리즘(Evolutionary algorithm)의 기법중의 하나인 유전자 프로그래밍(Genetic programming)을 이용하여 miRNA 유전자를 발굴하기 위한 알고리즘을 소개하고 있다. miRNA는 세포내에서 유전자의 전사를 중지시킴으로써 유전자의 발현을 직접적으로 조절하게 되는 작은 RNA 집단 중의 하나이다. 그러므로 miRNA를 유전체 데이터에서 동정해내는 작업은 생물학적으로 상당히 중요하다. 한편 유전체 데이터에서 miRNA를 동정해내는 알고리즘은 생물학적 실험에서의 시간과 비용을 상당히 절감할 수 있으며, 생물학적으로 miRNA를 동정하는 많은 어려움을 덜어주게 된다. 하지만 계산학적으로 miRNA의 동정은 1차 염기서열상의 통계적인 중요도가 부족하여 기존의 유전자 예측 알고리즘을 적용하기에는 어려움이 있다. 따라서 본 연구에서는 miRNA의 염기서열보다는 2차구조에서 더 많은 유사성을 갖는다는 점을 착안하여, 2차구조내에서 공통적인 구조를 찾아내고, 그 정보를 이용하여 miRNA를 동정해내는 방법으로 접근하였다. 이 알고리즘의 성능평가를 위해 우리는 test set을 이용하여 학습된 모델의 특이도(= 34/38)와 민감도(= 38/67)를 계산하였다. 평가결과 본 알고리즘이 기존의 miRNA 예측 프로그램보다 높은 특이도를 갖고 있으며, 유사한 수준의 민감도를 갖고 있음을 보여 주고 있다.

Introduction

MicroRNAs (miRNAs)는 약 70에서 120개의 염기로 이루어진 전구체에서 dicer에 의해 만들어 지는 non-coding RNAs (ncRNAs) 중 하나이다 [1, 21]. 이러한 miRNA의 주 기능은 특정 messenger RNA (mRNA) 상에 상보적으로 결합하여 protein의 발현을 직접적으로 억제하는 것이다 [2, 3, 4]. mRNA, transfer RNA (tRNA), ribosomal RNA(rRNA) 그리고 Iron Responsive Element (IRE)와 같이, 모든 RNA 분자들은 유전자 발현에 있어서 간접적인 역할을 할 것으로 알려져 왔다. 그러나 miRNA의 유전자 발현 조절 기능이 알려지면서, 세포내에서 RNA 분자들의 기능이 새롭게 조명되고 있다[20]. 최근 들어 이러한 miRNA 유전자들을 실험적으로 찾아내기 위해서 northern blot[5, 6], miRNP 분리[7], clone library[6, 8] 방법 등이 사용되고 있다. 하지만 이런 실험을 사용한 방법들은 시간과 비용이 많이 소모되는 어려움이 있다. 만약 특이도가 높은 알

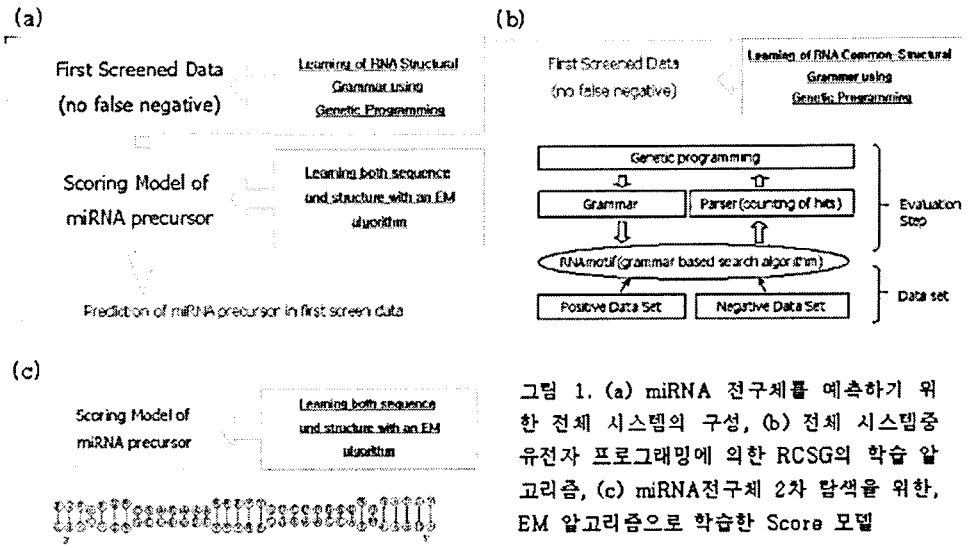

그림 1. (a) miRNA 전구체를 예측하기 위한 전체 시스템의 구성, (b) 전체 시스템중 유전자 프로그래밍에 의한 RCSG의 학습 알고리즘, (c) miRNA전구체 2차 탐색을 위한, EM 알고리즘으로 학습한 Score 모델

고리즘을 통해 miRNA의 후보군을 예측 한다면, 실험적인 시간과 비용을 획기적으로 줄일 수 있다. 최근에 몇몇 연구그룹에 의해 염기서열의 비교유전체학 방법과 통계적인 방법을 통해 miRNA 예측 방법을 발표했다[6, 9]. 하지만 이것은 진화적으로 가까운 종에서 이미 알려진 miRNA가 있어야만 사용 될 수 있는 한정적인 방법이다. 만약 miRNA의 2차 구조에서 진화적으로 보존된 공통구조를 찾아낸다면 범용으로 사용될 수 있는 알고리즘을 개발할 수 있다.

miRNA와 같은 ncRNA의 2차구조에서 공통적인 구조 정보를 추출하는 방법에는 RNA의 2차구조의 유사성을 직접적으로 분석할 수 있는 structural alignment 방법과 [10], 여러 개의 염기서열이 주어졌을 때 상동성이 높은 motif를 찾기 위해 HMMs 방법을 사용하는 것처럼, 여러 개의 RNA가 주어졌을 때 구조적 상동성을 갖는 모델을 학습하는 방법이 있다[11]. 우리는 유전자 프로그래밍을 사용하여 RNA 구조의 상동성을 학습하는, 두 번째 방법을 사용하였다.

현재까지 기능적으로 중요한 ncRNA를 유전체 내에서 찾아내기 위해 염기서열과 구조의 유사성이나 비교유전체학을 사용하는 방법이 소개되어져 왔으나[10, 12], 아직 해결해야 할 문제들이 많이 남아있다. 한편 RNA 구조 예측, 탐색 및 모델링을 위해 RNA 구조를 Stochastic Context Free Grammar (SCFG)로 표현하거나 리스트 구조 표현법을 사용한 연구가 진행되어 왔다[13, 14]. 이러한 연구는 모델링하기 힘들었던 pseudo-knot 같은 복잡한 RNA 구조를 모델링 가능하게 만들기도 하였다[15]. 또한 RNA를 데이터베이스에서 탐색하기 위해 RNA 구조를 컴퓨터 언어로 표현 하기도 하였다[16, 17]. 이러한 문법과 언어는 모두 계산학적으로 자주 사용되는 자료구조인 트리구조로 쉽게 표현될 수 있으며, 몇 가지 제약사항을 둔다면, 유전자 프로그래밍의 개체로 사용될 수 있다.

유전자 프로그래밍은, 유전 알고리즘의 하나로, 각 개체를 트리 형식으로 표현하며, 개체들에 대해 돌연변이(Mutation)나 교차(Crossover)와 같은 변이를 줌으로써 세대가 반복함에 따라 주어진 적합도 함수(Fitness function)에 근사하도록 하여 자동적으로 주어진 데이터를 학습하는 기계학습 방법이다[18]. 이 유전자 프로그래밍은 상위 레벨의 문제를 대변하는 유전자 프로그램으로 자동

적으로 학습하게 된다. 한편 유전자 프로그래밍은 생물학적 네트워크를 재구성하거나 보존된 염기서열 구간을 찾는 등의 방법에 사용되어져 왔다[19].

본 연구에서는 인간의 miRNA 전구체 데이터를 가지고 유전자 프로그래밍을 이용한 공통구조문법의 학습 방법을 소개하고, 학습된 공통 구조 문법을 바탕으로, miRNA 전구체 stem-loop의 scoring 모델을 이용한 miRNA 예측 알고리즘을 소개한다.

Materials and Methods

알고리즘의 개요

miRNA 유전자 예측을 위한 유전자 프로그래밍은 (그림 1a) 각 유전자 프로그램에 해당하는 트리의 노드를 재귀적 함수로 정의함으로써 공통 구조 문법 (RNA Common-Structural Grammar, RCSG)을 학습하게 된다. 이것을 위해 우리는 구조적 문법으로 표현 가능한 RNA 구조를 트리 구조로 변형할 수 있는 규칙을 개발하였다 (그림 3). 이렇게 변형된 트리구조는 유전자 프로그래밍에 의해서 공통 구조 문법으로 학습된다 (그림 1b). 학습된 RCSG는 유전체 데이터 내에서 miRNA의 후보를 동정하는 데 사용되며, 이렇게 동정된 miRNA 후보 중에서 2차 선별을 위해 우리는 miRNA 전구체의 Scoring 모델을 학습하여 사용한다(그림 1c).

RCSG 최적화를 위한 유전자 프로그래밍

일반적인 유전자 프로그래밍의 알고리즘은 5단계로 이루어진다. 그 과정은 다음과 같다. (1) population을 초기화 한다; (2) 적합도 함수로 모든 개체를 평가 한다; (3) 특정 선택 방법으로 부모를 선택한다; (4) 부모 개체에 변이를 주어 다음 세대를 만든다. (5) 종료 조건이 될 때 까지 (2)~(4)를 반복한다.

그러나 RNA 공통 구조 문법을 학습하기 위해서는 일반적인 유전자 프로그래밍의 알고리즘에서 응용된 형태의 개발이 필요하다. 그림 2는 pseudo-code로 표시된 RNA 공통 구조 문법을 학습하기 위한 유전자 프로그래밍 알고리즘 이다. 이 알고리즘에서는 구조에 대한 최적화 과정과 (그림 2(1)) 염기서열에 대한 최적화 과정을 (그림 2(2)) 두 번에 걸쳐 행하게 된다. 또한 구조의 최적화 과정에서는, 각 함수의 변수에 대한 전역 최적 해를 구하기 위해, 국지 탐색 (Local Search)을 실행하게 된다. 또한 개체인 함수 트리의 적합도 계산을 위해 트리에서 구조 문법, 또는 그 반대로의 전환을 위한 과정을 거치게 된다. 트리에서 문법으로 변환된 개체는 RNAMotif 프로그램에[16]을 의해 적합도가 계산 되어 진다.

```
begin                    /* Structural Learning */
    t = 0                /* generation */                          (1)
    initialize P(t)      /* population */
    convert P(t)         /* tree to grammar*/
    evaluate P(t)
    while (not termination-condition) do
    begin
      S = S + above(P(t))              /* Top group for Seq learning*/
      t = t+1
      select P(t) from P(t-1)          /* selection */
      crossover-mutate P(t) except Best /* genetic operators */
      convert P(t)
      evaluate P(t)                    /* fitness function */
      if(local search)
        while (not termination-condition) do
          j = j + 1
          Pj(t) = mutate P(t)
              if (evaluate P(t) < evaluate Pj(t) )
                  P(t) = Pj(t)
    end
end
```

```
w = wordwise(training data)
begin                                  /* Learning of Sequence */  (2)
    t = 0                              /* generation */
    initialize S(t) from S with w      /* population */
    convert S(t)
    evaluate S(t)
    while (not termination-condition) do
    begin
      t = t+1
      select S(t) from S(t-1)          /* selection */
      mutate S(t) for only seq. except Best  /* genetic operators */
      convert S(t)
      evaluate S(t)
    end
end
```

그림 2. RCSG를 학습하기 위한 유전자 프로그래밍 알고리즘의 pseudo-code. (1)은 구조 학습 과정의 알고리즘이며, (2)는 서열 학습 과정의 알고리즘이다.

개체 표현

구조 문법을 함수 트리로 변환하기 위해 그

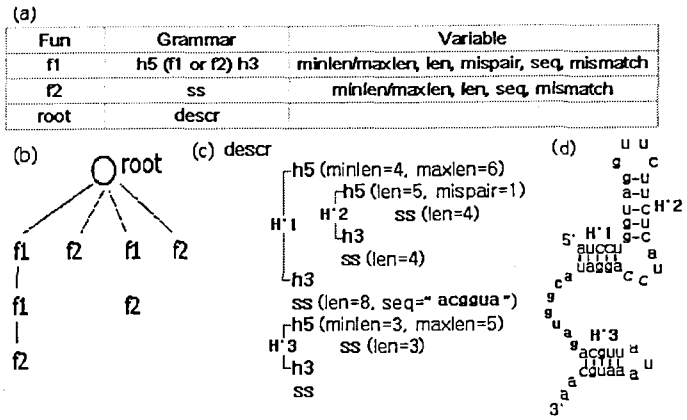

그림 3. (a): 구조 문법을 함수 트리로 바꾸기 위한 규칙과 함수 정의 (b): 함수트리 의 예 (c): b 함수 트리에 해당하는 구조 문법의 예 (d): c 구조 문법에 해당하는 RNA 구조의 예

림 3a와 같이 함수 f1, f2를 정의 하였다. 함수 f1은 helix구조와 (h5-h3) 자기 자신 (f1) 또는 f2를 부르는 재귀적인 형태로 표시되며, 함수 f2는 한 개의 가닥(single strand)을 의미하는 ss로 정의 하였다. 그리하여 함수 트리의 재귀적 전개에서 f2가 나오게 되면 그것은 말단 노드로 정해지게 된다.

각 함수에는 RNA 구조 문법이 가질 수 있는 변수 (minlen/maxlen, len, mispair, seq, mismatch)를 포함하고 있다. 여기서 'minlen/maxlen'은 최소길이와 최대길이를 의미하며, len은 특정 길이를 반환한다. 'mispair'는 helix 구조에서 mispair를 허용하는 개수를 의미하고, 'seq'는 그 구간에서 포함하는 염기서열이며, 'mismatch'는 염기서열이 포함된 구간에서 달라도 되는 염기서열의 수를 말한다. 한편 각 함수의 변수는 RNA 구조를 표현하는 매개변수 (parameter)로 정해져 있기 때문에 그 사용이 제한되어 있다. 즉 한 개의 함수에서 'maxlen/minlen' 과 'len'의 변수는 동시에 존재 할 수 없으며, 'mispair'의 수는 'len'의 값을 초과 할 수 없다. 또한 'mismatch'는 'seq' 변수와 함께 있어야 하며 그 값은 염기서열의 수를 넘어서는 안 된다.

개체군 초기화

개체군의 초기화는 개체인 함수 트리를 무작위적 생성으로 이루어진다. 함수 트리의 노드 개수와 트리의 깊이 (depth) 그리고 각 함수에 할당되는 변수의 종류와 그 값 모두 무작위 함수(random function)에 의해 정해진다. 초기 개체 수는 사용자에 의해 정해지도록 했다.

적합도 함수

유전자 프로그래밍 학습의 방향을 결정하는 적합도 함수는 각 세대에서 생성된 RNA 구조 문법이 RNAmotif 프로그램에 의해서 탐색되는 결과를 기초로 계산된다. 유전자 프로그래밍에 의해 세대별로 생성되는 RNA 구조 문법이 양성 데이터와 음성 데이터에서 탐색되는 개수에 의해 특이성과 민감도를 결정하게 되며, 이 두 가지 값은 적합도를 결정하게 된다(식 1). 또한 트리의 노드 수와 깊이를 이용해 얻은 복잡도를 적합도 함수에 적용함으로써 트리의 구조가 너무 작거나 커지는 것을 막을 수 있도록 조절하였다(식 3, 4). 더 나아가 특이도와 민감도에 상수 *spC* 와 *stC*를 곱하여 적합도 함수에서 특이도와 민감도에 대한 균형을 조절하도록 하였다(식 1, 2).

$$Fitness = spC*Specificity + stC*Sensitivity + Complexity \quad (1)$$

$$spC + stC = 1 \quad (2)$$

$$iComp = TreeDepth \times 10 + NodeNum \quad (3)$$

$$Complexity = \frac{1}{(NS+PS)^2} \times \frac{iComp}{(i-1)bestComp} \quad (4)$$

변이 (Variation) 와 선택 (Selection)

개체의 변이는 교차 (Crossover)와 돌연변이 (mutation)에 의해서 이루어진다. 개체에 대한 교차는 함수 트리간의 하위 트리(subtree)의 교환으로 이루어진다. 교차는 0.8의 확률로 일어나고, 교차지점은 무작위로 양쪽 트리에서 정해지며, 양쪽트리의 함수가 동일해야 교차가 가능하도록 하였다. 만약 다른 함수로 된 노드간의 교차가 일어난다면 구조문법에 치명적인 오류가 발생하기 때문이다. 개체에 교차 변이를 줌으로써 전체 개체군에 대한 다양성을 부여하게 된다.

개체에 대한 돌연변이는 함수에 속한 변수값의 변환으로 이루어진다. 돌연변이는 0.5의 확률로 일어나며, 변환되는 변수의 값은 현재 갖고 있는 값에 대한 포아송 분포 (Poisson Distribution)에 따라 결정된다.

이렇게 개체에 변이를 주어 새로운 개체를 만들었다면, 다음 세대로 전달될 개체를 선택해야 한다. 개체 선택의 방법에는 여러 가지가 있으나, 본 연구에서는 순위 선택 (Ranking selection) 방법을 통해 상위 50% (개체수)을 다음 세대로 전달하도록 하였다.

RCSG이용한 miRNA 전구체 후보군 탐색

최적화된 miRNA 전구체의 공통 구조 문법 (RCSG)을 찾았다면 다음으로 그것을 이용하여 유전체 데이터베이스 내에서 miRNA 전구체 후보를 찾는다. 물론 학습된 RCSG의 특이도가 상당히 높다라면 탐색의 결과가 최종의 결과가 될 수 있다. 하지만 특이도가 높은 RCSG가 학습되지 않았다면, 탐색된 RNA 유전자 후보 중 실제 찾고자 하는 목표인 것만을 골라내는 2차 탐색작업을 하게 된다. 이 전구체 후보는 2차 선별에서 라이브러리 (library)로 사용된다.

2차 선별을 위한 Scoring 모델

miRNA 전구체에 대한 2차 탐색작업을 위해, 우리는 miRNA 전구체에 대한 scoring 모델을 만들었다. 이 모델은 학습을 위해 주어진 각 miRNA 전구체 n개에 대해서 i=1부터 전구체 사이즈의 i=l까지의 transition score($S_{i,j}$)와 pairing score($P_{i,j}$)의 합을 최대화하는 방향으로 학습된다 (식 5). 이 scoring 모델은 알려진 인간의 miRNA 전구체들을 이용하여 학습을 하게 되며, 학습된 모델은 핵산의 IUPAC ambiguity 코드로 표시하게 된다 (표 1).

$$Score = \sum_{i=1}^{l}\sum_{j=1}^{n} \{S_{i,j} \cdot At\} + \sum_{i=1}^{l}\sum_{j=1}^{n} \{P_{i,j} \cdot Ap\} \quad (5)$$

여기서 At 와 Ap는 $S_{i,j}$ 와 $P_{i,j}$ 의 치우침 정도를 결정하는 상수이며, 이 상수 값에 의해 학습되는 모델이 서열의 염기 치환에 중점을 두는지, 염기의 상보결합 (pairing)에 중점을 두는지 결정하게 된다.

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

표 1. IUPAC 핵산 ambiguity 코드

학습 데이터

유전자 프로그래밍 과정과 scoring 모델 학습 과정에서 모두 human miRNA 전구체를 사용하게 된다. 이 데이터는 인터넷 사이트 <http://www.sanger.ac.uk/Software/Rfam/mirna/search.shtml>에서 모두 공개 되어 있다. 표 2에 나와 있듯이 이미 알려진 데이터 중에서 학습과정과 테스트 과정에 데이터를 나누

	분류	Data	데이터수
유전자 프로그래밍	Training	human miRNA precursors	50
	Test	human miRNA precursors	102
	negative		200
scoring 모델	Training	human miRNA precursors	85
	Test	human miRNA precursors	67
		mouse miRNA precursors	69
		negative set	700

표 2. 학습과 테스트 데이터

어 사용함으로써 본 알고리즘의 효율성을 검증하고자 하였다.

Results

miRNA 전구체의 RCSG 학습

본 연구에서는 유전자 프로그래밍을 응용하여 RNA 공통 구조 문법을 학습하는 알고리즘을 개발하여 human miRNA 전구체에 적용하였다. 학습된 human miRNA 전구체의 RCSG 결과가 그림 4에 표시되어 있다. 염기서열의 단어를 학습한 결과에서 최적합도를 갖는 RCSG는 변수 'mismatch'의 수가 최대일 때 해당된다. 그러나 본 연구에서는

```
descr (fitnss=0.906999, specfty=0.964286, sensty=0.391304)
    h5 ( mispair=3 )
        ss ( minlen=5, maxlen=24 )
        h5 ( len=5, mispair=1 )
            h5 ( minlen=8, maxlen=14, seq="acugacu", mismatch=1 )
                ss ( minlen=5, maxlen=27 )
                h5 ( mispair=5 )
                    ss ( minlen=5, maxlen=15 )
                h3
            h3
        h3
    h3                                                    (a)
```

```
descr (fitnss=0.740588, specfty=0.84058, sensty=0.84058)
    h5 ( mispair=5 )
        h5 ( mispair=1 )
            ss ( len=27, seq="gaguaaa", mismatch=0 )
        h3
        ss ( len=22 )
    h3                                                    (b)
```

```
descr (fitnss=0.905808, specfty=0.925926, sensty=0.724638)
    h5 ( minlen=1, maxlen=1 )
        h5 ( mispair=5 )
            h5 ( minlen=5, maxlen=15 )
                ss ( minlen=13, maxlen=14, seq="agcuggu", mismatch=4 )
                h5 ( mispair=5 )
                    ss ( minlen=16, maxlen=23 )
                h3
            h3
        h3
    h3                                                    (c)
```

그림4. 학습된 RCSG의 결과들 (a); 특이도가 가장 높은 RCSG (b); 학습이 진행 중인 민감도가 높은 RCSG (c); 특이도와 민감도가 높은 RCSG

데이터		50 miRNA 전구체
구조 학습	개체수	100 개
	Generation #	30 반복
	Local Search #	30 반복
염기서열 학습	개체수	1500 개
	Generation #	30 반복
상 수	*spC*	0.95
	stC	0.05

표3. 환경설정 과 매개변수 설정

높은 특이를 나타내는 RCSG를 찾는 것을목적으로 하였기 때문에 그림 4a의 결과가 최적의 RCSG라고 할 수 있다. 이것은 특이도가 0.96 민감도가 0.39이며 적합도는 0.91이었다. 민감도가 0.39로 낮게 나타났는데, 이 민감도는 'seq'의 단어와 'mismatch'의 수에 의존해 있기 때문이다. 본 연구의 장점은 그림 4c와 같은 적합한 RCSG의 결과를 여러 개 찾을 수 있다는 것이다. 그러므로 그림 4c와 같은 결과를 통해서 높은 민감도를 갖는 예측이 가능하며, 그림 4a와 같은 낮은 민감도의 결과를 보완 할 수 있는 것이다. 또한 특이도 보다 민감도가 중요할 때는 그림 4b와 같이 민감도가 높은 결과를 사용할 수 있다. 그림 4b의 결과는 그 구조문법에서도 확인 할 수 있듯이 그림 4a 나 4c 보다 더 일반적인 구조문법을 가지고 있는 것이다.

표 3에서는 이 실험의 매개변수와 상수값을 표시하고 있다. 구조 학습 과정에서는 총 개체수를 100으로 세팅하였으며 총 30세대를 학습하였다. 또한 변수의 국지 탐색을 위해 30세대를 다시 학습하였다. 염기서열 학습에서는, 구조 학습 과정에서 높은 적합도를 갖는 개체를 150개를 선택하고 각기 10개씩을 복제하여 총 1500개의 개체를 생성하였다. 그리고 30세대를 학습하였다. 앞에서도 언급했듯이 본 연구에서는 특이도에 더 중심을 두고 있기 때문에 상수 spC (specificity constant)의 값을 0.95으로 상수 stC (sensitivity constant)의 값 0.05 보다 훨씬 높게 정하였다.

그림 5은 구조학습과 서열학습의 세대별 최적합도와 특이도를 그린 그래프이다. 구조 학습 과정에서 최고 적합도는 0.93이었으며 특이도는 0.95이었다. 염기서열의 학습 과정에서 민감도의 감소로 전체 적합도는 0.91이었지만 특이도가 0.964로 상승하여 RNA 구조 학습 후 염기서열의 학습이 효과적임을 나타내는 결과이다.

우리는 학습된 RCSG의 평가를 위해 102개의 miRNA 전구체 테스트 세트를 사용하였다. 그 결과 77개의 miRNA 전구체를 찾아내어 민감도 0.76의 결과를 나타냈으며, 200개의 음성 테스트 세트에서 9개만이 잘못 찾아져 0.96의 높은 특이도를 보였다. 테스트 과정의 민감도가 학습 때보다 높았던 이유는 적합도가 높은 여러 개의 결과를 사용해 찾은 결과이기 때문이다.

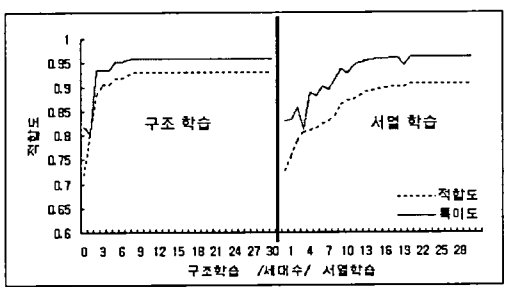

그림 5. 구조학습과 서열학습의 세대별 최적합도와 특이도의 변화

miRNA 전구체 Scoring Model

miRNA 공통 구조 문법이 학습되었다면 그것을 이용해 유전체 내에서 후보를 찾아내야 한다. 하지만 본 논문에서는 테스트 데이터를 이용한 Scoring 모델의 평가만을 보여준다.

우선 scoring 모델을 85개의 인간 miRNA 전구체 데이터를 이용해 학습하였으며, 음성데이터는 학습에 사용하지 않았다. 학습된 scoring 모델이 그림 6에 나타나 있다. 이 모델은 각 위치별로 가장 높은 점수를 갖는 ambiguity 코드로 표시되며, 그 코드로 상보적 결합 유무와 보존된 염기서열을 확


그림 6. 식 5를 최대화 하는 방향으로 학습한 miRNA 전구체 scoring모델

인 할 수 있다. 이 결과는 miRNA 전구체에서 Dicer 효소에 의해 인식되는 구조와 서열이 보존된 것으로 생각된다. 전체 miRNA 전구체에서 공통적으로 보전된 구조와 서열은 Dicer 효소에 의해서 인식되는 부위외에 RISC 단백질 복합체에 의해서 인식되는 부위일 것으로 예상된다.

Scoring Model의 평가

학습된 모델을 통해 나오는 점수의 분포 중에서 miRNA임을 예측하기 위한 경계 점수를 찾기 위해 총 700개인 음성데이터의 점수 분포를 확인 하였다. 그 결과 miRNA 예측 경계 점수로 310점을 정하였다. 음성데이터 중 310점을 넘는 데이터는 4개에 불과했으며, 학습 데이터 중 60%가 310점을 넘어 높은 특이도를 보이기 때문이다.

다음 학습된 scoring 모델의 평가를 위해 학습 데이터를 제외한 나머지 67개의 인간

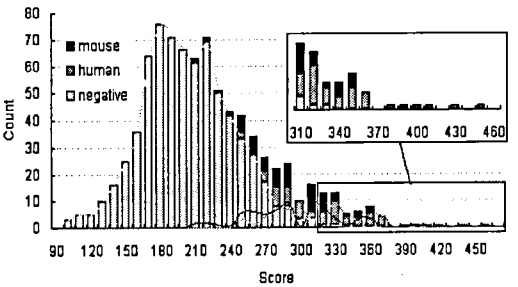

그림 7. scoring 모델을 이용한 테스트 데이터의 점수 분포, count: 그 점수구간의 속하는 전구체의 수, 짙은 검은색: 인간 miRNA 테스트 데이터, 짙은 회색: 생쥐의 miRNA 테스트 데이터, 연한 회색: 음성 데이터.

miRNA 전구체에 대한 점수 분포를 보았다(그림 7). 그 결과에서 67개중 38개의 인간 miRNA 전구체가 예측되어, 민감도 0.57, 특이도 0.89를 보였다. 이 결과는 miRseeker

[9]의 특이도 보다 훨씬 높은 수치이며, 민감도는 약간 높은 수준이다. 한편 우리의 miRNA 예측 알고리즘 이 다른 종의 miRNA도 예측 가능한지를 조사하기 위해, 인간 miRNA 전구체 scoring 모델을 사용하여 생쥐(mouse)의 miRNA 전구체에 대한 점수 분포를 확인해 보았다(그림 7). 그 결과 총 69개의 생쥐 miRNA중 28개가 예측 되었으며 민감도 0.41로 높은 수준을 보였다. 인간 miRNA의 예측율 보다는 높지 않았지만, 이것은 우리의 접근 방법이 기존의 알고리즘보다 더 일반적인 알고리즘임을 설명해 주는 것이다.

miRNA 전구체의 RCSG를 최적화 하는데 사용된 우리의 알고리즘은 다른 ncRNA에도 적용 될 수 있다. 우리는 miRNA 전구체 외에 진핵생물(eukaryote) 의 tRNA와 5s small rRNA에 대해서도 RCSG를 최적화여 높은 특이도와 민감도가 보이는 것을 확인하였다. 이것은 본 알고리즘이 다른 ncRNA을 예측하는 알고리즘으로도 활용 될 수 있음을 보여주는 중요한 결과이다 (본 논문에는 실지 않음).

Discussion

본 논문에서는 유전자 프로그래밍을 이용한 miRNA 유전자 예측 알고리즘을 소개하였다. 그 알고리즘을 통한 학습 결과 높은 특이도와 만족스러운 민감도를 보여주는 miRNA 공통 구조 문법을 찾을 수 있었다. 한편 본 알고리즘을 통해 최적합도를 갖는 공통 구조 문법 이외에 우수한 적합도를 갖는 문법들을 찾을 수 있었기 때문에, 그것들을 이용하여 높은 민감도로 miRNA의 예측을 가능하게 하였다. 또한 본 알고리즘이 miRNA의 예측 외에 다른 ncRNA의 예측에도 사용 될 수 있음을 알았다. ncRNA의 예측 알고리즘은 많이 소개 되고 있으나, 아직 풀어야 할 과제가 많이 남아 있고, 그에 대한 많은 연구가 진행되고 있다. 그 중에 본 연구에서 소개한 공통 구조 문법을 이용한 ncRNA 예측 방법은 전체 데이터에 대한 보존된 구조의 학습이라는 점에서 지금까지 연구되지 않았던 새로운 접근 방법이라 할 수 있다.

우리 연구를 통해 공헌 된 바는 세 가지로 요약하여 말할 수 있다. 첫째로, 유전자 프로그래밍을 통해 RNA의 구조를 나타내는 문법을 자동으로 생성할 수 있었다는 점이다. 실제로 복잡한 구조를 갖고 있는 RNA의 경우 그 문법을 생성한다는 것은 쉬운 일이 아니다. 하지만 RCSG 학습을 통해 쉽게 자동적으로 RNA 구조를 대변하는 문법을 생성할 수 있게 되었다. 둘째로 기존의 miRNA 예측 알고리즘보다 더 좋은 성능을 보이는 알고리즘의 개발이다. 본 알고리즘은 기계학습 방법을 사용하여 miRNA을 예측한 첫 번째 연구이며, 기존의 방법보다 로버스트 (robust)하며 더 일반적이라 할 수 있다. 또한 scoring 모델을 사용한 2차 선별을 하여 잘못된 예측 (false positive)의 수를 줄임으로써 예측 성능을 많이 높였다. 셋째로 본 알고리즘은 miRNA의 예측뿐 아니라 다른 ncRNA의 예측에 사용할 수 있다는 것이다. 세계적으로 최근 ncRNA 중에서도 smallRNA의 기능과 예측에 상당히 많은 관심을 갖고 있으며, 본 알고리즘이 그 smallRNA의 예측에 응용 될 수 있을 것이다.

이후의 연구로 RNA 공통 구조 문법의 더 세밀한 학습을 위해 점수에 관련된 문법이 추가된 상태에서 유전자 프로그래밍에 적용하는 작업을 진행하려고 한다. 점수와 관련된 문법으로 'if 문', 'loop'절 그리고 'case' 문을 예로 들 수 있다. 이러한 조건문이나 순환절을 트리 구조로 표현하기 위해서는 몇 개의 다른 변환규칙이 필요하다. 만약 점수와 관련된 문법을 표현하여 그것을 유전자 프로그래밍으로 최적화하게 된다면, 더 우수한 RCSG를 찾을 수 있으며, 그것을 통해 더 좋은 예측 프로그램을 만들 수 있을 것으로 예상된다. 또 다른 추가 연

구로 최적의 음성데이터 세트를 만들 수 있는 방법이 필요하다. RCSG의 학습은 양성 데이터와 음성데이터간의 점수로 최적화됨으로 음성데이터의 적절한 구성은 좋은 예측 성능을 보이는 데 중요한 요소라고 할 수 있다. 마지막으로 최근 본 알고리즘을 통해 인간 및 다른 진핵생물의 유전체 내에서 miRNA 유전자를 예측하는 연구를 진행하고 있으며, 좋은 결과를 기대하고 있다.

Acknowledgements

본 연구는 BK21-IT, NRL 그리고 IMT2000 생물정보학 프로그램에 의해 지원되었음.

References

[1]. Yan Zeng, Yi R. and Cullen B.R. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. PNAS 100:9779-9784 (2003).

[2]. Ambros V. miRNAs: Tiny regulators with great potential *Cell*, 107:823-826 (2001).

[3]. Gottesman S. Stealth regulation: biological circuits with small RNA switches *Genes & Development*, 16:2829-2842 (2002).

[4]. Zamore P. D. Ancient pathways programmed by small RNAs *Science*, 296:1265-1269 (2002).

[5]. Lagos-Quintana M., Rauhut R., Lendeckel W., and Tuschl T. Identification of novel genes coding for small expressed RNAs *Science*, 294:853-858 (2001).

[6]. Lim L. P., Glasner M. E., Yekta S., Burge C. B., and Bartel D. P. Vertebrate microRNA genes *Science*, 299:1540 (2003).

[7]. Dostie J., Mourelatos Z., Yang M., Sharma A., and Dreyfuss G. Numerous microRNPs in neuronal cells containing novel microRNAs *RNA*, 9:180-186 (2003).

[8]. Lagos-Quintana M., Rauhut R., Meyer J., Borkhardt A., and Tuschl T. New microRNAs from mouse and human *RNA*, 9:175-179 (2003).

[9]. Lai EC, Tomancak P., Williams R.W. and Rubin G. M. Computational identification of Drosophila microRNA genes *Genome Biology*, 4:R42 (2003).

[10]. Sakakibara Y. Pair hidden markov models on tree structures *Bioinformatics*, 19:i232-240 (2003).

[11]. Jih-H. Chen, Shu-Yun Le and Jacob V. Maizel. Prediction of common secondary structures of RNAs: a genetic algorithm approach *Nucleic Acids Research*, 28:991-999 (2000).

[12]. Zuker M. On Finding All Foldings of an RNA Molecule *Science*, 244:48-52 (1989).

[13]. Eddy S. R. and Durbin R. RNA sequence analysis using covariance models *Nucleic Acids Research*, 22:2079-2088 (1994).

[14]. Knudsen B. and Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history *Bioinformatics*, 15:446-454 (1999).

[15]. Cai L., Malmberg R.L., and Wu Y. Stochastic modeling of RNA pseudo-knotted structures: a grammatical approach *Bioinformatics*, 19:i66-i73 (2003).

[16]. 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*, 29:4724-4735 (2001).

[17]. Gary B. Fogel, V. William Porto, Dana G. Weekers, David B. Fogel, Richard H. Griffey, John A. McNeil, Elena Lesnik, David J. Ecker and Rangarajan

Sampath. Discovery of RNA structural elements using evolutionary computation *Nucleic Acids Research*, **30:5310-5317 (2002).**

[18]. Koza J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection MIT Press. (1992).

[19]. Koza J. R., Mydlowec W., Lanza G., Yu J. and Keane M.A. Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming Stanford Medical Informatics Technical Report SMI-2000-0851 (2000).

[20] Carrington J.C. and Ambros V. Role of MicroRNAs in Plant and Animal Delvelopment *Science* **301:336-338 (2003).**

[21] Lee Y., Jeon K., Kim S., Lee J.T., Kim S., Kim V.N. MicroRNA matruration: stepwise processing and subcellular localization *EMBO J.* **21:4663-4670 (2002).**