염기의 IUPAC 코드를 이용한 miRNA Scoring Model의 학습

이화진[012] 남진우[12] 장병탁[123]
생물정보학 협동과정[1]
바이오정보기술 연구소[2]
서울대학교 컴퓨터공학부 바이오지능 연구실[3]
{wjlee[0], jwnam, btzhang }@bi.snu.ac.kr

Learning miRNA scoring models using base IUPAC code

Wha-Jin Lee[012] Jin-Wu Nam[12] Byoung-Tak Zhang[123]
Graduate Program in Bioinformatics[1]
Center for Bioinformation Technology[2]
Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University[3]

요 약

miRNA(microRNA)는 길이가 약 22nt 정도 되는 작은 ncRNA로서 유전자 작용을 조절하는데 중요한 역할을 하는 것으로 알려져 있다. 다이서(dicer)에 의해 성숙한 miRNA(mature miRNA)를 계산학적(computational) 방법으로 학습하여 인간 miRNA의 구조를 예측하였다. miRNA에 관한 구체적인 기작은 아직 확실히 밝혀지지 않았기 때문에 서열 기반과 구조 기반 모두를 포함 하는 모델을 구현 하였으며 ambiguity code를 씀으로써 정보의 손실을 최소화 하도록 하였다. miRNA와 비슷한 구조를 가진 인간 EST로부터 데이터를 무작위 추출하여 실제 인간 miRNA 데이터와 비교함으로써 학습된 결과의 성능을 평가하였다.

1. 서 론

최근 작은 길이의 ncRNA가 동물, 식물, 균류에서 중요한 역할을 하는 것으로 인식되어지고 있다[1]. small RNA라고 불리우는 이것은 유전자 작용을 조절하고 세포 분열을 유도하며 유전자를 억제하는 등 유전자 발현 과정에 중요한 역할을 하는 것으로 밝혀지고 있다.

small RNA중 하나인 siRNA와는 비슷하지만, 다른 역할을 수행하는 작은 길이의 RNA가 발견되었고 이것을 microRNA (miRNA)라고 부르기로 한다. miRNA의 역할과 기능을 정의하기 위해 여러 방법이 시도 되어지고 있는데 이것은 siRNA와 많은 화학적, 기능적 유사성을 가지고 있다. siRNA처럼 다이서(dicer)에 의해 성숙한 miRNA (mature miRNA)가 만들어지고 같은 길이를 갖으며 5'인산기와 3'수산기를 가지고 있다[2]. 또한 효소 복합체인 RISC로 전달하는 것으로 알려져 있다[3].

이처럼 miRNA는 siRNA와 닮았지만, 기능은 조금 다른 것으로 알려져 있다. miRNA는 일종의 안티센스 올리고뉴클레오타이드(anitsense oligonucleotide)로서 mRNA에 상보적으로 결합해 번역을 막는다. 몇 가지 miRNA를 제외하고 기능이 정확하게 알려져 있지 않으나 그 다양성으로 보아 다수의 조절 경로에서 다양하게 활동하는 것으로 보인다[4].

miRNA는 새로운 형태의 유전조절 물질로서 다양하고 필수적인 기능을 가진 것으로 추측되고 있다. 오랜 시간에 걸친 실험으로 하나씩 발견되는 miRNA를 계산학적인 (computational) 방법을 통하여 예측함으로서 실험에 의한 시행착오를 줄여 시간을 단축시킴을 목적으로 하였다. miRNA 전구체(precursor)에 다이서가 와서 절단 한후 성숙한 miRNA를 목적 서열로 삼았고 서열 기반과 구조 기반 두가지 관점을 모두 수용하여 학습하였다. 마지막으로 miRNA와 비슷한 인간 EST와 비교하여 성능을 평가하였다.

2. 방 법

miRNA 전구체에서 실제로 기능하는 부분인 성숙한 miRNA(mature miRNA)들에 어떠한 공통점이 있는지 인간 miRNA 전구체를 가지고 학습해 보았다. 다음과 같은 실험 조건을 바탕으로 하였다.

실험조건

1. 성숙한 miRNA를 중심으로 3'방향으로 5mer, 5'방향으로 5mer정도를 포함하는 염기서열 35mer를 학습. miRNA 전구체 2차 구조에서 성숙한 miRNA의 상보서열도 함께 학습함
2. 염기서열(sequence)와 구조(structure)를 모두 포함할 수 있도록 함
3. Ambiguity code를 사용

miRNA라고 알려진 서열에 성숙한 miRNA는 약15mer 에서 30mer까지 길이가 다양하다. 그러므로 데이터를 추출할 때 다이서에 의해 성숙한 miRNA로 잘리는 영역을 중심으로 앞뒤로 5mer정도를 더하여 총 35mer를 학습하였다. 앞뒤로 약 5mer의 서열을 더하는 것은 성숙한 miRNA를 중심으로 어떤 보존된 영역이 있는지를 알아보기 위함이다. 또한, 아직까지 구조와 서열 어떤 것이 기능에 영향을 미치는 것인지도 알려지지 않았으므로 구조와 서열 모두를 포함할 수 있도록 하였으며, ambiguity code를 사용하여 대표 서열을 찾아내도록 하였다. 데이터는 인간 miRNA라고 알려진 서열 69개를 가지고 학습하였다.

2. 1) 서열 기반의 학습

학습할 데이터로부터 보존된 서열을 찾기 위해 [표 1]과 같은 점수표(score table)를 사용하였다. A→A, G→G, U→U, C→C등 변이(transition)가 일어나지 않는 경우는 점수를 많이 주어 대표하는 염기 서열 값을 찾을 수 있도록 하였고 그 외의 경우는 생화학적 특성을 고려하여 점수를 다르게 주었다. 성숙한 miRNA를 중심으로 35mer를 학습하였고, 이 서열에 상보적으로 붙거나 벌브(bulb)를 이루거나 루프(loop)를 만드는 등의 2차 구조를 이루는 영역 역시 학습하였다.

	A	G	U	C
A	10	4	-1	0.5
G	4	10	-1	0.5
U	-1	-1	10	4
C	0.5	0.5	4	10

[표 1] 변이에 따른 점수표 (transition score table)

2) 구조 기반의 학습

C와 G, U와 A, U와 G가 만나면 서로 상보적 이므로 한 쌍을 이루지만 그 외에는 짝을 이루지 않는다. 구조 기반의 학습은 데이터 집합의 구조를 학습하여 그것을 대표하는 구조를 갖는 서열을 찾아내는 것이다. 데이터 집합의 서열이 상보적 염기 서열 일 경우 학습할 서열도 상보적이면 점수(score)를 주고, 상보적이지 않을 경우 학습할 서열도 상보적이지 않으면 점수를 주도록 하였다. 즉, 데이터 집합의 서열과 학습되는 서열의 구조가 같으면 염기(base)의 종류에는 상관 없이 점수를 주었다. 서열 기반의 학습을 함으로서 데이터 집합을 대표할 후보 서열들을 추출하고 이 서열들의 구조의 유사성을 평가함으로써 데이터 집합에 가장 유사한 서열을 선택하였다.

3) 서열 기반과 구조 기반의 트레이드 오프(trade-off)

변이가 적게 일어난 서열에 점수를 높게 주어 순위를 매긴 후, 상보적 서열의 유무에 따른 구조 기반의 학습을 하여 miRNA를 대표하는 서열을 계산하도록 하였다. 그러므로 서열 기반과 구조 기반 중 어느 곳에 더 중점을 두느냐에 따라서 결과가 조금씩 다르게 나온다. 서열기반과 구조 기반의 관계는 식(1)과 같고 트레이드 오프(trade-off)의 관계에 있다.

$$Score = \sum_{i=1}^{l}\sum_{j=1}^{m} S_{i,j}At + \sum_{i=1}^{l}\sum_{j=1}^{m} P_{i,j}Ap \qquad (1)$$

$$1 = At + Ap \qquad (2)$$

데이터집합의 개수n과 성숙한 miRNA의 길이(length) l까지 차례로 변이 점수(transition score) $s_{i,j}$를 구하고 접합 점수(pairing score) $p_{i,j}$를 구한 후, 해당 변이 점수에 더해준다. At와 Ap는 서열 기반과 구조 기반의 기여도를 결정하는 상수이고 모델의 점수를 최대화하는 방향으로 결정한다.

4) Ambiguity code

[표 2]와 같은 ambiguity code를 사용하여 학습 할 때 점수를 계산하였다. A, T, C, U가 각각 많이 나오는 경우는 상관 없지만, 특정 위치에 두 개에서 네 개의 염기가 번갈아서 나타난 경우에 그것을 표현 할 수 있어야 한다. 이것은 ambiguity code의 사용을 통해서 해결 할 수 있다. 예를 들어 A→M으로 변이 될 경우, M은 A와 C를 나타내므로 점수는 A→A와 A→C로 변이 될 점수를 더한 후 2로 나누어 주었다. 이러한 방법은 해당 위치에 A로 나타내기에는 C가 너무 많고, C로 나타내기에는 A가 너무 많은 경우에 M으로 나타내어 모두를 만족 시킬 수 있다. 즉 학습된 서열의 문자가 M인 경우는 데이터 집합의 해당 위치에 A와 C가 각각 절반의 확률로 가장 많이 나온다는 것을 뜻한다. 구조 기반의 학습 시에도 ambiguity code를 사용하였는데 G와 M이 짝을 이룰 확률은 G와 A, G와 C가 짝을 이루는 점수를 더한 후 2로 나누었다. Ambiguity code의 사용은 서열 하나하나의 의미를 최대한 반영하면서 다양한 서열들을 대표하는 염기서열을 만들 수 있도록 해준다.

IUPAC-IUB/GCG 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
X/N	G or A or T or C	X
	not G or A or T or C	.

[표 2] Ambiguity code

3. 실험 결과

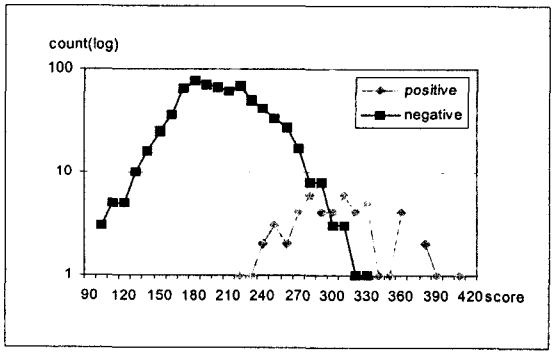
[그림 1]은 학습된 결과로 나온 miRNA 점수 모델(miRNA scoring model)이다. 어두운 원은 다이서에 의해 잘리어서 실제로 기능을 하는 부분이고 흰 원은 성숙한 miRNA에 5mer정도를 더한 것이다. 결과를 분석해 보면, 성숙한 miRNA가 시작하는 부분은 서로 상보적인 서열을 가지고

있고 성숙한 miRNA 영역에는 몇 개의 벌브(bulb)를 가지고 있다. 벌브를 이루는 서열을 중심으로 보존된 영역이 존재한다. 성숙한 miRNA가 끝나는 부분에는 loop이 존재하는 것을 볼 수 있다.

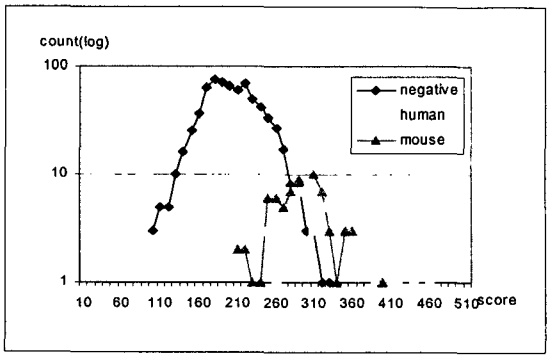
[그림 1] 학습된 miRNA 모델(model)

[그림 2]는 학습된 서열을 가지고 학습에 사용한 인간 miRNA 서열 69개와 인간 EST로 만든 700개의 부정적 데이터 집합의 점수를 내어 도표에 로그를 취하여 그린 것이다. 부정적 데이터 집합은 인간 EST중에서 miRNA와 비슷한 구조를 가진 서열을 Vienna RNA Package의 RNAfold라는 프로그램을 이용하여 무작위 추출하여 만들었다. 즉, miRNA와 거의 유사한 구조를 가지고 있지만 miRNA는 아닌 실제 mRNA 데이터를 뜻한다. 그래프를 보면 인간 miRNA는 평균 약 300정도의 점수를 가지고 있고 부정적 데이터 집합의 평균은 약 200정도의 점수를 가지고 있다. 두 데이터가 겹쳐지는 부분도 존재하지만 인간 EST에서 miRNA와 비슷한 구조만을 뽑은 것을 감안하고 전체 중에서 겹치는 수의 비율을 생각하면 그것은 소수 임을 알 수 있다.

[그림 3]은 인간 miRNA 테스트 데이터 서열 66개와 쥐 miRNA와 인간 EST의 점수표를 로그를 취하여 보여준다. 이것도 [그림 2]와 마찬가지로 부정적인 데이터 집합은 실제 miRNA보다 낮은 점수를 보여주고 있다. 인간 miRNA와 쥐 miRNA의 점수는 거의 비슷하지만 인간 miRNA가 오른쪽으로 약간 더 이동한 모습을 보여주고 있는데 이것은 인간 miRNA로 학습을 했기 때문에 더 높은 점수를 갖고 있는 것으로 보여진다. 각 종간의 miRNA는 공통으로 보존된 영역을 많이 가지고 있지만 기능에 맞게 조금씩 다르게 진화해 왔음을 추측할 수 있다.



[그림 2] 인간 miRNA 학습 데이터와 인간 EST의 점수 그래프



[그림 3] 인간 miRNA 테스트 데이터와 쥐 miRNA와 인간 EST의 점수 그래프

4. 결 론

본 논문은 miRNA를 계산학적인 방법으로 학습함으로써 인간 miRNA 모델을 만들어 보았다. 실제로 miRNA와 비슷한 인간 EST에서 추출한 서열과 비교해 봄으로써 성능을 평가하였다.

miRNA는 유전자 발현의 다양성을 알 수 있는 열쇠가 될 것으로 보여 본 연구가 학문적으로 큰 의미를 가질 것으로 보인다. 또한, 대사나 기능의 결함으로 인한 질병은 miRNA의 잘못된 작용으로 얻어질 수 있고, miRNA의 작용으로 병에 관련된 유전자를 억제시킬 수 있을 것으로 생각되어지기 때문에 신약 개발등에 응용 할 수 있을 것으로 보인다. 향후 다양하고 정확한 매개 변수(parameter)의 추가와 데이터 양의 증가는 더 정확한 예측을 할 수 있을 것으로 예상 된다.

감사의 글

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

참고문헌

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

[2] Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Billie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C.*elegans* developmental timing. *Cell* 106:22-23. 2001.

[3] Caudy, A.A, Myers, M., Hannon, G.J., and Hammond, S.M. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes & Dev.* 16:2491-2496. 2002.

[4] Lai,E.C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.*, 30, 363-364. 2002.