신경망을 이용한 microRNA target 예측

이화진[O123] 장병탁[123]
생물정보학 협동과정[1]
바이오정보기술 연구소[2]
서울대학교 컴퓨터공학부 바이오지능 연구실[3]
{wjlee[O], btzhang }@bi.snu.ac.kr

Identification of microRNA target using neural network

Wha-Jin Lee[0123] 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

microRNA(miRNA)는 ~22 nucleotide(nt)의 단일가닥 (single-stranded) RNA 분자로서 mRNA의 3'-untranslated region (3' UTR)에 상보적으로 결합하여 유전자 발현을 제어하는 새로운 조절물질이다. 지금까지 실험을 통해 1184 개의 miRNA가 알려져 있으나, miRNA에 의해 조절되는 target 유전자는 실험상의 어려움으로 아직까지 거의 알려지지 않았다. miRNA는 서열의 길이가 짧고 target과 느슨한 상보적 결합을 하기 때문에 기존의 서열 비교 방법으로 miRNA의 target을 찾는 것은 쉬운 일이 아니다. 본 논문은 신경망을 이용하여 mRNA의 3' UTR에서 miRNA가 결합하는 영역을 예측하였다. 신경망은 비선형의 데이터를 학습할 수 있어 miRNA target 예측에 적합하다. miRNA와 mRNA의 결합 영역을 다양하게 분석하였고 기존 예측 방법에 의한 결과와 비교하여 성능을 평가하였다.

1. 서 론

microRNA(miRNA)는 21-25 nucleotide(nt)의 RNA 분자로서 mRNA의 번역을 억제하여 진핵 생물의 유전자 발현을 직접 제어하는 역할을 한다. [1] 최초의 microRNA (primary miRNA/pri-miRNA)는 핵 안에서 Drosha라는 RNaseIII type 효소에 의해 70-90nt 정도의 stem-loop 구조로 만들어지고, 이후 세포질로 이동하여 Dicer라는 효소에 의해 21-25 nt의 성숙한 miRNA (mature miRNA)로 만들어진다.[2] 최근 이러한 miRNA를 동정하기 위해 유전체 비교 등의 계산학적인 방법[3]과 northern blot[4], miRNP 분리[5], clone library[3]등이 사용되고 있다. 이러한 실험을 통해 1184 개의 miRNA가 동정되었으나 실험적인 어려움으로 인하여 대다수의 miRNA들의 기능은 아직 밝혀지지 않은 상태이다. 성숙한 miRNA와 mRNA 상에 결합하는 위치는 3'-untranslated region (3'-UTR) 인데 이들은 완벽한 상보적 결합을 하지 않고 한 개 이상의 염기가 불일치(mismatch) 한다. 이처럼 miRNA는 target과 느슨한 상보적 결합을 하고 서열의 길이가 짧기 때문에 기존의 서열 비교 방법으로 miRNA의 target을 찾는 것은 쉬운 일이 아니다. 그래서 miRNA의 기능을 밝히기 위해 최근 생물 정보학 적인 (bioinformatics) 관점에서 접근한 miRNA target 예측 방법들이 발표되었다. [6][7][8] 그러나 이러한 방법들은 단순히 miRNA와 mRNA간의 자유 에너지(free energy)와 결합 빈도만을 통계적으로 비교하는데 그쳤기 때문에 좀 더 계산학적인 접근이 필요하다. 본 논문은 신경망을 이용하여 예쁜 꼬마 선충 (*Caenorhabditis elegans*) mRNA의 3' UTR에서 miRNA가 결합하는 영역을 예측하였다. 신경망은 불완전하고 잡음이 많은 입력에 강하고 특정 분야의 지식이나 휴리스틱 데이터를 네트워크 구조에 쉽게 반영할 수 있어 miRNA target 예측에 적합하다.

2. 방 법

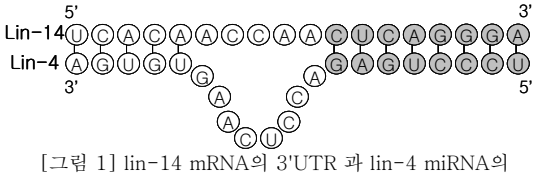
2.1. 데 이 터

학습 데이터는 실험적으로 증명된 18개의 예쁜 꼬마 선충 miRNA:target site 쌍을 사용하였다.[9] 사용한 데이터는 다음과 같다. lin-14/let-7 3쌍, lin-14/lin-4 7쌍, lin-28/lin-4 1쌍, lin-28/let-7 1쌍, lin-41/lin-4 1쌍, lin-41/let-7 2쌍, daf-12/let-7 3쌍이 사용되었다. Negative data로 사용될 miRNA sequence는 Rfam 데이터베이스로부터 예쁜 꼬마 선충의 miRNA 191개를 다운로드 받았고 mRNA 3' UTR은 EnsMart version 15.1을 사용하였다. 이미 실험적으로 증명된 miRNA/mRNA 3' UTR쌍에 seed부분과 상보적으로 5개 이상 결합하고 적당한 임계값 이상의 자유 에너지를 갖는 서열들을 무작위로 생성하여 negative data 81개를 만들어 사용하였다.

2.2. 학 습

miRNA seed이라 불리우는 miRNA의 5' 쪽 여덟 개의 염기가 miRNA가 target을 인식하는데 가장 중요한 것으로 보인다.[7][10] miRNA seed부분의 자유 에너지는 결합 여부를 결정하는 핵심 요소이며 miRNA seed를 제외한 나머지 염기 역시 이 결과를 조절한다.[10] 또한, G:U wobble pair는 열역학적인(thermodynamic) 안정성을 떨어뜨려 miRNA/target 결합을 방해한다.[10] 그리고 mRNA는 한가지 이상의 miRNA에 의해 발현이 억제되고 억제의 정도는 mRNA와 miRNA양과 관련이 있다.[10] 이와 같은 특징을 바탕으로 다음 열 한가지의 특징(feature)을 수치화 시켜 신경망의 입력으로 사용하였다.

(1) miRNA seed와 mRNA 3' UTR이 쌍을 이룰 때, 상보적으로 결합하는 염기의 개수
(2) miRNA seed과 mRNA 3' UTR의 자유 에너지(free energy)
(3) miRNA와 mRNA 3' UTR 결합 구조의 자유 에너지
(4) miRNA seed 영역의 G/U wobble pair 개수
(5) 결합하는 전체 RNA 분자 중 G/U wobble pair 개수
(6) miRNA/mRNA 3' UTR 결합 구조에서 상보적 결합을 하지 않은(mismatch) 염기의 수
(7) miRNA/mRNA 3' UTR 결합 구조에서 상보적으로 결합하는(match) 염기의 수
(8) 한 개의 miRNA가 결합하는 mRNA의 개수
(9) 한 개의 mRNA에 결합하는 miRNA 개수
(10) 한 개의 mRNA 에 miRNA가 결합하는 작용점의 개수
(11) 한 개의 miRNA가 결합하는 모든 mRNA안에서 작용점의 개수



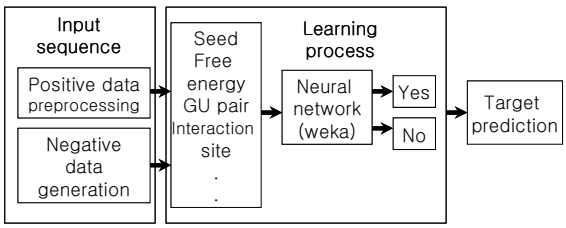
[그림 1] lin-14 mRNA의 3'UTR 과 lin-4 miRNA의 결합구조 예

[그림 1]은 lin-14 mRNA의 3' UTR에 lin-4 miRNA가 결합한 모습을 나타낸다. 회색 염기는 miRNA seed와 mRNA의 결합 부분이다. 이 부분은 자유 에너지를 따로 측정 하였으며 positive 데이터 중에 G/U wobble pair는 거의 발견되지 않았다. 이러한 mRNA/target RNA 2차 구조의 자유 에너지와 G/U wobble pair 개수, 결합한 염기의 개수, 결합하지 않은 염기의 개수를 수치화 하였다. RNA 2차 구조와 자유 에너지의 측정은 Vienna RNA 2차 구조 예측 프로그램 (Secondary Prediction Program) 을 사용하였다.

[그림 2]는 lin-14 mRNA의 3' UTR과 결합하는 let-7, lin-4영역을 보여준다. Lin-14와 결합하는 miRNA는 2개로 알려져 있으며 let-7은 7개, lin-4는 3개의 작용점을 갖고 있다. Let-7과 lin-4는 lin-28, lin-41과 같은 mRNA의 3' UTR과도 결합하는 것으로 알려져 있다.

[그림 2] lin-14 mRNA의 3'UTR과 let-7, lin-4의 예측 결합 빈도 예

신경망 학습에는 Weka 3.4를 사용하였으며 miRNA target인지 아닌지를 판별하는 분류기로는 다층 퍼셉트론(MLP: multi-layer perceptron)을 사용하였다. 입력 노드는 11개, 출력 노드는 ' 예/아니오' 의 2개이다. 은닉층은 한 개 층에 6개의 노드를 갖게 하였으며 학습율, 모멘텀, Decay는 각각 0.3, 0.2, True를 선택하였다. 그 외 항목들은 기본적인 설정 값을 사용하였고 5-fold cross validation으로 평가하였다. [그림 3]은 전체적인 프로그램의 흐름을 보여준다.



[그림 3] 프로그램 흐름도

3. 실 험 결 과

위와 같은 설정하에서, 5-fold cross validation 방법으로 평가한 결과를 [표 1]에서 보여준다. 특수도는 71%로 비교적 낮은것으로 보이는데 그 이유는 training data의 개수가 워낙 적기 때문인 것으로 보인다.

TP Rate	FP Rate	Precision	Recall	F-Measure	Target
0.714	0.012	0.938	0.714	0.811	Yes
0.988	0.286	0.93	0.988	0.958	No

[표 2] 5-fold cross validation을 수행한 신경망 학습 결과

Stark et al.은 그들이 제시한 방법론에 의해 miRNA target을 예측한 후, 가장 가능성 있는 miR-7과 mir-2a의 miRNA target을 직접 실험으로 증명하였다. [8] [표 2]는 그들의 예측 결과와 본 논문에 의한 예측 결과를 비교한 표이다. ΔG 과 Z_{max}는 stark et al.이 target을 선별하기 위해 제시한 기준 값이고 본 논문이 제시한 신경망에 의한 예측 결과가 옆에 있다. 이곳에 나열된 유전자는 서로 관련성이 높고 발현 시기가 비슷한 것이 모여 있어서 miRNA가 함께 관여할 가능성이 높다고 할 수 있다.

Mir-7	ΔG	Z_{max}	Neural network result
CG1489-RB	-38.7	7.47	Y
HLHm3	-37.3	7.03	Y
CG17657-RA	-35.3	6.39	Y
hairy	-35.0	6.29	Y
Tom	-34.5	6.13	Y
hep	-33.9	5.94	
CG8944-RA	-33.8	5.91	
CG10540-RA	-33.1	5.68	
CG10444-RA	-31.8	5.27	Y
m4	-31.5	5.17	Y

Mir-2a	ΔG	Z_{max}	Neural network result
CG1969-RB	-39.0	6.78	Y
CG4269-RA	-38.6	6.66	
reoper	-38.0	6.49	Y
Glaz	-34.3	5.42	Y
BG:DS05899.3	-33.5	5.19	
Scr	-33.2	5.1	
Hbs	-33.0	5.04	
amon	-32.8	4.99	
grim	-32.5	4.9	Y
CG7187-RA	-32.1	4.78	

그림 1. stark et al.의 결과와 신경망의 비교

4. 결 론

miRNA의 기능(function)을 연구하는데 있어서 한계점이 miRNA target을 찾는데 어려움이 많다는 점이다. 본 논문은 miRNA와 그것과 결합하는 miRNA target 간의 특징을 분석하여 정확하고 효율적으로 target 유전자를 찾고자 하였다.

신경망의 입력으로는 본 논문이 제시한 11가지 특성 외에도 mRNA 3' UTR의 결합 부위 근처의 프로파일, mRNA 3' UTR의 종간 보존도, miRNA/mRNA 결합 구조, bulge의 크기와 위치 등을 더 추가할 수 있고 siRNA 생성 프로그램을 miRNA target 예측에 이용될 수도 있다. 이 외에도 지금까지 제시된 다른 방법들의 실험 결과와 통계적인 비교를 통해 성능 검증이 필요한 것으로 보인다. 또한 다양한 케이스 스터디를 통해 환경이나 특정 상황의 지식을 학습하는 것도 좋은 성과가 있을 것으로 기대된다. cDNA나 miRNA, target 유전자에 관한 활발한 연구는 miRNA/target의 구조적인 지식을 더욱 풍부하게 하여 더 정확한 예측을 가능하게 할 것으로 생각된다.

감 사 의 글

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

참 고 문 헌

[1] Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, in press. (2004).

[2] Yoontae Lee, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim and V.Narry Kim, microRNA maturation: stepwise processing and subcellular localization. *EMBO Journal* 21:4663-70. 2002

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

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

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

[6] Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. MicroRNA targets in Drosophila, *Genome Biol.* 5:R1. 2003

[7] Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. Prediction of mammalian microRNA targets. *Cell* 115:787-798. 2003.

[8] Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. Identification of Drosophila MicroRNA Targets. *Plos Biology* 1:1-13. 2003.

[9] Banerjee, D., Slack, F., Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *BioEssays* 24 (2), 119-129. 2002.

[10] John G. Doench and Phillip A. Sharp, Specificity of microRNA target selection in translational repression, *Genes Dev.*, 18(5):504-11. 2004.