SVM과 위치 기반의 자질을 이용한 MicroRNA 목표 유전자 예측

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

MicroRNA Target Prediction using a Support Vector Machine and Position based Features

Sung-Kyu Kim[O,1,2] Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics[1]
Center for Bioinformation Technology[2]
Biointelligence Laboratory, School of Computer Science and Engineering[3],
Seoul National University, Seoul 151-742, Korea

요 약

MicroRNA (miRNA)는 작은 크기의 RNA 분자로서 동식물의 유전자 발현 과정을 직접적으로 조절하는 인자로 알려져 있다. MiRNA는 보통 목표 유전자의 3'-UTR 영역에 상보성을 갖고 결합함으로써 작용하며 특히 miRNA의 5' 부분의 8 nt 정도가 seed로서 중요하다고 알려져 있다. 반면 최근의 연구에 따르면 seed 부분의 서열의 조성 및 양상이 변화함에 따라 특이도가 결정됨을 알 수 있지만 기존의 컴퓨터를 이용한 miRNA 목표 유전자 예측 방법들은 이러한 정보를 활용하지 못한다. 본 논문에서는 열역학적인 수치와 서열의 조성뿐 아니라 miRNA:mRNA pair의 위치에 기반한 정보들을 학습에 자질로서 포함하여 목표 유전자를 예측한다. 그 결과는 위치 기반 자질이 학습 성능 향상에 중요하게 기여함을 보여준다.

1. 서 론

MicroRNA (miRNA)는 약 22 nucleotide (nt) 정도의 작은 크기를 갖는 RNA 분자로서 동식물의 유전자 발현 과정에서 messenger (m) RNA과 느슨한 상보적 결합을 함으로써 발현을 억제한다[1]. 현재까지 대략 1% 정도의 동물 유전자가 세포의 발생 과정에서 miRNA의 조절을 받고 있는 것으로 알려져 있으며 이들은 다수의 종간에 걸쳐 보존된 특성을 갖고 있어 miRNA의 중요성을 실감하게 한다[2]. 특히 알려진 몇몇 miRNA:mRNA pair들의 서열을 실험적으로 변형하여 행해진 실험으로부터 miRNA나 그 작용점의 변형이 일부 기능에 결정적인 영향을 미칠 수 있음이 알려지면서 이들에 대한 연구가 더욱 활발해지고 있다[2,3,4,5,6].

MiRNA 연구에는 miRNA와 그것이 조절하는 목표 유전자를 동정하고 그들의 작동 절차를 밝혀냄으로써 그 기능을 알아내는 것이 중요하다. 그러나 miRNA 기능 연구의 방법은 아직까지 초기 수준에 머무르고 있는 실정이며, 다만 지금까지의 연구를 통해 밝혀진 내용은 다음과 같다. MiRNA는 mRNA의 3'-untranslated region (3'-UTR) 영역에 다양한 형태의 불완전한 상보적 결합을 이룸으로써 작용한다. 또한 miRNA의 5' 영역의 약 8 nt에 해당하는 seed라 불리는 영역이 중요한 역할을 하며 seed 부분이 불완전한 결합을 하는 경우엔 3' 영역의 정보도 중요하게 영향을 미치는 것으로 알려져 있다[7,8,9,10].

지금까지는 실험적으로 목표 유전자를 동정한 이후 miRNA의 작용점을 규칙 기반이나 최소 자유 에너지 기법을 이용한 예측으로 알아내는 방법이 일반적이었다. 이러한 예측 방법에는 Mfold (http://www.bioinfo.rpi.edu/applications/mfold/)나 Vienna Package (http://www.tbi.univie.ac.at/~ivo/RNA/)를 이용한 열역학적 에너지 정보[7,8,9,10], 동적 프로그래밍[7], seed 부분의 pair 정보[7,8,10] 등이 이용되었다. 그러나 이런 예측 방법들은 대체적으로 잘못된 예측(False Positive)의 비율이 높다[9,10].

한편 최근의 연구에서는 다른 조건이 비슷하더라도 seed 부분의 염기쌍의 조성 및 양상의 차이에 따라 특이도가 변화함을 실험적으로 증명하였다[2]. 또한 목표 유전자 동정 시 서열에 변화를 주어 miRNA의 조절 기능이 달라지는 현상을 확인하는 방법을 주로 사용하는데[3,4,5,6] 이는 잘못된 예측을 방지하는데 좋은 데이터로서 사용할 수 있는 특이도를 높일 수 있는 정보이다. 이전의 예측 방법들은 이러한 정보를 예측에 활용하지 않았으며 따라서 잘못된 예측의 비율이 높을 수 밖에 없었다.

본 논문에서는 지금까지 알려진 정보들과 함께 seed 부분의 위치에 따른 pair 정보 역시 자질들로 포함하여 Support Vector Machine (SVM)을 이용한 학습에 사용하였다. 이를 통해 miRNA:mRNA pair의 특이도에 중요한 영향을 미치는 자질(Feature)들을 학습에 포함시켰다.

2. 방 법

2.1. 데이터

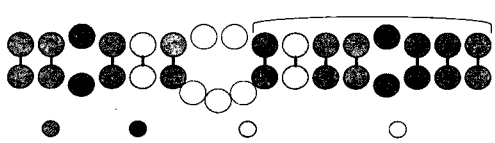
학습을 위한 데이터는 기존의 논문의 내용으로부터 실험적으로 증명된 내용들을 참고하여 얻었다. 실험에 사용된 miRNA 서열들은 Rfam (http://www.sanger.ac.uk/Software/Rfam/)의 miRNA 서열과 비교하여 검증하였고 3'UTR 서열 역시 UTRdb[11]로부터 얻은 서열과 일치하는 지 여부를 확인하였다. 특히 기존 논문에서는 목표 유전자만 실험적으로 밝히고 작용 지점에 대해서는 상보성에 기반하여 예측으로 제시하였지만, 이 결과는 잘못된 예측일 가능성이 있으므로 추후 실험적으로 검증이 된 것이 아닌 이상 사용하지 않았다.

이와 같은 방법으로 양성 52개, 음성 36개로 총 88개의 데이터를 얻을 수 있었다. 기존의 연구에서는 음성 데이터를 무작위로 생성하는 예가 있지만[10], 이는 예측 결과 내의 잘못된 예측의 비율을 측정하는 데는 유용하지만 학습에 직접 사용되기에는 불합리한 면이 있다. 왜냐하면 무작위로 생성되었지만 우연히 양성 데이터가 될 수 있는 조건을 갖도록 생성될 가능성이 있기 때문이다.

그러나 실제 목표 유전자 예측에서는 음성인 데이터를 양성이라고 예측하는 예가 많기 때문에 이를 줄이기 위해서는 어느 정도 많은 수의 음성 데이터가 필요하다. 따라서 음성 데이터를 임의로 생성해야 했는데, 이때 실험에 의해 let-7 miRNA의 목표 유전자 서열을 제거시킨 결과 발현이 억제되지 않는 예[6]를 이용하였다. 즉 이 예에서 let-7 miRNA의 작용지점이 없을 경우 let-7 miRNA가 UTR의 어느 곳에도 정상적으로 작용하지 못한다는 점을 이용하면 UTR의 서열을 임의로 약 25 nt 정도 잘라내어 강제로 let-7 miRNA 서열과 결합시켜 음성 데이터를 만들 수 있다. 이를 통해 얻은 약 1000개의 음성 데이터 중에서 seed 부분의 특성이 실제 양성 데이터와 비슷하고 열역학적 에너지가 낮은 47개만을 사용하였다. 결과적으로 사용한 데이터는 양성 52개, 음성 83개로 135개이다.

2.2. 자질

2.1의 데이터를 학습에 사용하기 위해, 필요한 자질들을 추출해야 한다. MiRNA와 목표 mRNA 서열은 그림 1과 같이 Vienna Package의 시뮬레이션을 이용해 2차원 결합모습을 얻는다. 이때 mRNA의 3'부분과 miRNA의 5'부분에 6개의 'L' 문자로 이루어진 linker 서열을 사용한다.



[그림 1] 목표 mRNA와 miRNA의 결합 예시

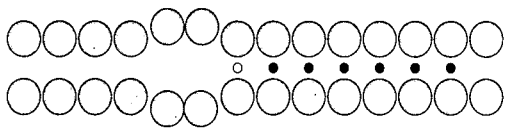
학습에는 총 31개의 자질을 사용하였으며 이들은 다음과 같은 두 개의 자질군으로 나눌 수 있다.

① miRNA:mRNA pair의 seed 영역, 비 seed 영역, 전체 영역 당 match, mismatch, GC, AU, GU, Gap 개수 및 자유에너지.

② 10 nt에 해당하는 miRNA 5'영역의 위치 별 match와 non-match의 이진 값.

①의 정보는 지금까지의 연구에서 주로 쓰인 정보[10]이며 이 중 seed 영역의 match 개수와 자유에너지가 중요하다고 알려져 있다. 여기서 총 21개의 자질을 얻는다.

반면 ②의 정보는, 이전 연구에서 사용하지 않았던, 이 논문에서 처음 도입된 정보이다. 10 nt에 대해서만 위치별 정보를 얻기 때문에 총 10개의 자질을 얻는다. 이전 연구[2]로부터 seed 부근의 pair에 대한 중요도를 그려보면 그림 2와 같이 표현할 수 있다.



[그림 2] seed 부근의 위치 별 pair의 중요도. 검은 원으로 표시된 부분은 pair가 match에서 mismatch로 변형되었을 때 miRNA가 기능하지 않게 되는 부분이고 빈 원은 영향이 있기도 하고 없기도 한 부분이다. 가위 표시된 부분은 영향이 없는 부분이다. 위치에 대한 인덱스 값은 숫자로 표시되어 있다.

이로부터 10개의 위치 별 match/mismatch 값을 match일 경우 1, mismatch일 경우 0으로 주었다. 이후의 위치는 서열에 따라 구조의 변화가 심하고 문헌에서도 seed에 비해 덜 중요하게 다루는 부분이므로 제외하였다.

2.3. 학습

학습은 커널 방법론 (Kernel method)을 이용한 통계적인 분류 방법인 SVM을 사용한다. 높은 차수를 갖는 비선형적인 문제 공간의 데이터를 자질 공간 (Feature space)에 투영한 뒤 자질들 간의 최적의 경계 면을 찾는 방식이다. 본 논문에서는 Sequential Minimal Optimization (SMO)를 이용한 SVM을 사용한다[12].

학습에는 Weka (version 3.4)의 SMO 분류기를 이용하였으며 파라메터를 변형하며 실험한 결과 Complexity parameter를 0.3, Polynomial kernel의 지수를 2.0으로 주었을 때 제일 좋은 결과를 냈다. 에러 측정 방법은 10 fold cross-validation을 사용하였는데, 적은 수의 데이터를 가지고 학습할 때에는 5 fold 보다 10 fold가 더 나은 결과를 낸다고 알려져 있다[13].

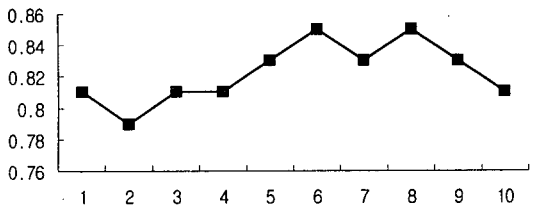
3. 결 과

실험결과는 표 1에서 보는 바와 같이 83%의 민감도와 87%의 특이도를 얻었다. 특히 자질군 ①과 ②의 정보를 모두 사용했을 때 ①의 정보만 사용하여 학습한 결과보다 민감도에서는 약 7%, 특이도에서는 약 5%정도의 성능 향상을 보였다. 이는 위치 기반의 자질들이 학습에 중요한 요소로서 기여함을 나타내는 결과라고 할 수 있다.

[표 1] SVM을 이용한 miRNA:mRNA pair 데이터 학습 결과

자질군	TP	FP	TN	FN	Sensitivity	Specificity
① + ②	43	11	72	9	83%	87%
①	40	15	68	12	77%	82%

위치 기반의 자질 각자가 갖는 상대적인 중요도를 알아보기 위하여 이들을 각각 하나씩 제외한 뒤 민감도를 산출한 뒤 그림 3의 간단한 그래프로 그려보았다. 그림에서 주의할 점은 각 위치의 자질을 제외한 실험 결과이므로 민감도가 낮을수록 해당 자질이 상대적으로 더 중요함을 의미한다.



[그림 3] 위치별 자질을 제외한 실험 결과의 민감도 그래프.가로축은 해당 숫자에 위치한 자질을 나타내고 세로축은 민감도 값을 나타낸다.

그림 3에서 보는 바와 같이 1~4, 10의 위치에 대한 자질은 5~9의 위치에 대한 자질보다 상대적으로 더 중요한 역할을 한다고 이해할 수 있다. 이는 그림 2에서 보인 내용과 비교해봤을 때 seed 앞부분의 역할이 중요하다는 점에서는 일치하지만 다섯 번째 이후의 위치를 고려하면 조금 다르다고 볼 수 있다. 즉 1~8에 해당하는 seed 부분의 중요도가 9, 10 위치에 비해 더 높아야 하겠지만 다섯 번째 이후가 대부분 덜 중요하다는 결과를 보인다. 이는 여러 가지로 해석할 수 있지만 miRNA:mRNA pair가 seed에만 의존하지 않는 경우가 있고 이전 연구[2]의 결과가 해당 실험에 편향되었을 가능성이 있음을 고려할 때 충분히 나타날 수 있는 결과이다.

4. 결론 및 토의

계산학적인 방법으로 miRNA의 목표 유전자를 찾는 문제는 실험적으로 알려진 miRNA:mRNA pair가 많지 않고, 각 경우에 공통점이 크지 않기 때문에 쉽게 풀 수 없는 문제라고 할 수 있다. 더욱이 대다수의 예측 프로그램이 miRNA의 짧은 길이 때문에 많은 수의 잘못된 양성 예측(False Positive) 결과를 낳고 있어서 예측 결과에 대해 신뢰하기가 어려운 실정이다.

위치 기반의 자질들은 학습의 결과를 상당히 향상시키는 결과를 냈지만 반면 이전 연구결과를 명쾌히 지지하지 못하는 단점을 보였다. 현재로서는 많은 수의 시험 가능한 데이터를 확보하는 것이 더 좋은 학습을 위한 최선이겠지만 이는 당분간은 현실적으로 불가능해 보인다. 따라서 miRNA:mRNA pair의 결합 특성을 좀더 잘 설명할 수 있는 자질을 창안하거나 결합 성향에 따라 분류를 한 뒤 적절하게 학습을 하는 방법을 도입해볼 수 있을 것이다.

감사의 글

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

참고문헌

[1] Batel, D.P. MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, 116:281-297, 2004.

[2] Brennecke J., Stark A., Russell R.B., Cohen S.M. Principles of microRNA-target recognition. *PLoS Biol.*, 3:e85, 2005.

[3] Vella M.C., et al. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev.*, 18:132-137, 2004.

[4] Nelson P.T., Hatzigeorgiou A.G., Mourelatos Z. miRNP:mRNA association in polyribosomes in a human neuronal cell line. *RNA*, 10:387-394, 2004.

[5] Poy M.N., et al. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432:226-230, 2004.

[6] Vella M.C., Reinert K., Slack F.J. Architecture of a validated microRNA::target interaction. *Chem. & Biol.*, 11:1619-1623, 2004.

[7] John B., et al. Human MicroRNA targets. *PLoS Biol.*, 2:e363, 2004.

[8] Lewis B.P., et al. Prediction of mammalian microRNA targets. *Cell*, 115:787-798, 2003.

[9] Stark A., et al. Identification of Drosophila MicroRNA targets. *PLoS Biol.*, 1:e60, 2003.

[10] Lee W.J., et al. Identification of C. elegans MicroRNA Targets Using a Kernel Method. *Genomics and Informatics*, 3:15-23, 2005.

[11] Mignone F., et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, 1:D141-146, 2005.

[12] Platt J.C. Fast training of support vector machines using sequential minimal optimization. *In Advances in kernel methods, support vector learning.* (MIT Press), 1999.

[13] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)* Montreal, CA, pp1137-1143, 1995.