공학석사학위논문

Support Vector Machine을 이용한 MicroRNA Target 예측

Prediction of MicroRNA Targets using a Support Vector Machine

2005년 2월

서울대학교 대학원
협동과정 생물정보학 전공
이 화 진

초 록

microRNA (miRNA)는 21-25 nucleotide (nt)의 single-stranded RNA 분자로서 mRNA의 3' untranslated region (3' UTR)에 상보적으로 결합하여 유전자 발현을 제어하는 최근에 새로이 발견된 조절 물질이다. 지금까지 실험을 통해 약 1345개의 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 miRNA/target간의 자유 에너지와 RNA 2차 구조, GU wobble pair, miRNA 5' 영역의 정보 등을 SVM의 입력으로 사용하여 mRNA의 3' UTR에서 miRNA가 결합하는 영역을 예측하였다. SVM은 복잡한 비 선형 데이터를 잘 분리해내고 불완전하고 잡음이 많은 입력에 강하기 때문에 miRNA target 예측에 적합하다. miRNA와 mRNA의 결합 영역을 다양하게 분석하였고 민감도 0.64, 특수도 0.98의 성능을 갖는 SVM을 구현하였다. 학습의 입력 값을 달리하여 각각의 특징이 결과에 미치는 영향을 분석하였고 기존 예측 방법에 의한 결과와 비교하여 성능을 평가하였다.

주요어 : MicroRNA Target, Support Vector Machine,
학번 : 2003-20649

목 차

그림 목록

표 목록

제 1 장 서 론

1.1 연구 배경

microRNA (miRNA)는 21-25 nucleotide (nt)의 RNA 분자로서 mRNA의 번역을 억제하여 진핵 생물의 유전자 발현을 직접 제어하는 역할을 한다 (Lai, 2003; Bartel, 2004). 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 (Lee et al., 2002). 이 성숙한 miRNA 분자는 세포질에서 유전자의 3' 쪽 비번역부위 (untranslated regions; UTR)에 상보적으로 결합하여 target mRNA의 번역을 조절한다. (Ambros, 2001; Ambros, 2003; Banerjee and Slack, 2002; Carrington and Ambros, 2003; Moss, 2002; Moss and Poethig, 2002)

이러한 miRNA는 1993년에 최초로 *C. elegans*에서 발생 과정을 조절하는 stRNA(small temporal RNA)인 *lin-4*와 *let-7*가 발견된 이후 (Lee et al., 1993; Wightman et al., 1993; Moss et al., 1997) 최근 세포 분화와 사멸, 초파리의 지방 대사 (Brennecke et al., 2003; Xu et al., 2003), 식물에서 꽃과 잎의 발생 과정 등에 (Aukerman and Sakai, 2003; Chen, 2003) miRNA가 관여하는 것으로 밝혀지고 있다. 또한 일부 miRNA의 염기 서열은 종간의 보존도가 매우 높아 중요한 생명 현상에 관여할 것으로 추측하고 있다(Pasquinelli et al., 2000; Aravin et al., 2001).

한편 miRNA의 위와 같은 기능을 이해하기 위해서는 miRNA와 반응하는 target mRNA를 찾는 것이 매우 중요하다. 그러나 miRNA의 target을 찾기 위한 고속 처리 실험 기술 (high-throughput experimental techniques)은 아직까지 알려지지 않은 상태이다. 또한 식물의 경우

miRNA은 그것의 target과 거의 완벽하게 상보적으로 결합하여 비교적 쉽게 miRNA target을 동정할 수 있지만(Rhoades et al., 2002) 동물의 경우에는 불일치(mismatch)와 벌지(bulge)를 허용하여 상보적인 결합을 하기 때문에 기존의 서열 (Sequence) 또는 서열의 상동성에 기반한 접근법으로 miRNA target을 찾기가 매우 어렵다. 그러므로 생물정보학적인(bioinformatics) 방법을 통한 접근법은 miRNA의 기능을 밝히는데 좋은 시도라 할 수 있다.

1.2 관련 연구

성숙한 miRNA는 mRNA 3' UTR에 불완전 결합하여 한 개 이상의 염기가 불일치(mismatch)하거나 루프(loop) 구조를 형성하고 서열의 길이가 짧기 때문에 기존의 서열 비교 방법으로 miRNA의 target을 찾는 것은 쉬운 일이 아니다. 그래서 최근 miRNA의 기능을 밝히기 위한 생물정보학적인 관점에서 접근한 miRNA target 예측 방법들이 발표되었다 (Enright et al., 2003; Lewis et al., 2003; Stark et al., 2003; Kiriakidou et al., 2004; Rehmsmeier et al., 2004). Stark et al.의 논문의 경우 최초로 miRNA target 서열을 찾기 위해 시도한 논문으로 miRNA/target간의 자유 에너지와 종간의 보존도를 토대로 miRNA target 후보(candidate)의 리스트를 제시하였다. 그러나 false positive 비율 등을 분석하지 않았기 때문에 이 방법을 이용할 경우 알고리즘의 성능 평가가 어렵다. Lewis et al.의 논문에서는 종간의 보존도가 있는 miRNA와 mRNA만을 사용했기 때문에 아직 다른 종에서 발견되지 않은 miRNA는 데이터로 사용되기 힘들다. 또한 miRNA와 mRNA 3' UTR의 결합 여부를 자유 에너지만을 기반으로 결정하기 때문에 정확하게 분류하기 어렵다. 이와 같이 지금까지의 방법들은 miRNA와 mRNA간의 자유 에너지(free energy)와 결합 빈도만을 통계적으로 비교하는데 그쳤기 때문에 false positive의 비율이 높다. 이 때문에 실제로 miRNA target을 동정하는데 유용하지 못하다.

1.3 연구 내용

본 논문에서는 miRNA/target 결합 구조의 특징을 크게 구조 정보, miRNA 5'의 8 nt, 자유 에너지, 그 밖의 부수적인 정보들 이렇게 네 가지로 나누어 분석하였다. miRNA/target의 RNA 구조 정보는 RNAdistance라는 RNA 2차 구조의 거리를 측정하는 프로그램과 Markov Chain을 사용하여 분석하여 수치화 하였다. 또한 miRNA/target 결합에 가장 핵심적인 요인이라 할 수 있는 miRNA 5' 8nt 정보를 이용하여 자유 에너지를 측정하고 RNA 2차 구조를 분석하였다. 특히 자유 에너지를 측정할 때에는 miRNA 3' UTR, miRNA 5' UTR, miRNA 전체 영역의 세 부분으로 나누어 추출하였다. 그 외에, 결합에 영향을 주는 것으로 알려진 GU wobble pair와 mRNA 3' UTR에서의 작용 지점의 개수 등을 참고함으로써 민감도를 높일 수 있었다. 이와 같이 다양한 특징을 분석하고 SVM (Support Vector Machine)을 이용함으로써 false positive가 높은 기존 알고리즘들의 단점들을 극복하고 정확하고 효율적으로 miRNA target 여부를 분류하였다. SVM은 기존의 기계 학습 이론에서 볼 수 없는 장점들과 함께 뛰어난 성능으로 인해 많은 관심을 끌고 있는데 입력 공간의 비선형적인 높은 차수를 특징 공간(feature space)에 선형적으로 투영함으로써 각 특징 사이의 최적의 경계면을 제시하는 효율적인 분류 방법이다. 이 방법은 통계적 분석에 데이터로서 전통적으로 많이 사용되어 온 벡터나 행렬뿐만 아니라 문자열이나 트리, 그래프와 같은 데이터에 좋은 성능을 보여 생물학적인 데이터의 분석에 적합하다고 알려져 있다. 위와 같이 학습된 모델을 바탕으로 예쁜 꼬마 선충(*Caenorhabditis elegans*), 인간(*Homo sapiens*)과 쥐(*Mus musculus*) mRNA의 3' UTR에서 miRNA target을 예측 하였다.

1.4 논문의 구성

2장에서는 miRNA와 target이 결합하는데 있어 특징들을 살펴보고 그 특징을 수치화하여 입력하는 과정을 설명한다. 또한 입력들의 분류 기법으로 사용한 SVM에 대해 간략하게 설명한다. 3장에서는 제안된 방법을 이용하여 성능을 분석하고 miRNA 검색 프로그램인 TargetScan과 비교하여 평가한다. 그리고 예쁜 꼬마 선충과 인간, 쥐에서 miRNA target을 검색한 결과를 보여준다. 4장에서는 결론을 맺는다.

제 2 장 Materials and Methods

2.1 데이터

학습 데이터(training data) 중 양성 데이터(positive data)는 실험적으로 증명된 43개의 예쁜 꼬마 선충 miRNA:target site 쌍을 사용하였다 (*lin-14/ cel-let-7, lin-14/ cel-lin-4, lin-28/ cel-lin-4, lin-28/ cel-let-7, lin-41/ cel-lin-4, lin-41/ cel-let-7, daf-12/ cel-let-7, hbl/ cel-lin-4, hbl/ cel-let-7, hid/dme-bantam, HLHm3/dme-miR-7, hiary/dme-miR-7, rpr/dme-miR-2, grim/dme-miR-2, Mtpn/mir-375*) (Banerjee and Slack, 2002; Lin et al., 2003; Stark et al., 2003; Poy et al., 2004).

학습 데이터 중 음성 데이터(negative data)는 랜덤 하게 생성한 miRNA 18~25mer와 3' UTR (ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data/)사이에 결합하는 1022쌍의 pair를 추출하여 사용하였다. 추출할 때 랜덤 서열의 5' 영역 8 nt 중 6개 이상이 상보적으로 결합하는 것과 그 영역을 포함하여 자유 에너지가 8.5kcal/mole 이하로 열역학적으로 (thermodynamic) 안정적인 것을 기준으로 하였다. 또한 추출된 서열 쌍 중에서 양성 데이터(positive data)에서 발견되지 않은 그림 1과 같은 서열 쌍은 제외 하였다. 그림 1은 mRNA와 miRNA가 서로 결합하다가 mRNA끼리 결합 구조가 생기거나 miRNA끼리 결합 구조가 생기는 서열 쌍을 말한다.

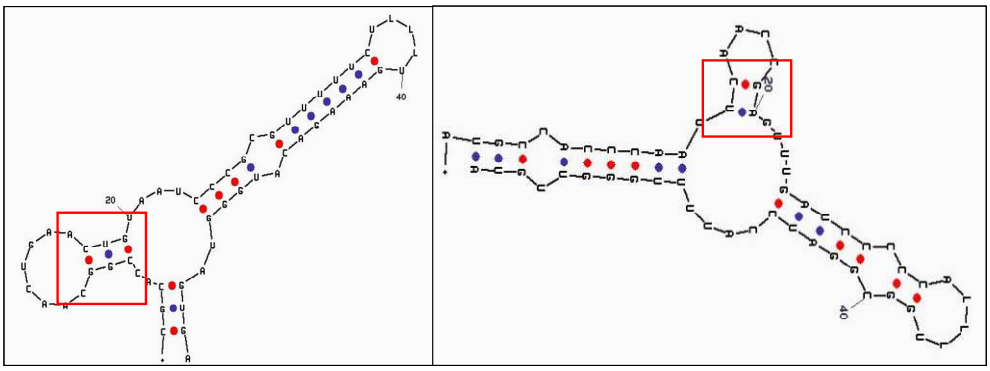

그림 1. miRNA/target에서 발견되지 않은 구조 (출처 : mfold 3.1)

테스트 데이터로는 지금까지 실험으로 증명된 miRNA/target 쌍과 간접적인 실험이나 서열 쌍 검색을 통해 miRNA/target 쌍이라고 믿고 있는 32개의 데이터를 사용하였다. 또한 본 논문에서 제시된 방법에 의해 Rfam 데이터 베이스의 예쁜 꼬마 선충의 miRNA 4개 (*lin-4, let-7, cel-miR-228, cel-miR-229, cel-miR-230, cel-miR-231*)와 사람(*Homo sapiens*)과 쥐(*Mus musculus*)에 공통적으로 보존되어 있는 miRNA *hsa-miR-199a*의 target을 예측해 보았다.

2.2 검증된 miRNA Target 서열의 특징들(Features)

Support Vector Machine의 입력으로 사용하기 위해 실험적으로 검증된 miRNA target의 특징들을(features) 분석해 보았다. 그림 2는 miRNA와 그에 해당하는 target 유전자의 mRNA 3' UTR 영역에 바인딩하는(binding) 구조를 나타낸다. 실험으로 검증된 miRNA target의 서열을 살펴보면 miRNA 3' 영역의 일치(match), miRNA의 벌지(bulge), mRNA 3' UTR의 벌지, miRNA 5' 영역의 일치, miRNA 5' flanking 영역 등의 구조를 공통적으로 갖는다.

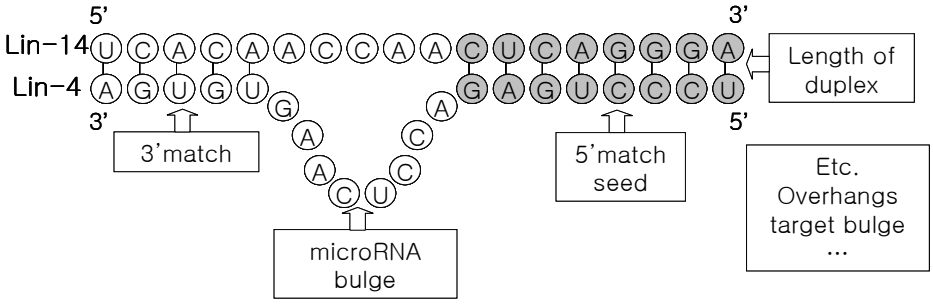

그림 2. miRNA/target 서열의 작용 구조

miRNA target을 인식하는데 miRNA의 5' 쪽 여덟 개의 염기가 가장 중요한 것으로 보인다(Doench et al., 2004). miRNA가 작용하는 target 서열의 패턴을 분석한 결과 miRNA 5' 영역의 보존도가 매우 높아 miRNA와 작용하는데 핵심 요소라고 알려져 있다. miRNA seed의 서열에 돌연변이(mutation)를 생성하여 서열 구조와 자유 에너지의 변화에 따른 miRNA target 서열의 발현량을 보면 자유 에너지가 약 −5.5kcal/mole 이상일 때 활성화 되는 것으로 나타나 이를 뒷받침 하였다(Doench et al., 2004). 이에 반해 miRNA 3' 영역의 경우 miRNA seed 영역 보다 영향을 덜 받는 것으로 나타났다. 그러나 miRNA seed 부분이 완벽한 상보적 결합을 한 경우에는 miRNA의 3' 영역에 불일치가(mismatch) 많아도 발현량이 일정하지만 miRNA의 5' 영역이 불일치를 허용한 상보적 결합일 경우 miRNA 3' 영역의 자유 에너지가 발현량에 큰 영향을 주는 것으로 나타났다. 이를 종합해 보면 miRNA seed의 자유 에너지가 mRNA의 발현 저해 여부를 결정 짓는 가장 큰 요인이지만 miRNA 3' 영역도 발현량에 영향을 미침을 알 수 있다(Doench et al., 2004). 또한 적절한 자유 에너지를 갖는 경우라도 miRNA 5' 영역의 G/U 염기 쌍은 발현량을 저해시키는 것으로 나타났다(Doench et al., 2004).

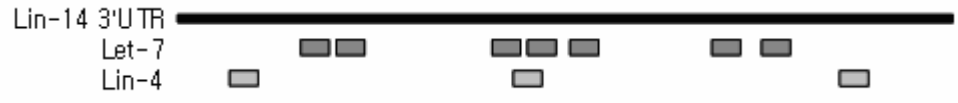

그림 3. lin-14 mRNA의 3'UTR과 *let-7, lin-4*의 예측 결합 빈도 예

그림 3은 *lin-14* mRNA 3' UTR이 *let-7, lin-4*과 결합하는 위치를 보여준다. *lin-14*와 결합하는 miRNA는 *let-7, lin-4*의 2개로 알려져 있으며 *let-7*은 7개, *lin-4*는 3개의 작용점을 갖고 있다 (Banerjee and Slack, 2002). 이처럼 miRNA와 작용하는 mRNA는 한가지 이상의 miRNA에 의해 발현이 억제되고, 한 개 이상의 작용점을 갖는다. 억제의 정도는 mRNA와 miRNA양과 관련이 있고 하나의 miRNA는 한 개 이상의 mRNA와 작용하는 것으로 보인다. 실제로 *let-7*과 *lin-4*는 *lin-28, lin-41*과 같은 mRNA의 3' UTR과도 결합하는 것으로 알려져 있다.

2.3 특징 선택 (Feature Selection)

위와 같은 특징을 바탕으로 학습 데이터의 특징(feature)을 선택하였다. 표 1은 SVM 입력에 사용된 열 가지의 특징(feature)을 보여준다.

(1) miRNA 5' 영역 8nt에서 mRNA 3' UTR과 상보적으로 결합하는 염기의 개수
(2) miRNA 5' 영역 8nt와 그와 결합하는 mRNA 3' UTR의 자유 에너지(free energy)
(3) miRNA/mRNA 3' UTR 결합 구조의 자유 에너지(free energy)
(4) miRNA 5' 영역 8nt의 G/U wobble pair 개수
(5) miRNA/mRNA 3' UTR 결합 구조에서 상보적 결합을 하지 않은(mismatch) 염기의 수

(6) miRNA/mRNA 3’ UTR 결합 구조에서 상보적으로 결합하는(match) 염기의 수
(7) miRNA 5’ 영역 8nt를 제외한 나머지 영역과 mRNA 결합구조의 자유에너지
(8) RNAdistance를 이용하여 RNA 2차 구조의 거리를 계산한 값
(9) miRNA/mRNA 결합 구조에 따른 Markov Chain probability를 계산한 값
(10) miRNA 5’ 영역과 mRNA 결합 구조에 따른 Markov Chain probability를 계산한 값

표 1. SVM의 입력

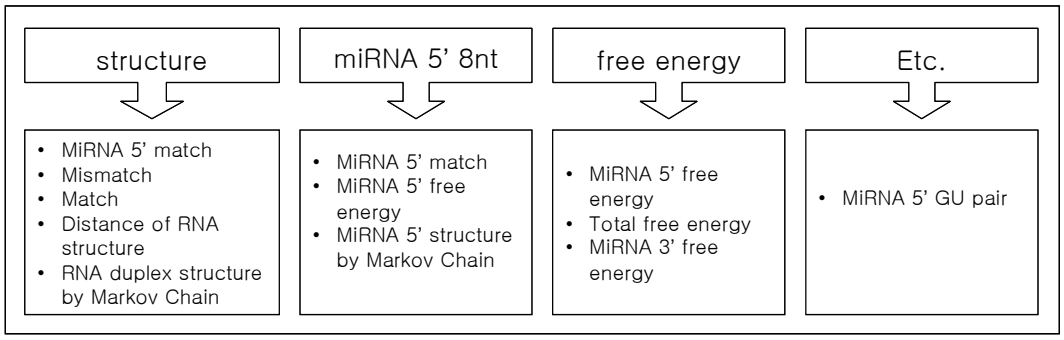

그림 4. 학습을 위한 특징 분석

표 1을 특징 별로 다시 정리하면 그림 4와 같이 (1) 구조 정보, (2) miRNA 5’ 영역의 8nt, (3) 자유 에너지, (4) 그 밖의 정보로 크게 네 가지로 나누어 생각할 수 있다. 첫 번째는 miRNA/target 구조 정보인데 이것을 추출하기 위해 다음 다섯 가지의 특징 정보를 이용하였다. miRNA 5’ 영역의 8nt 중에 일치(match) 개수, 전체 구조에서 불일치(mismatch)와 일치(match)의 개수, RNAdistance를 이용하여 2차 구조의 거리 측정 값, 구조 정보에 따른 Markov Chain probability 값이 그것이다. 둘째는 target 여부를 결정하는데 가장 중요한 특징으로 알려진

miRNA 5’ 영역의 8nt에 관한 정보인데 이것을 추출하기 위하여 miRNA 5’ 영역의 8nt에 해당하는 일치(match) 개수와 그것의 자유 에너지 정보를 사용하였다. 셋째는 target 여부를 결정하는 요소 중 하나인 열역학적 자유 에너지에 관한 특징 정보인데 이를 추출하기 위해 miRNA 5’ 8nt 영역, miRNA 3’ 영역, 그 둘을 포함하는 전체 영역, 이렇게 세 영역으로 나누어 자유 에너지를 각각 측정하였다. 마지막으로 그 밖의 요소로는 target 여부를 결정하는데 영향을 미치는 것으로 알려진 GU wobble pair의 개수에 관한 정보를 사용하였다.

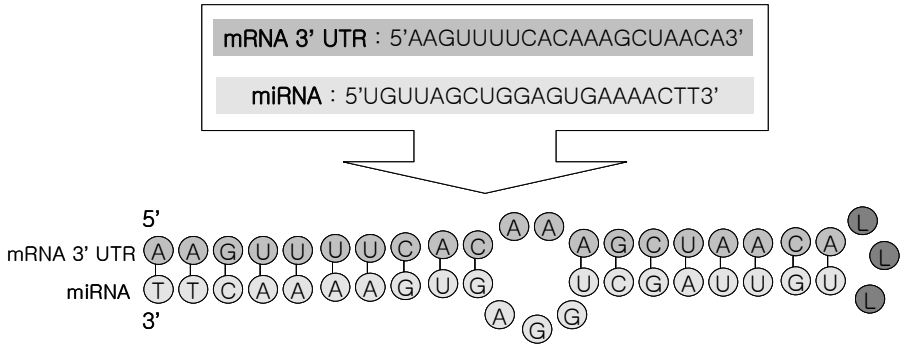

그림 5. mfold의 입력

표 1의 (2), (3)과 (4)에서 RNA 2차 구조와 자유 에너지의 측정은 mfold 3.1을 사용하였다(Mathews et al., 1999; Zuker, 2003). mfold의 입력은 그림 5와 같이 mRNA 바인딩 서열과 miRNA 사이에 “LLL”을 삽입하여 사용하였는데 “LLL”은 mfold에서 두 개의 분리된 RNA의 구조 예측을 위해 쓸 수 있다. 또한 표 1의 (4)와 (5)의 GU wobble pair의 개수와 (1), (7)과 (8)의 일치, 불일치의 개수는 mfold의 결과로 나온 구조로 계산하였다.

표 1의 (10)은 Vienna RNA Package의 RNAdistance라는 프로그램을 사용하였다(Shapiro, 1988; Shapiro et al., 1990; Fontana et al., 1993; Hofacker et al., 1994). RNAdistance는 RNA 2차 구조를 입력으로

받아 그것들의 유사 정도를 수치화 하여 결과로 보여주는 프로그램이다. 양성 데이터의 개수가 n_p 이고 새로운 2차 구조 서열 str_{query} 가 입력으로 들어왔을 때 각각의 양성 데이터의 2차 구조를 str_i 이라 하면 RNAdistance 프로그램을 이용한 점수(score)는 다음 수식 1과 같다.

$$score = \frac{\sum^{i=n_p} RNAdistance(str_i, str_{query})}{n_p}. \qquad (1)$$

표 1의 (11)은 그림 6의 구조에 따른 Markov Chain probability를 계산하였다. 그림 6과 같이 일치 상태(pair state)와 불일치 상태(mispair state)와 삭제 상태 (deletion state)의 총 세가지 상태로 구조를 구성하였다. 각각 상태에 따라 구조에 기반한 점수를 수식 (2)와 같이 구성하였다. i번째 위치에서 j와 같은 구조를 가질 때 $f(x_{ij})$ 는 그 위치에서 특정 구조에 대한 확률이고, $s(cons)$ 는 log값이 0이 되는 것을 방지하기 위한 임의의 매우 작은 수, $p(x_{ij})$ 는 백그라운드 확률(background probability)이다. 이에 따른 Markov Probability는 수식 (3)과 같은 조건부 확률을 가정할 경우 수식 (4)를 통해 구할 수 있다. 수식 (3)의 s와 t는 염기 쌍이 가질 수 있는 임의의 구조를 나타낸다.

pair
A/U C/G G/U U/A G/C U/G
mispair
A/C A/G A/A C/A C/U C/C U/C U/U G/A G/G
deletion
−/A −/U −/C −/G

그림 6. Markov Chain을 위한 세가지 상태

$$Score(x_{ij}) = \log(\frac{f(x_{ij}) + s(con)}{p(x_{ij})}) . \quad (2)$$

$$a_{st} = P[(x_i = t) | (x_{i-1} = s)] \quad (3)$$

$$P(x) = a_{0x_1} \prod_{i=1}^{L} a_{x_i x_{i+1}} \quad (4)$$

2.4 프로그램 진행 과정 (Program Procedure)

그림 7은 miRNA target 서열을 예측하는 프로그램의 진행과정을 나타낸다. 양성 데이터(positive data) 43쌍과 음성 데이터 (negative data) 1022쌍으로부터 구조 정보, 자유 에너지 정보 등 학습을 위한 특징 데이터를 추출한다. 추출된 데이터로부터 Weka라는 프로그램을 이용하여 SVM 분류 방법으로 학습하였다. Weka는 데이터 마이닝을 위한 기계 학습 알고리즘들을 Java로 구현한 소프트웨어이다. 실험 결과는 5-fold cross validation으로 제시 하였고 이를 바탕으로 테스트 데이터의 target 여부를 예측하였다.

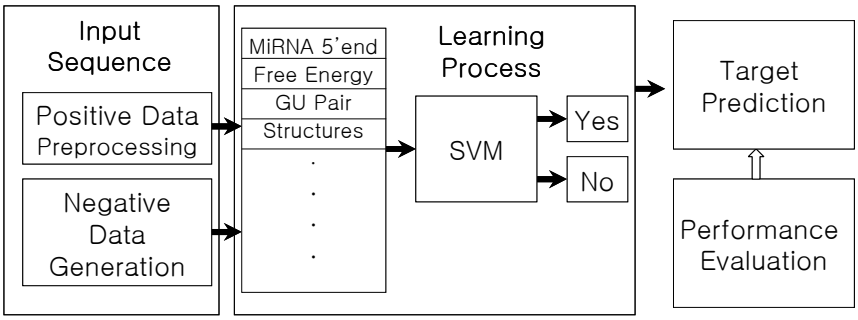

그림 7. 프로그램 진행 과정

2.5 SVM 소개

본 논문은 SMO(Sequential Minimal Optimization)를 이용한 SVM을 통해 miRNA target을 예측하였다. SVM은 데이터를 두 개의 카테고리로 분류하는 문제에 적합한 방법으로 1995년 Vapnik에 의해 제안되었으며 통계적인 학습 이론 Statistical Learning Theory에 기반한다. SVM은 기존의 통계적인 학습 이론(Statistical Learning Method)들에서 이용되는 경험적인 에러 최소화(Empirical Risk Minimization; ERM)와는 다른 구조적 위험성 최소화(Structural Risk Minimization; SRM)를 이용하여 에러를 줄여나가는 방법을 취하고 있다. 현재까지 Polynomial Machine, Radial Basis Function Machine, Two-layer Network Machine 이렇게 세가지 형태의 Kernel 함수가 많이 사용되고 있다. SVM 기법은 이런 Classifier 들을 이용하여 문제 공간의 비선형적인 높은 차수를 Feature Space에 선형적으로 투영하여 해석할 수 있도록 하며, 각 Feature 사이의 최적의 경계면을 제시한다. 다만 속도가 느리다는 단점이 있는데 본 논문에서는 SMO를 이용한 SVM을 통해 학습하여 이를 극복하였다. (Cortes and Vapnik, 1995; Vapnik, 1995; Platt et al., 1998; Keerthi et al., 1999)

SVM은 학습 데이터가 $\{(x_i, d_i) \mid i = 1,...,N\}$ 일 때 이것을 $d_i \in \{-1,1\}$ 로 분류하는 최적의 분리 경계면을 $f(x) = w^t x + b$ 로 놓는다. 이때 분리 경계면과 가장 인접한 점인 support vector와 $f(x)$ 의 거리를 $\frac{1}{\|w\|}$ 로 나타낼 수 있는데, SVM은 $\|w\|^2$ 을 최소화 하여 분리 간격을 최대화 하는 최적의 경계면을 찾는다. 이 문제는 수식 (5)와 같이 블록 최적화(convex optimization)문제로 나타낼 수 있다.

$$\min \frac{1}{2}\|w\|^2 \qquad (5)$$

$$\text{subject to } y_i(w \cdot z_i - b) \geq 1 \text{ where } i = 1,...,n.$$

이것을 라그랑즈 배수(Lagrange multiplier)로 유도하면 커널 함수가 $k(x_i, x_j)$일 때 수식 (6)과 같은 형식으로 바꿀 수 있다.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$
$$\text{subject to } \alpha_i \geq 0 \text{ where } i = 1,...,n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (6)$$

경계면의 hard-margin constraints를 유연하게 하기 위해 slack-variables ξ_i를 유도하면 수식 (5)로부터 수식 (7)과 같이 쓸 수 있다. 여기서 C는 empirical error와 복잡도를 결정하는 변수이다.

$$\min_{w,b\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad (7)$$

위 수식을 바탕으로 비선형 데이터를 분리하기 위한 모델은 수식 (8)과 같이 표현할 수 있다.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$
$$\text{subject to } 0 \leq \alpha_i \leq C \text{ where } i = 1,...,n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (8)$$

위 모델에서 slack-variables ξ_i로부터 라그랑즈 배수인 $\alpha_i \leq C \text{ where } i = 1,...,n$을 구할 수 있다. 본 논문에서는 커널 함수 $k(x_i, x_j)$로 polynomial machine을 사용했다.

학습에는 Weka(version 3.4)의 SMO를 분류기를 사용하였다. 양성

학습 데이터 43쌍, 음성 학습 데이터 1022쌍, 총 1065쌍을 입력으로 사용하였으며 입력 노드는 11개, 출력 노드는 예/아니오 2개이다. 복잡도에 관한 변수인 C값은 4, exponent는 5.0를 입력하였고 cross validation에 쓰이는 random seed 값을 임의로 주어가며 나머지는 기본 값을 사용하였다.

제 3 장 실 험 결 과

3.1 알고리즘의 성능

양성 데이터 43쌍과 음성 데이터 1022쌍을 입력으로 5-fold cross validation 방법으로 평가한 결과를 표 2에서 보여준다. 민감도는 64%, 특수도는 98%의 성능을 보인다.

TP	FN	FP	TN	Sensitivity	Specificity	PPV
25	14	23	999	0.64	0.98	0.52

표 2. training data의 5-fold cross validation 결과

표 3에서는 위 실험이 얼마나 정확한지 학습 데이터를 포함하여 지금까지 miRNA target이라고 실험으로 검증되었거나 예측된 miRNA/target 32쌍과 음성 데이터 1828쌍을 이용하여 본 논문에 의한 방법으로 성능을 평가하고 TargetScan (Lewis et al., 2003) 이라는 프로그램과 비교해보았다. TargetScan은 여러 종의 보존된 3' UTR 데이터베이스에서 서열 비교 방법에 의해 RNA/RNA 결합 구조를 열역학적 방법 기반으로 찾아나가는 모델링 방법으로 miRNA target을 예측하는데 가장 많이 사용되는 프로그램이다. 성능을 비교해놓은 표 3을 보면 민감도와 특이도의 경우 본 논문에 의해 제시된 방법은 TargetScan에 비해 거의 같거나 약간 더 좋은 성능을 보인다. 그러나 여기서 중요한 것은 false positive의 개수가 77에서 35로 절반 이상 줄었다는 점이다. 이 결과로부터 SVM classifier가 TargetScan보다 더 효율적이고 정확한 결과를 낸다는 점을 알 수 있다. 또한 성능을 비교하기 위한 32개의 양성 데이터 중 TargetScan을 개발한 후 그것의 검증을 위해 사용한 Lewis et al. 논문의 11개 데이터를 제외하면 본 논문에 의한 방법의 민감도는 0.74로 TargetScan의 민감도 0.53보다 성능이

매우 앞서는 것을 확인 할 수 있다.

Program	TP	FN	FP	TN	Sensitivity	Specificity	PPV
SVM	21[§]/14[†]	11/5	35	1793	0.66/0.74	0.98	0.38/0.29
TargetScan	20/10	12/9	77	1751	0.63/0.53	0.96	0.21/0.11

표 3. test data를 이용하여 TargetScan과의 성능 비교. "[§]"은 양성 데이터로 트레이닝 데이터를 포함하여 지금까지 miRNA target이라고 실험으로 검증되었거나 예측된 miRNA/target 32쌍을 사용한 경우이다. "[†]"은 TargetScan을 검증하기 위해 Lewis et al. 가 제시한 11개의 데이터를 제외한 결과이다.

표 4는 표 3의 양성 데이터 32쌍에 대한 결과를 구체적으로 보여준다. 'Actual result'는 miRNA/target쌍과 실제로 target 여부를 나타내고, 'SVM'은 본 논문의 알고리즘에 의해 예측된 결과이다. 또한 'TargetScan'은 TargetScan로 예측한 결과를 보여준다. 본 알고리즘에 의해 target이라고 판명된 경우 mRNA/miRNA의 서열과 2차 구조도 함께 보여주었다.

miRNA	target	miRNA 5' region ΔG	total ΔG	Sequence Vienna Package Type structure	Actual result	SVM	Target Scan	Reference
miR-13a	hb	-	-	-	yes	no	no	(Abrahante et al. , 2003)
miR-4	hb	-	-	-	yes	no	no	(Abrahante et al. , 2003)
miR-3	hb	-11.0	-21.9	tcgagacttaagatgtgagcccagtgtLLLucacugggcaaaagugugucuca ..(((((.....(..(..(((((((((.....)))))))))...)..))))))).	yes	yes	no	(Abrahante et al. , 2003)
miR-4	m4			-	yes	no	yes	(Lai, 2002)
miR-7	Tom	-9.5	-22.3	tcttagccgaatcattgtcttccaLLLuggaagacuagugauuuuguugu ...((((.((((((((((((((((.....)))))))).)))))))).)))).	yes	yes	yes	(Lai, 2002)
miR-14	Drice	-	-	-	yes	no	no	(Xu et al., 2003)
Lin-4	lin-14	-11.5	-17.5	tcacaaccaactcagggaLLLucccugagaccucaaguguga (((((.....(((((((((...)))))))).....)))))	yes	yes	yes	(Wightman et al., 1993; Ha et al., 1996)
Lin-4	lin-28	-11.5	-18.6	aaattgcactctcagggaLLLucccugagaccucaaguguga ...(..((((((((((((...))))))))).....)))..)	yes	yes	yes	(Moss et al., 1997)
Let-7	lin-41	-9.8	-24.2	tttttatacaaccattctgcctctLLLugagguaguagguuguauaguu ...((((((((((...(((((((.....)))))))..)))))))))..	yes	yes	no	(Reinhart et al., 2000; Slack et al., 2000)
Let-7	hbl-1	-9.6	-15.8	cagactatctcgcactttcattctacctcaLLLugagguaguagguuguauaguu ..((((((..((.((((.....(((((((.....)))))))..)))))).))))))	yes	yes	yes	(Abrahante et al. , 2003; Lin et al., 2003)
Batam	hid	-7.7	-17.3	atcatcatattcaaattggtctcaLLLugagaucauuuugaaagcugauu ...((((..((((((.((((((.....)))))))..)))))...)))).	yes	yes	yes	(Brennecke et al., 2003)
Lin-4	lin-41	-7.6	-19.6	gaatatttgaaatctcaggaaLLLucccugagaccucaaguguga ..(((((((..(((((((.......)))))))..))))))))..	yes	yes	no	(Slack et al., 2000)
Let-7	lin-14	-9.6	-23	ttattatgcaacaattctacctcaLLLugagguaguagguuguauaguu ..((((((((((....(((((((.....)))))))...)))))))))).	yes	yes	yes	(Reinhart et al., 2000)

Let-7	lin-28	-8.6	-20.2	taaaccatactaccacctacctccLLLugagguaguagguuguauaguu ..(((.((((.(((..(((((((.....)))))))..))).)))).)))	yes	yes	yes	(Moss and Tang, 2003)
miR-7	HLHm3	-9.5	-25.1	tgcaacaagatccgttgtcttccaLLLuggaagacuagugauuuuguugu .((((((((((....(((((((.....)))))))....))))))))))	yes	yes	yes	(Stark et al., 2003)
miR-7	hiary	-9.5	-22.8	taacagcaaatcagcaaaagtcttccaLLLuggaagacuagugauuuuguugu ..(((((((....((..((((((((.....))))))))..))..)))))))	yes	yes	yes	(Stark et al., 2003)
miR-2	reaper	-5.5	-25.7	ttactcatcaaagcgattgtgataLLLuaucacagccagcuuugaugagc ...((((((((((..((((((((...))))))))...)))))))))).	yes	yes	no	(Stark et al., 2003)
miR-2	grim	-5.6	-23.7	gctcaatcaaagcgcattgtgattLLLuaucacagccagcuuugaugagc (((((.(((((((...(((((((.....)))))))...)))))))))))	yes	yes	no	(Stark et al., 2003)
mir-375	Mtpn	-	-	-	yes	no	no	Poy et al., 2004
miR-26a	SMAD1	-5.7	-14.7	tgagccttgcatgtacttgaaLLLuucaaguaauccaggauaggcu ..((((((....(((((((((.....)))))))...))....))))))	yes	yes	yes	(Lewis et al., 2003)
miR-23a	SDF-1	-	-	-	yes	no	yes	(Lewis et al., 2003)
miR-23a	BRN-3b	-	-	-	yes	no	yes	(Lewis et al., 2003)
miR-101	ENX-1	-6.9	-18.8	gcttcaggaacctcgagtactgtgLLLuacaguacugugauaacugaag .((((((.....((((((((((...))))))))))..))))...))))))	yes	yes	yes	(Lewis et al., 2003)
miR-101	N-MYC	-	-	-	yes	no	yes	(Lewis et al., 2003)
miR-19a	PTEN	-5.7	-17.4	aactgttagggaattttacttgaaLLLuucaaguaauccaggauaggcu ..(((((..(((...(((((((((.....)))))))))))..)))))...	yes	yes	yes	(Lewis et al., 2003)
miR-34	Delta1	-11.2	-16.8	acatgccactcgctgcctLLLuggcagugucuuagcugguugu (((.((((((((((((.....)))))))....)).)))))))	yes	yes	yes	(Lewis et al., 2003)
miR-1	G6PD	-8.4	-14.5	tcagtgccacttgacattcctLLLuggaauguaaagaaguaugua ...((((..(((.(((((((.....)))))))..))).))))...	yes	yes	yes	(Lewis et al., 2003)

miR-1	BDNF	-8.2	-13.8	gggcatggtatttgagacattccaLLLuggaauguaaagaaguaugua ..((((((...(((...(((((((.....)))))))...))).)))))).	yes	yes	yes	(Lewis et al., 2003)
miR-34	Notch1	-	-		yes	no	yes	(Lewis et al., 2003)
miR-130	MCSF	-8.8	-18.3	cccctcatgaaggaagccattgcactgLLLcagugcaauguuaaaagggc .((((..(((.......((((((((((...))))))))))))))..)))).	yes	yes	yes	(Lewis et al., 2003)
miR-19a	MECP2	-	-	-	no	no	yes	(Lewis et al., 2003)
miR-34	VAMP2	-11.4	-18.0	gggtactagtcactgccaLLLuggcagugucuuagcugguugu ....((((((((((((.....))))))).....))))))...	no	yes	yes	(Lewis et al., 2003)

표 4. TargetScan과 SVM classifier의 양성 데이터 비교 결과

3.2 특징 분석 (feature analysis)

어떤 요인이 가장 성능에 영향을 미치는지 알아보기 위해 한가지 이상 특징을 제외하고 위와 같은 조건에서 실험해 보았다. 그림 8의 (1)부터 (10)은 표 1의 (1)부터 (10)을 의미하는 것으로 각 해당하는 번호의 특징을 제외하고 나머지 특징을 SVM의 입력으로 했음을 의미한다. 각각 조건에 따라 특이도는 거의 변하지 않은 상태에서 (97%~98%) 각 특징이 결과에 얼마나 영향을 미치는지에 대한 민감도의 변화를 그림 8에서 보여주었다. 각 경우의 민감도가 갖는 값이 낮을수록 결과에 큰 영향을 미치는 요인이라고 할 수 있다. 민감도가 가장 낮은 세 가지로는 (4) miRNA 5' 영역 8nt의 G/U wobble pair 개수 (1) miRNA 5' 영역 8nt에서 mRNA 3' UTR과 상보적으로 결합하는 염기의 개수 (10) miRNA 5' 영역과 mRNA 결합 구조에 따른 Markov Chain Probability를 계산한 값이다. 이 특징들이 miRNA target을 결정하는데 가장 중요한 요인이라 할 수 있는데 모두 miRNA 5' 영역에 관한 값임을 알 수 있다. 그림 9는 자유 에너지에 관한 값, 구조에 관한 값, miRNA 5' 영역에 관한 값들만 제외하고 성능의 변화를 살펴보았는데 이 실험에서도 miRNA 5' 영역을 제외했던 실험 성능이 가장 낮아서 이 특징이 가장 중요한 역할을 하는 것임을 알 수 있었다. 즉 이 실험을 통해 miRNA 5' 영역이 target 여부를 결정짓는 가장 중요한 요인이라고 생각 할 수 있다. 이것은 miRNA 서열에 변이를 주어서 3' UTR과 결합 정도를 분석한 결과, miRNA 5' 영역이 miRNA target을 결정하는데 가장 중요한 역할을 한다는 실제 실험을 (Doench et al., 2004) 지지하는 결과이다.

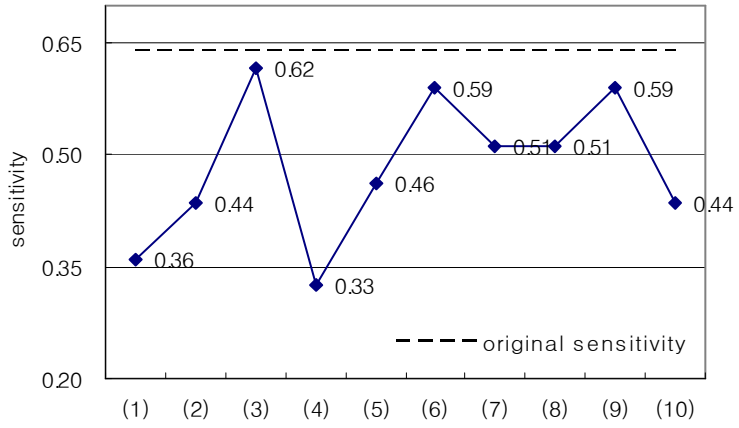

그림 8. 표 2에 나열된 각각의 요인이 SVM의 성능에 미치는 영향.

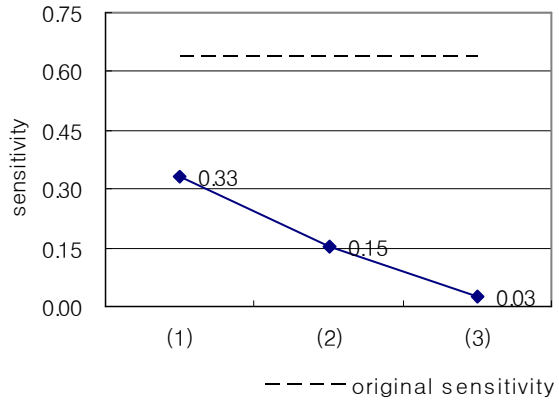

그림 9. (1) 자유 에너지에 관한 정보를 모두 제외하고 실험 (2) 구조에 관한 정보를 모두 제외하고 실험 (3) miRNA 5' 영역에 관한 정보를 모두 제외하고 실험 한 후 민감도에 따른 성능 변화

3.3 microRNA target 예측

본 논문에서 제시된 방법에 의해 예쁜 꼬마 선충과 인간, 쥐 데이터베이스에서 miRNA의 target을 예측해 보았다. 우선 target 유전자가 가장 많이 알려진 *lin-4*와 *let-7*의 target을 *C. elegans*에서 찾아보았다. 표 5는 *C. elegans* 3' UTR 데이터 베이스에서 본 논문의 알고리즘에 의해 *lin-4*와 작용하는 유전자 찾은 결과이다. 총 5개를 target 후보 유전자(candidate gene)로 찾았는데 그 중에서 *lin-4*와 작용한다고 검증된 4개가 모두 포함되어 있는 것을 확인할 수 있다.

genes	products	accession number	miRNA 5' G	total G
lin-41A	LIN-41A	AC: CC085391	-7.6	-19.6
lin-28	LIN-28	AC: CC013576	-11.5	-18.6
lin-14	LIN-14A	AC: CC013457	-11.5	-17.5
nhr-11	nuclear receptor NHR-11	AC: CC073419	-11.5	-16.5
hbl-1	hunchback-related protein	AC: CC073457	-7.6	-12.5

표 5. *C. elegans*에서 *lin-4*와 작용하는 유전자 목록. target이라고 알려진 *lin-41, lin-14, hbl*과 *lin-28*을 포함하고 있다.

표 6은 표 5과 마찬가지로 *C. elegans* 3' UTR 데이터 베이스에서 본 논문의 알고리즘에 의해 *let-7*과 작용하는 유전자를 찾아 자유 에너지가 높은 순으로 정렬한 결과이다. *let-7*도 마찬가지로 *let-7*과 작용한다고 알려진 5개의 유전자 *lin-41, daf-12, lin-14, hbl-1, lin-28*이 모두 포함되어 있었다.

genes	products	accession number	miRNA 5' ΔG	total ΔG
lin-41A	LIN-41A	AC: CC085391	-9.8	-24.2
daf-12	DAF-12 A2	AC: CC073740	-9.8	-23.4
lin-14	LIN-14A	AC: CC013457	-9.6	-23
hbl-1	hunchback-related protein	AC: CC073457	-8.6	-22.2
lin-28	LIN-28	AC: CC013576	-8.6	-20.2
unc-129	UNC-129	AC: CC054258	-9.8	-20
CePqM96	paraquat-inducible protein	AC: CC013365	-6.9	-19.8
-	nuclear receptor NHR-43	AC: CC125308	-7.8	-19.5
skr-21	SKR-21	AC: CC181290	-9.6	-18
mio	Mlx interactor	AC: CC125253	-9.8	-17.2
ces-2	CES-2	AC: CC013549	-9.6	-16.2
-	-	AC: CC013431	-9.8	-14.8
unc-16	UNC-16	AC: CC181137	-6.7	-14.6
wrk-1	immunoglobulin domain-containing protein WRK-1C	AC: CC181358	-9.8	-14.5
pip-1	PIP-1	AC: CC228823	-8.9	-14
daf-4	BMP receptor	AC: CC013410	-8	-13.9
daf-16	DAF-16	AC: CC046572	-8.6	-13.7
-	histone H1.Q	AC: CC085543	-9.6	-13.7
unc-2	high voltage activated calcium channel alpha-1	AC: CC230280	-7	-12.2
-	histone H1.1	AC: CC012659	-9.8	-11.6
-	sodium-calcium exchanger	AC: CC049470	-5.2	-10.1
unc-115	putative actin-binding protein UNC-115	AC: CC060593	-6.9	-9.7

표 6. *C. elegans*에서 *let-7*과 작용하는 유전자 목록

표 7은 target이 아직 알려지지 않은 *cel-miR-228, cel-miR-229, cel-miR-230, cel-miR-231*과 작용하는 유전자를 *C. elegans* 3' UTR 데이터 베이스에서 본 논문의 알고리즘에 의해 유전자를 예측한 결과이다.

genes	products	accession number	miRNA 5'ΔG	total ΔG
cel-miR-228				
unc-75	putative RNA-binding protein	AC: CC266578	-8.9	-25.6
vab-2	VAB-2	AC: CC085460	-8.8	-16.5
vab-10	VAB-10A protein	AC: CC231603	-7.5	-16.2
lin-9	LIN-9L	AC: CC103462	-7.5	-16.1
pcr55	transmembrane protein	AC: CC013339	-8.9	-15.7
pme-1	poly ADP-ribose metabolism enzyme-1	AC: CC181762	-9.1	-15.1
tim9b	small zinc finger-like protein	AC: CC084885	-8.9	-13.4
ehs-1	EHS-1	AC: CC126116	-7.5	-12.8
-	methuselah-like protein MTH-1	AC: CC279142	-6.9	-11.6
cel-miR-229				
-	Na/Ca,K-exchanger	AC: CC054777	-6.8	-13.2
cel-miR-230				
mab-21	mab-21 protein	AC: CC103635	-5.5	-13.6
cel-miR-231				
klp-12	kinesin like protein KLP-12	AC: CC121004	-7.7	-10.8
mom-1	MOM-1	AC: CC012616	-7	-21.6
let-413	LET-413 protein	AC: CC103637	-8.5	-12.7
let-23	tyrosine kinase	AC: CC013452	-8.5	-10.9

표 7. *C. elegans*에서 *cel-miR-228, cel-miR-229, cel-miR-230, cel-miR-231*과 작용하는 유전자 목록

miRNA는 종간의 보존도가 높아서 그것과 작용하는 유전자의 3' UTR 역시 종간에 보존되어 있을 것으로 생각된다. human과 mouse에서 공통적으로 발견된 *hsa-miR-199a*와 작용하는 유전자를 각각의 3' UTR 데이터베이스에서 찾아보았다. 표 8은 target이라고 예측한 유전자 중에서 human과 mouse에 공통적으로 작용하는 유전자의 목록을 보여준다. 이러한

표를 바탕으로 miRNA와 mRNA발현 정도, Gene Ontology(GO)를 이용하여 분석하면 더 정확한 예측을 할 수 있다.

gene	product	Human accession number Mouse accession number	miRNA 5' ΔG miRNA 5' ΔG	total ΔG total ΔG
Borg4		AC: CC108693;	−9.1	−19.5
		AC: CC106890;	−10.1	−15.6
	peptidylarginine deiminase type III	AC: CC078185;	−7.9	−24.4
		AC: CC076716;	−7.9	−17.9
hPAD−colony10	peptidylarginine deiminase type I	AC: CC108610;	−6.3	−14
		AC: CC076715;	−8.3	−13.2
sox11	SOX11	AC: CC078207;	−8.1	−16.2
		AC: CC047859;	−6.1	−20.8
ICAT	beta−catenin−interacting protein ICAT	AC: CC095058;	−6	−20.3
		AC: CC106789;	−7.1	−21.1
	dihydropyrimidinase related protein 4	AC: CC059253;	−6.8	−19.6
		AC: CC062600;	−7.5	−16.3

표 8. mouse와 human에서 hsa−miR−199a과 공통으로 작용하는 유전자 목록

제 4 장 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. 학습의 입력을 다르게 주어 miRNA seed의 정보와 자유 에너지가 결합 여부를 결정 짓는 매우 중요한 요소라는 것을 실험을 통해 증명하였고, 결합 여부를 결정 짓는 것은 결합 염기 쌍의 수보다 결합 자유 에너지가 더 큰 영향을 미치는 것을 알아냈다. 따라서 G/C 결합과 같은 단단한 결합 상태가 많을수록 더 안정적이었으며 G/U wobble pair는, 특히 miRNA seed 영역에는, 거의 나타나지 않았다. 그리고 miRNA/target의 2차 구조 정보와 SVM 분류 방법을 이용함으로써 false positive의 수를 급격히 줄일 수 있었다. 본 논문은 다른 종간의 보존도를 입력으로 사용하지 않았기 때문에 보존된 유전자가 밝혀지지 않은 경우에도 유용하다는 장점이 있다.

SVM의 입력으로는 본 논문이 제시한 열 가지 특징 외에도 mRNA 3’ UTR의 결합 부위 근처의 프로파일, mRNA 3’ UTR의 종간 보존도, miRNA/mRNA 결합 구조를 루프나 bulge의 크기와 위치, 이중 서열의 길이 등을 더 자세히 수치화 하여 입력에 추가할 수 있고 siRNA 생성 프로그램을 miRNA target 예측에 이용할 수도 있다. 또한 miRNA target여부는 miRNA와 mRNA 발현 정도, 생화학적인 방법, 유전적인 분석 등을 종합적으로 고려해서 판단해야 한다. 그리고 실제 실험을 통해 target 여부의 검증이 필요할 것으로 보인다.

miRNA가 target 유전자에 어떻게 작용하는지 기작에 관한 연구는 miRNA/target의 구조적인 지식을 더욱 풍부하게 하여 miRNA target의 기능을 밝히는데 더 정확한 예측을 가능하게 할 것으로 생각된다. 또한 다양한 케이스 스터디를 통해 환경이나 특정 상황의 지식을 학습하는 것도 좋은 성과가 있을 것으로 기대된다.

참고문헌

Abrahante, J.E., Daul, A.L., Li, M., Volk, M.L., Tennessen, J.M., et al., (2003) The *Caenorhabditis elegans hunchback*-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev Cell* **4**, 625-637.

Ambros, V., (2001) microRNAs: tiny regulators with great potential. *Cell* **107**, 823-826.

Ambros, V., (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* **113**, 673-676.

Aravin, A.A., Naumova, N.M., Tulin, A.A., Rozovsky Y.M., and Gvozdev, V.A., (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in Drosophila melanogaster germline. *Curr.Biol.* **11**, 1017-1027.

Aukerman, M.J., and Sakai, H., (2003) Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* **15**, 2730-2741.

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

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

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M., (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila. Cell* **113**, 25-36.

Carrington, J.C., Ambros, V., (2003) Role of microRNAs in plant and animal development. *Science* **301**, 336-338.

Chen, X. (2003). A MicroRNA as a translational repressor of APET-ALA2 in arabidopsis flower development. *Science* **303**, 2022-5

Cortes, C. and Vapnik, V., (1995) Support-vector network, *Machine Learning* **20**, 273-297

Doench, J.G. and Sharp, P.A., (2004) Specificity of microRNA target selection in translational repression, *Genes Dev.*, **18**, 504-11.

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

Fontana W., Konings D. A. M., Stadler P. F., Schuster P., (1993) Statistics of RNA secondary structures, *Biopolymers* **33**, 1389-1404.

Ha I., Wightman B., Ruvkun G., (1996) A bulged *lin-4/lin-14* RNA duplex is sufficient for *Caenorhabditis elegans lin-14* temporal gradient formation. *Genes Dev.* **10**, 3041-3050.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Mh. Chemie* **125**, 167-188.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K., (1999) Improvements to platt's SMO algorithm for SVM classifier design. *Neural Computation* **13**, 637-649.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., Hatzigeorgiou, A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* **18**, 1165-1178

Lai, E.C., (2003) MicroRNAs: runts of the genome assert themselves *Curr. Biol.* **13**, R925-R936.

Lai E.C., (2002) MicroRNAs are complementary to 3' UTR sequence motifs that mediate negative posttranscriptional regulation. *Nat Genet*

30, 363-364.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843-854.

Lee, Y., Jeon, K., Lee,J., Kim,S., and Kim, V., (2002) microRNA maturation: stepwise processing and subcellular localization, *EMBO Journal* **21**, 4663-70.

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

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., Slack, F.J., (2003) The *C. elegans* hunchback homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. *Dev. Cell* **4**, 639-650.

Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H., (1999) Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.* **288**, 911-940

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* **88**, 637-646.

Moss, E.G., (2002) MicroRNAs: hidden in the genome. *Curr. Biol.* **12**, R138-R140.

Moss, E.G., Poethig, R.S., (2002) MicroRNAs: something new under the sun. *Curr. Biol.* **12**, R688-R690.

Moss, E.G., Tang L., (2003) Conservation of the heterochronic regulator *lin-28*, its developmental expression and microRNA complementary sites. *Dev Biol* **258**, 432-442.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivassan, A., Fishman, M.,Hayward, D., Ball, E., et al., (2000) Conservation across animal phylogey of the sequence and temporal regulation of the 21 nucleotide *let-7* heterochonic regulatory RNA. *Nature* **408**, 86-89.

Platt, J., (1998) Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. Burges, and A. Smola, editors, Advances in kernel methods - support vector learning. MIT Press.

Poy, M.N., Ellasson L., Krutzfeldt J., Kuwajima S., Ma X., MacDonald P.E., Pfeffer S., Tuschi T., Rajewsky N., Rorsman P., Stoffel M., (2004) A Pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226-30.

Rehmsmeier, M., Steffen, P., Hochsmann, M., Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507-1517

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., et al., (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901-906.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* **110**, 513-520.

Shapiro, B.A., (1988) An algorithm for comparing multiple RNA secondary structures, *CABIOS* **4**, 381-393.

Shapiro, B.A., Zhang, K. (1990) Comparing multiple RNA secondary structures using tree comparison, *CABIOS* **6**, 309-318.

Slack F.J., Basson M., Liu Z., Ambros V., Horvitz H.R., et al., (2000) The

lin-41 RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol Cell* **5**, 659-669.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003) Identification of Drosophila MicroRNA Targets. *Plos Biology* **1**, E60

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855-862.

Xu, P., Vemooy, S.Y., Guo, M., and Hay, B.A., (2003) The *Drosophila* MicroRNA *Mir-14* suppresses cell death and is required for normal fat metabolism. *Curr. BioL.* **13**, 790-795.

Zuker, M., (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-15.

ABSTRACT

MicroRNAs (miRNAs) are a class of noncoding RNAs found in various organisms such as plants and mammals, however, most of the mRNAs regulated by miRNAs are unknown. miRNA targets in genomes can not be identified by standard sequence comparison since their complementarity to the target sequence is imperfect in general. In this paper, we propose a method for the efficient prediction of miRNA targets using a kernel method. To help in distinguishing the false positives from potentially valid targets, we elucidate the features shared by experimentally confirmed targets. The performance of our prediction method was evaluated by five-fold cross-validation. Our method showed 0.64 and 0.98 for sensitivity and specificity, respectively. Also, the proposed method reduced the number of false positives by half compared to the TargetScan. We investigated the effect of feature sets on the classification of miRNA targets. Finally, we predicted miRNA targets for several miRNAs in the *C. elegans* 3' untranslated region (3' UTR) database. The targets predicted by the suggested method will help in validating more miRNA targets and ultimately in revealing the role of small RNAs in the regulation of genomes. Our algorithm for miRNA target site detection will be improved by additional experimental-knowledge. Also, increasing the number of confirmed targets is expected to reveal general structural features that can be used to improve their detection.