다층 퍼셉트론 신경망을 이용한 microRNA의 목표 유전자 예측 및 조절 메커니즘 분별

이민수[O1], 남진우[1,2,3], 장병탁[1,2,3]
서울대학교 컴퓨터공학부 바이오지능 연구실[1]
서울대학교 바이오정보기술 연구센터 (CBIT)[2]
서울대학교 대학원 생물정보학 협동과정[3]
{mslee, jwnam, btzhang}@bi.snu.ac.kr

Prediction of microRNA Targets and Discrimination of microRNA Regulatory Mechanisms using Multilayer Perceptron Neural Network

Min-Su Lee[O1], Jin-Wu Nam[1,2,3], and Byoung-Tak Zhang[1,2,3]
Biointelligence Laboratory, School of Computer Science and Engineering[1]
Center for Bioinformatics Technology (CBIT)[2]
Graduate Program in Bioinformatics[3]
Seoul National University, Seoul 151-742, Korea

요 약

miRNA 유전체학의 중요한 이슈로 miRNA가 조절하는 목표 유전자를 예측하는 작업과 miRNA가 목표 유전자를 조절하는 메커니즘이 무엇인지 규명하는 것을 들 수 있다. 본 논문에서는 생물학적 특징들과 다층 퍼셉트론 신경망을 이용하여 miRNA의 목표 유전자를 예측하고 해당 miRNA 조절 메커니즘 타입을 분별해주는 시스템을 제안하고 실제 데이터를 사용하여 그 성능을 평가한다. 실험적으로 검증된 데이터를 사용하여 제안 시스템을 평가해본 결과, 다층 퍼셉트론 신경망을 사용할 경우 84.63%의 정확도로 miRNA의 목표 유전자를 예측할 수 있었고, 87.90%의 정확도로 miRNA가 목표 유전자를 조절하는 메커니즘을 분별할 수 있었다. 학습 데이터가 충분히 많아진다면 제안 시스템의 예측 성능은 더욱 높아질 것으로 예상된다.

1. 서 론

2001년부터 발견된 miRNA (microRNA), siRNA (small interfering RNA) 등과 같은 'small RNA'들은 20~30 뉴클레오티드(nt)로 구성된 짧은 핵산 길이에도 불구하고 유전자 발현을 제어하고 세포의 기능을 총괄 조정하는 역할을 수행한다는 것이 밝혀지면서 저널 Science에서 2002년 10대 과학 업적 중 하나로 'small RNA의 발견'을 선정할 정도로 주목받고 있는 연구 대상이다. 최근 유전체 수준에서 small RNA들을 발굴하고 [1-3], 그들이 제어하는 목표 유전자들을 예측하는 [4,5] 연구들이 활발하게 이루어지고 있다. small RNA에 관한 연구의 최종 목표는 small RNA의 발현 과정과 기능, 더 나아가 small RNA에 의해 조절되는 기전(mechanism)에 대해 이해함으로써 질병 발병, 분화 과정 등과 같은 생명 현상에 관한 비밀을 밝히고 신약개발이나 질병 치료에 이용하는 것이라 할 수 있다.

miRNA는 조직 특이적, 발생 특이적으로 발현되며 목표 유전자가 발현되어 단백질로 합성되는 과정에서 deadenylation을 촉진하여 안정성을 떨어뜨려 mRNA의 분해를 촉진시키거나 (PTD, post-transcriptional degra

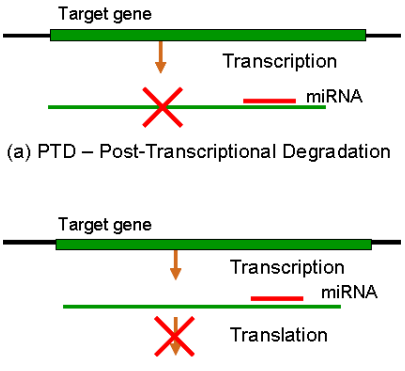

그림 1. miRNA의 두가지 조절 메커니즘

dation) (그림 1(a)), mRNA의 3' UTR(untranslated region) 부분에 상보적으로 결합하여 단백질로 번역되는 것을 억제함으로써 (TR, translational repression) (그림 1(b) 목표 유전자의 단백질 양을 조절한다.

miRNA 유전체학의 중요한 이슈로 miRNA가 조절하는 목표 유전자를 예측하는 작업과 miRNA가 목표 유전자를 조절하는 메커니즘이 PTD와 TR 중 무엇인지 규명하는 것을 들 수 있다.

miRNA의 목표 유전자를 예측하기 위해 miRNA가 조절하는 목표 mRNA의 유전자 위치 정보와 구조 정보 등의 생물학적 특징을 사용한 다양한 규칙 기반[6, 7] 또는 학습 기반[8] 방법들이 제안되었다. 그러나 이러한 miRNA 목표 유전자 예측 결과는 예측 민감도(sensitivity)는 높게 나오지만 특이도(specificity)가 낮다는 한계를 보이고 있다[5].

실험적으로 miRNA가 목표 유전자에 바인딩하는지 여부를 검증하는 *in vitro* reporter assay 방법은 *in vitro* 실험이므로 실제 세포 안에서는 서로 바인딩하지 않을 수도 있다는 한계점을 가지며 하나의 miRNA-목표유전자 쌍에 대해 실험하므로 수행 비용 및 시간이 많이 든다는 단점이 있다. Huang *et al*은 miRNA의 목표 유전자 예측 프로그램인 TargetScanS[6]의 결과로 산출된 각 miRNA의 목표 유전자의 신뢰성을 miRNA와 mRNA의 발현 데이터를 이용하여 평가함으로써 보다 높은 신뢰도의 miRNA의 목표 유전자들 선별하는 연구를 수행하였다[9, 10]. Huang *et al*의 연구는 miRNA 발현량 증가하면 목표 mRNA 발현량이 감소하는 경향을 보이는 PTD 메커니즘의 특징을 이용하여 miRNA의 목표 유전자 조절 여부에 대한 검증을 수행함으로써 PTD 타입에 해당하는 miRNA-목표 유전자 쌍들만 검증할 수 있다는 한계점을 가지고 있다.

PTD 메커니즘에 의해 miRNA가 목표 유전자를 조절하는 경우에는 직접적으로 목표 유전자의 mRNA 분해를 촉진시키므로 mRNA 발현량이 줄어들어 이를 발현 분석 기법으로 동정해볼 수 있는데 반해[3] miRNA가 목표 유전자를 TR 메커니즘으로 조절하는 경우에는 mRNA 발현량에는 변화 없이 mRNA가 단백질로 합성되는 것을 억제하므로, mRNA 발현 데이터의 변화 양상만으로도 예측된 miRNA-목표 유전자의 조절 메커니즘을 분별할 수 있을 것으로 생각할 수 있다. 그러나 현재 실험적으로 밝혀진 PTD 메커니즘들은 대부분이 microarray와 같은 유전자 발현 모니터링 기법을 사용하여 밝히고 있으므로, 목표 항목에 해당하는 TR/PTD 조절 메커니즘들을 분별하는 모델 구축하기 위해 발현 프로파일을 사용하여 학습할 경우에는 데이터 셋을 만들 때 사용된 속성이 다시 분류 속성으로 사용되는 오류를 범하게 된다. 따라서 발현 프로파일 이외의 다른 생물학적 특징들을 이용하여 각 miRNA이 어느 (PTD 또는 TR) 조절 메커니즘으로 목표 유전자를 조절하는지 여부를 밝히는 작업이 요구된다.

본 논문에서는 여러 가지의 생물학적 특징들과 다층 퍼셉트론 신경망 알고리즘을 사용하여 miRNA의 목표 유전자를 예측하고 그 조절 메커니즘까지도 함께 분별하여 제시해주는 연구를 수행한다.

2. 방법 및 시스템 구성

2.1. 시스템 구성

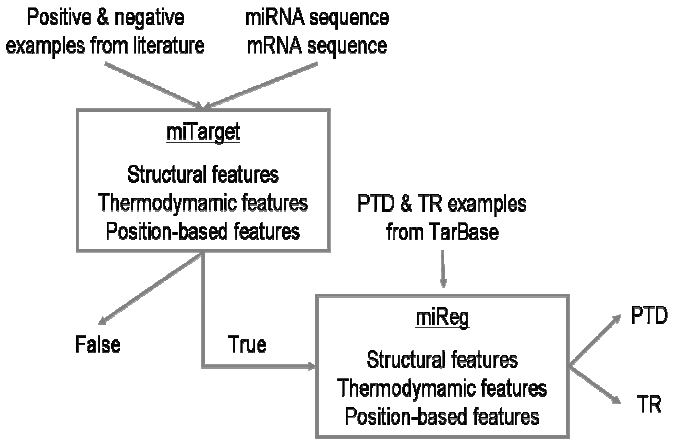

그림 2. 시스템 구성도

miRNA 목표 유전자 예측 및 miRNA 조절 메커니즘 분별 시스템은 그림 2와 같이 크게 miRNA 목표 유전자를 예측하는 모듈(miTarget)과 예측된 miRNA와 목표 mRNA 사이의 조절 메커니즘을 분별하는 모듈(miReg)로 구성되어 있다. miTarget 모듈에서는 문헌에서 수작업으로 모은 miRNA-목표 유전자 쌍 데이터와 miRNA의 mature 서열 데이터, miRNA와 상보적인 서열을 가지는 목표 유전자의 3'UTR 서열 데이터를 준비한 후, 각 miRNA-목표 유전자 쌍에 대해 생물학적 특징으로 사용될 수 있는 miRNA와 목표 유전자 mRNA가 결합했을 경우의 서열의 구조적, 열역학적, 서열위치상의 속성들을 계산하고, 이 속성들에 기반하여 miRNA에 대한 목표 유전자 예측 모델을 구축한다. miReg 모듈에서는 miTarget로부터 예측된 결과들 중 실험적으로 조절 메커니즘이 밝혀진 miRNA-목표 유전자 쌍 데이터를 이용하여 TR/PTD를 목표 항목으로 하는 분별 모델을 같은 생물학적 특징들을 사용하여 학습한다.

2.2. 데이터 셋

miRNA의 목표 유전자를 예측하는 모듈에서 사용한 학습 데이터는 9개의 문헌을 통해 실험적으로 검증된 152개 양성 miRNA-목표 유전자 쌍들과 83개의 음성 쌍들을 포함한 총 235개의 데이터로 구성되어 있다. 음성 데이터는 양성 데이터에 비해 분류 모델의 특이도에 더 영향을 끼치므로 충분한 음성 데이터의 확보가 중요하다. 그런데, 문헌에 기반하여 정리한 음성 데이터의 수가 효율적인 분류 모델을 구축하기에는 너무 작아 음성 데이터를 다음과 같은 방법으로 추가 생성하여 사용하였다.

miRNA에서 목표 mRNA와 결합하는 사이트(target sites)를 제거하면 더 이상 miRNA가 목표 mRNA를 조절하지 못한다. 따라서, 목표 mRNA 서열과 결합하는 사이트를 miRNA에서 제거한 후 남은 miRNA 서열과 목표 mRNA 서열에서 seed 부분이 짝을 이루는 부분이 존재한다면 이러한 데이터는 정확한 음성 데이터로 간주할 수 있다. 이러한 성질을 바탕으로 let-7이 조절하는 lin-41과 LIN-28의 서열 정보를 이용하여 seed 부분에서 4-mer 이상 매치하는 데이터들을 모으고 나머지들은 버림으로써 추가적인 163개의 음성 데이터들을 만들 수 있었다 [8]. 따라서 최종적으로 152개의 양성 데이터와 246개의 음성 데이터를 사용하여 miRNA의 목표 유전자를 예측하는 모델 학습에 사용하였다.

miRNA가 목표 유전자를 조절하는 메커니즘의 타입을 분별하기 위한 모델 학습을 위해, 실험적으로 밝혀진 miRNA 목표 유전자들을 수작업으로 정리한 TarBase [11]로부터 조절 메커니즘에 따라 PTD/TR로 분류되어 있는 204개의 TR 타입의 miRNA-목표 유전자 쌍과 386개의 PTD 타입의 miRNA-목표 유전자 쌍을 추가적으로 사용하였다.

서열정보에 기반하여 miRNA와 목표 유전자 사이의 구조적 특징, 열역학적 특징, 위치 기반 특징들을 계산하기 위해 miRBase[12]에서 mature miRNA 서열 정보를 Ensembl[13]에서 유전자의 3'UTR 서열 정보를 다운로드하여 사용하였다.

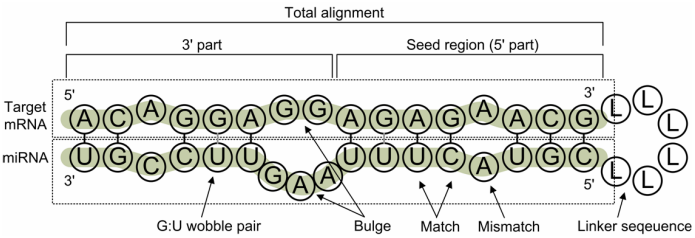

그림 3. miRNA와 목표유전자의 결합 구조도 [8]

2.3. 생물학적 특징 계산

구조적 특징 (structural features)

발현 프로파일상의 연관성과 더불어 기존에 사용되었던 구조적, 열역학적 특징과 위치기반의 특징이 함께 사용된다[8]. 구조적 특징에는 그림 3과 같이 전체 결합 부분과, seed부분, 3'부분으로 구별 지어, 각 match (GC, AU, GU, mismatch)의 개수와 GC+AU match, GU + mismatch의 개수를 각각 사용한다. 그리하여 총 18개의 구조적 특징을 사용하게 된다 (그림 3).

열역학적 특징 (thermodynamic features)

열역학적 특징은 각 결합부위의 자유에너지값을 사용한다. 전체 결합부분, seed부분, 3'부분으로 나누어 각각 자유에너지를 계산하고 그것을 특징으로 사용한다. 여기서는, 총 3개의 특징으로 구성된다 (그림 3).

위치 기반 특징 (position-based features)

miRNA-목표유전자 사이의 결합부위를 상대적 위치를 표시하고 그 위치에 따른 상태를 GC, AU, GU match 와 mismatch로 표시하여 각 특징으로 사용한다. 총 20개의 위치 기반 특징을 사용하게 된다 (그림 3).

2.4. 다층 퍼셉트론 신경망

단층 퍼셉트론이 복잡하게 얽혀있는 네트워크 구조로 구성되어 있는 다층 퍼셉트론 신경망 (MLPNN, multilayer perceptron neural network)은 크게 입력층, 은닉층, 출력층으로 나누어진다. 입력층에 들어간 입력신호는 각 연결강도(weight)와 곱해지고 각 노드(node)에서 더해진다. 출력층에서는 결과값과 실제값을 비교하여 오차가 작은 방향으로 노드간의 연결강도를 조절하며 델타 규칙(delta rule)을 이용하여 학습이 이루어진다.

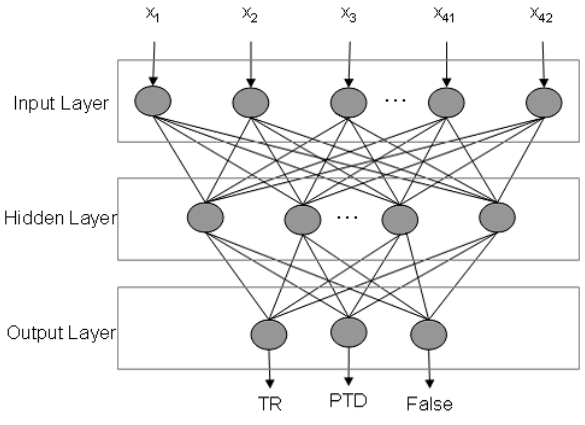

그림 4. 다층 퍼셉트론 신경망의 기본 구조

본 연구에서는 일반화된 델타 규칙이라고 불리는 역전파(back-propagation) 학습 알고리즘을 사용하여 다층 퍼셉트론 신경망을 학습하였다. 역전파 학습 알고리즘을 이용한 다층 퍼셉트론 신경망 학습 과정은 두 개의 순서에 의해 이루어진다. 첫째, 입력 벡터를 네트워크에 입력하면 이것이 네트워크의 전방향 (feed-forward)으로 전파되어 출력을 낸다. 이 출력과 목표 출력과의 차이에 미분계수를 곱하여 출력 노드에 대한 오차를 계산한다. 둘째, 오차신호가 네트워크의 역방향(backward)으로 전파되어 가면서 각각의 노드의 오차 신호가 계산되어, 이것을 바탕으로 연결강도를 수정한다.

본 논문에서는 다층 퍼셉트론 신경망 학습에 은닉층을 속성수와 목표항목 수의 평균값인 23개를 사용하였으며, 학습 반복 횟수를 500회로 제한하였다. 업데이트시의 연결강도에 적용되어 오차진동을 적게하여 수렴속도를 빠르게 해주는 momentum은 0.2로, 업데이트되는 연결강도의 학습 비율(learning rate)은 0.3으로 설정하고 학습하였다.

3. 결과 및 논의

miRNA와 유전자 쌍과 그들의 서열 데이터에 기반하여 계산된 생물학적 특징으로 구성된 학습 데이터는 Weka (version 3-5-5) 환경의 다층 퍼셉트론 신경망 모듈을 이용하여 학습되었다. 먼저 miRNA의 목표 유전자를 예측하는 모듈을 구축하였고, 그 결과를 이용하여 miRNA의 조절 메커니즘을 분별하는 모듈을 구축하였다.

제안 시스템에서의 MLPNN의 적합성을 보여주기 위해 k값이 1 또는 3인 KNN(K-nearest neighborhood), 나이브 베이즈, K2와 TAN 알고리즘을 이용한 베이지안 망 (BN, Bayesian network), RBF 커널을 사용한 SVM (support vector machine) 알고리즘들과 성능을 비교 평가하였다. 이 때 사용한 평가 기준들은 TP(true positive), FP(false positive), TN(true negative), FN(false negative) 데이터의 개수에 기반하여 계산되며 다음과 같은 식으로 표현된다.

$$Accuracy = \frac{(TP+TN)\times 100}{TP+FP+FN+TN}$$

$$Sensitivity = \frac{TP\times 100}{TP+FN}$$

$$Specificity = \frac{TN\times 100}{FP+TN}$$

$$MCC = \frac{(TP\times TN - FN\times FP)\times 100}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}$$

정확도(accuracy)와 MCC는 각 모듈에 대한, 예측 민감도, 특이도는 각 모듈에서의 각 목표 항목에 대한 예측 성능을 평가하는 기준으로 사용하였으며 10-fold cross validation을 이용하여 성능 평가를 수행하였다.

표 1. miRNA 목표유전자 예측 성능 평가

	Accuracy	MCC	Sensitivity	Specificity
MLPNN	84.62	67.62	80.92	86.94
KNN (k=1)	80.10	60.03	83.55	77.96
KNN (k=3)	81.86	64.37	88.16	77.96
NB	82.12	65.89	91.45	76.33
BN (k2)	83.12	64.76	80.92	84.49
BN (TAN)	85.39	71.05	90.79	82.04
SVM	82.12	66.38	92.76	75.51

먼저 표 1은 miRNA의 목표 유전자를 예측한 모듈의 성능을 보여주고 있다. MLPNN은 TAN 알고리즘을 이용한 BN보다 성능이 약간 떨어지기는 하지만 84.62%의 예측 정확도와 67.62%의 MCC를 기록하여 KNN, NB, SVM 알고리즘을 사용하여 구축한 모듈보다 높은 성능을 보여주었다. 다른 대부분의 알고리즘들은 MLPNN에 비해 예측 민감도는 높게 나왔다. 그러나 특이도가 조금 감소할 경우 많은 잘못된 예측 결과를 초래하므로 대용량의 유전체 데이터를 분석할 때는 보통 민감도보다 특이도를 중요시한다. 따라서 다른 알고리즘에 비해 민감도는 다소 낮지만 특이도가 높으면서 좋은 정확도와 MCC를 기록한 MLPNN이 목표 유전자를 예측하는 작업에서 좋은 성능을 보인다 할 수 있다.

표 2. miRNA 조절 메커니즘 분별 성능 평가

	Accuracy	MCC	Sensitivity (PTD)	Sensitivity (TR)
MLPNN	87.90	75.13	91.25	83.33
KNN (k=1)	84.03	67.14	90.38	75.40
KNN (k=3)	75.80	50.10	88.92	57.94
NB	70.25	38.46	77.55	60.32
BN (k2)	85.04	69.32	92.13	75.40
BN (TAN)	79.83	58.46	84.55	73.41
SVM	71.26	41.02	91.25	44.05

표 2는 특정 miRNA의 목표 유전자로 예측된 결과들 중 miTarget에 조절 메커니즘이 명세되어 있는 데이터를 뽑아 miRNA 조절 메커니즘을 분별하는 모델을 학습시킨 결과 성능을 보여준다. MLPNN을 사용한 제안 모델은 87.90%의 분별 정확도와 71.13%의 MCC를 기록하며 다른 알고리즘들에 비해 월등히 좋은 성능을 나타내었다. PTD에 대한 특이도가 TR에 대한 민감도와

같으므로 표 2에서는 각 클래스에서의 민감도만 표시하였다. 모든 알고리즘에서 전반적으로 TR에 비해 PTD 메커니즘에 대한 예측 민감도가 높게 나온 이유는 PTD 데이터의 양이 TR 데이터에 비해 상대적으로 많아 서열상에 존재하는 생물학적 특징을 보다 잘 학습하였기 때문인 것으로 예상된다. 학습 데이터셋이 보다 보강된다면 본 제안 시스템의 성능은 더욱 향상될 것이다.

성능평가 결과를 통해 본 연구에서 제안한 서열 데이터로부터 계산된 생물학적 특징과 학습 알고리즘을 이용하여 miRNA의 목표 유전자를 예측하고 miRNA의 조절 메커니즘을 분별하는 작업이 잘 동작함을 알 수 있으며, 특히 MLPNN 알고리즘이 miRNA 목표 유전자 예측 작업과 miRNA 조절 메커니즘 분별 작업에 있어 고르고 우수한 성능을 가짐을 보여주었다.

4. 결론

본 논문에서는 생물학적 특징들과 다층 퍼셉트론 신경망을 이용하여 miRNA의 목표 유전자를 예측하고 동시에 해당 miRNA 조절 메커니즘 타입을 분별해주는 시스템을 제안하고 실제 데이터를 사용하여 그 성능을 평가하였다. 실험적으로 검증된 데이터를 사용하여 제안 시스템을 평가해본 결과, 본 연구에서 사용한 생물학적 특징들과 기계학습 알고리즘을 이용한 miRNA 목표유전자 예측하는 작업과 miRNA의 조절 메커니즘을 분별하는 작업이 좋은 성능으로 수행됨을 보여주었다. 학습 데이터가 충분히 많아진다면 매우 좋은 성능으로 제안 시스템이 실용화 될 수 있을 것으로 예상된다.

miRNA의 목표 유전자를 예측함과 동시에 조절 메커니즘까지 분별해주는 제안 시스템은 miRNA 유전체학 연구와 miRNA에 의해 조절되는 유전자 네트워크 규명을 위한 연구 등에 매우 유용하게 사용될 수 있을 것이다.

참고 문헌

[1] Bartel, D. P., "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function", Cell, 116(2):281-297, 2004

[2] Nam, J. -W., K. -R. Shin, J. Han, Y. Lee, V. N. Kim, and B. -T. Zhang, "Human MicroRNA Prediction through a Probabilistic Co-learning Model of Sequence and Structure", Nucleic Acids Research, 33(11):3570-3581, 2005

[3] Lim, L. P., N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, "Microarray Analysis Shows that Some MicroRNAs Downregulate Large Numbers of Target mRNAs", Nature. 433(7027):769-73, 2005

[4] Rajewsky, N., "MicroRNA Target Predictions in Animals", Nature Genetics, 38:S3-S13, 2006

[5] Sethupathy, P., M. Megraw, and A. G. Hatzigeorgiou, "A Guide through Present Computational Approaches for the Identification of Mammalian MicroRNA Targets", Nature Methods, 3(11):881-886, 2006

[6] Lewis, B. P., C. B. Burge, D. P. Bartel, "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets", Cell, 120(1):15-20, 2005

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

[8] Kim, S. -K, J. -W Nam, J. -K Rhee, W. -J. Lee, and B. -T, Zhang, "miTarget: microRNA Target Gene Prediction using a Support Vector Machine", BMC Bioinformatics, 7(1):411-422, 2006

[9] Huang, J. C., Q. D. Morris, Brendan J. Frey, "Detecting MicroRNA Targets by Linking Sequence, MicroRNA and Gene Expression Data", LNCS 3909 pp.114-129, 2006

[10] Huang, J. C., Q. D. Morris, Brendan J. Frey, "Bayesian Learning of MicroRNA Targets from Sequence and Expression Data" To appear in Journal of Computational Biology

[11] Sethupathy, P., B. Corda, A. G. Hatzigeorgiou, "TarBase: A Comprehensive Database of Experimentally Supported Animal microRNA Targets", RNA 12(2):192-197, 2006

[12] Griffiths-Jones S., R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA Sequences, Targets and Gene Nomenclature", NAR, 34(Database Issue):D140-D144, 2006

[13] Birney, E., T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, *et al.* "An Overview of Ensembl", Genome Res. 14(5):925-928, 2004