생물학적으로 의미 있는 특질에 기반한 베이지안 네트웍을 이용한 microRNA의 예측

남진우[O,1,2] 박종선[4] 장병탁[1,2,3]
서울대학교 대학원 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터(CBIT)[2]
서울대학교 컴퓨터공학부 바이오지능연구실[3]
서울대학교 농생명학부[4]
{jwnam, btzhang}@bi.snu.ac.kr

microRNA prediction using Bayesian network with biologically relevant feature set

Jin-Wu Nam[O,1,2] Jongsun Park[4] Byoung-Tak Zhang[1,2,3]

Graduate Program in Bioinformatics[1]
Center for Bioinformation Technology (CBIT)[2]
Biointelligence Laboratory, School of Computer Science and Engineering[3]
College of Agriculture and Life Science[4]
Seoul National University, Seoul 151-742, Korea

요 약

MicroRNA (miRNA)는 약 22 nt의 작은 RNA 조각으로 이루어져 있으며 stem-loop 구조의 precursor 형태에서 최종적으로 만들어 진다. miRNA는 mRNA의 3'UTR에 상보적으로 결합하여 유전자의 발현을 억제하거나 mRNA의 분해를 촉진한다. miRNA를 동정하기 위한 실험적인 방법은 조직 특이적인 발현, 적은 발현양 때문에 방법상 한계를 가지고 있다. 이러한 한계는 컴퓨터를 이용한 방법으로 어느 정도 해결될 수 있다. 하지만 miRNA의 서열상의 낮은 보존성은 homology를 기반으로 한 예측을 어렵게 한다. 또한 기계학습 방법인 support vector machine (SVM) 이나 naive bayes가 적용되었지만, 생물학적인 의미를 해석할 수 있는 generative model을 제시해 주지 못했다. 본 연구에서는 우수한 miRNA 예측을 보일 뿐만 아니라 학습된 모델로부터 생물학적인 지식을 얻을 수 있는 Bayesian network을 적용한다. 이를 위해서는 생물학적으로 의미 있는 특질들의 선택이 중요하다. 여기서는 position weighted matrix (PWM)과 Markov chain probability (MCP), Loop 크기, Bulge 수, spectrum, free energy profile 등을 특질로서 선택한 후 Information gain의 특질 선택법을 통해 예측에 기여도가 높은 특질 25개 와 27개를 최종적으로 선택하였다. 이로부터 Bayesian network을 학습한 후 miRNA의 예측 성능을 10 fold cross-validation으로 확인하였다. 그 결과 pre-/mature miRNA 각 각에 대한 예측 accuracy가 99.99% 100.00%를 보여, SVM이나 naive bayes 방법보다 높은 결과를 보였으며, 학습된 Bayesian network으로부터 이전 연구 결과와 일치하는 pre-miRNA 상의 의존관계를 분석할 수 있었다.

서론

microRNA (miRNA)는 약 22 nucleotide (nt)의 작은 noncoding (nc)RNA의 한 종류이다 [1]. miRNA는 조직 특이적, 발생 특이적으로 발현하여 일반 유전자의 발현을 조절하는 전사 후 조절자로 알려지고 있으며, mRNA의 3' untranslated region (UTR) 부분에 상보적으로 결합하여 translation을 저해하거나 [1], deadenylation을 촉진하여 mRNA의 안정성을 떨어뜨려 분해를 촉진 시키는 것으로 밝혀졌다 [2-3].

miRNA는 RNA polymerase II에 의해 primary miRNA (pri-miRNA) 으로 전사된 후 RNase III 타입의 효소인 Drosha에 의해 stem-end 쪽이 절단 된 형태의 pre-miRNA으로 바뀐 후 세포질 밖으로 나오게 된다 [4-5]. pre-miRNA는 세포질에 존재하는 또 다른 RNase III 타입의 효소인 Dicer에 의해 loop end 쪽이 절단 된 후, 한 쪽의 functional strand 만 선택되어 최종적으로 mature miRNA 으로 만들어 지게 된다 [5].

let-7a 와 lin-4 가 발견된 후로 인간을 포함한 척추동물, 그리고 초파리 등을 포함한 곤충류, 예쁜 꼬마 선충을 포함한 선형동물에서부터 식물과 바이러스에 이르기 까지 다양한 종에서 수천 개의 miRNA가 보고가 되고 있으며 [6-9], 특히 인간의 유전체 내에서 약 300여개의 miRNA 가 발견 되었고, 그 보다 훨씬 많은 수가 여러 가지 예측 방법에 의해 예측이 된 상태이다 [9]. 그러나 아직까지 많은 수의 miRNA 가 발견되지 않고 있는 상태이며, 예측된 miRNA 중에서도 false positive가 많이 발견되고 있는 실정은 보다 정교하고 정확한 miRNA 예측 알고리즘을 요구 하고 있다.

miRNA의 발굴은 실험을 통한 방법과 생물정보학적인 방법을 통한 예측 등 크게 두 가지 방법을 통해 이루어지고 있다 [10]. 실험을 통합 방법은 대부분은 cloning 을 통한 방법이나 [11] northern-blot을 통한 방법이 [12] 이용되었지만, 최근 microarray를 이용하여 miRNA 발굴에 성공하기도 했다 [13]. 생물정보학을 이용한 방법은 homology를 기반으로 한 탐색 방법과 [14-15] 보존된 motif정보와 heuristic한 방법을 통해 예측

하는 방법 [16-17], 그리고 pre-miRNA의 보존된 서열과 구조정보로부터 데이터를 학습하여 예측하는 기계학습기법들이 사용되어 왔다 [18-20]. 이러한 생물정보학의 방법은 실험을 통해 발굴하기 힘든 조직 특이적이거나 발생특이적인 miRNA나 발현양이 적은 miRNA의 발굴에 크게 이바지 하고 있다 [10].

특히 hidden Markov model을 이용하여 구조와 서열을 동시 학습한 distant homology정보로부터 새로운 miRNA를 예측할 수 있는 기계학습 방법을 제안하였으며, 여러 가지 miRNA 예측을 위해 해결해야 할 문제들을 정의하여 해결하였다 [18]. miRNA 예측을 위해서는 우선 pre-miRNA의 예측이 우선되어야 하며, 그 이후로 functional strand를 결정하는 문제와 mature miRNA를 결정하는 문제가 해결되어야 한다. hidden Markov model이 제시하는 확률들은 이러한 문제를 동시에 해결할 수 있는 아이디어를 제공하였다 [18]. 또한 HMM을 기반의 알고리즘과 conservation 정보와 genome coordination 정보를 이용하여 clustered miRNA를 예측할 수 있는 시스템이 발표되었다 [21]. 또한 기계학습법을 이용한 예측의 성공으로 여러 가지 기계학습방법이 제안되기에 이르렀다. Support vector machine (SVM)을 이용한 방법들과 [20] naive bayes를 이용한 방법은 [19] miRNA와 음성데이터간의 특징들로부터 분류하기 위한 파라미터들을 학습하여, 그 학습된 모델들로부터 miRNA들을 성공적으로 예측하였다. 하지만, SVM은 좋은 예측성능을 보이긴 하지만, miRNA의 구조적, 서열적 특징과 관계를 분석할 수 있는 generative model을 제시해주지 못하는 한계를 갖고 있다. 또한 naive bayes는 각 특질간의 독립성을 전제로 하기 때문에 특질간의 관계가 있다고 예상되는 데이터에서는 적합하지 않으며, 적절한 generative model 을 제시해 주지 못한다. 또한, 이전의 방법들은 구조나 서열 등의 단순하거나 몇 가지의 특질들만을 이용하여 데이터를 학습하여, 좀 더 생물학적으로 의미 있는 특질들의 생성이 필요한 시점이다.

본 논문에서는 각 특질간의 causality를 분석할 수 있는 대표적인 generative model인 bayesian network을 이용하여 주어진 데이터로부터 모델의 구조와 파라미터를 학습한 후 생물학적으로 의미 있는 특질들간의 인과관계를 분석해 보고, 최종적으로 miRNA의 예측모델을 제시한다. 실제 실험을 위하여 인간의 miRNA 서열과, miRNA가 아니지만 miRNA와 유사한 음성데이터를 이용하여 모델을 학습하고 miRNA 예측에 대한 성능을 평가한다. 특히 본 연구에서는 pre-miRNA 와 mature miRNA의 예측을 두 단계로 진행하게 된다.

방법 및 데이터

데이터

실험을 위한 데이터는 pre-miRNA와 mature miRNA의 데이터로 나누어 사용한다. 우선 pre-miRNA는 miRBase 8.1 버전 중에서 321개의 인간 pre-miRNA를 사용하며 (http://microrna.sanger.ac.uk) 음성데이터는 인간의 genome 데이터에서 pre-miRNA와 비슷한 구조를 갖는 서열을 추출 하였고, 총 1459개의 음성데이터를 사용하게 된다. 한편 mature miRNA 데이터는 총 455개가 존재 하였으며, mature miRNA 의 음성데이터는 pre-miRNA에서 mature miRNA 가 아닌 임의의 위치에서 추출한 서열로 정의하였다. 음성데이터의 추출 방법에 대해서는 아래에 별도로 설명한다.

음성데이터의 추출

우선 pre-miRNA의 음성데이터를 추출하기 위해 몇 가지 criteria를 정하였다. (1) genome상의 서열 중에서 3' UTR 지역에는 miRNA가 거의 발견되지 않으므로, 3'UTR에서 추출되어야 할 것; (2) 서열의 길이를 90 nt으로 고정한다; (3) stem의 길이가 22 nt 이상이어야 한다; (4) stem-loop 내의 최대 bulge 사이즈가 15 nt 이하여야 한다; (5) free energy가 -25kcal 이하여야 한다; (6) loop size 가 3 nt 이상 20 nt 이하여야 한다. 위 6가지 조건에 맞는 서열을 찾기 위해 window를 옮겨 가면서 90 nt 서열을 mfold 프로그램으로 [22] 구조를 예측하여 criteria에 맞는 서열만을 추출한다.

다음 mature miRNA의 음성 데이터를 추출하기 위해 길이 110 nt의 extended pre-miRNA를 genome 서열에서 추출한 후 mature miRNA의 위치를 표시하고, 그 위치 외에 다른 위치에 있는 22 nt의 모든 서열을 추출하여 음성데이터로 사용하게 된다.

Biologically relevant feature set

주어진 데이터로부터 생물학적으로 의미 있는 특질을 추출하기 위해 몇 가지 사전 작업을 실시하였다. 우선 pre-miRNA의 processing mechanism과 관련한 보존 정보가 pre-miRNA의 서열과 구조상에 있을 것으로 판단되어 서열과 구조에 대한 profile을 작성하였다. profile은 세 가지 방법으로 이루어졌다. 첫 번째 position weighted matrix (PWM)를 pairwise 서열로부터 구축하였다. pairwise 서열은 구조상의 pair/mispair 정보뿐 아니라 1차 서열의 정보에 따라 모두 다른 24개의 state로 만든 후 해당 state의 score를 각 position정보로 제공하게 된다. position의 기준 위치는 mature miRNA의 5' end로 결정하였다. 그리하여 총 22개의 position에 대한 특질을 추출하였다. 두 번째 Markov property를 이용하여 이전 position에 대한 현 position의 확률을 테이블로 작성한 다음 확률 값의 곱의 로그 값을 취한 값을 특질로 사용한다. 세 번째로 각 포지션별 free energy profile을 작성하여 각 포지션별 값을 특질로 사용한다. mature miRNA의 5' end 위치를 기준으로 총 22개의 포지션의 free energy 값을 특질로 추출하였다. 또한 pre-miRNA의 서열상의 preference를 조사하기 위해 spectrum 정보를 특질로 추가 사용하였다. 2-mer와 3-mer의 염기의 조합에 따른 빈도를 측정하여 각각 16개 와 64개, 총 80개의 특질을 추출하고, 각 서열의 shannon entropy값과 GC ratio, 그리고 전체 free energy 값을 추가 특질로 사용한다. 그리하여 pre-miRNA를 위한 130개의 특질을 추출하게 되었다 (표 1).

또한, mature miRNA 대한 bayesian network을 학습하기 위해 전체 free energy값을 제외한 나머지 특질들은 pre-miRNA와 동일한 방법으로 추출하게 되며, free energy profile과 PWM은 mature miRNA 5' end를 기준으로 15개 위치에 대한 특질을 사용하여 총 115개의 특질을 사용하게 된다 (표 1).

No	Name	Description
1	PWM*	Position weighted matrix
2	MCP*	Markov chain prediction
3	Spectrum(2bp)	The distribution and frequency of two bases in each miRNA
4	Spectrum(3bp)	The distribution and frequency of three bases in each miRNA
5	Loop size	The number of bases in loop structure predicted by mFold
6	Bulge count*	The number of bulges predicted by mFold in miRNA
7	Energy profile*	Sequence of energy values in each base of miRNA
8	Shannon's entropy	$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i)$ where p={A,G,C,T}
9	Minimum free energy*	The minimum free energy value predicted by mFold
10	GC Ratio	The ratio of GC/AT

표 1. Feature set

Bayesian network 학습

Bayesian network은 의존 구조와 지역 확률 모델의 두 가지 요소로 구성되는 확률 모델이다 [23-24]. 의존 구조는 각 확률변수들이 어떻게 서로 연관되어 있는지를 순환 구조 없이 방향성 있는 연결선을 통해 표현된다. 각 확률변수는 부모로 생각되는 다른 변수들 중 가능한 하나의 비공집합

set에 의존하는 식 (1) 으로 표현되며

$$p(x_1,\ldots,x_n)=\prod_{i=1}^{n}p(x_i|Pa(x_i)) \quad (1)$$

여기서 $Pa(x_i)$는 x_i의 부모에 해당한다. 한편 지역 확률 모델은 각 변수들이 부모에 어떻게 의존하는 지를 조건부 확률 테이블로 표시하게 된다. 그리하여 각 조건부 확률 테이블은 부모 확률변수의 값이 주어졌을 때 한 변수가 한 특정한 값을 갖게 될 확률로 표시 된다.

Bayesian network의 학습은 모델의 구성요소에 따라 두 단계로 나누어진다. 첫 번째로 주어진 데이터로부터 network의 구조를 학습하게 된다. 최적의 구조를 찾는 작업은 확률 변수의 개수에 따라 기하급수적인 계산이 필요로 하는 어려운 작업이다. 본 연구에서는 greedy search 알고리즘의 하나의 K2 알고리즘을 사용 한다 [25]. 이때 최적화하기 위한 관측 값으로 Bayesian Dirichlet 값을 사용하게 된다:

$$p(S|D)\propto p(S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})}\right], \quad (2)$$

여기서 N_{ijk}는 현재 구조 S에서 각 부모의 j번째 인스턴스와 관계된 k state에서 변수 i를 갖는 데이터의 경우의 수로 표시된다. n은 변수의 총 수를 의미한다. N_{ij}는 한 변수의 모든 state의 합으로 계산 되며, N'_{ij}는 파라미터들의 prior 정보를 갖는 값을 의미한다. r_i 와 q_i 는 변수 i의 state의 개수와 인스턴스의 개수를 의미한다. 마지막으로 $p(S)$는 그 구조의 prior 확률을 의미한다. 식 2를 이용하여 K2 알고리즘이 구조를 찾게 되며, 각 변수는 랜덤 순서로 탐색하게 된다. 초기 구조는 naive bayes에 의해 학습된 구조로 시작하게 된다. K2 알고리즘은 반복적으로 각 변수에 대한 최적의 부모 변수를 찾게 되며 한 부모의 변수를 추가 했을 때 식 2의 값이 증가하지 않으면 멈추게 된다.

두 번째 학습은 결정된 network 구조의 지역 확률 값을 추정하는 단계이다. 이때 각 지역 확률 값을 추정하기 위해 조건부 확률 테이블을 이용하게 된다. 각 부모들의 인스턴스와 변수에 대해 하나의 파라미터 셋으로 구성된 조건부 확률 테이블이 주어지며, 각 파라미터 셋은 uniform한 Dirichlet prior로 주어진다:

$$p(\theta_{ij}|S)=Dir(\theta_{ij}|N'_{ij1},\ldots,N'_{ijk},\ldots,N'_{ijr_i}) \quad (3)$$

여기서 θ_{ij} 현재 구조에서 부모의 j 번째 인스턴스에 i 번째 변수를 갖는 현재 구조의 파라미터 셋을 의미한다. 그리하여 θ_{ij}는 각 부모의 현재 인스턴스가 주어 졌을 때 변수 x_i의 모든 값들에 대한 확률을 담고 있게 된다. Dir는 Dirichlet 분포에 해당하게 된다. 식 3은 데이터 D가 주어졌을 때 그 파라미터 셋 상에서 식 4처럼 Dirichlet posterior로 변환 될 수 있다:

$$p(\theta_{ij}|D,S)=Dir(\theta_{ij}|N'_{ij1}+N_{ij1},\ldots,N'_{ijk}+N_{ijk},\ldots,N'_{ijr_i}+N_{ijr_i}) \quad (4)$$

여기서 Dirichlet 분포의 maximum a posteriori (MAP) 방법을 통해 posterior를 요약할 수 있으며, 이 값은 각 조건부 확률 테이블을 채우기 위해 사용된다.

miRNA 예측 및 검증 방법

주어진 pre-miRNA, mature miRNA 데이터와 음성데이터 셋을 이용하여 Bayesian network을 학습 하는 과정에서 많은 수의 특질은 network 구조 탐색공간을 기하급수적으로 늘리기 때문에 pre-miRNA와 mature miRNA의 특질 130개와 115개 중에서 예측 성능에 크게 기여하는 특질을 먼저 선택해야 한다. 본 연구에서는 information gain 방법을 이용하여 특질을 추출 하였다. 이를 기반으로 학습된 bayesian network의 예측 성능을 측정하기 위해서 10-fold cross validation 을 수행한다. 성능 평가 실험과, 다른 알고리즘과의 비교 실험은 모두 Weka 3.4버전을 이용해 수행되었다.

결과

miRNA 예측 결과

우선 pre-miRNA의 특질을 선택하기 위해서 information gain 방법으로 10 fold cross-validation을 수행하여 평균값으로부터 등위를 매겼고, 이로부터 25 이내의 특질을 선택 하였다 (표 2). 각 편차는 10 fold cross-validation에서의 값의 표준편차를 표시한 것이다. 결과에서 PWM의 특질은 모두 선택 되었으며, 특히 15-20번째 위치의 PWM과 3번째 6번째 위치의 PWM이 예측 성능에 크게 미쳤고, Markov score 도 중요하게 작용하였다. 또 GC ratio와 free energy도 어느 정도 예측 성능에 기여를 하고 있었다. 반면 free energy profile과 spectrum, bulge size, loop size의 경우 크기 기여하는 특질로 선택되지 못했다.

Information Gain (편차)	Rank (편차)	특질 (위치)
0.667 +- 0.002	1.7 +- 0.9	PWM(3)
0.667 +- 0.002	2.1 +- 0.83	PWM(15)
0.667 +- 0.002	3.1 +- 1.64	PWM(6)
0.667 +- 0.002	3.8 +- 0.87	PWM(17)
0.667 +- 0.002	5.4 +- 1.2	PWM(20)
0.665 +- 0.006	5.7 +- 1.9	PWM(16)
0.667 +- 0.002	6.6 +- 0.8	PWM(11)
0.662 +- 0.002	7.9 +- 0.3	PWM(18)
0.66 +- 0.002	8.9 +- 0.3	PWM(2)
0.656 +- 0.003	9.9 +- 0.3	PWM(22)
0.643 +- 0.003	11.9 +- 0.7	PWM(12)
0.642 +- 0.003	12.1 +- 0.7	PWM(4)
0.638 +- 0.011	12.5 +- 1.8	PWM(19)
0.629 +- 0.003	14.4 +- 0.49	PWM(10)
0.626 +- 0.005	14.8 +- 1.54	MCP(1)
0.622 +- 0.003	16.4 +- 0.8	PWM(13)
0.619 +- 0.005	16.6 +- 0.92	PWM(9)
0.612 +- 0.013	17.4 +- 1.74	PWM(7)
0.601 +- 0.004	18.9 +- 0.54	PWM(21)
0.595 +- 0.002	20.5 +- 0.5	PWM(1)
0.586 +- 0.013	20.5 +- 0.81	PWM(5)
0.568 +- 0.008	21.9 +- 0.3	PWM(14)
0.53 +- 0.004	23 +- 0	PWM(8)
0.431 +- 0.05	24 +- 0	GC(1)
0.2 +- 0.006	25 +- 0	FREE(1)

표2. Information Gain 방법으로 선택된 Pre-miRNA의 특질

한편 mature miRNA의 특질들도 위와 동일 한 방법을 통해 선택하였으며 총 27개의 특질을 선택하였다 (표 2). 여기서는 PWM과 Markov score 그리고 spectrum 특질들도 선택되었으며, 이는 mature miRNA의 특질은 pre-miRNA 와 달리 1차 서열의 특질도 중요함을 의미한다. 특히 UG의 spectrum이 예측 성능에 많은 기여를 하고 있는 것으로 보이며. 이것은 mature 서열이 U로 많이 시작되는 사실과 일치함을 알 수 있다. PWM 특질에서는 10, 11번째 위치의 특질이 중요하게 기여를 하고 있는데 이것은 mature의 위치를 고려했을 때 bulge 구조가 많이 나타나는 위치임을 알

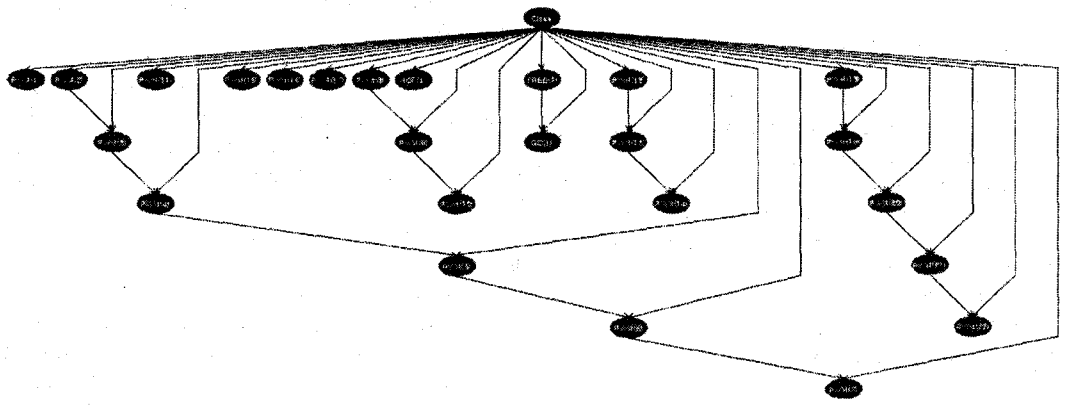

그림 1. pre-miRNA의 Bayesian network 모델

수 있다.

Information Gain score	Rank (deviation)	Feature
0.314 +- 0.001	2.3 +- 1.19	PWM(11)
0.314 +- 0.001	2.5 +- 1.02	PWM(10)
0.314 +- 0.001	2.5 +- 1.12	MCP(1)
0.313 +- 0.002	2.7 +- 1.1	PWM(4)
0.304 +- 0.002	6 +- 0.77	PWM(12)
0.303 +- 0.003	6.1 +- 1.76	PWM(0)
0.301 +- 0.001	7.7 +- 0.78	SP2(15)
0.299 +- 0.003	8.8 +- 2.36	PWM(9)
0.299 +- 0.001	8.9 +- 0.94	PWM(5)
0.299 +- 0.002	9.1 +- 1.37	PWM(8)
0.297 +- 0.004	9.8 +- 2.09	PWM(7)
0.291 +- 0.001	13.9 +- 1.14	PWM(13)
0.292 +- 0.004	13.9 +- 2.34	PWM(6)
0.29 +- 0.006	14 +- 2.45	PWM(-1)
0.29 +- 0.002	15 +- 1.26	PWM(2)
0.289 +- 0.003	15.4 +- 1.74	PWM(14)
0.287 +- 0.006	15.5 +- 2.46	PWM(3)
0.282 +- 0.003	17.6 +- 1.11	SP2(5)
0.277 +- 0.009	18.5 +- 1.2	PWM(1)
0.271 +- 0.003	19.9 +- 0.3	SP2(8)
0.246 +- 0.003	21.3 +- 0.46	PWM(15)
0.244 +- 0.005	21.6 +- 0.66	PWM(-2)
0.235 +- 0.002	23.3 +- 0.64	SP2(12)
0.233 +- 0.003	24.1 +- 0.54	SP2(9)
0.23 +- 0.002	24.6 +- 0.66	SP2(14)
0.204 +- 0.004	26.3 +- 0.64	SP2(11)
0.2 +- 0.003	27.2 +- 0.6	SP2(3)

표 3 Information Gain 방법으로 선택된 mature miRNA의 특질

위 특질들을 이용하여 Bayesian network의 구조를 학습한 후 조건부 확률 테이블로부터 파라미터를 추정하였다. Pre-miRNA와 mature miRNA에 대해서 각각 모델을 학습한 후 예측 성능을 측정하였다.

표 4는 각각에 대한 예측과 다른 방법과의 성능 비교를 보여주고 있다. 우선 pre-miRNA 예측에서 naive bayes는 특질의 선택 후 성능이 더 좋아 졌으며 반대로 SVM은 특질의 선택 후 성능이 더 나빠진 경우 이다. 결과를 비교해 보면, 특질을 선택한 후 Bayesian network의 성능이 가장 좋음을 알 수 있다.

	Accuracy	Sensitivity	Specificity
Naive Bayes	0.9836	0.9840	0.9836
Naive Bayes*	0.9904	0.9904	0.9904
Adaboost	0.9955	0.9744	1.0000
SVM	0.9983	0.9936	0.9993
SVM*	0.9955	0.9968	0.9952
Bayesian net*	0.9994	0.9968	1.0000

표 4. pre-miRNA 예측 성능 평가와 다른 알고리즘과의 비교

	Accuracy	Sensitivity	Specificity
Naive Bayes*	1	1	1
SVM*	1	1	1
Bayesian net*	1	1	1

표 5. mature miRNA 예측 성능 평가와 다른 알고리즘과의 비교

한편 mature miRNA의 예측은 좋은 특질들이 덕택에 예측 결과가 모두 완벽했으며, 결과에는 없지만 이전의 예측 프로그램을 똑같은 데이터로 실험했을 때는 이보다 좋지 못한 결과가 나온 것을 감안 하면, 생물학적으로 중요한 특질들이 만들어진 것으로 보여 진다.

Bayesian network으로부터 miRNA processing mechanism 분석

학습된 Bayesian network 모델은 generative model이기 때문에 생물학적인 의미를 해석 할 수 있다. 여기서는 miRNA의 processing mechanism과 관련해 중요한 위치 정보와, 그 위치간의 관계를 살펴 볼 수 있을 것이다. 이전의 연구에서 우리는 pri-miRNA processing과 관련한 분자적 요소를 free energy profile과 실험을 통해 보여주었다 [26]. 그 연구에서 miRNA가 정상적으로 processing 되기 위해서는 mature miRNA의 5' end를 기준으로 +1 번째의 위치가 열역학적으로 중요하며, stem 말단에 결합되지 않은 단일 가닥이 일정 길이만큼 존재해야 하고, mature의 중간 위치에는 free energy가 높은 구조가 나타나야 한다는 것을 보여주었다. 그림 1에서 pre-miRNA의 Bayesian network은 앞에서 언급한 연구 결과와 일치하고 있음을 보여주고 있다. 우선 free energy와 GC ratio간의 의존관계가 있음은 직관적으로 알 수 있다. PWM에서 12~13 번째 위치가 상관관계를 맺고 있으며, 18~22 번째 위치가 또한

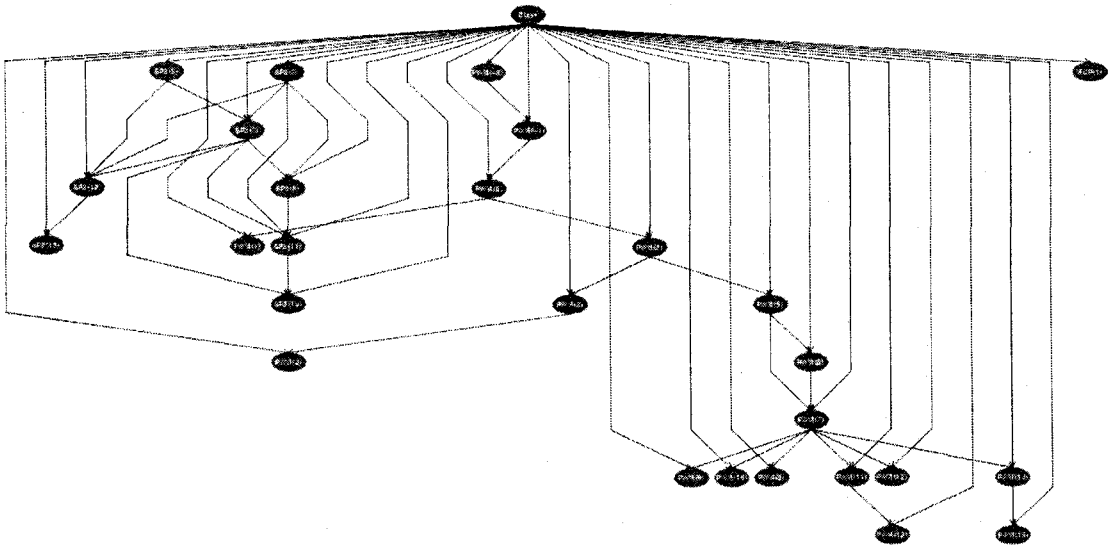

그림 2. mature miRNA의 Bayesian network 모델

의존관계를 맺고 있다. 또 8~10번째 위치와 2~4번째 위치가 서로 의존관계를 있음 보여 주고 있다. 3번째 위치가 5' end 임을 고려할 때 2~4번째 위치가 의존성을 보이는 것은, 그 위치에서 높은 free energy를 갖는 preference를 갖기 때문이며, 마찬가지로 12~13번째 위치에서 bulge 구조가 많이 나타나는 것은 그 위치들 사이에 의존성이 있음을 알 수 있다.

한편 mature miRNA의 Bayesian network도 위와 유사한 결과를 예상해 볼 수 있었다. 그림 2에서 -2~1번째 위치의 의존성은 거의 동일하게 나타났으며, 특이하게도 0번째 위치가 2번째 위치와 다시 2번째 위치는 5번째 위치와 그리고 다시 5번째 위치는 7번째와 8번째 위치와 의존성을 갖고 있었으며 7번째 위치는 9~14번째 위치와 각각 의존성을 맺고 있었다. 이러한 새로운 사실은 약 2~3 nt 떨어진 염기 pair가 서로 processing 기능에 연관되어 있음을 설명해주고 있다.

결론

miRNA 예측에 있어서 중요한 점은 물론 좋은 예측 성능이다. 하지만, 그것보다 더 중요한 것은 예측을 위해 만들어진 모델로부터 새로운 생물학적 지식을 이끌어 낼 수 있나는 것이다. 즉 잘 만들어진 모델은 예측 성능이 물론 좋을 것이며, 그 모델은 실제 생물학적은 현상을 잘 설명하고 있을 것이기 때문이다. 본 연구에서 사용한 Bayesian network을 비롯한 기계학습방법의 장점은 관측된 데이터로부터 실제 생물학적 현상에 맞는 모델을 자동으로 생성하고 탐색하고 최적화 할 수 있다는 것이다. 이것은 새로운 생물학적 사실을 도출 하는데 아주 유용한 점이라 할 수 있을 것이다. 본 연구에서는 실제 생물학적으로 의미 있는 특질들을 엄중히 발굴하여, miRNA 예측에 있어서 좋은 성능을 보이며, 또한 새로운 생물학적인 사실을 도출할 수 있는 generative model의 하나인 Bayesian network을 적용함으로써 도출된 특질들 간의 의존관계를 분석할 수 있었다.

사실 최근 miRNA 예측 알고리즘들은 유전체 내에서 새로운 miRNA 유전자를 계속해서 발굴해 내고 있고, 실험을 통한 검증을 통해 그 예측 시스템을 발전 시켜 나가고 있다. 예측 시스템의 개발에서 있어서 무엇보다 중요한 것은 그 시스템의 실험을 통한 검증일 것이다. 본 논문에서 제안한 Bayesian network 시스템은 pre-miRNA 예측과 mature miRNA 예측에서 동시에 좋은 성능을 보이고 있다. 하지만, 실험을 통한 검증이 없다면, 그 성능을 장담 할 수 없을 것이다. 다음 연구에서는 Bayesian network에서 발견된 새로운 사실들을 바탕으로 인공적인 miRNA의 설계와 실험적 검증을 하고자 한다. miRNA의 예측 시스템은 인공적인 miRNA 설계와 불가분의 관계에 있다. 즉 잘 만들어진 예측 시스템으로 설명될 수 있는 생물학적인 지식은 인공적인 miRNA의 설계에 중요한 정보로 제공 될 수 있기 때문이다.

감사의 글

이 논문은 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정연구실 사업(NRL)에 의하여 지원되었음.

참고논문

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

[2]. Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J. and Schier, A.F. (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**, 75-79.

[3]. Wu, L., Fan, J. and Belasco, J.G. (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A*, **103**, 4034-4039.

[4]. Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. and Kim, V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *Embo J.* **23**, 4051-4060.

[5]. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. et al.(2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415-419.

[6]. 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.

[7]. Ambros, V. and Moss, E.G. (1994) Heterochronic genes and the temporal control of C. elegans development. *Trends Genet*, **10**, 123-127.

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

853-858.

[9]. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140-144.

[10]. Kim, V.N. and Nam, J.W. (2006) Genomics of microRNA. *Trends Genet*, **22**, 165-173.

[11]. Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz, L.A., Jr., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E. et al. (2006) The colorectal microRNAome. *Proc Natl Acad Sci U S A*, **103**, 3687-3692.

[12]. Suh, M.R., Lee, Y., Kim, J.Y., Kim, S.K., Moon, S.H., Lee, J.Y., Cha, K.Y., Chung, H.M., Yoon, H.S., Moon, S.Y. et al. (2004) Human embryonic stem cells express a unique set of microRNAs. *Dev Biol*, **270**, 488-498.

[13]. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E.et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, **37**, 766-770.

[14]. Legendre, M., Lambert, A. and Gautheret, D. (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, *21*, 841-845.

[15]. Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X. and Li, Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, *21*, 3610-3614.

[16]. Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H. and Cuppen, E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21-24.

[17]. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S.and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, *434*, 338-345.

[18]. Nam, J.W., Shin, K.R., Han, J., Lee, Y., Kim, V.N. and Zhang, B.T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, **33**,3570-3581.

[19]. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C. and Showe, M.K. (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier machine learning for identification of microRNA genes. *Bioinformatics*, **11**, 1325-34.

[20]. Xue, C., Li, F., He, T., Liu, G.P., Li, Y. and Zhang, X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*. **6**, 310.

[21]. Nam, J.W., Kim, J., Kim, S.K. and Zhang, B.T. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res*, **34**, W455-458.

[22]. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406-3415.

[23]. Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman Publishers, San matteo, California.

[24]. Neapolitan, R. (2004) Learning Bayesian networks. Prentice Hall, Upper Saddle River, NJ.

[25]. Cooper, G. and Herskovits, E. (1992) A bayesian method for the induction of probabilistic networks from data. *Machine Learn.*, **9**, 309-347.

[26]. Han, J., Lee, Y.T., Yeom, K.H., Nam, J.-W., Heo, I.H., Rhee, J.-K., Shon, S.Y., Cho, Y.J., Zhang, B.-T. and Kim, V.N. (2006) Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell*, **125**, 887-901.