Mature microRNA 위치 예측 모델의 진화적 최적화

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

Evolutionary Optimization of Models for Mature microRNA Prediction

Jinhan Kim[O,1] Jin-Wu Nam[2,3] Byoung-Tak Zhang[1,2,3]

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

요 약

MicroRNA (miRNA)는 생체내에서 gene regulation에 관여하는 핵심 small RNA 중 하나이다. miRNA는 primary miRNA, precursor miRNA, mature miRNA의 과정으로 processing 된다. miRNA 최종 형태인 mature miRNA의 정확한 위치 예측은 miRNA 예측의 필수적인 부분이다. 본 논문에서는, 진화적 최적화 예측 모델 중 하나인 유전 알고리즘을 이용하여 mature miRNA의 정확한 위치 예측을 수행한다. 제시된 방법은 이미 알려진 mature miRNA 위치를 positive example로 하고 임의로 생성한 위치를 negative example로 하여 서로의 linear scoring function 적합성 함수의 값 차이가 최대한으로 되도록 예측 모델을 진화 시킨다. 유전 알고리즘을 이용한 진화적 최적화 모델로부터 mature miRNA 위치 예측에서 약 1.7nt 오차를 보여 기존의 방법 보다 개선된 성능을 보인다.

1. 서 론

microRNA (miRNA)는 생체내에서 gene regulation에 관여하는 핵심 noncoding (nc) small RNA 중의 하나이다 [1]. miRNA는 mRNA translation을 저해하거나 하지 못하게 하여 post-transcriptional gene regulation에 중요한 역할을 한다[1].

miRNA는 RNA polymerase II에 의해 전사되어 [2] 길이가 긴 transcript인 primary miRNA가 된 후 Drosha에 의해 non-looped 말단이 잘려져 stem-loop를 가지는 약 70nt의 precursor miRNA로 processing되고 세포내로 이동하여 약 22nt 길이의 mature miRNA로 최종 processing 된다 [3]. 이때 mature miRNA의 5'말단은 RNase III 타입 효소인 Drosha에 의해 결정되며, Drosha-DGCR8 complex의 pri-miRNA의 인식 메카니즘이 중요한 요소로 작용 한다 [4]. 최근 이러한 메카니즘에 대한 분자적 기반이 연구되면서 mature miRNA의 5' 말단의 예측에 대한 기대가 커지고 있다 [4].

현재까지 인간에서 330여개의 miRNA가 보고 되었고, 수백 종의 miRNA family가 동물에서 발견되었다. 또한 수십 개종의 miRNA family가 식물에서도 발견되었다 (http://microrna.sanger.ac.uk/sequences). 다른 ncRNA와는 달리 miRNA는 다른 family간에 homology가 거의 존재 하지 않는다 [5]. 그리하여 단순 similarity 탐색 방법으로 miRNA를 찾아내기는 쉽지 않다 [1].

miRNA의 예측의 첫걸음은 primary (pri-) miRNA상에서 mature miRNA의 위치 예측이다. 대개 mature miRNA는 *northern-blotting*을 통해 확인하게 되지만, 수많은 pri-miRNA 후보의 mature miRNA를 검증하는 것은 많은 시간과 비용을 요구 한다 [2]. 이에 computational 한 방법은 이 문제를 해결할 수 있는 좋은 접근 방법이라 할 수 있다. 이렇게 computational 한 방법으로 pre-miRNA를 예측하는 것에 대한 여러 논문이 발표 되어 왔다 [6, 7]. 하지만 이 논문들은 pre-miRNA 예측까지만 제시하고 있고 mature miRNA 위치 예측까지는 제시하지 않고 있다. 이에 mature miRNA 위치 예측의 연구의 필요성이 제기된다. 최근 Nam et al., Yousef et al.,은 각각 probabilistic *colearning model*과 Naive Bayes classifer 기계 학습 알고리즘을 이용하여 mature miRNA 위치를 예측하는 결과를 제시하였다 [8, 9].

본 논문에서는 또 다른 기계 학습 알고리즘인 유전

알고리즘 (GA)을 이용하여 평가 함수의 weight를 최적화함을 통해, mature miRNA 위치 예측이 좋은 성능을 나타냄을 보인다. 결과적으로 mature miRNA 위치 예측을 위한 최적화 된 평가 함수를 얻을 수 있고 각 항의 weight 경향을 봄으로써 각 feature가 mature miRNA 위치 예측에 얼마만큼의 영향도를 가지고 있는지도 제시할 수 있다.

2. Methods and Dataset

2.1 Dataset

학습 데이터중 positive data에 대해서는 실험을 통해 이미 검증된 인간 miRNA 126개를 대상으로 삼았다. negative data는 positive data의 mature miRNA 위치를 제외한 주위 10 base의 위치에서 추출하여 총 1260개를 구성하였다. 공정한 평가를 기대하기 위해 5-fold cross validation을 염두해 두고 각각의 positive data와 negative data를 5개의 임의의 subset으로 나누었다.

2.2 Feature sets

본 연구에서는 mature miRNA 위치를 예측하기 위해 총 네가지의 feature (Entropy, position weight matrix (PWM), Markov chain probability (MCP), *k*-mer spectrum)를 사용한다. Entropy는 식 1번과 같이 염기의 heterogeneous한 정도를 측정하는 shannon's entropy를 사용한다. 예를 들어 특정 염기가 반복 되어 있다면 entropy 값은 낮게 나올 것이다. PWM은 각 위치별 염기의 분포를 담고 있는, 데이터 수(n) 길이(l)인 행렬로, 위치별로 선호되는 염기를 고려한다. MCP는 이전 위치의 특정 염기와 현 위치의 특정 염기 사이의 correlation 정보를 담고 있다. MCP는 (mature 길이 - 1)만큼의 4x4 행렬로 구성된다. 마지막으로 *k*-mer spectrum은 mature miRNA의 *k*-mer의 분포를 표현하고 있다. 여기서는 3-mer의 spectrum만을 고려한다. 특별히 2개의 feature, 즉 GC ratio, mature miRNA의 길이는 그 영향이 미비하거나 없음을 보여 주기 위해 추가 되었다.

$$\text{Shannon's entropy} \quad -\sum_{i \in \{a,t,g,c\}} p_i \log_2 p_i \qquad (1)$$

p_i 는 전체 염기에서 각 염기가 차지하는 비율을 의미한다.

2.3 Genetic algorithms (GA) 과 목적함수

유전 알고리즘은 목적함수의 탐색공간에서 최적의 해를 찾기 위해, 생명체 진화 과정의 모델을 적용하여 병렬 탐색하는 학습 전략을 갖는다. 최근 생물정보학 분야에서 system 매개변수의 최적화, protein structure 예측, 최적 probe 디자인문제등에 많이 적용된 바 있다.

2.3.1 염색체 표현

각 염색체는 linear scoring function의 형태로 정의된 적합도 함수에서의 6개의 weight의 1차 서열로 구성된다. 각 weight는 (0, 1]의 범위로 제한되며 상대적인 크기는 각 feature의 중요도로 해석 될 수 있다.

2.3.2 알고리즘

본 연구에서는 전형적 구조의 단순 유전 알고리즘을 사용 하였으며 알고리즘은 아래와 같다. population size는 500이고 replacement는 10%로 설정하였다. 개체군의 초기화는 각 염색체 서열을 (0, 1] 범위의 무작위로 생성한다. 선택은 상위 10%에서 무작위로 2개를 선택한다. 교차는 선택된 염색체 2개를 일점 교차를 사용해 수행한다. 돌연변이는 30%의 확률로 값의 무작위 변화가 일어나도록 한다. 진화의 종료 조건은 세대수이다. 세대수는 5000으로 잡았다.

```
begin
t = 0  // generation
   initialize P(t)  // population
   evaluate P(t)
   while (not termination-condition) do
   begin
      t = t + 1
      select P(t) from P(t-1)  // selection
      crossover-mutate P(t)  // genetic operators
      evaluate P(t)  // fitness function
   end
end
```

2.3.3 적합도 함수

적합도 함수 (F)는 p개의 positive data와 q개의 negative data에 대해서 다음으로 정의되는 linear scoring function 평균의 차이로 정의된다 (식 2).

$$F = \frac{\sum_{i=1}^{p} S(x_i^+)}{p} - \frac{\sum_{j=1}^{q} S(x_j^-)}{q} \qquad (2)$$

$$S(x^+) = W^+_1 S^+_1 + W^+_2 S^+_2 + \ldots + W^+_n S^+_n \qquad (3)$$

$$S(x^-) = W^-_1 S^-_1 + W^-_2 S^-_2 + \ldots + W^-_n S^-_n \qquad (4)$$

$S(x^+)$와 $S(x^-)$는 positive data와 negative data의 linear scoring function을 의미하며, 가중치가 각각의 feature들에 곱해진 값들의 합으로 계산된다 (식 3,4). 여기서 x는 하나의 데이터를 의미한다. $W_1..W_n$은 가중치, $S_1..S_n$은 각각의 feature를 말하며, 여기서는 GC ratio, Entropy, PWM, MCP, k-mer Spectrum, mature miRNA length 총 6개의 feature들과 6개의 weight를 사용한다.

3. Results

3.1 Training Result

3.1.1 GA 학습 곡선

그림 1은 5-fold cross validation에서 5개의 training example에 대한 평균 학습 곡선을 보여주고 있다. 학습 결과는 평균 0.0462의 값으로 수렴했다 (그림 1). 약 3400 세대 근처에서 최적화가 이루어졌음을 알 수 있다.

3.1.2 Weight 최적화 결과와 Feature effects

각각의 feature가 mature miRNA 위치 예측에 얼만큼의 비율로 영향을 주는지 다음의 weight 값을 표시하는 그래프로 확인할 수 있다 (그림 2). 그림 2는 5-fold cross validation에서의 평균값과 표준편차 값을 이용하여 그려졌다. 예상했던 결과지만, Entropy, PWM, MCP, Spectrum이 major한 영향을 가지고 GC ratio, mature miRNA length는 영향이 별로 없음을 알 수 있다.

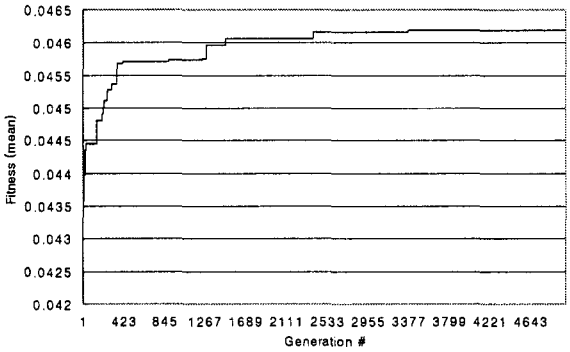

그림 1. 세대별 최적합도의 평균값 그래프

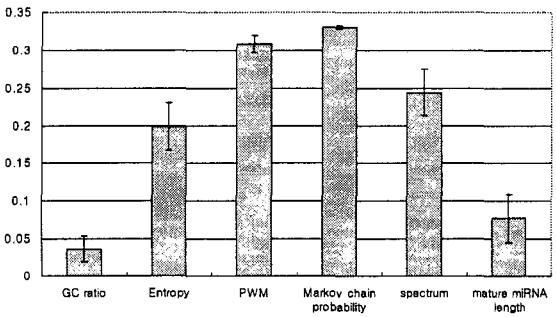

그림 2. Feature weight의 평균과 오차 그래프

3.2 Testing Result

정확도 결과를 구하기 위해 5-fold cross validation으로 test를 수행 하였다. 성능 평가의 방법은 다음과 같다. pre-miRNA를 제시하고 각 base의 처음부터 마지막 위치까지 window sliding을 하면서 각각의 학습 결과로 얻어진 weight를 적용하여 적합도 값을 계산한다. 그 중 적합도 값이 가장 높은 값을 갖는 위치와 이미 알고 있는 올바른 mature miRNA 위치와의 거리를 계산한다. 하나의 test set내에 구성된 pre-miRNA에서 mature 위치 예측에 대한 차이의 평균을 정확도 결과로 한다. 그 결과 5개의 test set에서 평균 1.763619 위치의 거리가 차이나는 성능을 보였다. 이것은 기존의 hidden markov model을 이용한 mature miRNA 예측에서 [8], 평균 1.9 nt의 오차가 있었음을 감안할 때 0.15nt 정도의 성능향상을 보이고 있다.

4. 결론

miRNA 위치 예측은 여러가지 문제를 복합적으로 가지고 있는 어려운 문제 중 하나이다. 위치 예측은 특히 functioal strand를 결정하는 데 큰 요소로 작용하기 때문에 좀더 정확한 예측이 요구되는 것이 사실이다. 하지만 본 연구에서는 약 1.7nt의 에러를 보이고 있다. 이것은 실험데이터의 오차, strand selection의 모호성, Drosha의 fidelity 문제에서 기인할지도 모르는 문제이다. 이후의 연구로 PWM과 MCP를 position별로 각각 score와 weight를 정의하여 좀더 세밀한 mature miRNA 위치 예측을 가능하게 할 것이며, 특히 structure에 기반한 pairwise 서열의 PWM과 MCP의 적용은 functional strand를 고려한 mature miRNA 위치 예측을 가능하게 해 줄 것이다. Drosha substrate의 5' 말단을 예측 하는 문제는, 현재 분자유전학에서 많이 적용되고 있는 siRNA를 shRNA vector상에서 artificial하게 디자인 할 수 있게 해주는 중요한 이슈이다. 정확한 위치 예측의 결과는 이러한 artificial shRNA의 생물학적, 의학적 응용에 크게 도움을 줄 것으로 기대된다.

감사의 글

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

References

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

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

[3] Lee, Y., Jeon, K., Lee, J.T., Kim, S. and Kim, V.N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21, 4663-4670, 2002.

[4] Han J., Lee Y., Yeom K.-H., Nam J.-W. , Hur I., Rhee J.-K., Son S., Cho Y. , Zhang B.-T. and Kim V.N. Molecular basis for the recognition and processing of primary microRNA by Drosha. *Cell* (In press).

[5] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ.miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34, Database Issue, D140-D144, 2006.

[6] Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B. and Bartel, D.P. The microRNAs of Caenorhabditis elegans. *Genes Dev*, 2, 2, 2003.

[7] Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. Computational identification of Drosophila microRNA genes. *Genome Biol*, 4, R42, 2003.

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

[9] Yousef M., Nebozhyn M., Shatkay1 H., Kanterakis S., Showe L.C. and Showe M.K. Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier machine learning for identification of microRNA genes. *Bioinformatics* Advance Access published on March 16, 2006.