하이퍼망 모델을 이용한 MircoRNA Strand 선택 예측

이지훈 [O1] 하정우 [2] 이제근 [1] 장병탁 [1,2]

서울대학교 생물정보학 협동과정 [1]

서울대학교 컴퓨터공학부 [2]

jhlee@bi.snu.ac.kr, jwha@bi.snu.ac.kr, jkrhee@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Prediction of MicroRNA Strand Selection using Hypernetwork Model

Ji-Hoon Lee[O1] Jung-Woo Ha[2] Je-Keun Rhee[1] Byoung-Tak Zhang[1,2]

Graduate Program in Bioinformatics, Seoul National University[1]

Department of Computer Science & Engineering, Seoul National University[2]

요 약

MicroRNA는 RNA로 전사된 유전자와의 상보결합을 통해 유전자 발현을 억제하는 조절인자이다. MicroRNA 생성과정에서 pre-microRNA의 3′ 또는 5′ 부근의 strand가 선택되어 mature 시퀀스가 되고 유전자 조절에 직접 작용하게 된다. 하지만 어떤 특징을 가진 strand가 선택 되는지에 대한 정확한 메커니즘은 아직 연구되어 있지 않다. 본 논문에서는 microRNA 시퀀스 정보를 바탕으로 하이퍼망을 구성하여 strand 선택 예측 모델을 구축하였다. 실험 결과 하이퍼망 학습을 통해 microRNA strand 선택에 중요한 영향을 미치는 시퀀스 특징을 찾을 수 있었고, strand 선택을 높은 정확도로 예측할 수 있음을 확인하였다.

1. 서 론

MicroRNA (miRNA)는 단일 가닥의 RNA 시퀀스로 뉴클레오티드가 19-25 개 정도 연결된 구조로 존재하며, RNA로 전사된 유전자들의 발현을 조절하는 물질이다[1]. 우선 miRNA 유전자가 전사되면 헤어핀 구조의 pri-miRNA 가 만들어 진다. 이후 RNase 인 Drosha에 의해 stem-end 쪽이 절단된 형태의 pre-miRNA로 바뀐 후 세포질 밖으로 나오게 된다. Pre-miRNA는 다시 다른 RNase인 Dicer 에 의해 loop end 쪽이 절단 된 후, 3’ 쪽의 돌출 된 구조인 RNA 이중가닥 구조를 가지게 된다. 이 두 개의 RNA strand 중 한개는 RNA-induced silencing complex (RISC) 단백질과 결합하여 RNA interference (RNAi) 작용을 하게 되고, 나머지 가닥은 사라지거나 상대적으로 적은 수만 남게 된다. 이렇게 miRNA 생성 과정 마지막 단계에서 3’ 혹은 5’ 쪽의 strand가 선택되는 것이 ‘strand 선택’ 이다[2-4].

일반적으로 Dicer에 의해 잘린 RNA 이중가닥의 각 strand는 RISC에 다른 확률로 결합된다[5]. 이러한 특징은 siRNA 생성 과정에서도 유사하게 볼 수 있는 것이다. 여기에서 RNA 이중가닥의 5’ 부분의 열역학적 안정성이 strand 가 RISC의 Argonaute (AGO) 단백질에 결합하는 작용에 영향을 미친다는 것은 연구되어 있다[6-7]. 또한 초파리를 사용한 연구에서 miRNA strand 선택은 AGO 단백질 종류에 따라 다르고 RNA 이중가닥의 특정한 자리의 상보결합 여부에 따라 다르다고 한다[8].

최근 사람과 초파리의 miRNA 시퀀싱에 의한 strand 선택 시퀀스 특징에 관해 연구한 사례가 있다[9]. 이 연구에서 사람과 초파리의 miRNA 뉴클레오티드 각 위치에서 염기의 구성 비율이 상당히 다름을 볼 수 있었다. 이처럼 현재까지의 strand 선택에 관한 연구 결과를 종합 해 볼 때 miRNA strand 선택에는 miRNA 자체의 시퀀스가 중요한 역할을 하고 있음을 알 수 있다.

하지만 기존 연구에서는 miRNA 시퀀스의 각 위치마다의 염기 서열 정보만을 특징으로 보았다는 단점이 있다. miRNA가 RISC 단백질에 인식이 될 때, 연결된 염기 서열의 묶음이 중요하게 작용할 수 있다. 본 연구에서는 다중 인자들간의 상호 연관 관계 분석이 가능한 하이퍼망 학습 모델을 사용하여 strand 선택에 관련된 시퀀스 특징을 찾을 뿐만 아니라 시퀀스

특징들의 상호 연관 관계 까지도 분석하고자 한다. 본 논문의 2장에서는 연구에 적용된 하이퍼망 학습 모델에 관해 설명하고, 3장에서는 컴퓨터 실험 절차와 결과에 대해 분석 하였다. 마지막 4장에서는 차후 연구 진행에 관한 논의와 결론을 제시한다.

2. 하이퍼망 모델

하이퍼망 H는 $H = (V, E, W)$의 형태로 표현되며 V는 하이퍼망을 구성하는 정점 v의 집합, E는 정점들의 묶음인 하이퍼에지 e의 집합, W는 하이퍼에지의 가중치 w의 집합을 의미한다[10]. 하이퍼망에서 정점은 학습데이터의 특징(feature)을 의미하며 하이퍼에지는 특징들간의 조합을 의미한다. 하이퍼에지는 3개 이상의 정점들과 연결이 가능하며, 이로 인해 하이퍼에지는 정점의 집합으로 표현될 수 있다. 하이퍼에지가 연결된 정점의 수를 cardinality 혹은 오더(order)라 표현하며 $n(e)$로 표현한다. 하이퍼망을 표현하면 다음과 같다[10-11].

$$X = \{x_1, x_2, \ldots, x_m\},\ Y = \{y_1, y_2, \ldots, y_p\},$$
$$di = \{x_{1i}, x_{2i}, \ldots, x_{mi}, y_i\},\ D = \{d_1, d_2, \ldots, d_n\},$$
$$V = \{v_1, v_2, \ldots, v_m\},\ E = \{E_1, E_2, \ldots, E_i\}$$

위 식에서 m은 특징의 종류, p는 부류의 종류, di는 한 데이터, D는 전체 데이터 집합이며 임의의 k개 정점으로 이루어진 하이퍼에지 E_i, E_i의 차수 $n(E_i)$, E_i의 가중치 값w_i는

$$Ei = \{ v_{i1}, \ldots, v_{ik}, w_i, y_i\},\ n(E_i) = |E_i - \{ w_i, y_i \}|$$

로 정의된다. 하이퍼망은 확률 그래프 모델로서 수학적으로 표현하면 다음과 같다.

$$P(D|W) = \prod_{n=1}^{N} P\left(x^{(n)}|W\right),$$

$$P\left(x^{(n)}|W\right) = \frac{1}{Z}\exp\left(\sum_{k=1}^{K}\frac{1}{C(k)}\sum_{i_1,i_2,\ldots,i_k} w_{i_1 i_2 \ldots i_k}^{(k)} x_{i_1} x_{i_2} \ldots x_{i_k}\right)$$

위 식에서 W는 가중치의 집합이며 하이퍼망은 가중치가 주어질 때 랜덤변수로 표현되는 데이터의 부분정보들의 확률분포의 형태로 표현된다[10-11].

하이퍼망의 생성과 학습은 다음과 같은 단계로 진행된다.

1) 주어진 데이터를 훈련데이터와 테스트 데이터로 분할한다.

2) 훈련 데이터로부터 샘플링 수(sampling rate)만큼 하이퍼에지를 생성하고 초기 하이퍼망 H를 구성한다.

3) 훈련 데이터의 특징값과 하이퍼에지를 구성하는 정점들 간의 값 비교를 통하여 H를 구성하는 각 하이퍼에지의 가중치를 산출하고 이를 H'로 정의한다.

이때 하이퍼에지 E_j의 가중치는 다음과 같이 계산된다

$$w_j = \frac{\alpha}{N_w} + \beta \times N_c$$

5) H'를 기반으로 훈련데이터의 예측 성능을 측정한다.

4) 가중치 값이 낮은 하이퍼에지를 모델에서 제거하고 제거된 만큼 새로운 하이퍼에지를 다시 생성한다.

6) 2)~5) 까지 단계를 정해진 반복 횟수만큼 반복한다.

7) 성능이 가장 좋은 모델 H''를 선택하고 테스트 데이터에 대하여 분류 성능을 측정한다.

위 식에서 N_w는 하이퍼에지와 매칭된 훈련데이터들에 대하여 하이퍼에지의 클래스와 훈련데이터의 클래스가 다른 경우의 수를 의미하며, N_c는 하이퍼에지와 훈련데이터의 클래스가 같은 경우의 수를 의미한다. 또한 α와 β는 임의의 상수로서 가중치의 정책을 결정하는 요인으로 작용한다. 즉 α가 상대적으로 큰 값을 가지면 하이퍼에지가 덜 틀릴수록 큰 가중치를 갖게 되고, 반대로 β가 크면 많이 맞출수록 가중치가 커진다. 실험적 결과에 따르면 $\alpha >> \beta$ 인 경우 예측 성능이 높은 경우가 많다.

3. 실험 방법 및 결과 분석

3.1 실험 데이터

실험에는 mirBase (www.mirBase.org) 의 mature miRNA 와 stem-loop 시퀀스 정보를 사용하였다. 발현되는 miRNA의 mature 시퀀스가 Positive 데이터로 사용되었고, Negative 데이터는 mature 시퀀스 가닥의 상보가닥에서 추출하였는데, pre-miRNA stem-loop 구조에서 mature 시퀀스가 있는 가닥의 반대편 시퀀스 전부에서 추출하였다. 즉 stem-loop 구조에서 loop 부분의 절반을 기준으로 mature 시퀀스의 반대편 시퀀스 전부를 Negative 정보로 사용하였다(그림 1).

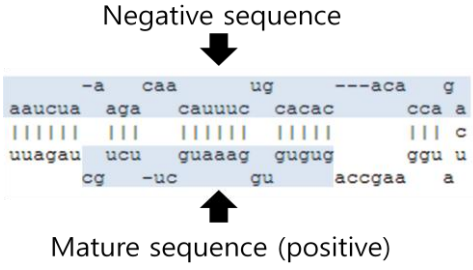

그림 1. miR-147의 stem-loop 구조도. Mature 시퀀스의 반대편 가닥의 시퀀스를 Negative 데이터로 사용하였다.

Positive 데이터는 발현하는 인간 mature miRNA 시퀀스 중 한쪽 가닥 시퀀스만 존재하는(pre-miRNA 양쪽

가닥 중 한 가닥만 발현) 382 개를 사용하였다. Mature 시퀀스가 같지만 loop 반대편 시퀀스의 종류가 다른 경우가 있기 때문에 총 Negative 데이터는 433개가 생성되었다.

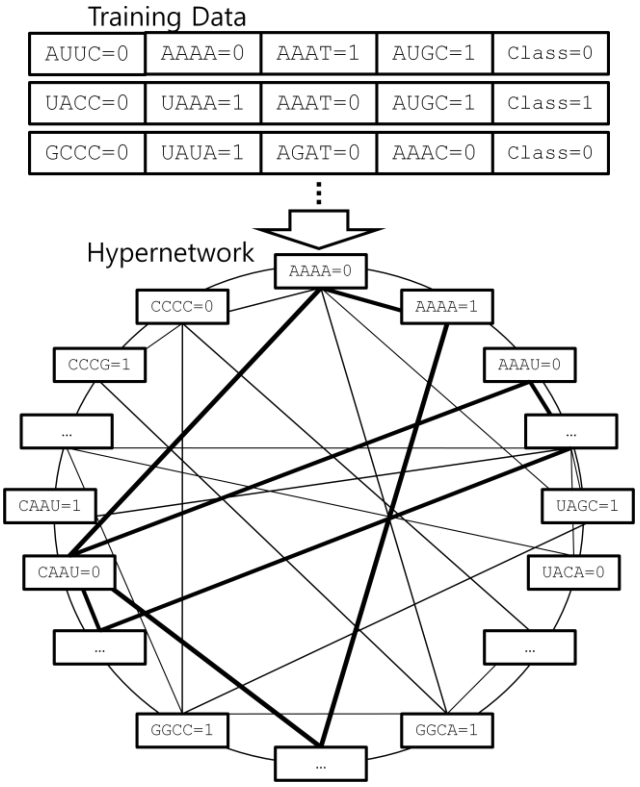

그림 2. 하이퍼망 모델. 오더 4인 하이퍼에지들이 모여 하이퍼망을 구성하며, 각각 하이퍼에지에는 4-mer 길이의 시퀀스 특징이 0과 1의 값을 가진다. 하이퍼에지는 학습을 통해 가중치가 변하며 하이퍼망에서 하이퍼에지의 굵기가 가중치의 크기를 표현한다.

3.2 하이퍼망 학습

하이퍼망의 학습을 위해 각 부류의 miRNA 시퀀스에서 존재하는 시퀀스 특징들은(4-mer, RNA 시퀀스, AAAA, AAAU, AAAG, …, CCCG, CCCC) 1의 값을 가지고, 존재하지 않는 시퀀스 특징들은 0의 값을 가지도록 하였다.(그림 2)

실험에 쓰인 하이퍼망 파라미터값은 다음과 같다(표 1).

표 1. 하이퍼망 학습에 사용된 파라미터

파라미터	값
샘플링 횟수	10
교체비율	10%
반복실험	10
오더	4
가중치 α	10
가중치 β	0.0001

각 부류의 miRNA 시퀀스 마다 10 번씩 샘플링을 하고, 추출된 시퀀스의 부류 정보를 추가한 10개의 하이퍼에지를 생성하였다. 오더는 4로 고정하였기 때문에 한 번 샘플링할 때 뽑히는 정점은 4개이다. 총 반복 실험은 10회 하였으며, 가중치가 낮은 하위 10%의 하이퍼에지들은 매 Epoch 마다 제거를 하고 새로 샘플링 하였다.

3.3 실험 결과 및 분석

하이퍼망의 학습 결과의 분류 성능을 비교하기 위해서 다른 몇몇 주요 알고리즘을 사용하여 비교 실험을 하였다(표 2).

표 2. 하이퍼망과 다른 기계학습 모델과의 분류 정확도 비교

기계학습 모델	성능 평균(10회)
Hypernetwork	**92.5306%**
SVM (Polynomial kernel)	**93.7423%**
Random Forest	**80.7362%**
Bayesian Network	**76.0736%**
Naïve Bayes	**93.3742%**

실험 결과 하이퍼망은 SVM과 Naïve Bayes에 비해 약간 낮은 성능을 보여주고 있지만 비교적 우수한 예측 정확도를 보여주고 있는 것을 알 수 있다.

하이퍼망은 학습을 통해 하이퍼에지의 가중치 정보를 알 수 있기 때문에 가중치가 높은 하이퍼에지에 포함되는 시퀀스 특징이 strand 선택에 중요한 역할을 한다는 사실을 알 수 있다. 표 3은 각 부류에서 가중치가 상위 5 이내의 연관된 시퀀스 특징들을 보여준다.

표 3. 가중치 상위 5개 이내의 하이퍼에지

하이퍼에지	Class
CCUU=0, CCGU=1, UAAG=0, GAGU=1	0
UCCG=1, GUUG=1, CAGC=0, UCAU=0	0
UCUC=0, CCGC=1, CCCA=1, ACUG=0	0
CGCC=1, GGUG=1, GUAA=0, UGUC=0	0
UCGA=0, UAAG=1, UUUC=0, UUGC=1	0
GAUA=0, UUGU=1, UGGA=1, AGUA=0	1
UGUC=0, GUGU=0, GGGA=1, GAGC=1	1
AUCA=1, CGAU=0, UUCU=0, AUAA=1	1
UUUC=1, AUGU=1, CGUG=0, GCGU=0	1
UAAA=1, AGAA=1, GGGA=0, CAUG=0	1

예를 들어 표 3의 첫 번째 하이퍼에지의 시퀀스특징들 중 CCGU와 GAGU는 발현되는 miRNA의 시퀀스에서 동시에 발견된 확률이 높음을 보여주며 값이 0인

CCUU와 UAAG는 발현된 miRNA에서 나타나기 힘든 시퀀스 특징이라고 볼 수 있다.

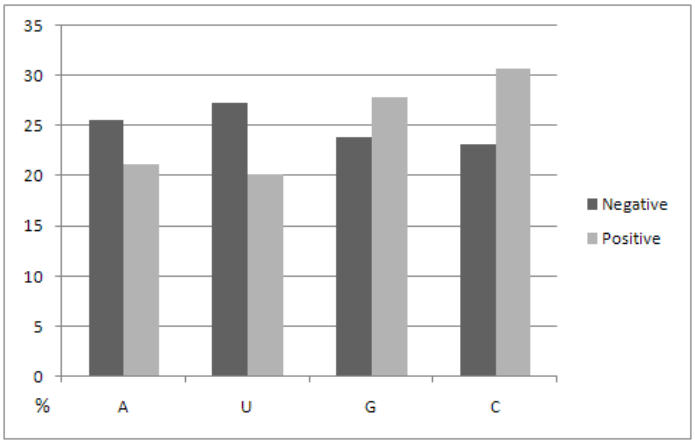

그림 3. 시퀀스 특징들의 염기 구성 비율. Positive 시퀀스 특징들의 경우 G와 C의 비율이 높은 반면, Negative 시퀀스 특징들은 A와 U의 비율이 높은 것을 보여준다.

그림 3은 가중치가 3 이상인 하이퍼에지에 포함되는 시퀀스 특징의 염기 구성 비율을 그래프로 나타낸 것이다. 단순히 염기들의 구성 비율을 본 것이지만 Positive 시퀀스 특징들에는 염기 G와 C의 비율이 비교적 상대적으로 높고 Negative 시퀀스 특징들에는 염기 A와 U의 비율이 높은 것을 알 수 있다.

그림 4는 가중치 3 이상인 하이퍼에지에서 추출된 시퀀스 특징들의 염기 개수 편향성 그래프를 보여준다. 다시 말해 각 시퀀스 특징들마다 특정한 한 염기를 몇 개 가지고 있는지를 조사한 것이다. Negative 시퀀스 특징들에서 1개의 A를 가지고 있는 시퀀스 특징은 60%라는 것이며 이하 2개, 3개, 4개를 가지고 있는 시퀀스 특징은 30%, 9%, 0.5%의 비율을 보여준다. 하지만 Positive 시퀀스의 경우에는 41%, 33%, 25%, 0%의 비율을 가지고 있어, Negative 시퀀스 특징과 비교적 다름을 알 수 있다. U와 G는 두 부류가 비교적 동일한 비율을 보여주고, C의 경우 Positive 시퀀스 특징들의 경우 C를 2개 이상 가진 비율이 비교적 많음을 알 수 있다.

4. 결론 및 논의

본 논문에서는 하이퍼망 학습 모델을 사용하여 RNA 시퀀스 정보를 통한 인간 miRNA strand 선택 예측 모델을 구축하고, 학습을 통해 찾아진 하이퍼에지의 가중치 정보를 활용해 strand 선택에 중요한 시퀀스 특징들을 찾았다. 하이퍼망은 기존의 다른 기계학습 방법과는 달리 학습 후 연관된 특징들의 정보를 분석하기가 쉽다. 예를 들어 성능 비교에 사용하였던 Bayesian 방법은 높은 확률을 갖는 시퀀스 특징을 개별적으로 확인할 수는 있지만 둘 이상 연관된 시퀀스 특징을 찾을 수는 없다. Tree 학습의 경우 시퀀스 특징

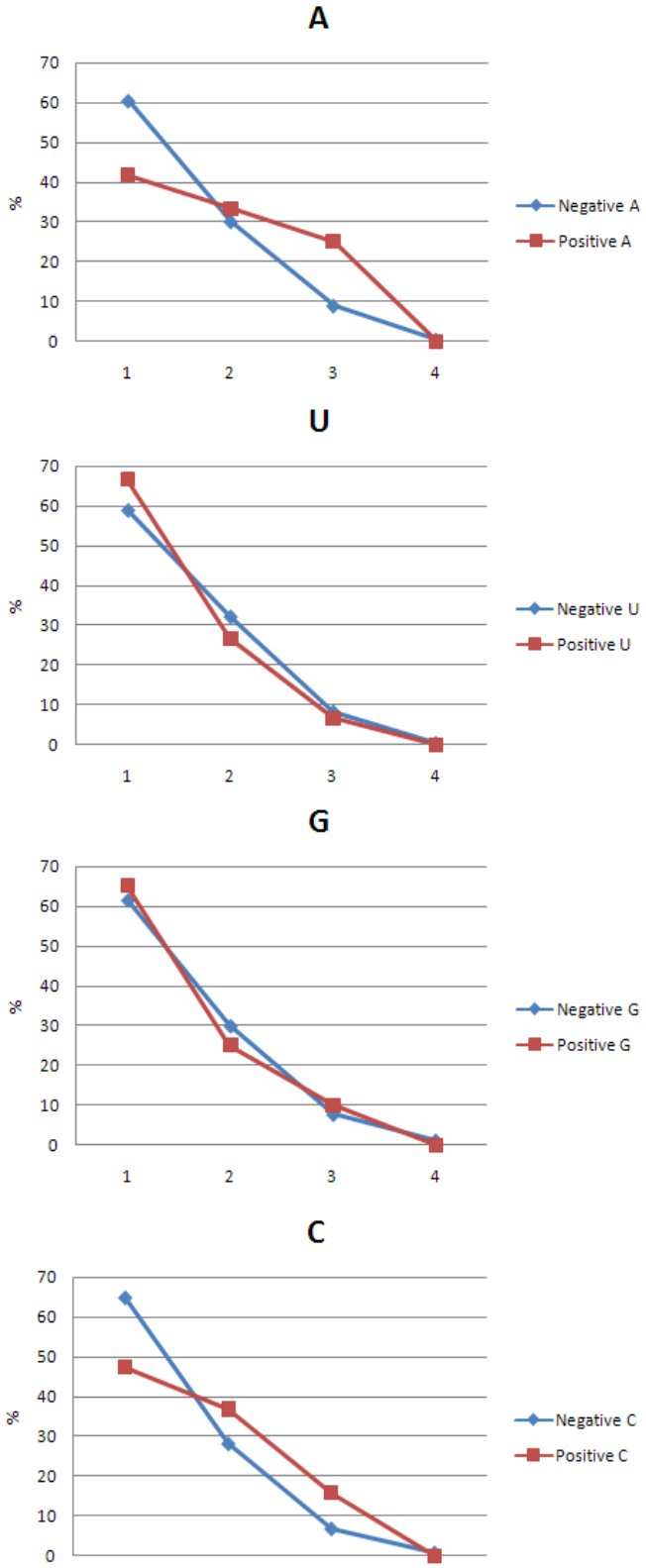

그림 4. 시퀀스 특징들의 염기 개수 편향성 그래프. U와 G의 경우 비슷한 염기 편향성을 보여주지만, A와 C의 경우 비교적 다른 편향성을 보여준다.

간의 상 하위간 정보를 얻을 수 있지만 평행적인

시퀀스 특징간의 연관정보는 얻을 수 없다. 하지만 하이퍼망은 중요한 시퀀스 특징들의 그룹으로 이루어진 하이퍼에지를 통해 miRNA strand 선택에 중요하다고 생각 되는 염기 서열을 알려준다. 추가적으로, 시퀀스 특징들에 대한 염기의 구성비와 염기 개수 편향성에 대한 분석을 통해 Positive 시퀀스와 Negative 시퀀스에 다른 특징이 있음을 확인할 수 있었다. 본 논문에서 찾아진 시퀀스 특징과 예측 모델을 사용해 앞으로 miRNA 시퀀스 strand 선택에 관한 생물학적 메커니즘을 밝히는 연구에 도움을 줄 수 있을 것이라 기대된다.

감 사 의 글

본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업의 일환으로 수행한 MARS (KI002138), 교육과학기술부 재원으로 한국연구재단의 지원(No. 2010-0017734), Xtran (No.314-2008-1-D00377), 한국장학재단의 지원(No. S2-2009-000-01116-1) 및 BK21-IT사업에 의해 일부 지원되었음.

참 고 문 헌

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

[2] Kim, V. N., Han, J. and Siomi, M. C., Biogenesis of small RNAs in animals, *Nat Rev Mol Cell Biol*, vol. 10, pp.126-139, 2009

[3] Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. and Filipowicz, W., Single processing center models for human Dicer and bacterial RNase III, *Cell*, vol. 118, pp.57-68, 2004

[4] Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. and Kim, V. N., The nuclear RNase III Drosha initiates microRNA processing, *Nature*, vol. 425, pp.415-419, 2003

[5] Khvorova, A., Reynolds, A. and Jayasena, S. D., Functional siRNAs and miRNAs exhibit strand bias, *Cell*, vol. 115, pp.209-216, 2003

[6] Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N. and Zamore, P. D., Asymmetry in the assembly of the RNAi enzyme complex, *Cell*, vol. 115, pp.199-208, 2003

[7] Farazi, T. A., Juranek, S. A. and Tuschl, T., The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members, *Development*, vol. 135, pp.1201-1214, 2008

[8] Okamura, K., Liu, N. and Lai, E. C., Distinct mechanisms for microRNA strand selection by Drosophila Argonautes, *Mol Cell*, vol. 36, pp.431-444, 2009

[9] Hu, H. Y., Yan, Z., Xu, Y., Hu, H., Menzel, C., Zhou, Y. H., Chen, W. and Khaitovich, P., Sequence features associated with microRNA strand selection in humans and flies, *BMC Genomics*, vol. 10, pp.413, 2009

[10] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, vol. 3, pp.49-63, 2008

[11] Ha, J.-W. and Zhang, B.-T., Evolutionary Hypernetwork Model for Higher Order Pattern Recognition on Real-valued Feature Data without Discretization, *Journal of KIISE : Software and Applications*, vol. 37, pp.120-128, 2010