위치 점수 행렬 혼합 모델을 이용한 microRNA 서열 특성 분석

이제근[O1,2] 장병탁[1,2,3]
서울대학교 생물정보학협동과정[1]
서울대학교 바이오정보기술 연구센터[2]
서울대학교 컴퓨터공학부[3]
{jkrhee[O], btzhang}@bi.snu.ac.kr

Analysis for microRNA sequences by the position-weight-matrix mixture model

Je-Keun Rhee[O1,2], Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics, Seoul National University[1]
Center for Bioinformation Technology, Seoul National University[2]
Graduate School of Computer Science and Engineering, Seoul National University[3]

요 약

특정한 기능을 하는 DNA 조각은 특정한 염기 서열들을 가진다. 이를 이용하여 특정 조각의 DNA 서열을 위치 점수 행렬을 이용하여 표현할 수 있다. 하지만 찾고자 하는 DNA 부분들이 완전히 밝혀진 것이 아닐 수 있다. 따라서 현재 알려진 정보만을 이용하여 위치 점수 행렬을 만들 경우, 실제 서열 패턴이 아닌, 편중된 정보가 얻어질 수 있다. 따라서 본 논문에서는 위치 점수 행렬의 혼합 모델을 이용하여, 각각의 특정 군집들을 대표할 수 있는 행렬들을 구분하여 구성하였다. 본 논문에서는 약 22개의 염기로 구성된 microRNA 서열 중, 초반부의 8개의 염기 서열정보를 이용하여, 이들 위치의 서열상의 특성을 확인해 보고자 하였다. miRNA 서열을 대표하기위한 위치 점수 행렬들은 구분하여 만들고, EM 알고리즘을 이용하여 학습한다. 학습 결과 얻어진 혼합 모델과 은닉 변수를 통해 microRNA들을 군집화하고, 각각의 군집에 속한 microRNA 서열의 특성을 확인한다.

1. 서 론

DNA는 아데닌(Adenine, A), 티민(Thymine, T), 구아닌(Guanine, G), 시토신(Cytosine, C)의 네 가지 염기로 구성된다. RNA의 경우에는 티민 대신 우라실(Uracil, U)이 들어가게 된다. 생체 내의 모든 기능은 DNA의 발현(expression)에 의해 일어나므로, 특정한 생체 내의 기능은 핵산(nucleic acid)의 염기 서열에 의해 결정되어 있다고 볼 수 있다. 따라서 특정한 기능을 하는 DNA, 혹은 RNA의 부분들이 세포 내에서 어떤 서열 조합으로 존재하는지 분석하는 작업은 매우 중요한 일이다. 인간 유전체 사업(human genome project) 등을 통해 유전체에 전체 대한 염기 서열이 거의 밝혀지기는 하였으나, 전체 유전체에서 특정한 기능을 하는 부분을 찾고, 이에 대한 염기 서열상의 특성을 확인하고 분석하는 작업은 계속 진행되고 있다. 본 논문에서는 현재까지 밝혀진 인간의 microRNA (miRNA) 서열을 이용하여 서열의 특성을 확인해보고자 하였다.

miRNA는 약 22개의 염기로 구성되어 있는 서열로 RNA가 단백질(protein)으로 발현되는 과정을 조절하는 물질로 알려져있다[1]. miRNA는 현재 인간의 경우 300개 이상이 알려져 있으나, 실제로는 약 1000개 가까운 수로 존재할 것으로 예측되고 있다[2, 3]. 따라서 현재 알려져 있는 miRNA의 수는, 예측되는 것의 절반에도 못 미치는 것으로, 이 정보만을 이용하여 miRNA 서열의 특성을 단정지어 이야기할 수는 없다. 또한 miRNA들이 각기 다른 유전자의 발현을 조절할 수 있으므로, 이들의 특성을 하나로 통합하여 분석하는 것 역시 좋은 방법은 아닐 것이다. 따라서 본 논문에서는 위치 점수 행렬(position weight matrix, PWM)의 혼합 모델(mixture model)을 이용하여, miRNA의 서열상의 특성을 대표하기 위한, 각 군집(cluster)별로 각기 다른 위치 점수 행렬을 구성한다. 위치 점수 행렬은 일반적으로 유전자 전사 조절 인자(transcription factor)들의 DNA 상의 결합 위치(binding site)를 표현하는 데에 많이 사용되고 있다[4]. 하지만 miRNA의 서열 정보 역시 위치 점수 행렬을 이용하여 표현 가능하다.

본 논문에서는 서열상의 공통적인 특성을 확인하기 위해 각 위치 점수 행렬들에 가중치(weight)를 부여한 선형 결합(linear combination)과, 특정 서열에 대한 표현 정도를 설명하기 위한 은닉 변수(hidden variable)를 사용한다. EM (expectation-maximization) 알고리즘을 통해 파라미터(parameter)들을 학습하고, 그 결과 얻어진 은닉 변수의 값을 이용하여 miRNA 서열의 군집화(clustering)가 가능하다. 각 군집들의 서열을 통해, miRNA 서열의 특성들을 확인해본다.

2. 실험 방법

2-1. 위치 점수 행렬

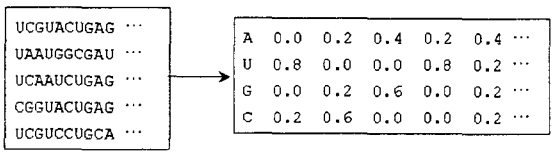

그림 1 위치 점수 행렬의 구성 예

위치 점수 행렬은 4 x L 크기의 행렬로 표현된다. L은 서열의 길이를 의미하는 것으로, 위치 점수 행렬의 각 열은 위치를 나타내며, 각 행은 차례로 특정 위치에서의 A, T, G, C의 빈도수를 이용한 발생 확률 값으로 표시된다. 그림 1은 위치 점수 행렬의 구성 예를 보여준다.

2-2. 위치 점수 행렬의 혼합 모델과 EM 알고리즘을 이용한 학습

본 논문에서 사용한 위치 점수 행렬의 혼합 모델 구성과 이 모델에 대한 학습 과정은, Hannenhalli와 Wang이 전자 조절 인자들의 결합 위치에 대한 연구에서 이용한 방법과 유사하다[5].

만일 j번째 위치 점수 행렬을 M_j, 각 위치 점수 행렬에 대한 가중치는 λ_j라고 할 때, m개의 위치 점수 행렬에 대한 선형 결합 형태의 혼합 모델은 식 (1)과 같이 표현되며, 각 가중치 λ_j의 총합은 식 (2)에서와 같이 1이 된다.

$$P(\mathbf{x}|\mathbf{M}) = \sum_{j=1}^{m} \lambda_j P(\mathbf{x}|\lambda_j, M_j) \quad (1)$$

$$\sum_{j=1}^{m} \lambda_j = 1 \quad (2)$$

식 (1)에서 $\mathbf{x}$는 전체 서열 정보를 의미한다. 혼합 모델에서 각 파라미터들의 값은 EM 알고리즘을 이용하여 학습된다. 즉, EM 알고리즘을 이용하여 위치 점수 행렬과 각각의 가중치에 대한 추정값을 계산하는 것이다. 이 과정에서 은닉 변수 z_{ij}로 이루어진 행렬 $\mathbf{Z}$를 사용한다. 은닉 변수 z_{ij}는 i번째 서열 x_i가 위치 점수 행렬 M_j에 의해 설명되는 정도를 표현하기 위한 변수이다. 즉, z_{ij}의 값이 1인 경우 서열 x_i가 위치 점수 행렬 M_j에 의해 완전히 설명된다는 뜻이며, z_{ij}의 값이 0인 경우는 이와 반대로 M_j에 의해서는 서열 x_i가 설명되지 않는다는 의미가 된다.

전체 n개의 서열 데이터에 대한 우도(likelihood)값은 식 (3)에 의해서 계산된다.

$$L(\mathbf{M}, \boldsymbol{\lambda}|\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{m} E(z_{ij}) \times \log(P(x_i|M_j)\lambda_j) \quad (3)$$

EM 알고리즘의 E 단계에서는 식 (3)의 우도 함수에 대한 기대값을 최대화하기 위하여, 은닉 변수 z_{ij}에 대한 기대값을 식 (4)를 이용하여 계산한다.

그림 2 miRNA 서열에 대한 웹로고

$$E(z_{ij}) = P(z_{ij}=1|x_i, M_j, \lambda_j) \quad (4)$$

$$= \frac{P(x_i|M_j)\lambda_j}{\sum_{l=1}^{m} P(x_i|M_l)\lambda_l}$$

M 단계에서는 식 (4)를 이용하여 E 단계에서 계산된 z_{ij}의 기대값을 이용하여 위치 점수 행렬 M과 가중치 λ에 대한 값을 추정하게 된다. 식 (5)는 가중치 λ에 대한 추정값을 계산하기 위한 식이며, 식 (6)은 위치 점수 행렬에 대한 추정값을 계산한다.

$$\hat{\lambda_j} = \frac{1}{n}\sum_{i=1}^{n} E(z_{ij}) \quad (5)$$

$$\widehat{M_j}[u,v] = \frac{\sum_{i=1}^{n} M_j(x_{iv}=u)E(z_{ij})}{\sum_{b\in A,U,G,C}(\sum_{i=1}^{n} M_j(x_{iv}=b)E(z_{ij}))} \quad (6)$$

$\widehat{M_j}[u,v]$는 j번째 위치 점수 행렬에서 위치 v에서 염기 u를 가질 확률값을 의미한다. x_{iv}는 i번째 서열의 v번째 위치를 뜨미하는 것으로 $M_j(x_{iv}=u)$는 서열의 v번째 특정 위치에서 염기 u가 나타날 확률값을 위치 점수행렬 M_j에서 얻은 값이다. 하지만 계산 과정에서 $E(z_{ij})$의 값이 0이 나오는 경우 분모가 0이 될 수 있다. 따라서 실제 실험에서는 식 (6)의 분모와 분자에 각각 10^{-19}의 작은 값을 더해주어 계산 과정상에서 에러가 발생하지 않도록 하였다.

최종 학습이 끝난 후 결정된 i번째 서열 x_i의 군집 $C(x_i)$는 식 (7)에서와 같이 은닉 변수 z_{ij}의 값에 의해 결정된다.

$$C(x_i) = \mathrm{argmax}_j(z_{ij}) \quad (7)$$

3. 실험 결과

본 논문에서는 m값을 5로 하여 5개의 위치 점수 행렬과 이에 대한 가중치 값을 이용하여 실험하였다. 가중치 λ_j의 초기값은 0과 1 사이의 값으로 랜덤으로 생성되었다. 또한 위치 점수 행렬의 초기값은 인간 염색체 19번에서 크기가 8인 서열을 임의로 300개씩 추출하여 이에 대한 위치 점수 행렬 만들어 사용하였다. 위치 점수 행렬의 초기값의 생성에 이용된 인간 염색체 서열은 UCSC의 Genome browser (http://genome.ucsc.edu/)

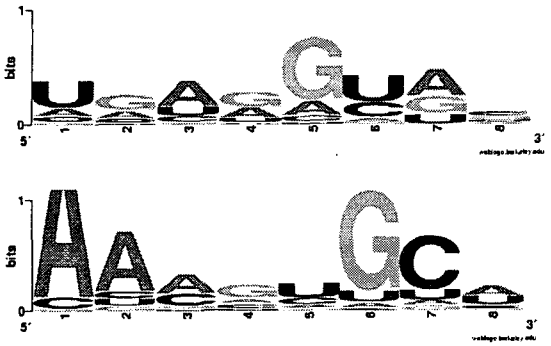

그림 3 군집 0과 군집 3에 대한 웹 로고

에서 받아 사용하였다.

miRNA들은 길이가 18~25로 다양한 크기로 존재하나, 본 실험에서는 1번부터 8번 위치의 서열만을 사용하였다. miRNA의 2~7번 위치는 seed라 불리는 부분으로, 목표 유전자(target gene) 등을 결정하는 데에 중요한 서열이므로, 본 실험에서는 seed 부위의 공통적인 특성을 확인하고자 하였다. miRNA 서열은 Sanger institute의 miRBase (http://microrna.sanger.ac.uk/sequences/)에서 받아 이용하였다.

그림 2는 miRNA 서열 전체에서 1번에서 8번 위치까지의 웹로고(WebLogo)를 보여준다[6]. 전체 miRNA를 이용하는 경우 miRNA에서 서열상의 특성을 확인하기 어렵다. 그림 3은 본 논문에서 설명한 혼합 모델을 이용하여 학습된 결과를 이용하여 만들어진 결과를 이용하여 만든 웹로고이다. 총 5개의 군집 중 대표적으로 그림 3에서는 군집 0번과 군집 3번에 대한 로고만을 보인다. 그림 2에서와는 달리 각 위치마다 특정한 서열이 많이 나타나고 있음을 알 수 있다. 또한 각 군집마다도 서로 구분된 특징이 보여짐을 알 수 있다. 그림 3의 예에서와 같이 군집 0번의 1번 위치는 U가 많은 수로 존재하고 5번 위치는 G가 많이 존재하고 있으나, 군집 3에서는 1번 위치에서는 A가 다른 염기에 비해 월등히 많이 존재하며, 5번 위치에서 G는 거의 나타나지 않음을 알 수 있다. 또한 군집 3의 6번 위치는 G의 염기가 특이적으로 많이 나타나고 있다. 이와 같이 각 군집들은 전혀 다른 형태의 서열 특성이 나타남을 확인할 수 있다.

표 1 각 군집별 miRNA 수와 해밍 거리에 대한 p-value

	miRNA 수	p-value
cluster 0	63	1.00E-141
cluster 1	84	2.39E-142
cluster 2	78	1.52E-86
cluster 3	43	7.04E-158
cluster 4	60	5.86E-153

각 군집에 대해 평가하기 위해, 각 군집에 있는 서열 데이터의 모든 쌍에 대해서 해밍 거리(Hamming distance)를 계산하고, 이 결과를 기존의 전체 miRNA 서열 전체에 대한 결과와 비교 검정하였다. 해밍 거리는 각 위치에서의 값이 같으면 0, 다르면 1로 둔다. 두 서열이 서로 유사할수록 0에 가까운 정수 값이 거리가 되고, 길이가 8인 두 서열이 서로 전혀 다른 서열인 경우 길이가 8의 값을 가지게 된다. 전체 miRNA에 대한 해밍 거리와 각 군집 내의 miRNA들의 해밍 거리의 두 샘플에 대한 검정은 t-검정(t-test)을 이용하였고, 그 결과는 표 1에 보인다. 각 군집 모두 매우 낮은 p-value를 보이는 것을 알 수 있으며, 서로 유사한 서열끼리 군집화되어 있다는 사실을 확인할 수 있다.

4. 결 론

본 논문에서는 EM 알고리즘을 통해 혼합 모델을 학습하고, 군집화 결과를 이용하여 miRNA 서열의 특성을 확인해보았다. 현재까지 알려져 있는 정보를 그대로 이용할 경우, 서열상의 공통적인 특성을 확인하기 힘들다. 하지만 이를 군집화하여 서열을 비교할 경우, 각 군집들 내에서는 서열들이 공통적인 특성을 가지고 있음을 확인할 수 있다. 특히 본 논문에서 분석한 miRNA의 1~8번 위치 중 2~7번 위치는 seed 부분으로 miRNA의 목표 유전자의 결정에 매우 중요하게 작용하는 서열이다. 따라서 이 서열의 특성을 확인하는 것은 miRNA의 목표 유전자 예측 및 기능 예측 등에도 유용한 정보로 사용될 수 있다.

향후 군집의 수를 변화시켜가면서, 보다 명확한 miRNA 서열상의 특성을 찾는 연구를 진행할 필요가 있다. 그리고 학습 결과 얻어진 각 군집에 속한 miRNA들 간의 위치 정보 및 기능 등에 대한 상호 관계에 대한 연구도 필요하다. 또한 miRNA 서열상의 특성에 따른 목표 유전자들과의 연관 관계를 분석하는 연구 역시 중요한 일이 될 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실사업(NRL)에 의해 지원되었음.

참고 문헌

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

[2] Berezikov E., *et al.*, Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120, 21-24, 2005.

[3] Bentwich I., *et al.*, Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, 37, 766-770, 2005.

[4] Stormo GD., DNA binding sites: representation and discovery. *Bioinformatics*, 16, 16-23, 2000.

[5] Hannenhalli S. and Wang LS., Enhanced position weight using mixture models, *Bioinformatics*, 21, Suppl. 1, i204-i212, 2005.

[6] Crooks GE, *et al.*, WebLogo: A sequence logo generator, *Genome Research*, 14, 1188-1190, 2004.