DNA 커널을 이용한 MicroRNA 목표 유전자 예측

노영균 [01] 김성규 [2] 김청택 [3] 장병탁[4]
서울대학교 인지과학 협동과정[1], 서울대학교 생물정보학 협동과정 [2]
서울대학교 심리학과[3], 서울대학교 컴퓨터공학부[4]
yknoh@bi.snu.ac.kr, skkim@bi.snu.ac.kr,
ctkim@snu.ac.kr, btzhang@bi.snu.ac.kr,

MicroRNA Target Prediction using DNA Kernels

Yung-Kyun Noh [01] Sung-Kyu Kim[2] Cheong-Tag Kim[3] Byoung-Tak Zhang[4]
Interdisciplinary Program in Cognitive Science[1], Graduate Program in Bioinformatics[2],
Department of Psychology[3], School of Computer Science and Engineering[4],
Seoul National University, Seoul 151-742, Korea

요 약

분류 방법으로서의 SVM(Support Vector Machine)은 커널 방법과 함께 사용됨으로써 그 유용성을 크게 향상시켰다. 커널 방법은 일반적으로 입력 데이터의 자질(feature)로 나타내는 공간으로부터 높은 차원의 공간으로 데이터를 사상(mapping)시키는 역할을 하게 되나, 기본적으로는 데이터간에 새로운 거리(metric)를 부여해주는 역할을 하는 것이다. 지금까지 나온 다양한 커널 방법은 구조화된(structured) 데이터에 대해 커널 형태로 거리를 부여하는 방법을 제시한다. 본 논문에서는 DNA의 작용을 모델링하여 만든 새로운 커널이 miRNA(micro RNA)와 mRNA(messenger RNA)쌍에 대한 발현 여부를 분류해 내기 위해 커널 형식으로 거리를 부여하는 방법을 보인다. 이 방법은 실리콘 컴퓨터가 아닌 실제 DNA분자로 실험할 수 있도록 설계된 것을 고려할 때 여러 종류의 DNA 코드를 분석하는 데 사용될 수 있는 새로운 분자컴퓨팅 방법이다.

1. 서 론

SVM(Support Vector Machine)이 최근 몇 년 사이에 큰 성공을 거둔 이유는 커널 방법의 장점을 나타내기에 적당한 구조를 지녔기 때문이다. 커널 방법은 데이터를 높은 차원의 공간에 사상(mapping)시켜 선형 분류방법의 VC차원을 크게 해 주는데 사용할 수 있을 뿐만 아니라 문자열이나 그래프 모델과 같은 구조화된 데이터에 대해서 적절한 커널을 적용시켜 줌으로써 데이터간의 거리(metric)을 부여시켜 줄 수 있다. 일반적으로 기계학습의 많은 문제들을 실제 데이터의 공간상의 위치가 아닌 데이터간의 내적값만을 가지고 풀 수 있는 문제로 환원시킬 수 있는데, 이것은 사상을 위한 함수가 명시적으로 정의되지 않더라도 데이터로부터 일정 조건을 만족하는 커널 행렬만 정의할 수 있는 경우 커널 행렬을 만드는 방법이 무엇이건 간에 기계학습 문제를 풀 수 있다는 것을 의미한다.

한편, 특성공간(feature space)의 쌍대공간(dual space)인 데이터 공간(data space)에서의 작업은 데이터의 양이 많은 경우 많은 계산량을 필요로 하고, 데이터x데이터의 크기를 갖는 커다란 커널 행렬을 유지해야 하는 부담을 갖는다. 이러한 가운데 최근 연구에서는 SVM의 경우를 직렬기계인 컴퓨터로 모든 계산을 순차적으로 해 내는 대신 칩 위의 아날로그 전자회로를 통해 한번에 실시간으로 계산해 내려는 시도가 있었다.[6]

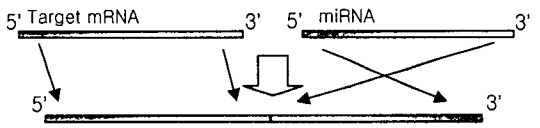
기계학습에 분자컴퓨팅을 응용한 연구로는 최근 RNA의 발현량을 통제함으로써 다양한 회로망을 구현할 수 있다는 연구결과가 있다[4]. 이 연구는 생화학적인 반응을 이용하여 신경망과 유사한 기능을 in vitro로 실현할 수 있는 가능성을 시사하는 점에서 의의가 있으나 작은 수의 노드로 구성된 망의 구성을 시도하고 있고 학습의 방법은 포함하고 있지 않다. 한편 본 연구진에서는 최근 DNA의 병렬적 혼성화 반응을 이용하여 양한정(positive definiteness) 조건을 만족하는 커널을 만들 수 있다는 것을 모사(simulation)하여 보여주었다[1]. 또한 이와 같은 방법이 실제 DNA의 hybridization 결과와 잘 일치함을 theorem proving 문제에 적용시킨 시뮬레이션 실험 결과를 통해 얻었다[2].

본 논문에서는 DNA 컴퓨팅 기반 커널 머신이 보다 실제적인 생물학 문제 해결에 적용될 수 있는지를 검토한다.

2. microRNA 목표 유전자 예측

DNA 커널이 직접적으로 사용되기 쉬운 분야는 생물학(biology)의 문제 중 DNA 순차열(sequence)의 분석을 통한 분류 문제이다. MicroRNA (miRNA)는 약 22 뉴클레오타이드(nucleotide) 정도의 작은 크기를 갖는 RNA 분자로서 동식물의 유전자 발현 과정에서 messenger RNA (mRNA)와 느슨한 상보적 결합을 함으로써 발현 억제 기능을 한다. 이 때, mRNA와 miRNA의 쌍이 가지는 순차열이 중요한 역할을 하는데, 본 연구진의 다른 연구는 miRNA의 mRNA 발현 억제 여부를 이들 RNA쌍이 가지는 영기 순차열에서 얻은 여러 가지 자질(feature)들을 이용하여 예측할 수 있음을 보였다.[3]

MicroRNA의 기능 여부는 miRNA와 mRNA쌍의 영기서열에 의해 결정된다. 실험에 필요한 miRNA:mRNA 쌍 서열 데이터는 문헌으로부터 직접 얻었으며 생물학 실험으로 검증된 것만을 사용하였다. DNA커널을 이용하기 위한 데이터를 생성하기 위해서 miRNA와 mRNA의 순차열(sequence)을 붙여서 하나의 순차열를 구성하게 되는데, 이 때 서로 상보적인 miRNA와 mRNA의 부분적인 영역이 의도하지 않게 서로 붙는 것을 방지하기 위해 miRNA 서열의 5'과 3'방향을 바꾸어 코딩한다. 이렇게 만들어진 순차열로 miRNA가 발현 억제기능을 하는지 분류한다.



<그림 1> 순서열 데이터를 만드는 방법.

이와 같이 구성한, 길이가 46mer(24mer의 mRNA, 22mer의 miRNA)로 같은 20개 데이터를 x_i로 표시하고, 각각의 x_i에 해당하는 y_i가 데이터의 억제 여부를 나타내는 값을 가지게 함으로써 학습을 위한 데이터 set

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{20}$$

가 만들어지게 된다. 이렇게 만들어진 데이터 set을 가지고 leave-one-out으로 20번의 학습과 테스트를 여러 온도에 대해서 3번씩 해 평균을 취해 보았다.

본 연구에서는 mRNA와 miRNA쌍의 데이터를 가지고 p-spectrum 커널을 통해 실험한 결과와 hybridization을 통해 형성된 커널을 통한 방법의 결과를 비교한다. 실험 결과를 통해 hybridization을 통해 새로 만들어진 커널이 적절히 데이터들간에 거리(metric)를 부여하는 커널이 됨을 보여준다.

3. DNA 커널

Hybridization을 통한 커널은 DNA의 작용을 모델로 했으며 DNA 순차열(sequence)에 적용시키기 가장 좋은 형태를 지니고 있으므로 DNA 커널이라고 부른다.

DNA커널의 구현은 DNA 순차열 데이터들에 대해 각각의 상보적(complementary) 순차열 데이터들을 생성시켜 전체를 hybridization시킴으로써 얻을 수 있다. 각 커널의 요소값 K_{ij}는 i번째 순차열과 j번째 상보적 순차열 데이터가 결합된 double strand의 개수를 통해 계산되며, 이 때 온도의 조절을 통해 simulated annealing을 시킴으로써 양한정 조건을 만족시키는 커널을 얻을 수 있게 된다.[1]

$$K_{ij} = |\{(dsDNA) : dsDNA(1) = DNA(i),\ dsDNA(2) = complementaryDNA(j)\}|$$

DNA커널의 커널값을 얻는 과정을 정리하면 다음과 같다.

1. 각각의 데이터에 해당되는 DNA와 그에 상보적인 DNA를 구성한다. 유사한 데이터일수록 영기배열이 비슷하다는 것을 가정한다.
2. 만들어진 DNA 모음을 시험관에서 섞고, 온도를 조절해서 DNA 가닥들이 서로 붙게 한다.
3. i번째 데이터에 해당되는 DNA 가닥과 j번째 데이터에 해당되는 상보적 DNA 가닥이 붙은 dsDNA의 양에 비례해 커널 행렬의 성분 K_{ij}를 구성한다.

이와 비교해 p-spectrum 커널은 두 개 string간의 유사도를 두 string이 공통으로 가지는 길이 p의 substring의 개수로부터 다음과 같이 정의한다.

$$K_{ij} = \sum_{u \in \Sigma^p} \phi_u^p(i)\phi_u^p(j)$$

단, $\phi_u^p(i) = |\{(v_1, v_2) : s = v_1 u v_2\}|, \quad u \in \Sigma^p$

SVM 분류기는 이렇게 얻어진 커널을 가지고 자질(feature)공간에서 데이터 set간 margin이 최대가 되도록 최적화된 파라미터 α들을 다음 식으로부터 구한다.

$$\vec{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} \alpha_i - 1/2 \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij}$$

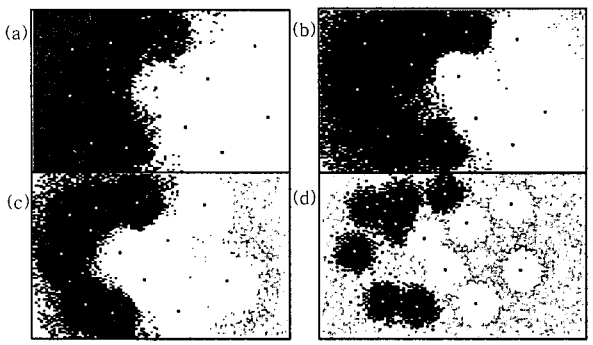
단, $\alpha_i \geq 0, i = 1,2,...,n, \sum_{i=1}^{n} \alpha_i y_i = 0, \quad y_i \in \{+1, -1\}$

이렇게 최적화된 파라미터를 가지고 테스트 데이터에 대해 다음의 연산을 함으로써 테스트 데이터에 대한 결과값을 얻을 수 있다.

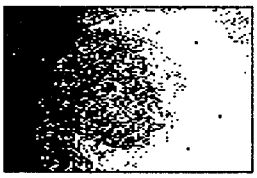
$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K_i(x)) \quad 단, f(x) \in \{+1, -1\}$$

DNA 커널을 통해 입력공간에서 선형적으로 분류되지 않는 학습 데이터가 있는 공간을 여러 가지 패턴으로 나눌 수 있게 되는데, <그림2>는 2차원 평면에서의 거리를 결합 에너지로 가정하여 데이터 공간이 분류 되는 것을 모사(simulation)한 예이다. <그림2>의 네 가지 경우는 모두 적절한 온도로 simulated annealing을 해주어 양한정 조건을 만족시키는 커널을 제작한 예인데, 이를 통해 simulated annealing온도를 조절해주는 것이 데이터 분류에 있어서 regularization 정도를 조절하는 역할을 함을 알 수 있다.

<그림3>는 simulated annealing을 통하지 않아 결과적으로 양한정 조건을 만족시키지 않은 커널을 사용했을 때, 데이터 공간을 잘 분류해 내지 못하는 것을 보여준다.



<그림 2>DNA커널을 통한 데이터의 분류. Simulated annealing 온도는 각각 (a)73℃→40℃, (b)73℃→60℃, (c)73℃→67℃, (d)74℃.

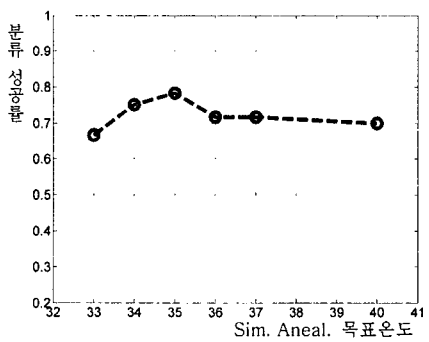


<그림 3> simulated annealing을 하지 않은 경우. 데이터 공간을 정확히 나누지 못함을 알 수 있다. hybridization온도는 40℃.

이 연구에서는 이러한 DNA커널을 이용하여 miRNA분자가 mRNA의 발현을 억제할지 여부를 예측할 수 있는지 분석하여 본다.

4. 실험 결과

다음의 그래프는 simulated annealing온도를 달리한 DNA 커널을 이용한 SVM분류의 결과를 보여준다.



<그림 4> simulated annealing의 목표온도를 달리한 DNA 커널을 이용한 SVM분류성능이다. DNA 커널의 성능이 p-spectrum 커널의 성능을 거의 따라간다.

Simulated annealing의 목표온도에 따라 regularization 정도가 달라지면서 분류 성공률에 차이가 남을 알 수 있다.

각각 p값을 달리한 p-spectrum 커널을 이용한 SVM분류 결과(분류 성공률)는 최대 .85로서 DNA커널의 결과가 p-spectrum 커널의 성능에 미치지는 못하는 것으로 나타났다. 하지만, 최대 .85를 제외하고 대부분 .75이상의 성공률을 보이지 못하는 것으로 나타나 DNA커널의 결과가 어느 정도 이상 p-spectrum 커널의 결과와 비교해 좋은 성능을 보임을 알 수 있다.

5. 결론 및 토의

모사를 통한 DNA커널이 p-spectrum커널의 성능을 능가하지 못하는 이유는 사용하는 분자의 수가 실제 분자의 수를 다 모사하지 못하고, 난수의 사용에 의한 잡음이 에러로 작용하기 때문이다.

하지만 DNA hybridization을 모사하여 커널을 생성하는 방법이 miRNA의 분류에 simulated annealing 온도에 따라 다양한 거리(metric)를 부여해 줄 수 있으며, 이 때문에 분류 성능이 차이가 나는 것을 보았다.

이 방법은 일반적으로 DNA 염기 서열 분석에 많이 쓰이는 hamming 거리를 통한 방법보다 더 풍부하고 자유롭게 거리를 부여할 수 있는 방법이다. 또한 이 방법은 실제 DNA를 통해 병렬성을 지닌 분자 컴퓨팅 기법으로 사용될 수 있다는 점에서 많은 가능성을 지니고 있다.

감사의 글

본 연구는 과학기술부 국가지정연구실사업, 산자부 차세대 연구개발사업, 그리고 산자부 뇌신경정보학 사업(김청택)에 의하여 지원 되었음.

참고문헌

[1] 노영균, 김청택, 장병탁. (2005) 개념학습을 위한 DNA컴퓨팅 기반 커널의 설계, *한국인지과학회 춘계학술대회*, 2005.

[2] 김준식, 김종찬, 노영균, 이동윤, 장병탁. (2005) DNA 컴퓨팅연산 과정의 통계 물리적 예측, *한국정보과학회 춘계학술대회 2005.*

[3] 김성규, 장병탁. (2005) SVM과 위치 기반의 자질을 이용한 MicroRNA 목표 유전자 예측, *한국정보과학회 춘계학술대회 2005.*

[4] J. Kim, J.J. Hopfield, & E. Winfree. (2004) Neural network computation by in vitro transcriptional circuits, *Advances in Neural Information Processing Systems 2004.*

[5] L.M. Adleman. (1994) Molecular Computation of Solutions To Combinatorial Problem. *Science, 266*, 1021-1024.

[6] S. Chakrabartty & G. Cauwenberghs. (2005) Sub-Microwatt Analog VLSI Support Vector Machine for Pattern Classification and Sequence Estimation, *Advances in Neural Information Processing Systems 2004.*

[7] Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, and H. Honda, (2004) Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma, *BMC Bioinformatics 2004, 5:120.*

[8] Schoelkopf, B., & Smola, A. (2001) *Learning with Kernels*, MIT Press.

[9] John S. & Nello C. (2004) *Kernel Methods for Pattern Analysis*, Cambridge University Press.